Sequential Extraction of Several Gene-sets with Proper Groups of Individuals for Gene Expression Data Analysis

International Conference on ***Statistical Data Mining for Bioinformatics Health Agriculture and Environment***, 21-24 December, 2012.

ISBN: **978-984-33-5876-9**

Sequential Extraction of Several Gene-sets with Proper Groups of Individuals for Gene Expression Data Analysis

Md. Bahadur Badsha[1]*,Nusrat Jahan[2], Md. NurulHaque Mollah[3] and Hiroyuki Kurata[1, 4]

[1]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan.

[2]Department of Applied Mathematics, University of Rajshahi, Rajshahi-6205, Bangladesh.

[3]Department of Statistics, University of Rajshahi, Rajshahi-6205, Bangladesh.

[4]Biomedical Informatics R&D Center, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka, 820-8502, Japan.

{MBB:mbahadur_stat_ru@yahoo.com; NJ: njahan3396@yahoo.com; MNHM: mnhmollah@yahoo.co.in

and HK: kurata@bio.kyutech.ac.jp}

* Corresponding author

Abstract: One of the ultimate goals of microarray gene expression data analysis in bioinformatics is to identify individual genes or gene-sets which influence the gene expression patterns. There are several research areas in bioinformatics, where data analysis offers a challenging statistical problem due to their high dimensionality with small sample of sizes. Clustering is one of the most popular statistical techniques to addressing these challenges. Nowak and Tibshirani (2008) proposed complementary hierarchical clustering (CHC) for sequential exaction of several gene-sets having relatively low expressions than highly expressed genes. However it produces misleading clustering results for sequential exaction of several gene-sets if there exist some contaminations (outliers) in the gene expression data, which is an important issue in gene expression data analysis research field. Therefore, in this paper we proposed a robust statistical clustering technique based on the value of tuning parameter β, we called β-CHC for sequential extraction of biologically important gene-sets has similar expression patters with proper groups of individuals the genes expression data analysis in bioinformatics from the robustness points of view. The proposed robust method reduces to the traditional method when we put the value of tuning parameter $\beta \to 0$. Simulation gene expression data clustering results show that the performance of the proposed method is better than performance of the traditional method in the case of data contaminations; otherwise, it shows almost equal performance.

Key words: Gene expression, complementary hierarchical clustering based on β (β-CHC), Minimum β-divergence, Robustness.

1 Introduction

Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting and utilizing information from biological sequences and molecules. It is the hybrid version of molecular biology, statistics and computer technology. Gene expression data analysis by statistical algorithms might be playing the significant task to reach the final goal. High dimensional gene expression microarray data throw the challenge to multivariate analysis and to develop effective ways to analyze gene expression data. Now a day's DNA microarray tools make it easy to monitor the millions of gene expressions simultaneously during important biological processes. One of the ultimate goals of microarray gene expression data analysis in bioinformatics is to identify individual genes or gene-sets which influence the gene expression patterns. However, millions of genes and their complex functions in gene expression analysis gradually increasing the challenges for interpreting the results from the high dimensional data. For gene expression data analysis, the data are arranged in a matrix form, where the rows represent the genes and the columns

represent the individuals. There are several research areas in bioinformatics, where data analysis offers a challenging statistical problem due to their high dimensionality with small sample of sizes. Clustering is one of the most popular statistical techniques to addressing these challenges. Many clustering methods have been proposed for gene expression data analysis (Terry Speed 2003). Differential regulation often means differential expression, and a number of useful methods for identifying differentially expressed (DE) genes or gene sets are available (Datta et al., 2004; Allison et. al., 2006; Barry et. al., 2008; Ho et al., 2007; Newton et. al., 2007; Tracy and Wilson 2011).Detailed discussions about statistical analysis of gene expression microarray data are given (Datta 2003).

