**A Feature Transformation Approach on Selected Features based on Differential Expression Analysis and Unsupervised Classification Using SOM**

J. Briso Becky Bell1, S. Maria Celestin Vigila2

1Research Scholar,Department of Information Technology, Kings Engineering College, Chennai, Tamil Nadu, India.

2Associate Professor,Department of Information Technology, Noorul Islam Centre for Higher Education,

Kumaracoil, Tamil Nadu, India.

E-mail:[1brisobell@gmail.com](mailto:1brisobell@gmail.com), [2celesleon@yahoo.com](mailto:2celesleon@yahoo.com)

**Abstract:- In the modern field of genomics there has been a lot of immense awareness in research, due to the formation of new and advanced data banks. These data banks act as a centre of all medical research biological views, which are functional and actively updated by various boards of bio-medical research council (NCBI). These data warehouses host a lot of multi-dimensional databases readily available for active analysis, algorithmic problem solving and efficient optimization to gain higher information. In this paper, it uses gene disease datasets to be acted on by differential gene analysis approaches such as Log2mean and negative binomial model to extract the most differential genes. Thus obtained genes undergo a feature transformation approachsuch as PCA to obtain highly significant genes. Then the data sets are learned using unsupervised learning based classification approach by using SOM based classifier to classify samples to give an efficient accuracy in predicting the categorical disease type samples. Thus this gives a comparison analysis of classification with feature transformation and classification without feature transformation. Similarly feature selection based classification is compared with transformed feature based classification. It also uses a hybrid approach of using feature transformation on selected genes and the categorical class samples while classification gives efficient accuracy on classification. The classification onL2M-NBIN based PCA transformed features shows better efficient accuracy than normal transformed PCA genes and selected L2M-NBIN based classification approaches.**

**Keywords:**Genomics, Bio-Informatics, Differentially Expressed Gene, Feature Construction, and MassSpectrometry.

**I.Introduction**

In the Recent scientific field,handlingof analysis withhigh numbered genesis the fact aroseto provide way for protein analysis.Genomics has advanced to deal with multiple different data types and now it is beingwell-resourced to assist proteomics. Currently,proteomics systems[1] are already most actively co-operating with bioinformatics web data systems for complete functional analysis and mining logical knowledge of compound datasets.

Thefeature transformation methods are usually used to reduce the dimensionality of data by changing data expressions into new gene features. Gene selection methods[2] without transformation of class variables, i.e when there is categorical information present in the data. Thisworkalso highlights on the importance ofbioinformatics for data mining andfunctional analysis of datasets gaininginsights at biologically interpretable results.The categorical classification method uses unsupervised learning for to classify the samples as type 1 and type2 samples. The neural network based methods are used to classify the samples precisely with efficient accuracy on prediction.

In modernday era,mass spectrometry has beenone of the powerful technologiesfor studying proteins in the discipline of proteomics on a large-scale [3]. When it is united with innovative experimental methods [4], and advanced computational stratergies [5]. Kumar and Choudhary [6] constructs a top-down three layer model first layer classifies protein sequence as an enzyme or non-enzyme, the next layer predicts main functional class, last layer predicts sub-function class using random forest and has high classification accuracy. It has poor performance.Gall et al., [7] has studied the properties of residues under four categories and analyzed its relations for predictions of disorder based on various web tools based algorithms to characterise regions which revealed that the vast majority of residues in the observed dataset are ordered and not observed regions are mostly disordered. In some methods, less efficient outcome occurs.Inorder to tackle the poor classification of samples, highly efficient approaches are such as feature transformation based classification and feature selection using prediction based classification is needed. By these approaches the efficiency of classification can be improved by having improved accuracy on classifying categorical class samples.

1. **Differential ExpressionAnalysis on Genes**

In the microarray platform, in order to analyze gene expression data,the disease data sets are the typical example of study for gene expression analysis (GEA) [8]. Both normalized and raw expression data are available on the GEO database [9].

The dataset is quite huge and provides a lot of the information related to genes that do not show any vital changes throughout the experiment. In order to make a way easier for finding the significant genes, firstly,one has to reduce the dimensional size of the dataset by filtering genes with low expression profiles,which do not show any vitalchanges. There arenumerous techniques for reducing this many genes to some subset of genes that contains the most differentially expressed genes.

Feature transformation is group of approaches that filters to form new features. These approaches use dimensionality reduction for transforminggene features which has a descriptive power i.e.,much easily processed than that of original gene features. Here in this case, much lowexpressivegene features can be filteredfor consideration while constructinglogical model.Theseapproaches are much different from the approaches in actual gene selection, where dimensionality reduction of gene features is done by selecting an optimal subset of predictive genesby computationof theobserved original data.The main objective of feature selection is reduced feature set so to observe good classification between two distinct categorical classes.

**II.Implementation**

The data from every microarray gene expression experiments ideallyhasthese components such as thesample information of experiment,gene annotations,data values and the class variablesinformation. From the data, construct the experiment to find the differentially expressed genes. We implement a novel approach for gene selection. It is done in two phases, at first the log2mean ratio based gene selection approach is operated on leukaemia disease dataset having 7129 genes and 72 samples and a subset of differentially expressed genes are obtained. Thus obtained reduced gene sample set is processed using the PCA approach in next phase in-order to get the top 10 significant genes. The significant genes are used for protein pathway analysis by identifying set of associated gene paths; one can identify the pathways reactions, enzymes and compounds relevant to the specific gene using a protein pathway database. Also enable for further retrieving the set of related disease and drugs from disease database. To study about the associated diseases and proteins specific to significant genes. The above specified novel gene analysis and protein pathway analysis system process is shown in figure 1.

Extraction of Most Significant Genes(PCA)

Sample Classification (SOM)

Performance Measures (MSE)

Gene DB

Differential Expression Analysis

(Log2 Mean Ratio with Negative Binomial)

Fig. 1Differential Gene expression analysis

**Differential Expression Analysis using Log2 Mean ratio with Negative Binomial Analysis**

A classic differential expression analysis [10] of DNA gene data has a process of normalizing the raw data and thento perform a statistical test to accept or reject the null hypothesis that two categories of samples show no major difference in gene expression.Here we use Log2 Mean ratio with Negative Binomial Analysis for separating the differentially expressed genes. By computing the fold change FCin each gene, one can easily look and separate the difference of the gene expression among two categories.This analysis assumes the read counts are modelled according a negative binomial distribution andit performs a hypothetical testing with suitable three potential options for specifying the kind of linkage amongst bothmean and variance. So byinitializing the link between mean and variance as identity, one can assume the mean is equal to the variance, and then thedataismodelled by Poisson distribution. For to determine, the gene expression in either of two categories are statistically different must reject the null hypothesis, that the two data categories come from distribution with equal mean. In order to find the fold change and mean value of each gene use equations 1 and 2, then identify the down- and up-regulated genes.Where, X is the category AML data expression and Y is the category ALL data expression and n and m their corresponding sample count.

(1)

(2)

**Principal Component Analysis PCA**

PCA is an analysis approach which can be used to rescale data in lesserdimensions;PCA[11] is a convenientmethodwhichis used to reduce the high dimensional datasets, such as genes from microarray data analysis. PCA is used to find signals from noisein other data. Fortunately, in these dataset with many gene variables, categories of sample variablesoften get together. The same governing principle for driving the behavior of the system is the particular one reason that more than one category might be the determiningfactor.There are only a few such driving forces in many systems,but an abundance of instrumentshelpsone to measure a dozens of system gene variables. Itoccurs whenone take advantage of this redundancy of data. One can solve the problem, by simplyreplacing a group of gene variables with few new genesvariables.

PCA is a much quantitative rigorously heldapproachwhich is used for achieving the simplification of information. Thisapproachcreates a new variableset, called principal components. The each principalcomponent is constructed as a linear combination of actualgene variablesand these are mainly orthogonal to oneanother, so there cannot be any redundant principal components. The principal components form as a whole orthogonal basis for space of data. The first PCwill always be at a single axis in space. While projecting each observation on the single axis, the subsequent values generate new variable,the variance of this variableisextreme among all possible selections of this first axis.

