Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-10

Identification of Marker Genes in Leukemia Cancer Types with Minimum Threshold Gene Expression Value and Quadratic Discriminant Analysis

Abstract- **Attempts are still in full swing for the detection of differentially expressed genes and using the genes for the detection of early cancer. Each method is with its own simplicity or complexity with the results, yet to portray some degree of confidence level in the designation of a gene as being differentially expressed or not expressed. However, no one statistic is universally optimal and there is seldom any basis or guidance that can direct toward a particular statistic of choice for such complicated gene data. Identifying a subset of genes that are expressed differentially in leukemia types from a large pool of candidates' genes generated by Golub *et al*(1999) in microarray experiment has been empirically demonstrated in the study. Principal component analysis on a Box-Cox transformed data exposed clusters of specific leukemia types. Taking into consideration the consistency of a gene expression the relative variance of each gene across the sample expression profiles is determined. A minimum threshold value of gene expression level from the data set developed a feature selection discriminant rule that discriminated genes into their specific leukemia types. Quadratic discriminant analysis provided encouraging results, validating identification of genes with probability of correct identification exceeding 85%. The group of differentially expressed genes, when subjected to principal component clustering fell into clusters of their own specific leukemia types.**

***Keywords*-principal component analysis, relative variance, quadratic discriminant analysis.**

I. Introduction

Micro array analysis is in progress for a last decade plus and in particular in an unclassified cancer to identify novel cancer subtype for subsequent validation and prediction, and ultimately to develop individualized prognosis and therapy. It also promotes to find clusters of genes that may be functionally belonging to known subtypes of cancer. These clusters of genes may then judiciously disclose the biological pathways and pathogenic aspects diseases. A comparison of gene expression (GE) levels in clusters, uncover the function and reactions of different genes. "A gene found that is differentially expressed in a cancerous tissue, then has the corresponding protein product (or an antibody to it) may be detectable in blood or urine, and could be the basis for a population screening for a population test" [1]. The fact to be considered is that there are genes that may be contributive in inflammation or growth which is the natural process in the body so cannot be taken as the likely genes for specific disease. Therefore identifying a group of genes that are expressed differentially in the different subtypes of a cancer from a big pool of all likely genes has been a cucumber some task achieved in the data under study. The site http://www.genome.wi.mit.edu/MPR contains the leukemia gene data set [3]. The gene data contain measurements corresponding to Acute Lymphoblastic Leukemia with T , B-cells(ALL-B, ALL-T) and Acute Myeloid Leukemia (AML) samples from Bone Marrow and Peripheral Blood. The 27 samples were taken from the bone marrow of the patients suffering from ALL with 19 samples of ALL-B and 08 of ALL-T cell lineages. The 11 samples taken from peripheral blood belonged to AML. The preprocessing steps [3] designed for GE data set were followed with thresholding and filtration. The threshold was a floor of 100 and ceiling of 16000; and genes were filtered if the ratio of maximum/minimum was less than or equal to 5 and a difference of maximum-minimum was less than or equal to 500 (maximum and minimum GE level across the samples). As an output of the preprocessing and screening steps each of the 38 samples squeezed from 5000 to 2299 genes forming a data matrix ($X_{2299\times38}$).

II. EXPLORING THE DATA SET

A. Clustering with Principal Component Analysis

Cluster and discriminant analysis for exploration GE data of tumors has been performed in literature [2]. For discriminant analysis, "or 'class prediction', they proportioned a 'weighted gene voting scheme' that turns out to be a variant of a special case of linear discriminant analyses". They used SOM to cluster sample expression profiles and found four clusters with two non-mutual exclusive clusters with respect to leukemia cancer types. In the study as an exploratory segment a scrutiny of the correlation matrix (R) of the high skewed $X_{2299\times38}$ patterned correlation matrix shadowed the presence of a three cluster structure in the data set. A principal component analysis (PCA) was applied on $X_{2299\times38}$ in the following four ways

(i) PCA using R on logarithmically transformed data (Method 1).

