Fundamental Toxicological Sciences (Fundam. Toxicol. Sci.)
Vol.3, No.2, 39-46, 2016

Original Article

Toxicogenomic prediction with graph-based structured regularization on transcription factor network

Keisuke Nagata[1,2], Yoshinobu Kawahara[2], Takashi Washio[2] and Akira Unami[1]

[1]Drug Safety Research Laboratories, Astellas Pharma Inc., 2-1-6 Kashima, Yodogawa-ku, Osaka, 532-8514, Japan
[2]The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan

(Received February 9, 2016; Accepted February 16, 2016)

ABSTRACT — Structured regularization is a mathematical technique which incorporates prior structural knowledge among variables into regression analysis to make a sparse estimation reflecting relationships among them. Abundance of structural information in biology, such as pathways formed by genes, transcripts, and proteins, especially suits well its application. Previously, we reported on the first application of latent group Lasso, a group-based regularization method, in toxicogenomics, with genes regulated by the same transcription factor treated as a group. We revealed that it achieved good predictive performances comparable to Lasso and allowed us to discuss mechanisms behind liver weight gain in rats based on selected groups. Latent group Lasso, however, does not lead to a sparse estimation, due to large group sizes in our analytical setting. In this study, we applied graph-based regularization methods, generalized fused Lasso and graph Lasso, for the same data, with regulatory networks formed by transcription factors and their downstream genes as a graph. These methods are expected to make a sparser estimation since they select variables based on edges. Comparisons showed that graph Lasso generated an accurate, biologically relevant and sparse model that could not have been possible with latent group Lasso and generalized fused Lasso.

Key words: Structured regularization, Transcription factor network, Generalized fused Lasso, Graph Lasso

INTRODUCTION

In regularized regression analysis, a regression parameter $\mathbf{w} \in \mathbb{R}^d$ 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:

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

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. While an estimation with L1-norm as a regularization term (known as Lasso) is known to generally leads to an accurate and sparse model (Tibshirani, 1996), Lasso ignores structural relationships of explanatory variables (e.g. biological pathways).

Recently, various kinds of regularization terms, collectively known as structured regularization terms, have been proposed to take such structural relationships into account as a prior knowledge in estimation of a regression model. Among such techniques is latent group Lasso (Jacob *et al.*, 2009; Obozinski *et al.*, 2011), which uses a group structure as a prior structural information and employs the following norm as a regularization term:

$$\mathrm{P}_{LGL}(\mathbf{w}) = \min_{\mathbf{v}^g \in \mathbb{R}^{p+1} \mid g \in \mathcal{G},\ v_i^g = 0 \text{ if } i \notin g} \sum_{g \in \mathcal{G}} d_g \|\mathbf{v}^g\|_2, \quad s.t. \sum_{g \in \mathcal{G}} \mathbf{v}^g = \mathbf{w},$$

where $\mathcal{G} \subset \mathcal{P}([1, p])$ is a set of groups (the power set of $\{1, \ldots, p\}$), $g \in \mathcal{G}$ is a group that is a subset of indexes of parameter $\mathbf{w}$, and $d_g \in \mathbb{R}_+$ is a weight for a group g. This norm tends to select explanatory variables as unions of groups. In our previous study (Nagata *et al.*, 2015), we applied latent group Lasso in toxicogenomics. Utilizing the TG-GATEs database, we built both Lasso and latent group Lasso classifiers to predict whether a chemical compound induces liver weight gain after 14-day repetitive treatments in rats based on transcriptomic data after 3-day repetitive treatments. We also used the transcription

Correspondence: Keisuke Nagata (E-mail: keisuke.nagata@astellas.com)

factor target gene sets from the MSigDB database as a prior structural information, since genes regulated by the same transcription factor are expected to be co-expressed and therefore should be selected to or discarded from a model together. Our study showed that latent group Lasso marked a good predictive performances comparable to Lasso and had an advantage over Lasso in that the selected groups by latent group Lasso enabled us to discuss biological mechanisms behind liver weight gain. However, it also revealed that latent group Lasso selected a much larger number of genes than Lasso, which deemed a disadvantage.

This setback of latent group Lasso might be attributable to large sizes of the groups that we used (often several hundred genes). Since many of the genes regulated by transcription factors regulate other genes functioning as transcription factors themselves, relationships among genes in the form of groups can be rearranged into a graph structure by drawing edges between transcription factors and their downstream genes. Doing so, structured regularization methods on a graph structure are applicable as well for our setting. Graph-based structured regularization methods such as generalized fused Lasso and graph Lasso select variables on edge basis, instead of group basis. Therefore, we expected that graph-based structured regularizations would lead to sparse as well as accurate and biologically relevant estimations which group-based structured regularizations cannot.

