International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 9 • Number 46 • 2016

Hierarchial Kmeans Clustering for Gene Expression in Parkinson's Disease

Ramkumar[a], R.B. Mishra[b], Babita Pandey[c] and Raja Ram Sah[d]

[a]*Department of Computer Science and Engineering, BIT Mesra, Ranchi*
[b,d]*Department of Computer Science and Engineering, IIT BHU, Varanasi*
[c]*Department of Computer Applications, Lovely Professional University, Phagwara, India*

Abstract: Disorders can be identified by gene expression changes that are reactive leading to diseases representing that there exists a sensitive connection of information flow between disease and genes which play a vital role in predicting disease progression or severity by comparing normal and disease affected samples. Further evolution led to determine the genes correlated with disease, causal communication between traits and genes and their co-expression networks which are successfully tested in animal and human models. Differential gene expression analysis for microarray data as a tool and with the use of highly effective computational methods is applied to the Gene expression datasets of Normal and MPTP (neurotoxic leading to Parkinson's Disease) treated mice brains, we are able to replace it with the traditional approaches which exhibit quite inefficient results when dealt with such huge amount of heterogeneous data. In this paper, we expand our research on microarray-based gene expression analysis of the brain in Parkinson's disease by implementing efficient computational methods based on Hierarchical k-means clustering.

Keywords: Parkinson's disease, Gene expression, Clustering, Correlation.

1. INTRODUCTION

Parkinson's disease is a deteriorating disease and idiopathic disease probably due to a complaint in the central nervous system by the death of dopamine-producing cells in the midbrain. The diseases with neuropsychiatric disorders frequently have significant contributions in genetics but identifying the relevant genes is getting difficult due to the complexities of human genetic analyses. In unicellular systems, useful understanding of gene networks has been obtained from high-throughput gene expression methodologies, exemplified by microarrays, gene chips, etc. Keen insights into their pathogenesis could be achieved by the extensive sampling of gene expression patterns. Gene expression is one of the primary solution in studying neuro degenerative disorders [1] like PD detecting the genetics of pathological phenotype and the neurons exhibiting regional vulnerability within the brain. Neuro degeneration and gene expression share a critical link caused due to selective uptake of toxin MPTP (1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) by dopaminergic neurons termed as neurotoxicity[2] leading to death of neurons in the Substantia nigra of brain and finally becomes a primary cause for PD.

T-statistic Distribution to find the optimum threshold using the *p*-values [6] i.e., significance based on differential gene expression values. Finding autocorrelation among the genes of same segments and cross-correlation between the genes across the segments in both normal and MPTP treated mice brains. Hierarchical *k*-means clustering have been used for clustering the genes based on gene expression values.

The crucial pathological feature of Parkinson's disease is the loss of dopaminergic neurons within the SNpc [3]. Neuronal loss in Parkinson's disease occurs in many other brain regions, including the locus ceruleus, nucleus basalis of Meynert, pedunculopontine nucleus, raphe nucleus, and dorsal motor nucleus of the vagus, amygdala, and hypothalamus. Lewy Pathology, which is an aggregation of abnormally folded proteins, has emerged as a common theme in neurodegenerative diseases, including Parkinson's disease. In Parkinson's disease, this protein was identified as α-synuclein following the discovery that mutations in its gene, SNCA, cause a monogenic form of the disease. Our paper has been organized as follows. Section 2 describes the Materials and tools used. Section 3 deals with computational methods. Section 4 deals with Implementation of method. Section contain the results and discussion. Section 6 concludes the paper.

1.1. Problem Description

In the current PD genetics nomenclature, 18 specific chromosomal regions are termed PARK. Six genes have been proposed to mediate autosomal dominant forms of Parkinson's disease [4]: SNCA, LRRK2, VPS35, EIF4G1, DNAJC13, and CHCHD2. Additional genes identified patient cohorts include ATP13A2, C9ORF72, FBXO7, PLA2G6, POLG1, SCA2, SCA3, SYNJ1, RAB39B, and possibly one or more genes affected due to microdeletion syndrome in 22q11.2. Parkin, PINK1, and DJ-1 are associated with autosomal recessive forms of Parkinson's disease. The diseases with neuropsychiatric disorders like Parkinson's disease and Alzheimer's disease frequently have significant contributions in genetics but identifying the relevant genes is getting difficult due to the complexities of human genetic analysis. In unicellular systems, useful understanding of gene networks has been obtained from high-throughput gene expression methodologies, exemplified by microarrays, gene chips, etc. Keen insights into their pathogenesis could be achieved by the extensive sampling of gene expression patterns. So in order to determine the characteristics and to gain information regarding the genetics of these diseases we need an extensive approach powered by efficient computational methods so that processing and extraction of required genetic, pathological information through gene expression analysis from the huge amount of microarray data.