(ii) PCA using Σ_{RV} (the relative variance covariance matrix of the $X_{2299\times38}$) on logarithmically transformed data (Method 2).

(iii) PCA using R on Box-Cox transformed data (Method 3).

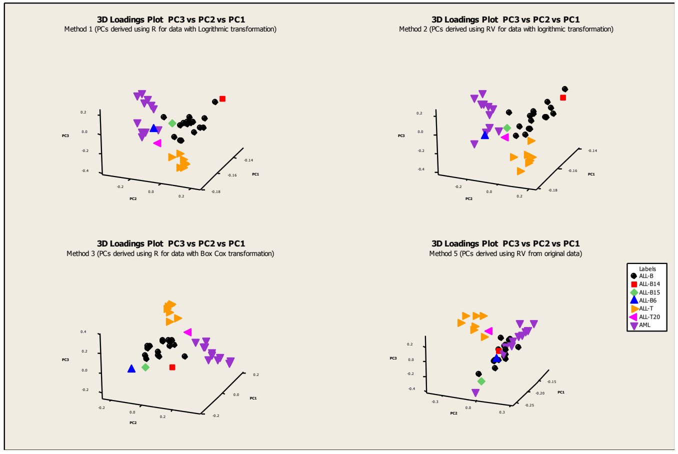
(iv) PCA using Σ_{RV} on original data matrix

Uzma Nawaz is Associate Professor of Statistics in The Women University Multan, Punjab Pakistan.

(Method 4).

The concept of Σ_{RV} was primarily introduced and applied in PCA in 1998 [3]

Principal component loadings were derived from all four methods. A three dimensional (3D) principal component loadings plot was drawn for each method used. Method 3 was proven effective in clearly depicting three mutually exclusive clusters of the specific leukemia types, the ALL-T, ALL-B and AML in the figure below


3D Loadings Plot PC3 vs PC2 vs PC1
Method 1 (PCs derived using R for data with Logrithmic transformation)

3D Loadings Plot PC3 vs PC2 vs PC1
Method 2 (PCs derived using RV for data with logrithmic transformation)

3D Loadings Plot PC3 vs PC2 vs PC1
Method 3 (PCs derived using R for data with Box Cox transformation)

3D Loadings Plot PC3 vs PC2 vs PC1
Method 5 (PCs derived using RV from original data)

Method 3 is the only effective method to expose the three biologically known clusters of the leukemia types existing in the data set with 2299 genes across the 38 samples. A feature of the study unlike taking few highly variant genes in literature so far to the best of knowledge.

II. METHODOLOGY

The methodology used in the study for finding discriminatory genes takes the relative variance of a gene at time and selects a handful of high variant genes. The GE levels showed abundance of a minimum GE level measured as “20” the minimum threshold value (Mth) which is later found as a discriminatory value for identifying marker genes in the leukemia type clusters. The continuous and steady research of classification tasks based on microarray data has shown that to classify two groups of sample a handy number of genes is sufficient [4-14]. Usually, genes few in numbers are studied for classification in a joint multivariate manner. Yet this is not the end at times the task of good classification is achieved with one or two genes only [10, 15, 14]. Initially the examination of one gene at a time was opted and a gene was ranked with its classification ability. In the second phase only genes with high ranks were segregated for further studies, including new confirmation experiments [16-18]. “Some information could be lost by not considering genes jointly, but focusing on single genes often simplifies the biological interpretation of the results [19]”. It has been already found that “many genes exhibit near-constant expression levels across tumor samples [4]”. These genes have a similar behavior across the tumor samples so are classified as meta genes or group of genes classified of a type. The similar behavior of a gene in statistical implication is the GE level is consistent across the samples. Thus relative variance (RV) the ratio of the standard deviation of a gene (σ_{RV}) to the mean (μ_{RV}) of the gene across all the samples is determined. The initial classification of genes is achieved with the application of the Garcia criterion [21]. A classification of RV into four groups with respect to the level of homogeneity of data and used Shapiro-Walks test the normality of the proposed classification. The Garcian criterion and its application is presented in Table 1.

