Reason- A Technical Journal
ISSN 2277-1654
Volume XIX 2020

METABOLIC PATHWAY OF HEREDITARY CANCER DISEASE FROM PPI-NETWORK OF DEGS DETECTED USING MEAN-OF-MEAN METHOD

Bandana Barman
Department of Electronics and Communication Engineering, Kalyani Government Engineering College, Kalyani- 741235, Nadia, West Bengal.
Email: bandanabarman@gmail.com

Paper received on: January 01, 2021, accepted after revision on: April 18 2021
DOI:10.21843/reas/2020/64-80/209273

Abstract: Uncontrolled growth of cells often results Cancer disease in human body. When it is eventually transmitted throughout the generations in a family, it is referred to as hereditary diseases. Metabolic pathway explains several chemical reactions occurred for growth of a disease and KEGG pathway analysis identifies key genes involved with that disease. At first, differentially expressed genes (DEGs) are detected from cancer gene microarray time series datasets using a new and simple method, Mean of Mean (MoM). The MoM concept is developed from Gregor Johann Mendel's First Law of Heredity or Segregation rule of heredity. Highly expressed (HG) and lowly expressed (LG) genes in two different groups of microarray dataset are identified first using MoM. Then DEGs are found by implementing intersection operation between HG and LG genes. Performance of MoM method is analyzed by Support Vector Machine classifiers (SVMs) on some binary class cancer microarray data samples. Then all results are compared with the performance of other statistical parametric and nonparametric hypothetic tests. It is noticed that performance of MoM is better than other statistical methods in almost all data sets. Finally, Protein-Protein Interaction Networks (PPINs) are constructed within identified DEGs using web based tool. Lastly, KEGG pathway analysis is performed for all proteins involved in PPINs to obtain list of key genes for growth of cancer disease.

Keywords: Hereditary Disease, Cancer, Mean of Mean (MoM), DEGs, SVM Classifier, PPI-Networks, Metabolic Pathway, KEGG Pathway.

1. INTRODUCTION

Genes are biomolecules consist of double stranded Deoxyrioboneuclic Acids (DNA) and proteins are formed from DNA after entire central dogma process happens. The central dogma consists of four hormonal steps: 1. Replication, 2. Transcription, 3. Splicing and 4. Translation. Protein can interact with other proteins, genes or biomolecules to construct a gene regulatory network (GRN). Each gene has its unique specialty. When a special character of a gene is carried throughout some generations, it is defined as hereditary characteristics. When an unhealthy condition of gene is carried out continuously in several generations, it is called genetic disorders or inherited diseases. It may be happened due to abnormal functionalities of single or a list of proteins induced from genes. If gene mutation happens very noticeably (drastic change or abnormal growth), then there is a

chance to occur cancer disease. The familial cancer is a hereditary disease which happens due to inherited gene mutation. It is found in literature that almost 8% of different cancers are occurred due to gene mutation or defects. Breast Cancer, ovarian cancer in women is very common in familial cancer. The other familial cancer names are Leukemia (acute lymphocytic leukemia (ALL) and acute myelogenous leukemia (AML), Chronic lymphocytic leukemia (CLL) Chronic adult leukemia, Chronic myelogenous leukemia (CML), Melanoma (Skin cancer), Pancreatic cancer, Prostate cancer, Lymphoma (occurred due to infection in immune system), Colorectal or Colon cancer, etc. Sometimes, cancer syndrome starts due to development of a number of multiple independent tumors or tumor like cell growth and lastly it increases risk of cancer due to abnormal growth of cell. It happens as protein-protein interaction network (PPIN) is constructed between proteins of normal and abnormal genes. It is also seen that though one inherited genetic defect is passed on to members within a family does not develop cancer for everyone as every person's genome sequence is different from each other.

DEGs detection from microarray data is important as it indicates the changes in expression level in 2-sample groups of data. Different parametric and nonparametric statistical methods are used to detect DEGs [4,6]. Statistical t-test, Pearson's correlation tests (corr), Analysis of variance 1 (ANOVA 1) tests are parametric hypothesis tests. Non-parametric statistical hypothesis tests are Permuted T-test (perm), Wilcoxon Ranksum test (RST), Modified Wilcoxon Ranksum test (ModRST or MRST), Significance Analysis of Microarray (SAM), Linear Models for Microarray Data (Limma), Shrink-t, Softthreshold-t (soft-t) tests. The other statistical tests to find DEGs are Kruskal-Wallis test (KW test), Ideal discriminator (i.e., ID) method, Kolmogorov-Smirnov test (KS test) [1]. The ranking analysis method (RAM) is also used to find DEGs [15]. The DEGs in microarray data in a principal component space is identified [11,19]. To detect DEGs for RNA-seq data some statistical methods [3, 14] are used in literature.

Detection of different activity of genes in a cell helps researchers to investigate the concept about normal and abnormal functions in cells. To understand the function of a disease in human body, microarray technology helps researchers a lot [20, 21]. Classification of several types of cancer with the development of tumor in an organ of human body is understood by genes activity with different patterns. Microarray technology helps researcher to design a model for treatment of cancer disease as culprit genes of a disease can be identified with this technology [5, 13].

Classification process organizes data into different categories by arranging them into groups or classes. It is done with effectiveness and efficiency of the data. By classification method, essential data can be identified and also can retrieve by finding the common characteristics in dataset. This is also important for risk management. Classification is important also for legal discovery. The classification is done based on, Qualitative, Quantitative, Geographical and Chronological or Temporal Base. There are three types of classification method.

Those are (1) one-way classification, (2) two-way classification and (3) multi-way classification [17, 10].

The machine learning process involves in data mining and it is also an artificial intelligent method. It finds data pattern, understand computer program for data pattern identification and also modify the program accordingly. A training dataset contains observations or instances of known members. The classification of new observations is done based on training dataset and by using machine learning algorithms. Those are supervised, semi-supervised, unsupervised and reinforcement learning. In supervised algorithm training data is input data with a known label. In unsupervised learning process, unlabeled input data with unknown result is used. Here, one model is designed with the structure of input data [9]. In semi-supervised learning algorithm, input data is the combination of unlabeled and labeled data. In reinforcement learning algorithm, input data is as stimulus to a model from an environment to which model respond and reacts [16].