Generalized fused Lasso takes two regularization terms and can be represented as follows:

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

where $G = (V, E)$ is a graph with vertices V and edges E (Tibshirani and Taylor, 2011). The first regularization term is L1-norm and the second is called the fused term, which penalizes pairwise differences between variables connected by edges (Xin *et al.*, 2014).

On the other hand, graph Lasso is an extension of latent group Lasso 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 variables that it connects).

While several authors reported applications of structured regularization techniques in biological fields (Ma *et al.*, 2007; Obozinski *et al.*, 2011; Silver *et al.*, 2012), there have been so far no direct implications of the graph-based structured regularization based on the transcription factor network in toxicogenomics. In this study, we compared the predictive performances, sparsity and biological relevance among Lasso, latent group Lasso, generalized fused Lasso and graph Lasso, when applied to the same setting as in our previous study. For generalized fused Lasso and graph Lasso, we converted the original group information into a graph structure and used it as a prior structural knowledge.

MATERIAL AND METHODS

We conducted this study based on our previous study (Nagata *et al.*, 2015). Note that the methods shown here included the same or similar explanations as (Nagata *et al.*, 2015).

Data sources

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). To construct a predictive model for liver weight gain after repetitive doses of compounds based on transcriptomic data of a shorter period, from the TG-GATEs database, we used gene expression data (n = 3 per group) one day after 3-day repetitive doses (hereinafter 4D) in the liver of rats as explanatory variables and liver weight data (n = 5 per group) one day after 14-day repetitive doses (15D) in rats as response variables for this study. For each compound, only the data of the highest dose group and its control group was used. Of the 150 compounds, we omitted one compound and analyzed the remaining 149 compounds because that one compound was found to have killed animals before 15D in the study and therefore no data is available for liver weight of 15D.

MSigDB is a collection of annotated gene sets (Subramanian *et al.*, 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. Hereafter, we call this set of groups the transcription factor target groups.

Because genes are represented as human Entrez IDs in MSigDB, we mapped probe set IDs of Affymetrix GeneChip Rat Genome 230 2.0 Array to Entrez IDs of their corresponding human homologue genes, following (Nagata *et al.*, 2015).

Rearrangement of groups into graph

For generalized fused Lasso and graph Lasso, we rearranged transcription factor target groups into a graph. Conversion process is depicted in Fig. 1. While the converted graph is directed, generalized fused Lasso and graph Lasso does not distinguish directions of edges.

Data preprocess

Prior to the analysis, we have preprocessed gene expression data and liver weight data. First, gene expressions were corrected and normalized by the MAS 5.0 algorithm (Hubbell *et al.*, 2002) to reduce inter-array variances (Welle *et al.*, 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 *et al.*, 2012). Secondly, values were averaged over individual animals included in each experimental 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 *et al.*, 2009). Finally, based on fold changes (fc) and p values (p) of the student's t-test conducted between a compound-treated group and its corresponding control group, experimental groups with liver weights satisfying $fc > 1$ and $p < 0.05$ were labeled as positive and otherwise as negative. In general, numerical parameters 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.

Data analysis

For Lasso, latent group Lasso and graph Lasso analyses, we used the MATLAB® (The MathWorks, Inc.) code based on the algorithm of (Meier *et al.*, 2008) and (Jacob *et al.*, 2009) available on Dr. Jacob's Homepage (http://cbio.ensmp.fr/~ljacob/). For generalized fused Lasso, we used the MATLAB® code based on the algorithm of (Xin *et al.*, 2014) available on Dr. Wang's Homepage (http://idm.pku.edu.cn/staff/wangyizhou/).

We followed the procedure of (Nagata *et al.*, 2015), which is based on the pathway analysis experiment for breast cancer data reported in (Jacob *et al.*, 2009; Obozinski *et al.*, 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. Secondly, internal 5-fold cross validations (internal CV), further splitting the training set, for each $\lambda \in [2^x | x = 0, -0.5, -1, \cdots, -12]$ for Lasso, latent group Lasso, and graph Lasso, or $(\lambda_1, \lambda_2) \in [2^x | x = 0, -2, -4, \cdots, -12] \times [2^x | x = 0, -2, -4, \cdots, -12]$ for generalized fused Lasso, were conducted to select the best $\lambda(\mathrm{s})$ based on the average balanced accuracy. Thirdly, the model was built on the training set with selected $\lambda(\mathrm{s})$ 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.

