Role of Permutations in Significance Analysis of Microarray and Clustering of Significant Microarray Gene list

Ms. Tejashree Damle[1] and Dr. Manali Kshirsagar[2]

[1] Department of Computer Technology, Yeshwantrao Chavan College of Engineering Nagpur, Maharashtra,India-441110

[2] Department of Computer Technology, Yeshwantrao Chavan College of Engineering Nagpur, Maharashtra,India-441110

Abstract

Microarray is the gene expression data that represent gene in different biological states. Methods are needed to determine the significance of these changes while accounting for the enormous number of genes. Significance analysis of microarrays (SAM) is a statistical technique for determining whether changes in gene expression are statistically significant. During the SAM procedure permutation of microarray data is considered to observe the changes in the overall expression level of data. With increasing number of permutations false discovery rate for gene set varies.

In our work we took microarray data of Normal Glucose Tolerance (NGT), and Diabetes Mellitus (DM Type II). In this paper we proposed the result of permutations during execution of SAM algorithm. The hierarchical clustering is applied for observing expression levels of significant data and visualize it with heat map.

Keywords: *Significance analysis of Microarray, False discovery Rate, Clustering*

1. Introduction

Significance Analysis of Microarray identifies genes with statistically significant changes in expression. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are deemed potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR). To estimate the FDR, nonsense genes are identified by analyzing permutations of the measurements. The threshold can be adjusted to identify smaller or larger sets of genes, and FDRs are calculated for each set.[1]

False discovery rate (FDR) control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. False discovery rate (FDR) is defined as the expected percentage of false positives among all the claimed positives. In practice, with the true FDR unknown, an estimated FDR can serve as a criterion to evaluate the performance of various statistical methods under the condition that the estimated FDR approximates the true FDR well, or at least, it does not improperly favor or disfavor any particular method. Permutation methods have become popular to estimate FDR in genomic studies.

SAM identifies statistically significant genes by computing a statistic dj for each gene j, which measures the strength of the relationship between gene expression and a response variable. This analysis uses non-parametric statistics, since the data may not follow a normal distribution. The response variable describes and groups the data based on experimental conditions. In this method, repeated permutations of the data are used to determine if the expression of any gene is significant related to the response. The use of permutation-based analysis accounts for correlations in genes and avoids parametric assumptions about the distribution of individual genes. [2]

In statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. In hierarchical clustering, clusters are defined as branches of a cluster tree. In this paper we had shown the hierarchical clustering for significant gene list.

2. Data Source

In this we take data from Mootha VK *et al.* (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are co-ordinately down regulated in human diabetes. *Nature Genetics*; Vol 34(3); 267-273. The disease model is Diabetes mellitus (Type II). The study involved 34 males, 17 with normal glucose tolerance (NGT), and 17 with Diabetes Mellitus (DM Type II).[7]

3. Permutations of Microarray Data

We used the permutation method is used to evaluate the change in FDR. Initially we had taken 50

permutations and calculated FDR value for different values .Figure shows the graph between Delta and FDR values and Table 1 shows the result of FDR values after 50 permutations.

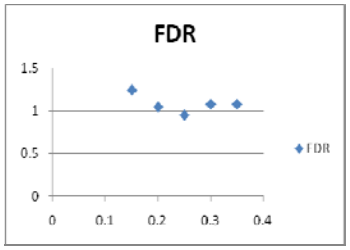

Figure 1: Scatter plot for FDR for 50 permutations

Table 1:FDR for 50 permutations

Delta	FDR
0.15	1.245826377
0.2	1.045813586
0.25	0.94939759
0.3	1.078313253
0.35	1.078616352

Similarly, Figure 2 through Figure 4 show the scatter plot between FDR and Delta after 100, 150 and 200 permutations respectively. Table 2 through Table 4 shows FDR values after 100, 150 and 200 permutations respectively.

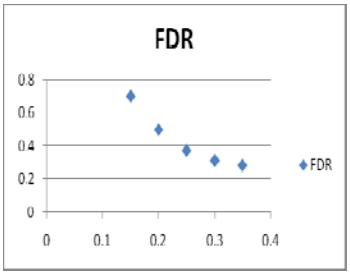

Figure 2: Scatter plot for FDR for 100 permutations

Table 2:FDR for 100 permutations

Delta	FDR
0.15	0.699464524
0.2	0.497350993
0.25	0.371975806
0.3	0.310810810
0.35	0.282178217

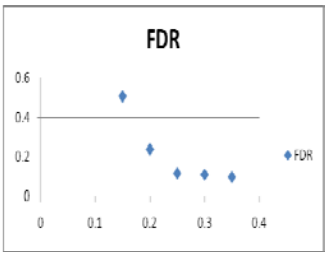

Figure 3: Scatter plot for FDR for 150 permutations

Table 3:FDR for 150 permutations

Delta	FDR
0.15	0.506134969
0.2	0.238993710
0.25	0.116071428
0.3	0.109589041
0.35	0.097285067

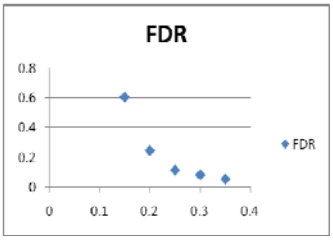

Figure 4: Scatter plot for FDR for 150 permutations

Table 4:FDR for 150 permutations

Delta	FDR
0.15	0.606500290 1
0.2	0.246981339
0.25	0.114081996
0.3	0.081920903
0.35	0.054621848

From this data set it has been observed that as the number of permutations increases the value of FDR decreases. Therefore, for getting correct FDR values permutation of microarray data set should be as maximum as possible. But, it can be said that after some permutations FDR has no change or very slight change. So, at this stage we can terminate permutations. For our dataset we fixed total 200 permutation and got FDR as 5% for delta=0.35. We obtained total 238 significant genes out of which 217 are falsely positive and 21 are falsely negative.