Support vector machines (SVMs) are supervised learning algorithm. SVMs analyze and recognize data with its patterns. It is used to classify and also for regression analysis. In linear separation of the training set, a separating hyperplane is defined as in Eq. (1).

$$hyperplane = x|\langle w, x\rangle + b = 0 \qquad (1)$$

w, b are normal vector and offset respectively. $\langle .,. \rangle$ is the inner product which is scalar or dot product; w, b are picked from training set to train data. Label of a new point is identified as follows;

1. The 'POSITIVE' classified Points are the points which are in hyperplane normal vector's direction, 2. The 'NEGATIVE' classified Points are the points which are in opposite direction of normal vector. The training set is separated by keeping maximal margin from separating hyperplane. The points, Support Vectors (SVs) are closest to separating hyperplane. Only SVs determine the position of hyperplane.

In this paper, a new method is proposed to find DEGs from microarray data. The algorithm is based on finding mean of mean (MoM) of datasets. Other state of art statistical methods, e.g. Statistital t-test, significance of microarray test (SAM), Wilcoxon Ranksum test, Pearsons correlation test (corr), and Analysis of variance (ANOVA 1) test on microarray data sets are also performed to detect DEGs. All these tests and the proposed MoM algorithm are implemented on different types of cancer microarray data to detect DEGs. Then efficiency of algorithms for identified DEGs are classified with support vector machines (SVMs) classifier to analyze performance of all testing methods. In next section, data preprocessing method is stated.

2. PRE-PROCESSING METHOD OF DATASETS

All methods are performed on two-class real life data sets, which are in 2D-matrix format. The matrix columns show genes and rows represent sample values. At first mean value and standard deviation (S.D.) of two class data are calculated. Then signal to noise

ratio, SNR of each column is calculated. The |SNR| is defined in Eq. (2).

$$|\mathrm{SNR}| = \left|\frac{\text{mean (cls 1)} - \text{mean (cls 2)}}{\text{S.D.(cls 1)} + \text{S.D.(cls 2)}}\right| \qquad (2)$$

After finding |SNR|, value of gene i.e. column is arranged in a descending order according to |SNR| value. High |SNR| means a wide range of data set value and low |SNR| means a low range of data value. Then normalization of data set with mean and standard deviation of total data set is done. It is defined in Eq. (3).

$$\text{Normalize Dataset} = \frac{DataSet - mean\ (DataSet)}{StandardDeviation} \qquad (3)$$

After normalization, all the tests to identify DEGs are implemented on scaled dataset.

3. PROPOSED ALGORITHM: DETECTION OF DEG WITH FINDING MEAN-OF-MEAN (MoM)

The concept of identifying the DEGs are developed from the first law of genetics by Gregor Johann Mendel which is shown in Fig. 1. According to law of hereditary, father has 46 chromosome of which 44 autosome and 2 sex chromosome (XY) whereas mother has 46 chromosome (44 autosome and 2 sex chromosome XX). Father's sex chromosome (XY) is divided and forms haploid gametes X and Y (i.e., Element (X) and Element (Y) in Fig.1). On the other hand, mother's sex chromosome is divided to form haploid ova X and X (i.e., both Element (X) in Fig.1.). When X chromosome containing gamete or sperm fertilizes ova containing X chromosome (means the crossover process in Fig.1.), the progeny will be XX (i.e., Offspring1 and 3 in Fig.1.), i.e., female characteristic. But if Y chromosome containing sperm fertilize X chromosome containing ova (Crossover process) then progeny will be XY (i.e., Offspring 2 and 4 in Fig.1.), i.e., male characteristic. So, determination of sex in the next progeny depends on chromosome comes from the father not from the mother.

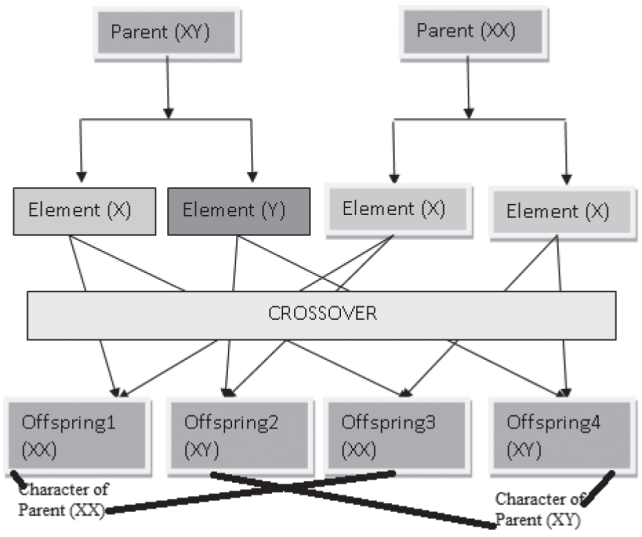


Fig.1. The first law of Genetics by Gregor Johann Mendel's

3.1 Concepts of Differentially Expressed Genes (DEG)

A gene is called differentially expressed (DE) if its pattern (spatial or temporal) of expression vectors is varying in different phenotypes. If a gene is in 'on' stage in a phenotype and 'off' stage in another phenotype, that particular gene is called differentially expressed. If genes have a varying dependence between different phenotypes, then those genes are called DEG.