2. MATERIALS AND TOOLS USED

The gene expression datasets [5] of Normal and MPTP treated mice brains is used. The dataset contains expression values and their respective gene id of about 8000 genes in each segment of the brain. The entire dataset can be downloaded from http://www.ncbi.nlm.nih.gov/geo under the series accession GSE30. The algorithm has been implemented in Matlab 8.0.

3. COMPUTATIONAL METHODS-HIERARCHICAL K-MEANS CLUSTERING

Clustering is applied in order to differentiate the abnormal genes, tissues, regions, etc. Clustering can reveal the hidden structures of biological data, and is particularly useful for helping biologists investigate and understand the activities of uncharacterized genes and proteins and further, the systematic architecture of the whole genetic network

1. Determination of gene expression data:
2. Calculation of similarity matrix:
3. Clustering the genes based on the similarity data or gene expression data.

In contrast, to partition based clustering, set of disjoint clusters is produced by the decomposing the data set, hierarchical clustering generates a hierarchical tree form of clusters called as a dendrogram. The branches of dendrogram record the formation of the clusters and similarity between the clusters. An agglomerative algorithm UPGMA (Unweighted Pair Group Method with Arithmetic Mean) developed by Eisen et. al.[7] represented the clustered data set graphically in which fluorescence ratio is measured and used to color the each cell of the gene expression matrix and its rows are re-ordered based on the consistent node-ordering rule and hierarchical dendrogram structure. After clustering, the original gene expression matrix is represented by a colored table (a cluster image) where groups of genes are represented by large contiguous patches of color represent that share similar expression patterns over multiple conditions. Hierarchical clustering groups genes together with similar expression pattern and provides a graphical representation of the dataset.

To characterize the data K-means method uses the centroids of clusters i.e., K prototypes and determined by minimizing the sum of squared errors. Initially, as the number of gene clusters is unknown for a gene expression data set. We test the algorithm repeatedly with different values of 'K'(number of disjoint subsets) to detect the optimal number of clusters. The fine-tuning process is impractical for a large gene expression dataset. Gene expression data typically contain an enormous amount of noise; however, the K-means algorithm forces each gene into a cluster representing its sensitivity to noise [8].

$$\mathrm{J}_k = \sum_{k=1}^{k} \sum_{i \in \mathrm{C}_k} \left(x_i - m_k\right)^2 \tag{1}$$

where $(x_1, ..., x_n) = \mathrm{X}$ is the data matrix and $m_k = \sum_{i \in \mathrm{C}_k} x_t / n_k$ is the centroid of cluster C_k and n_k is the number of points in C_k.

4. IMPLEMENTATION OF HIERARCHICAL K-MEANS CLUSTERING

To explore the relationships between brain regions find the correlations between each of segments concerning gene expression levels. Find most strongly differentially expressed genes between different parts of the normal brain and also in the MPTP-treated brain using *t*-statistic distribution to determine an optimum threshold from the obtained *p*-values which determines the significance of differentially expressed genes by its variance the so that large differences may indicate interesting genes involved in brain development. Brain possesses a high degree of bilateral symmetry. The correlation analyzes are spatial and provide information on the correlation between segments based on gene expression levels. Correlations determine the degree of expression of one gene is correlated with the other genes across the segments.

A subset of the gene expression data of normal and MPTP-treated brains in which genes are differentially expressed was extracted from the all parts of brain with a spatial expression correlation coefficient with certain threshold (say greater than 0.80) so that there exist at least one other gene to identify networks of coregulated genes conserved between normal and MPTP-treated brains. The region strongly expressed suggesting that the genes in particular cluster may be particularly important in specifying this part of the brain implying that although the spatially coregulated gene clusters maintain mutually dependent network within each brain, expression patterns are modified in the MPTP-treated brain compared with the normal brain for all the clusters. Spatial gene expression patterns representing strongly expressed genes are acquired and are shown in Figure 1.

By applying the hierarchical *k*-means clustering we can illustrate that the change in gene clusters due to change in gene expression values can be used to detect the genes related to PD in which Hierarchical clustering yields the structure formation of gene clusters and *k*-means can efficiently optimize the solution acquired from hierarchical clustering by reducing its computational complexity and gene cluster are formed based on the respective significance of the genes and their distances.