4. Hierarchical Clustering of Significant Gene List:

Hierarchical clustering arranges items in a hierarchy with a treelike structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram. In Spotfire, hierarchical clustering and dendrograms are strongly connected to heat map visualizations. We can cluster both rows and columns in the heat map. Row dendrograms show the distance or similarity between rows, and which nodes each row belongs to as a result of clustering. Hierarchical clustering can be visualize with heat map. The easiest way to understand a heat map is to think of a table or spreadsheet which contains colors instead of numbers. Heat maps are well-suited for visualizing large amounts of multi-dimensional data and can be used to identify clusters of rows with similar values, as these are displayed as areas of similar colour. A dendrogram is a tree-structured graph used in heat maps to visualize the result of a hierarchical clustering calculation. The result of a clustering is presented either as the distance or the similarity between the clustered rows or columns depending on the selected distance measure. [8]

The algorithm used for hierarchical clustering in Spotfire is a hierarchical agglomerative method. For row clustering, the cluster analysis begins with each row placed in a separate cluster. Then the distance between all possible combinations of two rows is calculated using a selected distance measure. The two most similar clusters are then grouped together and form a new cluster. In subsequent steps, the distance between the new cluster and all remaining clusters is recalculated using a selected clustering method. The number of clusters is thereby reduced by one in each iteration step. Eventually, all rows are grouped into one large cluster. The order of the rows in a dendrogram are defined by the selected ordering weight. The cluster analysis works the same way for column clustering.[8]

Figure 5 shows the heat map for the gene list that is obtained after 200 permutations. The heat map clearly shows the differentiation between NGT and DM Type II with status as up regulated and down regulated genes. The highlighted part shows the down regulated gene i.e. whose expression is down regulated in DM Type II, while rest shows up regulated genes.

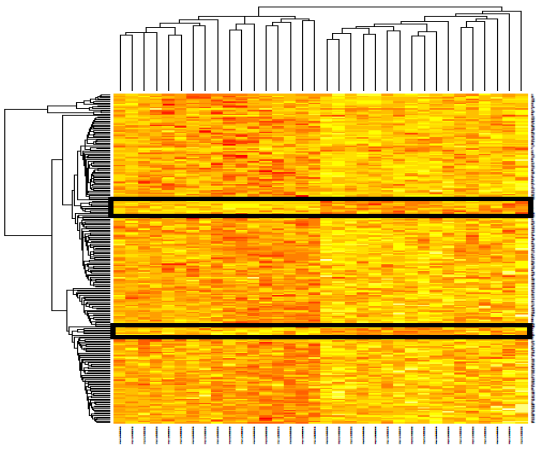

Figure 5: Heatmap for GeneSet

5.Conclusion

In our work, Microarray data is normalized to adjust the intensity levels . This data is used as an input and Significance analysis of Microarray algorithm is applied. During the implementation of this algorithm , number of permutations are applied on data to calculated false discovery rate. It has been found that FDR is more accurate as the number of permutations are increases.

Final significant gene list contains 238 genes ,out of which 217 are upregulated and 21 are downregulated genes.

References

[1] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu," Significance analysis of microarrays applied to the ionizing radiation response",PNAS, April 24, 2001, vol. 98, no. 9

[2] Gil Chu ,Jun Li y Balasubramanian Narasimhan , Robert Tibshirani ,Virginia Tusher, "Significance Analysis of Microarrays" Users guide and technical document, Department of Biochemistry, Stanford University, Stanford CA 94305,2002.

[3]Saravanakumar Selvaraj, Jeyakumar Natarajan," Microarray Data Analysis and Mining Tools" , Bioinformation 6(3): 95-99 (2011)

[4] Julia Sivriver1,*Naomi Habib and Nir Friedman,"* An integrative clustering and modeling algorithm for dynamical gene expression data*"*, Oxford University Press 2011.

[5]Peter Langfelder, Bin Zhang,Steve Horvath," Data Mining of Microarray Databases for the Analysis of Environmental Factors on Plants Using Cluster Analysis and Predictive Regression", Oxford University Press, Vol. 24 no. 5 2008, pages 719–720.

[6] Iiris Hovatta, Juha Saharinen, Katja Kimppa,, M. Minna Laine, Antti Lehmussola," DNA microarray data analysis", CSC – Scientific Computing Ltd.2005

[7] Vamsi K Mootha,Cecilia M Lindgren, Kerl-Fredrik Eriksson,Aravind Subramanian…..,"PGC-1α-responsive genes involved in oxidative phosphorylation are co-ordinately down regulated in human diabetes" *Nature Genetics*, *Nature Genetics*; Vol 34(3); 267-273,2003

[8] http://spotfire.tibco.com