3.2 Proposed Algorithm

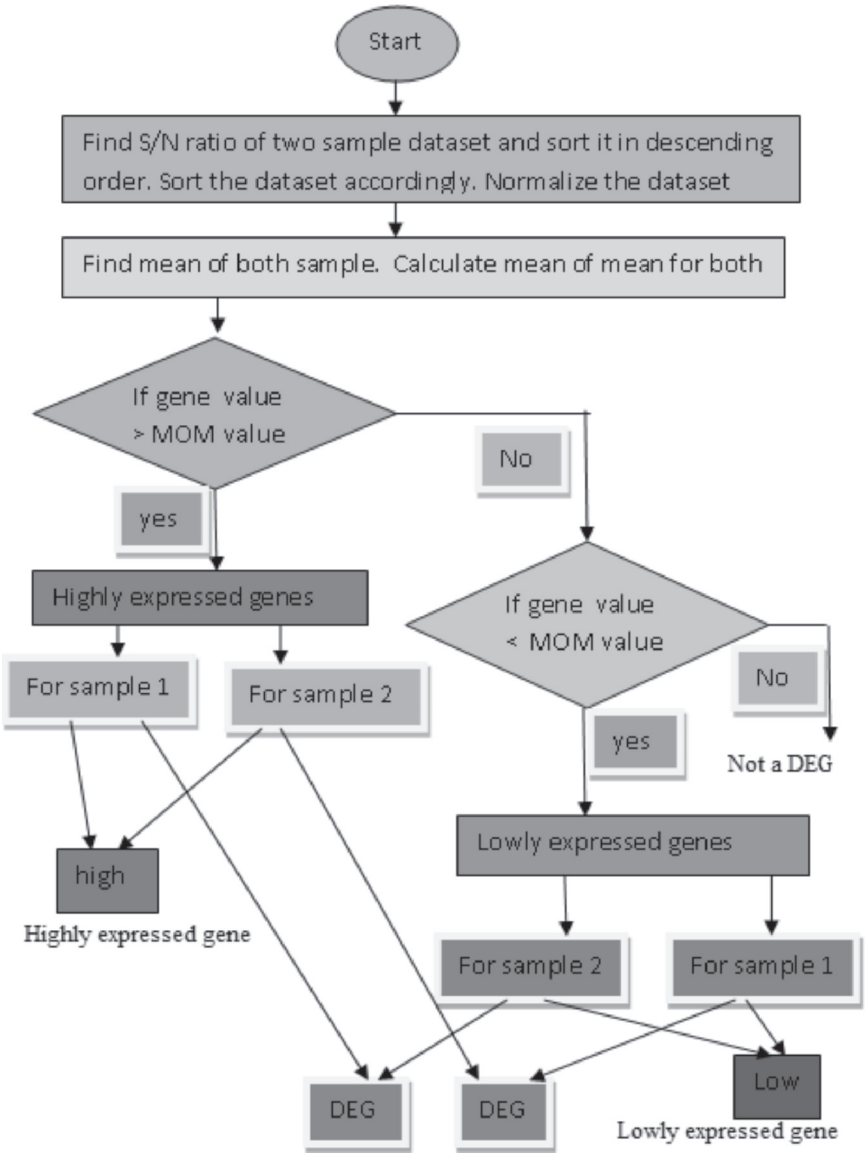
In proposed algorithm, two types of expression vectors (gene expression levels) viz., “high” and “low” in different phenotypes are considered. Based on these two expression categories DEGs are identified. The binary class (two-class) sample microarray data to perform proposed algorithm is taken. In Figure 2 a conceptual flowchart of proposed algorithm is shown. Data preprocessing and normalization are performed using the method discussed in Section 2. According to MoM method mean of each gene expression is calculated then mean of all genes mean value (MoM) is calculated. If a gene has expressed mean value > MoM value then that gene is called a Highly expressed gene otherwise that gene will be Lowly expressed i.e. these are as Elements shown in Fig.1. If one gene is highly expressed in both sample (i.e. Sample 1 and 2 in Fig.2) then that gene will be referred to as HIGH. When one gene is highly expressed in sample 1 but lowly expressed in sample 2, it is called a differentially expressed gene (DEG). If a gene is lowly expressed in both samples then that gene is referred to as LOW. When one gene is lowly expressed in sample 1 and highly expressed in sample 2 that gene is also mentioned as a DEG. The entire method is shown in the Fig.2. The DEGs, high and low genes are detected by the set theory operation; intersection (i.e., crossover process mentioned in Fig.1).

All steps of the proposed algorithm are as follows.

Algorithmic Steps:

Input: Binary Class (Two class) Sample microarray data, let X1 and X2;
Output: Differentially expressed genes (DEGs);
Step 1: After preprocessing, Normalized Samples are NX1 and NX2.FindMean, m of each gene of normalized samples.
m1= mean(NX1,1) and m2=mean(NX2,1);
Step 2: To find mean of all gene's mean value, M, i.e., MoM from two samples, M1=mean(m1), M2=mean(m2);
Step 3: If one gene's m > M then that gene is highly expressed, HEG.
Step 4: If one gene's m<M then that gene is lowly expressed, LEG.
Step 5: The HEGs and LEGs are separated from Sample 1 and 2.
Step 6: Intersection process is performed between HEGs and LEGs of sample 1 and 2.
Step 7: HIGH, LOW and DEGs are detected after performing Step 6.
Step 8: If sample 1's LEG becomes HEG in sample 2 then it is called upregulated DEG and If sample 1's HEG becomes LEG in sample 2 then that gene is called down regulated DEG.

In the next section, some state of the art statistical tests to detect DEGs are explained.



4. STATE-OF-ART STATISTICAL METHODS USED TO COMPARE THE PROPOSED ALGORITHM

In this section the principal of Statistical t-tests (T-test), Wilcoxon Ranksum test (RNK), Significance of Analysis of microarray (SAM), Pearson Correlation coefficient (CORR) and Analysis of Variance 1 (ANOVA 1) methods are explained briefly.

4.1 Statistical t-tests (T-test)

Standard statistical t-test finds a difference between two groups. Here, Statistical significance is calculated by finding difference of group's averages, sample size, and standard deviations. If each sample has same no. of genes (n), then each sample's distribution will have same variance. It is explained by Eqn. 4 and 5.

$$D = \frac{mean\ (X1) - mean\ (X2)}{Std\ (X1X2) Sqrt\ (2/n)} \quad (4)$$

$$Std(X1X2) = \text{Sqrt}(\tfrac{1}{2}(S^2_{X1} + S^2_{X2})) \quad (5)$$

Std(X1X2) is accumulative standard deviation of two sample groups. S^2_{X1} and S^2_{X2} are unbiased estimators of variances. Mathematically, t-score is calculated as written in Eq. (6).

$$T = \frac{mean\ (X) - \mu}{S/\text{sqrt}\ (n)} \quad (6)$$

mean(X), μ and S are sample mean, population mean, and standard deviation of sample respectively. False discovery rate (FDR) is estimated from list of findings. The FDR is computed using Eq. (7).

$$FDR = E\frac{V}{R} \quad (7)$$

Here, *V* is total false positives. *R* is total rejected null hypothesis and *E denotes* estimation. The *FDR* is retained below Quantile value (test significant q-value or threshold value).