Hierarchical clustering (HC) algorithm is the most widely used unsupervised statistical technique for analyzing microarray gene expression data. It becomes a very useful popular tool for analyzing microarray gene expression data from the research works of Eisen et al. (1998). In Hierarchical clustering, the number of classes is determined by cutting the tree structure at certain level chosen subjectively by the user. When applying HC algorithm to the gene expression data to cluster individuals or phenotypic outcomes, the HC algorithms produce clusters based on the highly differentially expressed (DE) genes those have very similar expression patterns. These types of highly DE genes sometimes may not be relevant in the biological process. Therefore, we have to need to explore another low expressed gene or gene-sets having important biological functions. In this context, Nowak and Tibshirani (2008) proposed the CHC for sequential exaction of several gene-sets having relatively low expressions than highly expressed genes. However it produces misleading clustering results for sequential exaction of several gene-sets if there exist some contaminations (outliers) in the gene expression data, which is an important issue in gene expression data analysis research field. In gene expressions microarray data are often contaminated by outliers due to the many steps involved in the experimental process from hybridization to image processing for producing data. For example, an outlying data value could occur because of scratches or dust on the surface, imperfections in the glass, imperfections in the array production (Gotterdo et al. 2006). Therefore, in this context we proposed a robust statistical clustering technique based on the value of the tuning parameter β we called β-CHC for sequential extraction of biologically important gene-sets has similar expression patters with proper groups of individuals by minimizing β-divergence(Minami and Eguchi 2002) for the genes expression data analysis. The proposed method reduces to the traditional method when we put the value of tuning parameter $\beta \rightarrow 0$.

2. Proposed Model and Parameter Estimation

Let us consider the dummy variable regression model in matrix notation written as

$$\boldsymbol{Y}_g=\boldsymbol{D}_k\boldsymbol{\delta}_{gk}+\boldsymbol{\epsilon}_{gk} \qquad (1)$$

Where $\mathbf{Y}_g$=gth gene expression, $\boldsymbol{D}_k$= dummy variables, $\boldsymbol{\delta}_{gk}$=the regression coefficient of the dummy variables regression model which would be estimated, $\boldsymbol{\epsilon}_{gk} \rightarrow N(o,\sigma_{gk}^2)$ is the error term. Then the probability density function (pdf) for the *i*-th component y_{gi} of $\boldsymbol{Y}_g$ for the *k*-th cut is given by

$$f\left(y_{gi}\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{y_{gi}-\mathbf{d}_{hi}\boldsymbol{\delta}_{gk}}{\sigma_{gk}}\right)^2\right\} \qquad (2)$$

where $\boldsymbol{\delta}_{gk}$ and σ_{gk}^2 are the parameter of the distribution. Then the β-likelihood function (Mollah et. al., 2007 and 2009) for parameter $\boldsymbol{\varphi}_{gk}=(\boldsymbol{\delta}_{gk},\sigma_{gk}^2)$ is given by

$$L_\beta\left(\boldsymbol{\varphi}_{gh}\right)=\frac{1}{\beta}\left[\frac{1}{nl_\beta(\boldsymbol{\varphi}_{gh})}\sum_{i=1}^{n}\mathrm{f}\left(y_{gi}\right)^\beta-1\right], \qquad (3)$$

where $l_\beta\left(\boldsymbol{\varphi}_{gk}\right)=\left[\int\left\{f\left(y_g/\boldsymbol{\varphi}_{gk}\right)\right\}^{\beta+1}dy_g\right]^{\beta/\beta+1}=\left(1+\beta\right)^{-\frac{\beta}{2(1+\beta)}}\left(2\pi\sigma^2\right)^{-\frac{\beta^2}{2(1+\beta)}}$.

Maximization of β-likelihood function is equivalent to the minimization of β-divergence for estimating the model parameters $\boldsymbol{\varphi}_{gh}$ (Mollah et. al., 2007 and 2009). The β-likelihood function is used in several statistical algorithms

for robustification (Mollah et. al., 2007 and 2009). It is highly robust against outliers. To estimate the parameters $\boldsymbol{\varphi}_{gk} = (\boldsymbol{\delta}_{gk}, \sigma^2_{gk})$, we minimize the β-divergence which equivalent to maximize the β-likelihood function (3) with respect to parameters $\boldsymbol{\delta}_{gk}$ and σ^2_{gk}, respectively. Then the maximum β-likelihood estimator for the parameters $\boldsymbol{\delta}_{gk}$ and σ^2_{gk} obtained iteratively as follows:

$$\boldsymbol{\delta}_{gk\beta}^{(t+1)} = \left[\sum_{i=1}^{n} f_\phi\left(y_{gi}\right)^\beta \ \mathbf{d}_{ki}^T \mathbf{d}_{hi}\right]^{-1} \sum_{i=1}^{n} f_\phi\left(y_{gi}\right)^\beta \ y_{gi}\mathbf{d}_{ki}^T$$

$$= \left[\mathbf{D}_k{}^T \left\{\mathbf{D}_k \# \left(\mathbf{W}_{g\beta}^t 1\right)\right\}\right]^{-1} \left\{\mathbf{D}_k \# \left(\mathbf{W}_{g\beta}^t 1\right)\right\}^T \mathbf{Y}$$

