Unsupervised Data Mining Applications on High Dimensional Gene Expression Time Series in Toxicogenomics

A Dissertation Presented

By

Ce Gao

to

The Department of Bioengineering

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the field of

Bioinformatics

Northeastern University
Boston, Massachusetts

August, 2015
ABSTRACT

Toxicogenomics, the study of adverse effects caused by toxicants to human health and environment via high-throughput genomics technologies, present promising alternatives to expensive and lengthy animal-based approaches in toxicity testing and risk assessment. Advances in toxicogenomics techniques now enable monitoring cellular activities continuously, under a large range of experimental and biological conditions and providing comprehensive and high-resolution information at molecular levels. The research interests in toxicogenomics center on key issues involving the quantification of molecular toxic effects, the linkage between molecular endpoints and phenotypic ones, the discernment of dose-response and pharmacokinetics relationships, as well as the integration of bioinformatics into predictive toxicology. In particular, the increasingly complex and voluminous toxicogenomics data pose great analytical challenges. The existing bioinformatics tools are incompatible with the high dimensionality and temporal dynamics of the data, possibly leading to unreliable and misinterpretation of the potential toxicity connotation. The objectives of this dissertation are to develop and demonstrate new or improved methodology that better address the challenges and limitations in high dimensional time series toxicogenomics data analysis for critical bioinformatics application such as toxicity mechanism identification, toxicants classification, and for predictive toxicology knowledge discovery. In this study, we develop new or improve bioinformatics data analysis algorithms so that they are capable of processing high dimensional time series toxicogenomics data, therefore better capture and reflect the dynamics of cellular response to toxicants. We also prove the potential and validity of the incorporation of various molecular disturbance/effect quantifiers into various functional toxicogenomics bioinformatics to provide quantitative insights into the toxicant-induced cellular molecular responses at individual gene, specific pathway and system levels. In addition, we demonstrate the effectiveness of unsupervised bioinformatics tools for mining new, more in depth, much-detailed and fundamental knowledge and understanding of toxicological information at molecular level. This research could generate new information to fill in the urgent knowledge gaps in toxicogenomics that present barriers to the realization of predictive toxicology and make contributions to several fields including toxicology, bioinformatics and environmental science.
ACKNOWLEDGEMENT

First, I would like to express my greatest gratitude to my advisor, Professor April Z. Gu, for her guidance in the past seven years.

I would also thank my committee members for their directions.

I am extremely grateful for Professor David Weisman for his directions in the beginning of my research. His generous support and encouragement helped me go through the darkest time. His demonstration of scientific curiosity and constant pursuit of new knowledge have set a great example for me and my future career.

I thank all my past collaborators, including Professor Vanni Bucci, Professor Jennifer Dy and her students, Professor Feng Long, Professor Qiang Cai, Professor Miao He for their help in the past. In addition, thank my current and former colleagues with the group, including Dr. Carla Cherchi, Na Gou, Dr. Jiaqi Lan, Nimet Yildirim, Man Hu, Nick Tooker for all the joyful time we spent together.

I thank my parents for their emotional and financial supports in the past few years, without them I could not survive graduate school.

My work was supported by the National Science Foundation (NSF, EEC-0926284, and CAREER CBET-0953633 and CBET-1440764), NIEHS PROTECT P42ES017198 and CDM/Diane and Bill Howard Scholarship.
# TABLE OF CONTENT

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

TABLE OF CONTENT ........................................................................................ iv

LIST OF FIGURES .............................................................................................. viii

LIST OF TABLES .................................................................................................. xii

Chapter 1 Introduction and Objectives ............................................................... 13
  1.1. Toxicogenomics: Progresses and Challenges ............................................. 14
    1.1.1. Toxicogenomics: Paradigm Shift in Toxicology .................................. 14
    1.1.2. Challenges in Toxicogenomics ............................................................ 15
    1.1.3. Bioinformatics Applications in Toxicogenomics ............................... 20
    1.1.4. Challenges in High Dimensional Time Series Toxicogenomics Data Analysis ........................................................................................................ 24
  1.2. Research Overview .................................................................................... 27
    1.2.1. Research Objectives and Overview ..................................................... 27
    1.2.2. Organization of the Dissertation ......................................................... 33

Chapter 2 Literature Overview: Trends and Challenges in High Dimensional Time Series Toxicogenomics Data Analysis .......................................................... 36
  2.1. Common Data Mining Algorithms in Toxicogenomics ............................... 37
    2.1.1. Toxicants Clustering ........................................................................... 37
    2.1.2. Pathway Analysis .............................................................................. 39
    2.1.3. Tensor Decomposition ...................................................................... 40
    2.1.4. Gene Network Analysis .................................................................... 42
  2.2. Bioinformatics Software Development .................................................... 47
    2.2.1. Software for Data Normalization and Processing ............................... 47
    2.2.2. Software for Experimental Design and Data Analysis ...................... 48
  2.3. Conclusion ............................................................................................... 48
Chapter 3 TimeVis: Software for Whole Cell Array Data Management, Analysis and Visualization

3.1. Introduction ............................................................................................................. 50
3.2. Materials and Methods ....................................................................................... 52
   3.2.1. Data Model and Database ................................................................. 52
   3.2.2. User Interface and Visualization ......................................................... 52
   3.2.3. Data Analysis Capability ..................................................................... 53
3.3. Conclusion .......................................................................................................... 54

Chapter 4 Analyzing High dimensional Toxicogenomic Data Using Consensus Clustering

4.1. Introduction .......................................................................................................... 55
4.2. Materials and Methods ....................................................................................... 58
   4.2.1. Data Generation ............................................................................... 58
   4.2.2. Data Preprocessing ........................................................................ 60
   4.2.3. Ensemble Clustering and Data Validation: Consensus Clustering (CC) .... 61
4.3. Results and Discussion ....................................................................................... 63
   4.3.1. Consensus Clustering Based on SOM ............................................. 63
   4.3.2. The Impact of High Dimensional Nature of the Data: the Resolution Power from Time-Series ........................................................................... 68
   4.3.3. Consistency Test for Consensus Clustering .................................... 71
   4.3.4. Impact of Gene Selection: Consensus Clustering Using Reordered or Subset of Total Genes ................................................................. 73
4.4. Conclusion .......................................................................................................... 75

Chapter 5 Toxicity Mechanisms Identification via Gene Set Enrichment Analysis of Time-Series Toxicogenomics Data: Impact of Time and Concentration

5.1. Introduction .......................................................................................................... 76
5.2. Materials and Methods ....................................................................................... 79
   5.2.1. Toxicogenomics Time Series Data Generation ................................ 79
   5.2.2. Data Pre-processing .......................................................................... 80
   5.2.3. Gene Set Enrichment Analysis with Two Different Ranking Metric ........ 81
   5.2.4. Software ............................................................................................. 84
5.3. Results and Discussion ....................................................................................... 84
   5.3.1. Chemical-Specific and Temporally dynamic Stress Response Gene Expression Profiles ............................................................. 84
   5.3.2. Impact of Time on GSEA for Toxicity Mechanisms Identification .... 90
5.3.3. GESA Analysis of Time Series Toxicogenomics Data-Comparison of Two Score Metrics ................................................................. 91

5.3.4. Impact of Concentrations on Toxicity Mechanisms Identification .................. 95

5.4. Conclusion .............................................................................. 99

Chapter 6 High dimensional Toxicogenomics Data Analysis Using Tensor Decomposition Model for Toxicity Identification and Characterization ................. 100

6.1. Introduction ........................................................................... 101

6.2. Materials and methods ............................................................ 102

6.2.1. Data Generation and Processing ............................................ 102

6.2.2. Tensor Decomposition Using PARAFAC (Parallel Factor Analysis) .......... 104

6.2.3. Pathway Analysis and Dose-Response ..................................... 105

6.3. Results and Discussion ............................................................. 106

6.3.1. Factorial Design Offers Dynamic Insights of Stress Responses ............... 106

6.3.2. Tensor Decomposition (PARAFAC) of High Dimensional Toxicogenomics Data ................................................................. 109

6.3.3. Pathway Analysis Based on Tensor Decomposition ............................. 114

6.3.4. The Investigation of Dose-Response Relationship and Temporal Dynamics Using PARAFAC Analysis ........................................ 117

6.4. Conclusion .............................................................................. 119

Chapter 7 Identifying Dose and Time Dependent Impacts on Gene Co-Expression Networks in Toxicogenomics .......................................................... 121

7.1. Introduction ........................................................................... 122

7.2. Materials and Methods ............................................................. 125

7.2.1. Toxicogenomics Gene Expression Time Series Data and Gene Annotation Generation ........................................................................ 125

7.2.2. Data Pre-processing ............................................................... 127

7.2.3. Network Inference and Analysis ............................................. 127

7.3. Results and Discussions ........................................................... 128

7.3.1. The Impacts of Dose on Gene Co-expression Network Topology .......... 128

7.3.2. Dose-Aware Mechanistic Analysis Using Gene Co-Expression Network ... 135

7.3.3. The Impacts of Exposure Time on Gene Co-Expression Network Topology ........................................................................ 140

7.4. Conclusion .............................................................................. 143

Chapter 8 Conclusions and Future Outlook ................................................................. 144

8.1. Conclusions ........................................................................... 145

8.2. Overview of Future Studies ...................................................... 147
REFERENCE .................................................................................................................. 150
APPENDIX ....................................................................................................................... 167
  Appendix A  Analyzing High Dimensional Toxicogenomics Data via Consensus Clustering  168
  Appendix B  Toxicity Mechanisms Identification via Gene Set Enrichment Analysis of Time-Series Toxicogenomics Data: Impact of Time and Concentration ............ 173
  Appendix C  High dimensional Toxicogenomics Data Analysis Using Tensor Decomposition Model for Toxicity Identification and Characterization .................. 175
  Appendix D  Identifying Dose and Time Dependent Impacts on Gene Co-Expression Networks in Toxicogenomics ................................................................. 176
LIST OF FIGURES

Figure 1-1 Dissertation research overview ................................................................. 32
Figure 3-1. User interfaces in TimeVis (a) Experiment information interface where user
defines independent and dependent variables regarding the overall hypothesis of a
specific study. (b) Time series visualization interface, where user query data under
different conditions and plot them for comparison....................................................... 53
Figure 4-1. SOM and consensus clustering results for all treatments: (a) Self organizing
map (SOM). Treatments are projected onto the map. The differences of gene expression
pattern among the treatments are transferred into topographical distance on the map.
Hexagon colors represent clusters, which are determined according to treatment distances
using the U-matrix. Average CI values shown for each cluster are arithmetic mean among
all pairs of treatments within the cluster; (b) Consensus clustering (CC). The consensus
matrix represented as a heatmap. A dendrogram is based on the CI’s, showing the inner
structure in the blocks along the main diagonal. The identified clusters are labeled in
different colors in the dendrogram, and dashed square in the heatmap. Concentration unit
is mg/L. ........................................................................................................................... 65
Figure 4-2. Average consensus index for each node in consensus clustering. Results are
from four different experiments that selected different time-series data subsets, including
Max- dataset containing the maximum gene expression signal observed for each reporter
during the 2-hr assay; 20th - dataset that uses only gene expression data at the middle time
point of the testing, the 20th time point, at 57 minutes of exposure; TELI- dataset
consisting of derived integrated endpoint-Transcriptional Effect Level Index (TELI).
Total represents the full dataset with all the time points. ............................................. 69
Figure 4-3. Dendrogram produced from consensus clustering of nanomaterials. .......... 71
Figure 4-4. Consensus clustering result for dataset perturbed with additional noise: the
noise is Gaussian with mean value of 0 and standard deviation of 0.5 ......................... 72
Figure 4-5. Consensus matrix based on the clustering result with dynamic genes library:
the scheme of re-sampling is changed to reconstruct the treatment vectors using genes
randomly selected from the original library. With this perturbed input, the pattern
differences between treatments change accordingly, which in turn impact the resolving
power of the clustering algorithm.................................................................................. 73
Figure 5-1. Normalized temporal variation in gene expression profiles of stress response
pathways or categories upon exposure to various levels of (a) MMC, (b) H2O2 and (c)
Pb2+. Each curve represents the mean temporal variations (measured as ln(I), I is the
altered gene expression level) profile of all the genes (3 replicates) in the specific stress
response pathway, with 95% confidence intervals indicated by the gray bands. x-axis top: various stress response pathways and categories; x-axis bottom: exposure time in hour; y-axis left: normalized temporal variations in altered gene expression level as ln(I), where I is induction factor normalized against the first time point. Pathway/Stress response categories abbreviation: General (general stresses), Redox (Redox stresses), Protein (protein stress), Cell (cell killing), DNA (DNA damage), Drug (drug resistance), Detox (detoxification).

Figure 5-2. Comparison of GSEA results based on toxicogenomics data at three different time points upon exposure to MMC (0.5 ng/L) (30, 65 and 100 minutes upon exposure). Genes were ranked by mean altered gene expression levels based on triplicate treatments. X axis bottom: displays p value in negative logarithmic scale.

Figure 5-3. Comparison of GSEA gene ranking results for altered gene expression results for (a) MMC at 0.5 µg/L, (b) H₂O₂ at 10mg/L and (c) Pb²⁺ at 0.125 µg/L using two different ranking metrics CPCA score (left) and TELI (right). Genes are positioned based on their ranking score in non-decreasing order from left to right based. Each vertical strip represents a gene; with color-code indicating its associated stress response pathway and the dot displaying its ranking score. The results are based on average of triplicate treatments. The color codes for each stress response categories or pathways labels are shown in (d).

Figure 5-4. Comparison of GSEA analysis gene ranking results with consideration of all six dose concentrations for (a) MMC, (b) H₂O₂ and (c) Pb²⁺ using two different ranking metrics CPCA score (left) and TELI (right) for each chemical. Genes are positioned in non-decreasing order from left to right based on the metrics. Each vertical strip represents a gene, with color indicating its associated pathway and dot displaying ranking score. The results are based on results (triplicate treatments) of all six dose concentrations for each chemical. The color codes for each stress response categories or pathways labels are shown in (d).

Figure 6-1. Overview of three-dimensional tensor of time series toxicogenomics data (altered temporal gene expression in response to toxicant 4NNP) and dissected different 2-dimensional profiles: (a) The cubic tensor data structure shows that the altered gene expression levels are indexed by 3 experimental factors: gene, time and concentration. The arrow in each dimension indicates the order of that factor (increasing order for time and concentration). The small cube in gray shade illustrates a specific data point measured for a 1st gene, at the 1st time point after exposure to the 1st (lowest) concentration of the given chemical. (b-d) By slicing the tensor from different perspectives, various two-dimensional profiles can be extracted. For example, by slicing horizontally, gene × time profiles can be obtained at each concentration. In these profiles, the gene expression values within each profile are represented by the elevation in the vertical axis and color (color scale is shown in the color bar to the right). The dimensions are not proportional to the number of levels for individual factors, and random levels for the factors are selected for visualizing sliced profiles.

Figure 6-2. Demonstration of tensor decomposition (toxicant: 4NNP). The original tensor (Column 1) was decomposed into two components (Column 2 and 3), plus a residual tensor (Column 4). The tensors are sliced horizontally for visualization, and selected slices are displayed here with each row indicates a concentration (low to high from top to bottom). The tensor is the corresponding sum of the component tensors and residual.
Figure 6-3. The effect coefficients of factors within each component (toxicant: 4NNP) and their time- and concentration-dependent patterns. In tensor decomposition, each produced component is the outer product of 3 vectors, which contains contributing effect of each factor to the final gene expression values. So the $r^{th}$ component can be expressed as $Y_r = a \circ b \circ c$, where $a$, $b$, and $c$ are the respective vectors containing the effects of gene, time and concentration, shown above. The lengths of the vectors equal to the number of levels for the corresponding factors. An individual gene expression value in the component can be expressed as $y_{ijk} = a_i b_j c_k$, where $a_i$, $b_j$, and $c_k$ represents the effects of the $i^{th}$ gene, $j^{th}$ time point, and $k^{th}$ concentration in this component, respectively. In (a), genes are sorted by their coefficients in Component 1.

Figure 6-4. Dose concentration-dependent effects of individual components for the four toxicants. The order of the components for each toxicant is the same as Table 6-1. Figure 6-5. Time-dependent effects of individual components for the four toxicants. The order of the components for each toxicant is the same as in Table 6-1.

Figure 7-1. Gene co-expression networks constructed using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused E. coli whole cell reporter library with exposure to arsenic of different doses: from (a) $1.4 \times 10^{-5}$ mg/L to (f) $1.4$ mg/L. The networks are displayed using circular layout where the same genes are located in the identical locations of both. For visualization simplicity the gene names are omitted, but the pathways to which the genes belong are denoted using different colors as shown in (g).

Figure 7-2. Network level dose-response pattern: the linear relationship between gene co-expression network size (number of edges), structure (average clustering coefficient) and toxicant arsenic dose concentrations. Gene co-expression networks constructed using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused E. coli whole cell reporter library with exposure to arsenic at different doses. The X axis is in logarithmic scale.

Figure 7-3. Temporal trends of altered gene expression profiles for different stress response pathway categories induced by exposure to arsenic at varying doses concentrations, using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused E. coli whole cell reporter library with exposure to arsenic of different doses: (a) $1.4 \times 10^{-5}$ mg/L, (b) $1.4 \times 10^{-3}$ mg/L and (c) $1.4$ mg/L. The temporal mean altered gene expressions of all genes in each specific stress response pathway are shown in solid curves using different color legend (d). The 95% confidence intervals for each curve are shown as the bands surrounding the curves with the lighter colors, indicating the range of variations of altered gene expression levels among all the genes in each stress response pathway.

Figure 7-4. Gene ranking and gene set enrichment analysis results for all stress response pathways at varying arsenic dose concentrations, using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused E. coli
whole cell reporter library with exposure to arsenic of different doses: (a) DNA damage, (b) general stress, (c) membrane stress, (d) protein stress, and (e) oxidative stress. The genes are ranked according to their degree (number of connections) in the gene co-expression network at each dose. Y axis left: genes represented by horizontal bars and listed based on ranking order (lower value means higher ranking order) as shown in vertical columns; Y axis-right: GSEA enrichment scores shown in solid circle: the higher the enrichment score is, the more enriched toward the top the specific gene set is; X axis top: The statistical significances of enrichment for the specific pathway at a given dose concentration (after corrected for multiple-comparison using false discovery rate Benjamini-Hochberg procedure). The possible significant enrichment (p<0.25, threshold determined according to Subramanian et al.) are highlighted in red. X axis bottom: arsenic exposure dose concentrations in mg/L (Subramanian et al., 2005).

Figure 7-5. The change of gene co-expression network size (number of edges, solid line) and structure (average clustering coefficient, dashed line) over time: the networks were constructed using altered gene expression data at each time points generated with stress response pathway ensemble of GFP-fused E. coli whole cell reporter library with exposure to arsenic of 1.4×10⁻³ mg/L. Y axis left: edge number of networks; Y axis right: clustering coefficient; X axis: time after exposure to arsenic. The time series are smoothed using moving average method with a window of 5.

Figure A-1. Cluster number determination by finding the peak cumulative distribution function (CDF) of area change.

Figure A-2. Consensus clustering using (a) the maximum expression values of each gene, (b) the 10⁻⁰ time points, (c) the 20⁻⁰ time points. Y axis: CI value.

Figure A-3. Consensus clustering using TELI values for each gene. Y-axis: CI value.

Figure A-4. Consensus clustering with dataset built with only redox stress genes. Y-axis: CI value.
LIST OF TABLES

Table 5–1. Comparison of GSEA results (p-values) using two different ranking metrics, namely CPCA score and TELI, respectively. ................................................................. 93
Table 5–2. Comparison of GSEA results with consideration of six dose concentrations using different ranking metrics, CPCA score and TELI ......................................................... 99
Table 6-1. Summary of GSEA (gene set enrichment analysis) results for the components of each chemical. The significance of enrichment (p value) for each stress response pathway is shown, after correction with false discovery rate method. ......................... 119
Table A-1. Specifications of chemicals and nanomaterials used in the study. ............ 168
Table A-2. Summary of the characterization data for all the nanomaterials. .......... 169
Table B-1. Genes included in the E. coli stress response ensemble assay library for toxicogenomics assessment. ........................................................................................................... 173
Table B-2. Model Chemicals Analyzed in This Study and Know Toxicity Mechanisms Reported .................................................................................................................. 174
Table C-1. Specifications of chemicals used in this study........................................ 175
Table C-2. Genes used in the study and their pathways. ........................................... 175
Table D-1. Genes and their pathways used in the study for dose-related study. ....... 176
Chapter 1

Introduction and Objectives
1.1. Toxicogenomics: Progresses and Challenges

1.1.1. Toxicogenomics: Paradigm Shift in Toxicology

There are an overwhelmingly large number of unregulated chemicals that have human exposure potential but little toxicity information (United States Government Accountability Office., 2005). Relying on the evaluations of adverse phenotypical effects on animals, current toxicity testing approaches are lengthy and resource-intensive (Ankley et al., 2006). There are thus urgent needs for the development of toxicity mechanism- and pathway-based cost-effective testing scheme as alternatives to the conventional whole-animal-based testing, then in time we hope to eventually build predictive models and tools to provide toxicological information that transcends the limits of data generation (Cunningham & Lehman-McKeeman, 2005; Hayes & Bradfield, 2005). Particularly, it demands for high-throughput and cost-efficient testing techniques for public health protection as well as environmental risk assessments (Hayes & Bradfield, 2005). For such reason, it is envisioned, as described in the Toxicity Testing in the 21th Century report by the National Research Council, that the dominant toxicity testing approaches will be shifting from animal-based in vivo tests to cell-based in vitro assays with robotic automation and mechanistic quantitative parameters (Krewski et al., 2010).

Toxicogenomics, which is the combination of genomics with toxicology and the foundation of in vitro assays, describes the study of molecular adverse effects caused by environmental or pharmaceutical chemicals to human health, through the utilization of transcriptomics, proteomics and metabolomics (National Research Council Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology, 2007). Based on
molecular biology and computational biology, toxicogenomics techniques target cellular responses and provide insightful data on toxicity mechanisms. Because of its high-throughput capacity, toxicogenomics assays are able to cover a large range of toxicant doses with significant lower cost and shorter time. In addition, the high-resolution and information-rich exposure fingerprints can then be used, in combination with bioinformatics tools, for toxicants and exposure effects classification, as well as for revealing and predicting underlying toxic mechanisms and mode of action. Due to abovementioned advantages, toxicogenomics has become a promising and ever-growing field since its initial definition in 1999 (Nuwaysir, Bittner, Trent, Barrett, & Afshari, 1999). Multiple programs and initiatives in various countries have been trying to incorporate in vitro assays into toxicity screening and evaluation, such as US EPA TOXCAST, US NIH Nation Toxicology Program High-throughput Screening Initiative, European Union REACH regulation, and so forth (Dix et al., 2007; R. Huang et al., 2008; Rogers, 2003).

1.1.2. Challenges in Toxicogenomics

The benefits of toxicogenomics information and tools in regulatory ecotoxicology are only starting to be elucidated. Despite the vast interests and rapid growth in toxicogenomics, there are still several remaining challenges the whole field is facing in the context of environmental and human health protection. Existing challenges limited their application for environmental regulatory and risk assessment (Ankley et al., 2006; Boverhof & Zacharewski, 2006; National Research Council Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology, 2007). First, the biological system is a rather complex system, and most cell-based assays can only target limited
number of cellular components. How to select appropriate identifiers and quantifiers of cellular targets to cover a range of MOAs and/or recipients relevant for the environmental samples/chemicals are still being investigated. Second, the key scientific challenge is to identify reliable patterns (signatures) and magnitude (intensities) to describe specific exposures and molecular effects. There will be significant statistical challenges in establishing qualitative and quantitative criteria for recognizing transcriptomic, proteomic, and metabolomics signatures of exposure. Third, there will be difficulty of extrapolating data from the simple biologic system of single cells to the complex interactions in whole animals. In addition, to further incorporate these in vitro bioassays into current environmental regulations and decision making frame, comparison and correlation of molecular endpoints with conventional whole-animal tests results have to be established. Lastly, mixture toxicity evaluation is essential since current threshold regulatory doses/concentrations based on points of departure may not be sufficient against combined exposures. We will further discuss some of these technical hurdles in the form of four important questions as follows.

1) How to meaningfully quantify the molecular effects caused by toxicants?

Although toxicogenomics techniques are now able to monitor cellular responses at the molecular level, there is still a lack of fundamental understanding and knowledge and therefore no consensus in the approaches on how to accurately determine and quantify the real alterations of gene, protein, or metabolite expression caused by toxicant (Ankley et al., 2006). Currently, most of the toxicogenomics studies have been focusing of qualitative instead of quantitative analysis of molecular endpoints (Krewski et al., 2010). One reason is the lack of comprehensive information about the baseline states of cells.
Such baseline information includes not only a single expression level for each gene in homeostatic state, but potentially also their temporal expression patterns, since gene expression is not static but may have cycles or rhythms (Bar-Joseph, Gitter, & Simon, 2012). The other reason is that, in order to derive a robust molecular endpoint to quantify external perturbation, it is necessary to have a thorough understanding of the underlying biology of the cell lines. However, prior knowledge in terms of gene and functional pathway annotation for this purpose is still quite limited right now (Ankley et al., 2006).