4.2 Wilcoxon Ranksum test (RST)

In Wilcoxon ranksum test, let two sample groups are *X*1 and *X*2. Population size is same for both samples, *n*. At first, ranking of values of two samples are done in an ascending order. Then, summation of all ranks, i.e. statistic (*W*), is calculated for group *X*1 using Eq. (8).

$$W = \Sigma(RankX1) \quad (8)$$

W is an integer value and is called R-statistics. Its value is < total of rank of *X*2 only when all *X*1's value are <*X*2 values so that all values of *X*1 will always be at the beginning of total list as shown in Eq. (9).

$$W >= 1 + 2 + \ldots\ldots + nX1 = [nX1(nX1 + 1) / 2] \quad (9)$$

Test-statistic, *W* will be maximum when *X*1 values placed after all *X*2 values, as in Eq. (10).

$$W =< (nX2 + 1) + (nX2 + 2) + \ldots_{+} (nX2 + nX1) \quad (10)$$

When *W* extreme value, *X*1 and *X*2 are different. If *W* is at its middle value then it is considered that *X*1 and *X*2 are indifferent. Normal approximation is needed for large sample size, n. When n is small, null distribution of test-statistic is exactly calculated. If population size in *X*1 and *X*2 are equal then RST statistic (test Statistic), *W* is the minimum of RSTs. When population size of *X*1 and *X*2 are not equal, test statistics is equal to sum of all ranks for identified low valued sample. If *W* has a small value then null hypothesis is rejected. Significance levels for small values in both samples are tabulated. The p-value is calculated using Eq. (11).

$$p\text{-}value = min(2 * min(RST\ statistic), 1) \quad (11)$$

If population size is *n* large value then a z-statistic is calculated using Eq. (12).

$$Z = \frac{(|T - meanW\ | - 0.5)}{Sqrt\ (varW\ )} \quad (12)$$

4.3 Significance Analysis of Microarray (SAM)

SAM is a permutative method and by performing SAM test, the significant genes are sorted out. It performs gene specific t-tests. It also calculates each gene's

d-statistic (as shown in Eq. (13)). The d-statistic finds strength of relationship within gene expression and response variable which describes and groups experimentally conditioned data.

$$\mathrm{d-statistics} = \frac{mean\ (sample\ 1) - mean\ (sample\ 2)}{Seg + S0} \quad (13)$$

Seg is standard deviation. *S0* is called 'fudge factor'. The value of*S*0 is selected for minimizing d-statistic's coefficient of variation. In SAM test, gene expression values between two groups are randomly shuffled. The d-value is computed for each randomized grouping. SAM allows a dynamically changing threshold for significance (by tuning parameter delta) and median number of significant gene which is obtained from all permutations results median False Discovery Rate (FDR) as shown in Eq. (14). SAM statistic is likely same as t-statistics.

False discovery rate(*FDR*)

$$= \frac{median\ \ of\ \ falsy\ \ called\ \ genes}{\text{number of genes called significant}} \quad (14)$$

4.4 Pearson Correlation coefficient (CORR)

The type of correlation can be categorized in three ways. Those are 1. Positive correlation, 2. Negative correlation, 3. No correlation. CORR, i.e., ρ statistically measures relationship strength of paired data as shown in Eq. (15). Where, two groups of data, i.e., *x* and *y*, having *n* samples, mean(*x)* and mean(*y)* are their mean value. δ*x* and δ*y* are standard deviation of two groups respectively.

$$\rho = \frac{\Sigma i=1:n\left((x - meanx\ )(y - mean\ \ y)\right)}{\delta x\, \delta y} \quad (15)$$

The CORR can only predict the connectivity between two variables, i.e., causality. The t-statistic for CORR, ρ is: t = /(Sqrt(1-2)/n-2), where (n - 2) is degree of freedom (df). For single group, df is one.

4.5 Analysis of Variance 1 (ANOVA 1)

The ANOVA 1 compares between means of three or more samples using F distribution. It tests null hypothesis in all groups. ANOVA calculates F-statistic which is a ratio of variance within means to the variance within sample groups. Grand mean is the weighted mean of all sample means as stated in Eq. (16).

$$\text{mean(Xgm)} = \Sigma n(\text{mean(x)}) \,/\, N \quad (16)$$

The between group variance, *SS*(*T*) defines interaction between samples as in Eq. (17) and Sum of Squares Between groups, SS(B) (Shown in Eqn.18.) means the interactive variation between samples. Mean Square Between groups, MS(B) means the interaction variance between samples. MS(B), is the quotient of between group variation when divided by DoF, S^2_b (DoF of each sample)). SS(W) means sum of squares within groups. It means variation because of differences within each sample as stated in Eq. (19).

$$SS(T) = \Sigma(x - \text{mean}(Xgm))\text{^}2 \quad (17)$$
$$SS(B) = \Sigma n(\text{mean}(x) - \text{mean}(Xgm))\text{^}2 \quad (18)$$
$$SS(W) = \Sigma d\ f.s^2 \quad (19)$$

Mean square within group, MS(W) is result of within group variation/its DoF,S^2_w. It is a weighted average of variances. Recall, F is a test statistic measures the ratio of two sample

variances and is found by dividing between group variance by within group variance.

$$F = S^2_b / S^2_w \quad (20)$$

When test statistic>F critical value, null hypothesis is rejected.

5. EXPERIMENTAL DATA SETS

The proposed algorithm is coded using Matlab R2017a and implemented on 10 sample microarray cancer disease datasets found in literature. Those datasets are; 1. Childhood ALL (GSE412) which has total 8274 genes with 110 samples, 2. Leukemia (Golub et al.) having 5147 genes and 72 samples, 3. Prostate cancer (Singh et al.) with 12533 genes and 102 samples, 4. DLBCL (Shipp et al.) having 7070 genes and 77 samples, 5. Medulloblastoma (GSE468) with 13 examples and total 1465 genes, 6. AML prognosis (GSE2191) containing 12625 genes with 54 samples, 7. Prostate cancer (GSE2443) dataset which has total 12627 genes with 20 samples, 8. Yeoh-2002-v1 having either T-ALL or B-ALL. Total samples are 43 T-ALL and B-ALL, 9. Armstrong-2002-v1 with a distribution of samples: 24 ALL and 48 MLL, 10. Golub-1999-v1 having 47 ALL and 25 AML Data.

