Gene Expression Analysis for Type-2 Diabetes Mellitus –A Case Study on Healthy vs Diabetes with Parental History

Chandra Sekhar Vasamsetty, *IACSIT Member*, Srinivasa Rao Peri, Allam Appa Rao, K. Srinivas, and Chinta Someswararao, *IACSIT Member*

***Abstract*—Both environmental and genetic factors have roles in the development of any disease. The quest for an understanding of how genetic factors contribute to human disease is gathering speed. Differential gene expression analysis plays an important role for the study of genetic factors causing diseases. We proposed a method for identifying differentially expressed genes causing Type-2 diabetes mellitus using microarray data for diabetes with parental history and healthy. This method focuses on identifying multivariate and univariate outliers using Mahalanobis Distance, Minimum Co-variance Determinant (MCD) and other statistical methods. For the identified inflammatory genes we performed the functional classification by using Gene Ontology and identified the pathways between these inflammatory genes using pathway analysis. This method is applied on microarray data from two samples one from diabetes with parental history and the other from healthy and identified 1579 genes which are differentially expressed and functional classification was preformed to these genes. Prior to analysis, the microarray data is normalized using Lowess Normalization method.**

***Index Terms*—Inflammatory genes, Gene ontology, Pathway analysis, Outliers, Mahalanobis distance, Type-2 diabetes mellitus.**

I. Introduction

Molecular Biology research evolves through the development of the technologies used for carrying them out. It is not possible to research on a large number of genes using traditional methods. DNA Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner [1]. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body [2].

Manuscript received January 25, 2011; revised April 23, 2011.

F. A. Author is with the Department of CSE, S R K R Engineering College, Bhimavaram, India (email: chandu.vasamsetty@gmail.com). S. B. Author is with the Professor in CS&SE, AU College of Engg.,,Visakhapatnam-530003,AndhraPradesh,India.

T. C. Author is with the Vice Chancellor, JNT University Kakinada, AP, India.

F. D. Author is with the 4Professor, Department of CSE, VR Siddartha College of Engg., Vijayawada, AP, India.

F. E. Author is with the Department of CSE, S R K R Engineering College, Bhimavaram, India (email: chinta.someswararao@gmail.com).

A typical microarray experiment involves the hybridization of an mRNA molecule to the the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each on the array indicates the expression level of the various genes.

A. Microarray Technique

An array is an orderly arrangement of samples where matching of known and unknown DNA samples is done based on base pairing rules. An array experiment makes use of common assay systems such as micro plates or standard blotting membranes. The sample spot sizes are typically less than 200 microns in diameter usually contain thousands of spots.

Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support (a microscope glass slides or silicon chips or nylon membrane). The spots can be DNA, cDNA, or oligonucleotides. These are used to determine complementary binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery. An experiment with a single DNA chip can provide information on thousands of genes simultaneously. An orderly arrangement of the probes on the support is important as the location of each spot on the array is used for the identification of a gene [3, 4].

B. Types of Microarrays:

Depending upon the kind of immobilized sample used construct arrays and the information fetched, the Microarray experiments can be categorized in three ways:

Microarray expression *analysis***:** In this experimental setup, the cDNA derived from the mRNA of known genes is immobilized. The sample has genes from both the normal as well as the diseased tissues. Spots with more intensity are obtained for diseased tissue gene if the gene is over expressed in the diseased condition. This expression pattern is then compared to the expression pattern of a gene responsible for a disease.

Microarray for mutation analysis: For this analysis, the researchers use gDNA. The genes might differ from each other by as less as a single nucleotide base. A single base difference between two sequences is known as Single Nucleotide Polymorphism (SNP) and detecting them is known as SNP detection Comparative *Genomic Hybridization***:** It is used for the identification in the increase or decrease of the important chromosomal fragments harbouring genes involved in a disease.

In this paper we used Microarray expression analysis for

identifying and classifying genes causing Type 2 Diabetes Mellitus (T2DM). The prevalence of T2DM is rising worldwide. While environmental factors, such obesity and lack of physical activity, play an important role to the rapid increase in the prevalence of T2DM, genetic factors are also important for the increased risk of T2DM. Studies have estimated that risk for diagnosed T2DM increases approximately two- to fourfold when one or both parents are affected.

