Methods used for Identification of Differentially Expressing Genes (DEGs) from Microarray Gene Dataset: A Review

Chanda Panse(Wajgi)#[1], Manali Kshirsagar[#2], Dipak Wajgi[*3]
[#]*Department of Computer Technology, YCCE, Nagpur, India 441110*
[*]*Department of Computer Engineering, SVPCE, Nagpur, India 441108*

Abstract

Genes contain blue print of living organism. Malfunctioning occurred in cellular life is indicated by proteins which are responsible for behavior of genes. Fixed set of genes decides behavior and functioning of cells. They guide the cells what to do and when to do. To analyze the insight of biological activities, analysis of gene expressions is necessary. Advanced technology like microarray plays an important role in gene analysis. It captures expressions of thousands of genes under different conditions simultaneously. Out of thousands of genes, very few behave differently which are called as Differentially Expressed Genes (DEGs). Identification of these most significant genes is a crucial task in molecular biology and is a major area of research for bioinformaticians because DEGs are the major source of disease prediction. They help in planning therapeutic strategies for a disease through Gene Regulatory Network (GRN) which is constructed from them. GRN is a graphical representation containing genes as nodes and regulatory interactions between them as edges. GRN helps in knowing how genes regulate each other and in sense maintain underlined state of art working of cells. Deregulation between genes is the cause of major genetic diseases. In this paper we have discussed many methods proposed by researchers for identifying differentially expressing genes based upon changes in their expressions patterns.

Keywords — *Gene Regulatory Network (GRN), gene expressions, differentially expressed gene.*

I. INTRODUCTION

Microarray technology monitors abundant amount of gene data in terms of gene expressions. Gene expressions are stored in terms of two dimensional matrix of size n X m where n indicates no of genes and m indicates no of samples. Since gene expressions are stochastic in nature it is necessary to observe changes in them over time. Understanding these changes and obtaining useful information from them is a major challenge in bioinformatics. Single gene chip contains of thousands of gene expressions. Before doing any downstream analysis like clustering or modeling of GRN on gene expressions, it is necessary to identify most significant genes. These genes called marker genes. Sudden changes in their expressions help to understand what goes wrong and where. They also help in classifying different types of tumors and act as starting point for studying certain systems like cell cycle, drug response etc. They also help in identifying abnormalities in biological activities. Thus assist in detection of disease and its diagnosis at early stage.

In this paper we have critically reviewed existing methods used for identification of differentially expressed genes along with their advantages and disadvantages. In this context, it has been found in the literature that, methods which were applied to static data were also extended for time course expressions [1]. However statistical validations of genes were not done as methods were originally designed for static data. Methods used for finding significant genes are classified into two categories i.e. parametric methods and non-parametric methods. Parametric class includes those methods which have fixed set of parameters based on model used whereas non-parametric methods do not have fixed set of parameters.

In case of non-parametric methods, parameters vary based on type of data. At early stages, significant genes are found using fold-change method which is parametric method where ratio of log of any two samples are taken for deciding differentially expressing genes but later it is proved to be inadequate due to dynamic variations in gene expressions [5]. To deal with this, many parametric statistical methods are used by researchers which are discussed in this section. Some of the most commonly used parametric statistical methods are ANOVA [2], RM-ANOVA [3], t-test [4], SAM [5], Empirical Bayesian method [6] etc. which are used for any type of datasets. Non-parametric method includes Wilcoxon rank sum test, hypothesis testing etc. Mostly parametric methods are used in the literature for finding significant genes. Tusher in [5] extended a method developed for static data known as SAM for time series data but couldn't validate it. Therefore we have classified these methods based on nature of microarray dataset used. Microarray datasets considered in the literature, for experimentation fall into two main categories: Replicated Microarray Dataset (RMD) and Non-

Replicated Microarray Dataset (NRMD). If in case gene expressions are not recorded correctly, they are captured more than one time to improve accuracy in case of RMD. On other hand in NRMD gene expressions are captured only once. This eventually sometimes leads to generation of missing value in gene expressions. This needs employment of imputation methods before further analysis. Some of the most commonly used methods for identification of DEGs are Fold Change (FC), ANOVA [2], RM-ANOVA [3], t-test [4], SAM [5], Empirical Bayesian method [6] etc. This review helps researchers to design more general algorithm for identifying DEGs apart from statistical methods.