The second PC is next axis in space, which is perpendicular to the previous axis. While the observations are projecting on to this second axis,itgenerates an extra new variable. Among all selections of the second axis, thevariance of this variable isextremelylarge.These set of PC’s is a large and similar to the actual set of variable and it is common for the first few PC’s to cross 80 percentin total variance foractual datawhile computingall the sum of variances.

(3)

(4)

Where, X is the data matrix, u - variable weights is a vector of length N containing all positive elements.The variable weights are the inverse of sample variance. The wcoeff is the observed weights coefficient found using equation 3, these coefficients are weighted,so the coefficient matrix is never orthonormal.In the first output,wcoeff, has the coefficients of firstPC.The score data matrixS is as same as the input data matrix X having same size. Similarly one can also get the component score by finding the orthonormal coefficients standardized through X, usingthe equation 4.The columns of score matrix correspond to samples and rows to genes in a score matrix.

**SOM Categorical Pattern Recognition**

Neural network models try to simulate the data processing which occurs in brain to formulate automated pattern recognition.Clustering data is an excellent approach for neural networks. This involves process of grouping similar data. It can be used for Bio-informatics analysis by grouping genes with related expression patterns.Start the process by simply arranging Q input vectors to be clustered as columns in an input matrix.Suppose consider clustering this set in5 two-element vectors.Then compute the class vectors of each of training inputs,these classificationsobserve the feature space populated by the known samples, and it classifies new samplessimilarly. The network output will be a 32x75 matrix, where each ith column represents the jth cluster for each ith input vector with a 1 in its jth element.The SOM topology of 16 neurons positioned in 4x4 hexagonal net,each neuron has the learning ability to represent a different category, with adjacent neurons representing similar categories.A hit calculates the categories of each sample and then it shows the number of samples in each category[12].

The easy conditions for simulating a network is when the network to be conditioned as static or that has no feedback. The standard algorithm for training the network is based on the minimization of an energy function representing the instantaneous error shown in equation 5. Where represent the desired network output for the input pattern and is the actual output of the neural network. Each weight is changed according to the rule in equation 6.



(5)

(6)

Where K is a constant of proportionality, E is the error function and represents the weights of the connection between neuron j and neuron i. The weight adjustment process is repeated until the difference between the node input and actual output are within some acceptable tolerance.



1. **Results and works**

The data structure of input gene data set contains data fieldsand header. The data field has a data matrix of expression levels from the MA experiment.The header field contains two fields namely, gene ID and samples of experiment.The dataset in this example includes of DNA gene data acquired in the leukaemia disease MA experiment. Every gene hasa specific role to play in regulation of protein pathways in human internal biology. The dataset contains47 biological (ALL) samples and 25 biological (AML) samples of the leukaemia disease. Inspecting read datatables for genomic features. The gene expression values are identified with sample IDs and gene IDs. To identify the differential changes in the gene expression values of ALL and AML samples and to compare the differentially expressed gene expression values ofthe two categoriesin data [13].

**Log2 Mean ratio with Negative Binomial Analysis**

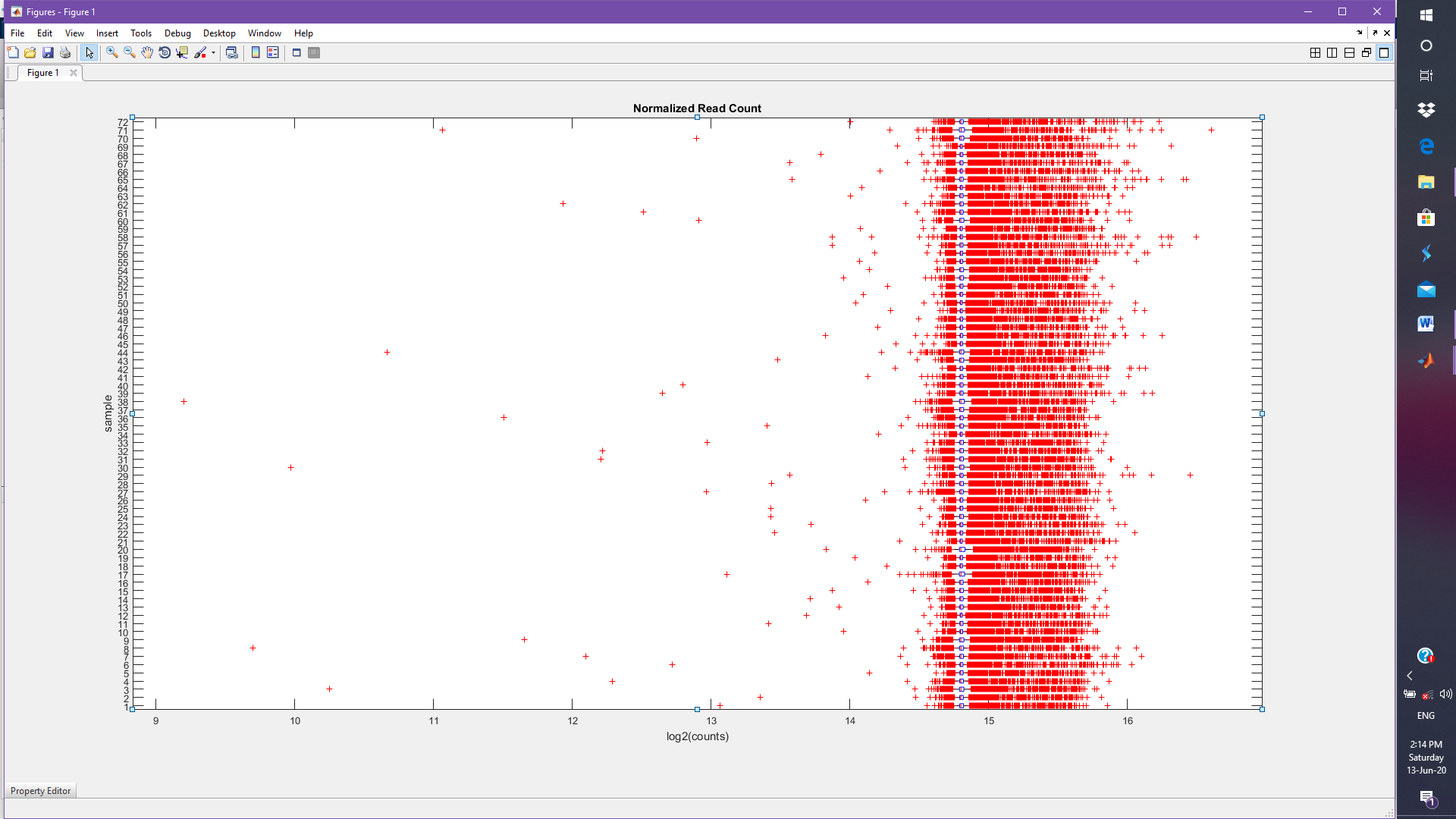
This demonstrates how to measure the size factor for data expression normalization and shows a way to find the differentially expressed genes bylog2 ratio and negative binomial model.Where the rows representthe genes and the columns represent the given samples and the expression values represent the specific expression change for a particulargene for a given sample.A best normalization procedurecomprises of calculating the actual size by taking a commonresize factor for each sample. Once by dividing every sample data by the commonresize factor, which brings all the data valued at a common scale, thus changing them as equivalent. The obtained normalized data is shown in figure 2.

Fig.2 Normalized Gene Expression Data

In order to assess the resize factor, consider the median of those of theobserved data expression ratiosin a pseudo reference sample,where thedata expressionmust be found by taking the geometric mean of each gene ineach and everysample. Then, convert the observed data expressionto a common scale and divide the experientialdata expressionin every sample by the equivalent resize factor.

For better characterization of the data, we represent mean and dispersion of normalized data expression. The variance is given by the sum of two terms for thedata expression as the variation throughthe samples and the un-certainty of assessing the expression while reading thedata expressions. The variance leads significantly for differentially expressed genes, where the shot noise leads for lowly expressed genes. One can showdispersion values with the mean of normalized data expressionin a log2 scale as plotted in figure 3.The small number of duplicates is never surprising,while expecting the dispersion scatters with some of the variance around it basically the true value.Some of the variance shows the true variationsbetween each of the gene in gene expressions samples and some of this variance shows the sampling variance.