(In matrix notation) (4)

$$\sigma_{gk}^{2\ (t+1)} = \frac{(1+\beta)\sum_{i=1}^{n} f_\phi\left(y_{gi}\right)^\beta \ \left(y_{gi} - \mathbf{d}_{gi}\boldsymbol{\delta}_{gh}\right)^2}{\sum_{i=1}^{n} f_\phi\left(y_{gi}\right)^\beta}$$

and $$= \sigma_{gk}^{2\ (t+1)} = (1+\beta)\left(\mathbf{Y}_g - \mathbf{D}_k\boldsymbol{\delta}_{gk}^{(t+1)}\right)^T \left[\left(\mathbf{Y}_g - \mathbf{D}_k\boldsymbol{\delta}_{gk}^{(t+1)}\right) \# \mathbf{W}_{g\beta}^t\right] \left[1^T \mathbf{W}_{g\beta}^t\right]^{-1} \quad (5)$$

where $$\mathbf{W}_{g\beta} = \left[\exp\left\{-\frac{\beta}{2}\left(\frac{y_{gi} - \mathbf{d}_{ki}\boldsymbol{\delta}_{gk}}{\sigma_{gk}}\right)^2\right\}\right]_{n\times 1},$$

which is called a β-weight vector (Mollah et. al., 2007 and 2009). It produces almost zero weight for outlying gene expressions.

3. Proposed Methodology

In microarray gene expression data analysis important is to identify sequential exaction of individual genes or gene-sets which influence the gene expression patterns from the robustness point of view. We develop complementary hierarchical clustering technique based on β (β-CHC)by minimizing the β-divergence (Minami and Eguchi 2002) for sequential extraction of important genes sets influencing patients groups in the genes expression microarray data analysis. First we calculated robust correlation matrix using robust covariance matrix (Mollah et al., 2009). Then we calculated robust dissimilarity matrix using robust correlation matrix. Perform clustering algorithms on this dissimilarity matrix. Then it is known as robust hierarchical clustering (RHC) (Mollah et al., 2009). The proposed β-CHC procedure performs within 3 different steps. First, apply RHC on the original dataset. After applying RHC the results can be represented by a dendrogram. For every cut between two heights in the dendrogram in a particular gene we can get group labels of samples. Then we fit a dummy variable linear regression model using this group labels for each gene. We estimate the model parameter by minimizing the β-divergence. Secondly, compute the residual matrix which we call modified data, and finally apply β-CHC on the modified data. We also calculate gene-sets which influence the gene expression clustering patterns using gene important (GI). GI shows that corresponding gene-set is important for this clustering pattern. The proposed method performs with a weight function called β-weight function (Mollah et al., 2007). It plays the key role on the performance of the proposed method. It produces almost zero weights for outlying gene expressions. Thus estimates become robust. When we put the value of tuning parameter $\beta \to 0$, it reduces to the classical CHC. The robustness of the proposed method we can be test using influence function (not shows this paper). The values of the tuning parameter β play a key role on the performance of the proposed method. It controls the balance between the robustness and efficiency of the estimators. Smaller β produces more efficient results than larger β, while larger β produces more robust results than smaller β. In the simulation study we select β using cross-validation Mollah et al., (2007).

Summarizes the proposed procedure are given below:

(i) Select an appropriate β by cross validation.

(ii) Compute $\boldsymbol{\delta}_{g\beta k}$ and $\sigma^2_{g\beta k}$ obtained iteratively using equation (4) and (5).

(iii) Compute robust correlation matrix using $\sigma^2_{g\beta k}$ matrix.

(iv) Compute robust dissimilarity matrix using robust correlation matrix.

(v) Perform HC using dissimilarity matrix.

(vi) HC results represented by a dendrogram, cutting these dendrogram in different height then fit the dummy variable linear regression model. Computing residuals matrix using dummy variable linear regression model. These residuals matrix called modified data.

(vii) Apply β-CHC on modified data.

It should be noted here that the proposed procedure reduce to the classical procedure for β=0.

