International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 18, 2019, pp 99-114
© 2019 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

Clustering Toxicogenomic Data using Probabilistic Latent Variable Model

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

[1]Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur-1706, Bangladesh

[2]Bioinformatics Lab., Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh

[3] Department of Veterinary and Animal Sciences, University of Rajshahi, Rajshahi 6205, Bangladesh

[*]Correspondence should be addressed to Mohammad Nazmol Hasan (nazmol.stat.bioin@bsmrau.edu.bd)

[Received June 3, 2019; Revised September 1, 2019; Accepted November 5, 2019]

Abstract

Toxicogenomics along with suitable statistical techniques can assess toxicity of chemical compounds (CCs) that is an effective approach to safety assessment of CCs or drugs. Therefore, grouping of CCs as well as genes is essential to understand the pattern of CCs and genes regarding toxicity issue. Probabilistic Latent Variable Model (pLVM) can group or cluster the variables and sampling elements simultaneously. That is the uniqueness of pLVM from the traditional statistical methods like K-means, hierarchical clustering, model based clustering etc. However, the number of latent class selection is a challenging job for valid result from pLVM. Because, pLVM clusters CCs and genes concurrently based on the latent class in the toxicogenomic dataset. In this study, we have used pLVM along with latent class selection methods for clustering CCs and genes. From the pLVM generated clusters we have discovered the toxic CCs and toxicogenomic biomarkers. The CCs in cluster five acetaminophen_Low, methapyrilene_High, nitrofurazone_Medium, acetaminophen_Medium, nitrofurazone_High, acetaminophen_Highis a group of toxic compounds and genes Gsta5, Gss, Mgst2, Gstp1, Gsr, Gclc, Gclc, G6pd, 1374070_at incluster five are the toxicogenomic biomarkers that are regulated by the mentioned toxic CCs. In this study we have also discovered significant gene-CC intersections using the logistic moving range chart (LMRC) on the pLVM generated gene-CC joint probability. The results obtained from the used method are also valid from the biological view point that has been verified from the literature. Therefore, the pLVM a probabilistic model can cluster CCs and genes simultaneously and more efficiently.

Keywords: Toxicogenomics, glutathione metabolism pathway, chemical compounds, pLVM, latent class.

AMS Classification: 92C50.

1. Introduction

Toxicogenomics is a *"omics"* technology stems from toxicology has been gripped a great attention recently because of its safety assessment capacity of chemical compound in the drug development pipeline. The swift progress and evolution on genomic- (DeRisi et al., 1996; Duggan et al., 1999), proteomic- (Lueking et al., 1999; Rubin and Merchant, 2000; Steiner and Anderson, 2000), and metabolomics- (Corcoran et al., 1997; De Beer et al., 1998) technologies enables the application of gene expression for understanding chemical and other environmental stressors' effects on biological systems. The technologies enrich the emerging toxicity assessment field toxicogenomics. Furthermore, administrating drug on animal, prospective toxicity can be discovered through the gene expression analysis of target organs before phenotypic variation occurred (Fielden et al., 2007; Uehara et al., 2008; Hasan et al., 2018).

The toxicogenomic experiment generates thousands of gene expression data under a wide range of treatment conditions (CCs together with multiple dose level and time points). Analysis of these enormous amount of data for identification of important genes and treatment conditions is a very complicated mission and requires very powerful statistical tools and data mining techniques. Clustering as well as pattern recognition techniques are also very useful in analysis of these types of high throughput microarray data (Valafar, 2002; Hasan et al., 2019b).Cluster analysis techniques is the most popular technique that explore relationships among treatment conditions or genes/biomarkers and the relationship between treatment conditions and genes by grouping them based on their similarity. The most popular clustering algorithms are hierarchical clustering three

(HCT) (Eisen et al., 1998) and k-means (Tavazoie et al., 1999). HCT results a tree-like dendrogram merging sequentially the most similar cluster sub-nodes. K-means is the most commonly used non-hierarchical clustering algorithm unlike in which samples are divided into previously defined k partitions or clusters based on their similarity measure. K-means algorithms and other non-hierarchical clustering algorithms perform poorly when random initial seeds are used but their performance is improved when the results from hierarchical methods are used to form the initial partition (PoDAa, 1989).

