frontiers in Genetics

ORIGINAL RESEARCH
published: 01 November 2018
doi: 10.3389/fgene.2018.00516

Robust Co-clustering to Discover Toxicogenomic Biomarkers and Their Regulatory Doses of Chemical Compounds Using Logistic Probabilistic Hidden Variable Model

Mohammad Nazmol Hasan[1,2], *Md. Masud Rana*[1], *Anjuman Ara Begum*[1], *Moizur Rahman*[3] *and Md. Nurul Haque Mollah*[1]*

[1] Bioinformatics Laboratory, Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh, [2] Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh, [3] Department of Veterinary and Animal Sciences, University of Rajshahi, Rajshahi, Bangladesh

Detection of biomarker genes and their regulatory doses of chemical compounds (DCCs) is one of the most important tasks in toxicogenomic studies as well as in drug design and development. There is an online computational platform "Toxygates" to identify biomarker genes and their regulatory DCCs by co-clustering approach. Nevertheless, the algorithm of that platform based on hierarchical clustering (HC) does not share gene-DCC two-way information simultaneously during co-clustering between genes and DCCs. Also it is sensitive to outlying observations. Thus, this platform may produce misleading results in some cases. The probabilistic hidden variable model (PHVM) is a more effective co-clustering approach that share two-way information simultaneously, but it is also sensitive to outlying observations. Therefore, in this paper we have proposed logistic probabilistic hidden variable model (LPHVM) for robust co-clustering between genes and DCCs, since gene expression data are often contaminated by outlying observations. We have investigated the performance of the proposed LPHVM co-clustering approach in a comparison with the conventional PHVM and Toxygates co-clustering approaches using simulated and real life TGP gene expression datasets, respectively. Simulation results show that the proposed method improved the performance over the conventional PHVM in presence of outliers; otherwise, it keeps equal performance. In the case of real life TGP data analysis, three DCCs (glibenclamide-low, perhexilline-low, and hexachlorobenzene-medium) for glutathione metabolism pathway dataset as well as two DCCs (acetaminophen-medium and methapyrilene-low) for PPAR signaling pathway dataset were incorrectly co-clustered by the Toxygates online platform, while only one DCC (hexachlorobenzene-low) for glutathione metabolism pathway was incorrectly co-clustered by the proposed LPHVM approach. Our findings from the real data analysis are also supported by the other findings in the literature.

Keywords: toxicogenomic biomarker, doses of chemical compounds (DCCs), co-clustering, outlying observations, logistic transformation, probabilistic hidden variable model (PHVM), logistic probabilistic hidden variable model (LPHVM)

Edited by:
Paul Jennings,
VU University Amsterdam,
Netherlands

Reviewed by:
Olivier Taboureau,
Paris Diderot University, France
Concetta Ambrosino,
University of Sannio, Italy

***Correspondence:**
Md. Nurul Haque Mollah
mollah.stat.bio@ru.ac.bd

Specialty section:
This article was submitted to Toxicogenomics, a section of the journal Frontiers in Genetics

Received: *02 July 2018*
Accepted: *12 October 2018*
Published: *01 November 2018*

Citation:
Hasan MN, Rana MM, Begum AA, Rahman M and Mollah MNH (2018) Robust Co-clustering to Discover Toxicogenomic Biomarkers and Their Regulatory Doses of Chemical Compounds Using Logistic Probabilistic Hidden Variable Model. Front. Genet. 9:516. doi: 10.3389/fgene.2018.00516

INTRODUCTION

Toxicogenomics studies combines toxicology with several *omics* technologies (genomics, transcriptomics, proteomics, and metabolomics) to assess the risk of toxins (small molecules, peptides, or proteins) and chemical agents (drugs, gasoline, alcohol, pesticides, fuel oil, and cosmetics) in organism (NRC, 2007; Afshari et al., 2011). Through integration of these *omics* technologies with bioinformatics, toxicogenomics can be used to suggest the molecular mechanism of toxicity. This can reduce the cost in terms of time, labor, compound synthesis, and animal use which are main limitations of traditional toxicology work (Nuwaysir et al., 1999; Chen et al., 2012). In drug discovery and development, it is also necessary to assess the doses of chemical compounds (DCCs) toxicity administering these DCCs on individuals for measuring drugs' safety. This assessment can be done by toxicogenomic biomarkers those are upregulated or downregulated by the influence of a set of DCCs on individuals. These toxicogenomic biomarkers can be identified from the extensive gene-treatment expression dataset of target organs of individuals (Fielden et al., 2007; Uehara et al., 2008; Igarashi et al., 2015).

An online toxicogenomic data analysis platform "ToxDB" increases its predictive power based on the pathway level gene expression data (Hardt et al., 2016). It calculates the pathway scores for a chemical compound to identify significant biomarker genes using t-statistic from different pathways. Nevertheless, there is no facility in this platform to study another interesting problem of relationship between gene groups and DCCs groups asserted by Afshari et al. (2011). To address this problem another online platform "Toxygates" produces co-clusters between genes and DCCs using hierarchical clustering (HC) (Nyström-Persson et al., 2017). But HC does not use two-way (gene-DCC) information simultaneously for co-clustering and it is sensitive to outlying observations (García-Escudero et al., 2010). Probabilistic hidden variable model (PHVM) has been developed for co-clustering between words and documents in a text mining problem (Hofmann, 2001). It uses two-way (row-column) information simultaneously during co-clustering. It was also successfully used in detecting hidden patters of biological profiling datasets (Joung et al., 2006; Bicego et al., 2010). Therefore, PHVM would be more effective approach than HC for co-clustering between genes and DCCs which is also supported by Joung et al. (2006). However, the PHVM algorithm is sensitive to outlying observations of gene expression. These outlying observations often occur in the gene expression dataset due to several steps involve in the data generating processes from hybridization to image analysis including scratches or dust on the surface, imperfections in the glass or imperfections in the array production (Gottardo et al., 2006; Upton et al., 2009). The outliers in the dataset may arise following Tukey–Huber contamination model (THCM; Agostinelli et al., 2015) or independent contamination model (ICM; Alqallaf et al., 2009). To overcome the robustness problems of conventional PHVM approach an attempt is made to propose logistic PHVM approach called as LPHVM for robust co-clustering between genes and DCCs to discover toxicogenomic biomarkers and their regulatory DCCs.

METHODS AND MATERIALS

Let us consider a toxicogenomic experimental design as described in **Figure 1** that reflects Japanese Toxicogenomics Project (TGP) (Uehara, 2010) experiment for a single time point from which the toxygates (Nyström-Persson et al., 2013) data were collected. According to this design, gene expression data of both treatment and control group of animal samples are assumed to be generated. Then the fold change gene expression data for a single time point are computed from the treatment and control group of animals. It can measure the actual treatment (DCCs) effects on the genes. The fold change gene-expression value of a gene is defined as follows:

$$Y_{tlq} = log_2\left(\frac{x_{tlq}}{x'_{tlq}}\right) = log_2\left(x_{tlq}\right) - log_2\left(x'_{tlq}\right), \qquad (1)$$

where Y_{tlq} is the fold change expression value of a gene for the qth ($q = 1, 2, 3$) sample under lth (l = Low, Middle, High) dose level of the tth ($t = 1, 2, \cdots, T$) chemical compound, x_{tlq} is the expression value of that gene of mentioned sample under the treatment group and $x_{tlq}{}'$ is the expression value of the same gene of the respective control sample. The effect of compound-dose combination or treatment/DCCs on the animal can be measured by $\overline{Y}_{tl.}$ which is the average fold change value over the samples. In this paper, our objective is to robust co-clustering between genes and DCCs to discover toxicogenomic biomarkers and their regulatory DCCs from the fold change gene expression data using the proposed LPHVM.