There were some attempts and efforts in trying to quantify molecular disturbance. For instance, the percentage or number of differentially expressed genes was used for the quantification of the molecular perturbation in various species (Menzel, Bogaert, & Achazi, 2001; Tilton et al., 2011). In recent years, our group and others have explored some novel ways to quantify molecular endpoints such as TELI (Transcriptional Effect Level Index) and PELI (Protein Effect Level Index) (Gou & Gu, 2011; Lan, Gou, Gao, He, & Gu, 2014). These studies demonstrated that molecular effect can potentially be quantified at both individual gene levels and pathway or biomarkers ensemble level.

2) How to link molecular endpoints with phenotypic ones at regulatory relevant level?

To integrate the molecular endpoints into regulatory decision making, it is necessary to understand the link between the short-term, lower level in vitro responses and long term, whole-animal level adverse outcomes (Ankley et al., 2006). Adverse outcome pathway (AOP) is a conceptual construct that provides a roadmap for establishing linkage between a direct molecular initiating event (e.g. a molecular interaction between a xenobiotic and a specific biomolecule) and an adverse outcome at a biological level of organization relevant to risk assessment (Ankley et al., 2010; National Research Council...
Committee on Toxicity Testing and Assessment of Environmental Agents, 2006). AOP is able to describe the known linkage along the continuum from the molecular event in which a chemical interacts with a biological target(s); following on through a sequential series of cellular, anatomical and functional changes in biological process, and ultimately culminating in an adverse outcome of relevance to human or ecological risk assessment (Chang, Smalley, & Conway, 2002).

The study, also known as phenotypic anchoring, may also help discern the real toxicological pathways from irrelevant biological processes that get mixed into the expression signals (Paules, 2003). Although promising, phenotypic anchoring has not received rigorous enough studies, partially due to the complexity of biological systems. In other words, there are possibly multiple pathways that will respond to stressors directly or indirectly, but it is difficult to distinguish them analytically.

So far, most phenotypic anchoring studies have been mainly linking qualitative results of molecular assays, such as binary results of positive and negative, to phenotypic endpoints, whereas studies showing quantitative links are very limited (Bugiak & Weber, 2010; Ung et al., 2010). In the previous studies of our group, there were evidences demonstrating quantitative correlations between molecular endpoints with phenotypic ones (Gou & Gu, 2011; Lan et al., 2014). However, it is still an ongoing effort in determining the real existence of such quantitative links, and the level of molecular quantification (genes, pathway or above). In recent years, there is a major interest for quantitative anchoring of *in vivo* and *in vitro* studies, initiated by public toxicological research project, such as ToxCast from US EPA (Kleinstreuer et al., 2014; Wambaugh et al., 2013). More importantly, the molecular endpoints in these studies are not limited to
adverse outcome pathways but include global gene expression profiles, the large volume of data and complexity of statistical correlation analysis calls for more sophisticated approaches from computation biology (Luo et al., 2005).

3) How to incorporate dose-response and pharmacokinetics modeling into toxicogenomics?

There are two types of models to consider so that in vitro testing results on a small range of chemical doses can be extrapolated and linked to in vivo testing result. Dose-response modeling involves the quantification of mechanistic information about toxicity pathways and integration of dosimetric information, which allows for dose extrapolation of a toxicant’s toxicity (Conolly, 2002; Krewski et al., 2010). However, to fully evaluate the toxic effects of toxicants, the dose-response data in toxicogenomics need also to encompass the information about biological processes of absorption, distribution, metabolism and elimination in animals, also known as physiologically based pharmacokinetic (PBPK) modeling (Collins, Gray, & Bucher, 2008). Pharmacokinetics information will compensate toxicogenomics for the extrapolation across species, and help estimate the adverse effects based on relatively low dose exposure samples (Krewski et al., 2010). To this date, the studies on these models are still insufficient for practical applications.

4) How to integrate molecular biology and bioinformatics to develop predictive toxicology?

The ultimate goal for toxicogenomics is the development of predictive toxicology, which is the study aiming at replacing animal-based toxicity tests with molecular level tests. Besides various -omics techniques, the possible approaches also include in silico
screening and modeling. However, to achieve such objective, it is imperative to integrate data generated from multiple platforms, system biology information as well as results from different simulation models, like quantitative structure-activity relationship (QSAR) models (Cunningham & Lehman-McKeeman, 2005; National Research Council Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology, 2007). Current bioinformatics methods do not yet adequately support such integration with respect to data standardization, comparison and analysis (Ankley et al., 2006).

1.1.3. Bioinformatics Applications in Toxicogenomics

The advances in high-throughput gene expression profiling assays, such as microarray, RNA-seq and whole cell array, allows for the simultaneous monitoring of the transcriptional activities of a large number of genes (Waters & Fostel, 2004). Utilizing such techniques in toxicogenomics can provide biological insights at the molecular level in a fast and less expensive way, serving as initial toxicity screening for guiding and prioritizing further toxicity evaluation, thus reducing the need for laboratory animal testing (Simmons, Fan, & Ramabhadran, 2009). These high-throughput toxicogenomics assays generate larger amount data that require proper data collection procedure, quality control protocols and analysis approaches. Particularly, the advances in the technologies now enable continuous measurement of real time gene expression that can better monitor the dynamic toxicant responding process, producing time series data. In addition, when multiple experimental factors are involved, such as chemicals, doses and genes, the data need to be organized in a high dimensional fashion. The resulting large volumes of data with convoluted structures require the development of more sophisticated analytical techniques, which present great challenges in data analysis and data mining.
In general, bioinformatics applications in toxicogenomics can be categorized into two groups: characterization of toxic substances and characterization of biomarker and biological systems (Afshari, Hamadeh, & Bushel, 2011).

1) Characterization of toxic substances

The core of toxicogenomics is to understand the characteristics of toxicants using genomics information collected from high-throughput assays. The particular applications include the following.

The early application of bioinformatics in toxicogenomics is to cluster the chemical samples into groups, based on the assumptions that toxicants with similar gene expression profiles may have close toxic mechanisms of action. Toxicological knowledge may be extracted by further analyzing the individual groups; and toxicity of unknown samples may be extrapolated from other known compound in the same cluster. The common clustering techniques used are hierarchical clustering (Eisen, Spellman, Brown, & Botstein, 1998), self-organizing maps (Tamayo et al., 1999) and principal component analysis (Yeung & Ruzzo, 2001). However, it is recognized early on that the experimental designs of toxicogenomics studies is different from those of other biological disciplines, e.g. cancer studies. The former often involves time series and dose-response related samples, in contrast to the experiment against control samples in the latter. Some improved clustering techniques are designed for toxicogenomics data, including (semi)-supervised clustering that harnesses the phenotypic data (Bushel et al., 2007), and biclustering which identifies of samples that share highly similar patterns for a subset of data features (Cheng & Church, 2000; Prelic et al., 2006). Still, some open issues exist, such as how to account for gene expression offset in the time or dose dimension, and how
to compare the groups statistically for clustering performance evaluation (Afshari et al., 2011).

Discerning mechanism of action of toxic responses is an integral part of risk assessment and toxicogenomics studies (Cunningham & Lehman-McKeeman, 2005). In comparison of identifying differentially expressed genes, mechanistic studies unravel the underlying biological events and processes in response to external stressors. The techniques developed for this purpose are generally called pathway analysis which includes gene ontology analysis, gene set analysis and gene network analysis (Khatri, Sirota, & Butte, 2012). The key component for these analytical methods is a good understanding of biology and hence reliable databases for gene annotations and other related biological knowledge, include gene ontology (Huntley et al., 2015), pathway information (Kanehisa & Goto, 2000), as well as some toxicogenomics databases (Ganter et al., 2005).

2) Characterization of biomarker and biological systems

From the perspective of molecular biology, understanding the underlying biological processes in the events of toxicological stresses would help design better assays and facilitate all the toxicant oriented bioinformatics applications to obtain more insightful conclusions. The applications in this category are in general about the characterizations of biomarkers and biological systems.

Despite the progress of high-throughput bio-assays, it is still cost-effective to focus on subsets of all genes for an organism. It is not practical to design in vitro high-throughput assays that can monitor every potential toxicological target and pathway. Therefore, it is necessary to develop an integrative approach based on indicative key
genes (biomarkers) in limited and conserved pathways associated with adverse outcome
(Simmons et al., 2009). Cellular stress response pathways, for example, are good
candidates in that they detect and repair damage caused by toxicants on primary cellular
infrastructure (Simmons et al., 2009). More importantly, stress response pathways proved
a manageable set of biomarkers to enable the development of high-throughput cell-based
assays (Gou, Onnis-Hayden, & Gu, 2010; Onnis-Hayden et al., 2009; Simmons et al.,
2009; Vollmer & Van Dyk, 2004). However, it is recognized that there are molecular
toxic effect and endpoints beyond stress responses that require other biomarkers to
capture. The set of biomarkers and their expression patterns can become biological
signature for toxicants or classes of toxicants. Compared to conventional individual genes
based biomarkers, advanced biomarker discovery take unique gene expression patterns
into account as well, where bioinformatics models and tools are instrumental
(Bartosiewicz, Jenkins, Penn, Emery, & Buckpitt, 2001; Zidek, Hellmann, Kramer, &
Hewitt, 2007). Furthermore, the selection of gene subset should also consider time and
dose dependent patterns as well (Peddada et al., 2003).

Because of the availability of high-throughput toxicogenomics data, bioinformatics
applications can focus on higher level of biological entities, such as gene pathways or
networks, which is often described as systems biology. For instance, pathway analysis
aims at identifying key functional gene groups and regulatory pathways evoked during
the toxicant exposure under a given condition (D. W. Huang, Sherman, & Lempicki,
2009; Khatri et al., 2012). Instead of focusing on differentially expressed genes
individually, pathway analysis approaches try to discern sets of related genes that share
common biological function or regulation. The results of pathway analysis are more
robust and less complex, compared to the individual analysis of a daunting number of
genes (Khatri et al., 2012). In the past, several particular pathway analysis tools have
been employed successfully in toxicogenomics, such as gene ontology analysis
(Ashburner et al., 2000) and gene set enrichment analysis (Subramanian et al., 2005). In
addition, due to the success in the theoretical study of complex systems and networks,
gene network analysis have been applied more and more in toxicogenomics to understand
the state of cellular activities and the changes of the biological system structure during
external perturbations (Aderem, 2005). The studies of gene network properties have been
used in the discovery of biomarkers and endpoints for quantify molecular effects in
toxicogenomics (Eschrich et al., 2009; Ideker & Krogan, 2012).

1.1.4. Challenges in High Dimensional Time Series Toxicogenomics Data
Analysis

The advances in toxicogenomics techniques, as well as the recognition of the cellular
dynamics with time and sensitivities to dose and other exposure conditions, motivate
more complicated experimental design that consider time, dose and other factors (Bar-
Joseph et al., 2012; Hayes & Bradfield, 2005). The generated data is not large in volume
but also complex in structure. As most current bioinformatics tools were designed
specifically to handle statics toxicogenomics data, they are insufficient for those new
high dimensional time series data (Ankley et al., 2006). In particular, there are several
major challenges on the analysis of such complex data, as discussed below.

1) Challenges of time series data analysis

It has long been recognized that cellular responses to toxicants are highly dynamic
and their global response profiles depends on time point of measurement (Bar-Joseph et
al., 2012). However, the efforts in illustrating the impact of time on toxic assay results
have been quite limited mostly due to the lack of time-series toxicogenomics data. This is because that the labor-intensiveness or high-cost associated with mainstream toxicogenomics techniques such as RNA-seq or microarray technologies prohibit measurements with high temporal resolution (Jayapal, 2012). An alternative approach is the use of whole cell arrays with transcriptional fusions of reporter genes, which allows for faster and lower-cost real-time measurement of temporal gene expressions for a large number of chemicals under various test conditions (Melamed, Elad, & Belkin, 2012; Timothy et al., 2012).

The time series gene expression data generated by such arrays calls for analysis approaches that are time-factor sensitive. Most current studies simply adopt strategies extended from those of static, time-independent experiments and resort to integrated endpoint-like quantities (Gou & Gu, 2011), which do not account for the dynamic nature of stress responses and lose temporal information by discarding all information other than endpoint measurements (Gao, Weisman, Gou, Ilyin, & Gu, 2012; Schliep, Costa, Steinhoff, & Schonhuth, 2005). In this dissertation, we proposed and demonstrated the application of several different strategies for addressing the challenges in analyzing high dimensional time series gene expression data.

2) Challenges of high dimensional data analysis

Toxicogenomics data are often highly dimensional in nature (Vladimirova & Ganter, 2008). This high dimensionality refers to either the large number of genes considered or the amount of factors impacting the result that include chemicals, exposure time, and dose concentrations, all of which can be problematic in toxicogenomic studies (Clarke et al., 2008; Donoho, 2000). In toxicogenomics, the number of genes is often several orders
of magnitudes higher than the quantity of samples surveyed in a typical high-throughput study. Several properties of the resulting high dimensional data space are detrimental to the performance of most statistical models, leading to a problem known as the “curse of dimensionality” (Donoho, 2000). Firstly, because the possible outcomes of a model grow exponentially in proportion to the number of genes, whereas the sample size is usually limited, the resulting data space becomes very sparse (Hinneburg, Aggarwal, & Keim, 2000). In addition, the sample data points that lie in such space become effectively equidistant to each other, therefore debasing the discriminant power of algorithms that use distance-based metrics to differentiate samples such as clustering methods (Xu, Damelin, Nadler, & Wunsch, 2010). Secondly, models with high dimensional inputs are prone to overfitting and local convergence, which makes difficult for them to generalize on unseen data, therefore compromising their predictive power (Somorjai, Dolenko, & Baumgartner, 2003). To avoid these issues, dimension reduction techniques are usually applied in the data pre-processing stage preceding the main bioinformatics analysis, and resampling-based approaches like bootstrapping are used to improve the reliability of the results (Gao et al., 2012; van der Laan & Bryan, 2001).

The other aspect contributing to the high dimensionality of toxicogenomics data is the increasing number of experimental factors or covariates impacting cellular response such as gene expression, also known as modality. Besides genes, these factors include the type of toxicants, their concentrations, as well as the time points at which the measurements were taken in a time course experiment. This more sophisticated factorial design provides a more thorough investigation of the dynamic response of an organism after exposure to a set of toxicants; meanwhile it also results in a more complex data structure. This makes
decoupling and interpreting the effect of each factor a very challenging task. The total set of data for the entire experiment with multiple samples of the toxicant, possibly from varying concentrations, is organized by stacking \( time \times gene \) matrices together into a three-dimensional array \( (time \times gene \times dose) \), also known as tensor, which is unsuitable as input for most statistical models (Clarke et al., 2008; Ponnapalli, Saunders, Van Loan, & Alter, 2011). Consequently, it is common to unfold the data tensor to lower-dimensional matrices or vectors to accommodate the models (Gao et al., 2012). This approach is not appropriate as it loses the degrees of freedom represented by those factors and treats them indifferently (Bar-Joseph, 2004; Omberg, Golub, & Alter, 2007). In addition, such method worsens the curse of dimensionality by increasing the number of features (Bandyopadhyay, Ganguli, & Chatterjee, 2011). The combined difficulties in both aspects of data high dimensionality hamper modern high-throughput data analysis and call for the introduction of new analytical approaches.

1.2. Research Overview

1.2.1. Research Objectives and Overview

The objectives of this dissertation are to propose and demonstrate the improved methodology that better address the challenges and limitations in high dimensional toxicogenomics data analysis for toxicity mechanism identification, toxicants classification, and for predictive toxicology knowledge discovery. In particular, the specific objectives of this research are:

1) Explore quantitative analytical methodology for toxicogenomics data analysis.

Most current toxicogenomics studies are qualitative rather than quantitative, limiting the possibility of cross-study comparisons and inter-species extrapolations, as well as
their applicability in regulatory decision making (Ankley et al., 2006; Krewski et al., 2010). Furthermore, focusing mostly on expression alternations of individual genes or descriptive conclusions about functional pathway, these studies did not develop quantitative molecular endpoints at systems biology level that allow for reliable dose-response evaluation. In this research, we explore the possibility to incorporate various molecular effect quantifiers/approaches while exercising different toxicogenomic data analysis, such as clustering, pathway analysis, data decomposition, as well as gene network analysis, involving experimental measurements of cellular molecular activities at individual gene, specific pathway or stress response system level. Through the establishment of quantitative analytical methodology, this study aims at seeking insights on whether molecular disturbance caused by toxicant can be quantified, and the level of such quantification, namely gene, pathway, or system levels. In the end, we are also trying to elucidate the quantification characteristics such as their dose-response relationships and temporal dynamics.

2) Develop data analysis methodology that incorporate time factor onto the design and data analysis in toxicogenomics studies

In terms of experimental design, the majority of studies are static, taking only a snapshot of the biological systems, partially due to the lack of proper data collection capability. Most toxicogenomics techniques are destructive, and cost- or resources-prohibitive for generating large amount of time series data (Zoppoli, Morganella, & Ceccarelli, 2010). Because of this, time-dependency and temporal dynamics of cellular response to toxicants have not yet been thoroughly investigated. Specifically, current bioinformatics approaches and software are not able to handle the high dimensionality of
time series data, or take temporal dynamics into account for toxicity evaluation. Our group has recently developed a quantitative toxicogenomics testing platform using whole cell array with green fluorescent protein (gfp) fusions to different gene promoters, which monitors temporal gene expression changes through measuring cellular GFP levels (Gou et al., 2010; Onnis-Hayden et al., 2009). It can generate high-resolution time series toxicogenomics data for a large number of environmental pollutants, therefore provide a unique database for exploration of bioinformatics tools/approaches that focus on capturing and revealing time-sensitive nature and dynamics of cellular response to toxicants. In this study, we incorporate and evaluated the importance of time factor in the analysis of toxicogenomics data, and develop various data collection, visualization, and analytical methods that integrate time series data.

Figure 1-1 shows the overview of this study that illustrate the specific objectives, the organization and relations among the particular tasks and studies included in this dissertation. Our high-throughput toxicogenomics assay platform using GFP-fused reporter whole cell library of *E. coli* K12, MG1655 (Gou et al., 2010; Onnis-Hayden et al., 2009) platform enables generation of high dimensional time series transcriptomics data to be used for further analysis. We first developed a comprehensive data collection, analysis and visualization platform and software that allow easy data management, data quality control, data visualization and customizable research inquiries. In this dissertation, I proposed and applied different unsupervised data mining techniques to improve data analysis of high dimensional toxicogenomics data in order to obtain more systematic, holistic, in-depth and finer-resolution toxicological information for toxicity identification and characterization. Three type of functional data analysis were performed and they
include major bioinformatics application, responding to the core challenges in toxicogenomics, namely, toxicant classification based on altered gene expression profiles signatures with temporal resolution (Chapter 4), Toxicity mechanism identification (Chapter 5, 6, 7), and quantitative dose-response examination (Chapter 6 and 7). In all these analysis, improved and novel approaches to incorporate temporal dimension of the time series toxicogenomics data were applied in various gene alteration, pathway activation and system network analysis. These analyses aim to evaluate and importance and impact of exposure time factor on the outcome, as well as to elucidate potential quantitative molecular disturbance patterns, and molecular dose-response relationships in toxicogenomics studies.

The outcome of this study will make contributions to several fields including toxicology, bioinformatics and environmental science in the following aspects:

1) Develop new or improves toxicogenomics data analysis algorithms and methods that is capable of processing high dimensional time series data, therefore better capture and reflect the dynamics of cellular response to toxicants.

2) Prove the potential and validity of incorporating various molecular disturbance/effect quantifiers into various functional toxicogenomics bioinformatics to provide quantitative insights into the toxicant-induced cellular molecular responses at individual gene, specific pathway and system levels.

3) Demonstrated the effectives of unsupervised bioinformatics tools for mining new, more in depth, much-detailed and fundamental knowledge and understanding of toxicological information at molecular level.
4) Generate new information to fill in the urgent knowledge gaps in toxicogenomics that present barriers to the realization of predictive toxicology and provide insights to key questions related to molecular effect quantification, molecular dose-response relationship characterization.
High Dimensional Time Series Toxicogenomics Data Analysis

Data Management, Analysis, Visualization Software Platform

- Functional Analysis
  - Toxicants Classification
  - Toxic Mechanisms Identification
  - Dose-Response Relationship

Quantification at Different Molecular Levels
- Gene Level
- Pathway Level
- System Level

Time Impact
- Temporal Dynamics Importance of Time

Figure 1-1 Dissertation research overview
1.2.2. Organization of the Dissertation

This dissertation consists of 8 chapters. The synopsis and content of each chapter is described below:

1. Chapter 1 provides background information and motivations of this study. In particular, this chapter first introduces the advantages and key challenges of toxicogenomics, as well as applications and technical barriers of bioinformatics in toxicogenomics. In the end, this chapter describes the overall objectives and organization of studies in this dissertation.

2. Chapter 2 reviews the trend and challenges in high dimensional time series toxicogenomics data analysis. Specifically, this chapter provides an overview of the current status of bioinformatics and data mining techniques in dealing with high dimensional toxicogenomics data, in the area of toxicant clustering, gene network analysis, pathway analysis and tensor decomposition. As an important component of bioinformatics applications, commonly used software as well software development are also reviewed.

3. Chapter 3 describes the development of a data management, analysis and visualization software platform based on a web application. This software aims assisting in the standardization of toxicogenomics assay data sharing, analysis and comparison, provides intuitive data management and visualization. More importantly, through powerful plugin system, it can help software developer easily extend the analytical functionality and integrate computational biology solutions into the toxicogenomics.
Chapter 4 applies consensus clustering technique to cluster toxicants according to their temporal gene expression profiles. The proposed technique is designed to overcome the shortcomings of traditional clustering on high dimensional time series toxicogenomics data, namely low reliability and difficulty in performance evaluation. In this chapter, the temporal gene expression profiles of various chemicals, especially nanomaterials, are used to classify the toxicants into mechanistically meaningful groups. Using newly developed performance evaluation strategy, the robustness of consensus clustering against noisy data is examined. In addition, data processed by different dimensional reduction methods are clustered and compared, in order to understand the importance of time factor in toxicogenomics analysis.

Chapter 5 discusses the possibility of integrating an innovative projection based data mining approach, common principal component analysis, into conventional pathway analysis tool, gene set enrichment analysis. Such improvement on conventional toxic mechanisms of action identification technique makes it possible to incorporate time factor into consideration while discern the significant gene pathways during cellular stress response. The results of pathway analysis with or without time factor awareness are compared, and the impact of time factor on mechanistic analysis in toxicogenomic is discussed.

Chapter 6 uses high dimensional data analysis technique, namely tensor decomposition (PARAFAC model in particular), to disintegrate the complicated toxicogenomics data into simpler components, in order to study the dose- and time-dependent response patterns. Using experimental results based on several
toxicants, the feasibility of applying tensor decomposition to high dimensional time series toxicogenomics data is explored. In addition, pathway analysis has also been conducted on the resulting components from the decomposition, to understand their mechanistic meaning and generally applicability of tensor decomposition techniques for toxic mechanisms identification.

7. Chapter 7 explores the possibility of deriving molecular endpoints for the studies of dose- and time-response relationships from the systems biology perspective. In this study, gene co-expression networks using toxicogenomics data under different conditions have been constructed. Molecular endpoints based on topological characteristics of the networks are analyzed using regression analysis against time and dose experimental factors. In addition, the possibilities of using network characteristics at different levels including individual genes or the whole network for pathway analysis have been investigated.

8. Chapter 8 summarizes the conclusions or major contributions of this study and provides some recommendations for future studies.
Chapter 2

Literature Overview: Trends and Challenges in High Dimensional Time Series Toxicogenomics Data Analysis
As the general trend of analytical applications in toxicogenomics has been summarized in last chapter, this chapter provides an overview of the current status of the specific bioinformatics and data mining techniques in dealing with high dimensional time series toxicogenomics data. High dimensional toxicogenomic data represents the responses of numerous transcripts, proteins, metabolites, or reporters to toxicants. These responses are functions of the specific cell line phenotypes, the specific toxicants and their dose concentrations, as well as the time points of measurement (Daxin, Chun, & Aidong, 2004). In addition, cellular responses can be affected by confounding factors such as the media composition, degree of mechanical agitation during treatment and the homogeneity of the toxicant in the growth medium. The variations in experimental design, as well as normal biological and technical measurement noise, make inter-study comparison and biological generalization difficult (Ioannidis & Khoury, 2011).