Here using dendrogram of resultant hierarchical *k*-means clustering, we can visualize both the segment clustering (column clustering) and the gene clustering (row clustering) as shown in Figure 2 representing normal brain and Figure 4 representing MPTP treated mice brain. The numbers of genes belonging to each cluster of normal and MPTP-treated brains are tabulated in Table 1 and Table 2 respectively. We can easily visualize that the change in gene expression level due to PD when compared with normal condition lead to change in gene clusters. Here we have the provision of dynamically selecting the clusters and its size using threshold or level cut in hierarchical *k*-means dendrogram. We can easily differentiate the genes by directly comparing the gene clusters of normal brain and MPTP treated mice brain

5. RESULTS AND DISCUSSION

Figure 1 shows Spatial expression patterns of the genes for the normal and MPTP treated mice brains. Based on the calculated optimum threshold of significance of genes by T-static distribution and correlation analysis we are able to plot the strongly differentially expressed individual genes or genes with similar significance considered as regions. The relative level of expression of any gene in any segment is read by looking along the relevant row and column, finding the intersection, and referring to the scales. The segment numbering in the columns of the matrix correspond to segments. The two clusters of genes are apparent, and although these have highly conserved patterns of expression within the normal and PD brains, these patterns are somewhat divergent between the two brains.

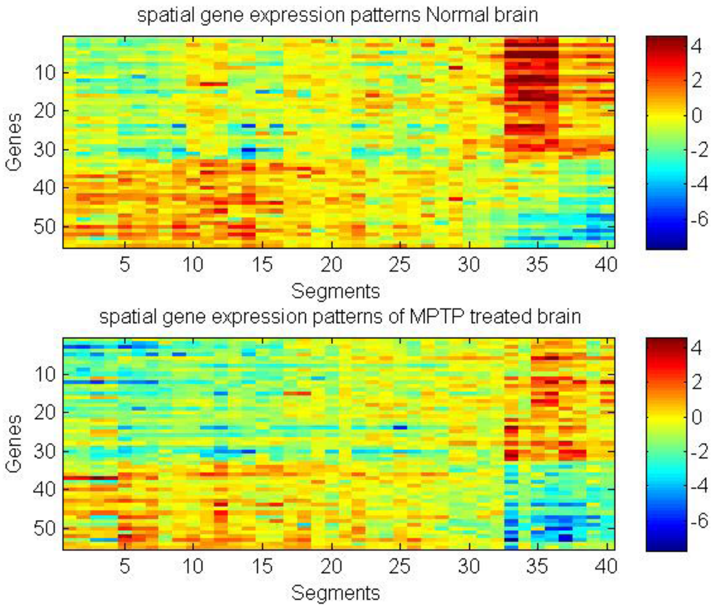

Figure 1: Spatial Gene Expression Patterns of normal and MPTP treated brains

Figure 2 shows Hierarchical K-means Dendrogram of normal Brain with abscissa representing the segments and ordinates represent the genes. We can visualize both segment clustering (columns) and genes clustering (rows) by looking relevant segments or genes linkage to form a cluster and are highlighted with different colors. In the plot, intensity of color represents the differential expression of genes (green as low and red as high).

Table 1 shows Segments clustering acquired from hierarchical *k*-means clustering represented in dendrogram of normal brain in Figure 2 and Table 2 shows number of genes in each cluster (when assumed for 16 clusters and 32 clusters) of normal brain using Hierarchical clustering.

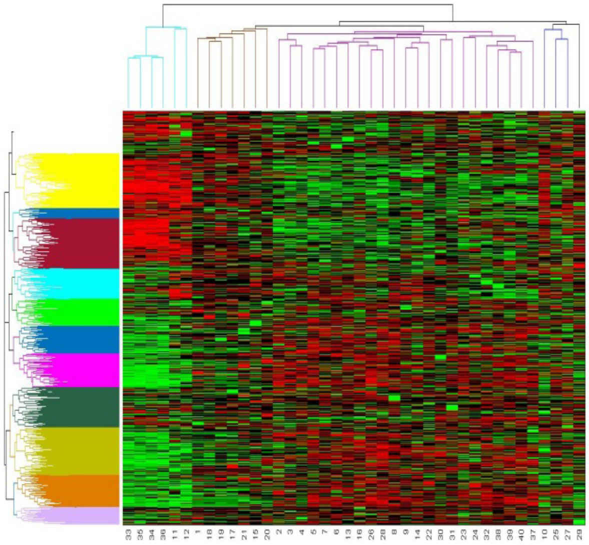

Figure 2: Dendrogram obtained from Hierarchical kmeans clustering of normal brain