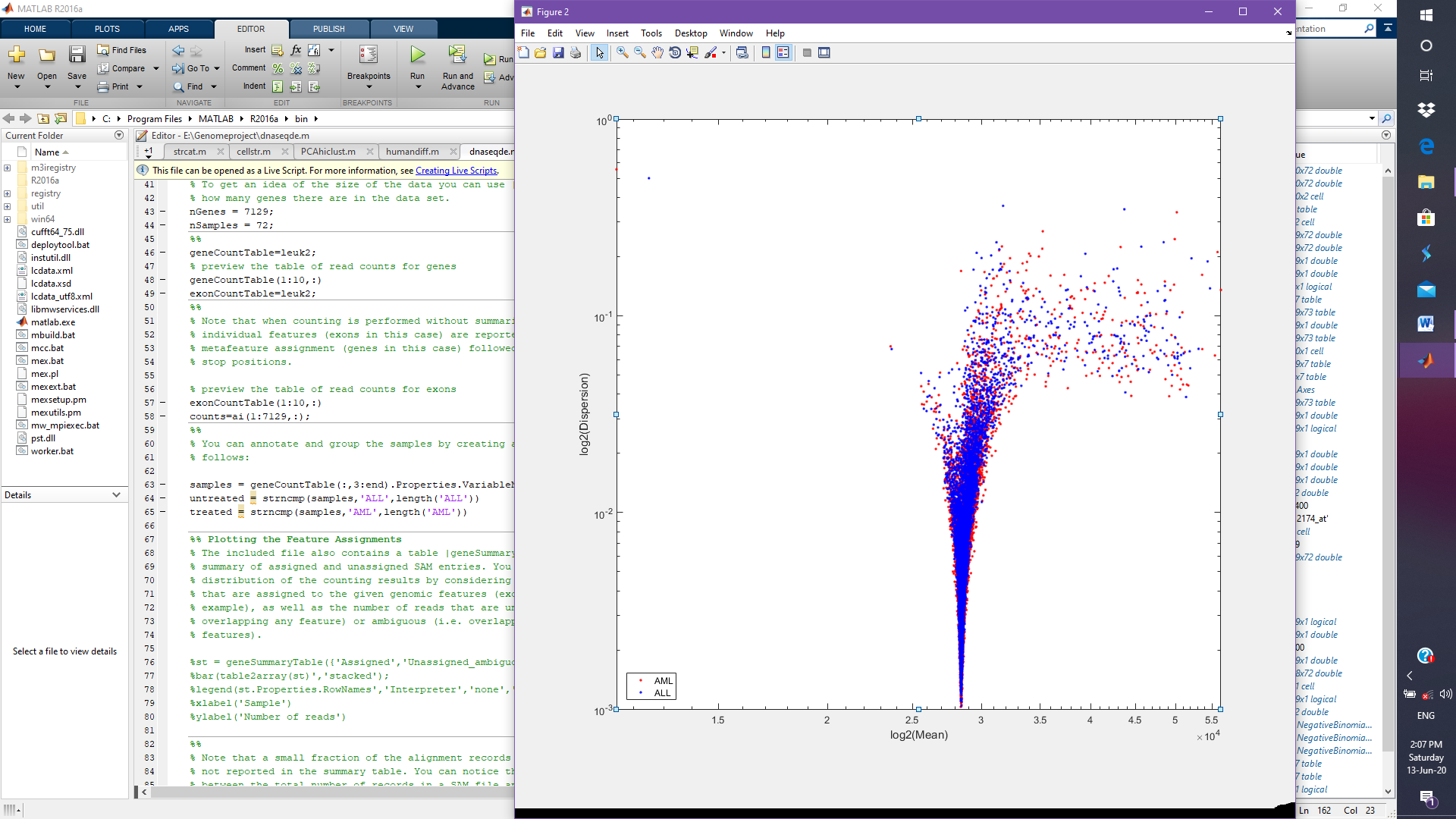


Fig 3: Mean vs Dispersion

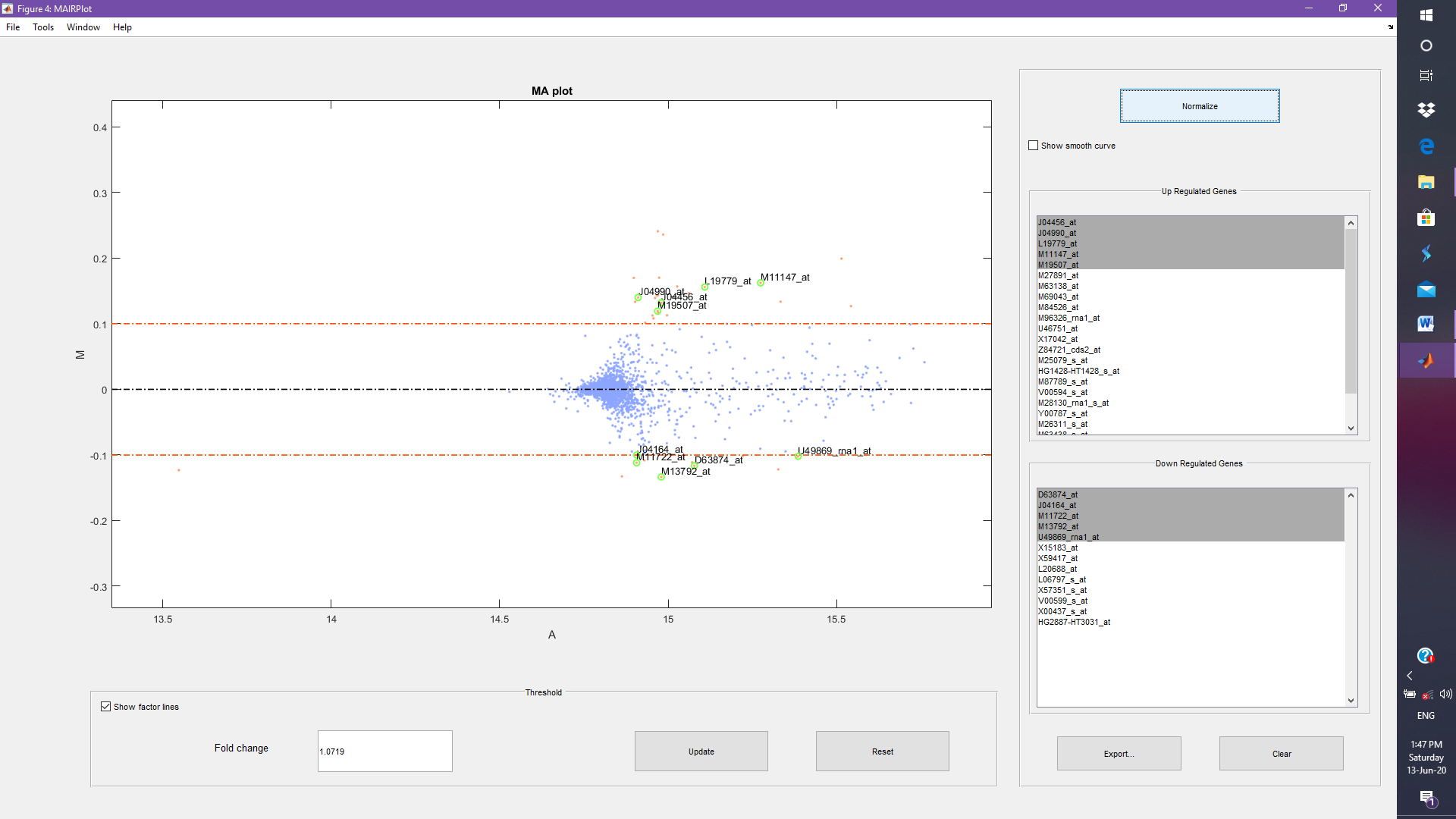


fig 4: Mean(M) vs Fold Change(A)

One can easily see the changes of the gene expression values between two categories, by means of calculated the fold change FConevery gene, that is the ratio ofchanges to data expression in the AML category over the changes to data expression in the ALL category. Normally this ratio isbased ona log2 scale, so allchangeare symmetric with respect to zero. While plotting the FC(A) and Mean (M) on a scatter plot, one can easily observe highly expressed variables as in figure 4.