In modern toxicogenomics identification of biological-related chemical compounds and genes is an important issue for assessing the chemical compounds or drugs' safety in the early stage of drug development (Hasan et al. 2018). That's why, quantification of chemical compounds and genes relationship is essential to know the effect of compounds on the regulation of specific gene (up or down regulations) that brings phenotypic changes finally. Simultaneously, clustering chemical compounds or genes based on their similar characteristics is also important for better understanding of compounds and genes characteristics. However, the statistical or data mining methods mentioned above or others traditional methods fail to clustering compounds or genes as well as quantifying the relationships between compounds and genes simultaneously. The topic modeling algorithms Probabilistic Latent Variable Model (pLVM) (Hofmann, 2001) and latent dirichlet allocation (LDA)(Blei et al., 2003) can perform these job. Among these models pLVM is the most popular model for dyadic data analysis and applied in various fields like text mining (Zhu et al., 2005), bioinformatics (Bicego et al., 2010; Chang et al., 2003; Joung et al., 2006, Hasan et al., 2018)which usually represent high dimensional data in terms of lower dimensional hidden class. On the other hand, identification of significant gene-CC interactions is very important issue for the biologist and drug developers (Zhu et al., 2005, Hasan et al., 2019a). Therefore, in this study, pLVM along with a set of hidden class selection techniques and logistic moving range chart (LMRC) (Hasan et al., 2019a) are used for clustering genes as well as CCs and identification of significant gene-CC interactions.

2. Materials and Methods

2.1 Datasets

Liver and the kidney are the main detoxification organs. In the liver, glutathione an enzyme which scavenges reactive oxygen species, is one of the major detoxification players (Nyström-Perssonet al., 2013). According to (Nyström-Perssonet al. 2013) *acetaminophen*, *methapyrilene* and *nitrofurazone* are the glutathione depleting compounds. Japanese Toxicogenomics Project (TGP) has been taken the scheme collecting high dimensional toxicogenomic (microarray gene expressions) data systematically since 2002 as a joint government-privet sector project (Uehara, 2010). Both *In vivo* and *in vitro* are the two main types of data that have produced by the TGP. The *in vivo* data, which was collected from *Rattus Norvegicus* at four time points (3hr, 6 hr, 9hr, 24hr) for each of four dose levels (control, low, middle, high) from two organs (liver, kidney). In this study, we consider *Rattus Norvegicus*'s liver expression data of 42 *glutathione* metabolism pathway genes after exposing 10 compounds including *acetaminophen*, *methapyrilene* and *nitrofurazone* along with three dose levels (low, medium, and high) at 24 hour time point from the TGP database.

2.2 Data Processing

In the pLVM the co-occurrence values of compounds and genes are assumed count value. Since each cell in the observed $n \times m$ gene-compound data matrix consisting of $G = g_1, g_2, \cdots, g_n$ genes and $C = c_1, c_2, \cdots, c_m$ compounds. The each and every cell of this data matrix represents the fold change expression value $ev(g_i, c_j)$ of the i^{th} gene and j^{th} treatment condition. These values are transformed into count value $\#(g_i, c_j)$ applying the following the formula: $\#\left(g_i, c_j\right) = \left(100 \times \left(\frac{1}{1+\exp(-ev(g_i, c_j))}\right)\right)$.Which was also used by (Hasan et al., 2018).

2.3 Probabilistic Latent Variable Model

In this study our main objective is to clustering chemical compounds (treatments) and genes on the basis of the transformed $n \times m$ gene-compound count data matrix where $\#\left(g_i, c_j\right)$ represents weight or frequency of the g_i^{th} gene under c_j^{th}compound or treatment condition using pLVM (Hofmann, 2001). In applying pLVM it is assumed that there prevail a set of unobserved latent classes underlying our gene-compound count data matrix. Introducing latent classes $Z = z_1, z_2, \cdots, z_l$ the model quantify the relationships between genes and latent classes as well as CCs and latent classes. The following are the probability definition and underlying assumptions accordingly: (1) $P(z_k)$ is the probability of the k^{th} latentclassand $\sum_{k=1}^{l} P(z_k) = 1$. (2) $P(G_i \backslash Z_k)$ is the probability of the i^{th} gene over the k^{th} latentclass and $\forall z_k$; $\sum_{i=1}^{n} P(G_i \backslash Z_k) = 1$. (3) $P(C_j \backslash Z_k)$ is the probability of the j^{th} CC over the k^{th} latentclass and $\forall z_k$; $\sum_{j=1}^{m} P\left(C_j \backslash Z_k\right) = 1$.Based on these definition and assumptions we obtain the co-occurrence of the gene-compound observed pair(g_i, c_j)considering latentclass$z_k$as follows:

$$P\left(g_i, c_j\right) = P(c_j)P(g_i \backslash c_j)$$

Where

$$P\left(g_i \backslash c_j\right) = \sum_{k=1}^{l} z_k\, P(g_i \backslash z_k) P(z_k \backslash c_j)$$