2.1. Common Data Mining Algorithms in Toxicogenomics

In this section, we introduce some data mining algorithms that are frequently used in toxicogenomics studies. We will survey the variations of different techniques, as well as their advantages and shortcomings.

2.1.1. Toxicants Clustering

In order to discover biologically meaningful patterns and relationships, and to identify biological similarities in the responses to different toxicants from the complex high dimensional omics datasets, computational data mining and statistical analysis techniques have been applied (Harper & Pickett, 2006). The most frequently used methods are unsupervised clustering algorithms such as hierarchical clustering and self-organizing maps (SOM) (Afshari et al., 2011; Daxin et al., 2004). These clustering algorithms
partition the input data into non-overlapping subsets, which represent the underlying biological similarities (D’Haeseleer, 2005; Jain & Dubes, 1988). Dendrogram visualization of hierarchical clustering has been widely applied in biological studies, however, the hierarchical clustering results can be sensitive to outliers, and the visual representation neither conveys the number of distinct biological clusters, nor the clear relationships between those clusters (Bar-Joseph et al., 2003; Inglese et al., 2006; Tamayo et al., 1999). Clustering by SOM is also conventionally used for discovering groups of co-expressed genes or classifying chemicals (Törönen, Kolehmainen, Wong, & Castrén, 1999). SOM is relatively resistant to noisy data and missing observations, both common in biological research (Mangiameli, Chen, & West, 1996). However, SOM algorithm can also be sensitive to the random process of map initialization, which makes comparisons between studies challenging (Halkidi, Batistakis, & Vazirgiannis, 2001).

These challenges in analyzing HTS toxicogenomic data point to the pressing need of a method for validation of clustering results and assessment of cluster robustness. Current validation schemes include internal approaches, which are based solely on analysis within a dataset, and external approaches, which compare clusters with known class labels (Datta & Datta, 2006; Gibbons & Roth, 2002). It is argued that the internal measures might not be suitable for biological data which are subject to high noise levels (Datta & Datta, 2003). To address clustering stability and validation, Monti et al. developed the consensus clustering method, which iteratively re-samples the input dataset and invokes a conventional clustering algorithm over the reconstructed dataset (Monti, Tamayo, Mesirov, & Golub, 2003). Consensus clustering aggregates the ensemble of clustering outputs into a single consensus result, which has a higher confidence level than a typical
single invocation of the underlying clustering algorithm. Over the past few years, several researchers have applied this approach in microarray-based gene expression research, and found that consensus clustering could identify reliable clusters and hidden patterns, and therefore provide a highly meaningful interpretation of the biological response (Nguyen, Nowakowski, & Androulakis, 2009; Seiler, Huang, Szalma, & Bhanot, 2010; Wilkerson & Hayes, 2010). Evaluation of the consensus clustering methods using a relatively large toxicogenomic dataset with temporal resolution has not yet been reported.

2.1.2. Pathway Analysis

Another functional toxicogenomics tool is pathway analysis. Pathway analysis is one family of bioinformatic tools for toxicity mechanisms elucidation, which aims at pinpointing key functional gene groups and regulatory pathways evoked during the toxicant exposure under a given condition (D. W. Huang et al., 2009; Khatri et al., 2012). Through shifting the focus from detecting differentially expressed genes individually to discerning sets of genes that share common biological function or regulation, pathway analysis catches the expression patterns on the higher pathway level, avoids results misinterpretation due to subjective expression thresholds for individual genes, and reduces the complexity of data analysis that deals with daunting number of genes.

Pathway analysis of high dimensional toxicogenomics data, such as time series data, faces great challenge, however, since most current techniques are mainly designed for the analysis of a biological system snapshots (Khatri et al., 2012). The commonly used pathway analysis techniques, such as the gene set enrichment analysis (GSEA), are designed to find differentially expressed set of genes sharing common functions or regulations (Mootha et al., 2003). In GSEA, genes ranked based on a certain metric,
which can be simply the expression level, or more complicated ranking methods based on various statistical analyses, namely, Pearson's correlation, Euclidean distance, or signal to noise ratio (Mootha et al., 2003; Subramanian et al., 2005). A typical pathway analysis of time series experiments would analyze expression changes at different time points individually or reduce the time series to an endpoint-like metric, both of which bear an implicit assumption that data at multiple time points are independent (Calvano et al., 2005; Grigoryev et al., 2010). Lacking the recognition of the inherent correlation within time series data, this approach may miss potentially important pathways or yield biased and inconsistent results that ignore dynamic patterns and time-sensitivity.

In addition to time factor, it is recognized that molecular toxicity response is also dose-dependent (Allen, Kavlock, Kimmel, & Faustman, 1994). Transition of dominant function or pathway at different dose concentrations has been observed in previous studies, which provided extra mechanistic information beyond traditional phenotypic dose-response curves (Ahlborn et al., 2008; M. E. Andersen, III, Bermudez, Willson, & Thomas, 2008; Daston, 2008; Mezentsev & Amundson, 2011). The ability to reveal if and how the molecular toxicity response conserves or change at various dose concentrations is therefore necessary and has not received adequate study.

2.1.3. Tensor Decomposition

Tensor decomposition is a group of models that are able to fit high dimensional data. These models decompose the data into several components, each representing a stereotypical pattern of variation in the data. Within each component, the effects of factors are fixed at all levels, and the factors are independent of each other and together contribute to the overall gene expression levels profile (Kolda & Bader, 2009; Mørup,
Tensor decomposition models generalize low-dimensional decomposition methods singular value decomposition (SVD) to tensor data (Omberg et al., 2007; Ponnapalli et al., 2011). These result in two benefits. First, as the data dimensionality is reduced, it is now possible to approximate the variance of the original dataset into the sum of the contribution of different components and random background noise (Vannieuwenhoven, Vandebril, & Meerbergen, 2012). Second, this approach helps addressing the problem of high modality of toxicogenomic data (Clarke et al., 2008). It is known that genes can work in different pathways and that are regulated by different signaling network. In this type of gene profiling we can only observe the overall expression outcome instead of the contributions of individual process (Miller, Wang, & Kesidis, 2008; Ransohoff, 2005).

Tensor decomposition, as demonstrated by its successful application to blind source separation in signal processing, is especially useful in such problem (Cichocki, 2009). By decomposing the data into several components that account for different biological mechanisms, such as pathways, signaling network, we can possibly avoid false conclusion based on spurious correlation or the apparent lack of any correlation. In light of these advantages, several studies in the past have utilized different tensor decomposition models for understanding DNA replication and transcription (Omberg et al., 2007; Omberg et al., 2009), stem cell processes (Yener et al., 2008). Among different variants of tensor decomposition models, the most common ones are Parallel Factor Analysis (PARAFAC) (Rasmus Bro, 1997) and Tuckers3 (Kolda & Bader, 2009; Omberg et al., 2007). While the latter is more flexible, we are often interested in obtaining not the best-fitting models but the most parsimonious and robust ones. PARAFAC is regarded as an adequate data decomposition method to capture
interpretable results and is widely used for data mining problems (Rasmus Bro, 1997; Yener et al., 2008).

2.1.4. Gene Network Analysis

The large amount of data generated by high-throughput bioassays pose analytical challenges as individual inspection of genes for differential expression is much too labor-intensive and often neglecting the big picture, namely the overall toxic mechanism (Khatri et al., 2012). This difficulty brings forth to bioinformatic methods focusing on higher level functional groups, including gene sets analysis (Subramanian et al., 2005) or gene networks analysis (Barabasi & Oltvai, 2004). These methods are able to reveal the overall physical or functional landscape of the whole biological system, and thus more suitable for toxicity mechanism-centered research (Ideker & Krogan, 2012; North & Vulpe, 2010). Among functional toxicogenomics tool, gene co-expression network (GCN), also known as gene correlation network, has enjoyed a wide use for the past decade (Stuart, Segal, Koller, & Kim, 2003). GCN is normally represented as a mathematical graph, where genes are depicted as nodes/vertices, and a significant co-expression relationship between genes considering a series of expression measurements is denoted as an edge (Roy, Bhattacharyya, & Kalita, 2014; Usadel et al., 2009). It bases on the “guilt-by-association” principle to infer functional relationships among genes, which are informative since past studies have already shown that co-expressed genes tend to be functionally connected (Oldham, Horvath, & Geschwind, 2006; Wolfe, Kohane, & Butte, 2005). On the hand other, another type of gene network, gene regulatory network (GRN), is a directed graph whose edges represent directed biochemical processes, such as reaction, activation or inhibition. Compared to GRN, GCN is simpler in structure and
does not attempt to draw causal relationship among genes (Carter, Brechbühler, Griffin, & Bond, 2004). Generally speaking, inferring GCN requires fewer experimental samples, puts less constrains to experiment designs and provides more robust results (Davidson & Levin, 2005; Karlebach & Shamir, 2008; Roy et al., 2014). The application of GCN includes understand cellular responses to external stresses in different organisms in the past studies (Dewey et al., 2011; Zheng & Zhao, 2013).

Nonetheless, much of the GCN methodology has been developed for steady state experiment type, namely a snapshot of the transcriptional system, much less literature is produced about time-course data (Zoppoli et al., 2010). This is partially due to the lack of continuous data collection capability because most of the assays are destructive and can only obtain measurements at a single time point. It is speculated that such type of experiment design overlooks the inherently dynamic nature of gene interaction, which could be extensively rewired facing different conditions (Bandyopadhyay et al., 2010; Califano, 2011; Mitra, Carvunis, Ramesh, & Ideker, 2013). In recent years, there are an increasing number of time-course experiments for gene expression profiling in different areas (Deferme et al., 2013; Geijer et al., 2012; Wen et al., 1998; Weng, Chen, Ma, Lai, & Ho, 2014). Particularly, whole cell array becomes a promising platform, as it can monitor the gene expression levels continuously in a non-destructive fashion (Elad, Lee, Belkin, & Gu, 2008; Elad, Lee, Gu, & Belkin, 2010). In the meantime, there have also been ongoing efforts to incorporate temporal information into bioinformatic analysis, including network biology (Bar-Joseph et al., 2012; Wang, Liu, Liu, Liang, & Vinciotti, 2009). Therefore, it is time now to develop dynamic gene co-expression network (DGCN) to take advantage of these progresses (Capobianco, 2012).
The readily available computation power nowadays accelerates the researches in data mining and bioinformatics, which has set stage for new development in GCN. For instance, the essential step in the co-expression network inference is to quantify the functional correlation (co-expression) relationships among genes (Bansal, Belcastro, Ambesi-Impiombato, & Di Bernardo, 2007). The most used metric (Pearson’s correlation coefficient) could only provide relatively good approximation, but is not designed to tackle the main characteristics of gene expression time series data: short duration, non-linear relations and time shift (Butte & Kohane, 1999; Ernst & Bar-Joseph, 2006; Ernst, Nau, & Bar-Joseph, 2005; Schmitt, Raab, & Stephanopoulos, 2004; Wang et al., 2009).

To design and select better temporal correlation metrics, several studies have been conducted, either from the particular discipline of bioinformatics (Riccadonna, Jurman, Visintainer, Filosi, & Furlanello, 2012; Wang et al., 2009; Yuan et al., 2011), or data mining in general (Fu, 2011; Kramer, Eden, Cash, & Kolaczyk, 2009; Morse & Patel, 2007; Song, Langfelder, & Horvath, 2012; Yin, Qi, Xu, Hung, & Song, 2014).

In addition, the establishment of a good network inference framework is not only a theoretical effort but also a practical one (B. Zhang & Horvath, 2005). It has been shown the performance of network construction can vary drastically depending on the data type and application (Elo, Jarvenpaa, Oresic, Lahesmaa, & Aittokallio, 2007). Although a plethora of GCN network inference methods have been published, it is hard to really evaluate their performances, which can be attributed to the lack of both prior knowledge and biological data (Kumari et al., 2012). However, with the fast advances of co-expression network research and the unprecedented rate of biological data accumulation, it is now possible to acquire large amount of experimental data or simulate synthetic data.
with high fidelity (Barrett et al., 2013; Obayashi et al., 2008; Schaffter, Marbach, & Floreano, 2011). Guidelines on assessing the inference efficiency are beginning to emerge (Stolovitzky, Monroe, & Califano, 2007). These progresses make it possible to systematically investigate the inference of dynamic gene co-expression network (Elo et al., 2007).

Biological network analysis has also been helped by the rapidly developing theory of complex network, which strives to uncover the governing principles for network formation and evolution (Albert & Barabasi, 2002; Strogatz, 2001). Especially, the realization that the architectural properties of biological networks share some similarities with other complex systems, such as Internet and social network, paves the way to import sophisticated analytical techniques into bioinformatics (Barabasi & Oltvai, 2004; L. Chen, Wang, & Zhang, 2009; Dorogovtsev & Mendes, 2013; Newman, 2003). Past studies have discovered the following important characteristics within different networks. First, the distribution of node degree (namely the number of nodes connected to one node) generally follows a power law, indicating only a minority of genes are highly connected in contrast to the remaining majority (Ma'ayan, 2009; van Noort, Snel, & Huynen, 2004). Previous, it has been found that such highly connected nodes (also known as hub) are associated with key cellular functionality (Jeong, Mason, Barabasi, & Oltvai, 2001). These findings thus lead researchers in search of hub genes in various networks for biomarkers (Eschrich et al., 2009; Langfelder, Mischel, & Horvath, 2013). Second, it has been shown that cellular organizations are highly modular and cellular networks are highly clustering (Barabasi & Oltvai, 2004; Mitra et al., 2013). Modules consisting of mutually connected genes are considered the functional building blocks of the cell (Alon,
Therefore, module detection from co-expression network becomes an integrated step to elucidate the underlying pathways in toxicogenomics (AbdulHameed et al., 2014; Davis et al., 2015; Williams et al., 2011). Third, the architecture and topological properties of biological networks could undergo massive changes in response to external changes (Bandyopadhyay et al., 2010; Califano, 2011). In other words, it is not only the differential expressions of individual genes but also the differential interactions among them that account for cellular adaption of outer environment. A new branch of network analysis, called differential network biology, has been focusing on developing statistical approaches to identify and test those differential interactions, as well as finding their correlation with different conditions including toxicant, dose and time point (Gill, Datta, & Datta, 2010; Ideker & Krogan, 2012). This is especially important for toxicogenomics as many of its applications, such as biomarker discovery and dose-response study, rely heavily on finding the relationship between molecular activity and experimental conditions (Boverhof & Zacharewski, 2006). On the other hand, looking for those conserved connections across species and/or conditions can point us to some common responsive strategies adopted during evolution (Minguez & Dopazo, 2011; Oldham et al., 2006; Stuart et al., 2003). The abovementioned understandings of biological networks form the foundations upon which several analytical tools and frameworks have been built (Carter et al., 2004; Tesson, Breitling, & Jansen, 2010; B. Zhang & Horvath, 2005). However, thorough evaluation of network properties in dynamic toxicological responses is still lacking in toxicogenomics. Applied well, they may bring about new insights in temporal co-expression relations. In particular, differential network analysis is suitable
for studying the dynamic changes of co-expression network structure along time, and to our best knowledge, whose application has not be seriously studied.

Besides directly used as a standalone procedure, network analysis can also be integrated into the workflows of the other functional analysis tools. For instance, gene co-expression network can work as a data transformation tool, mapping genome-wide high dimensional expression data to lower-dimensional connection data (Y. Q. Chen et al., 2008; Plaisier et al., 2009). Also, the significances of individual gene can be represented by different centrality metrics bear various meaning in network theory and can be used as input for tools requiring a ranked gene list, like gene set enrichment analysis (GSEA) (Subramanian et al., 2005).

2.2. Bioinformatics Software Development

2.2.1. Software for Data Normalization and Processing

High-throughput techniques in toxicogenomics generate a large amount of data with varying quality depending on the platforms and instruments. Generally, the raw data undergo a normalization process to reduce the systematic error due to experimental machinery. For example, for whole cell array, the measurements could suffer row, column or edge effects due to the evenness of temperature across the microplates (Qu, 2011). Such error could be reduced by proper experimental design in combination with statistical methods (Malo, Hanley, Cerquozzi, Pelletier, & Nadon, 2006). In the past, several software have provided solutions for data normalization and processing (Aichaoui et al., 2012; Boyer et al., 2010), but the overall results still depends on good practices in quality control (Shun, Lazo, Sharlow, & Johnston, 2011).
2.2.2. Software for Experimental Design and Data Analysis

Time series experiment is the proper design to study and model dynamic biological processes, such as developmental systems or stress responses (Bar-Joseph et al., 2012). In transcriptomics, there are quite abundant time series gene expression data from whole cell array, a non-destructive platform suitable for long time monitoring (Zaslaver et al., 2006). However, as research deepens, the experimental design becomes more complex as well, involving a range of external factors and rapidly increasing amount of measurement data. In practice, we find experimenters are in need of a platform to manage the information for further processing and analysis. Furthermore, data analysis for whole cell array is a highly interactive process, considering the wide range of organisms/strains used and the lack of definitive solutions. Several common activities include visually inspecting the data under different experimental conditions to form plausible hypotheses, and refining normalization methods to reduce possible artefacts and system errors (Shun et al., 2011). Several prior studies proposed solutions addressing different issues in whole cell array data analysis (Aichaoui et al., 2012; Boyer et al., 2010). But those software need to be individually managed which further increases the burden for both experimenters and bioinformaticians.

2.3. Conclusion

As the throughput of toxicogenomics techniques keep increasing, studies in the field are more and more data-intensive. The role of bioinformatics techniques becomes critical in every steps of the entire workflow from experimental design to data analysis and visualization. In the meanwhile, since powerful computational resources are more readily available now, the adaptations of advanced data mining algorithms into toxicogenomics
are ever-accelerating. In this chapter, we reviewed the application of toxicants clustering, pathway analysis, tensor decomposition and gene network analysis. Employed properly, these algorithms would facilitate the quantification of molecular effects under toxicant influences, as well as the elucidation of crucial biological pathways with toxicological significance. However, as research deepens and experimental designs grow more complex, some of the conventional data analysis methods turn out to be incompatible with the emerging high dimensional time series data type. As a result, important experimental factors are neglected and the analysis is no longer reliable and accurate. It is highly desirable to update those key bioinformatics tools in order to meet the requirements of sophisticated objectives in the toxicological investigations. Similarly, bioinformatics software development, an area that was helpful to the development and implementation of high-throughput techniques in the past, needs further improvement to meet the urgent analytical needs for various purposes, especially data normalization, processing, analysis and visualization. And all the challenges and opportunities discussed so far are the topics of this dissertation.
Chapter 3

TimeVis: Software for Whole Cell Array Data

Management, Analysis and Visualization
3.1. Introduction

Time series experiment is the proper design to study and model dynamic biological processes, such as developmental systems or stress responses (Bar-Joseph et al., 2012). In transcriptomics, there are quite abundant time series gene expression data from whole cell array, a non-destructive platform suitable for long time monitoring (Zaslaver et al., 2006). However, as research deepens, the experimental design becomes more complex as well, involving a range of external factors and rapidly increasing amount of measurement data. In practice, we find experimenters are in need of a platform to manage the information for further processing and analysis. Furthermore, data analysis for whole cell array is a highly interactive process, considering the wide range of organisms/strains used and the lack of definitive solutions. Several common activities include visually inspecting the data under different experimental conditions to form plausible hypotheses, and refining normalization methods to reduce possible artefacts and system errors (Shun et al., 2011). Several prior studies proposed solutions addressing different issues in whole cell array data analysis (Aichaoui et al., 2012; Boyer et al., 2010). But those software need to be individually managed which further increases the burden for both experimenters and bioinformaticians.

In this study, we present TimeVis, a data analysis application with both usability and extensibility in mind. It provides a user friendly interface to manage experimental information based on three-layer data organization model. A visualization tool enables user to query and plot the time series data quickly, with the functionality analogous to that of an oscilloscope to electronic signals. In the end, through a plugin system and
RESTful API, developer can access user input seamlessly and supply analytical solution in the future.

### 3.2. Materials and Methods

TimeVis is open-source software published under MIT license. It can be installed in most operating systems, and its interface resides in any modern browsers. In the back end, its architecture follows an extended Model-View-Controller (MVC) pattern with a plugin module as the portal to integrate other data analysis capabilities. In the front end, multiple data driven libraries like D3.js (Data-Driven Documents) is used to interact with users and visualize time series data.

#### 3.2.1. Data Model and Database

TimeVis organizes whole cell array data at 3 levels: experiment layer stores data regarding research objectives, namely types of experimental factors and instrumental measurements, layout layer records each factor’s level at every single well on the plates, plate layer keeps all the experimental measurements. As the data model is user-oriented and not necessarily in the same representations as the real tables in the database, an outer abstract layer built upon python package SQLAlchemy is used to map data model and database schema. Embedded inside is an implementation of SQLite based database to maximize portability, as it is included in Python implementation. However, it can be easily switched to other RDBMS through a database migration utility included in the package when the data volume is larger and requires a more sophisticated database.

#### 3.2.2. User Interface and Visualization

The application is user friendly as it provides the common online form and spreadsheet-like interfaces for data management. Uploading data in common file format
such as GMX or GMT is also supported. For data visualization, TimeVis transforms SQL queries to selection on multiple drop-down lists, and delivers the plotting outcome using flexible visualization tools.

![User interfaces in TimeVis](image)

**Figure 3-1.** User interfaces in TimeVis (a) Experiment information interface where user defines independent and dependent variables regarding the overall hypothesis of a specific study. (b) Time series visualization interface, where user query data under different conditions and plot them for comparison.

### 3.2.3. Data Analysis Capability

TimeVis provides 3 ways for developer to access the user input and extend the data analysis functionality. The first is to write a plugin for the application which can be automatically detected and used by the user through web interface. The second is to communicate with the application through its RESTful API which is well documented with plenty of examples. The third is to incorporate the TimeVis package in one’s own data analysis application.
3.3. Conclusion

We present TimeVis, an application that facilitates whole cell array experimental design, as well as data organization, visualization and analysis. TimeVis provides a three-layer data organization model that defines whole cell array researches in the perspectives of biological hypothesis, external factors and instrumental measurement. Through a user friendly web interface, large amount of data regarding all aspects of a study, especially the time series gene expression data, can be properly managed and stored. With a time series visualization tool, TimeVis delivers flexible data query and visual inspection capacity. A built-in plugin system makes it easy for bioinformaticians to extend the existing data analysis functionality per users’ request. To address the needs for rapid and iterative analytics development, TimeVis can also be integrated into other applications as a standalone Python package or through RESTful APIs.
Chapter 4

Analyzing High dimensional Toxicogenomic Data Using Consensus Clustering
4.1. Introduction

In order to discover biologically meaningful patterns and relationships, and to identify biological similarities in the responses to different toxicants from the complex high dimensional omics datasets, computational data mining and statistical analysis techniques have been applied (Harper & Pickett, 2006). The most frequently used methods are unsupervised clustering algorithms such as hierarchical clustering and self-organizing maps (SOM) (Afshari et al., 2011; Daxin et al., 2004). These clustering algorithms partition the input data into non-overlapping subsets, which represent the underlying biological similarities (D’Haeseleer, 2005; Jain & Dubes, 1988). Dendrogram visualization of hierarchical clustering has been widely applied in biological studies, however, the hierarchical clustering results can be sensitive to outliers, and the visual representation neither conveys the number of distinct biological clusters, nor the clear relationships between those clusters (Bar-Joseph et al., 2003; Inglese et al., 2006; Tamayo et al., 1999). Clustering by SOM is also conventionally used for discovering groups of co-expressed genes or classifying chemicals (Törönen et al., 1999). SOM is relatively resistant to noisy data and missing observations, both common in biological research (Mangiameli et al., 1996). However, SOM algorithm can also be sensitive to the random process of map initialization, which makes comparisons between studies challenging (Halkidi et al., 2001).

These challenges in analyzing HTS toxicogenomic data point to the pressing need of a method for validation of clustering results and assessment of cluster robustness. Current validation schemes include internal approaches, which are based solely on analysis within a dataset, and external approaches, which compare clusters with known class labels.
(Datta & Datta, 2006; Gibbons & Roth, 2002). It is argued that the internal measures might not be suitable for biological data which are subject to high noise levels (Datta & Datta, 2003). To address clustering stability and validation, Monti et al. developed the consensus clustering (CC) method, which iteratively re-samples the input dataset and invokes a conventional clustering algorithm over the reconstructed dataset (Monti et al., 2003). Afterward, CC aggregates the ensemble of clustering outputs into a single consensus result, which has a higher confidence level than a typical single invocation of the underlying clustering algorithm. Over the past few years, several researchers have applied this approach in microarray-based gene expression research, and found that CC could identify reliable clusters and hidden patterns, and therefore provide a highly meaningful interpretation of the biological response (Nguyen et al., 2009; Seiler et al., 2010; Wilkerson & Hayes, 2010). Evaluation of the CC methods using a relatively large toxicogenomic dataset with temporal resolution has not yet been reported.