By representing the variance as the sum of mean and by using a constant multiplied by squared mean, thedata expression ismodelledas perthe given distribution. Theconstant term is assessedusingevery row of data.According to the proposed distribution, the data are modelled by utilizing the variance as the sum of the meanand a relatively locally regressed non parametric smooth function of mean.Thus the output of negative binomial functionhas a variable vector containingp-values. Thep-value vectorshows the change in expression probabilityvalue ashigh as observed one occurs under null hypothesis.So the conditions have low effect on gene expression. We observe an enrichment of low values,in the histogram of the p-values, whereas other values are uniformly spread. While enrichingthe values equal to 1 isbecause ofvery low expression valueof genes.Select those genes with comparatively low expression valuefor to perceive morenumber of non-significant p-values across the range of 0.1 in uniform spread.So by using thresholding ofp-values to find, which are the FCvalues, which are more significant than others and are not suitable for this kind of data analysis, on basis ofmultiple testing based problem. The probability of getting a significant result merely by chances in rises with the number of tests occurs,whileperforming a many number of simultaneous tests. Thus, forto account multiple testing performances, dothesimplemodificationin thep-values.So, the probability of detecting least significant result remains below desired significance level,due to chance. The enriched p-values are plotted with the gene frequencies as shown in figure 5.The Benjamin-Hochberg modification is a statistical method whichgives an adjustedp-value toset a particular threshold of 0.1 for the adjusted p-valuesis equal to considering10false positives as desirable and then whileconsidering every gene with adjusted p-values below this threshold,one can easily identify the genes which are differentially expressed. For identifying the Up and down regulated genes [14], one can now easily separate thegenes by selecting an absolute FC value above a selected cut-off. The upand downregulated genes are found and are shown in table 2 and update MA Plot with genes having highest and lowest FA values as in figure 4. Out of 7129 genes the 24 genes are upregulated and 15 genes are down regulated and in total genes 358 are differentially expressed genes in leukaemia dataset. Out of 358 genes 5up-regulated and 5 down regulated genes are displayed in table 2.A better visual representation of the gene expressions and the respective significance is given, by plotting the mean andfold change in log2 scale and by colouring the data points according to that of corresponding adjusted p-values is graphically shown in figure 6. One can see the weakly expressed genes as the FDR is usually high because the low data expressions are subject untopoisson noise.