Table 1. Garcian Criterion and its Application

Groups	**Garcian Criterion evaluated (No of Genes)**
Group (I)	$RV \le \mu_{RV} - \sigma_{RV} : RV \le 0.4257$
Group(II) $G_{1(1998 \times 38)}$	$\mu_{RV} - \sigma_{RV} < RV \le \mu_{RV} + \sigma_{RV}$ $0.4257 < RV \le 1.6466$ (1998 genes) Medium RV
Group(III) $G_{2(176 \times 38)}$	$\mu_{RV} + \sigma_{RV} < RV \le \mu_{RV} + 2\sigma_{RV}$ $1.6466 < RV \le 2.2571$ (176 genes) High RV
Group(IV) $G_{3(125 \times 38)}$	$RV > \mu_{RV} + 2\sigma_{RV} : RV > 2.2571$ (125 genes) Very High RV

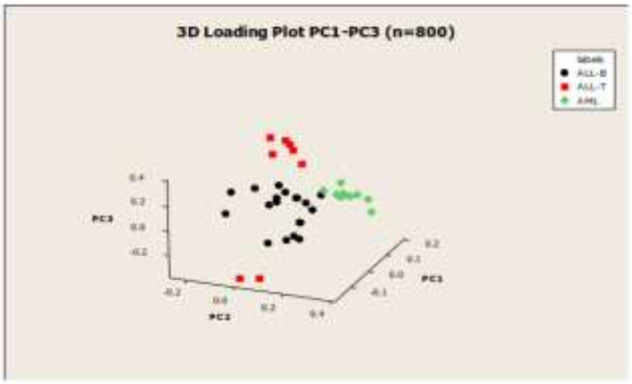
None of the genes fell in Group I the low RV category for GE levels high consistent and homogeneous across the sample expression profiles. Group II consisting of 1998 genes represent portion of data which is homogeneous but not as Group I was. Group III consist of 176 genes with high varying GE levels across the samples. It may be more likely to find genes in the group that can be the marker to the sub types of leukemia Cancer. Group IV has filtered 125 genes with extreme degree of inconsistency in their GE levels across the samples the outlying group of genes. Therefore it is highly likely that genes falling in category III and IV may be the functional genes which might be useful for the detection of early cancer type with their significant values of GE levels. Pooling the high variant genes (Group III and IV) result in a data matrix of X_{301*38}.

A. Role of Minimum Threshold Value Measured as Gene Expression level ‘20’

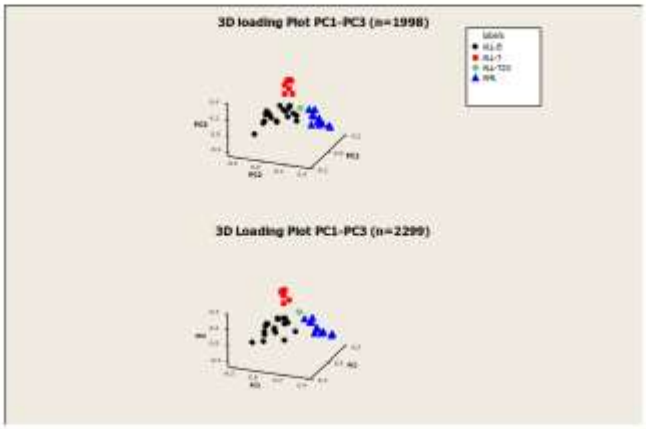
A gene with a uniform pattern of lowest GE level like the Minimum threshold (Mth) value measured as ‘20’ across the samples in a cluster or clusters may be functionally passive with respect to the particular leukemia type cluster but it may at the same time be functionally important and contributive to the disease with GE level other than Mth value in other cluster or clusters. The Mth GE level ‘20’ appears to be one of the major sources of variation in the data set. The role of Mth value has been explored by screening the genes with Mth values such that a data matrix X_{800*38} was extracted from the $X_{2299*38}$ which is devoid of the Mth value ‘20’ across all the 38 samples. A Box-Cox transformed extracted and main data matrices were checked for their data trends using the easy fit

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-10

software. The samples in the extracted data matrix followed the family the normal pattern whereas the later followed the Extreme Value distribution (EVD). PCA method 3 is applied on X_{800*38}. A 3D loading plot presented below exposes a non-distinctive cluster structure of the leukemia types.