A lot of studies have showed an excess paternal transmission of T2DM in different populations. Genetic factors, such as mitochondrial DNA mutations, and environmental mechanisms, such as intrauterine environment, have been proposed for the explanation of the excess paternal transmission of T2DM. In the present study, we performed gene expression profile in subjects with type 2 diabetes mellitus with parental history Versus Healthy using micro array data.

Searching for all of the available information about each gene of interest is very time consuming. This is hampered further by the wide variations in terminology. Gene Ontology (GO) is a collection of controlled vocabularies describing the biology of a gene product in any organism

There are 3 independent sets of vocabularies, or ontologies:

Molecular Function (MF)

e.g. ”DNA binding” and ”catalytic activity”

Cellular Component (CC)

e.g. ”organelle membrane” and ”cytoskeleton”

Biological Process (BP)

e.g.“DNA replication” and ”response to stimulus”

C. Normalization

In carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. In this study we will consider Lowess Normalization method for normalization.

The global locally weighted scattered plot smoothing (LOWESS) normalization is a good choice because it provides a good balance on the following three factors: ideally the center of the distribution of log-ratios should be zero, the log-ratios should be independent of spot intensity, and the fitted line should be parallel to intensity axis. It has been reported that the log2(ratio) values can have a systematic dependence on intensity, which most commonly appears as a deviation from zero for low-intensity spots. Locally weighted linear regression (LOWESS) analysis has been proposed as a normalization method that can remove such intensity-dependent effects in the log2(ratio) values (see M/A plots below). The easiest way to visualize intensity-dependent effects is to plot the measured log2(red/green) ratio or (M) for each element on the array as a function of the log2(red*green) product intensities or (A). LOWESS method detects systematic deviation in the “ratio-intensity” plot and corrects them by carrying out a local weighted linear regression as a function of the log2(intensity) and subtracting the calculated best-fit average log2(ratio) from the experimentally observed ratio for each data-point.

II. Literature Review

Several methods are available in the Literature to perform Differential Gene Expression Analysis to find the potential Genes causing various Diseases.

M. Kathleen Kerr et.al demonstrated [5] that ANOVA methods can be used to normalize microarray data and provide estimates of changes in gene expression that are corrected for potential confounding effects. This approach establishes a frame work for the general analysis and interpretation of micro array data.

The probability that a false identification is committed can increase sharply when the number of tested genes gets large. Correlation between the test statistics attributed to gene co-regulation and dependency in the measurement errors of the gene expression levels further complicates this problem. Anat Reiner et.al addressed [6] this problem by adapting the False Discovery Rate (FDR) controlling approach. Comparative analysis shows that all the four FDR controlling procedures control the FDR at the desired level.

D. L. Wilson et.al presented [7] two methods for the normalization of the micro array data to remove biases towards one or the other fluorescent dyes used to label each mRNA sample allowing for proper evaluation of differential gene expression. One method deals with smooth spatial trends in intensity across micro arrays. Second method deals with normalization of a new type of cDNA micro array experiment where large proportion of the genes on the microarrays is expected to be highly differentially expressed.

Hong-Ya Zhao et.al applied [8] a multivariate mixture model to model the expression level of replicated arrays, considering the differentially expressed genes as the outliers of the expression data. In order to detect the outliers of the multivariate mixture model, a statistical method based on the analysis of Kurtosis Coefficient (KC) is applied to the micro array data. They used the RT-PCR method and two statistical methods, Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) to verify the expression levels of outlier genes identified by KC algorithm.

Dan Nettleton et.al developed [9] a non parametric multivariate method for identifying gene categories whose multivariate expression distribution differs across two or more conditions. By comparing the performance to several existing procedures via the analysis of a real data set and showed that this method has good power for differentiating between differentially expressed and non- differentially expressed gene categories.