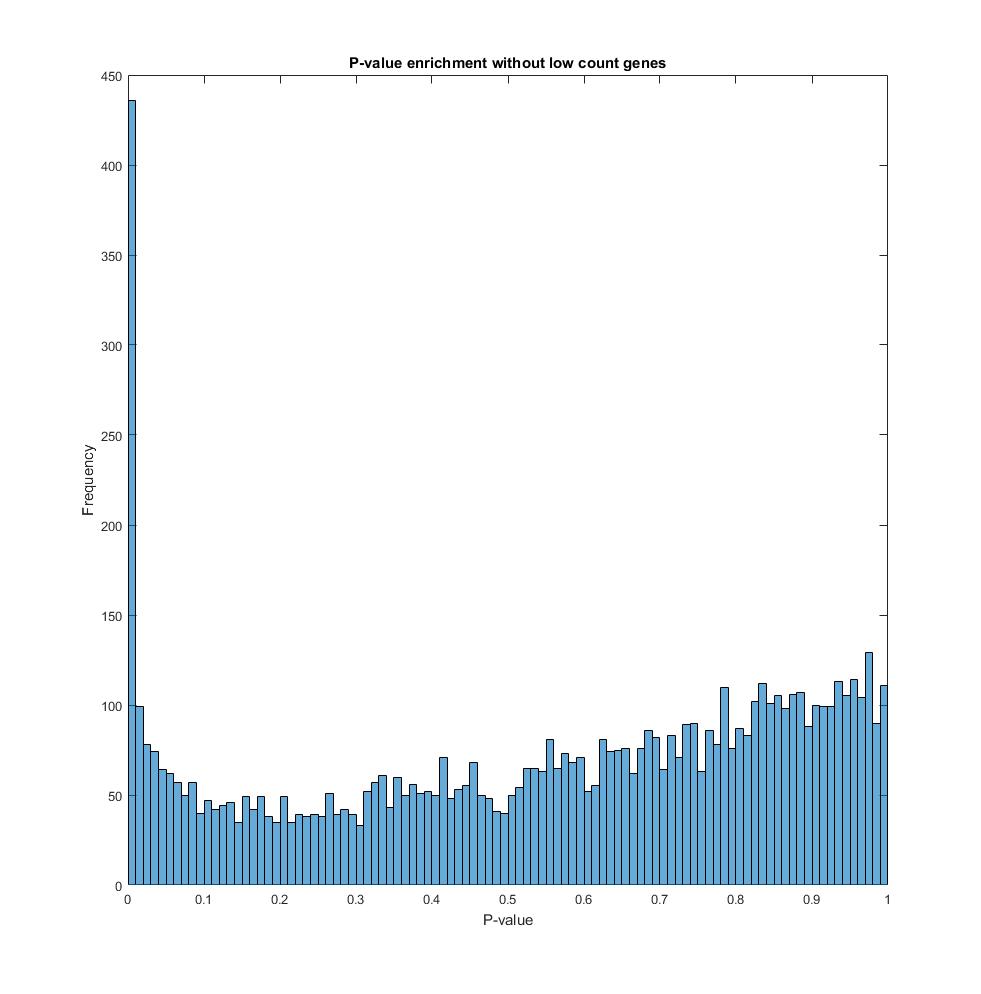


Fig.5 P-value Enrichment

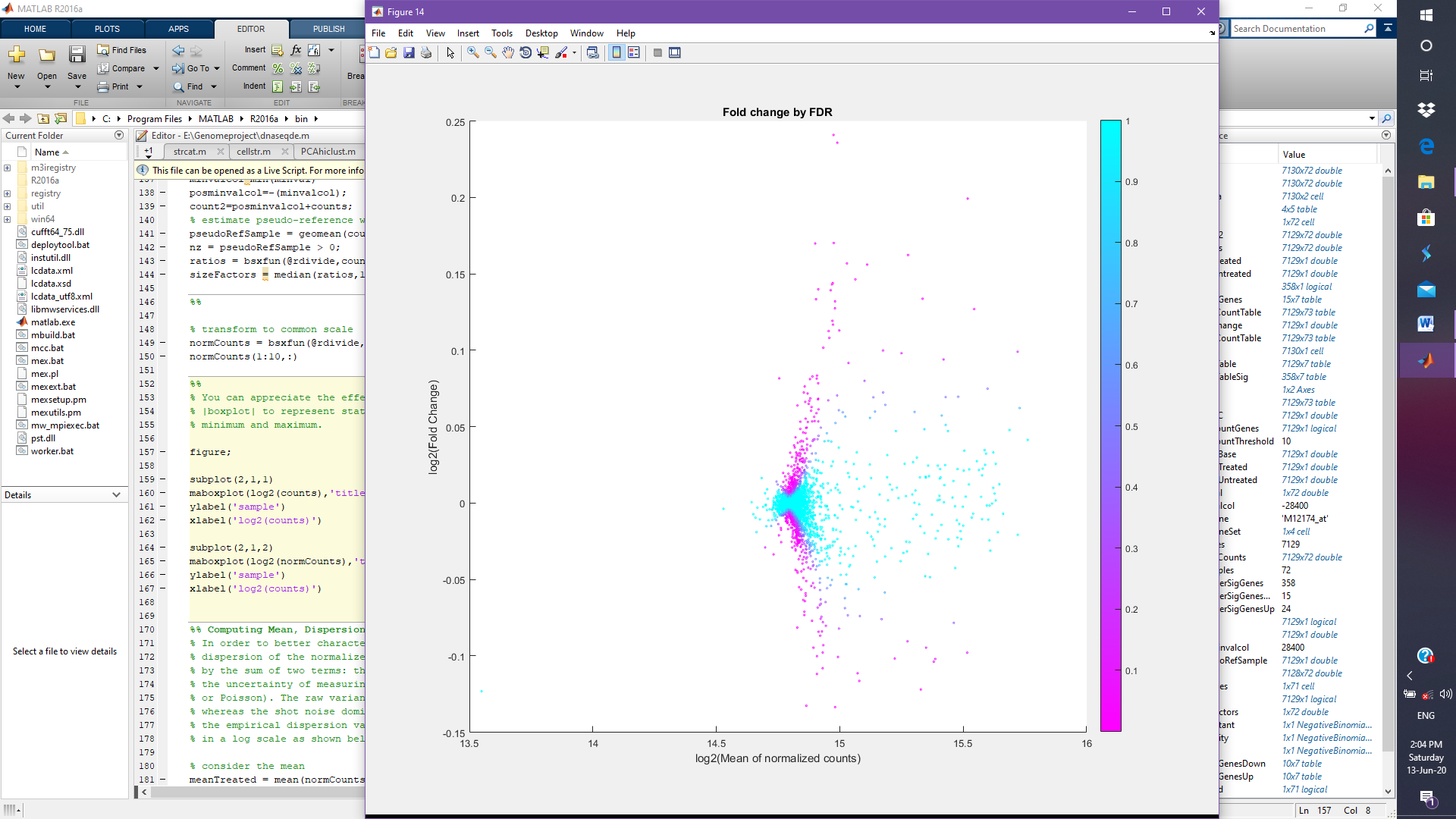
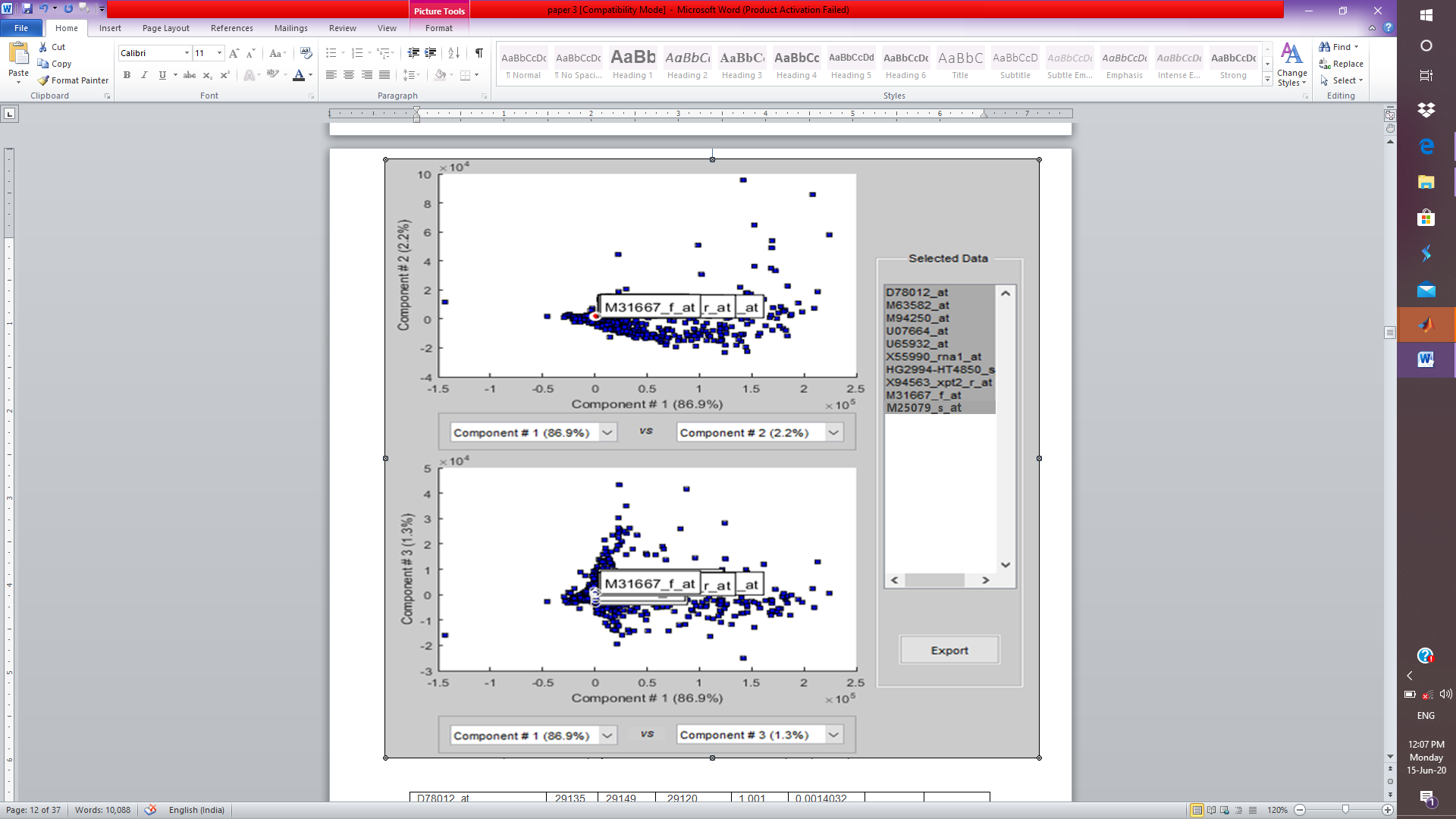


Fig. 6 Fold Change by FDR

Table2 Up-regulated and Down-regulated Significant Genes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **Gene-ID** | **MeanBase** | **MeanAML** | **MeanALL** | **FoldChange** | **Log2FC** | **P-value** | **P-adj** |
| UP gene | M19507\_at | 32131 | 33733 | 30530 | 1.1049 | 0.14393 | 6.4974e-07 | 0.00005 |
| M11147\_at | 39698 | 41929 | 37468 | 1.1191 | 0.16231 | 3.1569e-06 | 0.00020 |
| J04456\_at | 33434 | 35249 | 31619 | 1.1148 | 0.15683 | 3.1672e-06 | 0.00020 |
| L19779\_at | 35387 | 37299 | 33475 | 1.1142 | 0.15604 | 6.5259e-06 | 0.00038 |
| J04990\_at | 34236 | 35972 | 32501 | 1.1068 | 0.14639 | 2.6321e-05 | 0.00131 |
| Down gene | M13792\_at | 32344 | 30848 | 33840 | 0.91159 | -0.13354 | 7.7719e-05 | 0.00333 |
| D63874\_at | 34618 | 33221 | 36014 | 0.92246 | -0.11644 | 0.00097 | 0.02697 |
| M11722\_at | 30733 | 29542 | 31924 | 0.92537 | -0.11190 | 2.8985e-05 | 0.00143 |
| JO4164\_at | 34451 | 33121 | 35780 | 0.92570 | -0.11139 | 0.00158 | 0.03985 |
| U49869\_at | 42844 | 41329 | 44359 | 0.93169 | -0.10208 | 0.00451 | 0.09177 |