Applying Bayes' rule, the joint co-occurrence probability function can be written as

$$P\left(g_i, c_j\right) = \sum_{k=1}^{l} z_k\, P(g_i \backslash z_k) P(c_j \backslash z_k) P(z_k)$$

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

$$L(G,C) = \sum_{i=1}^{n} \sum_{j=1}^{m} \#(g_i, c_j) log P\left(g_i, c_j\right)$$

The widely used method for estimating the maximum likelihood parameters of probabilistic model is the Expectation- Maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm starts with a random set of initial parameter values and iterates both the expectation step (E-step) and maximization step (M-step) alternatively until a certain convergence criteria is satisfied. The E and M-step for the total likelihood can be given as follows:

E-step:

$$P(z_k \backslash g_i, c_j) = \frac{P(g_i \backslash z_k) P(c_j \backslash z_k) P(z_k)}{\sum_{k=1}^{l} P(g_i \backslash z_k) P(c_j \backslash z_k) P(z_k)}$$

M-step:

$$P(z_k) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \#(g_i, c_j) P(z_k \backslash g_i, c_j)}{\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} \#(g_i, c_j) P(z_k \backslash g_i, c_j)}$$

$$P(g_i \backslash z_k) = \frac{\sum_{j=1}^{m} \#(g_i, c_j) P(z_k \backslash g_i, c_j)}{\sum_{i=1}^{n} \sum_{j=1}^{m} \#(g_i, c_j) P(z_k \backslash g_i, c_j)}$$

$$P(c_j \backslash z_k) = \frac{\sum_{i=1}^{n} \#(g_i, c_j) P(z_k \backslash g_i, c_j)}{\sum_{i=1}^{n} \sum_{j=1}^{m} \#(g_i, c_j) P(z_k \backslash g_i, c_j)}$$

2.4 Number of Latent Class Selection Methods

Appropriate number of latent class selection for pLVM is an important task before the application of pLVM in data matrix since insufficient number of latent classes or topics is too coarse to recognize accurate clusters. Conversely, excessive number of latent classes could make the model more complex, drawing conclusions (Zhao et al., 2014). In our study, we have used a set of very popular methods for choosing optimum number of latent class. The number of latent classes which suggested by the maximum number of methods considered as the desired number of latent classes for the data matrix. The methods like Kaiser-Guttman rule, Parallel Analysis, Scree Test Optimal Coordinate, Scree Test Acceleration Factor:(Guttman, 1954; Kaiser, 1960; Horn, 1965; Montanelli and Humphrey, 1976; Cattell, 1966) are used in this study all are based on principal components eigen values of gene-compound correlation matrix.

2.5 Identification of Up-regulatory and Down-regulatory Gene-CCs Interactions

As we have mentioned earlier that clustering genes as well as CCs is very important issue in toxicogenomic studies. There are several studies in the literature used different clustering methods for this purpose (Hasan et al., 2019b, Hasan et al., 2018). However, these methods is not suitable for the identification of significant up-regulatory and down-regulatory gene-CCs interactions. This problem can be overcome using the logistic moving range chart (LMRC) (Hasan et al., 2019a). There are central line (CL) which represents the average value of the quality characteristics corresponding to in-control state, an upper control limit (UCL) and a lower control limit (LCL) in the LMRC. If the probability of fold change value of a gene corresponding to a compound or gene-compound interaction $P\left(G_i, C_j\right)$ plots outside the UCL or LCL we consider that interaction as significant up-regulatory and down-regulatory interaction. Where up-regulatory and down-regulatorygene-CCs interactions indicates a CC which influence a gene to be up-regulated and down-regulated respectively.