Huaizhen Qin et.al proposed [10] a computationally simple method for finding differentially expressed genes in small micro array experiments. This method incorporates a novel stratification based tight clustering algorithm, principal component analysis and information pooling. They applied this method to three real micro array data sets. Comprehensive simulation shows that this method is substantially powerful than the popular SAM and eBayes approaches.

Bogdan Done et al proposed [11] a technique that improves previous method for predicting novel GO annotations by extracting implicit semantic relationships between genes and functions. In this work, they use a vector space model and a number of weighting schemes. The

technique described is able to take into consideration the hierarchical structure of the Gene Ontology (GO) and can weight differently GO terms situated at different depths.

Purvesh Khatri et al proposed [12] an impact analysis approach that considers crucial biological factors to analyze regulatory pathways at systems biology level. This approach calculates perturbations induced by each gene in a pathway, and propagates them through the entire pathway to compute an impact factor for the given pathway. They proposed an alternative approach that uses a linear system to compute the impact factor. Their proposed approach eliminates the possible stability problems when the perturbations are propagated through a pathway that contains positive feedback loops. Additionally, the proposed approach is able to consider the type of genes when calculating the impact factors.

Monica chagoyen et al proposed [13] non-negative Independent Component Analysis (nnICA) for the classification of genes based on their associated functional annotations.

III. Analysis Performed

Data from three samples were hybridized on Human 40 K OchiChip Array. Gene expression values were obtained after quantification of TIFF images. Data has 40,320 X 3 data-points (or probes). Empty spots and control probes were removed before proceeding with data analysis.

A. Differential Expression Analysis

In any micro array study the primary objective is to assess MRNA transcript levels of samples under different experimental conditions. Which of the thousands of genes show significant difference in expression levels in the samples is the question of importance. Appropriate statistical techniques are required to furnish the accurate information on differentially expressed genes if there are no or limited replicates due to practical constraints in majority of the experiments.

For experiments with single sample in different conditions, we assume that the log intensity values of gene expression for the two samples are linearly related, following bivariate normal distribution, contaminated with outliers. In a contaminated bivariate distribution, the main body of the data is characterized by bivariate normal distribution and constitutes regular observations. The non-regular observations, described as outliers, represent systematic deviations. These outliers are often suspected as possible candidates for differential expression genes.

Here we use an exploratory approach consisting of two-stages to detect outliers from bivariate population and determining differentially expressed candidates from these outliers. The approach provides the fold-change value considering the scatter of observations and thereby provides up and down regulated genes across the samples.

B. Functional Classification

To determine biological significance of differentially expressed genes, functional classification was performed.

1) Gene Ontology

GO provides a dynamic controlled vocabulary and hierarchy that unifies descriptions of biological, cellular and molecular functions across genomes.

2) Pathway Analysis

To determine pathways associated with differentially expressed genes, pathway analysis was performed.

IV. Materials and Methods

A. Stage- I: Multivariate Outlier Detection:

Outlier detection is one of the important tasks in any data analysis, which describe abnormalities in the data. Many methods have been proposed in the literature for detecting univariate outliers based on robust estimation of location and scale parameters. The standard method for multivariate outlier detection involves robust estimation of parameters in the *Mahalanobis Distance* (MD) measure and then comparing MD with the critical value of c2 distribution. The values larger than the critical value are treated as outliers of the distribution.

1) Mehalanobis Distance:

The covariance matrix is used for the quantification of the size and shape of the multivariate data, which is taken into account in the Mehalanobis distance. For a multivariate sample Xij, where i = 1,2,3,...n (number of genes) and j = 1,2,3...p (number of samples), the Mahalanobis distance is defined as,

$$MD_i = ((X_{ij} - m)^T C^{-1} (X_{ij} - m))^{0.5}$$

where m is estimated multivariate location parameter and C is the estimated covariance matrix. The location and the covariance parameters are determined using Minimum Covariance Determinant estimation method. The MCD estimator is determined by that subset of observations of size h, which minimizes the determinant of the covariance matrix computed only from the h observations. The location estimator is the average of these h observations, whereas the scatter estimate is proportional to the variance covariance matrix.