****

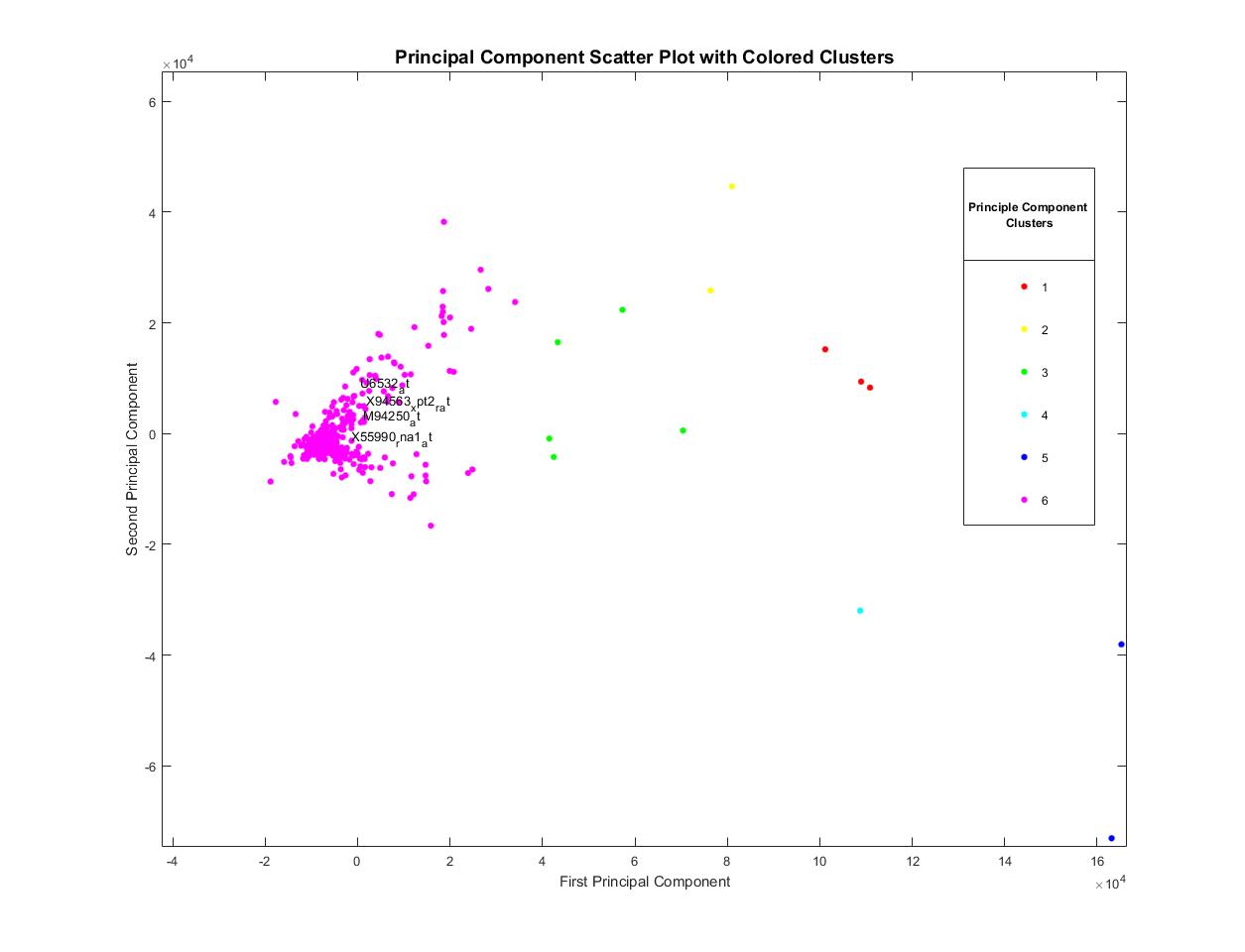
Fig. 7 Comparison Analysis of PC1 vs PC2 & PC1 & PC3

Fig.8 Principle Component Analysis with clusters

**Principle Component Analysis**

The Principle Component Analysis is done on the differentially expressed gene based expression data. The PCA is used to check the pair-wise correlation between the variables using correlation of AML, ALL data to find the relation of two categorical data. At first, find the PCA by using the inverse variances of the expression data as weights, on bi-categorical Gene Expression data. Using cluster value as minimum ie. C=6 in order to obtain the specified cluster of principle components. Successively plot component PCA scores, thus creating a map of first two principle components of relational score. As in figure 7, this plot shows the centered and scaled expression data projected onto the first two principal components, In the PCA scores,it always keeps the mean zero.PCA Removes particular genes, which is not unexpected as many of the filtering process remove much of the genes with less information or less variance. Selects the genes with highest PCA score likewise those are the gene that wouldappear in the middle of the scatter plot, plotted between top two principle components. Extract genes names by generating an index containing the gene id of every geneon gene expression set and get the names of the top most significant genes. By using clustering analysis of significant gene profiles on PCA based differentially expressed gene data shows linear separability of the AML samples from the ALL samples. Plot the coefficients of the first two principal components with six clusters as in figure 8. The retrieved most significant top 10 genes by Principal Component Analysis are listed in table 3.

Table 3 Significant Genes

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene-ID** | **PC1** | **PC2** | **PC3** |
| 'HG2994-HT4850\_s\_at' | 165248 | -38046.9 | 27280.35 |
| 'M25079\_s\_at' | 163121.2 | -73055.5 | 59313.54 |
| 'U07664\_at' | 110858.4 | 8300.774 | -33832 |
| 'U65932\_at' | 108940.3 | 9381.539 | -43395.2 |
| 'M94250\_at' | 108731.5 | -31969.7 | 27420.03 |
| 'X55990\_rna1\_st' | 101172.9 | 15210.85 | -31058.7 |
| 'M31667\_f\_at' | 81006.79 | 44642.03 | -3335.95 |
| 'X94563\_xpt2\_r\_at' | 76362.73 | 25825.04 | -2761.56 |
| 'M63582\_at' | 70414.36 | 559.8713 | -35405.5 |
| 'D78012\_at' | 57298.39 | 22350.04 | -12211.6 |

The PCA is trained using SOM for validating the known features. One can assign the clusters using the SOM, by computing the nearest node to each point ofdataset. Thus the PCA component clusters which are previously formed are

displayed with SOM weight vectors for 2PCs and 3PCs in figure 9 and 10 respectively.One can tell you how many data points are accompanying with each neuron, by seeing the figure. It is best if the data are fairly evenly distributed across the neurons.