3. Results and Discussions

3.1 Number of Latent Class Selection

We have applied the methods described in section 2.4 for selection of optimum number latent class in the dataset. The Kaiser-Guttman rule, Parallel Analysis, Scree Test Optimal Coordinate and Scree Test Acceleration Factor suggest that there are 9, 6, 6 and 1 latent classes in the dataset respectively (Figure 1). For this study, we assume that there are 6 latent classes or clusters in the dataset as it is suggested by maximum number methods.

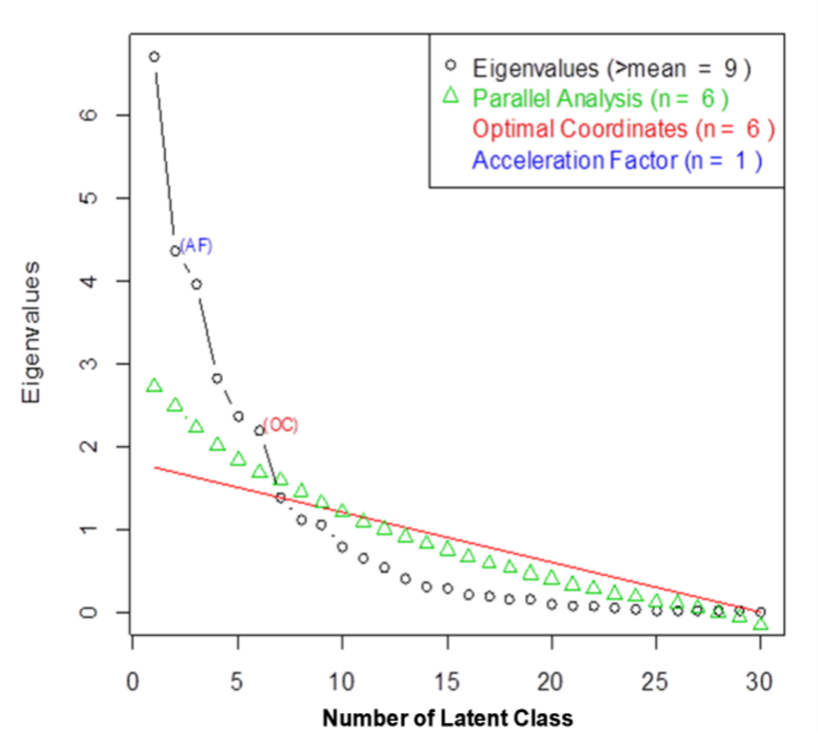

Figure 1: Kaiser-Guttman, Parallel Analysis, Scree Test Optimal Coordinate and Scree Test Acceleration Factor rule for optimum number of latent class selection based on eigen values of correlation matrix of the dataset.

3.2 Chemical Compound (CC) or Treatment Conditions Clustering

In the liver, detoxification process is always continuing and glutathione plays the major role in this process. It conjugates target toxic compounds and exports the conjugated compounds into bile ducts. To analyze drug induced glutathione depletion from TGP dataset, we find out the pattern of drugs along with their dose levels at 24 hour time point. A compound/drug give different probability over the latent classes and it will remain in that latent class where it gives maximum probability. For example, according to Figure 2 all dose levels of acetaminophen, medium and high dose level of nitrofurazone and only high dose level of methapyrilene are showed maximum probability in the latent class five. Therefore, they remain in cluster five. In the same way, all the CCs along with their dose levels were grouped that has been presented in Table 1. The cluster five (latent class five) is a cluster of highly toxic compounds because all the compounds belonging to this cluster are toxic compounds with their dose levels. Nyström-Perssonet al., (2013) also found the CCs in cluster five are glutathione depleting compounds.