The datasets are obtained as 2D matrices where rows represent total genes and columns represent sample with class. Preprocessing of the datasets is performed using Eq. (2) and normalization is performed by using Eqn. 3. Then the proposed algorithm, MoM is performed to detect DEGs. The performance of MoMis cross-validated using Support vector machines (SVMs) classifier. Then comparison between MoM results and other mentioned (in Section 4) statistical methods is done.

6. CROSS-VALIDATION BY SVMs

The support vector machines (SVMs) are used to classify binary class sample data and basic concept of SVMs is described in the Section 1.The support vector machines are supervised machine learning method. SVMs have a training set consists of class label and features. The goal of the classification is to identify the class of a new test data. Let, dataset of *G* genes (i.e. features) for total *n* observations *i.e.* mRNA samples

xi = *xi*_1, *xi*_2, ..., *xi*_*G* is the feature vector of sample, i.

yi = response or class of sample, i.

Where, *i* = 1, ... , *n*.

The separating hyperplane ($hyperplane = x|\langle w, x\rangle + b = 0$), separates positive from negative.

The DEG identification results are cross validated with supervised machine learning method SVMs [7, 8]. After that accuracy, specificity, sensitivity, F-score and area under curve (AUC) in ROC of all methods and proposed MoM algorithm are calculated. Let assume, in two class sample, P and N are the two classes. TP is true +ve, FP is false +ve, TN is true -ve and FN is false -ve [2].

Accuracy= $\frac{TP+TN}{TP+TN+FP+FN}$, Recall or Sensitivity= $\frac{TP}{TP+FN}$ Specificity= $\frac{TN}{TN+FP}$,

Fscore= (2**Precision* ∗*Precision)/(Precision*+*Recall)*, Precision= TP/(TP+FP) [18]. The AUC is the conversion of single value of ROC [12] as ROC plots true positive rate vs false positive rate.

7. EXPERIMENTAL RESULTS

Here, the results obtained after performing all experiments on publicly available ten different datasets are demonstrated. Table 1 shows comparative results all statistical methods and the proposed MoM algorithm by calculating accuracy, sensitivity, specificity, Fscore and AUC on ROC for all methods after performing cross validation by SVM classifier. It is seen that the proposed new algorithm, finding mean of mean (MoM) is a better testing algorithm among all methods. After investigating performance of proposed MoM algorithm, it is implemented and verified on Ovarian cancer dataset available from GEO website. Total genes present in this dataset are 54675. In the description of dataset it is mentioned that 15 samples were included in experiment with a 2x2 factorial design with 2 different cell lines (2008 and PEO4).

The dataset was taken as it was an experiment for determining Menopausal estrogen (E2) effects on tumor promotion for growth of ovarian cancer. MoM is implemented on dataset after preprocessing of dataset. After preprocessing total genes present in Ovarian cancer dataset is 21897. First DEGs are found between sample group 1 of PEO4 Estrogen Tx (Set A) and 2008 Placebo Control (Set B) and then between sample group 2 of 2008 Estrogen Tx (Set C) and PEO4 Placebo Control (Set D).

According to MoM algorithm all genes are divided into highly and lowly expressed genes i.e. HEG and LEG. After intersection operation (Crossover) total 1554 DEGs are found from sample group 1. From sample group 2 total 2343 DEGs are found. Then PPINs are generated from DEGs of sample Group 1 and 2 respectively using web based tool DAVID. Those are shown in Fig. 3 and 5 respectively. From 1554 DEGs in sample group 1, total 1114 gene IDs are found in DAVID tool database for human which produce proteins for generating PPIN (In Fig. 3.) and 27 genes are significant in KEGG pathways in cancer (in Fig. 4). From 2343 DEGs in sample group 2, 1787 genes are listed in DAVID database among which 11 genes are significant in long term protentiation (in Fig. 6) i.e. growth of cancer. In both PPINs proteins are considered as Nodes either colored or white. When one node is directly linked to another node the node is shown in colored. If a node reaches in higher iteration or depth then it becomes white. The connections are noted as edges which mean different functional links. Different colors of the edges mean type of evidence for association.

8. CONCLUSION

Identification of DEGs is important to get variations of a gene in the samples of microarray data. There are several statistical parametric and nonparametric tests to identify the DEGs. The proposed mean of mean (MoM) algorithm is a simple arithmetical algorithm to detect differentially expressed genes. After detecting DEGs by MoM algorithm, identified DEGs are classified with SVM classifier to analyze performance of this algorithm. The accuracy, speficity, sensitivity, Fscore, AUC for proposed MoM algorithm are compared with performance of statistical T-test, Wilcoxon ranksum test, SAM test, ANOVA 1, and

Pearson correlation test. It is found that MoM algorithm results good as compared with other statistical tests. Then MoM is implemented on sample overian cancer dataset and found DEGs. After that PPINs are constructed and KEGG pathways are developed with web based tool. It is found that some genes in the listed DEGs are associated in hereditary cancer disease.

REFERENCES

[1] Bandyopadhyay, S., Mallik, S. and Mukhopadhyay, A., A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data, IEEE/ ACM Transactions on Computational Biology and Bioinformatics, IEEE ACM, Vol. 11, No.1, 2013. DOI: 10.1109/TCBB.2013.147.

[2] Burges, C.J.C., A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, Vol. 2, pp.121–167, 1998.

[3] Chen, Z., Liu, J., Ng, H.K.T., Nadarajah, S., Kaufman, H.L., Yang, J.Y. and Deng, Y., Statistical Methods on Detecting Differentially Expressed Genes for RNA-Seq Data, BMC Systems Biology, Vol. 5, No.S1, 2011. DOI: 10.1186/1752-0509-5-S3-S1.

[4] Dudoit, S., Yang, Y.H. Callow, M.J. and Speed, T.P., Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments, Statistica Sinica, Vol. 12, No.1, pp.111-139, 2002.

[5] Ma, P., Zhong, W. and Liu, J.S., Identifying Differentially Expressed Genes in Time Course Microarray Data, Statistics in Biosciences, Vol. 1, No.144, 2009.

[6] Heydebreck, A.V., Statistical Tests for Differential Gene Expression, Technical Report.