In this study, we employed CC to analyze a set of high dimensional transcriptomic data set with temporal resolution, which was generated using our E. coli whole cell array HTS system treated with a diverse variety of toxicants including endocrine disrupting chemicals, heavy metals, antibiotics, and nanomaterials, representing a range of possible toxicity mechanisms (Gou et al., 2010). This HTS system contains a library of 91 GFP reporter genes representing 10 cellular stress response pathways, and it records real time transcriptional level responses (Gou & Gu, 2011; Gou et al., 2010; Onnis-Hayden et al., 2009). The diversity of chemical treatments is intended to explore a wide region of the reporter state space. For the CC experiments, we focused on SOM as the underlying clustering algorithm. We found that CC/SOM reliably differentiated distinct biological
responses between various toxicants at different doses, separating toxicants with distinct toxic mechanisms and uniting diverse toxicants that cause biologically similar responses. These observations support the hypothesis that robust ensemble clustering algorithms can produce meaningful results for chemical classification based on the toxic mechanisms of toxicants (Afshari et al., 2011). In addition, the CC analysis allowed us to evaluate the cluster robustness and sensitivity to a number of conditions that represent the common variations in HTS experiments, including noisy data, subsets of treatments, subsets of reporter genes and subsets of time points. The outcome of these experiments helps identify the key data features that impact the clustering results and, provides information and insights for standardizing the practice in this field.

4.2. Materials and Methods

4.2.1. Data Generation

The whole cell array used for HTS toxicity tests was constructed through transcriptional fusions of green fluorescent protein (GFP) to 91 stress response-related promoter genes in *E. coli* K12, MG1655 (Onnis-Hayden et al., 2009; Tina K. Van Dyk et al., 2001; Zaslaver et al., 2006). The selected genes cover a variety of genes involved in different known cellular stress response pathways, such as general stress, DNA damage, protein stress, redox stress, energy stress, heat shock, drug resistance, detoxification, cell killing and other functions (Gou & Gu, 2011; Gou et al., 2010). Each fusion was expressed from a low-copy plasmid, pUA66 or pUA139 that contains a kanamycin resistance gene and a fast folding *gfpmut2*, allowing for real-time measurement of the promoter activities (Onnis-Hayden et al., 2009).
The 11 chemicals tested included 4 model chemicals with known toxic mechanisms, such as mitomycin C (MMC), mercury (Hg), hydrogen peroxide (H$_2$O$_2$), and 4-nonylphenol (4NNP). The remaining are different nanomaterials, including carbon black (CB), nano silver particles (nAg), nano titanium dioxide rutile (TiO$_2$r), nano titanium dioxide anatase (TiO$_2$a), fullerene soot (F), single-walled nanotube (SWNT) and oxidized single-walled nanotube (SWNT_OX). Detailed characterization information for these nanomaterials is listed in Appendix Table A-1 and A-2. The nanomaterials were prepared in M9 medium with 1% of crude bovine serum albumin (BSA) as a dispersant (Gou et al., 2010). The stock solutions were dispersed in a 90W sonicator for at least 15 minutes to increase the homogeneity of the mixture. Other chemicals were dissolved in de-ionized water before application. For each chemical, 3-5 different dose concentrations were tested (See Table A-1).

The protocol to measure the temporal gene expression profile was described in our previous reports (Gou & Gu, 2011; Gou et al., 2010; Onnis-Hayden et al., 2009). In brief, the *E. coli* was cultivated in 96-well plates (Costar, Bethesda, MD, USA) in dark condition to avoid GFP photobleaching until exponential growth stage (OD$_{660}$ ~0.1) was reached. The toxicants at specific concentrations were added into the micro-plate wells, and the plate was placed into a micro-plate reader (Synergy Multi-Mode, Biotek, Winooski, VT) for simultaneous cell growth (absorbance, OD$_{660}$) measurement and fluorescent readings (GFP level, EX 485nm, EM 528nm). Measurements were taken every 3 minutes over 1 hour 45 minutes, with a total of 36 time points for each experiment. Two biological replicate experiments were performed for each treatment condition. Data for Hg, MMC, nAg and TiO$_2$ were reported previously (Gou & Gu, 2011;
Gou et al., 2010; Onnis-Hayden et al., 2009) and all other data were used for the first time in this study.

4.2.2. Data Preprocessing

The GFP and OD data were first corrected for background (medium control without bacteria, and bacteria control with promoter-less strains). The gene expression level was calculated as $P = \frac{GFP}{OD}$. Induction factor was calculated as the ratio of expression level between experiment groups and control groups (GFP-fused bacteria without toxicants added), $I = \frac{P_e}{P_c} = \frac{(GFP/OD)_{experiment}}{(GFP/OD)_{control}}$. The induction factor measures the gene expression alteration as a result of toxic effect of different toxicants. The natural log of the induction factor $\ln(I)$ was then calculated: when a certain gene is up-regulated, $\ln(I) > 0$; and when it is down-regulated, $\ln(I) < 0$ (Gibbons & Roth, 2002; Tina K. Van Dyk et al., 2001). To remove the signals below a noise floor, we filtered the inductor factor values between 0.67 and 1.5 (-0.4 < $\ln(I)$ < 0.4), and changed them to 0. In previous studies, we defined a toxicity endpoint-transcriptional effect level index (TELI) for time-series transcriptomic data, which is calculated with the altered gene expression effect level,

$$
\text{TELI}(\text{gene } p) = \int_0^T e^{\ln(I)} - 1 \, dt \quad \text{Exposure} \quad \text{Time}(t),
$$

where $I$ is the induction factor of the $p^{th}$ gene (Gou & Gu, 2011).

To construct the dataset to be analyzed, we defined a “treatment” as the gene expression time-series data corresponding to a specific chemical at a specific concentration, $T$. Each treatment is a 3276 ($=36 \times 91$) dimensional input row vector,

$$
T_n = (\ln(I_{1,1}^{(n)}), \ln(I_{1,2}^{(n)}), ... \ln(I_{1,36}^{(n)}), ... \ln(I_{p,q}^{(n)}), ... \ln(I_{91,1}^{(n)}), \ln(I_{91,2}^{(n)}), ... \ln(I_{91,36}^{(n)}))
$$

(1)
where, $\ln(I_{p,q}^{(n)})$ is the expression of the $p^{th}$ gene at the $q^{th}$ time point, under the $n^{th}$ treatment ($p = 1, 2, \ldots 91$, $q = 1, 2, \ldots 36$, and $n = 1, 2, \ldots 40$). Since the control results were involved in the pre-processing, we define the control vector with all elements being 0. Then we define a “sample” as the whole or part of the overall dataset that contained a total of 40 treatments.

To test the relevance of temporal pattern of the gene expression alternation on the clustering analysis, we compared the results of the full time-series with those using one selected time point data, and with those using integrated time-series data represented by the aggregated TELI value.

4.2.3. Ensemble Clustering and Data Validation: Consensus Clustering (CC)

Consensus clustering works through iteratively re-sampling and clustering the input data to generate a similarity matrix (consensus matrix) of sample, which can be used to predict the number of clusters and to assess cluster stability (Kim & Lee, 2007). The underlying assumption of CC is that high cluster stability, produced from perturbed input data, indicates high confidence in the resulting clustering. The stability of the clustering can be evaluated from the consensus matrix, where each entry is a consensus index (CI), which are calculated as

$$CI(i,j) = \frac{\sum_h M_{(h)}(i,j)}{\sum_h I_{(h)}(i,j)},$$

where $M_{(h)}(i,j)$ and $I_{(h)}(i,j)$ are the entries of connectivity matrix and indicator matrix in the $h^{th}$ re-sampling cycle, respectively. $M_{(h)}(i,j)$ is defined to be 1 if the $i^{th}$ and $j^{th}$
treatments are clustered together, and 0 otherwise; \( I_{(h)}(i,j) \) is defined to be 1 if the \( i^{th} \) and \( j^{th} \) treatments both appear in \( h^{th} \) re-sampling. The positive index, \( CI(i,j) \), is an indicator of similarity between treatments, and is within the range of \([0,1]\) following the above formula. If a consensus matrix consists largely of 1’s and 0’s, one may infer that the sample is well clustered, because the treatments are similar within their own clusters and distinct from those outside. The algorithm of consensus clustering was implemented in MATLAB (version R2011a).

To conduct a CC analysis, a re-sampling scheme and a basic existing clustering algorithm are first chosen. SOM, which has been effectively used for the exploratory analysis of gene expression data, was used as the underlying clustering algorithm for this study. A SOM is a special case of neural network consisting of nodes organized on a regular typically two-dimensional grid (Vesanto, Himberg, Alhoniemi, & Parhankangas, 1999). The high dimensional sample is projected onto the maps and clustered by proximity. For the re-sampling algorithm, we chose the bootstrapping with 1000 iterations, as this maintains the size of the original dataset, which is necessary for stable SOM performance. Finally, a dendrogram based on average linkage hierarchical clustering, whose distance function is I-CI, was generated to show the inner structure of the blocks along the main diagonal in the heatmap. The SOM algorithm in this paper, including map creation, initialization, training, finding best-matching units for treatments on the map using U-matrix, are implemented through using SOM toolbox, a function package for MATLAB (version R2011a). The SOM topology was optimized by SOM toolkit (version 2.0) with a two-dimensional hexagonal 8×4 grid (32 units) (Vesanto et al., 1999). All other parameters were at the default settings.
4.3. Results and Discussion

4.3.1. Consensus Clustering Based on SOM

SOM are frequently used to reveal patterns in gene expression datasets, in part, for their computational scalability to large datasets (Tamayo et al., 1999). By partitioning the input data into a small set of nodes, and by representing the nodes in two dimensions, the SOM intuitively conveys the notion of a similarity neighborhood between nodes. Moreover, it is straightforward to aggregate neighboring nodes into higher-level clusters and represent these aggregate clusters as distinct colors.

We first applied SOM to our full time-series dataset and found four distinct clusters, which are shown in Figure 4.1a. The resulting SOM patterns are mostly consistent with prior toxicological knowledge. The treatments of toxicants at their lowest dose concentrations aggregated closely around the untreated control experiments, indicating their molecular effects are close to detection limit. As dose concentrations increased, chemical-specific transcriptional level effects became more pronounced. For the same chemical, treatments with moderate and high dose levels tend to cluster closely together, as observed for SWNT (0.32 and 8 mg/L), TiO$_2$-a (10 and 50 mg/L), 4NNP (1 and 15 mg/L), Hg (0.01 and 0.05 mg/L), MMC (0.03 and 0.3 mg/L), H$_2$O$_2$ (1 and 5 mg/L) and fullerene (1 and 10 mg/L), indicating that there is conserved similarity in the response patterns for a given chemical at different dose levels. The four model chemicals with different known toxic mechanisms exhibited distinctive profiles and were separated into different clusters. MMC is a model genotoxicant that specifically leads to DNA damage (Khil & Camerini-Otero, 2002; Onnis-Hayden et al., 2009). H$_2$O$_2$ is a model oxidant that causes oxidative stress via reactive oxygen species (ROS) and ultimately lead to various
damages to cell (Cantoni, Brandi, Salvaggio, & Cattabeni, 1989). Examination of the states of various stress genes of *E. coli* in exposure to mercury suggests that it leads to redox stress and DNA damage in the cells [36]. Protein stress and toxicity of 4NNP to *E. coli* strain has been reported (Soares, Guieysse, Jefferson, Cartmell, & Lester, 2008; T. K. Van Dyk et al., 1995). The nanomaterials frequently clustered together and their distances to those model compounds indicate their toxic effects. Most nanomaterials clustered with H2O2, suggesting their dominant oxidative stress-related toxicity, which is consistent with literature reports (Bello, Hsieh, Schmidt, & Rogers, 2009; Gou & Gu, 2011; Gou et al., 2010; Reddy, Reddy, Himabindu, & Krishna, 2011).
Figure 4-1. SOM and consensus clustering results for all treatments: (a) Self organizing map (SOM). Treatments are projected onto the map. The differences of gene expression pattern among the treatments are transferred into topographical distance on the map. Hexagon colors represent clusters, which are determined according to treatment distances using the U-matrix. Average CI values shown for each cluster are arithmetic mean among all pairs of treatments within the cluster; (b) Consensus clustering (CC). The consensus matrix represented as a heatmap. A dendrogram is based on the CI’s, showing the inner structure in the blocks along the main diagonal. The identified clusters are labeled in different colors in the dendrogram, and dashed square in the heatmap. Concentration unit is mg/L.

To quantitatively evaluate the quality and consistency of the resulting SOM, we employed CC, which iteratively re-samples the input dataset to re-compute new SOMs, and measures the consistency of the resulting clusters. For a given pair of gene expression inputs, the consensus index (CI) represents the relative frequency that these inputs co-clustered together. A CI value of 1 indicates that the two inputs always co-clustered, while a CI element of 0 indicates that the inputs never co-clustered. The
theoretically ideal result occurs when all elements of the consensus matrix are either 1 or 0; however, in practice, re-sampling and re-clustering produces intermediate consensus values between 1 and 0. As shown in Figure 4-1a, in our study, 4 distinct clusters are found and the average CI values (arithmetic mean among all pairs of treatments within the cluster) vary widely, ranging from 0.39 to 0.87. This large variance underscores the sensitivity of SOM to perturbations of the input data, and illustrates one difficulty in comparing results between high-throughput screening toxicological studies. In addition to the stochastic effects of re-sampling, some of this variance may be explained by the inherent SOM random ordering of its input processing, coupled with inherent SOM sensitivity to initial conditions (de Bód, Cottrell, & Verleysen, 2002). The CI matrix, by indicating the frequencies of co-clustered input pairs, can also be interpreted as a matrix of pairwise similarity scores. Since two highly similar input vectors likely cluster together frequently, their CI score tends towards 1. By the same logic, two highly dissimilar input vectors would co-cluster and their CI score tends towards 0. The clustering results using this CC approach are illustrated as consensus matrix (Figure 4-1b). The second-level clustering reordered the rows and columns identically, such that blocks along the diagonal indicate high confidence consensus clusters. This approach revealed three distinctive clusters, labeled in different colors. The optimal cluster number was confirmed by finding the peak cumulative distribution function area as proposed by Monti et al. (Monti et al., 2003) (Appendix Figure A-1a). The average CI value for each cluster ranged from 0.56 to 1, which was higher than those with SOM (0.39-0.87), indicating that higher statistical robustness of the CC results. In addition, there are gradations in CI within each cluster. This subtlety may be lost in conventional single-pass
SOM clustering, in which neighboring nodes are clustered based on the aggregate similarities of the representative node vectors (Figure 4-1a). In addition to the improvements in reliability, resolution and visual presentation of the clustering results by CC approach compared to SOM results, the chemical-specific clusters are more consistent among chemicals with CC than the SOM results, and they appear in better agreement with prior understanding of the toxic mechanisms of these toxicants. For example, all SWNT treatments (except the one at lowest concentration) clustered as one group, which is consistent with the current knowledge that carbon nano tubes cause toxicity mainly via oxidative damage due to their strong ability to mediate electron transport and generate reactive oxygen species (ROS) (Vecitis, Zodrow, Kang, & Elimelech, 2010). More impressively, the CI revealed more clear distinction between original SWNT versus those that had surface modification-SWNT_OX. The surface characterization changes in the SWNT_OX resulted in the subtle changes in its toxic response profiles and this phenomena has been reported by toxicological studies (Liu et al., 2010; Yang, Mamouni, Tang, & Yang, 2010). Two nano-titanium oxides were tested, namely TiO$_2$-rutile (TiO$_2$-r) and TiO$_2$-anatase (TiO$_2$-a). TiO$_2$-a is known to be more toxic than TiO$_2$-r and they were found to lead to both oxidativs stress and DNA damage(Gou & Gu, 2011). Indeed, the clustering results showed that TiO$_2$ clustered more closer with both DNA-damager MMC and oxidant H$_2$O$_2$. TiO$_2$-r has denser arrangement of atoms and higher stability, which may explain its apparent lower toxicity at similar concentrations compared to TiO$_2$-a at transcriptional effect level, especially in nanosize (Jin et al., 2011). nAg, a metal nanoparticle clustered with toxic metal Hg at higher dose concentrations, indicating likely metal-specific toxic response and both of which are
known to lead to oxidative stress, DNA damage and protein stress (Gou & Gu, 2011). Consistent with the SOM results, the toxicants at the lowest dose levels (near detection limit) clustered most closely with controls as expected and the treatments for the same chemical at different dose concentrations all cluster together. The chemical-specific response profiles became more distinctive as the dose concentration increased. Overall, the results demonstrated that the stress-response pathways ensemble-based HTS toxicity assays yield chemical-specific and concentration-sensitive transcriptomic profiles. Moreover, statistically reliable clustering such as CC based on SOM algorithm is capable of identifying classes of chemicals according to their underlying toxic mechanisms.

4.3.2. The Impact of High Dimensional Nature of the Data: the Resolution Power from Time-Series

The datasets analyzed above take advantage of the full offset of measurement time points, which presumably provides substantial information that helps distinguish the treatment clusters. To test that hypothesis, we performed three more experiments using different features extracted from the time series. The first used gene expression data at the middle time point of the testing (the 20th time point, at 57 minutes after measurement began). The second experiment employed the maximum gene expression signal observed for each reporter during the 2-hr assay (Max), and the third experiment used a derived integrated endpoint-Transcriptional Effect Level Index (TELI). TELI integrates a reporter signal over time to reflect more of the temporal pattern of the reporter response than a single time point, effectively compressing a number of time points into a compact representation (Ahn et al., 2009; Gou & Gu, 2011). These selected data features mimic three conventional experiment designs.
Figure 4-2. Average consensus index for each node in consensus clustering. Results are from four different experiments that selected different time-series data subsets, including Max- dataset containing the maximum gene expression signal observed for each reporter during the 2-hr assay; 20th- dataset that uses only gene expression data at the middle time point of the testing, the 20th time point, at 57 minutes of exposure; TELI- dataset consisting of derived integrated endpoint-Transcriptional Effect Level Index (TELI). Total represents the full dataset with all the time points.

In CC analysis, the treatments with high consensus index from different nodes in the dendrograms eventually merge into a common root. Higher average CI value indicates higher similarity among the treatment within the node. To visualize the quality of clustering with different subsets of time-series toxicogenomics data, we analyzed the consensus matrices and calculated the average consensus indices for each node and ordered them in an increasing fashion. Figure 4-2 shows the comparison of average consensus index versus ordered node number curves based on dendrograms generated with the four different data subsets described above. Lower consensus index values indicated nodes close to the root, reflecting those connecting a group of treatments with
distinct transcriptomic profiles; while higher values indicate nodes that could represent a real cluster, a group of treatments exhibiting similar toxic response profiles. A good clustering produces nodes that have a distinct jump of the average node CI value, that is, a large difference between the intra and inter cluster values. Otherwise, nodes with similar levels of average consensus index reflect a weak consensus, a sign of lower resolution power.

Compared to the dataset with all the time points, the consensus indices for datasets using peak signal or individual time points showed a dramatic deterioration in resolution between treatments (Figure 4-2, Appendix Figure A-2 and A-3). These results indicate that the temporal differences between signal peaks contribute substantial information beyond the simple peak signal level. This finding underscores the importance of careful selection of sampling times for a given experimental system. Individual time points may not provide sufficient information to fully reflect the complex and dynamic biological response. Using maximum expression as a single feature removes any dependence on timing and synchronization, but may also sacrifice the rich biological information associated with time-series generated by cell based HTS assays. The results also indicated that temporal data compression using our proposed TELI concept followed by CC distinguished several prominent treatment clusters, and performed qualitatively better than the other two datasets without temporal resolution. As expected, CC with the full dataset performed the best among all four experiments. This result does not imply that temporal toxicogenomic data with such a high resolution (every 3 minutes in our assay) is necessary in all experimental systems. Rather, these experiments demonstrate the
sensitivity of clustering to sampling frequency and time points, and the importance of gene response dynamics in the interpretation of omics data.

### 4.3.3. Consistency Test for Consensus Clustering

In order to evaluate the sensitivity of the clustering results towards the variations in experimental design such as the number of treatments, we performed SOM-based CC to analyze the nanomaterials subset of our data (24 out of total 40 treatments). If the subset clusters similarly to the full dataset clustering, that would support the hypothesis that CC is resistant to treatment numbers and data size, and that clustering results can be meaningful.

![Dendrogram produced from consensus clustering of nanomaterials.](image)

**Figure 4-3.** Dendrogram produced from consensus clustering of nanomaterials.

The nanomaterial subset clustering results are shown in Figure 4-3 (the curve for determining the optimal cluster number is shown in Appendix Figure A-2b). The results are largely consistent with clustering of the full dataset. The low concentration treatments
again clustered near the untreated control. Interestingly, the carbon nanotube treatments, SWNT and SWNT_OX, remain distinct from the remaining toxicants and the surface-oxidized nanotubes were distinguishable from their original as discussed previously. Although consisting of the same carbon element, the distinguishable toxic effects among the three carbon-based nanomaterials evaluated (carbon nanotube, fullerene and carbon black), are suggestive of their nanostructure-dependent toxicological implications (Vecitis et al., 2010). Moreover, these results showed that perturbing the input dataset by taking a subset of treatments continued to produce relatively stable clusters, which indicate plausible classes of biological responses.

**Figure 4-4.** Consensus clustering result for dataset perturbed with additional noise: the noise is Gaussian with mean value of 0 and standard deviation of 0.5.

To further test the susceptibility of the clustering results to data noise level, we assessed the consistency of CC by adding computer-generated noise to the original dataset using MATLAB function (normrnd). This experiment evaluated the robustness of CC to the variance caused by multiple laboratories performing similar experiments but
lacking standardized protocols. Normally distributed random noise was generated with mean ($\mu$) value of zero and a range of standard deviation ($\sigma$) values. The results showed that the CC could still produce relatively consistent clustering result without significantly losing resolution power at $\sigma=0.5$, as shown in Figure 4-4. Although the differentiating quality was reduced as shown by the somewhat blurred clustering boundaries, the number of clusters and aggregates of chemicals were largely conserved. This result demonstrates that the CC analysis is relatively reliable and therefore may be applicable for clustering analysis with data generated from different laboratories if the noise level is not excessive.

4.3.4. Impact of Gene Selection: Consensus Clustering Using Reordered or Subset of Total Genes

![Consensus matrix](image)

**Figure 4-5.** Consensus matrix based on the clustering result with dynamic genes library: the scheme of re-sampling is changed to reconstruct the treatment vectors using genes randomly selected from the original library. With this perturbed input, the pattern differences between treatments change accordingly, which in turn impact the resolving power of the clustering algorithm.
To assess the sensitivity of the clustering results to the selection of gene reporters, we performed a CC perturbation experiment that performs bootstrap re-sampling of the 91 gene reporters while maintaining the full set of treatment conditions. The resulting consensus matrix, shown in Figure 4-5, differed from the previous results in having broadly higher average CI values among non-clustered treatments. Individual clusters were less apparent, blurring the earlier strong distinctions between treatment classes. This loss of clustering resolution suggests that the noise introduced by gene re-sampling allowed only strong, common responses between treatments to cluster, while obscuring the more subtle distinctions between treatments. Choosing an appropriate and sufficient set of gene reporters remains an open research question (Saeys, Inza, & Larrañaga, 2007). An ideal gene reporter set would clearly discriminate between a wide variety of treatment conditions, convey large changes in signal level, have low variance as well as a low noise floor, and carry little redundancy to minimize experimental time and cost.

To further explore sensitivity to gene selection, we performed CC using a subset of reporters involved in redox stress. The result, shown in Appendix Figure A-4, also shows a large but diffused cluster. The results from these two gene perturbation experiments suggest that the particular set of 91 reporters employed here cannot be greatly reduced while maintaining the resolution needed to distinguish between the treatments tested here. This conclusion, however, does not exclude the possibility that a different and smaller set of reporters might have the resolution necessary to clearly identify separate clusters of treatments. Nor does this result preclude using a larger set of reporters to discriminate between wider selections of toxicants. Of course, optimization in the selection of
pathways and genes to gain sufficient resolution power and with minimal redundancy is yet another challenge that is beyond the scope of this study.