For discussions of sparsity and biological relevance, we used all the samples as the training set and conducted no external CV, whereas internal CV for selecting $\lambda(\mathrm{s})$

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

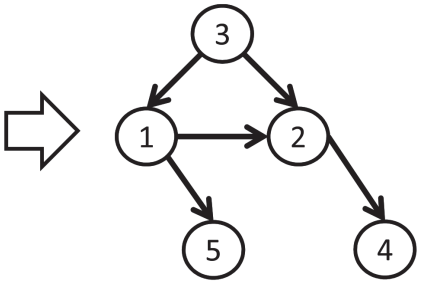

Fig. 1. Schematic of Rearrangement Process of Groups into a Graph. Each group has a transcription factor (not a group member itself) and a set of regulated genes. Transcription factors and regulated genes are both represented in common identifying numbers (human Entrez IDs in actual analyses). For each 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 this schematic. Repeating this process for all the three groups makes an entire graph as shown here.

were still conducted.

To make comparisons fair, similarly with (Nagata *et al.*, 2015), we prepared dummy groups, each of which includes genes that were not included in any groups (one gene per group), in prior to latent group Lasso and graph Lasso analyses.

Cross validation

K-fold cross validation, is one of the standard methods for evaluating predictive performances of classifiers. This method divide a dataset into equally-sized k partitions (1, 2, ..., k). In the first step, the first partition (1) is reserved as a test set and the other partitions (2, 3, ... k) are used as a training set to build a classifier. Once a classifier is built, it is validated for its predictive performances with a test set (the first partition in this case). k-fold cross validation repeats this steps k times changing a partition serving as a test set one by one. In the end, averaged predictive performance over k validation steps is regarded as the predictive performance of a classification algorithm.

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

Predictive performance

We used the following parameters (Carrillo *et al.*, 2014; Florkowski, 2008) to compare predictive performances.

Balanced Accuracy: (Sensitivity + Specificity) / 2
Sensitivity: True Positive / (True Positive + False Negative)
Specificity: True Negative / (True Negative + False Positive)

Student's t-test

For statistical comparison at discretization steps of liver weights between a compound-treated group and its corresponding control group for each compound conducted, the unpaired two tailed student's t-test without equal variance assumption was conducted.

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 absolute values exceeded a threshold. For each estimated parameter, we set a threshold as 1/1,000 of the maximum absolute value. We set a different criteria from our previous study (Nagata *et al.*, 2015), where the threshold was set as zero, since the previous threshold was too strict for the generalized fused Lasso code used in this study.

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) (Nagata *et al.*, 2015).

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 transcription factor target groups and the graph converted from these transcription factor groups, regardless of employed regularization methods. Note that edge groups temporarily prepared for graph Lasso were not used here. Also note that dummy groups in latent group Lasso and graph Lasso were not included.

Pathway analysis

Canonical Pathway analysis was conducted with QIAGEN's Ingenuity Pathway Analysis (IPA) software. Canonical pathway analysis answers how statistically significantly the pre-defined sets of molecules based on literature were affected, considering how many molecules a user-specified set and each pre-defined set share. In this study, we used only the genes that can be mapped to human Entrez IDs for the analysis.

Computer

We used a personal computer with Intel Xeon E5620 CPU (2.40 and 2.39 GHz processors) and 48.0 GB RAM for the analyses.

RESULTS

Predictive performance

We compared predictive performance of generated classifiers of Lasso, latent group Lasso, generalized fused Lasso, and graph Lasso in 5-fold cross validations (Table 1). All the four methods achieved almost equivalent performances in terms of balanced accuracy, while generalized fused Lasso scored lower sensitivity and higher specificity compared to the other methods.

Group sparsity

We compared gene-level, group-level, and edge-level sparsity of generated classifiers for Lasso, latent group Lasso, generalized fused Lasso, and graph Lasso (Table 2).

Table 1. Comparison of predictive performance.

	Lasso	LGL	GFL	GL
Balanced accuracy (%)	73 ± 4	74 ± 8	75 ± 7	75 ± 4
Sensitivity (%)	62 ± 4	62 ± 12	56 ± 15	67 ± 14
Specificity (%)	83 ± 9	86 ± 8	93 ± 3	83 ± 9

Predictive performance for generated classifiers of Lasso, latent group Lasso (LGL), generalized fused Lasso (GFL), and graph Lasso (GL) was compared in 5-fold cross validations. Values are shown as mean ± standard deviation (%).

