Doctoral Dissertation

# Study on Biologically Relevant Toxicogenomic Predictive Modeling Using Machine Learning

Keisuke Nagata

June 2016

Graduate School of Engineering
Osaka University

# Preface

This dissertation presents my research on exploration of biologically relevant toxicogenomic predictive modeling using machine learning techniques. The dissertation is the result of the research before and during the Ph.D. course at Division of Electrical, Electronic and Information Engineering, Information and Communication Technology, Graduate School of Engineering, Osaka University. The dissertation is organized as follows.

In Chapter 1, we describe the background, challenges, notations, data sources, analytical methods, and the outline of this dissertation. The main purpose of this dissertation is finding a method for building a toxicogenomic model that is accurate and biologically relevant. In this end, we apply three different families of methods, namely class association rule mining, group-based sparse regularization and graph-based sparse regularization, to build a toxicogenomic model that predicts liver weight gain based on earlier microarray data, and compare their predictive performances and biological relevancy with those of conventional methods.

In Chapter 2, we discuss the Classification Based on Association (CBA) algorithm and its predictive performances and interpretability compared with linear discriminant analysis (LDA). CBA is one of the Class Association Rule (CAR) mining algorithms, which integrate association rule mining (finding all the rules existing in the database that satisfy some constraints) and classification rule mining (discovering a small set of rules in the database that forms an accurate classifier) by focusing on mining a special subset of association rules, called class association rules (CARs).

In Chapter 3, we discuss the latent group Lasso (LGL) method and its predictive performances and biological relevancy compared with Lasso. LGL is a variant of group-based structured regularization techniques and uses a specialized norm as a penalty term in regression to select explanatory variables as unions of groups. To do so, we regard genes regulated by the same transcription factor as a group.

In Chapter 4, we discuss the generalized fused Lasso (GFL) and graph Lasso (GL) methods and their predictive performances and biological relevancy compared with Lasso and LGL. GFL and GL are variants of graph-based structured regularization techniques. In contrast to group-based techniques such as LGL, graph-based techniques select explanatory variables as a set of edges in a graph, not of groups, therefore expected to lead to sparser modeling. To do so, we rearrange the structure given as a set of groups into a graph structure by drawing edges between transcription factors and their downstream genes.

Finally, Chapter 5 concludes this dissertation, with additional discussion.

# Acknowledgement

I worked on this dissertation under the supervision of Prof. Takashi Washio. First, I would like to thank him for sharing his knowledge, encouraging me to focus on my research, and constantly giving pieces of advice on the research, presentation materials and, of course, this dissertation. In addition, my research cannot be done without the help of Assoc. Prof. Yoshinobu Kawahara. I appreciate his support in directing me throughout my research.

I am grateful to Prof. Babaguchi in Osaka University for helping me improving dissertation much better.

Besides, I would like to thank Prof. Komatani, Prof. Takine, Prof. Sampei, Prof. Miyaji, and Prof. Inoue in Osaka University for their support throughout semesters and reviewing this dissertation.

I also thank my colleagues in Astellas Pharma, Inc for their support.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background

New technologies such as DNA microarray and next-generation sequencer have allowed researchers to learn biological phenomena in genome or transcriptome levels. Especially in toxicology, these new technologies have led to a new subdiscipline, termed toxicogenomics. Toxicogenomics is concerned with the identification of potential human and environment toxicants, and their putative mechanisms of action, through the use of genomics resources (Nuwaysir, Bittner, Trent, Barrett, and Afshari, 1999). For example, by evaluating and characterizing differential gene expressions, in humans or animals, after exposure to drugs, it is possible to use complex expression patterns to predict toxicological outcomes and to identify mechanisms involved with or related to the toxic event (Suter, Babiss, and Wheeldon, 2004).

The concept of toxicogenomics-based approach is illustrated in (Figure 1.1). In a conventional preclinical toxicological study, experimental animals such as rats are repeatedly dosed with a chemical compound for a certain period of time (typically one week, four weeks, etc.) and then necropsied for toxicological evaluations. Incorporating toxicogenomics-based approach allows us to predict the outcome based on high dimensional data of DNA microarrays obtained after a shorter period of dosages (from one week to one day, for example), instead of waiting until adverse effects of compounds appear.

Traditionally, various techniques in machine learning such as k-nearest neighbors, linear discriminant analysis (LDA) and support vector machine (SVM) have been used to build such a predictive classifier (Phan, Quo, and Wang, 2006). In practical use, however, it is often insufficient to only predict an outcome (e.g. negative or positive), unless predictive accuracy is 100% or at least close to that. For example, a prediction that a new drug has some likelihood of carcinogenicity would only confuses doctors and researchers, if it cannot answer the reason in a biological manner. Therefore, we strongly need a method that can build a biologically relevant as well as accurate classifier to further utilize ever-increasing gene expression data.

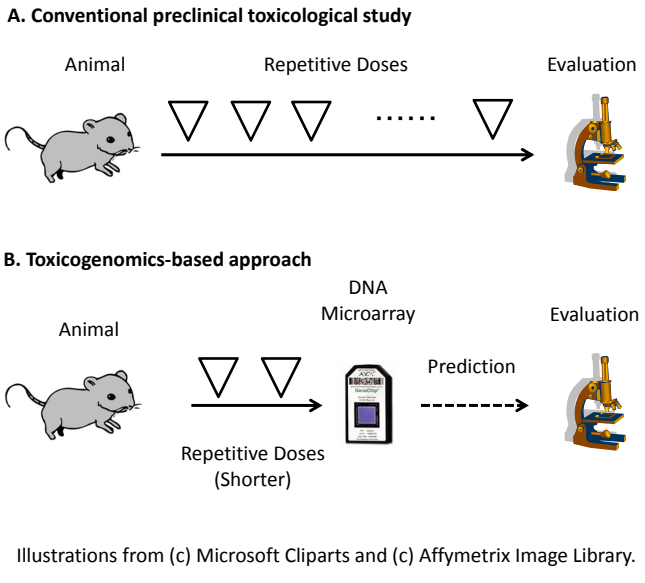


Figure 1.1: Conceptual diagram of toxicogenomics-based approach.

## 1.2 Challenges

Building a classifier that is accurate and biologically relevant at the same time is not necessarily an easy task. For example, while SVM achieves high classification accuracy, resulting classifiers are hard to interpret as variables are transformed nonlinearly into a feature space, and hence difficult to use in order to extract relevant biological knowledge from it (Ratsch, Sonnenburg, and Schafer, 2006).

Appropriate definitions of biological relevancy might vary depending on purposes of modeling. We defined that a biological relevant model is one that (1) mainly consists of known parameters reported to be involved with the outcome of interest and (2) not only classifies a sample but also gives an useful insight into biological mechanism behind that.

As was implied by the above example of SVM, biological relevancy of modeling methods is closely related to interpretability of their classifiers, since we first need to interpret a model in order to judge it as biological relevant or not. Although interpretability alone does not necessarily mean biological relevancy, we assume interpretability to be a necessary condition of biological relevancy. Besides, the difficulty in satisfying both accuracy and interpretability in modeling is fueled by the dilemma that predictive accuracy, interpretability, and computational demands often need to be traded off against one another, because algorithms often compromise one to gain performance in the other (Apte, Hong, Natarajan, Pednault, Tipu, and Weiss, 2003).

To tackle this challenge, we decide to clarify the characteristics of different machine learning techniques in building a biologically relevant as well as accurate toxicogenomic predictive model, focusing on three families of methods that are relatively new and applicable in this end: class association rule mining, group-based sparse regularization and graph-based sparse regularization. For this purpose, we apply them to build a toxicogenomic model that predicts liver weight gain based on earlier microarray data, and compare their predictive performances and biological relevancy with those of conventional methods. This enables us to reveal their advantages and disadvantages in toxicogenomic predictive modeling that is not only accurate, but also biologically relevant.

While hepatomegaly, including liver weight gain, without histological or clinical pathological alterations indicative of liver toxicity is usually considered an adaptive and non-adverse reaction, it appear to be correlated with the subsequent development of irreversible toxicity such as fibrosis, necrosis, vacuolization, fatty degeneration, and even neoplasia (Hall, Elcombe, Foster, Harada, Kaufmann, Knippel, Kuttler, Malarkey, Maronpot, Nishikawa, Nolte, Schulte, Strauss, and York, 2012). Therefore, early detections of liver weight gain based on gene expressions could help select safer compounds in pharmaceutical industry.

In the remainder of this chapter, we first explain the notations, data sources and analytical methods that are common throughout all the chapters. Then, we conclude the chapter with Summary of Contributions.

## 1.3 Notations

The notations used in this dissertation are explained in Table 1.1.

Table 1.1: Notations.

| Notation | Meaning |
|---|---|
| $\mathcal{P}([1,p])$ | the power set of $\{1,\cdots,p\}$ |
| $\|\|\mathbf{x}\|\|_p$ | the L$p$-norm of $\mathbf{x}\in\mathbb{R}^n$, defined as $(\sum_{i=1}^n \|\mathbf{x}_i\|^p)^{1/p}$. |
| $\|\mathbf{x}\|$ | the L1-norm of $\mathbf{x}\in\mathbb{R}^n$, the same as $\|\|\mathbf{x}\|\|_1$. |

## 1.4 Data Sources

### 1.4.1 TG-GATEs

For gene expression data and liver weight data before and after treatments of compounds in rats, we use the TG-GATEs database.

TG-GATEs is a toxicogenomic database developed by The Toxicogenomics Project (TGP), a joint government-private sector project organized by the National Institute of Biomedical Innovation, National Institute of Health Sciences and 15 pharmaceutical companies in Japan, and The Toxicogenomics Informatics Project (TGP2), a follow-on project from TGP organized by the National Institute of Biomedical Innovation, National Institute of Health Sciences and 13 companies. Gene expression and toxicity data in vivo (rats) and in vitro (primary cultured hepatocytes of rats and humans) after treatments of more than 150 compounds are stored in the TG-GATEs database. TG-GATEs is now released for public as Open TG-GATEs (http://toxico.nibio.go.jp).

### 1.4.2 MSigDB

For the list of transcription factors and their downstream genes as the structural information used in Chapter 3 and 4, we use the MSigDB database.

MSigDB is a collection of annotated gene sets (Subramanian, Tamayo, Mootha, Mukherjee, Ebert, Gillette, Paulovich, Pomeroy, Golub, Lander, and Mesirov, 2005) and publicly available on the Broad Institute's website (http://www.broadinstitute.org/gsea/msigdb/index.jsp). We used the 615 transcription factor target gene sets as groups, each of which shares a transcription factor binding site, from the motif gene sets (C3) of the MSigDB database. Of the 615 groups, we omitted those regulated by an unknown transcription factor.

Because genes are represented as human Entrez IDs in MSigDB, we map probe set IDs of Affymetrix GeneChip Rat Genome 230 2.0 Array to Entrez IDs of their corresponding human homologue genes with the aid of QIAGEN's Ingenuity® Pathway Analysis (IPA®. QIAGEN. Redwood City, CA, USA. http://www.ingenuity.com/products/ipa) software. We confirm that the gene sets in the C3 set of MSgiDB "contain genes that share a cis-regulatory motif that is conserved across the human, mouse, rat, and dog genomes," according to the website, and therefore such interspecies mapping is reasonable.

## 1.5 Analytical Methods

In this section, we describe the analytical methods that are common throughout all the chapters. We describe those peculiar to each chapter later in the corresponding chapters.

### 1.5.1 Modeling

To compare the predictive performances and biological relevancy of various methods, we build classifiers that predict whether a chemical compound induces increases in liver weight after 14-day repetitive treatments in rats based on transcriptomic data of 3-day repetitive treatments.

From the TG-GATEs database, we use gene expression data (n=3 per group) one day after 3-day repetitive doses (hereinafter 4D) in the liver of rats and liver weight data (relative liver weights calculated from body weights) (n=5 per group) one day after 14-day repetitive doses (15D) in rats for this study. For each compound, only the data of the highest dose group and its control group are used. Of 150 compounds, we omit one compound and analyze the remaining 149 compounds because that one compound is found to have killed animals before 15D in the study and therefore no data is available for liver weight of 15D.

### 1.5.2 Student's t-test

For statistical comparison of mean gene expressions or liver weights between a compound-treated group and its corresponding control group for each compound, the unpaired two tailed student's t-test without equal variance assumption is conducted. When gene expressions are compared between two groups, gene expressions are log-transformed with base of 2 prior to the test. Log transformations of gene expression data is known to result in more consistent statistical inferences and be often considered desirable, due to its large coefficient of variation (Long, Mangalam, Chan, Tolleri, Hatfield, and Baldi, 2001).

It is well known that the standard p-value method leads to the high rate of false positives when applied in repeated testing. This is the case when analyzing gene expression data collected via microarrays, as this usually involves testing from several thousands to tens of thousands of hypotheses simultaneously. While a number of adjustment procedures (e.g. controlling the false discovery rate) are available, they are often too conservative for microarray studies in that they can lead to low sensitivity (Pawitan, Michiels, Koscielny, Gusnanto, and Ploner, 2005), thus increasing the risk of missing true positives. In our studies, no adjustments are applied, taking it into consideration that even if false positive genes with no or little relevance for liver weights are detected by

statistical tests, the classification methods would discard many of them from a generated classifier. This is because the classification methods, including CBA and LDA, select genes not based on statistical differences, but based on predictive performance. Hence, the impact of such false positives is effectively marginalized while minimizing the risk of overlooking true important changes.

### 1.5.3 Pathway Analysis

We use QIAGEN's Ingenuity Pathway Analysis (IPA) software to understand what pathway or function an obtained set of genes was mainly involved with. IPA answers how statistically significantly a user-specified set of molecules are involved with pre-defined sets of molecules, considering how many molecules they share. IPA is based on "expertly curated biological interactions and functional annotations from millions of individually modeled relationships between proteins, genes, complexes, cells, tissues, drugs, and diseases" and "reviewed for accuracy by PhD scientists," according to QIAGEN's website (http://www.ingenuity.com/products/ipa).

## 1.6 Summary of Contributions

Below, we briefly summarize the contributions of each chapter:

- **Chapter 2:** We discuss the Classification Based on Association (CBA) algorithm and its predictive performances and interpretability compared with LDA. CBA is one of the Class Association Rule (CAR) mining algorithms, which integrate association rule mining (finding all the rules existing in the database that satisfy some constraints) and classification rule mining (discovering a small set of rules in the database that forms an accurate classifier) by focusing on mining a special subset of association rules, called class association rules (CARs) (Liu, Hsu, and Ma, 1998). This chapter is based on (Nagata, Washio, Kawahara, and Unami, 2014) (A-1).

- **Chapter 3:** We discuss the latent group Lasso (LGL) method and its predictive performances and biological relevancy compared with Lasso. LGL is a variant of group-based structured regularization techniques and uses a specialized norm as a penalty term in regression to select explanatory variables as unions of groups (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011). To do so, we regard genes regulated by the same transcription factor as a group. This chapter is based on (Nagata, Kawahara, Washio, and Unami, 2015) (A-2).

- **Chapter 4:** We discuss the generalized fused Lasso (GFL) and graph Lasso (GL) methods and their predictive performances and biological relevancy compared with Lasso and LGL. GFL (Tibshirani and Taylor, 2011) and GL (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011) are variants of graph-based structured regularization techniques. In contrast to group-based techniques such as LGL, graph-based techniques select explanatory variables as a set of edges in a graph, not of groups, therefore expected to lead to sparser modeling. To do so, we rearrange the structure given as a set of groups into a graph structure by drawing edges between transcription factors and their downstream genes. This chapter is based on (Nagata, Kawahara, Washio, and Unami, 2016) (A-3).