4.4. Conclusion

CC serves both as a validation algorithm for conventional clustering, as well as an ensemble clustering approach. This study has shown that CC performs well in reliability for our three-dimensional HTS toxicological data. These results support the hypothesis that, within reasonable limits, CC facilitates cluster comparisons between experiments with differing designs and variations. We also found that for our cellular stress response ensemble-based HTS transcriptomics assay platform, the size and composition of the reporter gene set are critical factors that affect the resulting coherency of clusters. These findings suggested a relatively low level of redundancy within the set of 91 reporters employed here to represent the stress-response pathway ensemble. We also demonstrated the value and importance of utilizing rich time-series data, which produced the highest level of cluster resolution. In summary, this study has demonstrated the value of CC, and illustrated its potential usefulness to analyze high dimensional toxicogenomic databases.
Chapter 5

Toxicity Mechanisms Identification via Gene Set Enrichment Analysis of Time-Series Toxicogenomics Data: Impact of Time and Concentration
5.1. Introduction

Pathway analysis is one family of bioinformatic tools for toxicity mechanisms elucidation, which aims at pinpointing key functional gene groups and regulatory pathways evoked during the toxicant exposure under a given condition (D. W. Huang et al., 2009; Khatri et al., 2012). Through shifting the focus from detecting differentially expressed genes individually to discerning sets of genes that share common biological function or regulation, pathway analysis catches the expression patterns on the higher pathway level, avoids results misinterpretation due to subjective expression thresholds for individual genes, and reduces the complexity of data analysis that deals with daunting number of genes. Pathway analysis of high dimensional toxicogenomics data, such as time series data, faces great challenge, however, since most current techniques are mainly designed for the analysis of a biological system snapshots (Khatri et al., 2012). The commonly used pathway analysis techniques, such as the gene set enrichment analysis (GSEA), are designed to find differentially expressed set of genes sharing common functions or regulations (Mootha et al., 2003; Subramanian et al., 2005). In GSEA, genes ranked based on a certain metric, which can be simply the expression level, or more complicated ranking methods based on various statistical analyses (i.e. Pearson's correlation, Euclidean distance, or signal to noise ratio) (Subramanian et al., 2005).

In addition to time factor, it is recognized that molecular toxicity response is also dose-dependent (Allen et al., 1994). Transition of dominant function or pathway at different dose concentrations has been observed in previous studies, which provided extra mechanistic information beyond traditional phenotypic dose-response curves (Ahlborn et al., 2008; M. E. Andersen et al., 2008; Daston, 2008; Mezentsev & Amundson, 2011).
The ability to reveal if and how the molecular toxicity response conserves or change at various dose concentrations is therefore necessary and has not received adequate study.

In this study, we investigated the application of an improved GSEA based on CPCA (Common Principal Component Analysis) score to identify the activation of specific stress response categories and pathways for three representative chemicals based on the temporal altered gene expression data with multiple (six) dose concentrations. Temporal altered gene expression profiles were generated using a high-throughput with gfp-fused reporters library called whole cell array consisting of genes covering all known stress response pathways in *E. coli* (Gou & Gu, 2011; Gou et al., 2010; Gou et al., 2014; Onnis-Hayden et al., 2009). Two score metrics were proposed to rank the genes that consider the temporal gene expression profile. One employs an integrated altered gene expression quantifier-TELI (transcriptional effect level index) that integrates altered gene expression magnitude over the exposure time (Gou & Gu, 2011). Another one applies common principal components analysis (CPCA) to generate scores for all the genes based on their contribution to the common temporal variation among treatments for a given chemical at different concentrations. Compound-specific stress response activation profiles indicative of toxicity mechanisms were obtained using both metrics. The results were compared to determine the suitability, as well as advantages and disadvantages of applying the time-aware metric instead of the static metric. The impact of time and dose concentration on the gene enrichment analysis results were revealed and discussed.
5.2. Materials and Methods

5.2.1. Toxicogenomics Time Series Data Generation

A high-throughput toxicogenomics assay was employed using GFP-fused stress response ensemble whole-cell library of *E. coli* K12, MG1655 (Gou et al., 2010; Onnis-Hayden et al., 2009), with each fusion expressed from a low-copy plasmid, pUA66, which contains a kanamycin resistance gene and a fast folding *gfpmut2*, allowing for real-time measurement of gene expression level changes (Onnis-Hayden et al., 2009; Zaslaver et al., 2006), which measures temporal gene expression changes of 106 genes involved in various cellular stress response pathways that are known to be highly conserved among species (Zaslaver et al., 2006). The selected stress response assay library covers 106 promoters (genes) a variety of genes involved in different known cellular stress response pathways known to be highly conserved among species, and they are categorized into seven groups including general stresses, protein stresses, redox stresses, cell killing, DNA damage, drug resistance and detoxification (Keseler et al., 2013) (See Appendix Table B-1 for list of genes and their pathways).

Three compounds with known toxic mechanisms were evaluated for demonstration and they are mitomycin C (MMC), hydrogen peroxide (*H*₂*O*₂), and lead nitrate (*Pb(NO₃)₂* or *Pb*²⁺). For each chemical, 6 sub-lethal concentrations were tested with 3 replicates each, resulting in a total of \(54 = 3(\text{chemicals}) \times 6(\text{concentrations}) \times 3(\text{replicates})\) treatments.

The protocol to measure the temporal gene expression profile was described in our previous reports (Gou & Gu, 2011; Onnis-Hayden et al., 2009). In brief, *E. coli* reporter strains were cultivated in 384-well micro-plates (Costar, Bethesda, MD, USA) in dark
condition to avoid GFP photo-bleaching until the early exponential growth stage (OD600~0.2) was reached. Samples of specific compounds were added into the wells, and the plate was placed into a micro-plate reader (Synergy Multi-Mode, Biotek, Winooski, VT) for simultaneous cell growth (absorbance, OD600) measurement, denoted as \( OD \), and fluorescent readings (GFP level, excitation 485nm, emission 528nm), denoted as \( GFP \). The data within first 20 minutes were volatile and prone to have large outliers, and hence were not used. Measurements were taken every 5 minutes over 1 hour 40 minutes since then, resulting in a total of 21 time points for every gene in every treatment. The time used in this study was adjusted accordingly, with the starting point as time 0.

5.2.2. Data Pre-processing

The \( GFP \) and \( OD \) data were corrected for various controls, including blank with medium control (with and without water samples) and promoterless bacterial controls (with and without water sample). The alteration in gene expression, also called induction factor \( I \), for a given gene at each time point due to sample exposure, was represented by the ratio of the cell normalized gene expression GFP level (normalized over cell concentration (OD)) in the experiments condition with water sample exposure to that in the control condition without any chemical exposure(Gou & Gu, 2011; Gou et al., 2014) The natural logarithm of the induction factor \( \ln(I) \) was then calculated for the following data analysis, where a gene is up-regulated if \( \ln(I) > 0 \) and down-regulated if \( \ln(I) < 0 \)(Tina K. Van Dyk et al., 2001).
5.2.3. Gene Set Enrichment Analysis with Two Different Ranking Metric

Two ranking methods based on transcriptional effect level index (TELI) and common principal component analysis (CPCA) were applied for gene set enrichment analysis (GSEA) and they are described as following:

1) Pathway analysis: Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a statistical procedure to determine whether a predefined set of genes is over-represented toward the top or bottom of a ranked gene list (Subramanian et al., 2005). In this study, different gene lists were generated using either CPCA score or TELI as ranking metric (see detailed description of CPCA and TELI in following sections). For a given list, the enrichment score ($ES$) for a specific gene set or pathway is calculated by walking down the ranked gene list, and increase the running-sum statistic if the gene from the pathway is encountered and decrease it otherwise. The $ES$ is the maximum deviation from 0 encountered in the walk corresponding to a weighted Kolmogorov-Smirnov-like statistic (Hollander & Wolfe, 1999). The statistical significance of the $ES$ was estimated by permutation test: the gene list is permutated 1,000 times, and then an $ES$ for each permutation was calculated to generate a null distribution for the $ES$. The $p$ value for original $ES$ was then calculated in relative to the null distribution. Because the $ES$ for multiple pathways were calculated, the multiple comparison errors were adjusted using false-discovery rate method (Subramanian et al., 2005).

2) Ranking Metric Based on Common Principal Component Analysis

We proposed to implement Common Principal Component Analysis (CPCA) for determination of ranking metric based on their contribution to the temporal variance
in altered gene expression level over exposure time (Jolliffe, 1986; Raychaudhuri, Stuart, & Altman, 2000). A treatment was defined as one specific test for one particular chemical at a given concentration. The raw data for one treatment is a matrix for $T$ time points (rows) and $G$ genes (columns). PCA uses linear orthogonal transformation to convert the genes into a new set of uncorrelated variables, called principal components (PC). The transformation is designed such that the first PC points to the direction with the highest possible variance in the data space formed by the raw data, and each succeeding PC points to the direction with highest possible variance remained while under the constraint of being orthogonal to all the preceding PC’s. Each PC is a linear combination of the original genes, namely $p_j = \sum_{i=1}^{G} l_{ij} g_i$, where $g_i$ denotes the $i^{th}$ gene in the set of $G$ genes, $p_j$ denotes the $j^{th}$ PC, and $l_{ij}$, called loading, represents the weight or contribution of the $i^{th}$ gene to the $j^{th}$ PC. Only the first several PC’s are retained for the purpose of dimension reduction and filtering out system noise, and they represent most of the variance of the raw data. We determined the number of PC’s retained by setting the threshold to be 70%, namely, the first $K$ PC’s whose sum of variances is larger than 70% of the total variance are retained. For each gene, its contribution to the total variance can be measured using its loadings on these PC’s, namely $s_i = \sum_{j=1}^{K} l_{ij}^2$, where $s_i$ is the metric of the contribution of the $i^{th}$ gene, $K$ is total PC’s retained. Since the variance discussed here indicates the temporal variation, the contribution of genes in their ranking metric was based on the assumption that higher temporal variation indicates higher transcriptional level alterations.
CPCA is a generalization of PCA to generate a set of common principal components (CPC) that agrees most closely to the data from multiple experiments. It is designed in such a way that the first CPC points to direction closest to all first PC’s from each individual treatment, and so forth. Like PC’s, each CPC is also a linear combination of the original genes, namely $q_j = \sum_{i=1}^{G} w_{ij} g_i$, where $q_j$ denotes the $j^{th}$ CPC, and $w_{ij}$, also called loading, represents the weight or contribution of the $i^{th}$ gene to the $j^{th}$ CPC.

Mathematically, CPC’s can be calculated as the eigenvectors of matrix, namely

$$H = \sum_{i=1}^{N} L_i L_i^T$$

where $N$ is the total number of treatments, $L_i$ is the matrix for the $i^{th}$ treatment containing all the loadings for the retained PC’s. Since the CPC’s present the directions of common temporal variance, each gene’s contribution to total temporal variance can be calculated using the new loadings, namely $t_i = \sum_{j=1}^{J} w_{ij}^2$, where $t_i$ is the metric indicating the contribution of the $i^{th}$ gene, $J$ is the total number of CPC’s and determined by the minimum number of PC’s retained among all treatment. In this study, the genes’ contribution to temporal variance, designated as CPCA score, was used as the ranking metric to differentiate genes with varying temporal activities.

3) Transcriptional Effect Level Index

In previous studies, we defined a toxicity endpoint-transcriptional effect level index (TELI) for quantifying gene expression data, which incorporates the number, magnitude and the cumulative temporal pattern of genes with altered expression(Gao et al., 2012; Gou et al., 2010) and is calculated as
where $i = 1, \ldots, G$ indexes one of the $G$ genes, $t = 1, \ldots, T$ indexes one of the $T$ time points, and $\ln(1) = 0$ refers to control with no altered expression magnitude. More detailed description and discussion of TELI can be found in our previous reports (Gou & Gu, 2011). In this study, we applied TELI as an alternative ranking matrix for GSEA analysis.

5.2.4. Software

The calculation of TELI and CPCA scores were implemented using MATLAB (MathWorks, MA, version 2013a). GSEA algorithm and visualization of this study is implemented using R (version 2.15.3) together with package ggplot2 (version 0.9.3).

5.3. Results and Discussion

5.3.1. Chemical-Specific and Temporally dynamic Stress Response Gene Expression Profiles

Examination of temporal variation in altered gene expression is indicative of various stress response pathway (categories) for the three chemicals at different dose concentrations elucidated chemical-specific and dose concentration-dependent patterns. Figure 5-1 displays the average temporal profile of altered gene expression level for gene assembles indicative of seven stress response categories in *E. coli* for three chemicals, namely MMC, H$_2$O$_2$ and Pb$^{2+}$, across 6 concentrations. The grey bands show the 95% confidence interval of the variances in the altered expression level among all the genes in a given pathway, with 3 replicates. The relatively narrow gray bands suggested that genes selected in each specific stress response pathway (or category) seemed to have high co-
expression tendency, which agrees with prior report of the high proportion of co-regulations of genes under environmental stresses (Chung, Bang, & Drake, 2006).

For the three chemicals tested, most stress response pathways exhibited distinct temporal patterns among the different chemicals, and yet, for a specific chemical most seemed to have conserved temporal trends among all concentrations with dose-dependent magnitude changes (Figure 5-1). The various temporal patterns for stress response pathways included a relatively constant expression level (such as cell killing pathway when exposed to MMC, Figure 5-1a); short impulses (such as detoxification pathway induced by H₂O₂, Figure 5-1b); a monotonic increasing/decreasing trend spanning over a large expression level range (such as drug resistance pathway in response to Pb²⁺, Figure 5-1c). These patterns were consistent with those observed in earlier studies, where both short impulse-like and long sustained gene expression patterns were found in response to environmental stimuli depending on the function of the specific pathway (Yosef & Regev, 2011). Although, there were some that showed changing temporal patterns as dose concentration increased, such as those in protein damage and detoxification pathways in response to MMC, which showed transitional pattern changes at higher doses. These patterns are likely reflections of cellular requirements for an immediate remedy or a gradual recovery seeing the severity of the stress, as well as balances among efforts/energy needed for homeostasis and stress responses. The chemical-specific yet conserved temporal trends among genes in a given stress response pathway suggests that these conserved patterns can be potentially chemical-specific indicators for further toxicity mechanism evaluation.
Figure 5-1. Normalized temporal variation in gene expression profiles of stress response pathways or categories upon exposure to various levels of (a) MMC, (b) H₂O₂ and (c) Pb²⁺. Each curve represents the mean temporal variations (measured as ln(I), I is the altered gene expression level) profile of all the genes (3 replicates) in the specific stress response pathway, with 95% confidence intervals indicated by the gray bands. x-axis top: various stress response pathways and categories; x-axis bottom: exposure time in hour; y-axis left: normalized temporal variations in altered gene expression level as ln(I), where I is induction factor normalized against the first time point. Pathway/Stress response categories abbreviation: General (general stresses), Redox (Redox stresses), Protein (protein stress), Cell (cell killing), DNA (DNA damage), Drug (drug resistance), Detox (detoxification).

The observed dose concentration-dependent changes in the magnitude of altered gene expression level suggest possible dose-response relationship, which is the central dogma
of toxicology (Altshuler, 1981). As cellular stress responses to environmental perturbation involve coordinating gene expression in both magnitude and timing, it is desirable to have a means to quantify how temporal pattern changes according to chemical dosage (Chechik & Koller, 2009). However, there is not yet a consensus on dose-dependent temporal patterns and molecular toxicity endpoints-based dose-response relationship (Allen et al., 1994). Several previous studies suggested the existence of relationship between dose and molecular toxicity endpoints at single gene or ensemble of stress response genes (Ahlborn et al., 2008; Allen et al., 1994; Altshuler, 1981; Burgoon & Zacharewski, 2008; Flora, Gupta, & Tiwari, 2012; Gao et al., 2012; Gou & Gu, 2011; Gou et al., 2010; Gou et al., 2014; Kærn, Elston, Blake, & Collins, 2005; Mezentsev & Amundson, 2011; Nevozhay, Adams, Murphy, Josić, & Balázsi, 2009; Peddada et al., 2003). In this study, all chemical concentrations applied were at sub-cytotoxic levels, where the temporal stress responses were likely at homeostasis stage. The dose-dependent patterns of altered gene expression profiles indicated variations in both magnitude and nature of the molecular stress response systems in response to toxicant at varying concentrations. The results also demonstrated that time-course experiments could be potentially used to delineate prototypical temporal activation and co-regulation of genes.

The above observation that aggregated expression dynamics at pathway level could be consistent, chemical-specific and concentration-dependent highlighted the suitability of pathway analysis for toxicity mechanisms studies. Focusing on the common behavior of genes from the same pathway could provide more robust and reliable results, because it reflects more cohesive biological effect and minimize misleading conclusions due to
likely errors and inconsistency in results of individual genes (Khatri et al., 2012). For example, cellular processes often involve sets of genes acting in concert rather than the significant expression of only one gene (Subramanian et al., 2005). For time series toxicogenomics data, the inconsistency in altered gene expression for individual gene among replicates may result from both inherent experimental system error and lack of proper data processing processes such as of temporal gene expression profile alignment (Aach & Church, 2001; Lin, Kaminski, & Bar-Joseph, 2008).

5.3.2. Impact of Time on GSEA for Toxicity Mechanisms Identification

Most conventional GSEA analyses have been performed with static gene expression profiling data, which captured the snapshots among the dynamic transcriptional systems with the underlying assumption of cellular states. The dynamic temporal differential expression profiles of stress response genes in response to a toxicant as shown in our study suggests that application of GSEA directly to isolated time points would neglect the important time factor and lead to inconsistent conclusions. To demonstrate the caveat, we simulated and compared the GSEA results for three single-time points using our time series data. Altered expression levels of genes after the array being exposed to MMC samples (0.5 μg/L) at 3 time points (30, 65 and 100 minutes) were isolated and used to identify enriched pathways, respectively. Figure 5-2 illustrates the statistical significances of 7 stress categories at different time calculated by GSEA. MMC is a known DNA-damage agent and often used as a model genotoxic compound in environmental toxicity studies (Onnis-Hayden et al., 2009). GSEA identified the significant enrichment of DNA damage and repair pathway, but only in the middle of the experiment at 65 minutes, whereas at the other two time points, other stress responses were more significant. Such
inconsistency resonates with dynamic nature of stress responses and highlights the importance of exposure time and the necessity to take temporal patterns into account factor for the identification of toxicity mechanisms. The temporal variability may be as meaningful as the expression level at a single time point.

**Figure 5-2.** Comparison of GSEA results based on toxicogenomics data at three different time points upon exposure to MMC (0.5 ng/L) (30, 65 and 100 minutes upon exposure). Genes were ranked by mean altered gene expression levels based on triplicate treatments. X axis bottom: displays \( p \) value in negative logarithmic scale.

### 5.3.3. GESA Analysis of Time Series Toxicogenomics Data-Comparison of Two Score Metrics

To incorporate the time impact into the GESA analysis, we proposed and employed two gene ranking metric, namely CPCA score and TELI score. As described above, CPCA analysis can be applied to rank genes based on their contribution to the temporal variation among different treatments. Another molecular endpoint-TELI was also applied for gene ranking, which measured the accumulative gene expression level alternation
normalized to exposure time and it represented the average expression level within the experiment period. Figure 5-3 shows the comparison of GESA ranking results with both metrics. Based on the gene ranking results, the statistical significances of enrichment for 7 stress response pathways are summarized in Table 4-1.

(a) GESA ranking results for MMC (0.5 µg/L) based on CPCA score (left) and TELI (right).

(b) GESA ranking results for H_2O_2 (10mg/L) based on CPCA score (left) and TELI (right).

(c) GESA ranking results for Pb^{2+} (0.125 µg/L) based on CPCA score (left) and TELI (right).
Figure 5-3. Comparison of GESA gene ranking results for altered gene expression results for (a) MMC at 0.5 μg/L, (b) H$_2$O$_2$ at 10mg/L and (c) Pb$^{2+}$ at 0.125 μg/L using two different ranking metrics CPCA score (left) and TELI (right). Genes are positioned based on their ranking score in non-decreasing order from left to right based. Each vertical strip represents a gene; with color-code indicating its associated stress response pathway and the dot displaying its ranking score. The results are based on average of triplicate treatments. The color codes for each stress response categories or pathways labels are shown in (d).

Table 5–1. Comparison of GSEA results ($p$-values) using two different ranking metrics, namely CPCA score and TELI, respectively.

<table>
<thead>
<tr>
<th>Pathways$^1$</th>
<th>MMC</th>
<th>H$_2$O$_2$</th>
<th>Pb$^{2+}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CPCA</td>
<td>TELI</td>
<td>CPCA</td>
</tr>
<tr>
<td>General Stresses</td>
<td>&gt;0.99</td>
<td>0.48</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Redox Stresses</td>
<td>&gt;0.99</td>
<td>0.48</td>
<td>0.20</td>
</tr>
<tr>
<td>Protein Stresses</td>
<td>&gt;0.99</td>
<td>0.55</td>
<td>0.99</td>
</tr>
<tr>
<td>Cell Killing</td>
<td>0.13</td>
<td>0.69</td>
<td>0.99</td>
</tr>
<tr>
<td>DNA Damage</td>
<td>&lt;0.01</td>
<td>0.72</td>
<td>0.99</td>
</tr>
<tr>
<td>Drug Resistance</td>
<td>&gt;0.99</td>
<td>0.72</td>
<td>0.81</td>
</tr>
<tr>
<td>Detoxification</td>
<td>&gt;0.99</td>
<td>0.39</td>
<td>0.02</td>
</tr>
</tbody>
</table>

1. Data were based on time series stress response altered gene expression profiling data after exposure to MMC (0.5 μg/L), H$_2$O$_2$ (10 mg/L) and Pb$^{2+}$ (0.125 μg/L). $p$-values are corrected for multiple comparisons using false-discovery rate method.

Assessment of the pathway enrichment analysis results against the prior knowledge of the toxicity mechanism of the three model chemicals allows us to evaluate the
performance and reliability of the two approaches for gene ranking. For treatments with MMC, DNA damage and repair pathway-related genes were significantly \( p<0.01 \) enriched based on CPCA score for ranking, which is consistent with known toxic mechanism of MMC as a genotoxicant. However, TELI based GESA analysis did not yield clear gene enrichment for DNA stress pathway. For \( \text{H}_2\text{O}_2 \), both CPCA and TELI-based GESA analysis seemed to yield comparable results, pointing to enriched general stress and detoxification response that are consistent with known toxic effects of \( \text{H}_2\text{O}_2 \) as a strong oxidant with wide and multiple cellular impacts. There was also discrepancy between GESA results for \( \text{Pb}^{2+} \) with either CPCA or TELI ranking method. CPCA score pinpointed activation of drug resistance pathway, whereas result based on TELI indicated more significant detoxification activities (Goering, 1992).

CPCA-based GESA gives more weight to genes that have large temporal variation during exposure time, and less weight to those that have low level but sustained level of expression. For example, a gene that exhibits a sharp impulse response would be ranked higher than one that has lower but sustained altered expression level over the exposure time period. This “high temporal variance high ranking” assumption is somewhat ambiguous. It is recognized that the temporal regulation processes, the expression pattern and gene/pathway involvement could be too complex to be represented by a simple variance. More drastic and higher magnitude of expression change has been often assumed to indicate likely involvement and responsiveness of a gene to external stimulus of toxicant. However, genes that exhibited sustained but lower level of altered expression may also be vital for certain pathway activities depending on the gene function and its maximum level of fold change (i.e., the maximum level of alteration may vary for
different genes) (Khatri et al., 2012). The overall results indicated that CPCA-based GSEA seemed to be able to identify some toxicity mechanisms that were missed by TELI-based GSEA, suggesting the likely higher importance of temporal dynamics of gene expression than the accumulative magnitude alone, at least in the case of stress responses. It should be pointed out that the CPCA applied here also extracted common variation information from multiple samples such as replicates treatments, which helped to further reduce systematic experiment errors.

5.3.4. Impact of Concentrations on Toxicity Mechanisms Identification

Toxic response of a given chemical as revealed by toxicogenomics data can vary with dose concentrations, as previously discussed. The molecular toxicological responses seemed to be dose-dependent; therefore toxicant concentration would impact the GESA results for toxicity mechanism identification (Altshuler, 1981; Burgoon & Zacharewski, 2008; Daston, 2008; Peddada et al., 2003). Here, for each chemical and at each concentration, we first applied CPCA-based and TELI-based GSEA, and then we applied second-level GESA analysis using CPCA-based method for data of all treatments at six different concentrations to rank genes that shared common ranking among all concentrations.

Figure 5-4 and Table 5-2 show the GSEA results and statistical significances analysis of seven stress response pathways. For MMC, GESA that considered the common genes among all dose concentrations led to similar conclusions as that with one single concentration. CPCA-based GSEA identified DNA damage as main molecular effect of MMC, whereas TELI-based GESA failed to isolate any significantly enriched pathway activity. For H$_2$O$_2$, the GESA analysis using data at various dose concentrations,
however, yielded different outcome from those based on one single dose concentration, indicating more dose-sensitive response patterns. Both CPCA-based and TELI-based GESA analysis considering multiple dose concentrations suggested significant activation of redox stress, the known toxic mechanism of oxidant H$_2$O$_2$, which was not detected by analysis with one single dose concentration (Table 5-1). In addition, TELI-based analysis with multiple concentrations indicated the significant activities of genes involved in DNA damage that was not enriched based on single concentration data and it was not captured by the CPCA-based analysis either. For Pb$^{2+}$, the CPCA-GESA identified dominant and significant enrichment of pathway that changed from mainly drug resistance at single concentration to DNA-damage for multiple concentrations, suggesting possible dose-depending shifts in molecular response patterns. The results for TELI-based GESA were consistent for both single and multiple concentrations, pointing towards detoxification pathway that involves genes like norR, fpr, tam, yeiG, yafN, yeaE, grxA, gst. These results are generally consistent with prior toxicological knowledge, as Pb$^{2+}$ is known to cause oxidative stresses that can consequently lead to cellular level multiple responses including DNA damage and detoxification (Flora et al., 2012).