Table 1
Segments clustering of Normal brain

Cluster Index	*Number of elements in cluster*	*Segments*
1	4	33, 35, 34, 36
2	2	11, 12
3	7	1,18,19,17,21,15,20
4	23	2,3,4,5,7,6,13,16,26,28,8,9,14,22,30,31,23,24,32, 38,39,40,37
5	4	10,25,27,29

Table 2
(a) Number of genes in each cluster (*n* = 16)

Cluster Index	*Number of elements in cluster*
1	141
2	3299
3	89
4	4422
5	106
6	70
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1

Table 2
(b) Number of genes in each cluster (*n* = 32)

Cluster Index	*Number of elements in Cluster*
1	61
2	3172
3	48
4	299
5	44
6	111
7	2654
8	53
9	55
10	71
11	281
12	128
13	57
14	268
15	158
16	127
17	73
18	61
19	141
20	89
21	106
22	70
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1

Table 3 shows segments clustering acquired from hierarchical *k*-means clustering represented in dendrogram of MPTP-treated mice brain. Table 4 shows number of genes in each cluster (when assumed for 16 clusters and 32 clusters) of MPTP-treated brain using Hierarchical clustering.

From Table 2, Cluster 4 with highest number of genes of 4422 for cluster size 16 in normal brain while with the cluster size 32 Cluster 2has highest number of genes of about 3272.From table 4, Cluster 2 with highest number of genes of 2272 for cluster size 16 in MPTP treated brain while with the cluster size 32 Cluster 9 has highest number of genes of about 1903.

Here by the keen observation we can easily determine the gene clusters based on Differential gene expression using hierarchical gene clustering which in turn lead to analysis of genes that cause PD by comparing the result obtained from the normal brain and MPTP treated brain gene expression data.

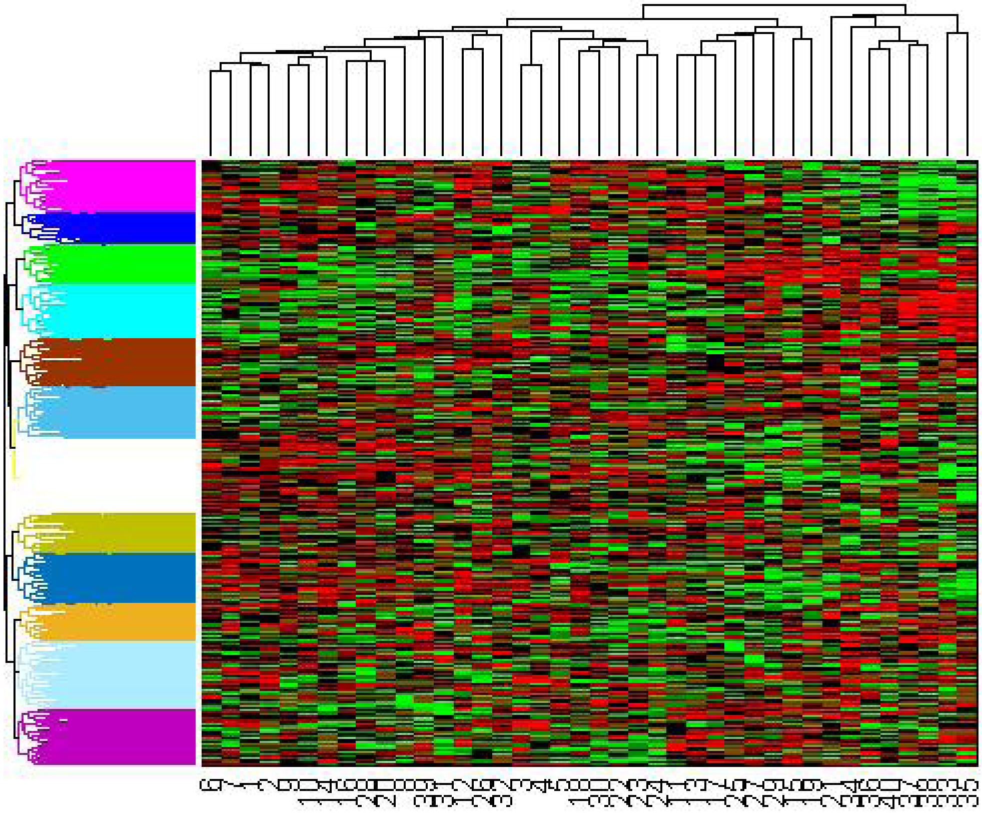

Figure 3: Hierarchical K-means Dendrogram of MPTP treated Brain

Figure 3 shows the Hierarchical K-means Dendrogram of MPTP treated Brain with abscissa representing the segments and ordinates represent the genes. We can visualize both segment clustering (columns) and genes clustering (rows) by looking relevant segments or genes linkage to form a cluster and are highlighted with different colors. In the plot, intensity of color represents the differential expression of genes (green as low and red as high).