The AML and ALL-B samples look like a big cluster. ALL-T group visibly ruptures into two sub clusters at the top and bottom of the figure). This scenario of cluster formation is entirely different from the presentation of 3 cluster structure in the 3D loading plots of $X_{2299*38}$ and $X_{1998*38}$.



B. The High Variant Group of Genes

The correlation pattern in the X_{301*38} the highly variant group of genes is in contrast from the positive correlation pattern of samples in Group II the homogeneous group of 1998 genes. The X_{301*38} possess a high abundance of the Mth value”20”. A gene that has the Mth value across all the samples in a cluster is not with the same pattern of GE levels in the other cluster or clusters. Using method 3 of PCA the 3D principal Component loading plot of the extracted matrix was successful in identifying the three mutually exclusive clusters of ALL-B, ALL-T and AML.

Apparently it is no cluster structure without Mth and a clear 3 cluster structure with abundance of Mth Values. For which the consistency of GE level in each of the three Clusters (C1, C2, and C3) checked with Garcia criterion of RV Table 4.

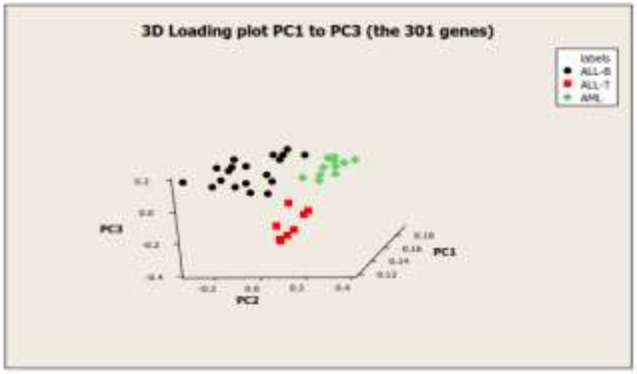


Table 4. The Garcia RV Criterion	$C1_{301\times19}$ (Genes)	$C2_{301\times08}$ (Genes)	$C3_{301\times11}$ (Genes)
$RV < \mu_{RV} - \sigma_{RV}$	RV<0.75 (43)	RV<0.24 (58)	RV<0.56 (51)
$\mu_{RV} - \sigma_{RV} < RV \le \mu_{RV} + \sigma_{RV}$	0.75<RV≤2.34 (207)	0.24<RV ≤1.55 (198)	.56<RV ≤2.00 (205)
$\mu_{RV} + \sigma_{RV} < RV \le \mu_{RV} + 2\sigma_{RV}$	2.34<RV≤ 3.14(40)	1.55<RV ≤2.20 (32)	2.00<RV≤ 2.72(40)
$RV > \mu_{RV} + 2\sigma_{RV}$	RV>3.14(11)	RV>2.20 (13)	RV>2.72 (05)

The RV of   genes is highly susceptible to the proportion of Mth value ‘20’ across the samples. The facts materialize as follows

a. With RV >1 and the proportion of ‘20’ more than 50% it was observed that atleast one sample possessed extreme high GE level or the RV was extremely high than 1 with GE level almost uniformly around ‘20’.

b. With RV >1 but with proportion of ‘20’less than or equal to 50% either there was a consistent pattern of GE levels across the samples or one or two of the samples were with very high GE levels.

c. For $RV \le 1$ irrespective of the proportion of ‘20’ either the GE levels of a gene were centered around ‘20’ or the GE levels were closely clustered around its mean mostly in two significant digits.