The above results showed that concentration affect molecular toxic response profiles and the pathway activation revealed by GESA analysis could yield different outcome depending on the ranking matrix employed, as well as on the toxicity nature of the chemical. The impact of concentration on pathway activation and toxicity mechanism identification may be more pronounced for chemicals that exhibit dose-dependent toxicity mechanisms, such as H$_2$O$_2$ or Pb$^{2+}$. Others may exhibit more chemical-specific and not so concentration-sensitive toxicity effects such as MMC. Designing and
implementing particular techniques, such as CPCA, presents the ability to extract and identify genes that show more common behavior among all the dose concentrations and indicate the conserved chemical-specific toxic effects. However, it can be argued that overly focuses on the commonality may overlook the dose-dependent specificity of toxicity nature. Although dose-dependent toxicity is the central dogma of toxicology at phenotypic level, whether this holds true for molecular toxicity as revealed by toxicogenomics studies remains largely unknown. On the other hand, there were some successful cases that applied molecular endpoints to address molecular dose-response relationships (Allen et al., 1994; Altshuler, 1981; Gou & Gu, 2011; Gou et al., 2014).

(a) GESA ranking results for MMC based on CPCA score (left) and TELI (right).

(b) GESA Ranking results for H₂O₂ based on CPCA score (left) and TELI (right).
Figure 5-4. Comparison of GSEA analysis gene ranking results with consideration of all six dose concentrations for (a) MMC, (b) H$_2$O$_2$ and (c) Pb$^{2+}$ using two different ranking metrics CPCA score (left) and TELI (right) for each chemical. Genes are positioned in non-decreasing order from left to right based on the metrics. Each vertical strip represents a gene, with color indicating its associated pathway and dot displaying ranking score. The results are based on results (triplicate treatments) of all six dose concentrations for each chemical. The color codes for each stress response categories or pathways labels are shown in (d).
Table 5–2. Comparison of GSEA results with consideration of six dose concentrations using different ranking metrics, CPCA score and TELI.

<table>
<thead>
<tr>
<th>Pathways</th>
<th>MMC</th>
<th>H₂O₂</th>
<th>Pb²⁺</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CPCA</td>
<td>TELI</td>
<td>CPCA</td>
</tr>
<tr>
<td>General Stresses</td>
<td>1.00</td>
<td>0.82</td>
<td>0.03</td>
</tr>
<tr>
<td>Redox Stresses</td>
<td>1.00</td>
<td>0.45</td>
<td>0.07</td>
</tr>
<tr>
<td>Protein Stresses</td>
<td>1.00</td>
<td>0.45</td>
<td>0.99</td>
</tr>
<tr>
<td>Cell Killing</td>
<td>0.50</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>DNA Damage</td>
<td>&lt;0.01</td>
<td>0.82</td>
<td>0.99</td>
</tr>
<tr>
<td>Drug Resistance</td>
<td>0.62</td>
<td>0.82</td>
<td>0.80</td>
</tr>
<tr>
<td>Detoxification</td>
<td>0.21</td>
<td>0.60</td>
<td>0.34</td>
</tr>
</tbody>
</table>

1. Data were based on time series data after exposure to MMC, H₂O₂ and Pb²⁺ at six different dose concentrations for each and in triplicates. *p*-values are corrected for multiple comparisons using false-discovery rate method.

5.4. Conclusion

In conclusion, we employed CPCA-based gene set enrichment analysis with two ranking metrics against high-dimension time series toxicogenomics data, with the aim to evaluate the impacts of time and concentration on the determination of significance of pathway activation and identification of chemical toxicity mechanisms. Our results demonstrated that both time and dose concentration impact the altered gene expression profiles, therefore the consequent GESA outcome. Comparison of the two gene ranking methods and metric indicated that choice of ranking score matrix may lead to inconsistent GESA results and toxicity evaluation conclusions as the results of differences in the underlying assumptions, logics and the weights given to the genes. Employment of ranking matrix that has improved ability to capture the temporal dynamics in gene expression pattern, as well as the consideration of dose-dependent toxic responses, as demonstrated in this study, is expected to potentially lead to more accurate identification of toxicity mechanisms of a chemical.
Chapter 6

High dimensional Toxicogenomics Data Analysis Using Tensor Decomposition Model for Toxicity Identification and Characterization
6.1. Introduction

Tensor decomposition is a group of models that are able to fit high dimensional data. These models decompose the data into several components, each representing a stereotypical pattern of variation in the data. Within each component, the effects of factors are fixed at all levels, and the factors are independent of each other and together contribute to the overall gene expression levels profile (Kolda & Bader, 2009; Mørup, 2011). Tensor decomposition models generalize low-dimensional decomposition methods singular value decomposition (SVD) to tensor data (Omberg et al., 2007; Ponnapalli et al., 2011). These result in two benefits. First, as the data dimensionality is reduced, it is now possible to approximate the variance of the original dataset into the sum of the contribution of different components and random background noise (Vannieuwenhoven et al., 2012). Second, this approach helps addressing the problem of high modality of toxicogenomic data (Clarke et al., 2008). It is known that genes can work in different pathways and that are regulated by different signaling network. In this type of gene profiling we can only observe the overall expression outcome instead of the contributions of individual process (Miller et al., 2008; Ransohoff, 2005). Tensor decomposition, as demonstrated by its successful application to blind source separation in signal processing, is especially useful in such problem (Cichocki, 2009). By decomposing the data into several components that account for different biological mechanisms (e.g. pathways, signaling network), we can possibly avoid false conclusion based on spurious correlation or the apparent lack of any correlation. In light of these advantages, several studies in the past have utilized different tensor decomposition models for understanding DNA replication and transcription (Omberg et al., 2007; Omberg et al., 2009), stem cell
processes (Yener et al., 2008). Among different variants of tensor decomposition models, the most common ones are Parallel Factor Analysis (PARAFAC) (Rasmus Bro, 1997) and Tucker3 (Kolda & Bader, 2009; Omberg et al., 2007). While the latter is more flexible, we are often interested in obtaining not the best-fitting models but the most parsimonious and robust ones. PARAFAC is regarded as an adequate data decomposition method to capture interpretable results and is widely used for data mining problems (Rasmus Bro, 1997; Yener et al., 2008).

In this study we investigate the environmental pollutants-induced toxicity through the application of tensor decomposition to high dimensional toxicogenomics time series data. Using gene expression data from various toxicants tested by our GFP-fused whole cell array platform, we show that the three-dimensional \( gene \times time \times concentration \) data structure is suitable to hold the information on gene expression’s dependency on multiple factors. We find that PARAFAC fits the original data well and is able to decompose it into several components. These components are simpler in that the effects of factors are decoupled and could be studied separately. Throughput further pathway analysis, the PARAFAC outputs provide better insights into possible mechanisms of action. The relationship between biological responses and toxicant concentration can also be quantified using PARAFAC in a more mechanistic way.

6.2. Materials and methods

6.2.1. Data Generation and Processing

We employed a high-throughput toxicogenomics assay that uses whole cell array of \textit{E. coli} K12, MG1655 (Gou et al., 2010; Onnis-Hayden et al., 2009) with green fluorescent protein (\textit{gfp}) fusions to different gene promoters, which allows monitoring
temporal gene expression changes through measuring cellular GFP levels. The library covers 100 genes involved in various cellular stress response pathways that are known to be highly conserved among species (Zaslaver et al., 2006). The genes are associated with 5 known cellular stress response pathways, such as general stress, protein stress, redox stress, DNA damage and membrane stress (Keseler et al., 2013). Each fusion is expressed from a low-copy plasmid, pUA66, which contains a kanamycin resistance gene and a fast folding gfpmut2, allowing for real-time measurement of gene expression level changes (Onnis-Hayden et al., 2009; Zaslaver et al., 2006).

A total of 4 toxicants were tested, 4-n-Nonylphenol (4NNP), trichloroacetic acid (TCA), carbaryl (CAR), mitomycin C (MMC). Based on prior studies, these chemicals can induce different stress responses and be used as model compounds to testify the efficacy of our proposed analytical methods (American Conference of Governmental Industrial Hygienists., 2001; Chitra, Latchoumycandane, & Mathur, 2002b; Dean Blevins, Lee, & Regan, 1977; International Agency for Research on Cancer & IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 1988a). For each chemical, 6 sub-cytotoxic concentrations (>95% survival percentage determined by a 2 hour growth inhibition test) were tested with 3 replicates each. Details of chemical information are listed in Table C-1. The protocol to measure the temporal gene expression profile was described in our previous reports (Gou & Gu, 2011; Onnis-Hayden et al., 2009). In brief, E. coli reporter strains were cultivated in 384-well microplates (Costar, Bethesda, MD, USA) in dark conditions to avoid GFP photo-bleaching until the early exponential growth stage (OD600~0.2) was reached. Inoculum of each of the compounds of interest was added to the wells, and the plate was placed into a micro-
plate reader (Synergy Multi-Mode, Biotek, Winooski, VT) for simultaneous cell growth (absorbance, OD600) measurement, denoted as OD, and green fluorescence readings (GFP level, excitation 485nm, emission 528nm), denoted as GFP. Measurements were taken every 5 minutes over a 2-hour period, resulting in a total of 25 time points for every gene in every treatment.

The raw GFP and OD data were first normalized among replicate plates against the average media controls (wells with growth media only), then were corrected by blank media control for potential interference of chemicals, and referred as GFP_corrected and OD_corrected. The gene expression level normalized by OD for each gene at every time point was calculated as \( P = \frac{GFP_{\text{corrected}}}{OD_{\text{corrected}}} \) and was further adjusted by subtracting the value attributed by non-promoter activities (promoter-less strain controls). The alteration in gene expression for a given gene at each time point due to chemical exposure relative to the control condition without any chemical exposure, also referred as induction factor I, is represented by as \( I = \frac{P_e}{P_c} \). Where, \( P_e = \frac{GFP/OD_{\text{experiment}}}{OD_{\text{control}}} \) as the normalized gene expression level in the experiments condition with chemical exposure, and \( P_c = \frac{GFP/OD_{\text{control}}}{OD_{\text{control}}} \) as the control condition without any chemical exposure. The natural logarithm of the induction factor \( \ln(I) \) was then calculated for further data analysis, where a gene was up-regulated if \( \ln(I) > 0 \) and down-regulated if \( \ln(I) < 0 \) (Tina K. Van Dyk et al., 2001). The data used in the final analysis is the mean value across three triplicates.

6.2.2. Tensor Decomposition Using PARAFAC (Parallel Factor Analysis)

For each chemical, the expression data are organized in a third-order tensor, namely a three-dimensional \((I \times J \times K)\) array, denoted as \( X \), where \( I, J, K \) are the total number of genes, time points and concentrations, respectively. Each individual expression value in \( X \)
can be represented as $x_{ijk}$, where the subscripts $i$, $j$, and $k$ indicates that the expression value is measured for the $i^{th}$ gene, at the $j^{th}$ time points, after exposure to the $k^{th}$ concentration of that chemical. We applied PARAFAC model to decompose this tensor into $R$ components and a residual tensor, all of which are of the same dimension as the original data. Each component $Y_r$ ($r = 1, 2, \ldots, R$) is the outer product of 3 vectors $a$, $b$, and $c$, of length $I$, $J$ and $K$, denoted as $Y_r = a \circ b \circ c$, meaning each element in $Y_r$ can be expressed as $y_{ijr} = a_i^r b_j^r c_k^r$, where $a_i^r$, $b_j^r$, and $c_k^r$ represents the effects of the $i^{th}$ gene, $j^{th}$ time point, and $k^{th}$ concentration on component $r$, respectively. The biological implication is that the expression value within each component is the product of gene, time, and concentration effects, and expression values measured at same level of a given factor would always have the same effect from that factor. The original tensor $X$ can be expressed as the sum of the components and residual, such that each element in $X$ can be expressed as $x_{ijk} = \sum_{r=1}^{R} a_i^r b_j^r c_k^r + e_{ijk}$. Importantly when the background noise effect is small (e.g. $e_{ijk} \to 0$) the data can be fully approximated by the decomposition of effects. In this project PARAFAC analysis was carried using the N-way toolbox (version 3.31) in MATLAB (Mathworks, MA, version 2013b). For more details in PRAFAC, please refer to Bro (Rasmus Bro, 1997).

6.2.3. Pathway Analysis and Dose-Response

A modified version of gene set enrichment analysis (GESA) was used to characterize the individual components after PARAFAC decomposition (Gao, 2015; Subramanian et al., 2005). In this method, a generalization of PCA, common PCA, was used to ranking genes based on their temporal variation. The gene labels were taken from online database
EcoCyc (Keseler et al., 2013). The $p$ values were corrected using a false-discovery rate method (Benjamini & Hochberg, 1995). We also investigate the relationship between concentration and the corresponding concentration effect within each component, namely dose-response relationship. The above algorithm and analysis was implemented using MATLAB (Mathworks, MA, version 2013b, including bioinformatics toolbox).

6.3. Results and Discussion

6.3.1. Factorial Design Offers Dynamic Insights of Stress Responses

The multiple-factor experimental design resulting in highly-dimensional datasets is common in several disciplines such as psychometrics (Kiers & Mechelen, 2001), chemometrics (Appellof & Davidson, 1981), neuroimaging (A. H. Andersen & Rayens, 2004), signal processing (Comon, 1994), and certainly bioinformatics (Omberg et al., 2009). In toxicogenomics studies, the temporal factor is a crucial one as the toxicological responses are usually of dynamic nature (Gao et al., 2012). Since snapshots at single time points cannot provide a comprehensive assessment of the system or be readily extrapolated to another time point, an increasing number of studies employed time course experiment design to study and model dynamic biological processes. Time course experiments are also beneficial for modeling because they provide extra information, which compensates the limited number of samples, compared to the large quantity of genes monitored, though sophisticated models taking time into account are needed in this case.

We applied GFP-fused whole-cell reporter array platform as a non-intrusive assay that enables high-resolution and relatively long-term continuous measurement of gene expression levels in response to toxicant(s) at a reduced cost. The high dimensional data
generated cause an increase of the complexity of data analysis and requires appropriate and efficient analytical approaches.

**Figure 6-1.** Overview of three-dimensional tensor of time series toxicogenomics data (altered temporal gene expression in response to toxicant 4NNP) and dissected different 2-dimensional profiles: (a) The cubic tensor data structure shows that the altered gene expression levels are indexed by 3 experimental factors: gene, time and concentration. The arrow in each dimension indicates the order of that factor (increasing order for time and concentration). The small cube in gray shade illustrates a specific data point measured for a 1st gene, at the 1st time point after exposure to the 1st (lowest) concentration of the given chemical. (b-d) By slicing the tensor from different perspectives, various two-dimensional profiles can be extracted. For example, by slicing horizontally, gene × time profiles can be obtained at each concentration. In these profiles, the gene expression values within each profile are represented by the elevation in the vertical axis and color (color scale is shown in the color bar to the right). The dimensions are not proportional to the number of levels for individual factors, and random levels for the factors are selected for visualizing sliced profiles.
An overview of our gene expression time series data is shown in Figure 6-1. The data is organized as a tensor, and can be visualized as a cube (Figure 6-1a), where three dimensions represent the three experimental factors: gene, time and concentration. Each element in the cube is the expression level for a specific gene at a given time after exposure to a toxicant of a certain concentration. It is difficult to visualize high dimensional data, but by slicing the cube into two-dimensional layers from different angles we can explore the impacts of factors on gene expression locally, as shown in Figure 6-1(b-d). For example, in Figure 6-1b, among the surface plots demonstrating the individual gene × time profiles at different concentrations, the topological aspects such as the locations of the peaks, troughs and trends are similar, whereas the relative expression intensity is increasing as the concentration increases. This results agrees with the common understanding that an increase in toxicant dose would generally result in an increase in toxicogenomic response (Imlay & Linn, 1987), and points out that this understanding can also be extended to temporal expression profile as well. It is worth noticing the shapes of these gene × time profiles differ among toxicants and can be conservative potentially, implying a fingerprint like behavior. In Figure 6-1c, the time × concentration profiles depict different gene expression time series at various concentrations for each different gene. The time series data of a gene’s expression reflects its functional role in toxicological response and the regulations it receives along the time (Bar-Joseph, 2004). It is interesting to observe that the gene expression time series is toxicant-concentration dependent, and there is intergenic variation in the response to the toxicant. This is likely due to toxicant’s different mechanistic influences to genes at varying concentrations and to the organism’s switching of responsive
strategies accordingly. In the end, when the data cube is sliced from the time perspective, the gene × concentration profiles are extracted as in typical static experiments. As expected for most of the genes, the expression levels increase along with the increasing toxicant concentration, however we also see that at distinctive time point, these patterns vary among genes. A possible explanation is that genes respond to external stimuli at varying pace, depending on the timing of the pathways and regulatory network they are part of. This may suggest the presence of complex non-linear interactions underlying the expression of specific genes such as a coherent feedforward loop (Mangan, Zaslaver, & Alon, 2003). This also confirms the fact that taking an isolated snapshot at an early stage in the response is not a good predictor of how toxicant concentrations impact slowly responding genes, and stresses the importance of taking into consideration the temporal element. In summary, the above results demonstrate that in a complex event such as a toxicological response, tensor representation is the proper data structure for the related data analysis, as the arrangement of data points reflects the factors’ mutual interactions without loss of information.

6.3.2. Tensor Decomposition (PARAFAC) of High Dimensional Toxicogenomics Data

In practice, there are two difficulties that prevent us from obtaining more useful insights from tensor data. First, the toxicogenomics data are possibly a mixture of signals from several sources, such as various stress response pathways or gene interaction networks. Thus it is hard to interpret the experiment results without separating it into individual underlying biological processes. Second, tensor itself can only offer qualitative impression, whereas a quantitative tool is in need to describe the effects of each factor. Here we employed PARAFAC model, a tensor decomposition technique, to overcome
these hurdles. PARAFAC decomposes original toxicogenomic data into several components, within each of which the factors are independent to one another and the effects of each factor are fixed for every level. Figure 6-2 demonstrates the decomposition result using one of the toxicants tested-4NNP. In this case, PARAFAC fits the original tensor with 2 components and a residual tensor for the noise. The active gene sets differ between the components. The temporal gene expression profiles are also slightly different—one with relatively earlier activation time and faster ascending trend compared to the other. On the other hand, these properties are similar within each component, and expression patterns are much simpler than the original data.

**Figure 6-2.** Demonstration of tensor decomposition (toxicant: 4NNP). The original tensor (Column 1) was decomposed into two components (Column 2 and 3), plus a residual tensor (Column 4). The tensors are sliced horizontally for visualization, and selected slices are displayed here with each row indicates a concentration (low to high from top to bottom). The tensor is the corresponding sum of the component tensors and residual tensor: each original data point is the sum of the corresponding data points from each component and residual tensor; and hence each original slice (two-dimensional gene × time profile) is the sum of the corresponding two-dimensional profiles from each
component and residual tensor. The altered gene expression values within each profile are depicted by the elevation in the vertical axis and color (color scale is shown in the color bar to the right).

The ability to decompose complex data into simple and coherent components paves the way for further functional analysis. It is well-acknowledged that gene expression can affect simultaneously the functioning of several pathways (Clarke et al., 2008) and most toxicogenomics techniques only record the whole expression level altogether without the ability to separate the sources (Kong, Vanderburg, Gunshin, Rogers, & Huang, 2008; Zinovyev, Kairov, Karpenyuk, & Ramanculov, 2013). In comparison with these other methods, PARAFAC, along with other tensor decomposition techniques, decomposes the original tensor data into several coherent components with high dimensional “spatial” regularities, because of the source separation capability from their signal processing origin (Cichocki, 2009) and allows detecting pattern regularities within the components that often have underlying biological implications. For example, in Column 2 and 3 of Figure 6-2, it is shown that the genes with substantial activities within the same component have the analogous temporal profiles except for magnitude differences, which could be due to the fact that these genes are under common kinetic regulation from an upstream gene (Bar-Joseph, 2004). Therefore, PARAFAC as well as other tensor decomposition techniques are valuable when the toxicological responses to a certain toxicant involve multiple processes, which is often the case. Furthermore, it allows uncovering some detailed toxicological mechanistic information by analyzing each single component, which is discussed later.
Figure 6-3. The effect coefficients of factors within each component (toxicant: 4NNP) and their time- and concentration-dependent patterns. In tensor decomposition, each produced component is the outer product of 3 vectors, which contains contributing effect of each factor to the final gene expression values. So the $r^{th}$ component can be expressed as $Y_r = a \circ b \circ c$, where $a$, $b$, and $c$ are the respective vectors containing the effects of gene, time and concentration, shown above. The lengths of the vectors equal to the number of levels for the corresponding factors. An individual gene expression value in the component can be expressed as $y_{ijk} = a_i' \times b_j' \times c_k'$, where $a_i'$, $b_j'$, and $c_k'$ represents the effects of the $i^{th}$ gene, $j^{th}$ time point, and $k^{th}$ concentration in this component, respectively. In (a), genes are sorted by their coefficients in Component 1.

Within a decomposed component, the effects of each factor are fixed and can be expressed as a vector of coefficients. The factors are independent of one another, and hence the overall expression is the product of 3 corresponding coefficients from each
factor, i.e., \( \text{expression} = \text{gene} \times \text{time} \times \text{concentration} \). This model ensures the shape of any two-dimensional slice is similar to all its parallel counterparts, which in turn make these profiles biologically relevant but for their expression intensity discrepancies. For example, as we discussed in Figure 6-2, all the \( \text{gene} \times \text{time} \) profiles within each component have similar patterns. The coefficients of three experimental factors in the two components are shown in Figure 6-3. For most of the genes, the effects in the two components have opposite signs showing the presence of two major opposing patterns affecting gene expression following toxicant addition. In the remainder of the genes the effect in one component is substantially large in absolute value while the other one is close to 0. On the other hand, the two components share similar temporal effect with only slight difference in activation stage and slope. The time effects in both components are similar with overall increasing trends. In addition, similar behavior is also observed when looking at the concentration effect vectors, where a general increasing trend (except for the last concentration values) strongly correlates with a corresponding increase in concentration. As discussed above, the components could correspond to regulatory or functional pathways, so a careful investigation of the effects of each factor in the components would be helpful to address the common toxicogenomic questions, such as toxicity mechanisms elucidation and quantification of toxicity effects like dose-response relationship.

We chose four chemicals, namely, 4-NNP, TCA, CAR and MMC, to demonstrate the application of tensor decomposition to extract biological close within but distinct without components. The tensor of gene expression data for each toxicant was first decomposed into several components using PARAFAC. To determine the number of components in
PARAFAC analysis, we referred to two commonly used metrics: the percentage variation of original data explained by the model which ranges from 0 to 100% and is analogous to the sum of the fraction of expected variance captured by each component in an SVD analysis (Alter, Brown, & Botstein, 2000), and core consistency ranging from 0 to 100 (R. Bro & Kiers, 2003). For both metrics, the higher value a model has, the better it performs, though it is sometimes difficult to obtain high scores in both cases. In practice, we tried to achieve a balanced point, namely the number of components that would make the model explain more than 70% of the original variation and have a core consistency larger than at least 50. The numbers of components for the toxicants after applying PARAFAC are shown in Table 6-1. For 4NNP and TCA, two components are able to explain around 70% of original variation, with a high core consistency. For CAR, though the model, the core consistency drops to 65, showing that these metrics are not necessarily correlated, explains more than 80% of variation. For MMC, though the component number is high (e.g. 4), both metrics are low than desired. This lower model performance can be attributed to several reasons, such as the inherent noise from the original data, or the suitability of PARAFAC to model the effect of this specific toxicant. For the latter, several other tensor decomposition methods like Turker3 are available, however their results are not so straightforward for interpretation which is beyond the scope of our discussion (Mørup, 2011).