[7] Lin, C.J., Support Vector Machines for Data Classification, Technical Report, National Taiwan University, February 9, 2004.

[8] Markowetz, F., Classification by Support Vector Machines, Technical Report, Max-Planck-Institute for Molecular Genetics, Berlin, 2003.

[9] Moore, A.W., Cross-Validation for Detecting and Preventing Overfitting, Technical report, Carnegie Mellon University.

http://www.cs.cmu.edu/~awm/tutorials.

[10] Mukherjee, S., Classifying Microarray Data Using Support Vector Machines, In: Berrar, D.P., Dubitzky, W. and Granzow, M. (Eds.) A Practical Approach to Microarray Data Analysis, Springer. DOI: 10.1007/0-306-47815-3_9

[11] Ospina, L. and Kleine, L.L., Identification of Differentially Expressed Genes in Microarray Data in a Principal Component Space, SpringerPlus, Vol. 2, No.60, 2013.

[12] Rakotomamonjy, A., Support Vector Machines and Area under Roc Curve, September 1, 2004.

[13] Smyth, G.K., Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments, Statistical Applications in Genetics and Molecular Biology, Vol. 3, 2004.

[14] Soneson, C. and Delorenzi, M., A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data, BMC Bioinformatics, Vol. 14, 2013.

[15] Tan, Y.D., Fornage, M. and Fu Y.X., Ranking Analysis of Microarray Data: A Powerful Method for Identifying Differentially Expressed Genes, Genomics, Vol. 88, No.6, pp.846-854, 2006.

[16] Vapnik, V.N., An Overview of Statistical Learning Theory, IEEE Transactions on Neural Networks, Vol. 10, 1999.

[17] Wang, J., Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications, Technical Report, Montclair State University, USA.

[18] Zhu, W., Zeng, N. and Wang, N., Sensitivity, Specificity, Accuracy, Associated Confidence Interval and Roc Analysis with Practical Implementations, Health Care and Life Sciences, Proceedings of NESUG 2010, 2010.

[19] Barman, B. and Mukhopadhyay, A., Extracting Biological Significant Subnetworks from Protein-Protein Interactions Induced by Differentially

Expressed Genes of HIV-1 Vpr Variants, International Journal of System Dynamics Applications, Vol. 4, No.4, pp. 35-51, 2015.

[20] Biswas, P. and Barman, B., An Approach to Identify Gene Markers Relating to Viral Carcinogenesis Using Data Mining Tools, Indian Science Cruiser, Vol. 33, No.2, pp.24-32, 2019. DOI: 10.24906/isc/2019/v33/i2/183890.

[21] Barman, B. and Mukhopadhyay, A., Detection of Differentially Expressed Genes in Wild Type HIV-1 Vpr and Two HIV-1 Mutant Vprs, FICTA, Vol. 1, pp.597-604, 2014

Table 1: Comparison of performance of proposed algorithm (MoM) with other standard statistical tests by support vector machines classifiers.

DataSets		T-test	Ranksum test	SAM test	ANOVA1 test	Corr test	Proposed Algorithm
childhood ALL (GSE412)	Accuracy	70.91	85.45	50.90	99.09	92.27	99.09
	Fscore	75	83.33	55.36	98.36	95.24	98.36
	Precision	68.92	88.64	46.30	100	100	100
	Sensitivity	76.12	87.30	53.45	99.17	97.56	99.17
	Specificity	85	91.67	51.67	100	100	100
	Area Under ROC	92.36	96.22	82.10	100	100	100
leukemia (Golub et al.)	Accuracy	72.20	73.61	76.39	98.61	97.22	84.72
	Fscore	79.60	80.41	81.72	98.95	97.87	88.17
	Precision	82.98	82.98	80.85	100	97.87	87.23
	Sensitivity	76.47	78	82.61	97.92	97.87	89.13
	Specificity	61.91	63.64	65.38	100	96	76.92
	Area Under ROC	93.12	94.30	94.56	100	99.95	96.21
prostate (Singh et al.)	Accuracy	91.18	88.23	96.08	93.14	90.20	94.12
	Fscore	90.91	87.76	96	93.07	90.00	93.75
	Precision	90	86	96	94	90	90
	Sensitivity	91.84	89.58	96	92.16	90	97.83
	Specificity	90.57	87.04	96.15	94.12	90.39	91.07
	Area Under ROC	97.34	95.91	98.12	97.82	95.33	98.91
DLBCL (Shipp et al.)	Accuracy	92.21	92.21	96.10	98.70	97.40	97.40
	Fscore	94.74	94.74	97.39	99.15	98.28	98.28
	Precision	93.10	93.10	96.55	100	98.28	98.28
	Sensitivity	96.43	96.43	98.25	98.31	98.28	98.28
	Specificity	80.95	80.95	90	100	94.74	94.74
	Area Under ROC	97.15	97.32	99.51	100	99.56	99.94
medulloblastoma (GSE468)	Accuracy	86.96	34.78	86.96	86.96	82.61	82.61
	Fscore	85.71	28.57	84.21	84.21	77.78	80
	Precision	90	30	80	80	70	80
	Sensitivity	81.82	27.28	88.89	88.89	87.50	84.62
	Specificity	91.67	41.67	85.71	85.71	80	84.62
	Area Under ROC	97.53	88.30	96.40	96.55	93.78	96.42
AML prognosis (GSE2191)	Accuracy	70.37	83.33	87.04	83.33	74.07	83.33
	Fscore	70.37	84.21	87.72	83.64	75	84.75
	Precision	67.86	85.71	89.29	82.14	75	89.29
	Sensitivity	73.08	82.76	82.76	85.19	75	80.65
	Specificity	67.86	84	88	81.48	73.08	86.96
	Area Under ROC	90.30	97.76	96.41	95.29	92.10	97.78