4. Results and Discussion

4.1 Artificially Generated Gene Expression Data in Absence of Outliers

To investigate the performance of the proposed method in a comparison of the classical method, we generated a microarray dataset for simulation study. The dataset is simulated by the models as displayed in **Fig.1**. In **Fig.1** the datasets have 4 different labels corresponding to the 4 different sets of DE genes, which are represented by the matrix ***X*** as shown in **Fig. 2(A)**. The rows of ***X*** are assumed the genes and columns of ***X*** the individuals. In **Fig. 1** the genes (1-20) in the table as assigned +14 (positive intensity) and -14 (negative intensity) are highly DE between the classes of individuals {1, 2, 3, 4, 5, 6, 7} and {8, 9, 10, 11, 12, 13, 14}. Similarly the genes (21-40)as assigned +8 (positive intensity) and -8 (negative intensity) are high medium DE between the classes of individuals {1, 2, 3, 4, 8, 9, 10} and {5, 6, 7, 11, 12, 13, 14}. Similarly the genes (41-60)as assigned +4 (positive intensity) and -4 (negative intensity) are low medium DE between the classes of individuals {1, 2, 5, 6, 10, 13, 14} and {3, 4, 7, 8, 9, 11, 12}. Finally the genes (61-80)as assigned +2 (positive intensity) and -2 (negative intensity) are low DE with the class of individuals {1, 3, 5, 7, 9, 11, 13} and negatively low DE with the class of individuals {2, 4, 6, 8, 10, 12, 14}, respectively. To randomize the gene expressions among the individuals, we randomly added the Gaussian noise with N (0, 1) to expression of each gene. All the simulation gene expression data analysis results computed by the R programming.

Now we would like to use the above groups of individuals corresponding to four gene-sets (1-20), (21-40), (41-60) and (61-80) as the true / reference results to examine the performance of classical and the proposed method for sequential extraction of gene-set with proper groups of individuals. Now we would like to sequential extraction of above gene-sets with corresponding individuals which influence the gene expression clustering patterns. When we perform HC, the clustering patterns {1, 2, 3, 4, 5, 6, 7} and {8, 9, 10, 11, 12, 13, 14} are generated by most highly DE gene-set (1-20) that have very similar expression patterns with GI greater than 0.8 for each gene in that gene-set (see **Fig.2(B)**).

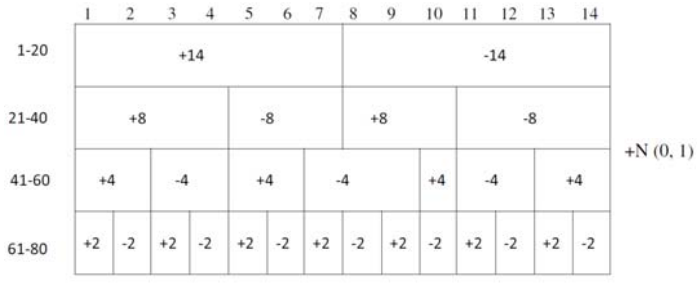

Fig. 1: Shows the 80×14 matrix of simulated gene expressions data with 4 different levels with respect to 4 different sets of DE genes. The rows of the matrix are assumed the genes and the columns of the matrix the individuals. To randomize the gene expressions among the individuals, we randomly added the Gaussian noise with N (0, 1) to the expression of each gene.

However, sometimes, these types of highly DE genes may be irrelevant with the organic processes under study. To overcome these problems, we need to explore another gene-set. For exaction of another gene-set we apply CHC using the output of HC as the input of CHC. At step-1 in CHC, we see that the clustering patterns {1, 2, 3, 4, 8, 9, 10} and {5, 6, 7, 11, 12, 13, 14}, are generated by high medium DE gene-set (21-40) that have closely related expression patterns with GI greater than 0.8 for each gene in that gene-set (see **Fig.2(C)**). If this high medium DE gene-set (21-40) also is not relevant to the biological processes, then CHC algorithm explores another low expressed gene-set in step-2. At step-2 in CHC, we see that the clustering patterns {1, 2, 5, 6, 10, 13, 14} and {3, 4, 7, 8, 9, 11, 12}, are generated by the low medium DE gene-set (41-60) that has closely related expression patterns with GI greater than 0.8 for each gene in that gene-set (see **Fig.2(D)**). Again if this low medium DE gene-set (41-60) also is not relevant to the biological processes, then CHC algorithm explores another low expressed gene-set in step-3. At step-3 in CHC, we see that the clustering patterns {1, 3, 5, 7, 9, 11, 13} and {2, 4, 6, 8, 10, 12, 14}, are generated by the low DE gene-set (61-80) that has closely related expression patterns with GI greater than 0.8 for each gene in that gene-set (see **Fig.2(E)**).