II. TYPES OF MICROARRAY DATASET

Microarray technology is a powerful tool which enables researchers to investigate and address issues in molecular biology by analysing gene expressions of thousands of gene in single reaction and in an efficient manner. A typical microarray chip preparation involves hybridization of an mRNA molecule to DNA template from which it is originated. An array is constructed from many DNA molecules. The amount of mRNA bound to each site on the array indicates the expression levels of various genes.

Before actually going into review of existing methods, first we will see short background of what is replicated time series and static microarray slide. There is a possibility that gene expressions are not consistently captured because of the factors like dust and scratches on glass slides, poor illumination. In order to have reliable and invariable gene expressions, experiment is replicated that is same set of experiment is repeated for several times. The author in paper [7] states the importance of replication in misclassification of genes. Replication is not a duplication of experiment. When microarray data from several replications are combined, there is reduction in false positive and false negative rate. Time series microarray captures multiple gene expressions at discrete time points (minutes, hours or days) whereas static microarray slide contains gene expressions for differential conditions only ones. As data is captured for different experimental conditions, it is not possible to have mathematical relationship between conditions. Therefore it is needed to have several replicates for such conditions. In the next section we will have brief review of methods applied on replicated time series data.

III. METHODS FOR REPLICATED MICROARRAY DATASET

Based on the limitations of existing methods described in [8][9] which would require to know DEGs in advanced, author in paper[10] compared 3 empirical Bayes methods i.e CyberT, BRB and Limma t-test on synthetic replicated gene expressions. It is found that CyberT maintains fixed False Positive Rate for data with unstable variance across intensities and BRB and Limma t-test also generate many false positive based on the pre-processing used on data without decreasing True Positive Rate(TPR). In order to extract useful biological knowledge from large microarray data, multivariate data analysis is needed as it reduces dimensions. Therefore Principal Component Space based algorithm is proposed in [11] where replicated microarray dataset of Tomato is used for finding DEGs. Genes which are close to a particular condition are considered as significant genes. If there exists strong relationship between gene and a condition then it is significantly expressing in that condition. Closeness measure is denoted by Cd which is given by

$$C_{\vec{d}_i}(\vec{v}) = I_{\propto \vec{d}_i}(\vec{v}) sign(\vec{v}, \vec{d}_i) \frac{\vec{v}.\vec{d}_i}{\vec{d}_i.\vec{d}_i} \vec{v}$$ Where

$I_{\propto \vec{d}_i}$ is a function which indicates whether gene belongs to direction di. Cd can have three values, 0, 1 and -1. 1 and -1 indicate strong relation between gene and specific condition and 0 indicates non-expressing gene in particular condition [11]. As replication of microarray chip is costly, very few methods exist for finding DEGs.

IV. METHODS FOR NON-REPLICATED MICROARRAY DATASET

In paper [12] author has adapted method from text categorization and information retrieval literature for classifying genes into diseased and normal category by identifying their discriminating capability. Author has decided a threshold value t which will classify a gene into one of the classes' i.e diseased or normal by observing gene expressions in both classes and discriminating V score is calculated as follows. Genes having highest value of V are the most discriminating genes.

$$V = (a + d) - (b + c)$$

Where a,b,c,d are the count of gene expression values of gene g having values greater than equal to or less than equal to threshold t. But some time it is very difficult to identify threshold value t because of stochastic nature of gene expressions. As microarray contains noise and lack of normal pattern of gene expressions, Olga [13] compared three model-free non-parametric methods i.e. t-test, rank-sum test and heuristic method based on high Pearson co-relation coefficient. Out of these three methods Wilcoxon Rank Sum Test performs better than others. But TPR rate depends on noise in the dataset and p-value selected. Another replication-free method is proposed in [14] based on significant temporal variation exhibited by genes in estrous cycle of rat mammary gland. The algorithm is designed in such a way that it will fit data on B-splins curve. But the algorithm is only applicable for time dependent biological processes such as cell cycle, circadian

rhythms, development patterns, hormonal fluctuations where gene expressions vary with time. Again the experiment was conducted on simulated data.

As clustering helps in finding biologically co-expressing genes, a method was discussed for non-replicated datasets in[15] based on pareto-optimal clustering using supervised learning in which genes are assigned to a cluster using SVM. But prior to clustering there is need of identification of significant genes which was done manually by observing changes in expressions of genes.

In order to observe the periodic nature in microarray gene expressions, average periodigram is used in [16]. It is a tool used in time series analysis for observing peaks in time domain. Upon noticing peaks, their significance is validated using g-test of Fisher [17] followed by calculation of p-value for each test and gene which reject null hypothesis is considered as DEG. But for the application of g-test minimum 40 measurements per gene are desirable and as p-value for each gene is calculated, the method is time consuming.