Table 2. Comparison of sparsity.

	Lasso	LGL	GFL	GL
Number of selected genes	83	2924	8380	108
Number of selected groups	0	7	3	0
Number of selected edges	0	6368	6996	73

Numbers of selected genes, groups, and edges for Lasso, latent group Lasso (LGL), generalized fused Lasso (GFL), and graph Lasso (GL) were compared.

As our previous study showed, Latent group Lasso led to a much larger number of selected genes (2,924) than Lasso (83). Latent group Lasso selected 7 groups and 6,368 edges, while Lasso selected no group or edges.

Generalized fused Lasso selected 8,380 genes, even larger than latent group Lasso. The numbers of selected groups and edges by generalized fused Lasso were 3 and 6,996, both comparable to latent group Lasso.

Graph Lasso selected 108 genes, much fewer than latent group Lasso and generalized fused Lasso and comparable to Lasso. Graph Lasso selected no group. The number of selected edges by graph Lasso was 73, in stark contrast to the other three methods.

Biological relevance

We further investigated 73 edges selected by graph Lasso. We extracted a subgraph that contains only the selected edges (Fig. 2). Then, we summarized the numbers of outbound and inbound edges for the selected genes included in the selected edges (Table 3).

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). Interestingly, we found that many of these genes are reportedly linked to oxidative stress. FOXO4 was activated by oxidative stress generated by H_2O_2, through nuclear translocation and transcriptional activation of FOXO4, in cultured cells (Essers *et al.*, 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 *et al.*, 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 *et al.*, 2010). Growth factors including FGFs (superfamily of FGF12) stimulated H_2O_2 production upon binding to their receptors (Truong and Carroll, 2012). Gene knockout mice of MAF, also known as c-MAF, showed downregulated GPx3, an antioxidant enzyme, in the kidney (Shirota *et al.*, 2006). CDKL5 is involved with oxidative stress observed in Rett syndrome with CDKL5 mutation (Pecorelli *et al.*, 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.

Although the selected genes by Lasso (83) and graph Lasso (108) shared the majority of genes (60) in common (Fig. 3), all of the 5 genes with the most outbound edges and 5 of the 7 genes with the most inbound edges (FGF12, POU2F1, MAF, HNRNPA0, and CDKL5) by graph Lasso 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 graph Lasso, but not with the selected genes by Lasso.

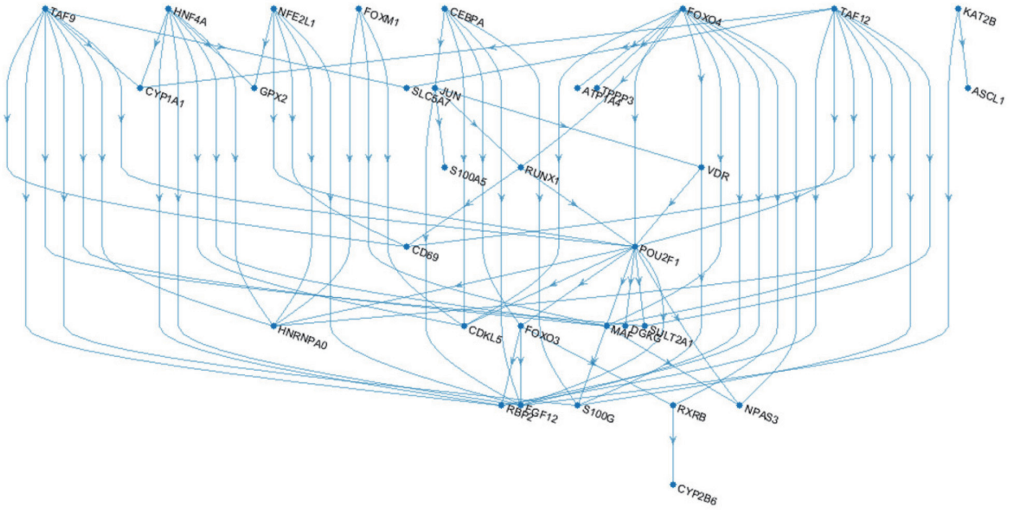

Fig. 2. Subgraph of Selected Edges by Graph Lasso. Each point represent a vertex (corresponding to a gene) labelled with its gene symbol. Each line represents an edge with an arrow showing a direction from a transcription factor to its downstream gene.