# Chapter 2

# Class Association Rule Mining

## 2.1 Introduction

In this chapter, we introduce the Classification Based on Association (CBA) algorithm and compare its predictive performances and interpretability with LDA in our toxicogenomic modeling.

CBA is one of the Class Association Rule (CAR) mining algorithms, which integrate association rule mining (finding all the rules existing in the database that satisfy some constraints) and classification rule mining (discovering a small set of rules in the database that forms an accurate classifier) by focusing on mining a special subset of association rules, called class association rules (CARs) (Liu, Hsu, and Ma, 1998).

One of the advantages of CAR mining algorithms over conventional methods (especially SVM) is its interpretability, because classifiers are generated as a set of simple rules without much sacrifice of accuracy (Pach, Gyenesei, and Abonyi, 2008). Another advantage is that CAR mining algorithms can be applied not only to linearly separable cases, but also to linearly inseparable cases, where LDA or other linear classification methods are not applicable (Sampson, Parker, Upton, and Hurst, 2011). SVM can handle linearly inseparable cases by mapping original data into a suitable feature space, but with loss of interpretability. Besides, especially when applied to gene expression data, CAR mining algorithms, which predict a class label based on specific sets of differentially expressed genes that are actually observed in training samples, are expected to generate more biologically relevant classifiers, because it is generally not individual genes but sets of genes that collectively define phenotypes such as drug responses (Bateman, El-Hachem, Beck, Aerts, and Haibe-Kains, 2014).

While applications of CBA and its variants in biological research were reported in several reports (Chiu, Chen, Yuan, and Lin, 2006; Kianmehr and Alhajj, 2008; Tamura and D'Haeseleer, 2008; Dua and Kidambi, 2010; Paul, Groza, Hunter, and Zankl, 2014), there had been no reports with direct implication for toxicogenomics until our study (Nagata, Washio, Kawahara, and Unami, 2014). Toxicogenomics is unique in that the number of variables to be analyzed is usually

far much greater (more than 30,000 genes) than for other applications and this high dimensionality makes it difficult to analyze its data with conventional methods.

## 2.2 Classification Based on Association (CBA)

Here, we explain the basic concept of CBA based on (Liu, Hsu, and Ma, 1998) with examples in our setting.

Let $D$ be the dataset, a set of records $d$ ($d \in D$). Let $I$ be the set of all non-class items in $D$, and $Y$ be the set of class labels in $D$. In our study, the non-class item is a pair of gene ID and its discretized expression (Inc or Dec) (Inc: Increased, Dec: Decreased), and the class label is a pair of a target parameter (RLW: relative liver weight) and its discretized value (Inc or NI, or Dec or ND) (NI: Not Increased, ND: Not Decreased). Therefore, the set of class labels $Y$ in our study is either (RLW, Inc), (RLW, NI) or (RLW, Dec), (RLW, ND) . We say that a record $d \in D$ contains $X \subseteq I$, or simply $X \subseteq d$, if $d$ has all the non-class items of $X$. Similarly, a record $d \in D$ contains $y \in Y$, or simply $y \subseteq d$, if $d$ has the class label $y$.

A rule is an association of the form $X \rightarrow y$ (e.g. (Gene_01, Inc), (Gene_02, Dec) $\rightarrow$ (RLW, Inc)). For a rule $X \rightarrow y$, $X$ is called an antecedent of the rule and $y$ is called a consequence of the rule. A rule $X \rightarrow y$ holds in $D$ with confidence $c$ if $c\%$ of the records in $D$ that contain $X$ are labeled with class $y$. A rule $X \rightarrow y$ has support $s$ in $D$ if $s\%$ of the records in $D$ contain $X$ and are labeled with class $y$.

The objectives of CBA are (1) to generate the complete set of rules that satisfy the user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*) constraints, and (2) to build a classifier from these rules (class association rules, or CARs). The original CBA algorithm of (Liu, Hsu, and Ma, 1998) consists of two parts, a rule generator (called CBA-RG) and a classifier builder (called CBA-CB), each corresponding to (1) and (2).

The key operation of CBA-RG is to find all rules $X \rightarrow y$ that have support above *minsup*. A rule is called frequent if it satisfies *minsup*. Otherwise, a rule is called infrequent. For all the rules that have the same antecedent, the rule with the highest confidence is chosen as the possible rule (PR) representing this set of rules. If there are more than one rules with the same highest confidence, one rule is randomly selected. If the confidence is greater than *minconf*, the rule is accurate. The set of CARs thus consists of all the PRs that are both frequent and accurate. The CBA-RG algorithm effectively searches for all the CARs in a dataset based on the Apriori algorithm (Agrawal and Srikant, 1994), assuming the downward closure property that for any $X$, $X$ is frequent if and only if any subset $x$ of $X$ is frequent. Instead of CBA-RG, the Coenen's CBA program that we used for

our study is implemented with the Apriori-TFP algorithm (Coenen, Goulbourne, and Leng, 2004; Coenen, Leng, and Ahmed, 2004), a variant of the Apriori algorithms that utilizes a tree-structured data representations for a higher performance.

The operation of the latter part, CBA-CB, is described as follows in (Liu, Hsu, and Ma, 1998). "Given two rules, $r_i$ and $r_j$. $r_i \succ r_j$ (also called $r_i$ precedes $r_j$ or $r_i$ has a higher precedence than $r_j$) if

1. the confidence of $r_i$ is greater than that of $r_j$, or
2. their confidences are the same, but the support of $r_i$ is greater than that of $r_j$, or
3. both the confidences and supports of $r_i$ and $r_j$ are the same, but $r_i$ is generated earlier than $r_j$.

Let $R$ be the set of generated rules and $D$ the training data". CBA-CB is "to choose a set of high precedence rules in $R$ to cover $D$". A generated classifier is of the form, $<r_1$, $r_2$, ..., $r_n$, *default_class*$>$, where $r_i \in R$ and $r_a \succ r_b$ if $b > a$. In classifying a sample with a unknown class label, the first rule that satisfies the sample will classify it. If there is no rule that applies to the sample, it takes on the default class, *default_class*. Below is a simple example of classifiers.

Example:
(Gene_01, Inc), (Gene_02, Dec) $\rightarrow$ (RLW, Inc)
(Gene_01, Inc), (Gene_03, Inc) $\rightarrow$ (RLW, Inc)
(NULL) $\rightarrow$ (RLW, NI)

In this example. each line corresponds to a rule included in the classifier. The rule with the (NULL) antecedent means the default rule of this classifier. When a sample, (Gene_01, Inc), (Gene_03, Inc) with an unknown class label (it is unknown whether RLW is Inc or NI), is classified, the classifier answers (RLW, Inc), as the second rule first satisfies the sample. In another case, where a sample, (Gene_01, Inc), (Gene_02, Inc), is classified, the classifier answers (RLW, NI), as none of the rules except the default rule satisfies the sample and thus the default rule is applied.

## 2.3 Analytical Methods

### 2.3.1 Software

**CBA:**

In courtesy of Dr. Frans Coenen, we used a CBA program available on the LUCS-KDD

website, which is implemented according to the original algorithm by (Liu, Hsu, and Ma, 1998), except that CARs are first generated using the Apriori-TFP algorithm instead of the CBA:A-RG algorithm.

**LDA:**

We used the lda function in the MASS library of R. R's lda function is implemented based on Rao's LDA (Rao, 1948; Venables and Ripley, 2002), also known as Fisher-Rao LDA, which generalized Fisher's LDA (Fisher, 1936) to multiple classes.

### 2.3.2 Data Process

**CBA:**

Prior to the CBA analysis, we have preprocessed gene expression data in the liver (4D) and liver weight data (15D) of rats after repetitive doses for 149 compounds from the TG-GATEs database. First, gene expressions were corrected and normalized by the MAS 5.0 algorithm (Hubbell, Liu, and Mei, 2002) to reduce inter-array variances (Welle, Brooks, and Thornton, 2002). Liver weights were transformed into relative liver weight, a ratio of liver weight divided by body weight to avoid large variations in body weight skewing organ weight interpretation (Hall, Elcombe, Foster, Harada, Kaufmann, Knippel, Kuttler, Malarkey, Maronpot, Nishikawa, Nolte, Schulte, Strauss, and York, 2012). Secondly, values were averaged over individual animals included in each group. Then, for each compound-treated group, a fold change was calculated as a ratio of an average value of a treatment group divided by an average value of its corresponding control group, to reduce inter-study variances (Cheng, Shen, Song, Luo, and Tseng, 2009). Finally, we discretized gene expressions and relative liver weights based on their fold changes (fc) and p values (p) of the student's t-test conducted between a compound-treated group and its corresponding control group, according to the criteria shown below.

**Gene expression data:**

If fc $> 2$ and p $< 0.05$, assign "Inc" (Increased).
If fc $< 0.5$ and p $< 0.05$, assign "Dec" (Decreased).
Otherwise, assign "NC" (Not Changed).

**Liver weight data:**

1. When a classifier for increased liver weight was built:
   If fc $> 1$ and p $< 0.05$, assign "Inc" (Increased).

Otherwise, assign "NI" (Not Increased).

2. When a classifier for decreased liver weight was built:
   If fc $< 1$ and p $< 0.05$, assign "Dec" (Decreased).
   Otherwise, assign "ND" (Not Decreased).

Discretization thresholds for gene expressions combined with fold changes and statistical test (e.g. student's t-test) have often been applied in microarray data analysis and is reported to be better than p-value alone (McCarthy and Smyth, 2009). In general, numerical parameters obtained in toxicity studies are judged to be increased or decreased, based essentially on statistical comparison with contemporary controls and, if available, additionally on historical data (Festing and Altman, 2002). In this study, we discretized liver weights based only on statistical tests, as no historical data was available.

Before proceeding to CBA, gene expressions discretized as "NC" in each group were discarded from the data, because we were interested only in genes with increased or decreased expressions. We then analyzed the data with CBA, with discretized gene expressions as non-class items and discretized liver weights as class labels.

**LDA:**

Prior to the LDA analysis, the data were preprocessed as in the CBA analysis, except that gene expressions were not discretized. Before proceeding to LDA, the feature selection step was conducted to reduce the number of genes, because classical LDA requires the total scatter matrix to be nonsingular, while the matrix can be singular when the sample size (149) does not exceed the number of features (genes) (more than 30,000) (Ye, Xiong, Li, Janardan, Bi, Cherkassky, and Kambhamettu, 2006), and tends to overfit and become less interpretable in the presence of many irrelevant and/or redundant features (Gu, Li, and Han, 2011). Based on the previous reports on microarray data analysis (Kondoh, Ohkura, Arai, Hada, Ishikawa, Yamazaki, Shindoh, Takahashi, Kitagawa, Matsubara, and Yamamoto, 2007; Shi, Bugrim, Nikolsky, Nikolskya, and Brennan, 2008), we selected only the genes that were up-regulated (fc $> 2$ and p $< 0.05$) or down-regulated (fc $< 0.5$ and p $< 0.05$) in the groups with increased or decreased liver weight when compared to the not-increased or not-decreased groups, respectively.

### 2.3.3 Comparison of Predictive Performances

To compare predictive performances of CBA and LDA, we conducted 10-fold cross validation (Ambroise and McLachlan, 2002) for each method. And then, we evaluated accuracy, sensitivity, and specificity averaged over 10 validations. These parameters are defined as follows (Florkowski, 2008).

**Sensitivity:** True Positive / (True Positive + False Negative)

**Specificity:** True Negative / (True Negative + False Positive)

**Accuracy:** (True Positive + True Negative) / Total

### 2.3.4 Pathway Analysis

We conducted canonical pathway analysis with IPA software using all the genes included in our CBA-generated classifier. During the analysis, used genes are mapped to their corresponding molecules and matched up against the molecules in pre-defined sets of molecules.

### 2.3.5 Computer

We used a personal computer with Intel Core i5-3320M 2.6 GHz CPU and 4 GB RAM for the analyses in this chapter.

## 2.4 Results

### 2.4.1 Selection of Minimum Support and Confidence

In CBA, a user must specify two parameters: minimum support (*minsup*) and minimum confidence (*minconf*). There is no universal criteria for these parameters. In our study, we assumed that lower *minsup* and higher *confidence* are basically desirable. That is to say, a rule is considered useful, if the rule $X \rightarrow y$ satisfies a large fraction of records that matches the rule antecedent $X$, even if the number of records that matches $X$ is small. This is because a drug-induced response (or more generally biological response) is considered to be not caused by a single mechanism. Rather, it is expected that there are several different mechanisms, thus different gene expression patterns, finally leading to the target drug-induced response, and that each gene expression pattern occurs in a relatively low frequency among the dataset even if the dataset contains an enough number of

records with the target drug-induced response. If set too strict, however, there is a risk of missing useful rules with few exceptions for too high *minconf*, or of selecting accidental rules with only a few satisfying records for too low *minsup*. Moreover, *minsup* is also limited by computational resources, as the lower the *minsup* is set, the higher the computational demand is, in terms of both time and memory.

To explore the ideal settings of *minsup* and *minconf*, we first evaluated accuracy of CBA classifiers for increased liver weight in 10-fold cross validations under various combinations of *minsup* and *minconf* (Table 2.1). Times needed to finish the cross validations under each setting were also shown (IM: failed due to insufficient memory). First, we fixed the *minsup* at 10% and changed the *minconf* from 50% to 100%. While the *minconf* at 90% marked the highest accuracy (79%), there were no obvious differences or tendency in accuracy among the different *minconfs*. Next, we fixed the *minconf* at 90% and changed the *minsup* from 20% downward. Lowering the *minsup* remarkably improved accuracy, but prolonged computational time at the same time. The accuracy reached at 83% with *minsup* at 8%. We tried with *minsup* at 7%, but failed to finish the computation due to memory insufficiency. Similar tendencies were also confirmed when assessing accuracy of classifiers for decreased liver weight under different *minsups* and *minconfs* (data not shown).

Table 2.1: Effects of various *minsup*s and *minconf*s.

(A) When *minsup* was fixed at 10%

| *minsup* (%) | *minconf* (%) | Average accuracy (%) | Total time (s) |
|---|---|---|---|
| 10 | 50 | 77 | 0.61 |
| 10 | 80 | 76 | 0.59 |
| 10 | 90 | 79 | 0.58 |
| 10 | 100 | 77 | 0.58 |

(B) When *minconf* was fixed at 90%

| *minsup* (%) | *minconf* (%) | Average accuracy (%) | Total time (s) |
|---|---|---|---|
| 20 | 90 | 0 | 0.42 |
| 15 | 90 | 9 | 0.42 |
| 10 | 90 | 79 | 0.58 |
| 8 | 90 | **83** | 22.37 |
| 7 | 90 | IM | IM |

Based on these results, we adopted the *minsup* at 8% and *minconf* at 90% for the following analyses.

### 2.4.2 Predictive Performance

We compared predictive performance of classifiers between CBA and LDA with 10-fold cross validation (Table 2.2). Direction specifies which direction (increased or decreased) a classifier of relative liver weight was built for. TP, FN, FP, and TN are the average numbers of true positive, false negative, false positive, and true negative records in a test set. Acc, Sen, and Spe are accuracy, sensitivity, and specificity, respectively. Hold is an average number of records in a test set that did not match any rules except the default rule (only for CBA-DR). When increased liver weight was targeted (i.e. when a classifier for increased liver weight was built), CBA outperformed LDA in all of the three criteria: accuracy (83% for CBA vs. 75% for LDA), sensitivity (82% vs. 72%), and specificity (85% vs. 75%). When decreased liver weight was targeted, CBA scored better accuracy (86% vs. 73%) and sensitivity (22% vs. 6%), while LDA marked better specificity (90% vs. 95%).