To examine the performance of the proposed method with this same dataset ***X,*** first we compute the dendrogram of β-divergence based RHC (Mollah et. al., 2009) with the tuning parameter β=0.005. Then we apply proposed approach using the output of RHC as the input of β-CHC with β =0.005.We selected the tuning parameter β by cross-validation (CV) (Mollah et. al., 2007). In our proposed method we have to select β in every step during our analysis. The clustering results of the proposed method in absence of outliers shown in **Fig. 2(F), 2(G), 2(H) and 2 (I).** Now, we compare those clustering results with the classical CHC results as mention above. We observed that it shows almost equal performance.

The proposed method performs with a weight function called β-weight function. We calculate the β-weight plot for each gene for both methods classical CHC (β-CHC with β=0) and proposed β-CHC (with β=0.005. We observed the β-weight plot in **Fig. 3** for absence of outliers**;** we see that for both methods classical CHC (β-CHC with β=0) and proposed β-CHC (with β=0.005) are produces almost same β-weight for each gene. **Figs. 3 (A), (B), (C)** are the β-weight plot for each gene using classical CHC method. Similarly **Figs. 3 (D), (E), (F)** are the β-weight plot for each gene using classical proposed method. The gray line indicates that whose genes are outliers or not. Under the gray line genes are consider (outlying gene) outliers and this weight multiply by corresponding genes intensity. That means if any genes consider as outlying gens (under the gray line) then it multiply by very small weight then became a reasonable. This is the main key of our proposed method.

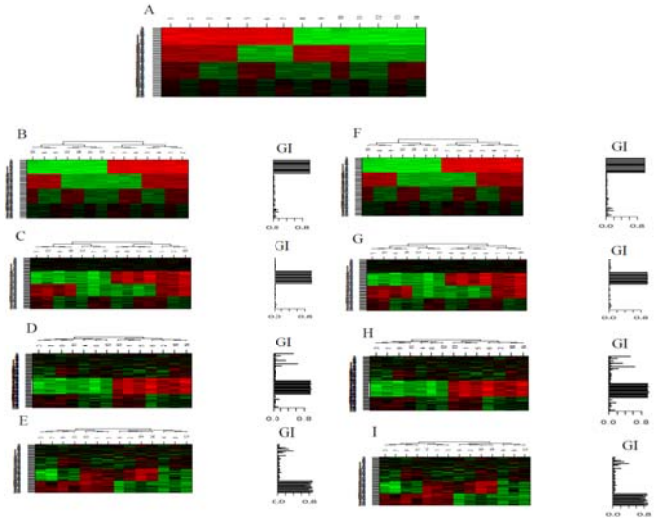

Fig. 2: A comparison between CHC and β-CHC by simulation study in absence of outliers. (A) Artificially generated Data (X) (data generating model is given in Fig.1), (B) HC, (C) CHC (step-1), (D) CHC (step-2), and (E) CHC (step-3) results. (F) RHC, (G) β-CHC (step-1), (H) β-CHC (step-2)and (I) β-CHC (step-3) results.