For biological processes which are periodic in nature, Fast Fourier Transform based method is proposed in [18] whose performance is affected by small time series experiments and missing values.

By observing the behavior of genes in microarray, author Ping Ma in [19] proposed a technique in which the time series gene chip is classified into two classes of genes. Genes which do not interact with time are kept in PDE (Parallel Differentially Expressing) class and those with time interaction are deposited in NPDE (Non-Parallel Differentially Expressing) class. After classification a functional ANOVA mixed-effect model is applied for identifying NPDE genes. For some cases of datasets, the method had identified NPDE genes as PDE. As above methods are examining individual gene, time required for the analysis is more. To overcome this, a new Functional Principal Component (FPC) approach is developed in [20] which observe the temporal trajectories of gene expressions which are modeled using few basic functions which require lesser number of parameters as compared to earlier methods. This method is suffering with increase in false positive rate. Another method based on Fourier transform is proposed in [21]. In this method genes which are not differentially expressed are filtered out on the basis of Fourier coefficients. As number of genes gets reduced model-based clustering is applied on them. But disadvantage of this method is that for each gene it is needed to calculate Fourier coefficient. Another approach is proposed for identifying DEGs in non-replicated time-course data by inculcating FPCA into hypothesis testing framework [22]. Other method includes a geometrical approach for identification of DEGs in which co-relation between genes is considered. It is multivariate approach. It reframes linear classification method to identify DEGs by defining a separate hyper-plane, the orientation of which decides DEGs. Comparative analysis of statistical methods is done in [23] along with various tools and packages used for identifying DEGs mentioned in it. It is observed that none of the statistical method is the best. The choice of method depends on the type of dataset selected for the experimentation. In Fold change method a log ratio of expressions under two conditions is taken. Its drawback is that statistical variance is not considered in [24]. FC method is subject to bias if the data have not been properly normalized. The danger of false positive and false negative is illustrated in [25] by Tanaka after strictly considering only FC. The shortcomings of FC are removed by t-test by considering variance. Two sample t-statistics is calculated as follows.

$$t = \frac{\overline{X_1} - \overline{X_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where s is sample variance and n_1 and n_2 are number of observations in each condition. But small sample size is the major obstacle in it[26]. Again it is observed that variance is not same within each group. Therefore Welch proposed t-test for unequal variance [27] by correcting degree of freedom for unequal variance as follows.

$$v = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \Bigg/ \frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}$$

Based on the limitation of small sample size, Wilcoxon rank sum test has used as alternative method for testing differential expressions

Many of these methods are based on hypothesis testing. Author in paper [32] applied hypothesis test on each gene to determine as to whether it falls under the category of significant or non-significant gene by observing its variance in population average versus time curve. Apart from this Angelini [33] have used Bayesian approach to identify highly expressed genes. Selection of method depends on nature of gene expressions i.e. Static and Dynamic. In static type, gene expressions are measured at single time point for multiple subjects whereas in dynamic type, multiple time points are used for capturing gene expressions.

While searching for the significant genes, it is quite possible that redundant genes get extracted from the dataset. To avoid this PSO based approach is

proposed in[28] which will identify non-redundant disease related genes. In this paper, multi-objective function is used which will optimize multiple goals i.e minimization of sensitivity and minimization of specificity. Redundancy is avoided in this approach by selecting the solution which gives mutually maximally dissimilar genes. As size of solution is (n+n*s), where n is number of genes and s is number of samples, time required is more and author has just given 2 or 3 biomarker genes for different cancer related datasets.

As replication of time series microarray expressions is costly and it is not possible to fit irregularly varying gene expressions into predefined models, a general approach is proposed in[30]. It is based on Partial Energy ratio for Microarray (PEM). PEM statistic is incorporated into the permutation based SAM framework for significance analysis. This method suffers with low samples as signal smoothing is not possible for dataset consisting of fewer samples. Again a general method is proposed to overcome the drawback of modelling methods in [PNAS]. This method uses cubic splines which are set of cubic polynomials used for fitting time series and noisy data and each gene is represented by spline but noise which makes the process computationally intensive. A Ranking and Combination based method is described in [34] which ranks genes based on feature such as variance, deviation, correlation and probability. After sorting genes rank-wise, combination is done.

Thus we have reviewed many methods for replicated and non-replicated datasets. Table I shows results of all the reviewed methods in terms of count of significantly identified genes in particular dataset on which that method is applied. In PSO-based method, author has confirmed 6,1,2 and 5 genes as significant genes for various cancer types. Total count of genes is not known for ANOVA and Wilcoxon test. It is observed that on an average 10%-15% of the genes can have discriminating behaviour and hence are differentially expressed.