DISCUSSION

While generalized fused Lasso did not lead to a sparse estimation that latent group Lasso cannot achieve, graph Lasso 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 generalized fused Lasso 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 generalized fused Lasso renders it especially suitable for image processing such as a case reported in (Xin *et al.*, 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 graph Lasso 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 *et al.*, 2002) through inductions of antioxidant enzymes (mainly phase 2 detoxifying enzymes) (Xu *et al.*, 2008). While hepatomegaly without histological or clinical pathological alterations indicative of liver toxicity is usually considered an adaptive and non-adverse reaction, certain degrees of liver weight increase correlated with the subsequent development of irreversible toxicity such as fibrosis, necrosis, vacuolization, fatty degeneration, and even neoplasia (Hall *et al.*, 2012). Therefore, the selection of edges by graph Lasso proved to be biologically reasonable, since the generated model consisted of many oxidative-related genes.

Interestingly, while Lasso and graph Lasso selected the majority of genes in common, most of the selected genes by graph Lasso 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 graph Lasso, 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 compounds inducing liver weight gain. Although our previous study also showed that the selection of groups by latent graph Lasso 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 graph Lasso has another advantage. While microarray gives us a rich source of

Table 3. Selected genes included in the selected edges by graph Lasso.

Gene	Outbound Edges	Inbound 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 total of 32 selected genes included in selected edges by graph Lasso are listed in order of the number of outbound edges.

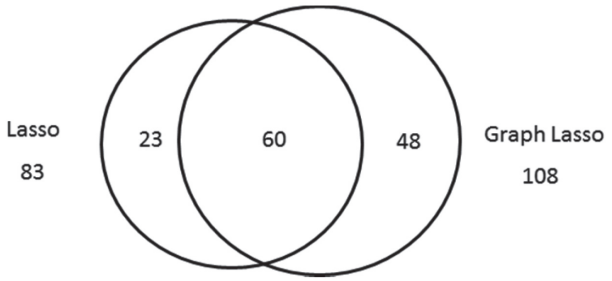

Fig. 3. Overlap of Selected Genes between Lasso and Graph Lasso. The numbers of selected genes by Lasso and graph Lasso are shown in a Venn diagram. The sizes of the circles do not exactly represent the numbers.

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 RT^2 Profiler™ 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.

ACKNOWLEDGMENTS

We wish to thank Dr. Laurent Jacob (University of California, Berkeley) and Dr. Wang (Peking University) for kindly making their MATLAB codes publicly available.

Conflict of interest---- The authors declare that there is no conflict of interest.

REFERENCES

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**, 1655-1661.

Das, S.K. and Vasudevan, D.M. (2005): Effect of ethanol on liver antioxidant defense systems: Adose dependent study. Ind. J. Clin. Biochem., **20**, 80-84.

Essers, M.A., Weijzen, S., de Vries-Smits, A.M., Saarloos, I., de

Ruiter, N.D., 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**, 4802-4812.

Festing, M.F. and Altman, D.G. (2002): Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J., **43**, 244-258.

Florkowski, C.M. (2008): Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. Clin. Biochem. Rev./ Australian Association of Clinical Biochemists, **29 Suppl. 1**, S83-S87.

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**, 971-994.

Hubbell, E., Liu, W.M. and Mei, R. (2002): Robust estimators for expression analysis. Bioinformatics, **18**, 1585-1592.

Jacob, L., Obozinski, G. and Vert, J.-P. (2009): Group Lasso with Overlap and Graph Lasso. International Conference on Machine Learning (ICML 26).

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**, 208-222.

Lankoff, A., Banasik, A. and Nowak, M. (2002): Protective effect of melatonin against nodularin-induced oxidative stress. Arch. Toxicol., **76**, 158-165.

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**, 40448-40460.

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

Nagata, K., Kawahara, Y., Washio, T. and Unami, A. (2015): Toxicogenomic prediction with group sparse regularization based on transcription factor network information. Fundam. Toxicol. Sci., **2**, 161-170.

Obozinski, G., Jacob, L. and Vert, J.-P. (2011): Group Lasso with Overlaps: the Latent Group Lasso approach. Technical Report. arXiv:1110.0413.

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**, 368-371.

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**, 501-506.

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**, 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. USA, **102**, 15545-15550.

Tibshirani, R. (1996): Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B, **58**, 267-288.

Tibshirani, R.J. and Taylor, J. (2011): The solution path of the generalized Lasso. Ann. Statist., **39**, 1335-1371.

Truong, T.H. and Carroll, K.S. (2012): Redox regulation of epidermal growth factor receptor signaling through cysteine oxidation. Biochemistry, **51**, 9954-9965.

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.; J. Tech. Methods Pathol., **88**, 1068-1078.