prostate cancer (GSE2443)	Accuracy	100	65	100	100	100	95
	Fscore	100	58.82	100	100	100	94.74
	Precision	100	50	100	100	100	90
	Sensitivity	100	71.43	100	100	100	100
	Specificity	100	61.54	100	100	100	90.91
	Area Under ROC	100	90.38	100	100	100	99.80
Yeoh-2002-v1	Accuracy	90.73	91.13	97.18	96.77	99.60	98.00
	Fscore	94.40	94.66	98.30	98.06	99.76	98.77
	Precision	94.63	95.12	98.54	98.54	100	98.05
	Sensitivity	94.17	94.20	98.06	97.58	99.51	99.50
	Specificity	73.81	75.61	92.86	92.68	100	91.30
	Area Under ROC	95.73	95.64	99.52	99.38	100	99.57
armstrong-2002-v1	Accuracy	98.61	98.61	88.89	97.22	98.61	97.22
	Fscore	97.87	97.87	81.82	95.65	97.87	95.65
	Precision	95.83	95.83	75	91.67	95.83	91.67
	Sensitivity	100	100	90	100	100	100
	Specificity	97.96	97.96	88.46	96	97.96	96
	Area Under ROC	99.87	99.92	97.58	99.84	99.82	99.75
golub-1999-v1	Accuracy	94.44	93.06	72.22	97.22	97.22	97.22
	Fscore	95.56	94.62	78.72	97.82	97.87	97.87
	Precision	91.49	93.62	78.72	95.74	97.87	97.87
	Sensitivity	100	95.65	78.72	100	97.87	97.87
	Specificity	86.21	88.46	60	92.60	96	96
	Area Under ROC	99.36	99.80	90.47	94.77	99.61	99.58

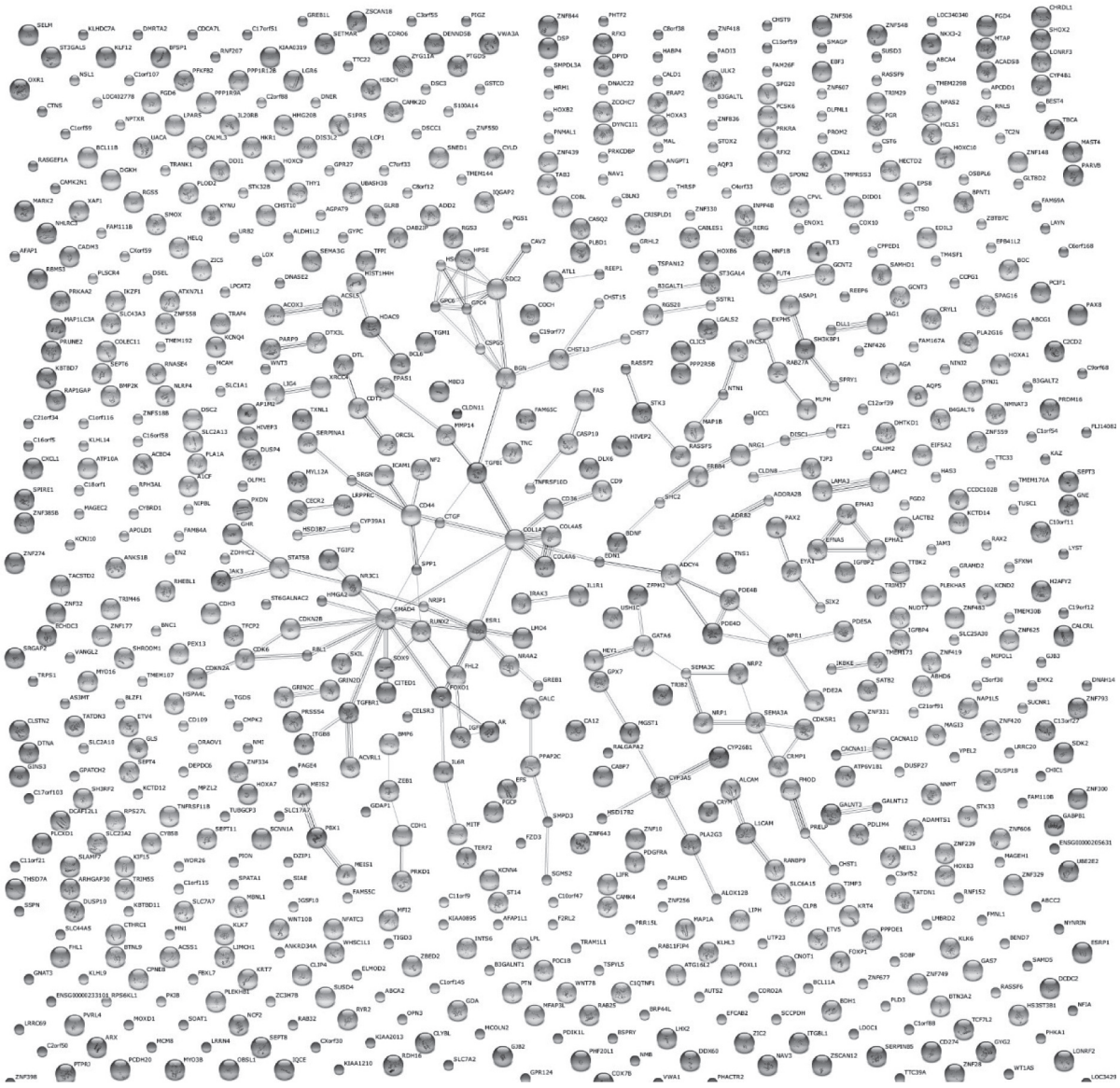


Fig.3. The PPI-Network evidence view of DEGs of Sample Group 1.