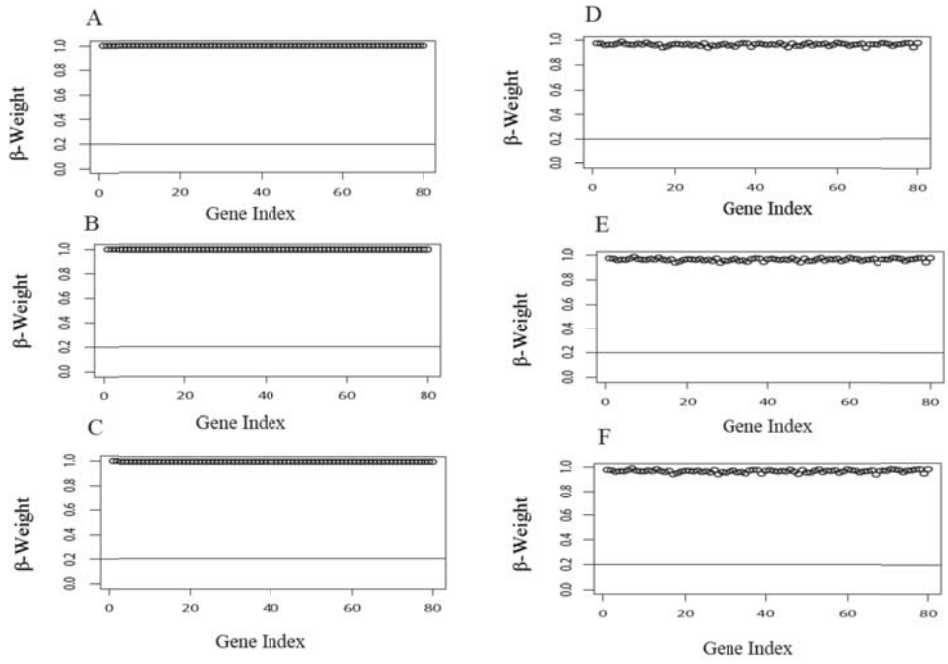

Fig. 3: β-Weight plot for each gene by CHC and β-CHC in absence of outliers.(A) CHC (step-1), (B) CHC (step-2), and (C) CHC (step-3) results.(D) β-CHC (step-1), (E) β-CHC (step-2)and (F) β-CHC (step-3) results.

In absence of outliers both methods classical CHC (β-CHC with β=0) and proposed β-CHC (with β=0.005) are produces same results, there is no genes under the gray line (see **Fig. 3**). This indicates that there are no outlying genes in this data set as shown in **Fig. 2 (A)** (data generating model shown in **Fig. 1**)**.** Finally, we found that the proposed RCHC (with β =0.005) method produces almost the same results as the classical CHC (β-CHC with β =0) in absence of outliers, which is also equivalent to the true / reference results as early mentioned.

4.2 Artificially Generated Gene Expression Data in Presence of Outliers

To examine the robustness of the proposed method in a comparison of the classical method, we added some outliers in the last (81-100) rows of the data matrix $\boldsymbol{X}$. **Fig. 4(A)** shows the data matrix in presence of outliers ($\boldsymbol{X}^*$). Now we would like to recover our previous results as discussed above from this contaminated data set $\boldsymbol{X}^*$. We can perform two methods classical and proposed for sequential exaction of several gene-sets with proper groups of individuals mention above on this contamination data. For classical CHC method first we perform HC as the input of CHC. **Fig. 4 (B)** shows the clustering results of HC. We observed that this result contradict the previous HC results in absence of outliers. Similarly, for explore several gene-sets to avoid HC problem first we can apply classical CHC. **Figs. 4 (B), 4(C), 4(D), 4(E)** show the clustering results of CHC (step-1), CHC (step-2), and CHC (step-3), respectively. We observed those figures; we see that those results are contradicting the previous CHC results in absence of outliers. Which indicates that the classical CHC highly affected by outliers. The classical CHC produces misleading clustering results in presence of outliers. The classical CHC cannot explore several gene-sets with proper groups of individuals as before in absence of outliers.

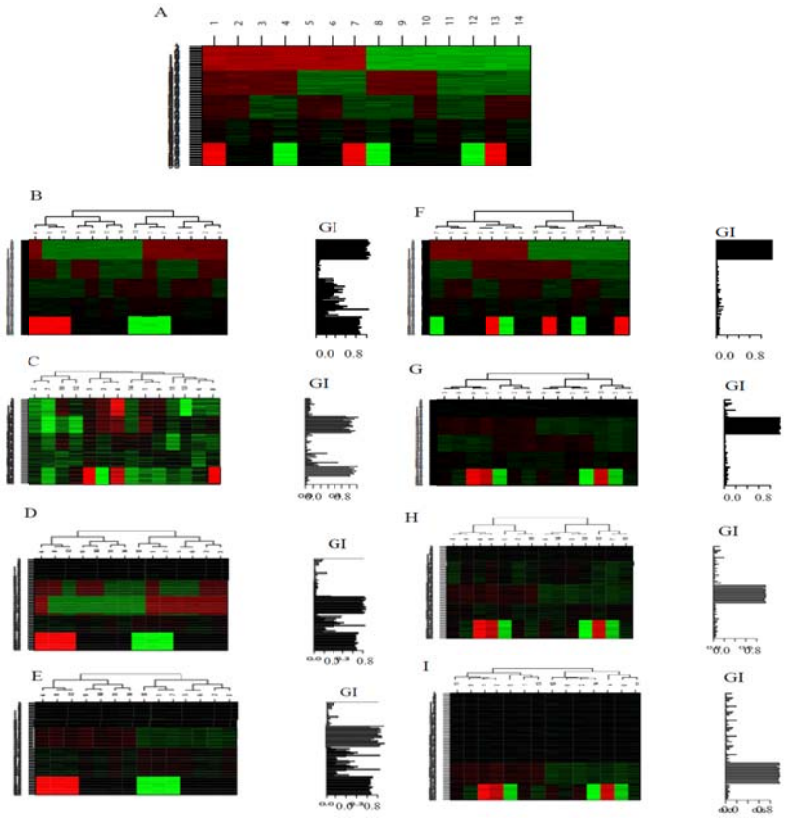

Fig. 4: A comparison between CHC and β-CHC by simulation study in presence of outliers. (A) Outlier Data (X*), (B) HC, (C) CHC (step-1), (D) CHC (step-2), and (E) CHC (step-3) results. (F) RHC, (G) β-CHC (step-1), (H) β-CHC (step-2) and (H) β-CHC (step-3) results.