We also compared between CBA and CBA-DR (our modified version of the original CBA) (Table 2.2). CBA-DR does not predict if a sample does not match any rule except the default rule in a classifier, and, in turn, return a 'hold'. When increased liver weight was targeted, CBA-DR marked lower accuracy (83% for CBA vs. 79% for CBA-DR) and specificity (85% vs. 29%) and higher sensitivity (82% vs. 100%). When decreased liver weight was targeted, CBA-DR marked lower sensitivity (22% for CBA vs. 0% for CBA-DR) and higher accuracy (86% vs. 95%) and specificity (90% vs. 100%).

Note that accuracy, sensitivity, and specificity for CBA-DR in (Table 2.2) were calculated excluding 'hold' samples, and that Totals are not integers since the number of records in the original data set was 149 and thus cannot be divided by 10, the number of trials for the cross validation in this study.

### 2.4.3 Interpretability

We compared the form of generated classifiers between CBA and LDA (Figure 2.1), when all the records were used as a training set for increased liver weight. CBA tells us a set of rules, arranged in order of confidence. Each rule consists of an antecedent, which is an itemset in the form of (non-class attribute, its discretized value), and a consequence in the form of (class attribute, its class label), shown after "->" here.

Table 2.2: Comparison of predictive performance.

| Method | Direction | Total | TP | FN | FP | TN | Hold | Acc (%) | Sen (%) | Spe (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| CBA | Inc | 14.9 | 4.4 | 1.1 | 1.4 | 8.0 | - | **83** | 82 | **85** |
| LDA | Inc | 14.9 | 2.7 | 1.0 | 2.8 | 8.4 | - | 75 | 72 | 75 |
| CBA-DR | Inc | 14.9 | 4.4 | 0 | 1.4 | 0.8 | 8.3 | 79 | **100** | 29 |
| CBA | Dec | 14.9 | 0.2 | 0.7 | 1.4 | 12.6 | - | 86 | 22 | 90 |
| LDA | Dec | 14.9 | 0.2 | 3.3 | 0.7 | 10.7 | - | 73 | 6 | 95 |
| CBA-DR | Dec | 14.9 | 0 | 0.7 | 0 | 12.6 | 1.6 | **95** | 0 | **100** |

On the other hand, LDA tells us a single discriminative function (fd), which is a polynomial of non-class attribute values with their coefficients. Coefficients in a discriminative function of LDA reflect discriminative power of each non-class attribute (gene, here), with higher positive values and lower negative values meaning larger contributions to each corresponding class label of a class attribute (liver weight, here).

### 2.4.4 Biological Relevancy

To look into how biologically reasonable the CBA-generated classifier is, we conducted the canonical pathway analysis for the set of genes selected in the classifier when all the records were used as a training set for increased liver weight (Table 2.3). For brevity, only top 10 pathways in order of $-\log p$ are shown. $p$ is a value representing statistical significance in the analysis. A smaller $p$ value (thus a larger $-\log p$ value) means that the pathway is more statistically significantly involved. The numbers of the total, increased (upregulated), and decreased (downregulated) molecules in each pathway are also shown. Corresponding Genes indicate the corresponding rat genes for the increased or decreased molecules included in the pathway. Because LDA itself, in contrast to CBA, does not explicitly select a set of genes in building a classifier, we did not compare CBA with LDA here.

We could assume that the most significant pathways involved with the genes in our classifier were mainly drug metabolism-related ones, such as Xenobiotic Metabolism Signaling, LPS/IL-1 Mediated Inhibition of PXR Function, PXR/RXR Activation etc.

Figure 2.2A is an excerpt around the NRF2 molecule from the illustration of the Xenobiotic Metabolism Signaling pathway, exported from IPA. NRF2 is a key modulator of oxidative stress

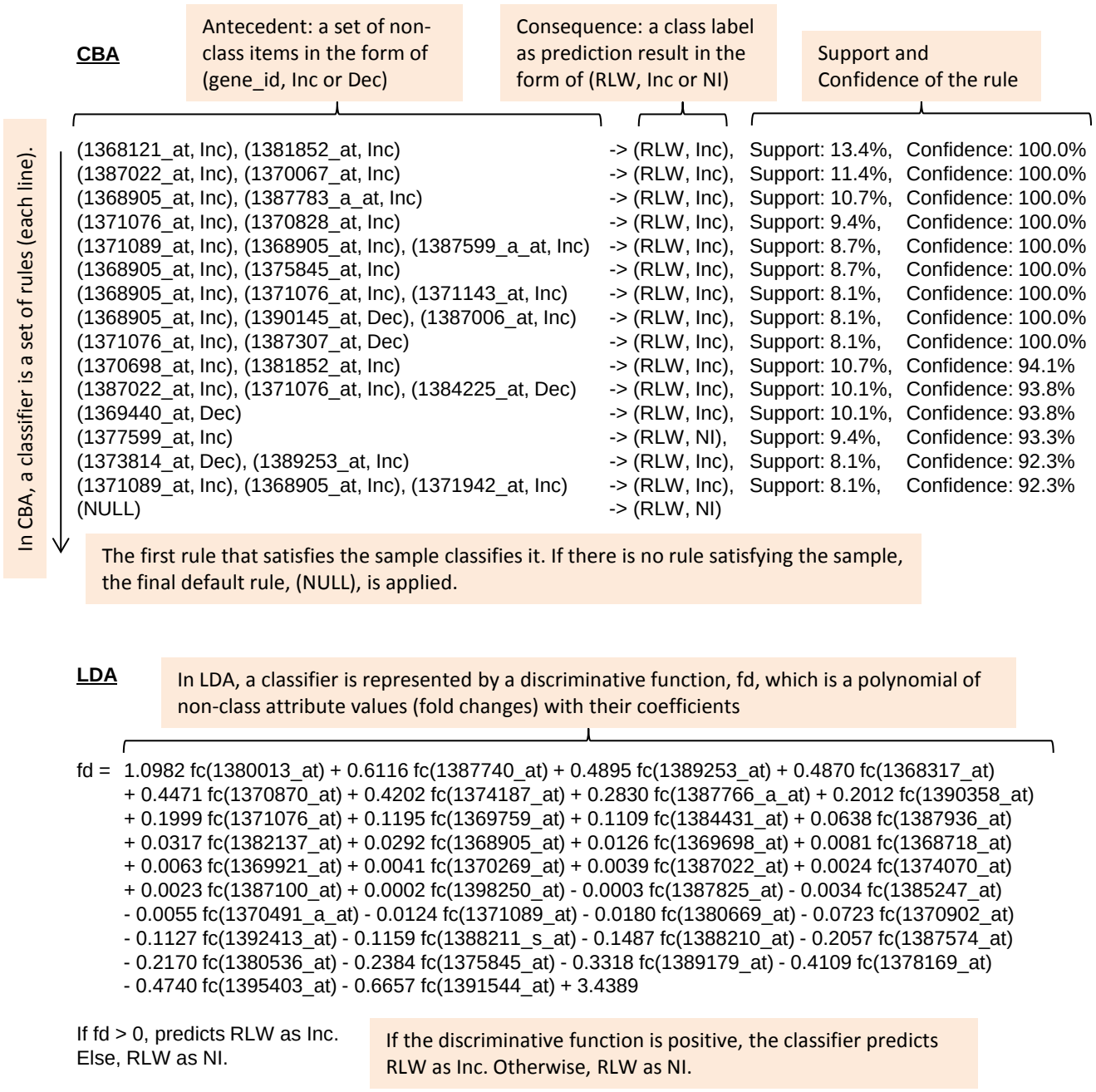


Figure 2.1: Comparison of forms of classifiers between CBA and LDA.

responses. In response of oxidative stress, NRF2 is released into the nucleus and up-regulates downstream antioxidant enzymes, mainly drug metabolism enzymes. Actually, the genes of drug metabolism enzymes such as GST, NQO, and UGT downstream of NRF2 were included in our classifier, suggesting the induction of drug metabolism enzymes triggered by NRF-2-dependent

Table 2.3: Canonical pathway analysis of CBA classifier.

| Pathway Name | $-\log p$ | Molecules | | | Corresponding Genes |
|---|---|---|---|---|---|
| | | Total | Inc | Dec | |
| Xenobiotic Metabolism Signaling | 8.96 | 219 | 8 | 0 | Gsta3, Aldh1a1, Ugt2b1, Nqo1, RGD1559459, Cyp2b2, Ces2c, Sult2a2 |
| LPS/IL-1 Mediated Inhibition of RXR Function | 5.07 | 178 | 4 | 1 | Abccg8, Gsta3 |
| PXR/RXR Activation | 3.95 | 58 | 3 | 0 | Aldh1a1, Cyp2b2, Sult2a2 |
| Aryl Hydrocarbon Receptor Signaling | 2.94 | 127 | 3 | 0 | Gsta3, Aldh1a1, Nqo1 |
| Nicotine Degradation III | 2.77 | 37 | 2 | 0 | Ugt2b1, Cyp2b2 |
| Melatonin Degradation I | 2.75 | 38 | 2 | 0 | Ugt2b1, Cyp2b2 |
| Serotonin Degradation | 2.67 | 42 | 2 | 0 | Aldh1a1, Ugt2b1 |
| Superpathway of Melatonin Degradation | 2.67 | 42 | 2 | 0 | Ugt2b1, Cyp2b2 |
| NRF2-mediated Oxidative Stress Response | 2.66 | 159 | 3 | 0 | Gsta3, Akr7a3, Nqo1 |
| Nicotine Degradation II | 2.65 | 43 | 2 | 0 | Ugt2b1, Cyp2b2 |
| Histidine Degradation III | 2.00 | 6 | 0 | 1 | Hal |

oxidative stress responses.

Figure 2.2B shows overlapping among the canonical pathways detected as significant, where each node corresponds to a pathway and each edge corresponds to the existence of molecules shared between two pathways, with a line width representing the number of molecules. Color depth of nodes corresponds to the $-\log p$ value. The pathways were divided into three clusters. The largest cluster consists of drug metabolism-related pathways as described above. Interestingly, two other clusters, histidine degradation-related and gluconeogenesis-related, were also detected with no overlap between the drug metabolism-related cluster and them.

We then summarized Affymetrix probe IDs, gene symbols and gene names for each gene in our classifier and divided them into four categories, drug metabolism, gluconeogenesis, histidine degradation and the other (Table 2.4), based on the canonical pathway analysis. Direction specifies the direction of change (Inc or Dec) in the classifier. NA means that information was not available. Of 22 genes, 10 genes were drug metabolism-related.

Our classifier was shown again, with genes converted from Affymetrix probe IDs to gene symbols and colored according to their category (Figure 2.3). The CBA classifier, the same as one in Figure 2.1, is shown again, with the genes converted from Affymetrix probe IDs to gene symbols and colored according to their category (Red: drug metabolism-related. Blue: gluconeogenesis-related. Green: histidine degradation-related. Black: Other). The mostly drug metabolism-related

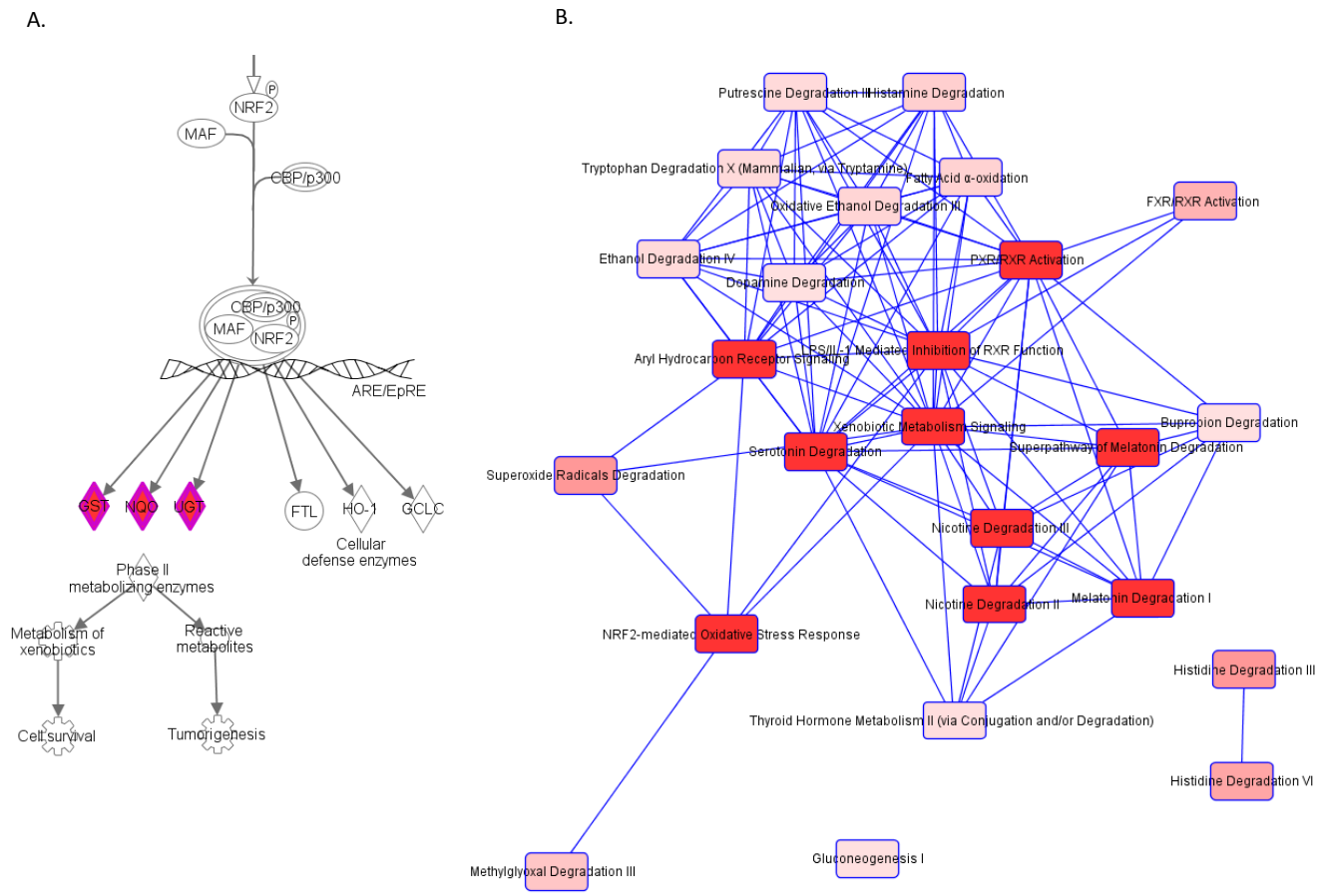


Figure 2.2: Canonical pathway illustrations of CBA classifier.

nature of our classifier was confirmed, as most of the rules in the classifier included drug one or more metabolism-related genes (shown in red).