B. Stage-II: Univariate Outlier detection:

Let S denote the original set of observations.

Let S_{out} and S_{in} be the subsets of S containing outlier and inlier observations respectively. Thus, $S_{out} \cup S_{in} = S$ and $S_{out} \cap S_{in} = \{\emptyset\}$, i.e. the two subsets are mutually exclusive. We denote:

$S_{out} = \{(\log_2(X_{i1}), \log_2(X_{i2})) / MD_i > c \text{ for } i=1,2,3...n\}$ and

$S_{in} = \{(\log_2(X_{i1}), \log_2(X_{i2})) / MD_i < c \text{ for } i=1,2,3...n\}$

where 'c' is the cut-off for a given quantile and n is the total number of genes.

We define a statistic:

$$Z = \log_2 (X_2 / X_1) = \log_2(X_2) - \log_2(X_1)$$

which is the log of the ratio of intensity values for different genes for the two samples. Here X1 is treated as reference, while X2 is treated as test sample. The statistic provides a measure of differential expression (DE) of genes across the samples. The genes showing at least k-fold change (usually k=2, i.e. Z=1) across the samples are considered to be DE genes. The appropriate choice of k is important since it influences the number of DE genes. Here we propose a rationale for selecting k for a given percentage of bivariate outliers.

We generate values for the statistic for the entire set as,

$$Z = \{ z_i, i = 1,2,3,\ldots n \}$$
$$= \{ \log_2(X_{i2} / X_{i1}); \ i = 1,2,3,\ldots n\}$$

The statistic is used to obtain Mahalanobis distance measure as,

$$MD_i^* = \left(\frac{Zi - m}{Se} \right)^2 \text{ for } i = 1,2,3,\ldots n$$

The transformed distance measure is supposed to follow chi-square distribution with one degree of freedom. The empirical distribution function of MD* could be obtained and compared with that of the cumulative distribution of chi-square with one degree of freedom. A cut-off could be selected for MD* such that the observations greater than the cut-off could be declared as outliers. We search for an optimal cut-off, so that the univariate subset of outliers does not include any of the bivariate inliers. In other words, if Rout is a subset of univariate outliers and Sin the subset of bivariate inliers of S, then the optimal cut-off could be obtained as,

$$C^*_{opt} = \inf [C_i^* / R_{out} \cap S_{in} = \{\varnothing\}]$$

The optimal cut-off could be obtained programmatically thereby yielding a set of univariate outliers that overlap with a subset of multivariate outliers.

The cut-off value could be used in Mehalanobis distance measure to obtain the z-value as,

$$Z = (S_e) \sqrt{C^*_{opt}} + m$$

This z-value determines the log fold change resulting into bivariate outliers that could be the potential candidates for differential expression.

V. Application

In the present context, there are two individuals, one from each of the categories namely diabetes with parental history (D&PH) and healthy (H). The expression levels of 39400 genes for each individual were obtained and compared pair wise. Prior to analysis, the data for each combination was normalized using Lowess normalization.

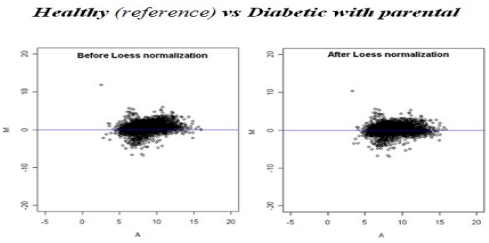

Figure 1 shows MA-plots showing scatter of expression values before and after Lowess normalization for healthy

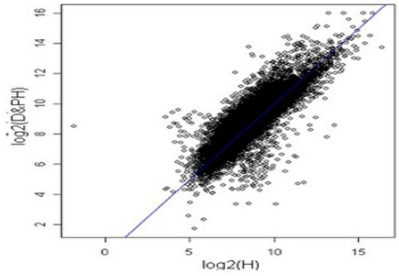

Figure 2 shows Scatter plot of log intensities for healthy vs. diabetic with parental history comparison after Lowess normalization.