Logistic Transformation of Fold Change Gene Expression Data

There are two ways to obtain robust estimates in presence of outlying observations (1) applying the robust methods (2) applying conventional methods on the modified dataset. The modification of the outlier contaminated dataset can be done deleting the outlying observations from the dataset or applying transformation on the dataset. Nonetheless, application of robust methods is complicated than using the conventional methods and deletion of outlying observations loses the information of the dataset. Hence, transformation is the better option for reducing outlier effects. Several authors (Box and Cox, 1964; Atkinson, 1982; Carroll, 1982) have been proved that transformation based robust methods outperform the conventional methods in reducing outlier effects. Thus, in this paper we consider logistic transformation for reducing outlier effects from the dataset. Before application of logistic transformation in the dataset we have taken average value ($\overline{Y}_{tl.}$) of the fold change gene expression (Y_{tlq}) over the samples. We denote this average value by $F\left(G_i, C_j\right)$ for the convenience of further use. In toxicogenomic data the expression profile of a subset of genes is highly correlated across a subset of conditions/treatments (Madeira and Oliveira, 2004; Bicego et al., 2010; Afshari et al., 2011). Interestingly, in the

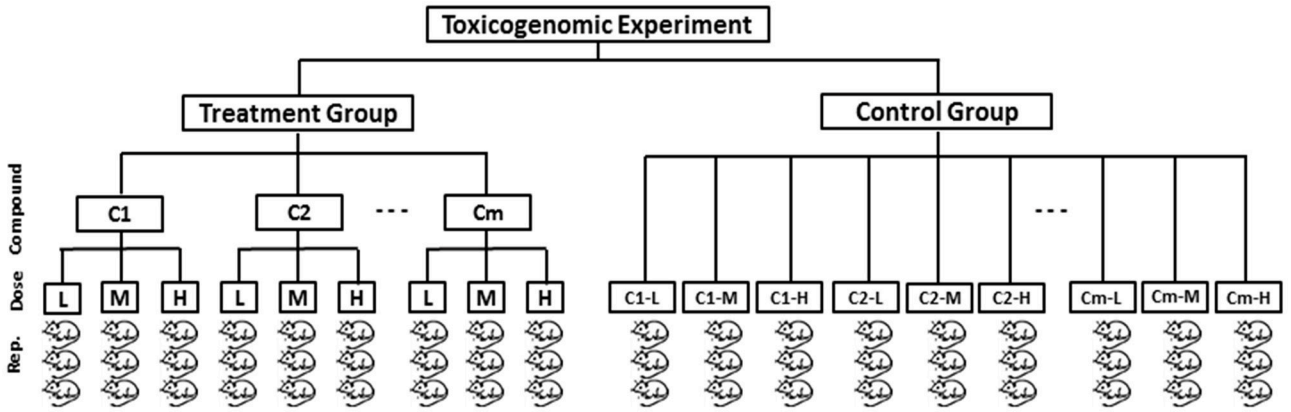

FIGURE 1 | A typical toxicogenomic experimental model for a single time point according to which gene expression data of the animal samples can be collected. In the figure there is a treatment group of animals and a control group of animals from which the fold change gene expression data can be obtained.

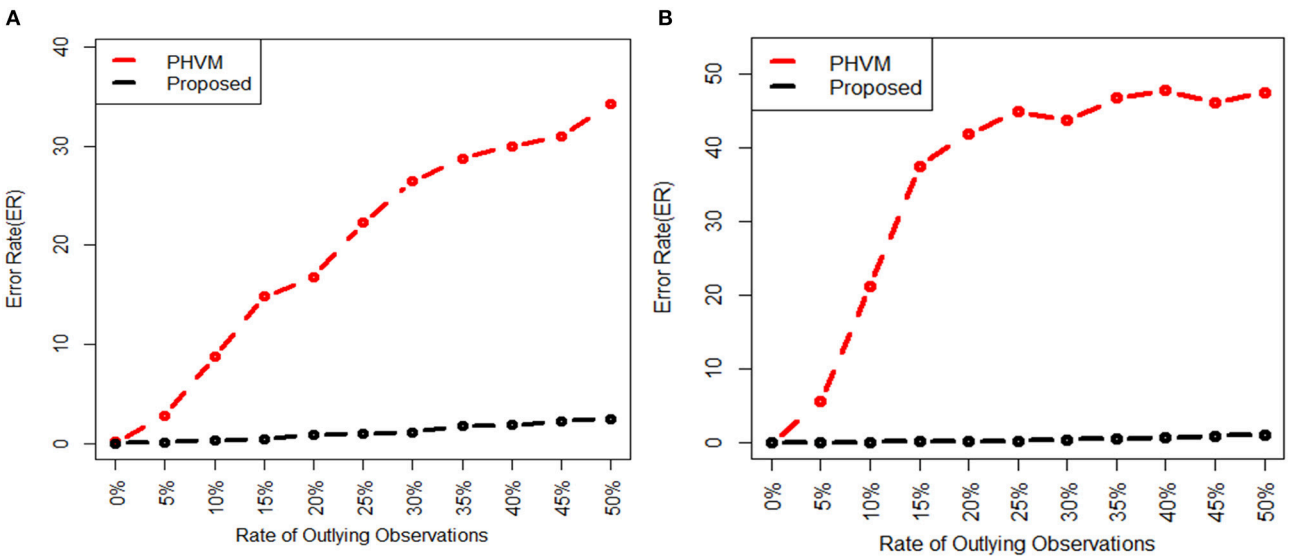

FIGURE 2 | Average gene and doses of chemical compounds co-clustering ER are plotted against the rate of outliers, when each of the data sets are simulated 100 times and outliers in the datasets are introduced using THCM. In the figure **(A)** for D_1 dataset and **(B)** for D_2 dataset.

gene expression or average fold change gene expression data there is a subset of genes which consists an upregulated and a downregulated clusters of genes which is highly correlated over a subset of DCCs. Therefore, we take absolute of the average fold change expression data to merge upregulated and downregulated clusters of genes into a single cluster/subset which are regulated by a subset of DCCs. Thereafter, the subset of genes forms a co-cluster with its regulatory subset of DCCs. Since in this study, we consider all the biomarker and non-biomarker genes (genes are not affected by DCCs) in a pathway, the non-biomarker genes make another co-cluster together with non-regulatory DCCs (which do not affect the expression patterns of the genes in a specific pathway). The term co-cluster refers to the clustering of correlated row (genes) and column (DCCs) simultaneously. Now we apply logistic transformation on the $(|F(G_i, C_j)|)$. If there are extreme values of $|F(G_i, C_j)|$ the logistic transformation bring them within the range of 0–1. The other transformation methods like Box-Cox family of power transformation returns unbounded value for the extreme one. The observed $n \times m$ (gene-DCCs) fold change gene expression data matrix consisting of $G = (G_1, G_2, \ldots, G_n)$ genes and $C = (C_1, C_2, \ldots, C_m)$ DCCs is transformed using logistic function

$$\#(G_i, C_j) = \left(\frac{1}{1 + \exp(-|F(G_i, C_j)|} \right) \times 100$$

Similar to other works (Joung et al., 2006; Bicego et al., 2010) we assume the transformed value $\#(G_i, C_j)$ as the count value for applying PHVM.