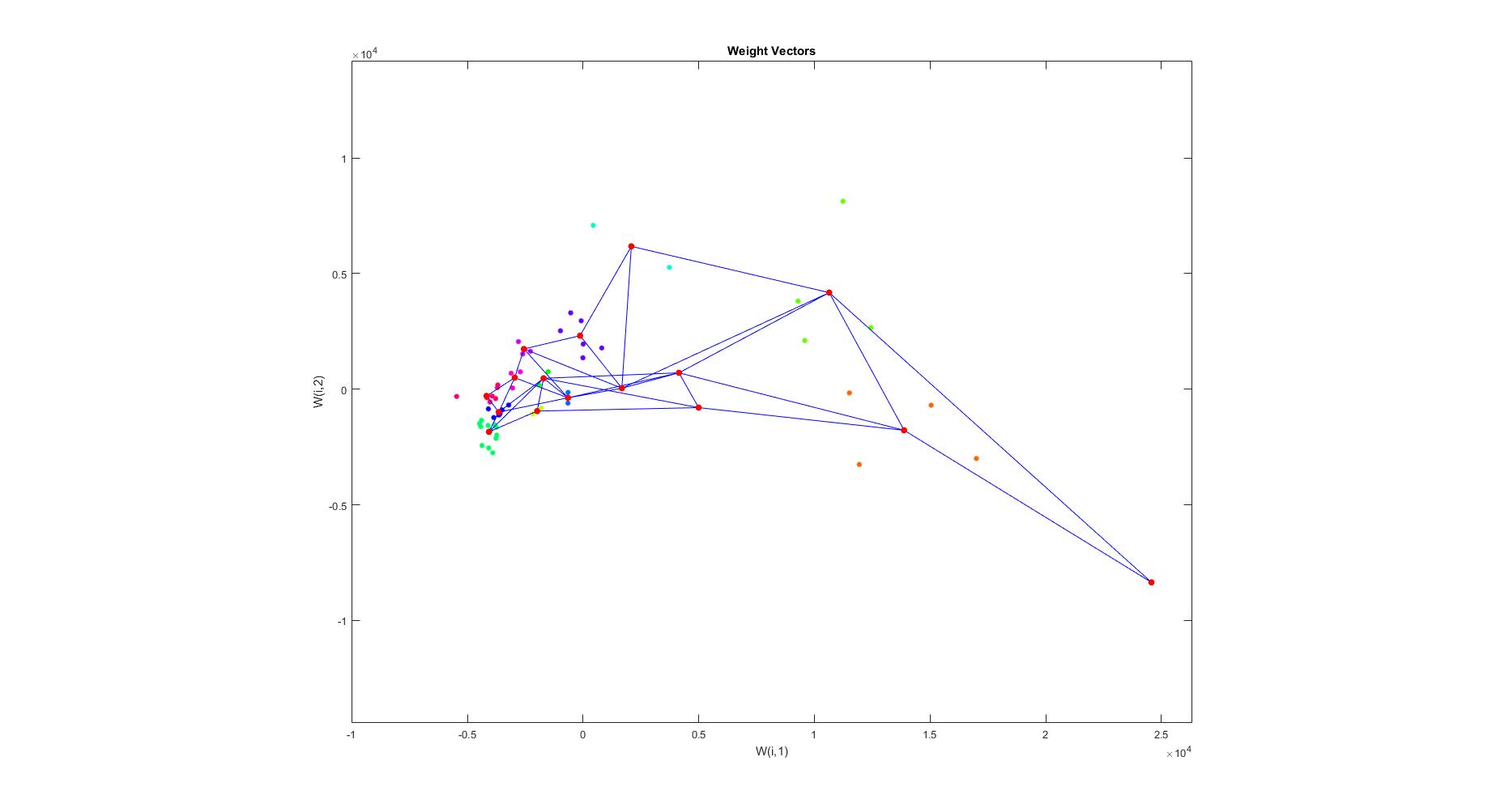


Fig.9 PCA Trained in SOM with weight for 2PC

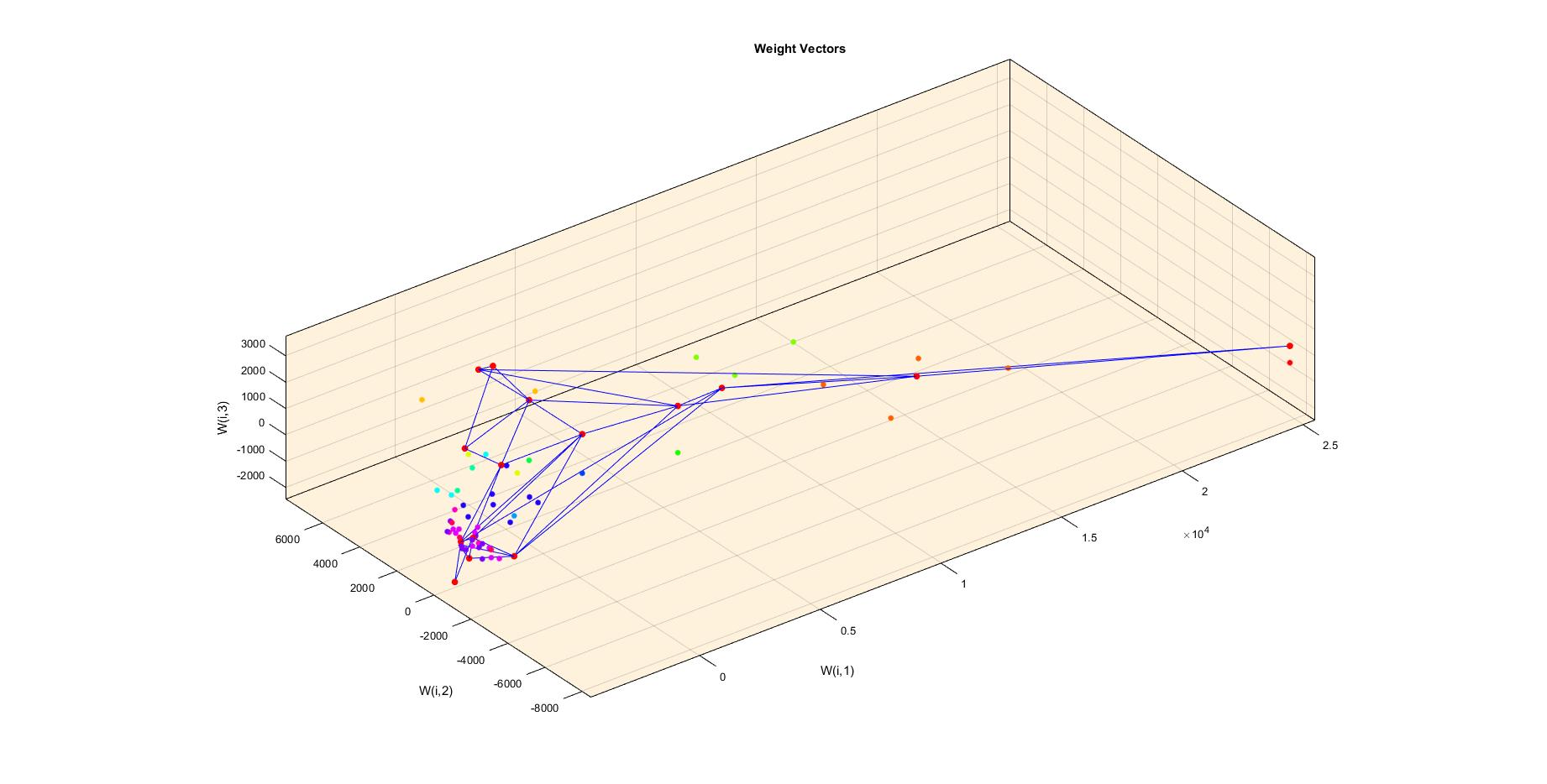


Fig.10 PCA Trained in SOM with weight for 3PC

Table 4 Training Test and Validation Data Samples for Classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Class Category** | **Training** | **Validation** | **Test** |
| Leukemia | 72 | 36 | 18 | 18 |
| 47 (ALL) | 23 | 12 | 12 |
| 25 (AML) | 13 | 6 | 6 |

The SOM learner model for training the input samples which can assess the category of samples is shown in Figure 11, having 4 inputs to form neurons as 2 neurons in output layer depicting the learned trained samples in the feature space. The data expression and class categorical samples are divided into 36 training sets, 18 validation sets and 18 test sets as shown in the Table 4

After learning the performance is verified in training, testing and validation samples without PCA using MSE is shown in Figure 12. The Gradient, Performance, and Learning Rate of PCA data is given in Figure 13 and classification details of SOM is given on Table 5. The net has high classification of categorical samples across the feature space. Thus by using cross validation based sample set selection and observing various categorical separations throughout the learned neural weights, PCA based L2M-NBIN gives highly predictive sample classification rate.

**Output**

**Hidden**

**4**

**Input**

**W**

**b**

**~~ꭍ~~**

**Output**

**W**

**b**

**T**

Fig.11 SOM Network Model

Table 5 SOM Sample Classification Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TruePos.(TP)+TrueNeg. (TN)** | **Training** | **Validation** | **Test** | **All** | **Accuracy (TP+TN)** |
| % Samples (72) | 70% | 15% | 15% | 100% | 100% |
| PCA | 70%+26% | 45.5%+36.4% | 45.5%+45.5% | 62.5%+30.6% | 93.1% |
| L2M-NBIN | 58%+42% | 72.7%+9.1% | 72.7%+9.1% | 62.5%+31.9% | 94.4% |
| L2M-NBIN-PCA | 64%+34% | 72.7%+22.1% | 72.7%+27.3% | 71.4%+26.4% | 97.8% |