In this context, we apply proposed method β-CHC for sequential exaction of several gene-sets with proper groups of individuals. First we select the tuning parameter β=0.05 using cross validation as before. **Figs. 4 (F), 4(G), 4(H), 4(I)** show the clustering results of β-CHC (step-1),β-CHC (step-2), and β-CHC (step-3), respectively. We observed these figures; we see that the proposed method correctly explores several gene-sets with proper groups of individuals as before in absence of outliers also. The proposed method are precisely classified the contaminated data without being affected by outliers.

Also observed β-weight plot for each genes in different steps (three steps) in **Fig. 5**.**Figs. 5 (A), (B), (C)** are the β-weight plot for each gene using classical CHC method and **Figs. 5 (D), (E), (F)** are the β-weight plot for each gene using classical proposed method in presence of outliers. The classical CHC (β-CHC with β=0) are produces same weight for each genes and there is no genes under the gray line (see **Fig. 5(A), (B), (C)**). But interesting in **Figs. 5 (D), (E), (F)** we see clear that in presence of outliers for β-CHC (with β=0.05) are produces different weight for different genes. The proposed method is found that 20 (81-100) genes are under the gray line, which indicates that those 20 genes are outlying genes (see **Figs. 5 (D), (E), (F)**). Now we match our results with the given data set (**Fig. 4 (A)**); we confirm that those 20 genes are actually outlying genes and those genes are added in the last row of the data matrix (**Fig. 4 (A)**).