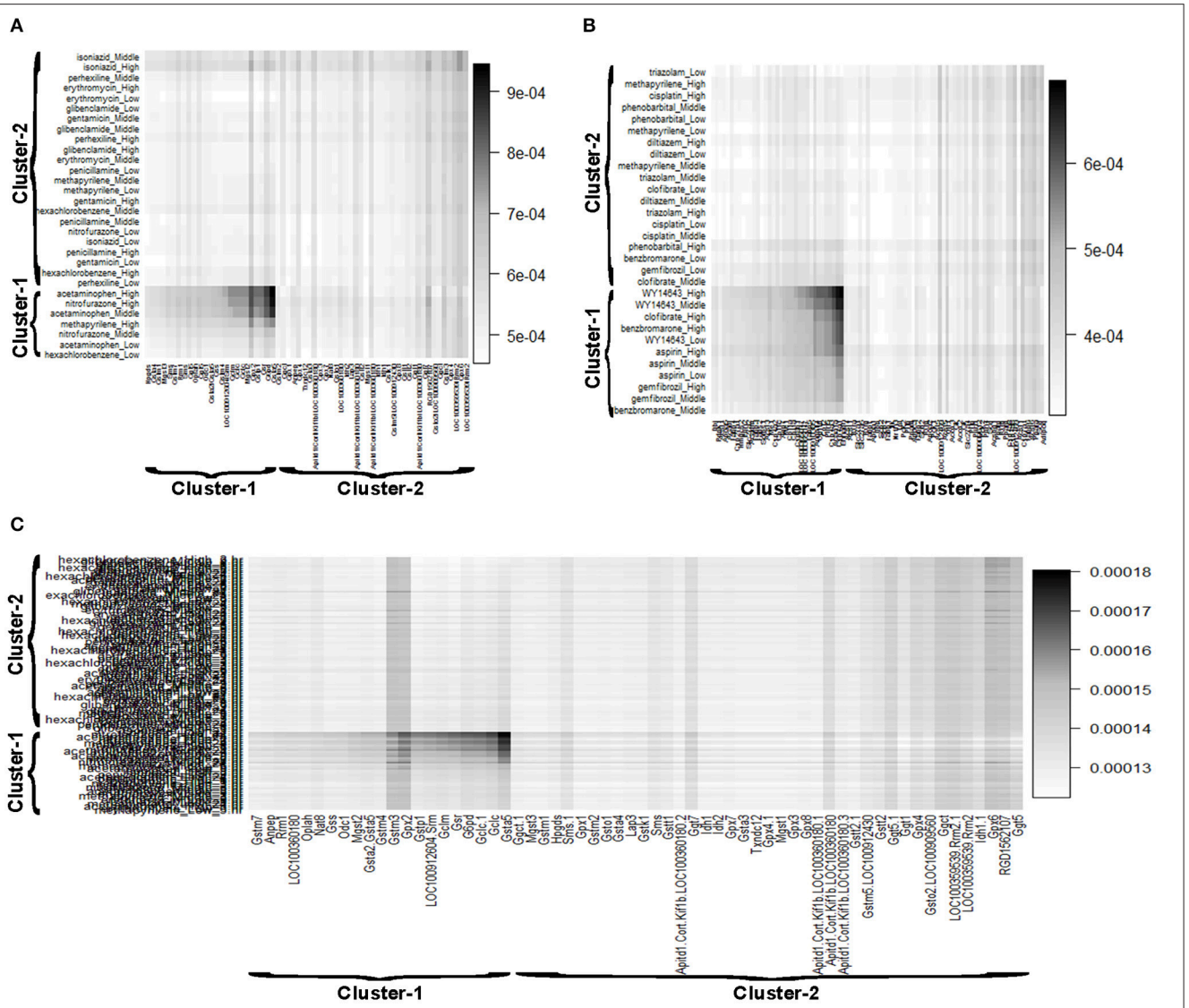

FIGURE 3 | Gene and doses of chemical compounds co-clustered view retrieved from the LPHVM generated gene and DCCs joint probability. **(A)** Represents glutathione metabolism pathway dataset at 24 h time point. **(B)** Represents PPAR signaling pathway dataset at 24 h time point. **(C)** Represents glutathione metabolism pathway dataset for all time points of Toxygates data.

Number of Co-clusters (k) Prediction

As we see from the previous section "logistic transformation of fold change gene expression data" in toxicogenomic dataset there are hidden patters or co-clusters between genes and DCCs. Thus the number of clusters in the DCCs is equal to the number of clusters in the genes. Before applying PHVM it is required to know the number of co-clusters in the dataset. Therefore, in this study, we consider gap statistic (Tibshirani et al., 2001) the most popular and reliable algorithm for predicting the number of co-clusters in the dataset. We use R function "fviz_nbclust" which required packages "factoextra" and "NbClust" (Malika et al., 2014) in order to predict number of co-clusters in the dataset via gap statistic. The detail algorithm of gap statistic is given in the **Supplementary Material**.

Robust Co-clustering Using Logistic Probabilistic Hidden Variable Model

In order to perform robust co-clustering between genes and DCCs we propose LPHVM approach. We define LPHVM as the application of PHVM on the count valued dataset which is obtained transforming absolute value of the fold change gene expression data by logistic transformation. For this standpoint, let us consider $n \times m$ gene-DCC count valued fold change gene expression data matrix consisting of $G = (G_1, G_2, \ldots, G_n)$ genes and $C = (C_1, C_2, \ldots, C_m)$ DCCs. LPHVM assumes that there prevail a certain number of unobserved hidden co-clusters or clusters underlying the gene-DCC count valued data matrix. We have estimated the number of co-clusters (k) in the dataset using gap statistic algorithm proposed by Tibshirani et al. (2001).

Introducing the hidden variable H = (H_1, H_2, ..., H_k; r = 1, 2, ..., k) the model quantifies the relationships $Pr(G_i|H_r)$, $Pr(C_j|H_r)$, and $Pr(G_i, C_j)$. The following are the probability definition and underlying assumptions of LPHVM accordingly: (1) $Pr(H_r)$ is the probability of the rth co-cluster/cluster and $\sum_{r=1}^{k} Pr(H_r) = 1$. (2) $Pr(G_i|H_r)$ is the probability of the ith gene over the rth co-cluster and $\forall H_r$; $\sum_{i=1}^{n} Pr(G_i|H_r) = 1$. (3) $Pr(C_j|H_r)$ is the probability of the jth DCC over the rth co-cluster and $\forall H_r$; $\sum_{j=1}^{m} Pr(C_j|H_r) = 1$. (4) $Pr(G_i, C_j)$ is the joint probability of the ith gene and the jth DCC and $\sum_{i=1}^{n}\sum_{j=1}^{m} Pr(G_i, C_j) = 1$. Based on these definition and assumptions we obtain the joint probability of the gene-DCC observed pair (G_i, C_j) considering hidden co-cluster H_r as follows:

$$Pr(G_i, C_j) = Pr(C_j)\, Pr(G_i|C_j)$$

Where,

$$Pr(G_i|C_j) = \sum_{r=1}^{k} Pr(G_i|H_r) Pr(H_r|C_j)$$

Applying Bayes' rule, the gene-DCC joint probability $Pr(G_i, C_j)$ can be written as

$$Pr(G_i, C_j) = \sum_{r=1}^{k} Pr(G_i|H_r)\, Pr(C_j|H_r)\, Pr(H_r)$$

So as to estimate the parameters of the model, we need to maximize the total likelihood of the observations:

$$L(G, C) = \sum_{i=1}^{n}\sum_{j=1}^{m} \#(G_i, C_j)\ log\ Pr(G_i, C_j)$$

We have applied the widely used Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for estimating the maximum likelihood parameters of the proposed model. The EM algorithm starts with a random set of initial parameter values and iterates both the expectation (E-step) and maximization (M-step) step alternatively until a certain convergence criteria is satisfied. For this study, we have taken the values of initial parameters from dirichlet distribution and the stopping condition for EM estimation was set to <0.00001 (difference between two log likelihood of successive EM iteration). The E and M-step for the total likelihood can be given as follows:

E-step:

$$Pr(H_r|G_i, C_j) = \frac{Pr(G_i|H_r)\, Pr(C_j|H_r)\, Pr(H_r)}{\sum_{r'=1}^{k} Pr(G_i|H_{r'})\, Pr(C_j|H_{r'})\, Pr(H_{r'})}$$

M-step:

$$Pr(H_r) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \#(G_i, C_j) Pr(H_r|G_i, C_j)}{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{r'=1}^{k} \#(G_i, C_j) Pr(H_{r'}|G_i, C_j)}$$

$$Pr(G_i|H_r) = \frac{\sum_{j=1}^{m} \#(G_i, C_j) Pr(H_r|G_i, C_j)}{\sum_{i'=1}^{n}\sum_{j=1}^{m} \#(G_{i'}, C_j) Pr(H_r|G_{i'}, C_j)}$$

$$Pr(C_j|H_r) = \frac{\sum_{i=1}^{n} \#(G_i, C_j) Pr(H_r|G_i, C_j)}{\sum_{i=1}^{n}\sum_{j'=1}^{m} \#(G_i, C_{j'}) Pr(H_r|G_i, C_{j'})}$$

Once the parameters $Pr(G_i|H_r)$ and $Pr(C_j|H_r)$ have been estimated the genes and DCCs are clustered independently and co-clustered simultaneously. The gene (G_i) and DCC (C_j) will belong to co-cluster r if

$$Pr(G_i|H_r) = argmax_{r'} Pr(G_i|H_{r'});\ i = 1, 2, \cdots, n;\ r = 1, 2, \ldots, k \text{ and}$$
$$Pr(C_j|H_r) = argmax_{r'} Pr(C_j|H_{r'});\ j = 1, 2, \ldots, m;\ r = 1, 2, \ldots, k$$

At the same time, if the gene (G_i) and the DCC (C_j) is grouped into a co-cluster (r) and this pair has the highest joint probability $Pr(G_i, C_j)$ in that co-cluster (**Figure 3**).

Extraction of Toxicogenomic Biomarker Genes and Their Regulatory Doses of Chemical Compounds

As described in section "logistic transformation of gene expression data" the biomarker genes form co-clusters with their respective regulatory DCCs. Additionally, the non-biomarker genes in a pathway form another co-cluster with non-regulatory DCCs. The LPHVM grouped the genes and DCCs simultaneously to their respective co-clusters. Zhu et al. (2005) has shown that the PHVM generated co-occurrence probabilities between correlated genes and chemical compounds which co-occur more frequently are higher than others. Biological relationship among these correlated genes and chemical compounds is also stronger. Therefore, we ranked the co-clusters based on the average LPHVM generated joint probability $(Pr(G_i, C_j))$ of gene-DCC within the co-clusters. The co-cluster having largest average joint probability contains most important biomarker genes and their regulatory DCCs and so on. The non-biomarker genes and non-regulatory DCCs in a dataset of a particular pathway are filtered in a co-cluster by LPHVM which have the smallest average joint probability. Except this co-cluster (co-cluster having smallest average joint probability) others are the co-clusters of biomarker genes and their regulatory DCCs and we define these co-clusters as biomarker co-clusters. We extract the toxicogenomic biomarker genes and their regulatory DCCs from these biomarker co-clusters.

Up/Down-Regulated Biomarker Genes and Ranking of Doses of Chemical Compounds

The biomarker co-clusters consisting of biomarker genes and their regulatory DCCs are separated from the whole gene-DCC fold change data matrix which is discussed in the previous section. Within this co-clustering matrix a subset of biomarker genes may be upregulated corresponding to a subset of DCCs or downregulated corresponding to another subset of DCCs. These can be observed from the average fold change value ($\overline{Y}_{tl.}$) of the co-clustering matrix. For example, a biomarker is define as up or down-regulated gene corresponding to the l^{th} dose level of the t^{th} chemical compounds if $\overline{Y}_{tl.} > 0$ or $\overline{Y}_{tl.} < 0$. Then this dose of chemical compound is said to be a regulatory DCC. Furthermore, for ranking the biomarker

gene regulatory DCCs and their relationships with biomarker genes we have separated a sub matrix of biomarker genes and their regulatory DCCs (biomarker co-clusters) from the LPHVM generated gene-DCC joint probability $Pr\left(G_i, C_j\right)$ matrix. The biomarker gene regulatory DCCs are ranked according to their average joint probability value over all biomarkers. We also rank the relationships among biomarker genes and their regulatory DCCs based on their joint probability. The ranking is made considering the formula:

$$\left(\frac{Z_{j/i,j}}{\max\left(Z_{j/i,j}\right)}\right) \times 100$$

where Z_j is the average joint probability of a DCC over the biomarkers or $Z_{i,j}$ is the joint probability of gene G_i and DCC C_j within the biomarker co-clusters.

Robustness of the Proposed Algorithm

We investigate the robustness of the proposed (LPHVM) algorithm and conventional PHVM using simulated datasets in absence and presence of outliers in the dataset based on the co-clustering /clustering error rate (ER). The genes and DCCs which are considered in one co-cluster/cluster in the simulated data are incorrectly assigned in another co-cluster/cluster by the PHVM or LPHVM is considered as the miss co-clustered/clustered observations. The ER is the percentage of miss co-clustered/clustered observations which is calculated as:

$$\left(\frac{tolal\ miss\ co-clustered/clustered\ observations}{Total\ observations}\right) \times 100$$

Computational Steps of LPHVM at a Glance

For detecting the toxicogenomic biomarker genes and their regulatory DCCs from the pathway level toxicogenomic dataset using LPHVM the following steps are to be considered for desired outputs:

Step 1: Obtain gene expression data of treatment and control group of animals from the toxicogenomic experiment (**Figure 1**). Thereafter, compute fold change gene expression data using Equation (1) and then make it absolute.

Step 2: Apply logistic transformation on the dataset obtain from step 1 and assume the transformed value as count value.

Step 3: Estimate the number of co-clusters in the dataset which is obtained from step 2.

Step 4: Obtain robust co-clusters applying PHVM on the dataset obtained from step 2 using the number of co-clusters which we get from step 3.