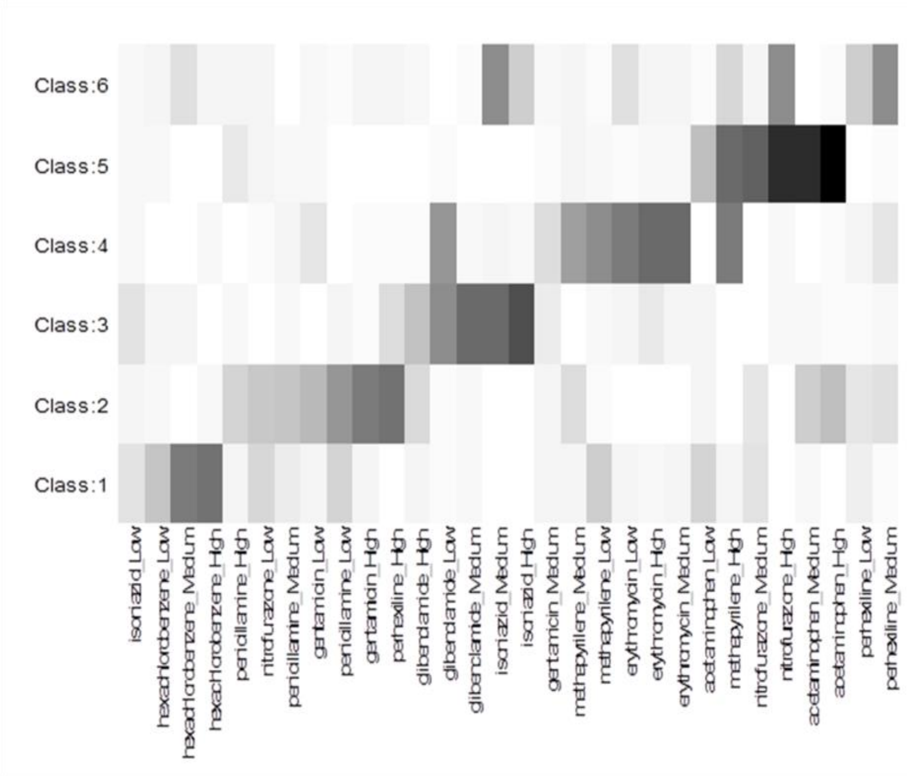

Figure 2: CCs clustering based onlatent classes or plot of $P(C_j\backslash Z_k)$. In the plot the darker block represent the maximum probability and a CC along with its dose level will remain in that class where it has the maximum probability compare to the other classes.

Table 1: Cluster membership of different CCs with their dose level.

Class:1	Class:2	Class:3	Class:4	Class:5	Class:6
isoniazid_Low, hexachlorobenzene_Low, hexachlorobenzene_Medium, hexachlorobenzene_High	penicillamine_High, nitrofurazone_Low, penicillamine_Medium, gentamicin_Low, penicillamine_Low, gentamicin_High, perhexiline_High	glibenclamide_High, glibenclamide_Low, glibenclamide_Medium, isoniazid_Medium, isoniazid_High	gentamicin_Medium,methapyrilene_Medium,methapyrilene_Low, erythromycin_Low, erythromycin_High, erythromycin_Medium	acetaminophen_Low, methapyrilene_High, nitrofurazone_Medium, acetaminophen_Medium, nitrofurazone_High, acetaminophen_High	perhexiline_Low, perhexiline_Medium

3.3 Gene Clustering

In our study, the toxic effect of considered CCs including three common glutathione depleting compounds *acetaminophen*, *methapyrilene*and *nitrofurazone* under different conditions (dose level and 24 hourtime point) were studied over the 42 genes/probes which belong to the *glutathione* metabolism pathway. The clusters of these genes are given in Figure 3 which is generated by pLVM. Cluster five contains the genes Gsta5, Gss, Mgst2, 1371942_at, Gclc, Gstp1, Gsr, Gclc and G6pdwhich have showed same pattern in response to toxic compounds (Table 2, cluster 5). It has also proved by (Nyström-Persson et al., 2013) that the top four ranked probes are Mgst2 (microsomal glutathione S-transferase 2), G6pd (glucose-6-phosphate dehydrogenase), Gsr (glutathione reductase) and Gclc (glutamate–cysteine ligase) that are influenced by the toxic CCs in (cluster 5 (Table 1)). Gclc is known to accelerate glutathione synthesis, and Gsr and G6pd are involved in the conversion of glutathione from the oxidized form to the reduced one. Mgst2 is glutathione-S-transferase, which is the main enzyme for detoxification of toxic compounds by the conjugation reaction.