The distribution of log fold change values was obtained and the outliers were detected for the optimum cut-off value (c*). Figure 1.4 shows the thresholds for 2.36-fold change, thereby providing the up and down regulated genes. Out of 3940 outlier genes, 1211 were detected as up-regulated, while 368 were detected as down-regulated genes with respect to the healthy (H) individual. Thus, for healthy vs. diabetic with parental history comparison, 1579 genes were found to be differentially expressed out of 39400, which amounts to 4% of the total genes under study. This is 2.73% less than the number of genes obtained for 2-fold change thresholds.

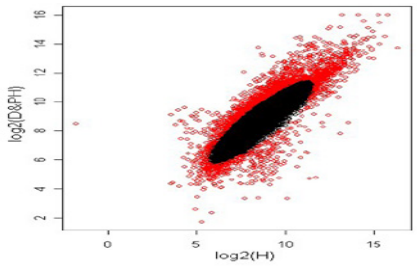

Figure 3 shows Bivariate outliers based on Mahalanobis distance measure for p=0.10 for healthy vs. diabetic with parental history comparison

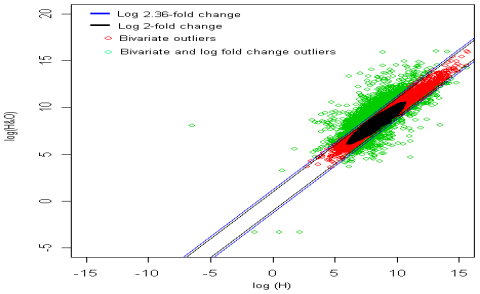

Figure 4: The thresholds for 2.36 and 2 fold change values. The green spots are the differentially expressed outlier genes for healthy vs. healthy with obesity comparison.

VI. Results

TABLE1.GENES INVOLOVED IN INFLAMMATORY RESPONSE

Condition	Inflammatory genes (Differentially expressed)
Diabetes with family history vs healthy individual **(D&PH vs H)**	*ALK, GCH1, IFIH1, IFIT1, IL11RA, ITGB2, MAP3K4, MMP19, MMP3, RPS27A, SLK, TNFRSF12A, UBC*

A. Gene Ontology Analysis

Molecular Function: Genes involved in NADH dehydrogenase(ubiquinone) activity, glutamate dehydrogenase[NAD(P)+]activity,CDP-diacylglycerol-glycerol-3-phosphate-3-phosphtidyltransferase activity are upregulated in D&PH with respect to H. Gene involved in protein kinase B binding, enzyme inhibitor activity, acyl-CoA oxidase activity, phosphatidylinositol transporter activity, acyltransferase activity are downregulated in D&PH with respect to H.

Biological Process**:** Genes involved in synaptic vesicle membrane organization and biogenesis, polysaccharide metabolic process, regulation of growth rate, nucleosome assembly are upregulated in D&PH with respect to H. Genes

involved in immune response, regulation of glycolysis are downregulated in D&PH with respect to H.

Cellular Component: Genes localized in cohesin core heterodimer, oligosaccharyl transferase complex, nucleosome, respiratory chain complex II are upregulated in D&PH with respect to H. Genes localized in isoamylase complex, protein kinase CK2 complex, proteasome activator complex, 6-phosphofructokinase complex are downregulated in D&PH with respect to H.

B. Pathway Analysis

Genes involved in Inositol phosphate metabolism, Starch and sucrose metabolism, Nitrogen metabolism, Oxidative phosphorylation, Androgen and estrogen metabolism, Glycan biosynthesis and metabolism pathways, Metabolism of cofactors and vitamins pathways, MAPK signalling pathway, ECM-receptor interaction, Neuroactive ligand-receptor interaction, Regulation of actin cytoskeleton, Cell communication pathways, Nervous system pathways, Neurodegenerative disorders pathways are upregulated in D&PH Vs H. Genes involved in Glycolysis / Gluconeogenesis, Propanoate metabolism, Carbon fixation, Biosynthesis of steroids, Fatty acid metabolism, Histidine metabolism, Phenylalanine metabolism, Tyrosine metabolism, Urea cycle and metabolism of amino groups, Cell cycle, Insulin signalling pathway, PPAR signaling pathway, Antigen processing and presentation are downregulated in D&PH Vs H.