Step 5: Calculate average joint probability of gene-DCC within the co-clusters and ranked them.

Step 6: Separate the co-clusters of biomarker genes and their regulatory DCCs from the co-cluster which have smallest average joint probability of gene-DCC.

Step 7: The genes and DCCs in the separated co-clusters which we get from step 6 are the toxicogenomic biomarkers and their regulatory DCCs.

Step 8: A biomarker gene obtains from step 7 may be upregulated corresponding to a DCC or downregulated corresponding to another DCC. A biomarker gene is said to be a up or down-regulated if its average fold change value corresponding to the lth dose level of the tth chemical compound is $\overline{Y}_{tl.} > 0$ or $\overline{Y}_{tl.} < 0$.

Simulated Datasets

To investigate the performance of the proposed LPHVM algorithm over the conventional PHVM we have simulated two sets of pathway level fold change gene expression data $D_1(n = 50 \times m = 30)$ and $D_2(n = 50 \times m = 60)$ imitating the toxicogenomic experiment given in **Figure 1**. Alongside these a pathway level dataset considering all time points of toxygates data are analyzed in the real data section. According to this experiment the fold change gene expression data (Y_{tlq}) have been generated using the following model:

$$Y_{tlq} = \begin{array}{lccc} & \textbf{DCCs group-1} & \textbf{DCCs group-2} & \textbf{DCCs group-3} & \\ \text{Gene group-11} & +F11 & 0 & 0 & \\ \text{Gene group-12} & -F12 & 0 & 0 & \\ \text{Gene group-21} & 0 & +F21 & 0 & +N(0,\sigma^2) \\ \text{Gene group-22} & 0 & -F22 & 0 & \\ \text{Gene group-3} & 0 & 0 & 0 & (2) \end{array}$$

In the above model, $+F11$and $+F21$ represent the fold change expression values for upregulated genes under the DCCs group 1 and 2, respectively. Similarly, $-F12$, and $-F22$ represent the fold change expression values for the downregulated genes under the DCCs group 1 and 2, respectively. The 0s represent there is no compound effects on the respective gene group and $N(0, \sigma^2)$ represents the random error term generated from normal distribution with mean 0 and variance σ^2. Now if we take absolute value of the fold change gene expression data generated from the above data generating model (2), the fold change gene expression data $+F11$ and $-F12$ will merge into a single gene group-1 and make a co-cluster with their correlated DCCs group-1. Accordingly, $+F21$ and $-F22$ will merge into a single gene group-2 and make a co-cluster with their correlated DCCs group-2. The rest of the genes which are not regulated by any DCCs make a gene group-3 and the DCCs that do not regulate the expression pattern of genes make a DCCs group-3. The gene group-3 and DCCs group-3 together will make another co-cluster. These co-clusters can be retrieved by the LPHVM. In the simulated datasets n represents the number of genes (G_i; $i = 1, 2, \ldots, n$) and m represents the number of DCCs (C_j; $j = 1, 2, \ldots, m$). The data generation procedures for D_1 and D_2 datasets are given in the **Supplementary Material.**

Real Datasets

Several studies proved that molecular network or pathway based analysis improved the predictive power of gene expression data (Yildirimman et al., 2011; Hofree et al., 2013). Hardt et al. (2016) also analyzed the pathway level data from *in vitro* and *in vivo* experiment of human and rat model. Presently, pathway based analysis in cancer research has also advanced promptly since pathway level analysis able to produce more stable biomarkers (Kim, 2017). Since performance of any method cannot be measured without known dataset. Besides the simulation study, to investigate the performance of the proposed method compare to other existing methods we use two known datasets of glutathione metabolism and PPAR signaling pathways. The fold change expression data of the TGP experiment for glutathione metabolism and PPAR signaling pathway for some selected DCCs of the respective pathway at 24 h time point have been downloaded from toxygates (https://toxygates.nibiohn.go.jp/toxygates/#columns). Because the compounds' toxicity at 24 h time point is more visible compare to other time points (Nyström-Persson et al., 2013). Alongside these a dataset consisting of glutathione metabolism pathway genes and glutathione depleting and non-glutathione depleting compounds (Nyström-Persson et al., 2013) for all time points is also considered for analysis to know about the toxicity of DCCs in other time points.

RESULTS

Simulation Study

We investigate the performance of our proposed method (LPHVM) by comparing it with the conventional PHVM using simulated datasets D_1 and D_2 in absence and presence of outlying observations for robust co-clustering between genes and DCCs to discover biomarker genes and their regulatory DCCs. The number of co-clusters/clusters for both of the simulated datasets is estimated as 3 via gap statistic as per the datasets are simulated (**Figure S1**). For calculating average co-clustering and clustering ER we have simulated each of the datasets 100 times. Every time of data simulation outliers are introduced in the dataset using the data contamination methods THCM and ICM at the same time ER are calculated for PHVM and LPHVM applying these methods on the datasets. The description of the data contamination by outliers, THCM and ICM are given in the **Supplementary Material**. Here it should be mentioned that in the case of THCM we have contaminated the simulated datasets by 5–50% rate of outliers. Similarly, in the case of ICM we have considered the range of probability of at least one component of the dataset is to be contaminated is 0.14–0.60 for D_1 dataset and 0.165–0.5962 for D_2 dataset. **Figure 2** visualizes the average co-clustering ER between genes and DCCs for datasets D_1 and D_2 in absence and presence of outliers when the datasets are contaminated by outliers using the THCM. The **Table 1** shows the average co-clustering ER between genes and DCCs in absence and presence of outliers for the simulated datasets D_1 and D_2 when the datasets are contaminated by outliers using ICM. **Figure S2** and **Table S1** in the Supplementary Material show the average clustering ER for gene and DCCs. It is observed from the mentioned figures and tables that in absence of outlier both of the proposed LPHVM and conventional PHVM approaches produce 0 ER. However, in presence of outlaying observations in the datasets the proposed approach produce far smaller ER than the conventional approach for both of the data contamination methods (THCM and ICM). The simulated data structure, structure of the data when row (gene) and column (DCCs) entities are randomly allocated and proposed method recovered structure of the data are given in the Supplementary Material (**Figures S3**, **S4)** for the datasets D_1 and D_2. From these figures it is observed that the proposed algorithm is efficient for co-clustering between genes and DCCs of the pathway level fold change gene expression data. **Figure S3C** represents the dataset D_1 where all the genes and DCCs are grouped into three co-clusters (co-clusters 1, 2, and 3) and within co-cluster average joint probability of gene-DCC are given in **Table 3**. From where it is found that co-cluster-1 produces the smallest average joint probability of gene-DCC. Therefore, co-cluster 2 and 3 are the co-cluster of biomarker genes and their regulatory DCCs for the dataset D_1. Similarly, for D_2 dataset co-cluster-3 produces the smallest average joint probability of gene-DCC (**Table 3**). Thus, co-cluster 1 and 2 are the biomarker co-clusters consisting of biomarker genes and their regulatory DCCs. The biomarker genes and their regulatory DCCs that we get from the biomarker co-clusters of the simulated datasets are given in the **Table S9**. Ranking of the biomarker regulatory DCCs are performed based on the biomarker gene-DCC joint probability matrix of biomarker co-clusters following the raking method described in sub section (Up/Down-regulated Biomarker Genes and Ranking of Doses of Chemical Compounds). The results are given in

TABLE 1 | Average values of the gene and doses of chemical compounds co-clustering ER for the simulated datasets D_1 and D_2 when each of the datasets are simulated 100 times and contaminated by outlier using ICM.