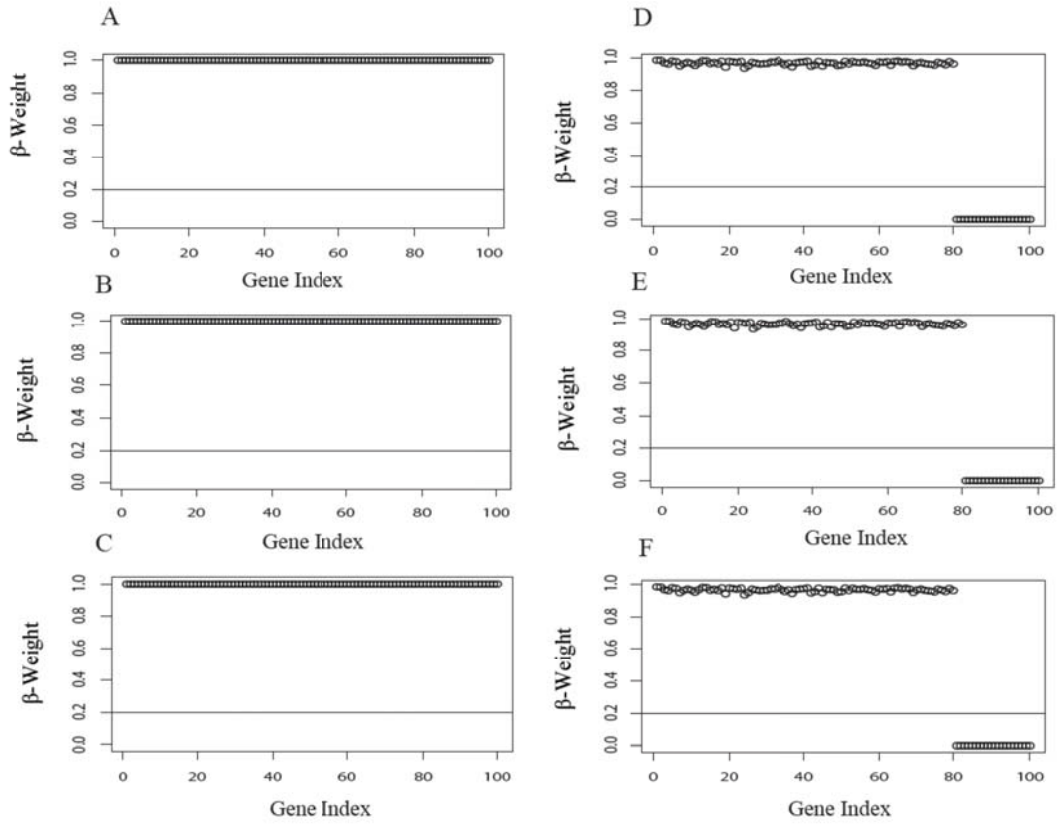

Fig. 5: β-Weight plot for each gene by CHC and β-CHC in presence of outliers.(A) CHC (step-1), (B) CHC (step-2), and (C) CHC (step-3) results.(D) β-CHC (step-1), (E) β-CHC (step-2)and (F) β-CHC (step-3) results.

5. Concluding Remarks

Complementary hierarchical clustering (CHC) important for sequential exaction of several gene-sets having relatively low expressions than highly expressed genes. However it produces misleading clustering results if there exist some contaminations (outliers) in the gene expression data, which is an important issue in gene expression data analysis research field. Therefore, in this paper we developed a robust statistical clustering technique based on the value of tuning parameter β we called β-CHC for sequential extraction of biologically important gene-sets has similar expression patters with proper groups of individuals by minimizing β-divergence for the genes expression data analysis in bioinformatics from the robustness points of view. The proposed robust method reduces to the traditional method when we put the value of tuning parameter $\beta \to 0$. The values of the tuning parameter β play a key role in the performance of the proposed method. It controls the trade off between the robustness and efficiency of the estimators. Smaller β produces more efficient results than larger β, while larger β produces more robust results than smaller β. We selected the tuning parameter β using cross-validation. To investigate the performance of the proposed method in a comparison of the classical approach, we consider synthetic gene expression microarray dataset. Simulation gene expression data clustering results show that the performance of the proposed method is better than performance of the traditional method in the case of data contaminations; otherwise, it shows almost equal performance.

References

1. Allison, D.B., Xiangqin, G.P., and Mahyar, S.: Microarray data analysis from disarray to consolidation and consensus. *Nat. Rev. Genet*., 7:55-65, 2006.

2. Barry T., Nobel, A.B., and Fred A. W.: A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat*., **2**: 286-315, 2008.

3. Datta, S., Satten, G. A., Benos, D. J., Xia, J., Heslin, M. J., and Datta S.: An empirical bayes adjustment to increase the sensitivity of detecting differentially expressed genes in microarray experiments. *Bioinformatics*,**20(2):**235-242, 2004.

4. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* of the United States of America, **95**: 14863–14868, 1998.

5. Gottardo, R., Raftery, A. E., Yeung, K.Y., and Bumgarner, R.E.: Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples. *Biometrics,***62**:10–18, 2006.

6. Ho Y., Leslie, C., Marcel, D., and Giovanni, P.:Statistical methods for identifying differentially expressed gene combinations. In: Ochs MF, editor. Gene Function Analysis, *Methods in Molecular Biology Series,* **408**:171-191, 2007.

7. Minami, M., and Eguchi, S.: Robust blind source separation by beta-divergence. *Neural Computation*, **14**:1859–1886, 2002.

8. Mollah, M. N. H., Minami, M., and Eguchi, S.: Robust Prewhitening for ICA by the Minimum β-Divergence and its application to Fast ICA. *Neural Processing letters*, **25**: 91-110, 2007.

9. Mollah, M. N. H., Mari, P., Komori, O., and Eguchi, S.: Robust Hierarchical Clustering for Gene Expression Data Analysis. *Proceedings of the 2nd international conference on bioinformatics and system biology,* Leipzig, Germany, 2009.

10. Mollah, M. N. H., Sultana, N., Minami, M., and Eguchi, S.: Robust extraction of local structures by the minimum beta-divergence method. *Neural Network,* **23**: 223-238, 2010.

11. Nowak, G., and Tibshirani, R.: Complementary Hierarchical Clustering*, Biostatistics,* **9, 3**: 467-483, 2008.

12. Newton, M.A., Fernando, A. Q., Johan A. B., Srikumar, S., and Paul, A.: Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat*., **1**: 85-106, 2007.

13. Tracy, L. B., and Jason, W.: Proportion statistics to detect differentially expressed genes: a comparison with log-ratio statistics. *BMC Bioinformatics*,**12**:228, 2011.

14. Terry, S.: Statistical Analysis of Gene Expression Microarray Data. *Chapman & Hall,* 2003.