## 2.5 Discussion

When increased liver weight was targeted, CBA outperformed LDA in all of the three criteria: accuracy, sensitivity, and specificity. In contrast, when decreased liver weight was targeted, both CBA and LDA scored low sensitivities and high specificities. These tendencies are attributable to the low frequency of decreased liver weight in the data set. For such a data set, a classifier returning a negative answer (i.e. 'No' for decreased liver weight) with a high frequency, regardless

Table 2.4: Details and category of the genes in our CBA classifier.

| Affymetrix Probe ID | Gene Symbol | Direction | Gene Name |
|---|---|---|---|
| **Drug metabolism** | | | |
| 1368121_at | Akr7a3 | Inc | Aldo-keto reductase family 7, member A3 |
| 1381852_at | RGD1559459 | Inc | Similar to expressed sequence AI788959 |
| 1387022_at | Aldh1a1 | Inc | Aldehyde dehydrogenase 1 family, member A1 |
| 1368905_at | Ces2C | Inc | Carboxylesterase 2C |
| 1371076_at | Cyp2b2 | Inc | Cytochrome P450, family 2, subfamily b, polypeptide 2 |
| 1371089_at | Gsta3 | Inc | Glutathione S-transferase alpha 3 |
| 1387599_a_at | Nqo1 | Inc | NAD(P)H dehydrogenase, quinone 1 |
| 1370698_at | Ugt2b1 | Inc | UDP glucuronosyltransferase 2 family, polypeptide B1 |
| 1387006_at | Sult2a2 | Inc | Sulfotransferase family 2A, DHEA-preferring, member 2 |
| 1371942_at | Gstt3 | Inc | Glutathione S-transferase, theta 3 |
| **Glucogenesis** | | | |
| 1370067_at | Me1 | Inc | Malic enzyme 1, NADP(+)-dependent, cytosolic |
| **Histidine Degradation** | | | |
| 1387307_at | Hal | Dec | Histidine ammonia-lyase |
| **Other** | | | |
| 1387783_a_at | Acaa1b | Inc | Acetyl-Coenzyme A acyltransferase 1B |
| 1370828_at | Zdhhc2 | Inc | Zinc finger, DHHC-type containing 2 |
| 1375845_at | Aig1 | Inc | Androgen-induced 1 |
| 1371143_at | Serpina7 | Inc | Serpin peptidase inhibitor, clade A, member 7 |
| 1390145_at | Dmxl2 | Dec | Dmx-like 2 |
| 1384225_at | (NA) | Dec | (NA) |
| 1369440_at | Abcg8 | Dec | ATP-binding cassette, subfamily G, member 8 |
| 1377599_at | Lpin1 | Inc | Lipin 1 |
| 1373814_at | R3hdm2 | Dec | R3H domain containing 2 |
| 1389253_at | Vnn1 | Inc | Vanin 1 |

of predictivity, can score a good specificity but a poor sensitivity. Except for such an imbalanced data set, CBA succeeded in building a better predictive classifier than LDA in this study. This superiority of CBA over LDA is considered to reflect the non-linear nature of the data set. Generally, a drug-induced response (or more generally biological response) is considered to be caused not by the single mechanism, but by several different mechanisms. Thus, there are several different, not necessarily linearly separable, gene expression patterns that finally lead to the same response (e.g. increased liver weight). In this light, CBA is likely to build a better classifier for a data set in toxicology, or more broadly biology, than LDA, as CBA can captures linearly inseparable patterns

| | | | |
|---|---|---|---|
| (Akr7a3, Inc), (RGD1559459, Inc) | -> (RLW, Inc), | Support: 13.4%, | Confidence = 100.0% |
| (Aldh1a1, Inc), (Me1, Inc) | -> (RLW, Inc), | Support: 11.4%, | Confidence = 100.0% |
| (Ces2c, Inc), (Acaa1b, Inc) | -> (RLW, Inc), | Support: 10.7%, | Confidence = 100.0% |
| (Cyp2b2, Inc), (Zdhhc2, Inc) | -> (RLW, Inc), | Support: 9.4%, | Confidence = 100.0% |
| (Gsta3, Inc), (Ces2c, Inc), (Nqo1, Inc) | -> (RLW, Inc), | Support: 8.7%, | Confidence = 100.0% |
| (Ces2c, Inc), (Aig1, Inc) | -> (RLW, Inc), | Support: 8.7%, | Confidence = 100.0% |
| (Ces2c, Inc), (Cyp2b2, Inc), (Serpina7, Inc) | -> (RLW, Inc), | Support: 8.1%, | Confidence = 100.0% |
| (Ces2c, Inc), (Dmxl2, Dec), (Sult2a2, Inc) | -> (RLW, Inc), | Support: 8.1%, | Confidence = 100.0% |
| (Cyp2b2, Inc), (Hal, Dec) | -> (RLW, Inc), | Support: 8.1%, | Confidence = 100.0% |
| (Ugt2b1, Inc), (RGD1559459, Inc) | -> (RLW, Inc), | Support: 10.7%, | Confidence = 94.1% |
| (Aldh1a1, Inc), (Cyp2b2, Inc), (1384225_at, Dec) | -> (RLW, Inc), | Support: 10.1%, | Confidence = 93.8% |
| (Abcg8, Dec) | -> (RLW, Inc), | Support: 10.1%, | Confidence = 93.8% |
| (Lpin1, Inc) | -> (RLW, NI), | Support: 9.4%, | Confidence = 93.3% |
| (R3hdm2, Dec), (Vnn1, Inc) | -> (RLW, Inc), | Support: 8.1%, | Confidence = 92.3% |
| (Gsta3, Inc), (Ces2c, Inc), (Gstt3, Inc) | -> (RLW, Inc), | Support: 8.1%, | Confidence = 92.3% |
| (NULL) | -> (RLW, NI) | | |

Figure 2.3: Our CBA classifier shown with categroized gene symbols.

residing in the data set.

We also compared between CBA and CBA-DR, our modified version of the original CBA. When increased liver weight was targeted, CBA-DR marked lower accuracy than CBA. Interestingly however, CBA-DR marked 100% sensitivity. This can be said as follows: if CBA returns an ”Inc” answer for liver weight and we know the default rule is not applied in the classification process, we can say that liver weight would be increased with higher confidence than if we don’t know whether the default rule is applied or not. In addition, we can also infer how reliable the classification is in CBA when non-default rule is met, based on its support and confidence. Therefore, CBA offers not only a classification result, but also additional information regarding reliability of classification. This can be another advantage of CBA over LDA, which returns only a classification result.

In terms of interpretability, while both CBA and LDA give us information regarding important genes which can discriminate increased liver weights well, LDA does not take the concept of co-expression into account. For example, in our setting, a rule (1368905_at, Inc) occurred 6 times in the CBA-generated classifier. This rule, however, always occurred with other rules, reflecting the pattern actually observed in the training data set. Therefore, even if the gene, 1368905_at, is highly increased in an unknown sample, it does not necessarily mean increased liver weight. Such co-expressed pattern was not taken into account by LDA. Besides, while coefficient values are useful to infer importance of each gene in LDA, the final prediction is determined by the total of

all the terms in a polynomial, not by a single or small set of genes. The classification process of CBA is much simpler and easy to understand, because each rule is as simple as a single or small set of genes and the prediction is determined once a rule is satisfied, regardless of the other genes. This characteristic of CBA makes a generated classifier easy to understand, even for a non-expert user, because a CBA-generated classifier can be expressed also in a natural language (e.g. ”If gene A is increased and gene B is decreased, then the classifier predicts liver weight to be increase”), not in a mathematical equation as is case in LDA.

Canonical pathway analysis with IPA revealed that the genes included in our CBA-generated classifier for increased liver weight were mostly drug metabolism-related ones. This is reasonable as inductions of hepatic drug metabolizing enzymes are well known to induce hepatocellular hypertrophy (Ennulat, Walker, Clemo, Magid-Slav, Ledieu, Graham, Botts, and Boone, 2010), of which increases in liver weight is the most sensitive indicator (Hall, Elcombe, Foster, Harada, Kaufmann, Knippel, Kuttler, Malarkey, Maronpot, Nishikawa, Nolte, Schulte, Strauss, and York, 2012). CBA succeeded in building a biologically relevant classifier without any prior knowledge such as literature. Intriguingly, the classifier included genes with other functions such as gluconeogenesis and histidine degradation, which are not directly related to increased liver weight or hepatocellular hypertrophy. While it is unclear whether these genes were actually causal or not, CBA can be used to look for genes with an unknown function but high correlation for a specified outcome as well as to build a biologically reasonable classifiers. In addition, it was also considered to be an advantage that CBA automatically selects a small set of genes to build a classifier, while LDA does not.

## 2.6 Conclusion

We applied the CBA algorithm to the TG-GATEs database to build a predictive classifier of increased or decreased liver weight for an unknown compound. We compared the generated classifiers between CBA and LDA, and showed that CBA is superior to LDA in terms of both predictive performances and interpretability.

# Chapter 3

# Group-based Sparse Regularization

## 3.1 Introduction

In this chapter, we introduce the latent group Lasso (LGL) method and compare its predictive performances and biological relevancy with Lasso in our toxicogenomic modeling. LGL is a variant of group-based structured regularization techniques and uses a specialized norm as a penalty term in regression to select explanatory variables as unions of groups (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011).

To apply LGL on our modeling, we regarded genes regulated by the same transcription factor as a group based on the MSigDB database, since such genes are expected to be co-expressed and therefore should be incorporated to or discarded from a model together. While several papers reported applications of structured regularization techniques in biological fields (Ma, Song, and Huang, 2007; Obozinski, Jacob, and Vert, 2011; Silver, Montana, and Alzheimer's Disease Neuroimaging, 2012), there had been no direct implication of LGL in toxicogenomic analyses until our study (Nagata, Kawahara, Washio, and Unami, 2015).

## 3.2 Latent Group Lasso

In regularized regression analysis, a regression parameter $\mathbf{w} \in \mathbb{R}^p$ is estimated from a given $n$-sample dataset of explanatory variables $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]^T$ and their corresponding response variables $\mathbf{Y} = [y_1, \cdots, y_n]^T$ by solving an optimization problem:

$$\min_{\mathbf{w}} \mathrm{L}_{\mathbf{w}}(\mathbf{Y}, \mathbf{X}) + \lambda \mathrm{P}(\mathbf{w}), \tag{3.1}$$

where $\mathrm{L}_{\mathbf{w}}$ is a loss function with a parameter $\mathbf{w}$, P is a regularization (or penalty) term, and $\lambda > 0$ is an arbitrary penalty parameter.

Among the most standard choices for a regularization term P is the L1-norm, known as the Lasso regularization, which tends to estimate non-important components of w as zero (Tibshirani,

1996). While an estimation with the Lasso regularization is known to generally leads to an accurate and sparse model (Tibshirani, 1996), Lasso ignores structural relationships of explanatory variables (e.g. biological pathways).

To take such structural relationships into account as a prior knowledge in estimation of a regression model, various kinds of regularization terms, collectively known as structured regularization terms, have been proposed and used. Latent group Lasso (LGL) (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011) is one of those structured regularization methods and uses a group structure as a prior structural information and employs the following norm as a regularization term:

$$\mathrm{P}_{LGL}(\mathbf{w}) = \min_{\mathbf{v}^g \in \mathbb{R}^p \mid g \in \mathcal{G},\, \mathbf{v}_i^g = 0 \text{ if } i \notin g} \sum_{g \in \mathcal{G}} d_g ||\mathbf{v}^g||_2,\ s.t.\ \sum_{g \in \mathcal{G}} \mathbf{v}^g = \mathbf{w}, \tag{3.2}$$

where $\mathcal{G} \subset \mathcal{P}([1, p])$ is a set of groups, $g \in \mathcal{G}$ is a group that is a subset of indexes of parameter $\mathbf{w}$, and $d_g > 0$ is a weight for a group $g$. This norm tends to select explanatory variables as unions of groups.

## 3.3 Analytical Methods

### 3.3.1 Software

For LGL and Lasso analyses, we used the MATLAB® (The MathWorks, Inc.) code based on the algorithm of (Meier, Geer, and Bühlmann, 2008; Jacob, Obozinski, and Vert, 2009) available on Dr. Jacob's Homepage (http://cbio.ensmp.fr/ ljacob/).

### 3.3.2 Data Process

We used the 615 transcription factor target gene sets as groups from the motif gene sets (C3) of the MSigDB database. Of the 615 groups, we omitted those regulated by an unknown transcription factor.

We basically followed the procedure of the pathway analysis experiment for breast cancer data reported in (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011). Since the data set is unbalanced, we used the balanced logistic loss function, which weighs each positive case with the proportion of negative cases and each negative case with the proportion of positive cases respectively, and the balanced accuracy (Obozinski, Jacob, and Vert, 2011). We did not weight

groups ($d_g = 1$ for $g \in G$) as the unweighted version outperformed the weighted version in (Obozinski, Jacob, and Vert, 2011).

For discussions of group sparsity and biological relevance, we used all the samples as the training set and conducted no external CV, whereas internal CV for selecting  were still conducted. Note that the gene-filtering process was also conducted before the analysis, instead of the beginnings of each external CV step.

The experiment of (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011) omitted all the genes that were not included in any groups in prior to the analyses, to ensure fair comparisons between LGL and Lasso, as LGL requires that all the genes belong to at least one group. This process, however, downplays the difference between the two methods, since the group information is used even for the Lasso analysis. To better compare the difference between the two methods, instead of omitting genes that were not included in any groups or unable to be mapped properly, we fabricated 'dummy' groups, each of which includes genes that were not included in any groups (one gene per group), in prior to the LGL analysis. This approach allows us to rightly demonstrate the effects of prior knowledge, as the Lasso analysis does not use group information, while still keeping the comparison between the two methods fair because the same set of genes are used for both methods.

### 3.3.3 Comparison of Predictive Performances

According to (Jacob, Obozinski, and Vert, 2009; Obozinski, Jacob, and Vert, 2011), to estimate the generalized predictive performances, we conducted a 5-fold cross validation on the data set (hereafter external CV). First, in each step of the external CV, we filtered 10,000 genes based on correlations with the discretized liver weights. This type of filtering is common practice in microarray data analysis (Obozinski, Jacob, and Vert, 2011). Besides, we confirmed that the results were robust to numbers of filtered genes. Secondly, internal 5-fold cross validations (hereafter internal CV), further splitting the training set, for each $\lambda \in [2^x | x = 0, -0.5, -1, \cdots, -12]$ were conducted to select the best $\lambda$ based on the average balanced accuracy. Thirdly, the model was built on the training set with selected  and evaluated for its predictive performances (the balanced accuracy, sensitivity, and specificity) on the test set. Finally, predictive performances were averaged over each external CV step.

Following Jacob's implementation, we adopted the randomized balanced cross validation approach, which randomly distributes cases into partitions so that each partition has the proportions of positive and negative cases as close as possible to those in the whole data set. To ensure repro-

ductivity, we fixed the random seed in MATLAB to 0 at the beginning of the process.

Parameters of predictive performances are defined as follows (Carrillo, Brodersen, and Castellanos, 2014; Florkowski, 2008).

**Sensitivity:** True Positive / (True Positive + False Negative)

**Specificity:** True Negative / (True Negative + False Positive)

**Balanced accuracy:** (Sensitivity + Specificity) / 2

### 3.3.4 Sparsity

For comparison of sparsity between methods, we evaluated two different metrics for generated classifiers: the number of selected genes and the number of selected groups.

We regarded covariates (genes) with non-zero coefficients as selected covariates (genes) and used it for discussion of covariate-level sparsity of generated classifiers.