Dataset	Method	Probability of at least one component in the dataset to be contaminated (ε)						
		0.00	**0.14**	**0.26**	**0.36**	**0.45**	**0.53**	**0.60**
D_1	PHVM	0.175	24.675	28.950	32.912	33.500	35.125	38.487
	Proposed	0.025	0.387	0.612	0.725	1.0	1.862	2.500
		0.00	**0.165**	**0.3031**	**0.4187**	**0.5154**	**0.5962**	
D_2	PHVM	0.00	25.390	26.563	29.554	32.172	39.754	
	Proposed	0.00	0.163	0.945	1.481	1.600	2.072	

TABLE 2 | Upregulated and downregulated biomarker genes and their regulatory doses of chemical compounds for real life datasets.

Dataset	Biomarker genes	Biomarker gene regulatory DCCs
Glutathione metabolism pathway	Gsta4, Gstm1, Sms, Rrm1, Odc1, Gsta2/Gsta5, Gss, Gstm4, LOC100912604/Srm, Gclm, Gclc, Mgst2, Gstp1, Gsr, Gpx2, G6pd, Gsta5, Hpgds, Mgst3, Gstm7, Oplah, Ggt5	hexachlorobenzene_Low acetaminophen_Low nitrofurazone_Middle methapyrilene_High acetaminophen_Middle nitrofurazone_High acetaminophen_High
PPAR signaling pathway	Dbi, Acsl1, Acadl, Hmgcs2, Plin2, Slc27a2, Acadm, Fads2, Fabp3, Me1, Sorbs1, Acsl3, Cyp4a2, Aqp7, Cpt1a, Cyp8b1, OC100365047, LOC100910385, Angptl4, Cpt1b, Cpt2, Plin5, Cyp4a3, Acaa1a, Cyp4a1, Ehhadh, Pdpk1, Apoa5, Fabp4, Cyp27a1, Cpt1c, Fabp5	benzbromarone_Middle gemfibrozil_Middle gemfibrozil_High aspirin_Low aspirin_Middle aspirin_High WY14643_Low benzbromarone_High clofibrate_High WY14643_Middle WY14643_High

TABLE 3 | Average values of the Gene and DCCs joint probabilities within the co-clusters generated by the proposed LPHVM algorithm for the simulated and real life datasets.

Dataset	Co-cluster-1	Co-cluster-2	Cocluster-3
D_1	0.0006095721	0.0010120670	0.0010117088
D_2	0.0005162618	0.0005163485	0.0003147069
Glutathione metabolism pathway	0.0006196723	0.0005331547	
PPAR signaling pathway	0.0004471087	0.0003704091	

TABLE 4 | Biomarker genes regulatory doses of chemical compounds ranking for real datasets (glutathione metabolism and PPAR signaling pathway).

Dataset	Doses of chemical compounds	Percent score
Glutathione metabolism pathway	acetaminophen_High	100.00
	nitrofurazone_High	99.59
	acetaminophen_Middle	95.98
	methapyrilene_High	88.66
	nitrofurazone_Middle	82.24
	acetaminophen_Low	77.84
	hexachlorobenzene_Low	74.57
PPAR signaling pathway	WY14643_High	100.00
	WY14643_Middle	97.59
	clofibrate_High	93.25
	aspirin_High	92.91
	benzbromarone_High	92.25
	WY14643_Low	91.19
	aspirin_Middle	87.93
	aspirin_Low	86.41
	gemfibrozil_High	85.51
	gemfibrozil_Middle	84.52
	benzbromarone_Middle	79.07

the Supplementary Material (**Table S10**) for both D_1 and D_2 datasets.

Analysis of Glutathione Metabolism Pathway Data

Reactive oxygen species (ROS) are produced by living organisms as a normal product as a result of normal cellular metabolism. However, in presence of environmental pollutants or toxic chemical the production of ROS increased dramatically. It is highly reactive molecules and can damage cell structures such as carbohydrates, nucleic acids, lipids, and proteins and alter their functions. In the liver, glutathione is an important antioxidant; a major detoxification player which scavenges ROS. Thus imbalance in the abundance of ROS and glutathione/antioxidant in favor of ROS in the liver in presence of toxic chemicals/drugs causes' drug induced liver injury. Subsequently, gene expression changes occur simultaneously in response to the glutathione depletion or after the glutathione depletion (Gao et al., 2010; Birben et al., 2012; Nyström-Persson et al., 2013). In order to identify glutathione depletion related biomarker genes and their regulatory DCCs as well as to investigate the performance of the proposed LPHVM approach we use known fold change gene expression dataset of glutathione metabolism pathway. The fold change gene expression dataset consists 62 glutathione metabolism pathway genes, three glutathione depleting compounds (acetaminophen, methapyrilene, and nitrofurazone) and seven non-glutathione depleting compounds (erythromycin, hexachlorobenzene, isoniazid, gentamicin, glibenclamide, penicillamine, and perhexilline) (Nyström-Persson et al., 2013) along with the dose levels (low, middle, and high) for 24 h time point. The number of co-clusters which is required in applying LPHVM for this dataset is estimated as 2 (**Figure S1**) via gap statistic. **Figure 3A** shows actual co-clusters in the glutathione metabolism pathway dataset. The genes and DCCs in the co-clusters are given in the **Table S2**. The average joint probabilities of gene-DCC within the co-clusters are 0.0006196723 and 0.0005331547 (**Table 3**), respectively for co-cluster-1 and co-cluster-2. Thus, Co-cluster-1 is the co-cluster of biomarker genes and glutathione depleting DCCs as it produces highest average joint probability. The biomarker genes and their regulatory DCCs in co-cluster-1 are given in **Table 2**. Additionally, the upregulated and downregulated biomarker genes corresponding to their regulatory DCCs are presented in the **Figure S7A**. For the same dataset the clustering results (heatmap) produced by toxygates are given in **Figure S5** where glibenclamide-low, perhexilline-low, and hexachlorobenzene-medium dose level are incorrectly co-clustered whereas only hexachlorobenzene-low dose is incorrectly co-clustered by the proposed LPHVM approach according to Nyström-Persson et al. (2013). The biomarker genes in co-cluster-1 are functionally annotated by the online database DAVID (Huang da et al., 2009) and the results are given in the **Tables S5, S6**. The results show that the biomarker genes are significant in different biological

TABLE 5 | Top 20 (ranked) biomarker gene and their regulatory doses of chemical compound relationships for glutathione metabolism pathway and PPAR signaling pathway datasets.