VII. CONCLUSIONS

Gene Expression Analysis was performed for two subjects on Healthy Vs. Diabetic with Parental History by using the statistical methods *Mehalanobis Distance, Minimum Covariance Determinant* and *Lowess Normalization* method to find out univariate and multivariate outliers using microarray Data. It was found that 1579 differentially expressed genes were identified out of 39400 genes tested between healthy vs. diabetic with parental history. For these differentially expressed genes, we perform molecular function, Bio-logical process and cellular component analysis was performed using Gene Ontology and also performed the pathway analysis to find out the pathways associated with these differentially expressed genes.

ACKNOWLEDGMENT

The authors acknowledge Ocimum Biosolutions, Hyderabad, India for analysis of the microarray data using their proprietary microarray analysis tool Genowiz.

REFERENCES

[1] John Ten Bosch, Chris Seidel, Sajeev Batra, Hugh Lam, Nico Tuason, Sepp Saljoughi, and Robert Saul,"Validation of Sequence-Optimized 70 Base Oligonucleotides for Use on DNA Microarrays", *OPERON a QIAGEN COMPANY,*2000.

[2] Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ., "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays", Department of Molecular Biology and Genomics and Department of Infectious Diseases, *Pfizer Global Research and Development, Ann Arbor, MI* 48105, USA,2000.

[3] Dr. Susanne Schröder1, Dr. Jaqueline Weber2, and Dr. Hubert Paul1 MWG Biotech AG ," 50 nucleotide long probes on microarrays enable high signal intensity and high specificity.*" microarray Development1and Department of Bioinformatics* 2, Anzinger Str. 7, 85560 Ebersberg, Germany,2000.

[4] Angela Relogio, Christian Aschwager, Alexandra Ritcher, Wilhelm Ansorge and Juan Valcarel.,"Optimization of Oligonucleotide - based DNA microarrays",2000.

[5] M. Kathleen Kerr, Mitchell Martin, and Gary A Churchill," Analysis of Variance for Gene Expression Microarray Data", *Journal of Computational Biology,* Vol.7, No.6,pp.819-837 ,2000 .

[6] Anat Reiner, et.al. "Identifying differentially expressed genes using false discovery rate controlling procedures", Bioinformatics-Oxford University press, vol.19, No.3,pp.368-375,2003.

[7] D.L. Wilson, et.al. "New Normalization methods for cDNA microarray data", Bioinformatics-Oxford University press, vol.19, No.11, pp.1325-1332,2003

[8] Hong-Ya Zaho, et.al,"Identification of Differentially Expressed Genes with Multivariate Outlier Analysis", *Journal of Biopharmaceutical Statistics*, vol. 14, Issue 3, pp.629-646,2004.

[9] Dan Nettleton, Justin Recknor and James M. Reecy," Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis", vol. 24, No. 2, pp.192-201,2008.

[10] Huaizhen Qin, Tao Feng, et.al, "An efficient method to identify differentially expressed genes in microarray experiements, " Oxford University press, vol. 24 no. 14 , pages 723-729,2008.

[11] Bogdan Done, Purvesh Khatri, Arina Done, Sorin Draghici,"Detriotpredicting Novel Human Gene Ontology Annotations Using Semantic Analysis" *IEEE/ACM Transactions On Computational Biology And Bioinformatics*,2010.

[12] Purvesh Khatri,Sorin Draghici,Adi L. Tarca,Sonia S. Hassan,Roberto Romero,"A system biology approach for the steady-state analysis of gene signaling networks", CIARP'07 Proceedings of the Congress on pattern recognition 12th Iberoamerican conference on Progress in pattern recognition, image analysis and applications,2007.

[13] Monica chagoyen, hugo fernandes, jose m. Carazo and alberto pascual-montano,"Functional Classification Of Genes Using Non-Negative Independent Component Analysis " *mathematics in industry,*volume12,pp.571-57,2008.