We defined a new metric, coverage of a group $g$ under the model parameter $\mathbf{w}$, as following, and used it for discussion of group-level sparsity of generated classifiers:

$$coverage(g) \equiv \frac{|support(\mathbf{w}^g)|}{|g|}, \tag{3.3}$$

where $support(\mathbf{w}^g)$ is the set of covariates $i \in [1, p]$ such that $w_i \neq 0$ and $i \in g$.

In our study, we distinguished pre-filtering coverage (denoted $coverage\_pre(g)$) and post-filtering coverage ($coverage\_post(g)$), as genes were filtered in each external CV step and thus the group sizes $|g|$ were different before and after filtering. We regarded $coverage\_post(g) = 1$ as the criteria that the group $g$ was considered to be selected in a generated classifier. This means that all the filtered covariates (genes) of the group have non-zero coefficients. Dummy groups were excluded from this discussion.

### 3.3.5 Pathway Analysis

We conducted canonical pathway and tox function analyses with IPA software using all the selected genes and the transcription factors upstream of the selected groups in our classifier as the user-specified set. During the analysis, used genes are mapped to their corresponding molecules and matched up against the molecules in pre-defined sets of molecules.

### 3.3.6 Computer

We used a personal computer with Intel Xeon E5620 CPU (2.40 and 2.39 GHz processors) and 48 GB RAM for the analyses in this chapter.

## 3.4 Results

### 3.4.1 Predictive Performance

We compared predictive performance of generated classifiers between LGL and Lasso with 5-fold cross validation (Table 3.1). Values are shown as mean $\pm$ standard deviation (%). BAcc, Sen, and Spe are balanced accuracy, sensitivity, and specificity, respectively. Balanced accuracy, sensitivity, and specificity were $74 \pm 8$ %, $62 \pm 12$ %, and $86 \pm 8$ % (mean $\pm$ standard deviation) for LGL. On the other hand, the Lasso scored $72 \pm 4$ %, $62 \pm 4$ %, and $83 \pm 9$ %, respectively. Our comparison under the same condition showed that the predictive performance of LGL was comparable to that of Lasso.

Table 3.1: Comparison of predictive performance.

| Method | BAcc (%) | Sen (%) | Spe (%) |
|---|---|---|---|
| LGL | $74 \pm 8$ | $62 \pm 12$ | $86 \pm 8$ |
| Lasso | $72 \pm 4$ | $62 \pm 4$ | $83 \pm 9$ |

### 3.4.2 Group Sparsity

We compared covariate-level and group-level sparsity of generated classifiers between LGL and Lasso (Table 3.2). The result showed that the number of selected genes in the LGL classifier (3,214) was far larger than in the Lasso classifier (86). In addition, the number of selected groups based on the post-filtering coverage criteria for LGL is 35, while that for the Lasso is 0 (i.e. no group was selected).

### 3.4.3 Biological Relevancy

We further looked into the groups selected by the LGL classifier (Table 3.3). The 35 groups adopted by LGL are listed in order of pre-filtering coverage. Groups are represented by the

Table 3.2: Comparison of sparsity.

| Method | Number of selected genes | Number of selected groups |
|---|---|---|
| LGL | 3,214 | **35** |
| Lasso | **86** | 0 |

names of their upstream transcription factors (TFs). Groups with a TF involved with the Wnt/$\beta$-Catenin/TCF signaling pathway are labelled with '+'. Note that there are groups with the same TF (e.g. TCF3), corresponding to different binding motifs. Groups regulated by two or more TFs are shown with '/'.

We found that all of the 5 groups regulated by TCF3 (TCF7L1) only, except for those regulated by two or more transcription factors including TCF3 (e.g. TAL1 / TCF3), were selected. In addition, 2 of the 5 groups regulated by TCF8 (ZEB1) were selected. These transcription factors are members of TCF/LEF family and known to play a key role in the Wnt/$\beta$-Catenin/TCF signaling pathway (Kolligs, Bommer, and Goke, 2002). Other transcription factors are also reported to be involved in the Wnt/$\beta$-Catenin/TCF signaling pathway. LEF1 binds to $\beta$-Catenin in response to Wnt pathway activation and mediates the expression of downstream genes (Valdivia, Young, Hawkins, Stickney, Cavodeassi, Schwarz, Pullin, Villegas, Moro, Argenton, Allende, and Wilson, 2011). EVI1 expression is increased by $\beta$-Catenin activation (Wei, Zeve, Suh, Wang, Du, Zerwekh, Dechow, Graff, and Wan, 2011). PAX3 is a downstream target of the Wnt/$\beta$-Catenin signaling (Zhao, Gan, Stokes, Lassiter, Wang, Chan, Han, Pleasure, Epstein, and Zhou, 2014). MYOD1 induction from the alveolar RMS cells is significantly decreased in response to human recombinant Wnt3a (Annavarapu, Cialfi, Dominici, Kokai, Uccini, Ceccarelli, McDowell, and Helliwell, 2013). Regulatory sites for MAZ have been reported upstream of MT1-MMP, which is up-regulated in colon carcinomas mediated by a direct interaction of $\beta$-Catenin/TCF4 complex (Hodar, Assar, Colombres, Aravena, Pavez, Gonzalez, Martinez, Inestrosa, and Maass, 2010). ALX4 can bind LEF1 and may modulate the $\beta$-Catenin/LEF1 signaling pathway (Church and Francis-West, 2002).

Further, we conducted pathway analyses for the selected genes with the transcription factors of the selected groups in the LGL classifier, using QIAGEN's Ingenuity Pathway Analysis (IPA) software (Table 3.4). Top 10 liver-related annotations are listed in ascending order of p values. Molecules is the number of molecules of each annotation that are included in our set. Tox Function analysis revealed that the two annotations under the "Liver Hyperplasia / Hyperproliferation" category ranked first and second, with 1,009 ($p < 10^{-12}$) and 996 ($p < 10^{-11}$) molecules, far distant

Table 3.3: List of selected groups by the LGL classifier.

| Group | Pre-filtering coverage | Post-filtering coverage | Wnt/$\beta$-Catenin/TCF |
|---|---|---|---|
| HOXA3 | 0.47 | 1.00 | |
| ALX4 | 0.43 | 1.00 | + |
| SF1 | 0.41 | 1.00 | |
| CUTL1 | 0.40 | 1.00 | |
| ESRRA | 0.38 | 1.00 | |
| PAX5 | 0.38 | 1.00 | |
| MAZ | 0.37 | 1.00 | + |
| NFAT | 0.37 | 1.00 | |
| SP1 | 0.37 | 1.00 | |
| MLLT7 | 0.35 | 1.00 | |
| PAX4 | 0.35 | 1.00 | |
| TCF8 | 0.35 | 1.00 | + |
| TAF | 0.35 | 1.00 | |
| LEF1 | 0.34 | 1.00 | + |
| TCF3 | 0.34 | 1.00 | + |
| MAZ | 0.34 | 1.00 | + |
| TCF3 | 0.33 | 1.00 | + |
| NR1H4 | 0.33 | 1.00 | |
| TFAP4 | 0.33 | 1.00 | |
| SREBF1 | 0.33 | 1.00 | |
| MYOD1 | 0.33 | 1.00 | + |
| REPIN1 | 0.33 | 1.00 | |
| TCF3 | 0.32 | 1.00 | + |
| MEIS1 | 0.32 | 1.00 | |
| TCF3 | 0.31 | 1.00 | + |
| TCF8 | 0.31 | 1.00 | + |
| TCF3 | 0.31 | 1.00 | + |
| MEIS1 / HOXA9 | 0.31 | 1.00 | |
| TFAP4 | 0.30 | 1.00 | |
| MYOD1 | 0.29 | 1.00 | + |
| UBP1 | 0.29 | 1.00 | |
| EVI1 | 0.25 | 1.00 | + |
| EVI1 | 0.22 | 1.00 | + |
| PAX3 | 0.20 | 1.00 | + |
| EVI1 | 0.19 | 1.00 | + |

from the third place with 56 molecules ($p < 10^{-6}$). In addition, the third and fourth places are under the "Liver Proliferation" category, also suggesting proliferative activity in the liver. Canonical Pathway analysis confirmed that the Wnt/$\beta$-Catenin/TCF signaling pathway ("Wnt/beta-catenin Signaling") is significantly involved ($p < 10^{-5}$) (Table 3.1). The molecules included in our set are outlined in purple.

Table 3.4: Pathway analysis with IPA tox function analysis.

| Category | Annotation | p value | Molecules |
|---|---|---|---|
| Liver Hyperplasia / Hyperproliferation | Liver Tumor | $3.84 \times 10^{-13}$ | 1,009 |
| Liver Hyperplasia / Hyperproliferation | Liver Cancer | $2.97 \times 10^{-12}$ | 996 |
| Liver Proliferation | Proliferation of Liver Cells | $5.49 \times 10^{-7}$ | 56 |
| Liver Proliferation | Proliferation of Hepatocytes | $1.37 \times 10^{-6}$ | 46 |
| Liver Cirrhosis | Cirrhosis | $1.55 \times 10^{-6}$ | 49 |
| Liver Damage | Damage of Liver | $1.29 \times 10^{-5}$ | 66 |
| Liver Necrosis / Cell Death | Cell Death of Liver | $1.82 \times 10^{-5}$ | 63 |
| Liver Fibrosis | Fibrosis of Liver | $2.83 \times 10^{-5}$ | 35 |
| Liver Hypoplasia | Hypoplasia of Liver | $4.29 \times 10^{-5}$ | 17 |
| Liver Necrosis / Cell Death | Necrosis of Liver | $5.35 \times 10^{-5}$ | 61 |

## 3.5 Discussion

While the LGL classifier was comparable to the Lasso classifier in terms of predictive performance, the generated classifiers were quite different in the numbers of adopted genes and groups. While the Lasso classifier selected only 86 genes out of the total of 31,099 genes, the LGL classifier selected 3,214 genes. Besides, while the Lasso selected no group, the LGL selected 35 groups out of the total of 439 groups. Thus, in terms of sparsity, the Lasso built a far much more compact model than the LGL.

Although compactness is itself a virtue of Lasso, it is often difficult for researchers to infer the mechanism underlying toxicological responses based on the generated classifier, since no relationships among genes are taken into account in modeling. This was clearly demonstrated by the fact that the Lasso classifier selected no group in our study. In stark contrast, the LGL classifier gave us information in the form of transcription factor networks involved with toxicological responses of interest. In this study, we found that transcription factors with known ties to the Wnt/$\beta$-Catenin/TCF signaling pathway had been preferably selected in the model, thus suggesting that the pathway plays some roles in liver weight gain in rats and could be useful transcriptomic markers to predict it.

Actually, there have been several reports that collectively suggest that the Wnt/$\beta$-Catenin/TCF signaling pathway is related to liver weight gain in rodents. (Tao, Lehwald, Jang, Baek, Xu,

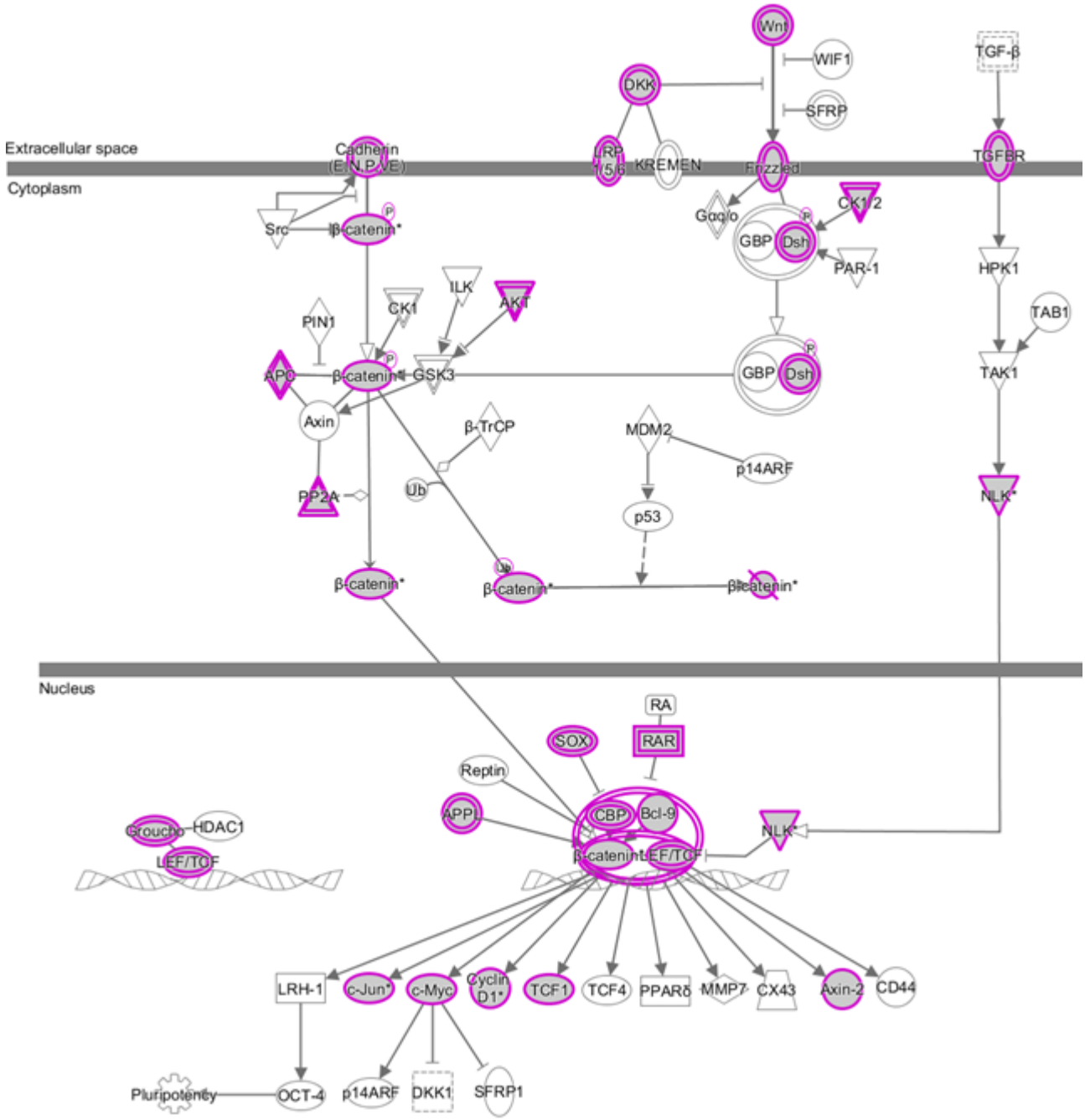


Figure 3.1: Pathway Analysis with IPA canonical pathway analysis.