6.3.3. Pathway Analysis Based on Tensor Decomposition

As we discussed above, the components from PARAFAC could potential account for different pathways activated during the toxicological responses after exposure to the toxicant. Therefore we employed GSEA (gene set enrichment analysis), which is a
specific technique for pathway analysis, to each individual component to detect the relevant pathway in it and tested our hypothesis. We selected a modified GSEA applicable to time series data based on our prior study (Gao, 2015) because it is more suitable to analyze gene expression dynamics and able to detect temporally active pathways, which is very important in the time course experiment design context. The results of GESA, namely the \( p \) values showing the significance of enrichment for each pathway, are shown in Table 6-1. The significantly enriched pathways with small \( p \) values would be more active during the toxicological response and could explain the mechanisms of toxicity for the specific chemical. Because of the noisy nature of biological data and the limited size of our library, we set \( p < 0.3 \) as a reasonable cutoff threshold, close to the one used by Subramanian et al (Subramanian et al., 2005). Among all the toxicants, 4-n-Nonylphenol (4NNP) is a known endocrine disruptor (Bonefeld-Jorgensen, Long, Hofmeister, & Vinggaard, 2007a; Chitra et al., 2002b). The two components for 4NNP both have significantly general stress. In addition, the second component also recognizes protein damage relate stress, which is consistent with prior knowledge (H. R. Andersen et al., 1999b; Petit, Le Goff, Cravedi, Valotaire, & Pakdel, 1997b). It is known that trichloroacetic acid (TCA) can cause redox stress (American Conference of Governmental Industrial Hygienists., 2001; Hassoun, Cearfoss, & Spildener, 2010b) as well as DNA damage. In the result, the first component for TCA shows significant redox stress. The second one does not identify a single enriched stress; however several pathways, namely general stresses, DNA damage and redox have \( p \) values close to threshold. Studies have shown that carbaryl (CAR) is related to DNA damage (Dean Blevins et al., 1977), as its 3rd component indicates. In addition, the 1st
component shows that CAR might cause general stress in the cell, which is also congruent with previous finding (Jansen et al., 2013). In the end, mitomycin C (MMC) is a known agent to induce DNA damage (International Agency for Research on Cancer & IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 1988a). While three out of four components for MMC indeed display strong signs that DNA damage exists. Detailed examination of genes with each component revealed distinct DNA-damaging mechanism each component represent. For example, Component 4 includes a few genes responding to double strand break, such as recE, ruvA and umuD that are have high effects (Kanehisa & Goto, 2000).

The above observations show that the pathway analysis of PARAFAC is generally consistent with the known mechanisms of action for the toxicants. It is possible that the selection of biomarkers indicative of these stress response categories or pathways need further evaluation, and/or they are equally enriched and hence failed to be recognized in current library setup, which could be improved by re-evaluating and incorporating more genes in the future. Nevertheless, it provides a set of extra information about pre-defined possibly activated pathways. The existence of more than one component suggests the presence of certain cellular response pattern (such as so called pathways) working currently when the cell is facing external toxicant, and these cellular response patterns each have their own unique three-dimensional $\text{gene} \times \text{time} \times \text{concentration}$ expression 3D shape. At the same time, similarities and differences of significantly enriched pathways among the components inferred for a specific toxicant also provide interesting insights.
6.3.4. The Investigation of Dose-Response Relationship and Temporal Dynamics Using PARAFAC Analysis

The independency of factors with each other within the components makes tensor decomposition an invaluable tool to analyze toxicant concentration effects or temporal dynamics. Figure 6-4 shows the concentration effect coefficients from the four analyzed toxicant. The results demonstrate that each component has varying dose-dependent response pattern, with some components have nearly constant effects at all concentrations, while other components exhibit more dose-dependent response patterns. For 4NNP and TCA, both identified components showed concentration-sensitive trends. CAR seems to cause general stress at all concentrations and it started to exert more DNA-damaging effects (component 3) only at higher doses. For MMC, the known DNA-damaging agent, further decomposition leads to finer-resolution understanding of the different DNA damaging mechanisms that revealed by biomarkers indicative if different DNA damage repair pathways.
Figure 6-4. Dose concentration-dependent effects of individual components for the four toxicants. The order of the components for each toxicant is the same as Table 6-1.

Figure 6-5. Time-dependent effects of individual components for the four toxicants. The order of the components for each toxicant is the same as in Table 6-1. Similarly,
PARAFAC can isolate the time effect coefficients, as shown in Figure 6-5. The time effects between components can be similar (Figure 6-5 a), sharing a rapid increase then a period of steady state. The components can also opposite trends in terms of time effect (Figure 6-5 b-d). If as we speculated earlier, different components may correspond to different toxicity mechanisms, then it is also possible these mechanisms have different priorities and cadences during the stress response process.

Table 6-1. Summary of GSEA (gene set enrichment analysis) results for the components of each chemical. The significance of enrichment ($p$ value) for each stress response pathway is shown, after correction with false discovery rate method.

<table>
<thead>
<tr>
<th>Chem.</th>
<th>Com.</th>
<th>Stress Response Categories or Pathways</th>
<th>Model performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DNA stress</td>
<td>General stress</td>
</tr>
<tr>
<td>4NNP</td>
<td>1</td>
<td>0.97</td>
<td>0.16*</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.76</td>
<td>0.27*</td>
</tr>
<tr>
<td>TCA</td>
<td>1</td>
<td>0.99</td>
<td>0.17*</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>CAR</td>
<td>1</td>
<td>0.88</td>
<td>0.00*</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.00</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.00*</td>
<td>1.00</td>
</tr>
<tr>
<td>MMC</td>
<td>1</td>
<td>0.06*</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.00*</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.53</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.30*</td>
<td>0.99</td>
</tr>
</tbody>
</table>

* $p \leq 0.3$

6.4. Conclusion

To summarize, the advances of toxicogenomics techniques lead to complex experimental design that involve multiple interacting factors along with high dimensional tensor data. Tensor decomposition provides us with the ability of detecting the effects of such multi-factorial set-up and identifying multiple biological modules/pathways during
toxicological response. It can be further incorporated into data analysis workflow and offer help to other research topics such as pathway analysis, dose-response relationship or temporal dynamics investigation. Our results show that PARAFAC, one of the tensor decomposition algorithms, is valuable in analyzing high dimensional gene expression time series data with time and dose concentrations dimensions. Compared to other techniques such as Tucker3 (Conesa, Prats-Montalban, Tarazona, Nueda, & Ferrer, 2010) and HOSVD (Omberg et al., 2009), PARAFAC is simple and its results are easy to interpret (Kolda & Bader, 2009). It helps provide more detailed analysis of possible mechanisms of actions, and investigate the does-response relationship for each of them. In general, tensor decomposition is a useful set of tools in the toxicogenomic studies.
Chapter 7

Identifying Dose and Time Dependent Impacts on Gene Co-Expression Networks in Toxicogenomics
7.1. Introduction

The emerging field of toxicogenomics (gene-expression profiling technology), which simultaneously examines the toxicity response and global molecular status of an organism that has been exposed to pollutant(s), promises a new ground for evaluating toxic effects, understanding toxicity mechanisms, and obtaining pollutant-specific molecular fingerprints (or biomarkers) for compound classification and identification (Boverhof & Zacharewski, 2006; National Research Council Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology, 2007; Simmons et al., 2009; Waters & Fostel, 2004). The large amount of data generated from toxicogenomics assays, however, pose analytical challenges and demands for data analysis approaches that not only examine individual gene expression changes but also reveal information at higher functional levels, including gene sets analysis (Subramanian et al., 2005), pathway analysis or gene networks analysis (Barabasi & Oltvai, 2004). These methods are able to reveal the overall physical or functional landscape of the whole biological system, and thus more suitable for toxicity mechanism-centered research (Ideker & Krogan, 2012; North & Vulpe, 2010).

Among functional toxicogenomics tool, gene co-expression network, also known as gene correlation network, has enjoyed a wide use for the past decade (Stuart et al., 2003). Gene co-expression network is normally represented as a mathematical graph, where genes are depicted as nodes, and a significant co-expression relationship between genes is denoted as an undirected edge between nodes (Roy et al., 2014; Usadel et al., 2009). It bases on the “guilt-by-association” principle to infer functional relationships among genes, which are informative since it has been shown that co-expressed genes tend to be
functionally connected (Oldham et al., 2006; Wolfe et al., 2005). Comparing to gene regulatory network, gene co-expression network requires fewer experimental samples, puts less constrains to experiment designs and provides more robust results (Davidson & Levin, 2005; Karlebach & Shamir, 2008; Roy et al., 2014). In toxicogenomics, gene co-expression network has already been applied to understand cellular responses to external stresses in different organisms in the previous studies (Dewey et al., 2011; Zheng & Zhao, 2013). However, current gene co-expression network methodology has been developed focusing on steady state experiment type, namely a snapshot of the transcriptional system, and much less literature has touched upon time series data generated from continuous, time-course experiments (Zoppoli et al., 2010). It has been recognized that experiment design and data analysis that focus on cellular status at a snapshot or isolated time points may overlook the inherently dynamic nature of cellular response and gene interaction, which could be extensively rewired facing different conditions (Bandyopadhyay et al., 2010; Califano, 2011; Mitra et al., 2013). Advances in toxicogenomics technologies such as whole-cell reporter array enables real time monitoring of gene expression levels in a non-destructive fashion (Elad et al., 2008; Elad et al., 2010). The availability of time series toxicogenomics data has recently motivated an increasing number of studies that investigate time-course gene expression profiling in different areas (Deferme et al., 2013; Geijer et al., 2012; Wen et al., 1998; Weng et al., 2014).

Biological network analysis has benefitted from the rapidly developing theory of complex network, which strives to uncover the governing principles for network formation and evolution (Albert & Barabasi, 2002; Strogatz, 2001). Especially, the realization that the architectural properties of biological networks share some similarities
with other complex systems, such as Internet and social network, paves the way to import sophisticated analytical techniques into bioinformatics (Barabasi & Oltvai, 2004; L. Chen et al., 2009; Dorogovtsev & Mendes, 2013; Newman, 2003). Past studies have discovered that the topological properties of biological networks could undergo massive changes in response to external changes (Bandyopadhyay et al., 2010; Califano, 2011). In other words, it is not only the differential expressions of individual genes but also the differential interactions among them that account for cellular adaption of outer environment. A newly developed branch of network analysis, called differential network biology, has been focusing on developing statistical approaches to identify and test those differential interactions, as well as finding their correlation with different conditions including toxicant, dose and time point (Gill et al., 2010; Ideker & Krogan, 2012). This is especially important for toxicogenomics, because many of its applications, such as biomarker discovery and dose-response study, rely heavily on finding the relationship between molecular activity and experimental conditions (Boverhof & Zacharewski, 2006). The abovementioned understandings of biological networks form the foundations upon which several analytical tools and frameworks have been built (Carter et al., 2004; Tesson et al., 2010; B. Zhang & Horvath, 2005). But to our knowledge there is still no systematic investigation of the impact of time and dose factors on gene co-expression networks and their potential applications in toxicogenomics studies so far.

In this study, we conducted time course gene expression profiling experiments and combined gene co-expression network analysis with gene set enrichment analysis (GSEA) to identify dose and time dependent patterns in the networks and inferred toxicity mechanisms and insights across different exposure conditions. Gene expression time
series data with a range of toxicant doses have been collected, upon which different networks have been statistical inferred/constructed. By examining the topology of these different networks, we found evidence showing quantitative toxicant dose-dependent impacts on gene networks in terms of both size, namely the number of gene-gene interactions, and structure, namely characteristics such as average clustering coefficient. In the meanwhile, by analyzing time series data in a similar fashion, we found gene co-expression networks derived at different time points undergo quantifiable size and structure changes over time as well. Furthermore, applying pathway analysis to the networks demonstrates varying manifestations of toxic mechanisms at different doses or time points.

7.2. Materials and Methods

7.2.1. Toxicogenomics Gene Expression Time Series Data and Gene Annotation Generation

A high-throughput toxicogenomics assay was employed that uses whole cell array of *E. coli* K12, MG1655(Gou et al., 2010; Onnis-Hayden et al., 2009) with green fluorescent protein (*gfp*) fusions to different gene promoters, which monitors temporal gene expression changes through measuring cellular GFP levels, as described in our previous reports(Gou & Gu, 2011; Onnis-Hayden et al., 2009). In brief, *E. coli* reporter strains were cultivated in 384-well microplates (Costar, Bethesda, MD, USA) in dark condition to avoid GFP photo-bleaching until the early exponential growth stage (OD600 ≈ 0.2) was reached. Samples of specific compounds were added into the wells, and the plate was placed into a microplate reader (Synergy Multi-Mode, Biotek, Winooski, VT) for simultaneous cell growth (absorbance, OD600) measurement, denoted as *OD*, and fluorescent readings (GFP level, excitation 485nm, emission 528nm), denoted as *GFP*. 
Measurements were taken every 5 minutes over 2 hours, resulting in a total of 25 time points for every gene in every treatment.

We examined both dose impact and temporal dynamics of cellular response to arsenic, a heavy metal toxicant with known toxicological information. To explore the dose impact, a reporter library of 100 genes was employed that cover various cellular stress response pathways known to be highly conserved among species (Zaslaver et al., 2006). The genes are associated with 5 known cellular stress response pathways, including DNA damage, general stress, membrane stress, protein stresses, oxidative stress (Keseler et al., 2013) (See Appendix Table D-1 for list of genes and their pathways). Each fusion is expressed from a low-copy plasmid, pUA66, which contains a kanamycin resistance gene and a fast folding gfpmut2, allowing for real-time measurement of gene expression level changes (Onnis-Hayden et al., 2009; Zaslaver et al., 2006). Six sub-lethal concentrations (1.4×10^{-5} mg/L, 1.4×10^{-4} mg/L, 1.4×10^{-3} mg/L, 1.4×10^{-2} mg/L, 0.14 mg/L and 1.4 mg/L) of Arsenic (III, As for short, Fisher Chemical) were tested, with 3 replicates of the whole library for each concentration. To investigate the temporal dynamics of cellular response to arsenic, an extended reporter library consisting of 1760 genes for the same E. coli strains was used (See Appendix Table D-2 for this list of genes). Ten replicates were performed for sufficient statistical power with Arsenic (III) exposure concentration at 1.4×10^{-3} mg/L. We extracted programmatically the functional gene ontology terms for the genes from UniProt GOA (gene ontology annotation) database using its QuickGO web services (Binns et al., 2009; Huntley et al., 2015).
7.2.2. Data Pre-processing

The GFP and OD data were first corrected for background against controls with E. coli strain without GFP fusion and growth medium without bacteria, respectively. Since GFP signal is proportional to E. coli cell numbers, the population normalized GFP signal was calculated as \( P = \frac{GFP}{OD} \). Induction factor, which measures the gene expression alteration, was calculated as the ratio of normalized expression levels between experimental and control groups (gfp-fused E. coli strains without toxicants added), \( I = \frac{P_e}{P_c} \), where \( P_e \) and \( P_c \) represents for experimental and control groups, respectively. The natural logarithm of the induction factor \( \ln(I) \) was then used for the following data analysis, where when a certain gene is up-regulated if \( \ln(I) > 0 \) and down-regulated if \( \ln(I) < 0 \) (Tina K. Van Dyk et al., 2001).

7.2.3. Network Inference and Analysis

Gene co-expression networks under different conditions (specific dose or time points) were statistical inferred/constructed using software ARACNE, which uses an information-theory based analysis method (Margolin et al., 2006). The only parameter other than default was the significance level used to determine threshold for establishing gene-gene connections, which was set at 0.05. The networks were analyzed using Python package NetworkX (Schult & Swart, 2008). The network visualization were achieved using Cytoscape (Cline et al., 2007). All the other visualizations were carried out using Python package matplotlib (Hunter, 2007). Statistical analyses including gene set enrichment analysis were implemented using Python packages numpy and scipy, with critical routine implemented in C (Oliphant, 2007; Subramanian et al., 2005). The whole
analysis workflow was organized into several IPython notebooks to ensure analytical reproducibility (Pérez & Granger, 2007).

7.3. Results and Discussions

7.3.1. The Impacts of Dose on Gene Co-expression Network Topology

(a) Concentration: $1.4 \times 10^{-3}$ mg/L
(b) Concentration: $1.4 \times 10^{-4}$ mg/L
(c) Concentration: $1.4 \times 10^{-3}$ mg/L
(d) Concentration: $1.4 \times 10^{-2}$ mg/L
Figure 7-1. Gene co-expression networks constructed using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused *E. coli* whole cell reporter library with exposure to arsenic of different doses: from (a) $1.4 \times 10^{-5}$ mg/L to (f) 1.4 mg/L. The networks are displayed using circular layout where the same genes are located in the identical locations of both. For visualization simplicity the gene names are omitted, but the pathways to which the genes belong are denoted using different colors as shown in (g).
As gene co-expression network reflect the state of transcriptional activities, analyze and compare networks provide us a way to analyze and compare cellular responses to different conditions from the systems biology perspective. Conceivably, differential network analysis, which studies the network differences, is analogous to differential expression analysis in several aspects but for the subject. We can thus formulate some hypotheses about networks similar to their counterparts about genes. For example, dose-response relationship has always been an important topic for its role in risk assessment and mechanistic analysis (Hu, Kapoor, Zhang, Hamilton, & Coombes, 2005). If dose has been proven to be able to influence gene expression profiles, here it is only natural for us to suspect the same impact will still exist at network level (Peddada et al., 2003). In Figure 7-1, we showed two gene co-expression networks constructed using gene expression data under different Arsenic concentrations. With the identical genes in both networks, it is clear upon visual inspection that the density of gene-gene connections varied. Such result makes the above hypothesis seem plausible qualitatively, and gives us confidence to seek further quantitative proofs.

We further utilized some metrics describing networks structural characters, namely number of edges and average clustering efficiency. The former measures the size of a network and the latter quantifies the tendency of genes to cluster together in a network. Figure 7-2 (a) and (b) depict the relations between these two metrics and toxicant doses. In both cases, we can see significant linear correlations between the toxicant doses (in logarithmic scale) with both metrics. It is easy to understand that under the influences of external stressor, the cells can possibly switch out of homeostasis and activate their various machineries to remedy the damages. Therefore, we can postulate that
transcriptional activities will increase accordingly, which has already been confirmed before (Gou & Gu, 2011). What is more relevant in this scenario is the fact that genes rarely work in isolation but rather are connected, which might be due to the conservative strategy during evolution (Albert, 2005). The result, especially average clustering coefficient here could be a demonstration that cells also respond to increasing stresses through increasing collaboration between different genes or other levels of functional groups, in addition to simply escalate transcriptional activities levels. Not surprisingly, such finding is consistent with the conclusion from prior study on other type of gene networks (Q. Zhang & Andersen, 2007). There are some other points worth noticing here. First, we only select two metrics as an example, while in reality many more characteristics of networks and corresponding metrics could also be used, possibly producing even more informative results. Second, the high modeling error (low $R^2$) for average clustering efficient might mean a possibility of other confounding factors, which is only to be expected in such a complex system as cells. And hence we can foresee more studies in this front.
Figure 7-2. Network level dose-response pattern: the linear relationship between gene co-expression network size (number of edges), structure (average clustering coefficient) and toxicant arsenic dose concentrations. Gene co-expression networks constructed using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused E. coli whole cell reporter library with exposure to arsenic at different doses. The X axis is in logarithmic scale.

To further understand the upward trend in the curves in Figure 7-2, we plot the average gene expression time series for each stress response pathways in Figure 7-3 at three selected toxicant doses. The results show quite flat expression levels for all functional pathways under low-dose condition, whereas drastic fluctuations for some pathways under high-dose condition. It is possible that when the toxicant concentration is low, the stress response related gene expression activities are kept minimal. In such situation, the expression patterns can be quite random, leading to low correlation among gene expression profiles, and hence low gene-gene connections in the co-expression network. On the other hand, high-dose condition may cause the cells to respond to the
harsher environment more vigorously. Multiple functional pathways will be activated, and those genes participating in such actions will share co-expressed patterns as time goes by, hence the high connectivity in the network. In the meanwhile, more genes taking active part in the functional pathways would explain the higher average clustering coefficient too. Therefore, as the dose of toxicant ascends, the networks have more connections among genes, and genes tend to cluster more closely. To some extent, such results illustrate the advantages of time course experiments in toxicogenomics studies.

\[ (a) \ 1.4 \times 10^{-5} \text{ mg/L} \]
(b) $1.4 \times 10^{-3}$ mg/L

(c) $1.4 \times 10^{-1}$ mg/L
Figure 7-3. Temporal trends of altered gene expression profiles for different stress response pathway categories induced by exposure to arsenic at varying doses concentrations, using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused *E. coli* whole cell reporter library with exposure to arsenic of different doses: (a) $1.4 \times 10^{-5}$ mg/L, (b) $1.4 \times 10^{-3}$ mg/L and (c) 1.4 mg/L. The temporal mean altered gene expressions of all genes in each specific stress response pathway are shown in solid curves using different color legend (d). The 95% confidence intervals for each curve are shown as the bands surrounding the curves with the lighter colors, indicating the range of variations of altered gene expression levels among all the genes in each stress response pathway.

7.3.2. Dose-Aware Mechanistic Analysis Using Gene Co-Expression Network

Gene set enrichment analysis (GSEA) is commonly used and powerful statistical procedure to interpret gene expression profiling result for identifying genes sharing biological functions that collectively have significant expression levels(Subramanian et al., 2005). Particularly, GSEA is used to determine if a predefined gene set (pathway) is over-represented toward the top of a ranked gene list(Subramanian et al., 2005). Usually genes are ranked by differential expression levels, but in gene co-expression network analysis, it has been shown that genes differ topologically. In other words, some genes are in more “central” positions, and there are some metrics quantify such “centrality”.

(d) Color legend
Among them, degree centrality is one of the commonly used ones. Degree centrality of a gene node measure the number of connections a gene has in the network, normalized by the total number of genes. Such metric has been used to evaluate the (relative) importance of a node in the different types of networks for a long time (Freeman, 1979). Potentially it can also be employed to rank genes and provide some functional insights in gene co-expression networks (Koschützki & Schreiber, 2008). Here, we proposed and demonstrated a new approach that incorporates the centrality measure of gene co-expression network analysis into pathway analysis.

Figure 7-4 shows gene set enrichment analysis (GSEA) results for different specific stress response pathway and categories at varying Arsenic dose concentrations, where genes are ranked by their degree centralities in different networks at each dose. Each vertical column lists the ordered genes in a specific stress response pathway according to their ranking at a given expose dose centration and the enrichment scores are shown as solid circle. Those significantly enriched pathways at a certain dose have their $p$ values marked in red. The results revealed dose concentration-dependent and stress response pathway-specific patterns and trends in gene set enrichment analysis outcome, which reflects the extent of involvement of genes and pathways in a particular exposure condition. For example, genes related to protein stress response were significantly enriched at both low and high exposure concentrations, suggesting protein damaging effect of arsenic ($p<0.25$, threshold determined according to Subramanian et al. (Subramanian et al., 2005)). In addition, membrane stress and oxidative stress are also significantly enriched at certain dose concentrations. These results are generally consistent with previous knowledge that Arsenic (III) has been shown to have oxidative
stress as its main toxic mechanism, which can also induce other stresses such as membrane stress (Bernstam & Nriagu, 2000; Takeuchi et al., 2007). The result in Figure 7-4 revealed finer grained detail in toxic mechanism assessment, which seems to exhibit quite concentration-sensitive shifts and dynamics in cellular response pathways and functions. It is still not yet clear if molecular level effects can be indeed quantified and, if yes, how to quantify such impacts in relevance to toxicity at different biological organizational levels. If such observation is true and general, it implies that dose-response pattern do exist at molecular level (gene or pathway) (Altenburger, Scholz, Schmitt-Jansen, Busch, & Eschert, 2012; National Research Council, 2006) and more cautions should be taken during the elucidation of molecular toxic mechanisms, especially considering toxicant doses. Clearly, dose-aware mechanistic analysis requires the precision of the molecular screening techniques, other than conventional phenotypical approaches. Note that all the above discussion is based on the presumption that degree centrality bear biological meaning, which has been generally shown to be true (Albert, 2005). On the other hand, there are several other centrality metrics with plausible definition of node importance too (Azuaje, 2014). Hence, a more thorough investigation in this matter in the future may include testing multiple centrality metrics for functional study.
Figure 7-4. Gene ranking and gene set enrichment analysis results for all stress response pathways at varying arsenic dose concentrations, using time series altered gene expression data generated with stress response pathway ensemble of GFP-fused *E. coli* whole cell reporter library with exposure to arsenic of different doses: (a) DNA damage, (b) general stress, (c) membrane stress, (d) protein stress, and (e) oxidative stress. The genes are ranked according to their degree (number of connections) in the gene co-expression network at each dose. Y axis left: genes represented by horizontal bars and listed based on ranking order (lower value means higher ranking order) as shown in vertical columns; Y axis-right: GSEA enrichment scores shown in solid circle: the higher the enrichment score is, the more enriched toward the top the specific gene set is; X axis top: The statistical significances of enrichment for the specific pathway at a given dose concentration (after corrected for multiple-comparison using false discovery rate Benjamini-Hochberg procedure). The possible significant enrichment (p<0.25, threshold determined according to Subramanian et al.) are highlighted in red. X axis bottom: arsenic exposure dose concentrations in mg/L (Subramanian et al., 2005).
7.3.3. The Impacts of Exposure Time on Gene Co-Expression Network Topology

As discussed above, cellular responses are not static in terms of gene expression levels. It seems reasonable that, as time goes by, cells will deviate from normal homeostasis and activate possible defense mechanisms. If a gene expression profile technique has small enough time resolution and long enough experiment duration, it is possible to capture the dynamics of the various biological processes happened after the cells are exposed to an external toxicant. In this study, we tried to see if the whole cell array platform fits the requirement. And if so, we would investigate the impact of time factor on gene co-expression networks. We focused on one dose ($1.4 \times 10^{-3}$ mg/L) of As (III) with an extended library of 1760 genes. In order to provide sufficient statistical power for more confident network construction at each time point, we employed 10 replicates of the whole library and analyzed the networks generated at different time points across 2-hour experiment duration at 5-mins interval.