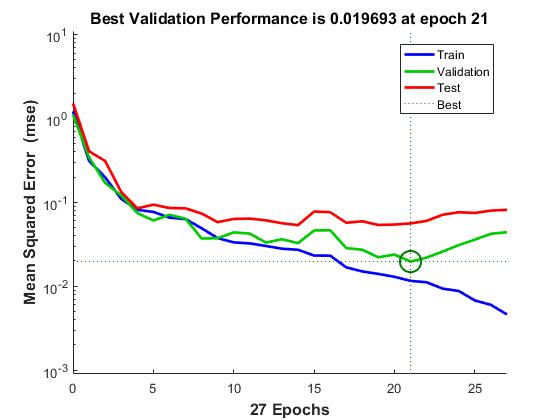


Figure 12: Performance of L2M-NBIN on SOM without PCA

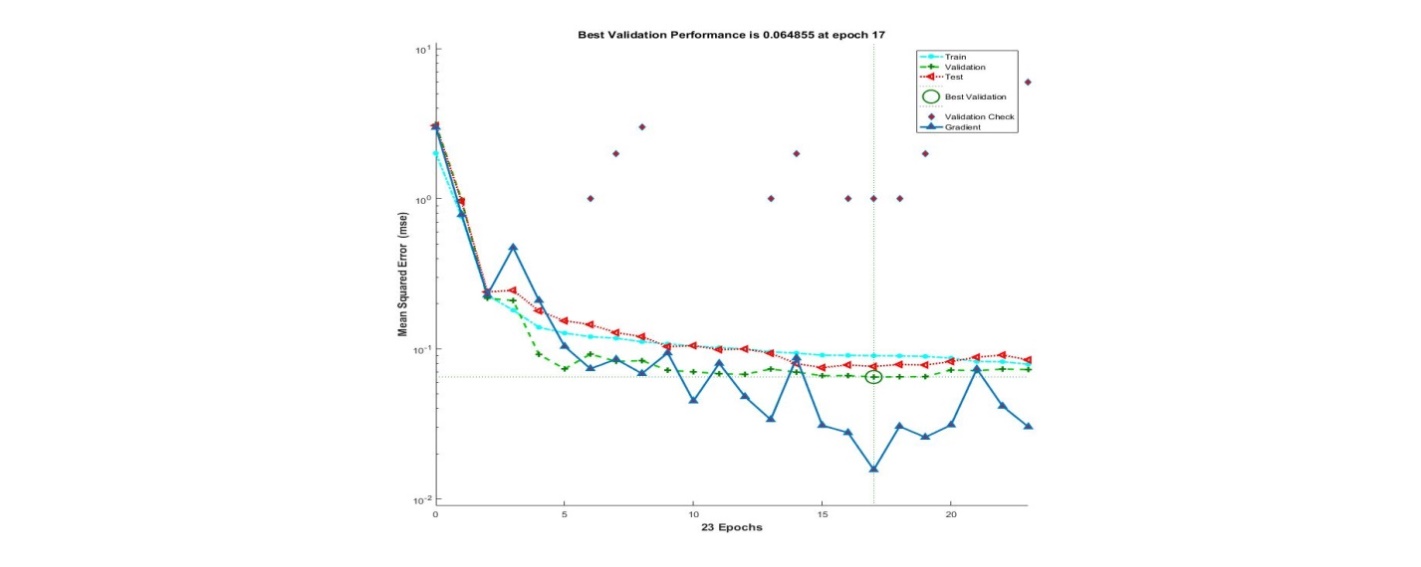


Figure 13: Performance of L2M-NBIN -PCA on SOM

When comparing gene selection with transformed features based classification, it gives high accuracy over other approaches as given in Table 6. The conventional PCA transformation based SOM Classification shows lesser accuracy rate. Thus the Differential expression analysis L2M-NBIN selects significant genes effectively and PCA also transforms data for better classification accuracy, while learning the transformed gene based leukaemia data expression of binary class samples in SOM classifier it gives effective accuracy, than non-transformed L2M-NBIN selection based leukaemia data expression.

Table 6. Comparison of Accuracy Classification Rate

|  |  |
| --- | --- |
| **Other Approaches** | **Accuracy (%)** |
| DWLB [15] | 92 |
| CS [16] | 96.9 |
| PCA with SOM | 93.1 |
| L2M-NBIN with SOM | 94.4 |
| L2M-NBIN-PCA with SOM | 97.8 |

**III.Conclusion**

This paper gives a clear framework of understanding on gene data handling system and with the use of gene expression data the most differentially expressed genes are found.Thus sub-selected genes are assigned for intense PCA analysis, so to extract the most significant gene. Inaddition to that, the gene can be further referred to many bioinformatics based databases for advanced pathway analysis, also making way forassociated disease information retrieval options possible. Based on the result statistics our leukaemia dataset had a high significant gene extraction, as the Log2Mean withNegative-Binomial model has efficient processing on gene expression data. The PCA has a high feature transformation capacity, as it obtains a good proportionate cluster of principle components. Further the genes provide high learning capability on the SOM classifier by showing distinct hits between two categorical classes.

**References**

|  |  |
| --- | --- |
| [1] | Pradipta M. and Paul S., “Scalable Pattern Recognition Algorithms” Springer Applications in Computational Biology and Bioinformatics, Vol. 22, No. 01, 2014. |
| [2] | Maji P. and Pal S. K., “ Fuzzy Rough sets for information measures and selection of relevant genes from microarray data”, IEEE Transaction on Systems Man and Cybernetics, Vol. 40, No. 3, pp.741-752, 2010. |
| [3] | Meinshausen N., and Yu B., “Lasso-type recovery of sparse representations for high-dimensionaldata”, Annals of Statistics, Vol.37, pp.:246-270, 2009. |
| [4] | Guyon I., Saffari A., Dror G., Cawley G., “Model Selection: Beyond the Bayesian/Frequentist Divide. Machine Learning Research, Vol.11, pp.:61-87, 2010. |
| [5] | Jianzhen, Yongjin, “Discovering disease-genes by topological features in human protein–protein interaction network”, Bioinformatics on System Biology, Vol. 22, No. 22, pp. 2800-2805, 2007. |
| [6] | Kumar C. and Choudhary A.,“A Top-down Approach to Classify Enzyme Functional Classes and Sub-classes Using Random Forest”, EURASIP Journal on Bioinformatics and Systems Biology, Vol. 1, pp. 1–14, 2012. |
| [7] | Gall T. L., Romero P. R., Cortese M. S., Uversky V. N. and Dunker A. K.,“Intrinsic Disorder in the Protein Data Bank”, Journal of Biomolecular Structure and Dynamics, Vol. 24, No. 4, pp. 325–341, 2007. |
| [8] | Tuv E., Borisov A., Runger G., Torkkola K., “Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination”, Machine Learning Research, Vol.10, pp.:1341-1366, 2009. |
| [9] | Leukaemia datasets in Gene Expression Omnibus database, “NCBI-National Centre for Biotechnology Information”, http://www.ncbi.nlm.nih.gov. |
| [10] | Buerger, Florian B., Jens P., Konstantin A., Kathrin P., Walter N., and Eberhard K. “Analyzing the basic principles of tissue microarray data measuring the cooperative phenomena of marker proteins in invasive breast cancer” Arxiv Preprint, Vol. 18, No. 3, pp.:1-28, 2018. |
| [11] | Christopher M.,“Pattern Recognition and Machine Learning” Springer Sci., 4th edn. 2006. |
| [12] | Liang Y. and Kelemen A., “Temporal Gene Expression Classification with Regularised Neural Network”, International Journal of Bioinformatics Research and Applications, Vol. 1, No. 4, pp.: 399–413, 2005. |
| [13] | Briso J., Maria S., “Genic Disorder Identification and Protein Analysis Using Soft Computing Methods”, Springer, Vol.07, No.12, pp.: 73-81, 2018. |
| [14] | Primrose et al., “Principles of Gene Manipulation and Genomics”, Blackwell, Edn.5, Oxford, 2006. |
| [15] | Bose S. S. C., Sivanandam N. and Sundar P. P., “Design of Ensemble Classifier Using Statistical Gradient and DynamicWeight Logit Boost for Malicious Tumor Detection”, Journal of Ambient Intelligence and Humanized Computing, Vol. 12, No. 6, pp. 6713–6723, 2021. |
| [16] | Sampathkumar A., Rastogi R., Arukonda S., Shankar A., Kautish S. and Sivaram M., “An Efficient Hybrid Methodology for Detection of Cancer-Causing Gene Using CSC for Micro Array Data”, Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No. 11, pp. 4743–4751, 2020. |