These facts lead to the development of a feature selection discriminant rule defining three functional groups of gene. The Dormant (Drt) group of genes (a passive gene for the particular disease type), Functionally dominant (Fd) genes and Sporadically dominant (Spd) genes (a type of active gene group with one extremely high GE level may be due to some sudden unknown epidemiological factor).The logical summarization is given below as

If $\bar{g}_i \le 99$, is classified as “Drt” ELSE If $(OR(AND(\bar{g}_i \le 99, RV > \mu_{RV} + 2\sigma_{RV})$, $AND(\bar{g}_i > 99, MthValue'20' > 72\%))$, classified as ‘Spd’ ELSE ‘Fd’) where $\bar{g}_i$ is the mean of the ith gene across all samples. The 301 genes in each of the cluster were classified with the developed feature selection discriminant rule.

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-10

C. *Discrimination Analysis*

The feature selection discriminant rule is validated with quadratic method of discriminant analysis (QDA) since the variance covariance matrices (Σ) of the functional groups were not homogeneous. The homogeneity of the three functional groups was tested with Log Determinant method. Larger the log determinants the more variable are the functional groups. Table 3. Show large differences in the log determinants of the functional groups asserting the variance covariance matrices not homogeneous.

Table 3 Log Determinant Method for the Covariance Matrices of the Three Functional Groups

	Log $\lvert\Sigma\rvert$ ALL-B	Log $\lvert\Sigma\rvert$ ALL-T	Log $\lvert\Sigma\rvert$ AML
Drt	166.958 (144)	63.259 (184)	92.518 (123)
Fd	246.707 (134)	116.082 (97)	163.373 (156)
Spd	195.993 (23)	89.996 (20)	114.417 (22)

( ) gives the classified no of genes.

"Box M test is quite powerful for moderate to large samples meaning it may find trivial differences between the Variance covariance matrices" [22]. The homogeneity of variance covariance matrices is further tested with Box M test.

Table 4. Test Results of Box's *M*

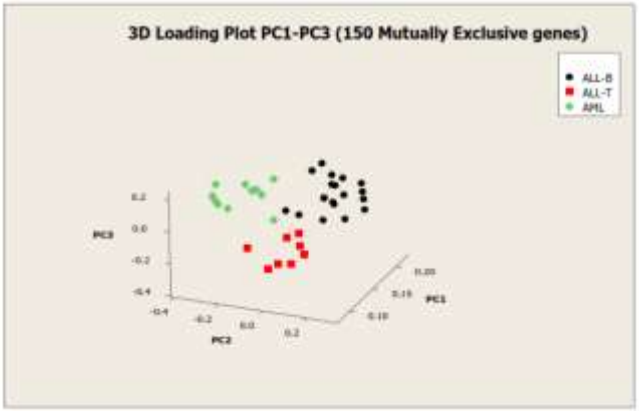
	Box's M	F-test	df1	df2	Sig
$C_{1_{301\times19}}$	9197.949	19.577	380	14827.571	0.000
$C_{2_{301\times08}}$	7733.913	96.731	72	8603.435	0.000
$C_{3_{301\times11}}$	7603.459	50.404	132	10262.942	0.000

The main objective of performing DA independently for each cluster is to find the Drt genes in each cluster so as to segregate specifically the Fd or Spd genes for each leukemia type.

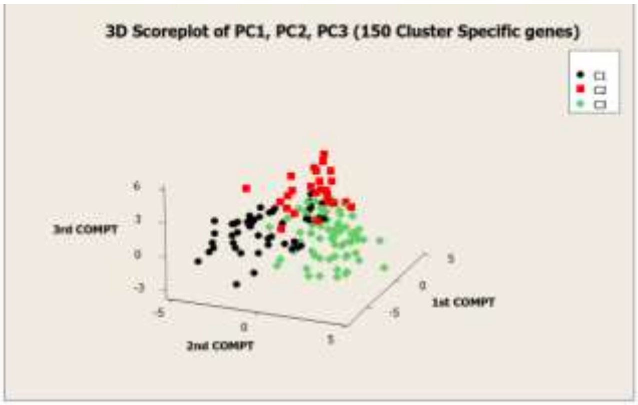
Table 5. QDA Output for Each Functional type of Gene.