The temporal patterns of cellular system response were revealed by network analysis derived parameters, namely size (represented by number of edges) and structural characteristics (represented by average clustering coefficient), as a function of exposure time (Figure 7-5). Both network size and structure exhibited time-dependent patterns, following a generally similar trend, with initial increasing, peaking at around 1 hour after exposure to toxicant, and then showing downward trend. Based on our assumption that gene co-expression networks reflect the states of cellular activities, we may infer that such result is a demonstration of the initial activation of cellular stress responses and the winding down of these supposed costly activities after a period of time. Such result also suggests the temporal resolution of whole cell array technique is fine enough.
Figure 7-5. The change of gene co-expression network size (number of edges, solid line) and structure (average clustering coefficient, dashed line) over time: the networks were constructed using altered gene expression data at each time points generated with stress response pathway ensemble of GFP-fused *E. coli* whole cell reporter library with exposure to arsenic of 1.4×10^{-3} mg/L. Y axis left: edge number of networks; Y axis right: clustering coefficient; X axis: time after exposure to arsenic. The time series are smoothed using moving average method with a window of 5.

The time-dependent gene co-expression networks topological differences motivated us to further explore the possibility of seeking functional information from those networks using pathway analysis, namely identifying the enriched pathways at each time based on network centrality metric. The GSEA analysis was performed for 10 selected biological processes based on gene ontology terms (limited to molecular function aspect) at different time points and the results are summarized in Table 7-1. These 10 biological processes include some essential molecular activities and each contain high enough number of genes (>10% of total 1760 genes) to allow statistically meaningful GSEA.
analysis. The gene sets associated with certain gene ontology terms that are significant enriched are highlighted in red. The results indicated that at different exposure time points, arsenic induces varying cellular responses and metabolic pathway activities. The significantly enriched functional GO terms (pathways) enriched at different time points suggested the dominant molecular response at the time. The significant enriched pathways were initially DNA binding only at 20 minutes, then transited into protein binding, and at around 90 minutes began to include multiple pathways such as oxidoreductase activity, ATP binding, and catalytic activity. Combining Table 7-1 and figure 7-5 above, it seemed to indicate that cellular responses are highly time-dependent with dynamics involvement of varying genes and pathways at different exposure time points. The cell seemed to exhibit more specific molecular activity at early exposure time and then gradually started to show more wide range of cellular effects that peaked at 60 minutes, as indicated by the network structure parameters peak, with enriched GO terms being more spread (more GO terms with relatively lower $p$ values), or in other words, more genes are affected at this time point. Then with extended exposure beyond 90 minutes, different specific GO activities enrichment was observed again. The biological explanation for these interesting trends is not yet fully understood. So overall, the results clearly showed that the molecular activities and toxicity mechanism are time-depend as well, and at different exposure time, the dominant molecular activities vary. This emphasizes again that cautions should be taken when interpreting toxicity mechanism at different time points.
7.4. Conclusion

In this study, we proposed and applied co-expression network analysis in combination with gene set enrichment analysis to analyze and compare the characteristics of networks constructed under different exposure conditions to arsenic, namely varying doses concentration or exposure time points. We employed GFP-fused *E. coli* whole cell reporter libraries to generate time series toxicogenomics data that captured the dynamics molecular transcriptional responses at varying dose concentrations. The results demonstrated dose-response patterns and temporal dependency of co-gene expression network size and structure. Using different analysis of co-expression networks and gene set enrichment analysis together, we reveal more detailed, in depth and system-level cellular molecular activities, therefore inferred toxicity response and mechanism, as a function of either dose or exposure time condition changes. This highlighted the need to have research paradigm shift from current dominantly individual gene inspection to systems biology. As high-throughput techniques show promises and become more predominant, it is desired to take the full advantage of the comprehensiveness of the data from higher levels of investigation, rather than put too much emphasis on isolated data points. In the same time, it also calls for a more prudent toxicity assessment strategy, namely one that is able to consider multiple factors, like doses, time after exposure, while evaluating toxicity of certain substances. We would like to point out the fact in this study the two factors of interest are studied separately, while it is highly possible they have statistical interactive effects. We hope refer interested researchers to another study of ours in which multiple factors could be analyzed simultaneously.
Chapter 8

Conclusions and Future Outlook
In this dissertation, I proposed and applied different unsupervised data mining techniques to improve data analysis of high dimensional time series toxicogenomics data in order to obtain more systematic, holistic, in-depth and finer-resolution toxicological information for toxicity identification and characterization. The novelty of this study resized on the following two aspects. First, we emphasized on developing new or improved bioinformatics methodologies that capture the temporal dynamics of molecular responses to enable the evaluation of the impact and importance of time factor in toxicological research. Second, we also incorporated various molecular effect quantifiers into the toxicogenomics data analysis for more quantitative and comparative analysis, which have never been explored previously. In addition to the specific conclusions drawn in the individual chapters (which are not repeated here), this chapter summarizes more general insights in bioinformatics applications for high dimensional temporal toxicogenomics data analysis and research. The needs in future research and suggested follow-up work are also discussed.

8.1. Conclusions

In this section, I summarized some general insights concluded while investigating different bioinformatics applications in response to the analytical challenges in high dimensional time series toxicogenomics data analysis.

- The data mining techniques studies proposed and employed in this dissertation considered temporal resolution of the gene expression data, and revealed the time dependent and dynamic nature of cellular activities. The applications of these time-aware bioinformatics techniques led to better description and quantification of molecular effects and therefore more comprehensive understanding of toxicant-
induced molecular responses captured by toxicogenomics data. The results demonstrated the impact and importance of time factor and dynamic nature of cellular response and thus it is critical to include temporal information for the characterization and toxic mechanisms identification of toxic substances.

- High-throughput genomic techniques and proper experimental design enhance the systematic understanding in toxicogenomics studies. The advances of techniques such as whole cell array make it much simpler to monitor the molecular activities inside cells continuously, which help the design of more sophisticated experiments that can simultaneously take into account several factors that could have significant influences on the toxicological conclusions. This improved assay resolution and better experimental design can guide toxicogenomics research into more precise examination of functional pathways and biological systems.

- Rich functional insights can be obtained from analyzing high dimensional toxicogenomics data with appropriate analytical approaches. In contrast to conventional tools that take low dimensional data as input, techniques like common principal component analysis (CPCA), parallel factor analysis (PARAFAC) and gene co-expression network analysis could examine inner relations among dimensions, producing more systematic understanding of dose- or time-dependent gene expression patterns, as well as more accurate mechanistic conclusions.

- Different Molecular effect quantifiers that quantify individual gene expression alternation magnitude, pathway activation extent or system response
characteristics proposed in this study have been shown to be valid and therefore can be employed and incorporated into various toxicogenomics analysis.

- Good analytical software facilitates experimental design, data management, analysis and visualization. The development of analytical software will not only provide data mining solutions, but also help researchers organize their studies properly, prompt better practices in the design of experiments. In addition, such software platform will also help process the raw data in a more reliable manner, and improve the reproducibility of the overall research.

### 8.2. Overview of Future Studies

Toxicogenomics data are often affected by many experimental factors such as exposure time, chemical dose, biomarker/genes selection and other experimental conditions. The advances in toxicogenomics technologies now allow high-throughput and quick generation of large amount data with multiple dimensions that consider above factors. Thus proper and compatible data presentation and bioinformatics tools are needed. In addition, flexible platforms that allow easy customization are also desired. Appropriate and effective bioinformatics tools for data visualization, data quality check and data analysis can also be used to guide experimental design and optimization. To follow up and continue the work performed in this study, some possible directions for future study are recommended as following:

- Further extend the data analysis software Timevis into a more comprehensive platform that can also be used to assist in the experimental design, quality control, outlier detection, data normalization and processing. Automated data visualization in combination with reliable statistical methods and normalization algorithms will
provide an overall evaluation of data quality and help reduce potential systematic errors.

- Exploring alternative algorithms to better reflect time-dependent features such as the employment of ARMA model. Better feature extraction techniques may be used for dimension reduction, such as statistical models like autoregressive-moving-average model may be promising to incorporate time-dependent nature of the gene expression profiles (Box, Jenkins, & Reinsel, 2011). Some high dimensional clustering algorithm, including biclustering and triclustering, can also be explored (Tanay, Sharan, & Shamir, 2002; Zhao & Zaki, 2005).

- Gene ranking is essential step in many analyses such as gene enrichment analysis, pathway activation analysis and biomarkers identification etc. Investigation and comparison of multiple gene ranking approaches, particularly those that consider temporal dynamics and features of high dimensional toxicogenomics data, will be beneficial to reveal the advantages and limitations associated with reach approach. Furthermore, an ensemble of multiple ranking methods may be used together depending the emphasis of the particular study.

- Further explore the applications of different tensor decomposition models besides PARAFAC to identify key components and underlying the integrated/combined cellular response to toxicants. Consider that the high dimensional toxicogenomics data from different toxicants may have components that have non-linear relationships; more advanced models can be examined.
Integrating more topological network characteristics and centrality metrics into functional analysis in toxicogenomics. More detailed study on the conservative gene-gene connections may reveal some basic functionality in the cells.
REFERENCE


improve drug candidate selection and to understand mechanisms of chemical toxicity and action. Journal of Biotechnology, 119(3), 219-244.


recombinant yeast for trout estrogen receptor and trout hepatocyte cultures. 
Journal of Molecular Endocrinology, 19(3), 321-335.
PLOS Genetics, 5(9), e1000642.
generalized singular value decomposition for comparison of global mRNA 
expression from multiple organisms. PLOS One, 6(12), e28072.
to summarize microarray experiments: application to sporulation time series. 
Paper presented at the Pacific Symposium on Biocomputing, Honolulu, HI
oxidative stress and cytotoxicity by carbon nanomaterials is dependent on 
physical properties. Toxicology and Industrial Health, 27(1), 3-10.
Riccadonna, S., Jurman, G., Visintainer, R., Filosi, M., & Furlanello, C. (2012). DTW- 
MIC coexpression networks from time-course data. arXiv preprint 
chemicals policy": A review. Risk Analysis, 23(2), 381-388.
Roy, S., Bhattacharyya, D. K., & Kalita, J. K. (2014). Reconstruction of gene co-
expression network from microarray data using local expression patterns. BMC 
Bioinformatics, 15(Suppl 7), S10.
bioinformatics. Bioinformatics, 23(19), 2507-2517.
generation and performance profiling of network inference methods. 
Bioinformatics, 27(16), 2263-2270.
expression time-courses. IEEE/ACM Transactions on Computational Biology and 
Bioinformatics, 2(3), 179-193.
interaction networks through time-lagged correlation analysis of transcriptional 
data. Genome Research, 14(8), 1654-1663.
using NetworkX. Paper presented at the Proceedings of the 7th Python in Science 
Conferences (SciPy 2008), Pasadena, CA.


APPENDIX
Appendix A  Analyzing High Dimensional Toxicogenomics Data via Consensus Clustering

Table A-1. Specifications of chemicals and nanomaterials used in the study.

<table>
<thead>
<tr>
<th>Chemical Name</th>
<th>Symbol</th>
<th>Dominant Toxic Mechanisms¹</th>
<th>Concentration (mg/L)</th>
<th>Manufacturer</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitomycin C</td>
<td>MMC</td>
<td>DNA stress²</td>
<td>0.0003, 0.003, 0.03, 0.3</td>
<td>Fisher Scientific, Pittsburgh, PA</td>
<td></td>
</tr>
<tr>
<td>mercury</td>
<td>Hg²⁺</td>
<td>Protein stress, oxidative stress</td>
<td>0.001, 0.01, 0.05, 0.1</td>
<td>Fisher Scientific, Pittsburgh, PA</td>
<td>Mercury Nitrate</td>
</tr>
<tr>
<td>hydrogen peroxide</td>
<td>H₂O₂</td>
<td>Oxidative stress, membrane stress³, DNA stress</td>
<td>0.1, 1, 5, 10</td>
<td>Fisher Scientific, Pittsburgh, PA</td>
<td></td>
</tr>
<tr>
<td>4-Nonylphenol</td>
<td>4NNP</td>
<td>Protein stress, oxidative stress</td>
<td>0.1, 1, 5, 15</td>
<td>Fisher Scientific, Pittsburgh, PA</td>
<td></td>
</tr>
<tr>
<td>carbon black</td>
<td>CB</td>
<td>oxidative stress</td>
<td>1, 5, 50</td>
<td>Fisher Scientific, Pittsburgh, PA</td>
<td></td>
</tr>
<tr>
<td>nano silver particles</td>
<td>nAg</td>
<td>oxidative stress, DNA damage, membrane stress, protein stress</td>
<td>1, 5, 10, 50, 100</td>
<td>NanoDynamics Inc., Buffalo, NY</td>
<td>Diameter: ~60nm</td>
</tr>
<tr>
<td>nano titanium dioxide</td>
<td>TiO₂_r</td>
<td>very low toxicity</td>
<td>1, 10, 50</td>
<td>Sigma-Aldrich, MO</td>
<td>Thickness: ~10nm (40nm laterally)</td>
</tr>
<tr>
<td>rutile</td>
<td>TiO₂_a</td>
<td>oxidative stress, DNA damage, membrane stress</td>
<td>1, 10, 50</td>
<td>NanoStructured &amp; Amorphous Materials, Houston, Texas</td>
<td>~10nm</td>
</tr>
<tr>
<td>fullerene soot</td>
<td>F</td>
<td>low oxidative stress</td>
<td>1, 10, 50</td>
<td>M.E.R.Co, Tuscon, AZ</td>
<td>Diameter: 1-2 nm; length: 5-30μm; pre-sonicated</td>
</tr>
<tr>
<td>single-walled nanotube</td>
<td>SWNT</td>
<td>membrane stress and oxidative stress</td>
<td>0.32, 1.6, 8</td>
<td>Nantero Inc., Woburn, MA</td>
<td></td>
</tr>
<tr>
<td>oxidized single-walled nanotube</td>
<td>SWNT_OX</td>
<td>membrane stress and oxidative stress</td>
<td>0.1, 1, 5</td>
<td>Nantero Inc., Woburn, MA</td>
<td></td>
</tr>
</tbody>
</table>

1. The toxic mechanisms are based on our previous studies on *E. coli* (Gou & Gu, 2011; Onnis-Hayden et al., 2009).
2. The dominant mechanisms of a toxicant (if identified) are in bold font and underlined.
3. Membrane stress includes those genes related to membrane transporter/proin, multidrug efflux system, lipid and lipopolysaccharide biosynthesis, flagella biosynthesis, and proton/electron transporter.
Table A-2. Summary of the characterization data for all the nanomaterials.

<table>
<thead>
<tr>
<th>Chemical Name</th>
<th>Abbreviation</th>
<th>SSA (m² g⁻¹)</th>
<th>zeta-potential (mv)</th>
<th>PdI</th>
<th>ROS (mg⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>carbon black</td>
<td>CB</td>
<td>110.6</td>
<td>-27.3</td>
<td>0.23</td>
<td>300.17</td>
</tr>
<tr>
<td>nano silver particles</td>
<td>nAg</td>
<td>8.4</td>
<td>-26.2</td>
<td>0.45</td>
<td>381.80</td>
</tr>
<tr>
<td>nano titanium dioxide rutile</td>
<td>TiO2_r</td>
<td>189.9</td>
<td>-24.1</td>
<td>0.30</td>
<td>252.13</td>
</tr>
<tr>
<td>nano titanium dioxide anatase</td>
<td>TiO2_a</td>
<td>274.2</td>
<td>-33.9</td>
<td>0.35</td>
<td>256.57</td>
</tr>
<tr>
<td>fullerene soot</td>
<td>F</td>
<td>194.3</td>
<td>-39.1</td>
<td>0.15</td>
<td>235.07</td>
</tr>
<tr>
<td>single-walled nanotube</td>
<td>SWNT</td>
<td>510.5</td>
<td>-8.8</td>
<td>0.33</td>
<td>925.43</td>
</tr>
<tr>
<td>oxidized single-walled nanotube</td>
<td>SWNT_OX</td>
<td>1154</td>
<td>-29.5</td>
<td>0.41</td>
<td>1000.10</td>
</tr>
</tbody>
</table>

SSA: Specific Surface Area; PdI: Polydispersity Index; ROS: Reactive Oxygen Species. ROS production are measure by Dichlorofluorescin diacetate (DCFH) assay, and expressed in H₂O₂ equivalent unit.
Figure A-1. Cluster number determination by finding the peak cumulative distribution function (CDF) of area change.
Figure A-2. Consensus clustering using (a) the maximum expression values of each gene, (b) the 10\textsuperscript{th} time points, (c) the 20\textsuperscript{th} time points. Y axis: CI value.
Figure A-3. Consensus clustering using TELI values for each gene. Y-axis: CI value.

Figure A-4. Consensus clustering with dataset built with only redox stress genes. Y-axis: CI value.
Appendix B  Toxicity Mechanisms Identification via Gene Set Enrichment Analysis of Time-Series Toxicogenomics Data: Impact of Time and Concentration

Table B-1. Genes included in the *E. coli* stress response ensemble assay library for toxicogenomics assessment.

<table>
<thead>
<tr>
<th>Category</th>
<th>Gene selected</th>
<th>Known functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Redox stress</strong></td>
<td><em>soxS, soxR, oxyR, inaA, dps, ahpF, katG, sodA, ahpC, katE, ytfE, katE, sodB, sodC</em></td>
<td>Increased levels of superoxides, increased levels of peroxides, any other conditions which alter the redox potential of the cell.</td>
</tr>
<tr>
<td><strong>Detoxification</strong></td>
<td><em>norR, fpr, tam, yeiG, yafN, yeaE, grxA, gst</em></td>
<td>Organisms protect themselves against the harmful effects of toxic compounds</td>
</tr>
<tr>
<td><strong>Cell killing</strong></td>
<td><em>dinJ, slyA, yeeV, ylfG, relB</em></td>
<td>Related to the killing of own cells or another organism</td>
</tr>
<tr>
<td><strong>General stress</strong></td>
<td><strong>Cold shock</strong> <em>cspA, cspB</em></td>
<td>Temperature downshift</td>
</tr>
<tr>
<td></td>
<td><strong>General function</strong> <em>phoB, crp, cdaR, ydeO, ybgI, gadX</em></td>
<td>Disturbance of the biochemical and biophysical homeostasis of the cell.</td>
</tr>
<tr>
<td>Chemicals</td>
<td>Symbols</td>
<td>Concentrations</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>mitomycin</td>
<td>MMC</td>
<td>5 ng/L, 50 ng/L, 500 ng/L, 5 μg/L, 50 μg/L,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500 μg/L</td>
</tr>
<tr>
<td>hydrogen peroxide</td>
<td>H₂O₂</td>
<td>2.5 mg/L, 5 mg/L, 10 mg/L, 20 mg/L, 40 mg/L,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>80 mg/L</td>
</tr>
<tr>
<td>lead nitrate</td>
<td>Pb²⁺</td>
<td>0.125 μg/L, 1.25 μg/L, 12.5 μg/L, 12.5 μg/L,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12.5 mg/L</td>
</tr>
</tbody>
</table>
## Appendix C  
High dimensional Toxicogenomics Data Analysis Using Tensor Decomposition Model for Toxicity Identification and Characterization

### Table C-1. Specifications of chemicals used in this study.

<table>
<thead>
<tr>
<th>Chemicals</th>
<th>Symbols</th>
<th>Concentrations</th>
<th>Suggested possible mechanisms of toxicity</th>
<th>Manufacturers</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-n-Nonylphenol</td>
<td>4NPN</td>
<td>280 ng/L, 2.8 μg/L, 280 μg/L, 2.8 mg/L, 28 mg/L</td>
<td>Protein stress (H. R. Andersen et al., 1999a; Petit, Le Goff, Cravedi, Valotaire, &amp; Pakdel, 1997a), redox stress (Bonefeld-Jorgensen, Long, Hofmeister, &amp; Vinggaard, 2007b; Chitra, Latchoumycandane, &amp; Mathur, 2002a)</td>
<td>Sigma-Aldrich, MO</td>
</tr>
<tr>
<td>Trichloroacetic acid</td>
<td>TCA</td>
<td>80 μg/L, 800 μg/L, 8 mg/L, 80 mg/L, 800 mg/L, 8 g/L</td>
<td>DNA damage (Pereira, Kramer, Conran, &amp; Tao, 2001), redox stress (American Conference of Governmental Industrial Hygienists., 2001; Hassoun, Cearfoss, &amp; Spildener, 2010a)</td>
<td>Fisher BioReagents, PA</td>
</tr>
<tr>
<td>Carbaryl</td>
<td>CAR</td>
<td>100 ng/L, 1 μg/L, 10 μg/L, 100 μg/L, 10 μg/L</td>
<td>DNA damage (Dean Blevins et al., 1977), General stress(Jansen et al., 2013)</td>
<td>Sigma-Aldrich, MO</td>
</tr>
<tr>
<td>Mitomycin C</td>
<td>MMC</td>
<td>5 ng/L, 50 ng/L, 500 ng/L, 5 μg/L, 50 μg/L, 500 μg/L</td>
<td>DNA damage (International Agency for Research on Cancer &amp; IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 1988b)</td>
<td>Research Products International, IL</td>
</tr>
</tbody>
</table>

### Table C-2. Genes used in the study and their pathways.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Gene Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein related stress</td>
<td>clpB, dnaJ, dnaK, entC, grpE, lon, rpoD, yfe</td>
</tr>
</tbody>
</table>
Appendix D  Identifying Dose and Time Dependent Impacts on Gene Co-Expression Networks in Toxicogenomics

Table D-1. Genes and their pathways used in the study for dose-related study.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Gene Selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA damage related stress</td>
<td><em>ada</em>, <em>dinB</em>, <em>dinG</em>, <em>dnaQ</em>, <em>lea</em>, <em>ftsK</em>, <em>mug</em>, <em>mutH</em>, <em>mutM</em>, <em>mutS</em>, <em>mutT</em>, <em>mutY</em>, <em>nfo</em>, <em>polA</em>, <em>polB</em>, <em>recA</em>, <em>recE</em>, <em>recN</em>, <em>recX</em>, <em>rnt</em>, <em>ruvA</em>, <em>sbmC</em>, <em>ssb</em>, <em>sulA</em>, <em>umuD</em>, <em>uvrA</em>, <em>uvrC</em>, <em>uvrD</em>, <em>uvrY</em>, <em>ybfE</em>, <em>yebG</em>, <em>yjiW</em>, <em>ykJG</em></td>
</tr>
<tr>
<td>General stress</td>
<td><em>bolA</em>, <em>crp</em>, <em>cspA</em>, <em>cspB</em>, <em>dinJ</em>, <em>gadX</em>, <em>inaA</em>, <em>otsB</em>, <em>slyA</em>, <em>uspA</em>, <em>uspB</em>, <em>ydeO</em>, <em>ydgL</em></td>
</tr>
<tr>
<td>Membrane stress</td>
<td><em>amiC</em>, <em>bacA</em>, <em>cls</em>, <em>dacA</em>, <em>dacB</em>, <em>fepB</em>, <em>flgM</em>, <em>fsr</em>, <em>midK</em>, <em>motA</em>, <em>mrcB</em>, <em>ompC</em>, <em>phpG</em>, <em>sanA</em>, <em>sbmA</em>, <em>ycgE</em>, <em>yedW</em>, <em>yfjX</em>, <em>cmr</em>, <em>cruR</em>, <em>emrA</em>, <em>emrE</em>, <em>marC</em>, <em>marR</em>, <em>yajR</em>, <em>zntA</em></td>
</tr>
<tr>
<td>Protein related stress</td>
<td><em>clpB</em>, <em>dnaJ</em>, <em>dnaK</em>, <em>entC</em>, <em>grpE</em>, <em>lon</em>, <em>rpoD</em>, <em>yfeE</em></td>
</tr>
<tr>
<td>Oxidative stress</td>
<td><em>ahpC</em>, <em>ahpF</em>, <em>dps</em>, <em>katE</em>, <em>katG</em>, <em>oxyR</em>, <em>sodA</em>, <em>sodB</em>, <em>sodC</em>, <em>soxR</em>, <em>soxS</em>, <em>cyoA</em>, <em>fpr</em>, <em>grxA</em>, <em>gst</em>, <em>norR</em>, <em>tam</em>, <em>yajN</em>, <em>yeaE</em>, <em>yeiG</em></td>
</tr>
</tbody>
</table>