Omary, and Sylvester, 2013) reported that $\beta$-Catenin knockdown mice showed severer liver injury than wild-type counterparts after feeding with 3,5-diethoxycarbonyl-1,4-dihydrocollidine (DDC), an oxidative stress inducer. They concluded that the Wnt/$\beta$-Catenin signaling is required for hepatocyte protection against oxidative stress-induced apoptosis via inhibition of FoxO3. Oxidative

stress inducers are well known to induce liver weight gain, and antioxidants counteract it (Das and Vasudevan, 2005; Lankoff, Banasik, and Nowak, 2002). This is thought of as the result of an adaptive response mediated by NFE2L2 (Nrf2), a key player in the antioxidant defense system, in which antioxidant enzymes (mainly phase 2 detoxifying enzymes) are induced to combat oxidative stress (Xu, Hellerbrand, Kohler, Bugnon, Kan, Werner, and Beyer, 2008), therefore leading to liver weight gain, hepatocellular hypertrophy, cell proliferation, and/or hepatocarcinogenesis (Hall, Elcombe, Foster, Harada, Kaufmann, Knippel, Kuttler, Malarkey, Maronpot, Nishikawa, Nolte, Schulte, Strauss, and York, 2012). While no direct link have so far been reported between the Wnt/$\beta$-Catenin/TCF signaling pathway and liver weight gain, there are also reports that the Wnt signaling pathway interacts NFE2L2 to regulate antioxidant metabolism in hepatocytes (Rada, Rojo, Offergeld, Feng, Velasco-Martin, Gonzalez-Sancho, Valverde, Dale, Regadera, and Cuadrado, 2015). Taken together, the selection of groups in the latent group Lasso classifier deems reasonable from the biological point of view. The pathway analyses with QIAGEN's IPA also suggested that the adopted genes and the transcription factors of the adopted groups are significantly related to the hepatocellular proliferative activity and Wnt/$\beta$-Catenin/TCF signaling pathway, thus supporting our hypothesis.

Therefore, our LGL classifier is not only accurate but also informative, as we can discuss underlying mechanisms behind liver weight gain based on transcription factor networks. This is clearly an advantage of LGL over Lasso, although both methods generated classifiers with comparable predictive performances.

In our study, we applied the LGL technique on toxicogenomic data with transcription factor networks as prior knowledge to predict liver weight gain in rats. This approach is, however, not limited to this setting. For example, we can apply the same analysis with another type of genetic network such as biological pathways reported in literature (Chuang, Lee, Liu, Lee, and Ideker, 2007; Obozinski, Jacob, and Vert, 2011; Silver, Montana, and Alzheimer's Disease Neuroimaging, 2012) and user-defined gene clusters based on statistical metric (Ma, Song, and Huang, 2007). (Obozinski, Jacob, and Vert, 2011) used the canonical pathways from MSigDB for the LGL analysis with microarray data. Given that enrichment analysis with such canonical pathways is very common in microarray data analysis, this approach might seem straightforward. But even when there are pathways that are actually activated or inactivated, it does not necessarily mean that the majority of the genes in those pathways are up- or down-regulated together. In fact, such a case is rare, because those pathways consist of various kinds of interactions among genes, proteins, and the other molecules including ones that have no direct effects on gene expressions (e.g. activation via phosphorylation). With this in mind, our choice of transcription factor network, where genes

whose expressions are regulated by the same transcription factor form a group together, seems reasonable in that genes under the same transcription factor are expected to be co-regulated.

It is also possible to choose another penalty term such as group Lasso (Bach, Jenatton, Mairal, and Obozinski, 2012). The reason why we used LGL, not group Lasso, lies in that group Lasso selects intersections of complements of groups, while LGL selects unions of groups as non-zero covariates (Bach, Jenatton, Mairal, and Obozinski, 2012). That is to say, group Lasso eliminates a covariate if one of the groups including it is discarded, while LGL eliminates a covariate if all of the groups including it are discarded. This property renders LGL especially suitable for an analytical setting where groups greatly overlap with each other and variables belonging to several groups should be selected if at least one of those groups is selected. This is the case of transcription factor network, where a gene is usually regulated by several transcription factors, and therefore we utilized LGL, instead of the group Lasso.

Back to the compactness issue again, although the LGL classifier gives us useful information that the Lasso cannot, it is also true that the Lasso classifier has an advantage in that it involved much smaller set of genes to predict the outcome. This advantage is especially evident when we use the model to screen many compounds in pharmaceutical industry, because we can run a screening test using custom microarrays loaded with that small set of genes of interest, thus making the screening process much effective in terms of both cost and time. Therefore, in such a circumstance, we propose first building a classifier with LGL, and then eliminating genes with no or little contribution to prediction based on techniques such as the forward or backward selection methods. This procedure would create a model satisfying compactness and informative.

## 3.6 Conclusion

We applied the LGL method to the TG-GATEs and MSig databases to build a predictive classifier of liver weight gain for an unknown compound. We compared the generated classifiers between LGL and Lasso, and showed that LGL is comparable to LDA in terms of predictive performances. Besides, we demonstrated that LGL enabled us to discuss the mechanism behind liver weight gain based on the selected groups.

# Chapter 4

# Graph-based Sparse Regularization

## 4.1 Introduction

As seen in the previous chapter, while latent group Lasso (LGL) can generate an accurate and biological relevant model, it does not necessarily lead to a sparse model, especially if group sizes used as a structural information are large.

In this chapter, we introduce the generalized fused Lasso (GFL) and graph Lasso (GL) methods and compare their predictive performances and biological relevancy with Lasso and LGL in our toxicogenomic modeling. GFL and GL are variants of graph-based structured regularization techniques which select explanatory variables as a set of edges on a graph. In contrast to group-based techniques such as LGL, graph-based techniques are expected to make a sparser estimation since they select variables based on edges (consisting of 2 explanatory variables), not on groups.

We apply GFL and GL under the same analytical setting as the previous chapter, except that we treat the regulatory networks formed by transcription factors and their downstream genes as a graph in this chapter, instead of groups. Our study, (Nagata, Kawahara, Washio, and Unami, 2016), report the first application of graph-based structured regularization in toxicogenomics.

## 4.2 Generalized Fused Lasso

GFL takes two regularization terms and can be represented as follows:

$$\min_{\mathbf{w}} \mathrm{L}_{\mathbf{w}}(\mathbf{Y}, \mathbf{X}) + \lambda_1 \sum_{i=1}^{p} |\mathbf{w}_i| + \lambda_2 \sum_{(i,j)\in E} |\mathbf{w}_i - \mathbf{w}_j|, \tag{4.1}$$

where $G = (V, E)$ is a graph with vertices $V$ and edges $E$ that represents structural relationship among explanatory variables $\mathbf{w}_i$, such that each vertex $i \in V = \{1, \cdots, p\}$ corresponds to $\mathbf{w}_i$ and each edge $(i, j) \in E \subset \{1, \cdots, p\} \times \{1, \cdots, p\}$ corresponds to existence of relationship between $\mathbf{w}_i$ and $\mathbf{w}_j$ (Tibshirani and Taylor, 2011). The first regularization term is L1-norm and the second

is called the fused term, which penalizes pairwise differences between coefficients of explanatory variables connected by edges (Xin, Kawahara, Wang, and Gao, 2014).

## 4.3 Graph Lasso

GL is an extension of LGL for a graph $G = (V, E)$, where $\mathcal{G} = E$ (i.e. each edge of the graph is regarded as a group consisting of two explanatory variables that it connects). By doing so, selecting a group of variables in LGL can be equivalently viewed as selecting a edge, or more precisely two variables connected by that edge, in an original graph.

## 4.4 Analytical Methods

### 4.4.1 Rearrangement of Groups into Graph

For GFL and GL, we rearranged the groups that we used in Chapter 3 into a graph. Conversion process is depicted in Figure 4.1. Each group has a transcription factor (itself not a group member) and a set of genes regulated by it. Transcription factors and regulated genes are both represented in common identifying numbers (human Entrez IDs in actual analyses). For each group, edges from a vertex corresponding to the transcription factor to every vertex corresponding to the regulated genes are drawn. For example, two edges (1→2 and 1→5) are to be made for the first group in Figure 4.1. Repeating this process for all the three groups makes an entire graph as shown here. While the converted graph is directed, GFL and GL do not distinguish directions of edges.

| Transcription Factor | Regulated Genes |
|---|---|
| 1 | 2, 5 |
| 2 | 4 |
| 3 | 1, 2 |

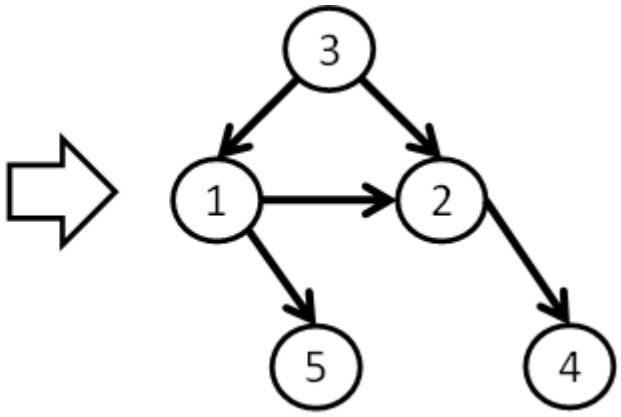


Figure 4.1: Schematic of Rearrangement Process of Groups into a Graph.

### 4.4.2 Software

For Lasso, LGL and GL analyses, we used the MATLAB® (The MathWorks, Inc.) code based on the algorithm of (Meier, Geer, and Bühlmann, 2008; Jacob, Obozinski, and Vert, 2009) available on Dr. Jacob's Homepage (http://cbio.ensmp.fr/ ljacob/).

For GFL analysis, we used the MATLAB® code based on the algorithm of (Xin, Kawahara, Wang, and Gao, 2014) available on Dr. Wang's Homepage (http://idm.pku.edu.cn/staff/wangyizhou/).

### 4.4.3 Data Process

The procedure is the same as Chapter 3, except for the following points.

To make comparisons fair, we prepared dummy groups, each of which includes genes that were not included in any groups (one gene per group), in prior to LGL and GL analyses.

### 4.4.4 Comparison of Predictive Performances

The procedure is the same as Chapter 3, except for the following points.

Internal 5-fold cross validations (internal CV) were conducted for each $\lambda \in [2^x | x = 0, -0.5, -1, \cdots, -12]$ for Lasso, LGL, and GL, or $(\lambda_1, \lambda_2) \in [2^x | x = 0, -2, -4, \cdots, -12] \times [2^x | x = 0, -2, -4, \cdots, -12]$ for GFL.

### 4.4.5 Sparsity

For comparison of sparsity among methods, we evaluated three different metrics for generated classifiers: the number of selected genes, the number of selected groups, and the number of selected edges.

The number of selected genes is the count of covariates in an estimated parameter whose coefficients exceeded a threshold. For each estimated parameter, we set a threshold as $\pm$ 1/1,000 of the maximum absolute value of coefficients. We set a different criteria from Chapter 3, where the threshold was set as zero, since the previous threshold was too strict when applied for the GFL code used in this chapter.

The number of selected groups is the count of groups whose post-filtering coverage was 1 (i.e. all the filtered genes of the group were selected). The coverage is defined in Chapter 3.

The number of selected edges is the count of edges in the graph whose connected vertices (corresponding to genes) were both selected. We calculated the number of selected groups and edges based on the same groups and the graph converted from it, regardless of employed regularization

methods. Note that edge groups temporarily prepared for GL were not used here as groups. Also note that dummy groups in LGL and GL were not included in this calculation.

### 4.4.6 Pathway Analysis

We conducted canonical pathway analysis with IPA software using the genes selected in our classifiers as the user-specified sets. During the analysis, used genes are mapped to their corresponding molecules and matched up against the molecules in pre-defined sets of molecules.

### 4.4.7 Computer

We used a personal computer with Intel Xeon E5620 CPU (2.40 and 2.39 GHz processors) and 48 GB RAM for the analyses in this chapter.

## 4.5 Results

### 4.5.1 Predictive Performance

We compared predictive performance of generated classifiers of Lasso, LGL, GFL, and GL in 5-fold cross validations (Table 4.1). Values are shown as mean $\pm$ standard deviation (%). BAcc, Sen, and Spe are balanced accuracy, sensitivity, and specificity, respectively. All the four methods achieved almost equivalent performances in terms of balanced accuracy, while GFL scored lower sensitivity and higher specificity compared to the other methods.

Table 4.1: Comparison of predictive performance.

| Method | BAcc (%) | Sen (%) | Spe (%) |
|---|---|---|---|
| Lasso | $73 \pm 4$ | $62 \pm 4$ | $83 \pm 9$ |
| LGL | $74 \pm 8$ | $62 \pm 12$ | $86 \pm 8$ |
| GFL | $75 \pm 7$ | $56 \pm 15$ | $95 \pm 3$ |
| GL | $75 \pm 4$ | $67 \pm 14$ | $83 \pm 9$ |

### 4.5.2 Group Sparsity

We compared gene-level, group-level, and edge-level sparsity of generated classifiers for Lasso, LGL, GFL, and GL (Table 4.2). As in Chapter 3, LGL led to a much larger number of selected genes (2,924) than Lasso (83). LGL selected 7 groups and 6,368 edges, while Lasso selected no group or edges.

GFL selected 8,380 genes, even larger than LGL. The numbers of selected groups and edges by GFL were 3 and 6,996, both comparable to LGL.

GL selected 108 genes, much fewer than LGL and GFL and comparable to Lasso. GL selected no group. The number of selected edges by GL was 73, in stark contrast to the other three methods.

Table 4.2: Comparison of sparsity.

| Method | Number of selected genes | Number of selected groups | Number of selected edges |
|---|---|---|---|
| Lasso | **83** | 0 | 0 |
| LGL | 2,924 | 7 | 6,368 |
| GFL | 8,380 | 3 | 6,996 |
| GL | **108** | 0 | **73** |

### 4.5.3 Biological Relevance

We further investigated 73 edges selected by GL. We extracted a subgraph that contains only the selected edges (Figure 4.2). Each point represent a vertex (corresponding to a gene) labeled with its gene symbol. Each line represents an edge with an arrow showing a direction from a transcription factor to its downstream gene. Then, we summarized the numbers of outbound and inbound edges for the selected genes included in the selected edges (Table 4.3). The total of 32 selected genes included in selected edges by GL are listed in order of the number of outbound edges.

Assuming that the genes with many outbound or inbound edges play key roles in liver weight gain in rat, we focused on top 5 genes (FOXO4, TAF9, TAF12, POU2F1, and HNF4A) with regard to the number of outbound edges, and top 7 genes (FGF12, POU2F1, MAF, HNRNPA0, RBP2, S100G, and CDKL5) with regard to the number of inbound edges. We selected genes up to 5th rank for each category. Note that the numbers are different between the categories since RBP2, S100G, and CDKL5 are tied 5th rank for the number of inbound edges, and that POU2F1 appears in both categories.

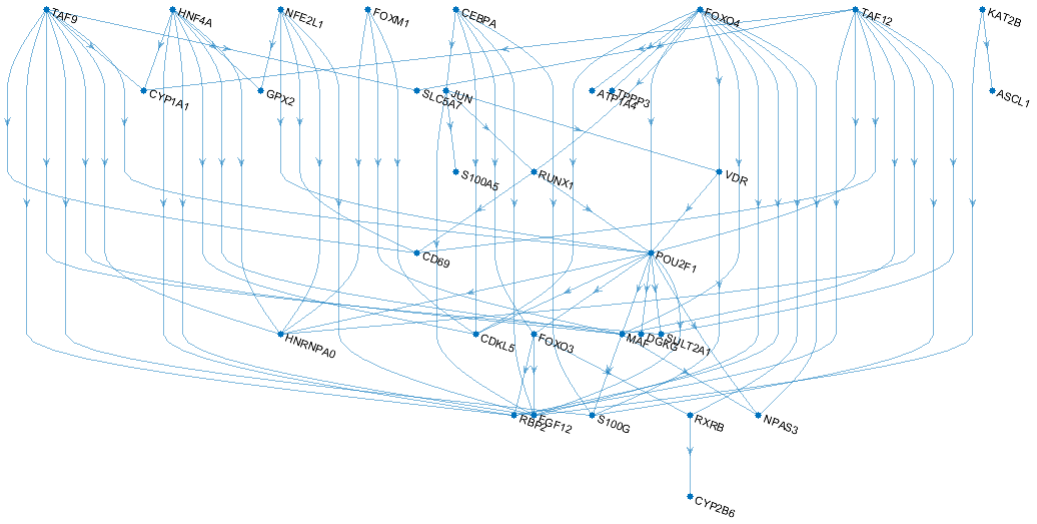


Figure 4.2: Subgraph of Selected Edges by GL.