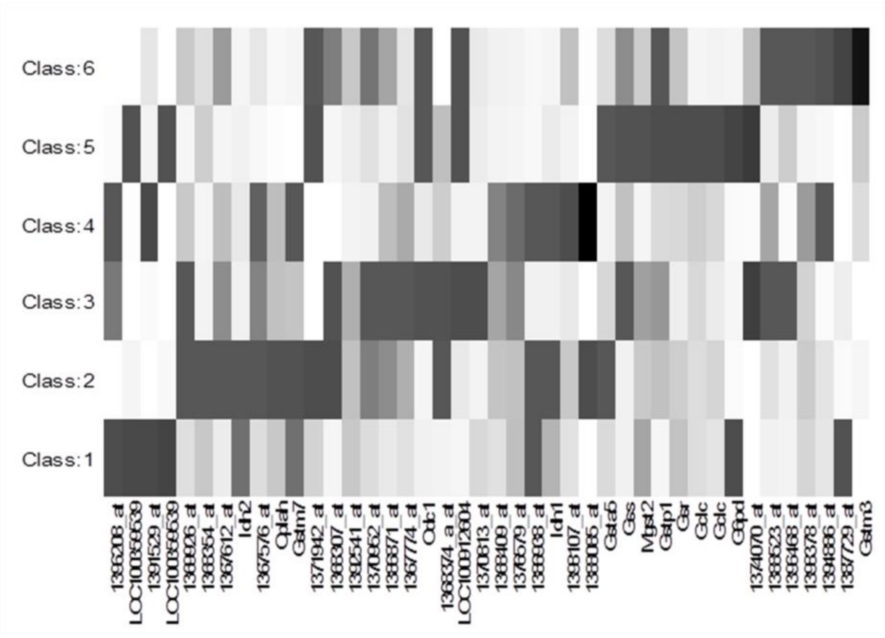

Figure 3: Genes/probs clustering based on hidden class or plot of $P(G_j \backslash Z_k)$. In the plot the darker block represent the maximum probability and a gene/probswill remain in that class where it has the maximum probability compare to the other classes.

Table 2: Cluster membership of gene/prob

Class:1	Class:2	Class:3	Class:4	Class:5	Class:6
1396208_at, LOC100359539, 1391529_at, LOC100359539	1369926_at, 1368354_at, 1367612_at, Idh2, 1367576_at, Oplah, Gstm7, 1371942_at, 1368307_at,	1392541_at, 1370952_at, 1386871_at, 1367774_at, Odc1, 1368374_a_at LOC100912604, 1370813_at	1368409_at, 1376579_at, 1386938_at, Idh1, 1398107_at, 1388085_at	Gsta5, Gss, Mgst2,Gstp1, Gsr, Gclc, Gclc, G6pd, 1374070_at	1388523_at, 1396468_at, 1398378_at, 1394886_at, 1387729_at, Gstm3

3.4 Identification of Up-regulatory and Down-regulatory Gene-CCs Interactions

In this section we have schemed the likelihood $P\left(G_i, C_j\right)$ of the gene-CCs interactions of CCs along with their dose levels and 42 glutathione metabolism pathway genes using the LMRC for identification of significant gene-compound interactions. The top 20 significant up-regulatory and top 20 significant down-regulatory gene-CCs interactions are presented in Table 3 and Table 4 respectively. In Table 3 the gene-CCs interactions which produce larger likelihood $(P\left(G_i, C_j\right) > UCL)$are the toxic CCs and up-regulated biomarker genes. In Table 4 the gene-CCs interactions which produce smaller likelihood $(P\left(G_i, C_j\right) < LCL)$ are the toxic CCs and down-regulated biomarker genes.

Table 3: Top 20 up-regulatory CCs-gene relationships.

CCs-Gene	Likelihood	CCs-Gene	Likelihood
nitrofurazone_High:1374070_at	0.00176589	methapyrilene_Medium:1388085_at	0.00139783
nitrofurazone_High:Gstm3	0.00174019	isoniazid_Medium:Gstm3	0.00136946
acetaminophen_High:1374070_at	0.00161524	acetaminophen_Medium:G6pd	0.00136647
perhexiline_Medium:Gstm3	0.00149566	erythromycin_Low:1388085_at	0.00134732
acetaminophen_Medium:1374070_at	0.00148971	nitrofurazone_High:Gstp1	0.00132410
erythromycin_Medium:1388085_at	0.00148157	nitrofurazone_Medium:G6pd	0.00131778
erythromycin_High:1388085_at	0.00144884	methapyrilene_High:1388085_at	0.00131558
methapyrilene_High:Gstm3	0.00139796	acetaminophen_High:G6pd	0.00131529
acetaminophen_High:Gsr	0.00130968	acetaminophen_High:Gclc	0.00130383
acetaminophen_High:Gstp1	0.00130360	acetaminophen_High:Gclc	0.00128999