	True Group		
Put into group	**Drt**	**Fd**	**Spd**
Drt in ALL-B	**139**	**18**	**00**
Fd in ALL-B	**02**	**113**	**00**
Spd in ALL-B	**03**	**03**	**23**
Classified total 301	**144**	**<u>134</u>**	**<u>23</u>**
Correct proportion of identification	**139/144 =0.97**	**113/134 =0.84**	**23/23 =1.00**
Total Correct Proportion	**275/301=0.91**		
Drt in ALL-T	**175**	**07**	**00**
Fd in ALL-T	**01**	**62**	**00**
Spd in ALL-T	**08**	**28**	**20**
Classified of Total 301	**184**	**<u>97</u>**	**<u>20</u>**
Correct proportion of identification	**175/184 =0.95**	**62/97 =0.64**	**20/20 =1.00**
Total Correct Proportion	**257/301=0.85**		
Drt in AML	**115**	**07**	**00**
Fd in AML	**01**	**126**	**00**
Spd in AML	**07**	**23**	**23**
Classified of total	**123**	**<u>156</u>**	**<u>23</u>**
Correct Discrimination	**115**	**<u>126</u>**	**<u>23</u>**
Correct proportion of identification	**115/123 =0.94**	**126/156 =0.81**	**23/23 =1.00**
Total Correct Proportion	**243/301=0.87**		

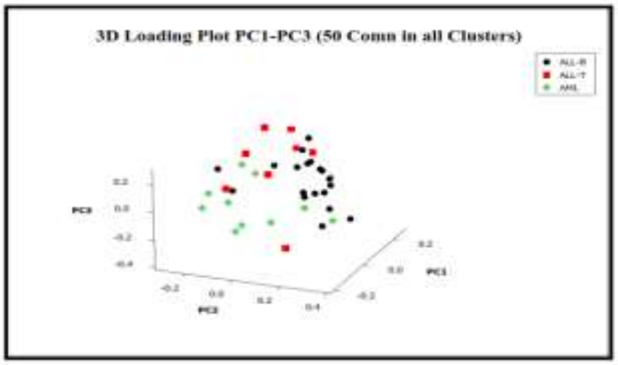
At least 85% of the genes are correctly identified in their functional groups in each of the three clusters. Apart from the Drt genes the Spd or Fd genes are not different in their role to the leukemia diseased samples being the contributors to deleterious mutations. Among the high variant gene group 30 genes were commonly Drt in all the three clusters. 50 genes were commonly Fd/Spd in all clusters. 71 Fd/Spd genes were common in any two of the three clusters. 49 Fd or Spd genes in $C_{1_{301\times19}}$, 27 in $C_{2_{301\times08}}$ and 74 in $C_{3_{301\times11}}$ i.e 150 genes were purely mutually exclusive in being specific to their cluster type. The 3D loading plot of 150 mutually exclusive genes singularly identify the three leukemia types.

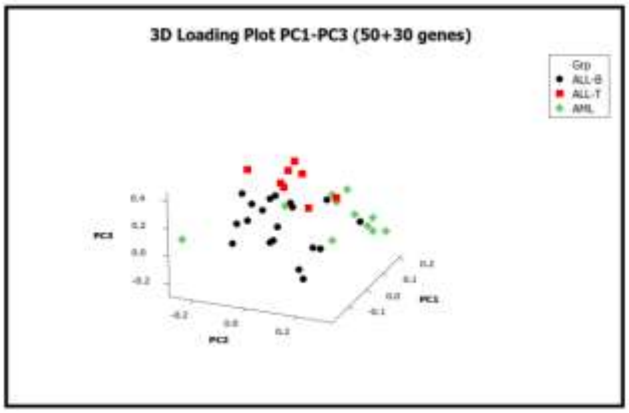


The 3D Component score plot below asserts that the three clusters are mutual exclusive with respect to their genes. However the slight mingling in the middle of the plot may be attributed to less than 100% correct identification.



The 50 functional and 30 Drt genes common across the three clusters do not make any show of specific gene or cluster in the respective 3D presentation below.