Table 3
Segments clustering of MPTP treated brain

Cluster Index	*Number of elements in cluster*	*Segments*
1	16	6,7,1,2,9,10,14,16,28,20,8,39,31,12,26
2	8	32,3,4,5,18,30,22,23,24
3	8	11,13,17,25,27,29,15,19
4	6	21,34,36,40,37,38
5	2	33,35

Here we are processing the differential gene expression data by selecting the cluster size of 16 and 32 to fit the genes in respective cluster as tabulated in Table 2 and table 4 for normal and MPTP treated brains respectively. Here we can visualize the segments clustering as tabulated in Table 2 and Table 4 for normal and MPTP treated brains respectively. In order to increase the accuracy of clustering we need to decide the optimum number of clusters. If the cluster size is made small then we may lose data related to the genes with low entropy. If the cluster size is made too large then it may lead to addition of genetic noise or redundant data. Here we can eliminate this problem by inducing iteration and replications methods to the hierarchical *k*-means algorithm.

Table 4
(a) Number of genes in clusters (*n* = 16)

Cluster Index	*Number of elements in cluster*
1	144
2	2272
3	72
4	1941
5	146
6	372
7	46
8	94
9	328
10	147
11	1557
12	191
13	372
14	90
15	240
16	124

Table 4
(b) Number of genes in clusters (*n* = 32)

Cluster Index	*Number of elements in Cluster*
1	80
2	248
3	121
4	1652
5	71
6	1351
7	102
8	1903
9	127
10	126
11	96
12	94
13	97
14	49
15	75
16	178
17	194
18	105
19	79
20	99

Cluster Index	*Number of elements in Cluster*
21	48
22	140
23	56
24	63
25	151
26	144
27	72
28	146
29	46
30	94
31	90
32	240

6. CONCLUSION

Here by using Hierarchical *k*-means clustering we can illustrate that change in gene expression values lead to change in gene clusters and by the efficient preprocessing techniques to find the significance of the genes by deciding optimum threshold can improve the efficiency of clustering there by improving the identification of genes that are related to Parkinson's disease. In future further improvement can be achieved by deploying clustering algorithms like SOM and feature reduction techniques like Principal Component Analysis, Subset selection and Classification algorithms like Support Vector Machine to increase the accuracy of prediction.

REFERENCES

[1] Francesco Camastra, Maria Donata Di Taranto and Antonino Staiano, *"Statistical and Computational Methods for Genetic Diseases": An Overview*, Hindawi Publishing Corporation, Article ID :954598, 2015.

[2] Langston J.W., Ballard P., Tetrud J.W., and Irwin I. "Chronic parkinsonism in humans due to a product of meperidine analog synthesis". Science, 219:979-980, 1983.

[3] Patrick A. Lewis, Mark R. Cookson, "Gene expression in the Parkinson's disease brain", Brain Research Bulletin 88, pp:302– 312, 2012

[4] Christine Klein, Ana Westenberger, "*Genetics of Parkinson's Disease*", Cold Spring Harb Perspect Med, doi: 10.1101/cshperspect.a008888, PMCID: PMC3253033, 2012.

[5] Vanessa M. Brown, Alex Ossadtchi, Arshad H. Khan, Simon Yee, Goran Lacan, 1William P. Melega, Simon R. Cherry, Richard M. Leahy and Desmond J. Smith. "Multiplex Three-Dimensional Brain Gene Expression Mapping in a Mouse Model of Parkinson's Disease", Cold Spring Harbor Laboratory (ISSN 1088-9051/01), 2002.

[6] Tsai C-A, Chen Y-J, Chen JJ. "Testing for differentially expressed genes with microarray data". Nucleic Acids Research. 2003;31(9):e52.

[7] Eisen, Michael B., Spellman, Paul T., Brown, Patrick O. and Botstein, David. "Cluster analysis and display of genome-wide expression patterns". Proc. Natl. Acad. Sci. *USA*, 5(25):14863–14868, December 1998.

[8] Sherlock G. "Analysis of large-scale gene expression data". Curr Opin Immunol, 12(2):201–205, 2000.