Table 4: Top 20 down-regulatory CCs-gene relationships

CCs-Gene	Likelihood	CCs-Gene	Likelihood
isoniazid_High:LOC100359539	1.0174e-05	hexachlorobenzene_Low:1388085_at	2.3627e-04
isoniazid_Medium:LOC100359539	2.1876e-05	acetaminophen_High:1387729_at	2.3628e-04
isoniazid_High:LOC100359539	2.6849e-05	acetaminophen_High:1396208_at	2.5267e-04
isoniazid_Medium:LOC100359539	3.7102e-05	perhexiline_High:G6pd	2.6009e-04
hexachlorobenzene_Medium:1388085_at	1.2631e-04	isoniazid_High:1371942_at	2.6400-04
perhexiline_High:LOC100359539	1.5065e-04	perhexiline_High:1391529_at	2.6746e-04
hexachlorobenzene_High:1374070_at	2.0341e-04	perhexiline_Medium:LOC100359539	2.7331e-04
perhexiline_High:LOC100359539	2.2569e-04	perhexiline_High:1396208_at	2.7414e-04
isoniazid_High:G6pd	3.4345e-04	acetaminophen_Low:1388085_at	3.3456e-04
nitrofurazone_High:Oplah	3.5856e-04	nitrofurazone_High:1396208_at	3.5602e-04

4. Conclusion

Correlated genes and compounds clustering simultaneously to the biological processes is one of the main objectives of toxicogenomic studies (Hasan et al., 2018). The Probabilistic Latent Variable Model (pLVM) and latent class selection strategies for the gene-CCdataset can successfully cluster genes as well as CCs. The CCs having same pattern of toxic effect over genes were grouped in the same cluster and the genes which have the same pattern of response to the CCs were belongs to the same cluster. Nonetheless, the pLVM cannot separate the significant up-regulatory and down-regulatory gene-compound interactions from the equal-regulatory interactions. But identification of significant up-regulatory and down-regulatory interactions between genes and chemical compounds or drugs are the cornerstone in toxicogenomic studies as well as in drug discovery and development (Hasan et al., 2019a; Zhu et al., 2005). Therefore, logistic moving range chart (LMRC) (Hasan et al., 2019a) were used for the identification of significant up-regulatory and down-regulatory gene-compound interaction/relationships. The results that have been got were validated from biological viewpoint. Thus, pLVM, number of latent class selection methods and LMRC can give information of toxicity study for toxicogenomic study and drug development. The limitation of the method in analyzing toxicogenomic data is that, it consider the equal number of clusters for the CCs and genes. However, in practical there may not the same number of clusters for CCs and genes. There should be done more research in this regard.

Acknowledgements: We are grateful to higher education quality enhancement project (HEQEP) because this study was financed by HEQEP Sub-Project (CP: 3603, W2, R3), Dept. of Statistics, Rajshahi University, Rajshahi, Bangladesh.

References

[1] Bicego, M., Lovato, P., Ferrarini, A. and Delledonne, M. (2010). Biclustering of expression microarray data with topic models, International Conference on Pattern Recognition, pages 2728-2731.

[2] Blei, D. M., Ng, A.Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022.

[3] Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

[4] Chang, J. H., Chi, S. W. and Zhang, B. T. (2003). Gene Expression Pattern Analysis via Latent Variable Models Coupled with Topographic Clustering, Genimics& Informatics, Vol. 1(1), pages 32-39.

[5] Corcoran, O., Spraul, M., Hofmann, M., Ismail, I. M., Lindon, J. C., and Nicholson, J. K. (1997). 750 MHz HPLC-NMR spectroscopic identification of rat microsomal metabolites of phenoxypyridines. J. Pharm. Biomed. Anal. 16: 481- 489.

[6] De Beer, R., Van den Boogaart, A., Cady, E., Graveron Demilly, D., Knijn, A., Langenberger, K. W., Lindon, J.C., Ohlhoff, A., Serrai, H., and Wylezinska-Arridge, M. (1998). Absolute metabolite quantification by in vivo NMR spectroscopy: V. Multicentre quantitative data analysis trial on the overlapping background problem. Magn. Reson. Imaging. 16: 1127-1137.