Glutathione metabolism pathway			PPAR signaling pathway		
Chemical compound and dose combination	**Biomarker gene**	**Ranking score**	**Chemical compound and dose combination**	**Biomarker gene**	**Ranking score**
acetaminophen_High	Gsta5	100.00	WY14643_High	Ehhadh	100.00
nitrofurazone_High	Gsta5	96.26	WY14643_High	Cyp4a1	97.29
acetaminophen_Middle	Gsta5	91.69	WY14643_Middle	Ehhadh	95.32
acetaminophen_High	G6pd	90.85	WY14643_Middle	Cyp4a1	93.17
acetaminophen_High	Gpx2	89.67	WY14643_High	Acaa1a	92.41
nitrofurazone_High	G6pd	89.48	clofibrate_High	Ehhadh	88.93
nitrofurazone_High	Gpx2	89.29	WY14643_Middle	Acaa1a	88.47
acetaminophen_Middle	Gpx2	86.05	clofibrate_High	Cyp4a1	87.34
acetaminophen_Middle	G6pd	85.91	benzbromarone_High	Ehhadh	87.04
acetaminophen_High	Gsr	85.19	WY14643_High	Cyp4a3	86.68
acetaminophen_High	Gstp1	83.54	WY14643_Low	Ehhadh	86.65
nitrofurazone_High	Gsr	83.25	WY14643_High	Plin5	85.99
nitrofurazone_High	Gstp1	81.53	benzbromarone_High	Cyp4a1	85.67
acetaminophen_High	Mgst2	80.46	WY14643_Low	Cyp4a1	85.17
acetaminophen_High	Gclc	80.38	WY14643_High	Cpt2	84.46
methapyrilene_High	Gsta5	80.23	WY14643_High	Cpt1b	84.45
acetaminophen_Middle	Gsr	79.71	WY14643_High	Angptl4	83.99
acetaminophen_High	Gclm	79.56	aspirin_High	Ehhadh	83.60
methapyrilene_High	Gpx2	79.47	WY14643_Middle	Cyp4a3	83.54
nitrofurazone_High	Gclc	78.93	aspirin_High	Cyp4a1	83.10

functions or processes including glutathione metabolism pathway. Ranking of biomarker gene regulatory DCCs and top 20 gene-DCCs relationship along with their ranking score for glutathione metabolism pathway dataset are given in **Tables 4**, **5**. From the tables it is observed that acetaminophen_High, nitrofurazone_High, and acetaminophen_Middle dose etc. are the most important glutathione depleting compounds and Gsta5, G6pd, Gpx2, Gsr, Mgst2, Gstp1, Gclc etc. are the most important biomarker genes. The detail ranked relationships results are given in **Table S12**. Besides this we have analyzed the same dataset considering all time points (3, 6, 9, and 24 h) by LPHVM to know about toxicity mechanism of the glutathione depleting compounds in other time points. The co-clusters produced by LPHVM are given in **Figure 3C**. The detail analyzed results of this dataset are given in **Tables S4**, **S11**. The proposed LPHVM identified 25 genes for the dataset at 24 h time points and 21 genes for the dataset where all time points are considered as biomarker in the glutathione metabolism pathway among which 18 are common.

Analysis of PPAR Signaling Pathway Data

Peroxisome proliferator-activated receptors (PPARs) $PPAR_{\alpha}$, $PPAR_{\beta/\delta}$, and $PPAR_{\gamma}$ are transcription factors which are activated by ligand/drug. They regulate the expression of target genes in response to endogenous and exogenous ligands/chemicals. The PPAR ligands may produce toxicity via receptor-dependent and/or off-target-mediated mechanism(s) (Peraza et al., 2006). To discover PPARs regulated biomarker genes and their regulatory DCCs as well as to investigate the performance of the proposed LPHVM approach we consider known dataset consisting 88 PPAR signaling pathway genes and PPARs related gene regulatory compounds (WY-14643, clofibrate, gemfibrozil, benzbromarone, and aspirin) (Kiyosawa et al., 2006) and some other randomly selected compounds (cisplatin, diltiazem, methapyrilene, phenobarbital, and triazolam) along with their dose levels low, middle and high. The number of hidden co-clusters for this dataset is 2 estimated via gap statistic (**Figure S1**). The LPHVM generates co-clusters of the PPAR signaling pathway dataset which is shown in **Figure 3B**. The average joint probabilities of gene-DCC within co-clusters are 0.0004471087 and 0.0003704091 where co-cluster-1 has the larger value than the co-cluster-2. Therefore, co-cluster-1 is the biomarker co-cluster of biomarker genes and their regulatory DCCs. The non-regulated genes and non-regulatory DCCs consist in co-cluster-2. The detail co-clustering results are given in the **Table S3**. The biomarker genes and their regulatory DCCs in co-cluster-1 are given in **Table 2**. Additionally, up/down-regulated biomarker genes corresponding to their regulatory DCCs are depicted in the **Figure S7B** For the same dataset the toxygates co-clustering result using HC given in **Figure S6** which shows that acetaminophen-middle and methapyrilene-low are incorrectly co-clustered whereas our proposed method properly co-cluster the DCCs (**Table 2**) according to the statement of Kiyosawa et al. (2006). Biomarker genes in co-cluster-1 are functionally

annotated via DAVID the results are given in the **Tables S7**, **S8**. WY14643-High, WY14643-Middle and clofibrate-High are the top most DCCs for regulating PPARs related biomarker genes for detail see **Table 4**. Top 20 (ranked) relationships between biomarker genes and their regulatory DCCs are given in **Table 5** from where it is observed that Ehhadh, Cyp4a1, Acaa1a, Plin5 etc. are the most important biomarker genes and WY14643_High, clofibrate_High, benzbromarone_High, aspirin_High etc. are their important regulatory DCCs in PPAR signaling pathway. The detail results of these relationships are given in the **Table S13**.

DISCUSSION AND CONCLUSIONS

Identification of biomarker genes and their regulatory DCCs is one of the most important tasks in the toxicogenomics studies as well as in drug design and development as mentioned before. In this article, we have proposed a robust co-clustering approach based on logistic probabilistic hidden variable model (LPHVM) to detect important biomarker genes and their regulatory DCCs. The proposed LPHVM approach is robust against outlying gene expressions and more flexible and effective than the application of one-way classical clustering approaches (e.g., k-means, fuzzy, HC, etc.) for co-clustering. The proposed method produces robust results by using the logistic transformation of fold-change gene expression data into the conventional PHVM approach. The logistic transformation reduces unusual/outlying observations into the reasonable space without changing the original hidden patterns of genes and DCCs in the dataset. Thus the proposed LPHVM approach produces robust results.

We investigated the performance of the proposed LPHVM method in a comparison with the traditional PHVM and Toxygates online computational platform using simulated and real life TGP gene expression data, respectively. The simulation results showed that the proposed method improves the performance over the conventional PHVM in presence of outlying observations; otherwise, they perform equally. We also demonstrated the performance of the proposed method in a comparison with the online computational platform "Toxygates" using the real life pathway based fold change gene expression datasets collected from the "Toxygates" database. We observed that three DCCs (glibenclamide-low, perhexilline-low, and hexachlorobenzene-medium) for glutathione metabolism pathway dataset as well as two DCCs (acetaminophen-medium and methapyrilene-low) for PPAR signaling pathway dataset were incorrectly co-clustered by the Toxygates online platform, while only one DCC (hexachlorobenzene-low) for glutathione metabolism pathway was incorrectly co-clustered by the proposed LPHVM approach. Our findings from the real life data analysis are also supported by the other findings in the literature (Kiyosawa et al., 2006; Nyström-Persson et al., 2013). Thus the proposed LPHVM outperform over the classical PHVM and "Toxygates" online coputational platform to detect toxicogenomic biomarkers and their regulatory DCCs.

DATA AVAILABILITY

The demo data and R-code are provided at http://www.bbcba.org/softwares/rCoClust.zip.