Interestingly, we found that many of these genes are reportedly linked to oxidative stress. FOXO4 was activated by oxidative stress generated by H2O2, through nuclear translocation and transcriptional activation of FOXO4, in cultured cells (Essers, Weijzen, Vries-Smits, Saarloos, Ruiter, Bos, and Burgering, 2004). POU2F1, also known as OCT1, was dynamically phosphorylated following exposure of cells to oxidative stress, and was essential for a normal post-stress transcription response (Kang, Gemberling, Nakamura, Whitby, Handa, Fairbrother, and Tantin, 2009). Inactivation of HNF4A in cells resulted in an increase of oxidative stress, thus suggesting that HNF4A plays a key role in anti-oxidative defense mechanisms (Marcil, Seidman, Sinnett, Boudreau, Gendron, Beaulieu, Menard, Precourt, Amre, and Levy, 2010). Growth factors including FGFs (superfamily of FGF12) stimulated H2O2 production upon binding to their receptors (Truong and Carroll, 2012). Gene knockout mice of MAF, also known as c-MAF, showed down-regulated GPx3, an antioxidant enzyme, in the kidney (Shirota, Yoshida, Sakai, Kim, Sugiura, Oishi, Nitta, and Tsuchiya, 2006). CDKL5 is involved with oxidative stress observed in Rett syndrome with CDKL5 mutation (Pecorelli, Ciccoli, Signorini, Leoncini, Giardini, D'Esposito, Filosa, Hayek, De Felice, and Valacchi, 2011). Therefore, at least 3 of 5 genes with the most outbound edges and 4 of 7 genes with the most inbound edges are reported to be involved with oxidative stress.

The numbers of selected genes by Lasso and GL are shown in a Venn diagram in (Figure 4.3). Note that the sizes of the circles do not exactly represent the numbers. Although the selected genes by Lasso (83) and GL (108) shared the majority of genes (60) in common, all of the 5 genes with

Table 4.3: Selected genes included in the selected edges by GL.

| Gene | Outbound edges | Inboud edges |
|---|---|---|
| FOXO4 | 11 | 0 |
| TAF9 | 9 | 0 |
| TAF12 | 9 | 0 |
| POU2F1 | 8 | 6 |
| HNF4A | 7 | 0 |
| NFE2L1 | 5 | 0 |
| CEBPA | 5 | 0 |
| JUN | 4 | 1 |
| FOXO3 | 3 | 2 |
| FOXM1 | 3 | 0 |
| MAF | 2 | 6 |
| VDR | 2 | 2 |
| RUNX1 | 2 | 2 |
| KAT2B | 2 | 0 |
| RXRB | 1 | 2 |
| FGF12 | 0 | 8 |
| HNRNPA0 | 0 | 6 |
| RBP2 | 0 | 5 |
| S100G | 0 | 5 |
| CDKL5 | 0 | 5 |
| CD69 | 0 | 4 |
| SULT2A1 | 0 | 3 |
| CYP1A1 | 0 | 3 |
| NPAS3 | 0 | 3 |
| GPX2 | 0 | 2 |
| SLC5A7 | 0 | 2 |
| CYP2B6 | 0 | 1 |
| S100A5 | 0 | 1 |
| ASCL1 | 0 | 1 |
| TPPP3 | 0 | 1 |
| DGKG | 0 | 1 |
| ATP1A4 | 0 | 1 |

the most outbound edges and 5 of the 7 genes with the most inbound edges (FGF12, POU2F1, MAF, HNRNPA0, and CDKL5) by GL were not selected by Lasso.

Canonical pathway analysis showed that the ”NRF2-mediated Oxidative Stress Response” pathway was significantly ($p < 0.05$) involved with the selected genes by GL, but not with the selected genes by Lasso.

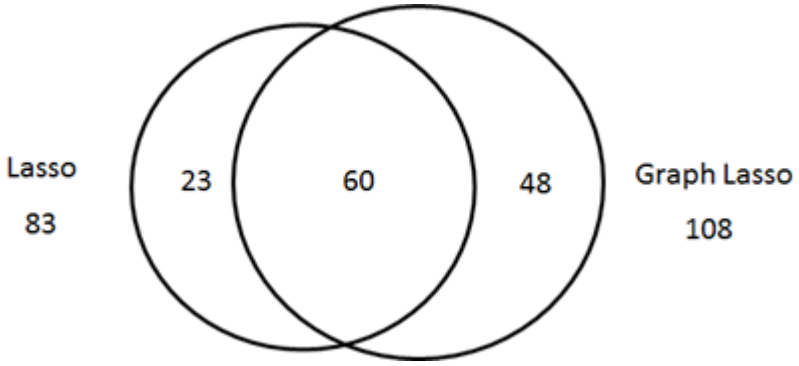


Figure 4.3: Overlap of Selected Genes between Lasso and GL.

## 4.6 Discussion

While GFL did not lead to a sparse estimation that LGL cannot achieve either, GL succeeded in building a sparse, accurate, and biologically relevant model for prediction of liver weight gain in rats based on microarray data and transcription factor network information. The reason behind this difference is that GFL makes a 'smooth' estimation (i.e. connected variables in a graph tend to be assigned close values) and this does not necessarily mean a sparse estimation. This characteristic of GFL renders it especially suitable for image processing such as a case reported in (Xin, Kawahara, Wang, and Gao, 2014), where selections of few spatially connected regions in a brain image would help doctors understand a model and make a diagnosis of Alzheimer's disease based on that. However, this behavior of generalized fused Lasso is not as attractive in our case as in image processing, since selected genes do not need to be interconnected with each other.

Edge-based sparse selection of genes by GL allowed us to infer that the mechanism behind liver weight gain is related to oxidative stress. It is well established that oxidative stress induces liver weight gain (Das and Vasudevan, 2005; Lankoff, Banasik, and Nowak, 2002) through inductions of antioxidant enzymes (mainly phase 2 detoxifying enzymes) (Xu, Hellerbrand, Kohler, Bugnon, Kan, Werner, and Beyer, 2008). Therefore, the selection of edges by GL proved to be biologically reasonable, since the generated model consisted of many oxidative-related genes.

Interestingly, while Lasso and GL selected the majority of genes in common, most of the selected genes by GL with the most outbound or inbound edges, which led us to the oxidative stress as a putative mechanism, were not selected by Lasso. In addition, canonical pathway analysis suggested that the oxidative-related pathway was involved with the selected genes by GL, but not with the selected genes by Lasso. Taken together, with Lasso, it would have been much more difficult, if not impossible, to infer the oxidative stress as a putative mechanism shared among many com-

pounds inducing liver weight gain. Although our previous study also showed that the selection of groups by LGL suggested the involvement of oxidative stress in the process of liver weight gain, it was easier to reach the same conclusion with graph Lasso, as the number of selected genes were much limited. We should be cautious because the inferred mechanism is only hypothetical and has yet to be confirmed by additional in vivo and/or in vitro studies. Nonetheless, the hypothesis induced from our approach would be valuable because it can pave the way for further experiments.

Sparse estimation brought by GL has another advantage. While microarray gives us a rich source of information that is useful for discussing putative mechanisms behind biological responses and constructing a discriminative model as in this study, screening many compounds in drug development based on constructed models with microarray is expensive and labor-intensive. If we need to evaluate at most 100 or so genes, we can use quantitative real-time PCR-based technologies such as RT2 Profiler$^{TM}$ PCR Arrays (QIAGEN) and TaqMan$^{®}$ Gene Expression Array Cards and Plates (Thermo Fischer Scientific), instead of microarray. Doing so, we can remarkably reduce the cost, labor, and time needed to select safer compounds.

Our approach is not limited to prediction of liver weight gain in rats from microarray, but can be applied to other cases where a graph structure is available. Especially, when structure information is given in the form of groups but their sizes are large, as is often the case in biological applications, and the groups can be rearranged into a graph, our graph-conversion technique would dramatically reduce the size of generated models while keeping accuracy intact.

## 4.7 Conclusion

We applied the GFL and GL methods to the TG-GATEs and MSig databases to build a predictive classifier of liver weight gain for an unknown compound. We compared the generated classifiers among Lasso, LGL, GFL, and GL, and showed that GL generated an accurate, biologically relevant and sparse model that could not have been possible with LGL and GFL.

# Chapter 5

# Discussion and Conclusion

## 5.1 Discussion on Entire Study

We have introduced three different types of machine learning methods - class association rule mining, group-based sparse regularization, and graph-based sparse regularization - and applied them to build a accurate and biologically relevant toxicogenomic models for liver weight gain in rats. Each of them has their advantages and disadvantages.

While regularization methods build a model by solving a continuous optimization problem, class association rule mining takes a combinatorial approach to find useful patterns from data and build a model with them. By doing so, class association rule mining can better handle nonlinear samples. This is especially the case in biology because a biological response is usually triggered not by a single mechanism, but by several different mechanisms. This heterogeneity renders combinatorial approaches suitable for biological applications. However, combinatorial approaches has its downside, since the number of combinations grow rapidly as the sizes of dataset or search space increase. As in Chapter 2, we can optimize a computational time by configuring *minsup* parameter in CBA, but at the expense of detection sensitivity. In our case of liver weight gain in rats, where there are an enough number of positive samples, we can successfully build an accurate classifier by finding an optimal *minsup*. But it cannot always be possible, especially when there are only insufficient positive samples.

In terms of interpretability, class association rule mining can build a highly interpretable model that is understandable for even non-expert users. As discussed in Chapter 1, interpretability is an important prerequisite of biological relevancy, since we cannot judge a model as biologically relevant or not, unless we can derive biological knowledge from it. However, interpretability does not necessarily lead to biological relevancy. In fact, class association rule mining methods, including CBA, does not take any outside information (e.g. literature, experimental results, etc.) into account.

This is where structural regularization methods come into play, as they explicitly employ out-

side structural information in the process of modeling. While an abundance of such information in biology makes them an attractive choice, it also poses a problem: what kind of information should we use? As discussed in Chapter 3, (Obozinski, Jacob, and Vert, 2011) used the canonical pathways from MSigDB for the LGL analysis with microarray data. However, taking the fact into account that such pathways consist of various kinds of interactions including ones with no direct effects on gene expressions, our choice of transcription factor network seems more biologically reasonable. The important reminder here is that the choice of outside structural information should be conducted by an expert who knows both machine learning and application domain well. Accuracy and biological relevancy of classifiers generated by structural regularization methods significantly depend on them.

Although our initial goal of accurate and biological relevant modeling was satisfied with group-based structural regularization, graph-based structural regularization generally leads to a sparser model. One of problem of group-based structural regularization methods such as LGL is that they are highly affected by the sizes of groups and the extent of overlapping. Graph-based structural regularization such as GFL and GL overcomes this difficulty, as it handles edges on a graph instead of groups. If smoothness between connected variables is not required, as in our case, GL is better in building a sparse model. Actually, we demonstrated that GL achieved sparsity almost comparable to Lasso, while retaining accuracy.

Although we obtained satisfying results, our study also reveals remaining issues that can be further improved. First, more computationally efficient algorithms are needed for class association rule mining when applied to cases where the number of positive samples is more restricted. This might be achievable by taking outside knowledge into account, as do structural regularization methods. For example, if we prune rules that do not include any pairs of genes with known relationships, we would be able to not only significantly reduce computation time, but also build a more biologically relevant model than the current algorithms.

Secondly, applicabilities of other types of biological information than our transcription factor network should be explored. Recently, it has been increasingly clear that the molecules called small non-coding RNAs such as micro RNAs (miRNA) and short interfering RNAs (siRNA) play an important role in regulating gene expressions in eukaryotes. While classical transcription factors regulate gene expressions at the pre-transcriptional level (i.e. before genes are transcribed into mRNAs), small non-coding RNAs regulate at the post-transcriptional level (i.e. after genes are transcribed into mRNAs). These non-coding RNAs silence cytoplasmic mRNAs by either repressing translation or promoting degradation (Valencia-Sanchez, Liu, Hannon, and Parker, 2006). Therefore, we can assume that genes regulated by the same non-coding RNA are simultaneously

down-regulated or up-regulated, as did we for genes regulated by transcription factors. Incorporating the non-coding RNA networks would further enrich modeling process with structural regularization.

## 5.2 Conclusion

Throughout this dissertation, in an aim to clarify the characteristics of different machine learning techniques in building a biologically relevant as well as accurate toxicogenomic predictive model, we explored the three different types of machine learning methods that are relatively new and applicable in this end: class association rule mining, group-based sparse regularization, and graph-based sparse regularization.

First, we compared CBA, a class association rule mining method, with LDA, and showed that it had not only a better predictive performance, but also more interpretablily than LDA.

Secondly, to explicitly take into account outside biological information, we compared LGL, a group-based structural regularization method, with Lasso and showed that it achieved predictive performance comparable to Lasso and was biologically relevant for liver weight gain. However, LGL had an downside in that it selected much more genes than Lasso.

Finally, to further achieve sparsity, we compared GFL and GL, graph-based structural regularization methods, with Lasso and LGL. We successfully demonstrated that GL built an accurate, sparse, and biologically relevant model for liver weight gain.

Our study also reveals remaining issues. First, more computationally efficient algorithms are needed for class association rule mining when applied to cases where the number of positive samples is more restricted. For example, if we prune rules that do not include any pairs of genes with known relationships, we would be able to not only significantly reduce computation time, but also build a more biologically relevant model than the current algorithms.

Secondly, applicabilities of other types of biological information than our transcription factor network should be explored. Non-coding RNAs silence cytoplasmic mRNAs by either repressing translation or promoting degradation. Therefore, we can assume that genes regulated by the same non-coding RNA are simultaneously down-regulated or up-regulated, as did we for genes regulated by transcription factors. Incorporating the non-coding RNA networks would further enrich modeling process with structural regularization.

# References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules.

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99 (10), 6562–6.

Annavarapu, S. R., Cialfi, S., Dominici, C., Kokai, G. K., Uccini, S., Ceccarelli, S., McDowell, H. P., and Helliwell, T. R. (2013). Characterization of Wnt/beta-catenin signaling in rhabdomyosarcoma. *Lab Invest*, 93 (10), 1090–9.

Apte, C., Hong, S., Natarajan, R., Pednault, E., Tipu, F., and Weiss, S. (2003). Data Intensive Analytics for Predictive Modeling. *IBM Journal of Research and Development.*, 47 (1), 17–23.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured Sparcity through Convex Optimization. *Statistical Science*, 27 (4), 450–468.

Bateman, A. R., El-Hachem, N., Beck, A. H., Aerts, H. J., and Haibe-Kains, B. (2014). Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci Rep*, 4, 4092.

Carrillo, H., Brodersen, K. H., and Castellanos, J. A. (2014). Probabilistic Performance Evaluation for Multiclass Classification Using the Posterior Balanced Accuracy. *Advances in Intelligent Systems and Computing*, 252, 347–361.