III. DISCUSSION

An empirical demonstration of PCA on Box-Cox transformed data using correlation matrix is successfully shown to cluster the 38 leukemia samples each with 2299 genes. The salient features of the Golub *et al.* (l999) data have been explored and scrutinized such that the data itself has provided a discriminatory footing for identifying the marker genes. The data in the study is available on the website with accession number of each gene. These results are not known in literature before to the best of our knowledge. The simple methodology of finding the feature selection rule along with appropriate DA discriminant may be applied on other gene data sets like Alon *et al*. (1999) of over 6500 human genes in 40 tumor and 22 normal colon tissue samples.

REFERENCES

[1] Pepe, M.S., Longton, G., Anderson, G.L., and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59,** 133-142.

[2] Golub T.R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeck,M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, j.R., Caligiuri, M.A., Blommfield, C.D., and Lander, E.S.(1999). Molecular classification of Cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

[3] Wajid R.A and Ali A (1998) Relative Variance Covariance as an alternative to covariance and correlation matrix in Principal Component Analysis. Journal of Statistical Sciences 17, 1,45-51.

[4] Dudoit, S., Fridlyand, J., snd Speed, T.P. 92002a).Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Standard Association 97, 77-87.

[5] Antoniadis, A S Lambert-Lacroix, F Leblanc (2003). Effective dimension reduction methods for tumor classification using gene expression data", Bioinformatics, 19:563- 570.

[6] Chiaromonte, F., and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. Math. Biosci. 176(1), 123–144.

[7] Ding, C.H.Q. (2003). Unsupervised feature selection via two-way ordering in gene expression analysis. Bioinformatics 19, 1259–1266.

[8] Jaeger J, Sengupta R, Ruzzo W. L (2003), Improved gene selection for classification of microarrays", Pacific Symposium on Biocomputing, 8:534.

[9] Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., and Mallick, B.K. (2003). Gene selection: A Bayesian variable selection approach. Bioinformatics 19, 90–97.

[10] Li, H., and Hong, F. (2001). Cluster-Rasch models for microarray gene expression data. Genome Biol. 2(8), research 0031.

[11] Li, J., Liu, H., Downing, J.R., Yeoh, A.E.J., and Wong, L. (2003). Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics 19, 71–78.

[12] Li, W., and Yang, Y. (2002a). How many genes are needed for a discriminant microarray data analysis? In Methods of Microarray Data Analysis: Papers from CAMDA'00, 137–150, Kluwer Academic, NY.

[13] Lu, J., Hardy, S., Tao, W.L., Muse, S., Weir, B., and Spruill, S. (2002). Classical statistical approaches to molecular classification of cancer from gene expression profiling. In Methods of Microarray Data Analysis: *Papers from CAMDA'00*, 97–107, Kluwer Academic, NY.

[14] Model, F., Adorjan, P., Olek, A., and Piepenbrock, C. (2001). Feature selection for DNA methylation based cancer classification. Bioinformatics 17(Suppl. 1), S157–S164.

[15] Nguyen, D.V., and Rocke, D.M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18, 39–50.

[16] Xiong, M., Li, W.J., Zhao, J., Jin, L., and Boerwinkle, E. (2001). Feature (gene) selection in gene expression-based tumor classification. Mol. Genet. Metabolism 73(3), 239–247.

[17] Siedow, J.N. (2001). Making sense of microarrays. (meeting report), Genome Biol. 2(2), reports 4003.

[18] Broberg, P. (2003). Statistical methods for ranking differentially expressed genes. Genome Biology. 4, R41.

[19] Siedow, J.N. (2001). Making sense of microarrays. (Meeting report), Genome Biol. 2(2), reports 4003.

[20] Smyth, G.K., Yang, Y.H., and Speed, T. (2003). Statistical issues in cDNA microarray data analysis. In Functional Genomics: Methods and Protocols, Methods in Molecular Biology Series, vol. 224, 111–136, Humana Press, Totowa, NJ.

[21] Garcia, C.H. (1989). Tables for classification of the coefficient of variation. Piracicaba: IPEF, 12p. (Technical Circular, 171).

[22] Paul R. Swank (2008). Internet response to the query to use of Box's Test for non-normal data. Professor and Director Research child learning Institute University of Texas Health Centre Housten.