TABLE I. REVIEW TABLE SHOWING COUNT OF SIGNIFICANT GENES IDENTIFIED BY VARIOUS METHODS

Method	**Dataset**	**Actual count of genes**	**Count of significant genes**
t-test [35]	Prostate cancer	27575	9985
SAM [5]	Human lymphoidblastoid cell lines	6800	180
ANOVA [2]	Human Liver and skeletal smaples(male) Placenta(female)	-	1274(male) 1886(female)
RM-ANOVA[3]	Breast cancer	1900	55
Empirical Bayes [6]	Lever cancer	6810	127
Wilcoxon(RST) [13]	Lung tumor	-	91
PEM [PEM]	Yeast cell cycle	800	104
PSO-based [28]	Prostate cancer Diffuse B-cell lymphoma ALL (child-ALL) CML	12533 7070 8280 12625	6 1 2 5
Fourier Transform [29]	Yeast cell cycle	4489	2227
FPCA [20]	C. Elegance	2430	1982
B-spline[14]	Estrous cycle	21044	871

IV. CONCLUSION AND FUTURE SCOPE

Thus we have reviewed statistical as well as non-statistical methods used for finding DEGs. Identification of differentially expressing genes is the main area of research in Bioinformatics. Though many methods exist in the literature but none of them is the best. Model-based methods are not suitable for any type of datasets because it is not feasible to fit stochastic gene expressions into those models. In these model-based methods computational cost increases as number of genes increases because number of mathematical expressions increases with respect to count of genes. Also many statistical methods are not discussing about missing values before finding DEGs. Curve based methods like b-splines fit individual gene expressions into curve. This results in increasing computation overhead and time requirement.

Many tools have been developed based on statistical methods. These include SAM, ANOVA,t-test etc. Since genes expressions do not have specific pattern, a generalized algorithm is needed which will find significant genes and one has to explore the comparison between statistical and non-statistical method by considering same dataset. Also methods for which same genes are observed as significant will be compared on other factors like time and computational cost.

REFERENCES

[1] Tusher V.G., Tibshirani R., and Chu G, "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," Proceeding National Academy of Sciences USA, vol. 98, 2001

[2] Kerr,M.K., Martin,M. and Churchill,G.A. "Analysis of variance for gene expression microarray data", Journal of Computational. Biology., 7,2000.

[3] Ola EiBakry,M.Omair Ahmad and M.N.S. Swamy, "Identification of Differentially Expressed Genes for Time-Course Microarray Data Based on Modified RM ANOVA", IEEE/ACM Transactions on Computational Biology and Bioinformatics,vol.9, 2012

[4] Thomas JG, Olson JM, Tapscott SJ, Zhao LP, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles", Genome Research vol.11,2001

[5] Tusher VG, Tibshirani R, Chu G, "Significance analysis of microarrays applied to the ionizing radiation response", Proc Natl Acad Sci USA, vol.98, 2001

[6] Efron B, Tibshirani R, Gross V, Tusher V G, "Empirical Bayes analysis of a microarray experiment", Journal of American Statistic Association, vol. 96,2001

[7] Lee ML, Kuo FC,Whitemore GA and Sklar J. "Importance of replication in microarray studies: statistical methods and evidence from repetitive cDNA hybridization", Proceeding National Academic Science,USA, vol.97,2000

[8] Qin LX, Kerr KF, "Empirical evaluation of data transformations and ranking statistics for microarray analysis", Nucleic Acids Res, vol.32, 2004

[9] Sioson AA, Mane SP, Li P, Sha W, Heath LS, Bohnert HJ, Grene R, "The statistics of identifying differentially expressed genes in Expresso and TM4: a comparison", BMC Bioinformatics vol.7,2006

[10] Carl Murie,Owen Woody,Anna Y Lee and Robert Nadon, "Comparison of small n statistical tests of differential expression applied to microarrays". BMC Bioinformatics,vol.10,2009

[11] Luis Ospina and Liliana Lopez-Kleine , "Indentification of differentially expressed genes in microarray data in a principal component space", SpringerPlus, vol.2,2013

[12] Hisham Al-Mubaid and Noushin Ghaffari, "Identifying the Most Significant genes from Gene expression Profiles for Sample Classification", Proceeding ,IEEE conference on Granular Computation,2006