Cheng, C., Shen, K., Song, C., Luo, J., and Tseng, G. C. (2009). Ratio adjustment and calibration scheme for gene-wise normalization to enhance microarray inter-study prediction. *Bioinformatics*, 25 (13), 1655–61.

Chiu, S. H., Chen, C. C., Yuan, G. F., and Lin, T. H. (2006). Association algorithm to mine the rules that govern enzyme definition and to classify protein sequences. *BMC Bioinformatics*, 7, 304.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3, 140.

Church, V. L. and Francis-West, P. (2002). Wnt signalling during limb development. *Int J Dev Biol*, 46 (7), 927–36.

Coenen, F., Goulbourne, G., and Leng, P. (2004). Tree structures for mining association rules. *Data Mining and Knowledge Discovery*, 8 (1), 25–51.

Coenen, F., Leng, P., and Ahmed, S. (2004). Data structure for Association Rule Mining: T-trees and P-trees. *Ieee Transactions on Knowledge and Data Engineering*, 16 (6), 774–778.

Das, S. K. and Vasudevan, D. M. (2005). Effect of ethanol on liver antioxidant defense systems: Adose dependent study. *Indian J Clin Biochem*, 20 (1), 80–4.

Dua, S. and Kidambi, P. C. (2010). Protein structural classification using orthogonal transformation and class-association rules. *Int J Data Min Bioinform*, 4 (2), 175–90.

Ennulat, D., Walker, D., Clemo, F., Magid-Slav, M., Ledieu, D., Graham, M., Botts, S., and Boone, L. (2010). Effects of hepatic drug-metabolizing enzyme induction on clinical pathology parameters in animals and man. *Toxicol Pathol*, 38 (5), 810–28.

Essers, M. A., Weijzen, S., Vries-Smits, A. M. de, Saarloos, I., Ruiter, N. D. de, Bos, J. L., and Burgering, B. M. (2004). FOXO transcription factor activation by oxidative stress mediated by the small GTPase Ral and JNK. *EMBO J*, 23 (24), 4802–12.

Festing, M. F. and Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J*, 43 (4), 244–58.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7 (2), 179–188.

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev*, 29 Suppl 1, S83–7.

Gu, Q., Li, Z., and Han, J. (2011). Linear discriminant dimensionality reduction. Springer, 549–564.

Hall, A. P., Elcombe, C. R., Foster, J. R., Harada, T., Kaufmann, W., Knippel, A., Kuttler, K., Malarkey, D. E., Maronpot, R. R., Nishikawa, A., Nolte, T., Schulte, A., Strauss, V., and York, M. J. (2012). Liver hypertrophy: a review of adaptive (adverse and non-adverse) changes–conclusions from the 3rd International ESTP Expert Workshop. *Toxicol Pathol*, 40 (7), 971–94.

Hodar, C., Assar, R., Colombres, M., Aravena, A., Pavez, L., Gonzalez, M., Martinez, S., Inestrosa, N. C., and Maass, A. (2010). Genome-wide identification of new Wnt/beta-catenin target genes in the human genome using CART method. *BMC Genomics*, 11, 348.

Hubbell, E., Liu, W. M., and Mei, R. (2002). Robust estimators for expression analysis. *Bioinformatics*, 18 (12), 1585–92.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group Lasso with Overlap and Graph Lasso.

Kang, J., Gemberling, M., Nakamura, M., Whitby, F. G., Handa, H., Fairbrother, W. G., and Tantin, D. (2009). A general mechanism for transcription regulation by Oct1 and Oct4 in response to genotoxic and oxidative stress. *Genes Dev*, 23 (2), 208–22.

Kianmehr, K. and Alhajj, R. (2008). CARSVM: a class association rule-based classification framework and its application to gene expression data. *Artif Intell Med*, 44 (1), 7–25.

Kolligs, F. T., Bommer, G., and Goke, B. (2002). Wnt/beta-catenin/tcf signaling: a critical pathway in gastrointestinal tumorigenesis. *Digestion*, 66 (3), 131–44.

Kondoh, N., Ohkura, S., Arai, M., Hada, A., Ishikawa, T., Yamazaki, Y., Shindoh, M., Takahashi, M., Kitagawa, Y., Matsubara, O., and Yamamoto, M. (2007). Gene expression signatures that can discriminate oral leukoplakia subtypes and squamous cell carcinoma. *Oral Oncol*, 43 (5), 455–62.

Lankoff, A., Banasik, A., and Nowak, M. (2002). Protective effect of melatonin against nodularin-induced oxidative stress. *Arch Toxicol*, 76 (3), 158–65.

Liu, B., Hsu, W., and Ma, Y. (1998). Integrating Classification and Association Rule Mining, 80–86.

Long, A. D., Mangalam, H. J., Chan, B. Y., Tolleri, L., Hatfield, G. W., and Baldi, P. (2001). Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem*, 276 (23), 19937–44.

Ma, S., Song, X., and Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8, 60.

Marcil, V., Seidman, E., Sinnett, D., Boudreau, F., Gendron, F. P., Beaulieu, J. F., Menard, D., Precourt, L. P., Amre, D., and Levy, E. (2010). Modification in oxidative stress, inflammation, and lipoprotein assembly in response to hepatocyte nuclear factor 4alpha knockdown in intestinal epithelial cells. *J Biol Chem*, 285 (52), 40448–60.

McCarthy, D. J. and Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25 (6), 765–71.

Meier, L., Geer, S. v. d., and Bühlmann, P. (2008). The group lasso for logistic regression. *J. Roy. Stat. Soc. B*, 70 (1), 53–71.

Nagata, K., Kawahara, Y., Washio, T., and Unami, A. (2015). Toxicogenomic prediction with group sparse regularization based on transcription factor network information. *Fundamental Toxicological Sciences*, 2 (4), 161–170.

Nagata, K., Kawahara, Y., Washio, T., and Unami, A. (2016). Toxicogenomic prediction with graph-based structured regularization on transcription factor network. *Fundamental Toxicological Sciences*, 3 (2), 39–46.

Nagata, K., Washio, T., Kawahara, Y., and Unami, A. (2014). Toxicity prediction from toxicogenomic data based on class association rule mining. *Toxicology Reports*, 1, 1133–1142.

Nuwaysir, E. F., Bittner, M., Trent, J., Barrett, J. C., and Afshari, C. A. (1999). Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog*, 24 (3), 153–159.

Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group Lasso with Overlaps: the Latent Group Lasso approach.

Pach, F. P., Gyenesei, A., and Abonyi, J. (2008). Compact fuzzy association rule-based classifier. *Expert Systems with Applications*, 34 (4), 2406–2416.

Paul, R., Groza, T., Hunter, J., and Zankl, A. (2014). Inferring characteristic phenotypes via class association rule mining in the bone dysplasia domain. *J Biomed Inform*, 48, 73–83.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21 (13), 3017–24.

Pecorelli, A., Ciccoli, L., Signorini, C., Leoncini, S., Giardini, A., D'Esposito, M., Filosa, S., Hayek, J., De Felice, C., and Valacchi, G. (2011). Increased levels of 4HNE-protein plasma adducts in Rett syndrome. *Clin Biochem*, 44 (5-6), 368–71.

Phan, J. H., Quo, C. F., and Wang, M. D. (2006). Functional genomics and proteomics in the clinical neurosciences: data mining and bioinformatics. *Prog Brain Res*, 158, 83–108.

Rada, P., Rojo, A. I., Offergeld, A., Feng, G. J., Velasco-Martin, J. P., Gonzalez-Sancho, J. M., Valverde, A. M., Dale, T., Regadera, J., and Cuadrado, A. (2015). WNT-3A regulates an Axin1/NRF2 complex that regulates antioxidant metabolism in hepatocytes. *Antioxid Redox Signal*, 22 (7), 555–71.

Rao, R. C. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10 (2), 159–203.

Ratsch, G., Sonnenburg, S., and Schafer, C. (2006). Learning interpretable SVMs for biological sequence classification. *BMC Bioinformatics*, 7 Suppl 1, S9.

Sampson, D. L., Parker, T. J., Upton, Z., and Hurst, C. P. (2011). A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches. *PLoS One*, 6 (9), e24973.

Shi, W., Bugrim, A., Nikolsky, Y., Nikolskya, T., and Brennan, R. J. (2008). Characteristics of genomic signatures derived using univariate methods and mechanistically anchored functional descriptors for predicting drug- and xenobiotic-induced nephrotoxicity. *Toxicol Mech Methods*, 18 (2-3), 267–76.

Shirota, S., Yoshida, T., Sakai, M., Kim, J. I., Sugiura, H., Oishi, T., Nitta, K., and Tsuchiya, K. (2006). Correlation between the expression level of c-maf and glutathione peroxidase-3 in c-maf -/- mice kidney and c-maf overexpressed renal tubular cells. *Biochem Biophys Res Commun*, 348 (2), 501–6.

Silver, M., Montana, G., and Alzheimer's Disease Neuroimaging, I. (2012). Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Stat Appl Genet Mol Biol*, 11 (1), Article 7.

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

Suter, L., Babiss, L. E., and Wheeldon, E. B. (2004). Toxicogenomics in predictive toxicology in drug development. *Chem Biol*, 11 (2), 161–171.

Tamura, M. and D'Haeseleer, P. (2008). Microbial genotype-phenotype mapping by class association rule mining. *Bioinformatics*, 24 (13), 1523–1529.

Tao, G. Z., Lehwald, N., Jang, K. Y., Baek, J., Xu, B., Omary, M. B., and Sylvester, K. G. (2013). Wnt/beta-catenin signaling protects mouse liver against oxidative stress-induced apoptosis through the inhibition of forkhead transcription factor FoxO3. *J Biol Chem*, 288 (24), 17214–24.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B.*, 58 (1), 267–288.

Tibshirani, R. J. and Taylor, J. (2011). The Solution Path of the Generalized Lasso. *Annals of Statistics*, 39 (3), 1335–1371.

Truong, T. H. and Carroll, K. S. (2012). Redox regulation of epidermal growth factor receptor signaling through cysteine oxidation. *Biochemistry*, 51 (50), 9954–65.

Valdivia, L. E., Young, R. M., Hawkins, T. A., Stickney, H. L., Cavodeassi, F., Schwarz, Q., Pullin, L. M., Villegas, R., Moro, E., Argenton, F., Allende, M. L., and Wilson, S. W. (2011). Lef1-dependent Wnt/beta-catenin signalling drives the proliferative engine that maintains tissue homeostasis during lateral line development. *Development*, 138 (18), 3931–41.

Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes and Development*, 20 (5), 515–524.

Venables, W. and Ripley, B. (2002). Modern Applied Statistics with S. 4th. Springer.

Wei, W., Zeve, D., Suh, J. M., Wang, X., Du, Y., Zerwekh, J. E., Dechow, P. C., Graff, J. M., and Wan, Y. (2011). Biphasic and dosage-dependent regulation of osteoclastogenesis by beta-catenin. *Mol Cell Biol*, 31 (23), 4706–19.

Welle, S., Brooks, A. I., and Thornton, C. A. (2002). Computational method for reducing variance with Affymetrix microarrays. *BMC Bioinformatics*, 3, 23.

Xin, B., Kawahara, Y., Wang, Y., and Gao, W. (2014). Efficient Generalized Fused Lasso and its application to the diagnosis of Alzheimer's Disease. *Proceedings of the 28th AAAI Conference on Artificial Intelligent*, 2163–2169.

Xu, W., Hellerbrand, C., Kohler, U. A., Bugnon, P., Kan, Y. W., Werner, S., and Beyer, T. A. (2008). The Nrf2 transcription factor protects from toxin-induced liver injury and fibrosis. *Lab Invest*, 88 (10), 1068–78.

Ye, J., Xiong, T., Li, Q., Janardan, R., Bi, J., Cherkassky, V., and Kambhamettu, C. (2006). Efficient model selection for regularized linear discriminant analysis.

Zhao, T., Gan, Q., Stokes, A., Lassiter, R. N., Wang, Y., Chan, J., Han, J. X., Pleasure, D. E., Epstein, J. A., and Zhou, C. J. (2014). beta-catenin regulates Pax3 and Cdx2 for caudal neural tube closure and elongation. *Development*, 141 (1), 148–57.

# List of Publications

**A. Research Papers** (Peer-reviewed)

- Nagata, K., Washio, T., Kawahara, Y. , and Unami, A. (2014). Toxicity prediction from toxicogenomic data based on class association rule mining. *Toxicology Reports*, 1, 1133-1142. **(A-1)**

- Nagata, K., Kawahara, Y., Washio, T. , and Unami, A. (2015). Toxicogenomic Prediction with group sparse regularization based on transcription factor network information. *Fundamental Toxicological Sciences*, 2 (4), 161-170. **(A-2)**

- Nagata, K., Kawahara, Y., Washio, T. , and Unami, A. (2016). Toxicogenomic prediction with graph-based structured regularization on transcription factor network. *Fundamental Toxicological Sciences*, 3 (2), 39-46. **(A-3)**

- Araki, T., Iwasaki, N., Ishiguro, N., Sakamoto, A., Nagata, K., Ohbuchi, M., Moriguchi, H., Motoi, M., Shinkyo, R., Homma, T., Sakamoto, S., Iwase, Y., Ise, R., Nakanishi, Y., Uto, M., and Inoue, T. (2016). Requirements for human iPS cell-derived hepatocytes as an alternative to primary human hepatocytes for assesing absorption, distribution, metabolism, excretion and toxicity of pharmaceuticals. *Fundamental Toxicological Sciences*, 3 (3), 89-99. **(A-4)**

**B. Presentations in Scientific Conferences** (Not peer-reviewed)

- Nagata, K., Tamura, K., Unami, A., and Miyamae, Y. (2013). ToxWorkshop: an extendable workflow software for data processing of massive diverse data. *21st Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2013)*. Berlin, Germany. **(B-1)**

- Yamaura, Y., Kanki, M., Nagata, K., Sasaki, D., Ono, M., Unami, A., and Nakagawa, S. (2015). Serum miR-206 as a useful biomarker of skeletal muscle injury compared with conventional biomarkers. *Society of Toxicology 54th Annual Meeting (SOT2015)*. San Diego, CA, USA. **(B-2)**

- Kazusa, K., Watanabe, Y., Ishikawa, N., Nagata, K., Sato, Y., Fujimori, C., and Okada, A. (2015). The usefulness of the xCELLigence system using human iPS-derived cardiomyocytes in identification of CV liabilities in the early stage of drug development. *The 42nd Annual Meeting of the Japanese Society of Toxicology (JSOT2015)*. Kanazawa, Japan. **(B-3)**

- Nagata, K., Washio, T., Kawahara, Y., and Unami, A. (2015). New toxicogenomic predictive model for decreased reticulocytes based on gene expressions in liver of rats built with class association rule mining. *23rd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2015)*. Dublin, Ireland. **(B-4)**

- Araki, T., Iwasaki, N., Nagata, K., Ohbuchi, M., Moriguchi, H., Motoi, M., Shinkyo, R., Sakamoto, S., Iwase, Y., and Ishiguro, N. (2015). Enzyme activities of human iPS-derived hepatocytes required for ADME research. *30th JSSX (Japanese Society for the Study of Xenobiotics (JSSX2015)) Annual Meeting*. Tokyo, Japan. **(B-5)**

- Okamura, A., Shirakawa, T., Nagata K., and Okada, A. (2016). Potential use of microelectrode arrays (MEA) using human iPS-derived neuronal cells to evaluate neurotoxicity. *Society of Toxicology 55th Annual Meeting (SOT2016)*. New Orleans, LA, USA. **(B-6)**