AUTHOR CONTRIBUTIONS

MH and MM worked together to develop the algorithm. MH analyzed the data and drafted the manuscript. MM coordinated and supervised the project. MMR, AB, and MR attended at the meeting regarding this article as well as read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

We are grateful to the authority of the bioinformatics laboratory, department of statistics, University of Rajshahi, Rajshahi, Bangladesh for their co-operation and giving chance to do this research work in their laboratory.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00516/full#supplementary-material

REFERENCES

Afshari, C. A., Hamadeh, H. K., and Bushel, P. R. (2011). The evolution of bioinformatics in toxicology: advancing toxicogenomics. *Toxicol. Sci.* 120, S225–S237. doi: 10.1093/toxsci/kfq373

Agostinelli, C. C., Leung, A., Yohai, V. J., and Zamar, R. H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 24, 441–461. doi: 10.1007/s11749-015-0450-6

Alqallaf, F., Van, A. S., Yohai, V., and Zamar, R. (2009). Propagation of outliers in multivariate data. *Ann. Stat.* 37, 311–331. doi: 10.1214/07-AOS588

Atkinson, A. C. (1982). Regression diagnostics, transformation and constructed variables. *J. R. Stat. Soc. Ser. B* 44, 1–36.

Bicego, M., Lovato, P., Ferrarini, A., and Delledonne, M. (2010). "Biclustering of expression microarray data with topic models," in *International Conference on Pattern Recognition* (Washington, DC: IEEE Computer Society), 2728–2731. doi: 10.1109/ICPR.2010.668

Birben, E., Sahiner, U. M., Cansin Sackesen, C., Serpil Erzurum, S., and Omer Kalayci, O. (2012). Oxidative stress and antioxidant defense. *World Allergy Organ. J.* 5, 9–19. doi: 10.1097/WOX.0b013e3182439613

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26, 211–252.

Carroll, R. J. (1982). Two examples of transformations when there are possible outliers. *Appl. Stat.* 31, 149–152. doi: 10.2307/2347978

Chen, M., Zhang, M., Borlak, J., and Tong, W. (2012). A decade of toxicogenomic research and its contribution to toxicological science. *Toxicol. Sci.* 130, 217–228. doi: 10.1093/toxsci/kfs223

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38.

Fielden, M. R., Brennan, R., and Gollub, J. (2007). A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by non genotoxic chemicals. *Toxicol. Sci.* 99, 90–100. doi: 10.1093/toxsci/kfm156

Gao, W., Mizukawa, Y., Nakatsu, N., Minowa, Y., Yamada, H., Ohno, Y., et al. (2010). Mechanism-based biomarker gene sets for glutathione depletion-r elated hepatotoxicity in rats. *Toxicol. Appl. Pharmacol.* 247, 211–221. doi: 10.1016/j.taap.2010.06.015

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. A. (2010). A review of robust clustering methods. *Adv. Data Anal. Classif.* 4, 89–109. doi: 10.1007/s11634-010-0064-5

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biomatrics* 62, 10–18. doi: 10.1111/j.1541-0420.2005.00397.x

Hardt, C., Beber, M. E., Rasche, A., Kamburov, A., Hebels, D. G., Kleinjans, J. C., et al. (2016). ToxDB: pathway-level interpretation of drug-treatment data. *Database* 2016:baw052. doi: 10.1093/database/baw052

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* 42, 177–196. doi: 10.1023/A:1007617005950

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Huang da, W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Igarashi, Y., Nakatsu, N., Yamashita, T., Ono, A., Ohno, Y., Urushidani, T., et al. (2015). Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* 43(Database issue), D921–D927. doi: 10.1093/nar/gku955

Joung, J. G., Shin, D., Seong, R. H., and Zhang, B. T. (2006). Identification of regulatory modules by co-clustering latent variable models: stem cell differentiation. *Bioinformatics* 22, 2005–2011. doi: 10.1093/bioinformatics/btl343

Kim, S. (2017). Identifying dynamic pathway interactions based on clinical information. *Comput. Biol. Chem.* 68, 260–265. doi: 10.1016/j.compbiolchem.2017.04.009

Kiyosawa, N., Shiwaku, K., Hirode, M., Omura, K., Uehara, T., Shimizu, T., et al. (2006). Utilization of a one-dimensional score for surveying chemical-induced changes in expression levels of multiple biomarker gene sets using a large-scale toxicogenomics database. *J. Toxicol. Sci.* 31, 433–448. doi: 10.2131/jts.31.433

Madeira, S. C., and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEE Trans. Comput. Biol. Bioinform.* 1, 24–45. doi: 10.1109/TCBB.2004.2

Malika, C., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data Set. *J. Stat. Softw.* 61, 1–36. doi: 10.18637/jss.v061.i06

NRC (2007). *National Research Council of the National Academies: Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment*. Washington, DC: National Academies Press.

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, 153–159. doi: 10.1002/(SICI)1098-2744(199903)24:33.0.CO;2-P

Nyström-Persson, J., Igarashi, Y., Ito, M., Morita, M., Nakatsu, N., Yamada, H., et al. (2013). Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform. *Bioinformatics* 23, 3080–3086. doi: 10.1093/bioinformatics/btt531

Nyström-Persson, J., Natsume-Kitatani, Y., Igarashi, Y., Satoh, D., and Mizuguchi, K. (2017). Interactive toxicogenomics: gene set discovery, clustering and analysis in Toxygates. *Sci Rep*. 7:1390. doi: 10.1038/s41598-017-01500-1

Peraza, M. A., Burdick, A. D., Marin, H. E., Gonzalez, F. J., and Peters, J. M. (2006). The Toxicology of ligands for peroxisome proliferator-activated receptors (PPAR). *Toxicol. Sci.* 90, 269–295. doi: 10.1093/toxsci/kfj062

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc.* 63, 411–423. doi: 10.1111/1467-9868.00293

Uehara, T. (2010). The Japanese toxicogenomics project: application of toxicogenomics. *Mol. Nutr. Food Res.* 54, 218–227. doi: 10.1002/mnfr.200900169

Uehara, T., Hirode, M., Ono, A., Kiyosawa, N., Omura, K., Shimizu, T., et al. (2008). A toxicogenomics approach for early assessment of potential non-genotoxic hepatocarcinogenicity of chemicals in rats. *Toxicology* 250, 15–26. doi: 10.1016/j.tox.2008.05.013

Upton, G. J. G., Sanchez-Graillet, O., Rowsell, J., Arteaga-Salas, J. M., Graham, N. S., Stalteri, M. A., et al. (2009). On the causes of outliers in Affy metrix GeneChip data. *Brief. Funct. Genomic Proteomic.* 8, 119–212. doi: 10.1093/bfgp/elp027

Yildirimman, R., Brolén, G., Vilardell, M., Eriksson, G., Synnergren, J., Gmuender, H., et al. (2011). Human embryonic stem cell derived hepatocyte-like cells as a tool for *in vitro* hazard assessment of chemical carcinogenicity. *Toxicol. Sci.* 124, 278–290. doi: 10.1093/toxsci/kfr225

Zhu, S., Okuno, Y., Tsujimoto, G., and Mamitsuka, H. (2005). A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics* 21(Suppl. 2), 245–251. doi: 10.1093/bioinformatics/bti1141

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.