[13] Olga G. Toyanskaya, Mitchell E. Garber, Patrick O. Brown, David Botstein and Russ B. Altman, "Nonparametric methods for identifying differentially expressed genes in microarray", Bioinofrmatics, vol.18,2002

[14] Stephen C Billupus, Margaret C Neville, Michael Rudolph, Weston Porter and Pepper Schedin, "Identifying significant temporal variation in time curse microarray data without replicated", BMC Bioinformatics, vol.10,2009

[15] Ujjwal Maulik, Anirban Mukhopadhyay,Sangmitra Bandopadhyay, "Combining Pareto-optimal clustering using supervised learning for identifying co-expressed genes", BMC Bioinformatics, vol.10,2009

[16] Sofia Wichert, Konstantinos Fokianos and Korbinian Strimmer, "Indentifying periodically expressed transcripts in microarray time series data", Bioinformatics, vol.20,2004

[17] Fisher R.A. "Tests of significance in harmonic analysis", Proceeding Royal Society Publishing, vol.125,1929

[18] Jerry Chen, and Paul Paolini, "Fourier Analysis of Time Course Microarray data and its Relevance to Gene Expressions Dynamics", Proceeding ACSESS , 2008

[19] 19. Ping Ma, Wenxuan Zhong, Jun S. Liu, "Identifying Differentially Expressed Genes in time Course Microarray Data", Statistic in Bioscience, vol,1. 2009

[20] Chen Kun, Wang, Jane-Ling, "Identifying Differentially Expressed Genes for Time-course Microarray Data through Functional Data Analysis", Statistic in biosciences,vol.2,2010

[21] Jaehee Kim, Robert Todd Ogden and Haseong Kim, " A method to identify differential expression profiles of time-course gene data with Fourier transform", BMC Bioinformatics, col.14,2013

[22] Shuang Wu, Hulin Wu, "More powerful significant testing for time course gene expression data using functional principal component analysis approaches", BMC Bioinformatics, vol.14, 2013

[23] J. Sreekumar and K.K. Jose, "Statistical tests for identification of differentially expressed genes in cDNA microarray experiments", Indian Journal of Biotechnology, vol.7,no.10,2008

[24] Biju J., Anuparna S and Govindswami K, "Microarray - chipping in genomics", Indian Journal of Biotechnology ,vol.1,2002

[25] Tanaka T.S.,Jaradat S.A., Lim M.K., Kargul G.J.,Wang X., "Genome-wide expression profiling and mid-gestation placenta and embriyo using a 15000 mouse development cDNA microarray", Proceeding National Academy of science USA, vol.97, 2002

[26] Devore J. And Peck R. "Satistics: The exploration and analysis of data" 3rd edition, Duxury Press, Pacific Grove,CA,1997

[27] Welch B.I., "The significance of the difference between two means when population means are unequal",Biometrika, vol.29,1938

[28] Anirban Mukhopadhyay and Monalisa Mandal, "Identifying Non-Redundant Gene Markers from Microarray Data: A Multiobjective variable Length PSO-Based Approach", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.11, 2014

[29] Jaehee Kim, Robert Todd Ogden and Haseong Kim, "A method to identify differential expression profiles of time-course gene data with Fourier transform", BMC Bioinformatics, vol.14,no.310,2013

[30] Xu Han, " PEM: A General Statistical Approach for Identifying Differentially Expressed Genes in the Time-Course cDNA Microarray Experiment Withour Replicate", Journal of Bioinformatics and Computational Biology,2006

[31] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K. Gifford and Tommi Jaakkola, "Comparing the Continuous Representation of time-seriess expressions profiles to identify Differentially expressed genes", PNAS, vol.100,no.18,2003

[32] J.D. Storey,W.Xiao, J.T. Leek, R.G.Tompkins, and R.W. Davis, "Significance Analysis of Time Course Microarray Experiments", Proceeding National Academy of science USA, vol. 102,2005

[33] C. Angelini, D. De Canditiis, M. Mutarelli and M. Pensky, "A Bayesian approach to estimation and testing in time-course microarray Experiments", Statistical application in Genetics and Molecular Biology,vol.6,2007

[34] Han-Yu Chuang, Hongfang Liu, Stuart Brown, Cameron McMunn-Coffran, "Identifying Significant Genes from Microarray Data", Fourth IEEE Symposium on Bioinformatics and Bioengineering, 2004

[35] Khalid Raza and Rajni Jaiswal, "Reconstruction and Analysis of Cancer-specific Gene Regulatory Networks from Gene Expression Profiles", International Journal on Bioinformatics & Biosciences, Vol. 3, No. 2, pp. 25-34, 201