[7] Dempster, A., Laird, N. and Robin, D. (1977). Maximum likelihood from incomplete data via the EM Algorithm, .J. Roy. Stat. Soc., B, 39, 1-38.

[8] DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y.A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer [see comments]. Nat. Genet. 14: 457-460.

[9] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. Nat. Genet. 21: 10-14.

[10] Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America, 95(25):14863-14868.

[11] Fielden, M. R. et al. (2007). A gene expression biomarker provides early prediction and mechanistic assessment of hepatic tumor induction by nongenotoxic chemicals. Toxicol. Sci., 99, 90-100.

[12] Guttman, L. (1954). Some necessary conditions for common factor analysis. Psychometrika, 19, 149-162.

[13] Hasan, M. N., A. A., Rahman, M. R. R. and Mollah, M. N. H., (2019a). Robust identification of significant interactions between toxicogenomic biomarkers and their regulatory chemical compounds using logistic moving range chart. Computational Biology and Chemistry, Volume 78, February 2019, Pages 375-381,

[14] Hasan, M. N., Malek, M. B., Begum, A. A., Rahman, M., and Mollah, M. N. H. (2019b). Assessment of Drugs Toxicity and Associated Biomarker Genes Using Hierarchical Clustering. Medicina, 55, 451. doi:10.3390/medicina-55080451w

[15] Hasan, M. N., Rana, M. M., Begum, A. A., Rahman, M. R. R., and Mollah, M. N. H. (2018). Robust co-clustering to discover toxicogenomic biomarkers and their regulatory doses of chemical compounds using logistic probabilistic hidden variable model. Front. Genet. 2018, 9, 516. doi: 10.3389/fgene.2018.00516.

[16] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42(1-2):177-196.

[17] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.https://doi.org/10.1016/j.compbiolchem.2018.12.020

[18] 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, vol. 22 no. 16, pages 2005-2011.

[19] Kaiser, H. F. (1960). The application of electronic computer to factor analysis. Educational and Psychological Measurement, 20, 141-151.

[20] Lueking, A., Horn, M., Eickhoff, H., Bussow, K., Lehrach, H., and Walter, G. (1999). Protein microarrays for gene expression and antibody screening. Anal Biochem, 270: 103-111.

[21] Montanelli, R. G. and Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. Psychometrika, 41, 341-348.

[22] Nyström-Persson, j., Igarashi, Y., Ito, M., Morita, M., Nakatsu, N., Yamada, H., Mizuguchi, K. (2013). Toxygates: interactive toxicity analysis on a hybrid microarray and linked data platform, Bioinformatics, 23, 3080-3086.

[23] PoDAa. (1989). Discriminant analysis and clustering. Statistical Science, 4(1):34-69.

[24] Rubin, R. B., and Merchant, M. (2000). A rapid protein profiling system that speeds study of cancer and other diseases. Am. Clin. Lab. 19: 28-29.

[25] Steiner, S., and Anderson, N.L. (2000). Pharmaceutical proteomics. Ann. N.Y. Acad. Sci. 919: 48-51.

[26] Tavazoie, S. Hughes, J.D., Campbell M.J., Cho, R.J. and Church, G.M. (1999). Systematic determination of genetic network architecture. Nature genetics, 22(3):281-285.

[27] Uehara T., (2010). The Japanese toxicogenomics project: Application of toxicognomics, Molecular Nutrition Food Research, 54, 218-227.

[28] Uehara, T. et al. (2008). A toxicogenomics approach for early assessment of potential non-genotoxichepatocarcinogenicity of chemicals in rats. Toxicology, 250, 15-26.

[29] Valafar, F. (2002). Pattern recognition techniques in microarray data analysis: A survey, Annals of the New York Academy of Sciences, vol. 980, pp. 41-64.

[30] Zhao, W., Zou, W., and Chen, J.J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics, 15(Suppl 11):S11.

[31] Zhu, S., Okuno, Y., Tsujimoto, G. and Mamitsuka, H. (2005). A probabilistic model for mining implicit 'chemical compound-gene' relations from literature, Bioinformatics, Vol. 21 Suppl. 2 2005, pages ii245-ii251.