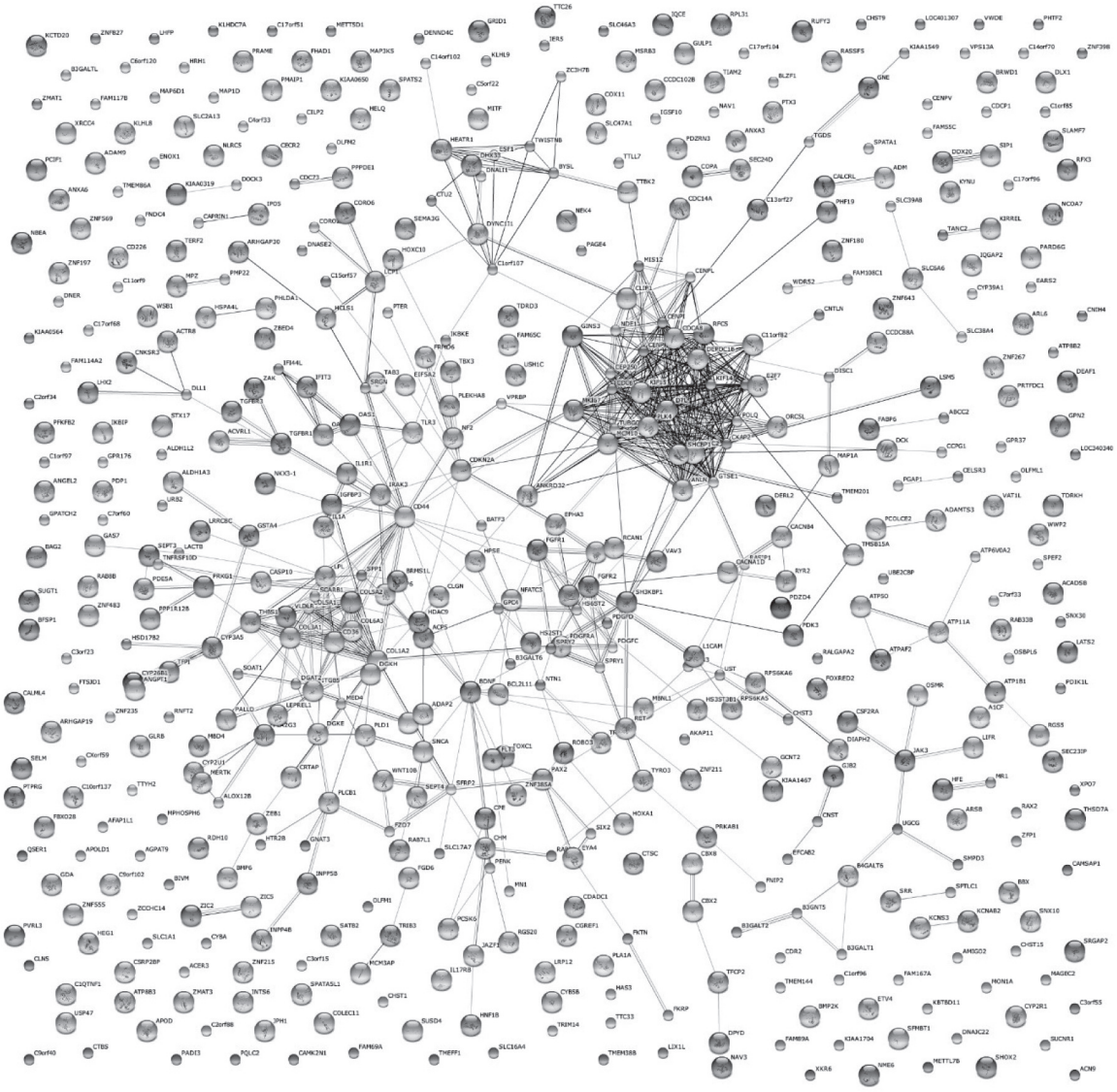


Fig.4. The PPI-Network evidence view of DEGs of Sample Group 2

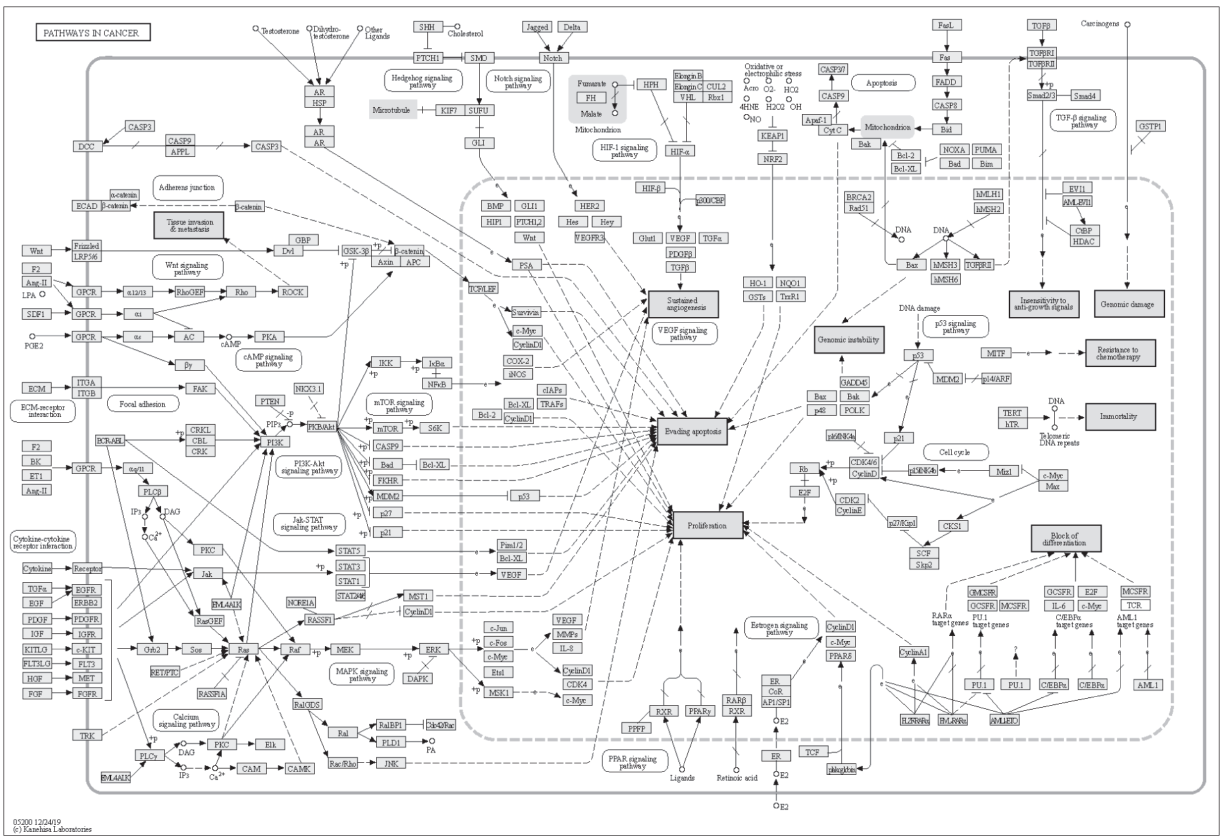


Fig.4. KEGG pathway of 27 DEGs which are significant in Cancer, identified from sample group 1.

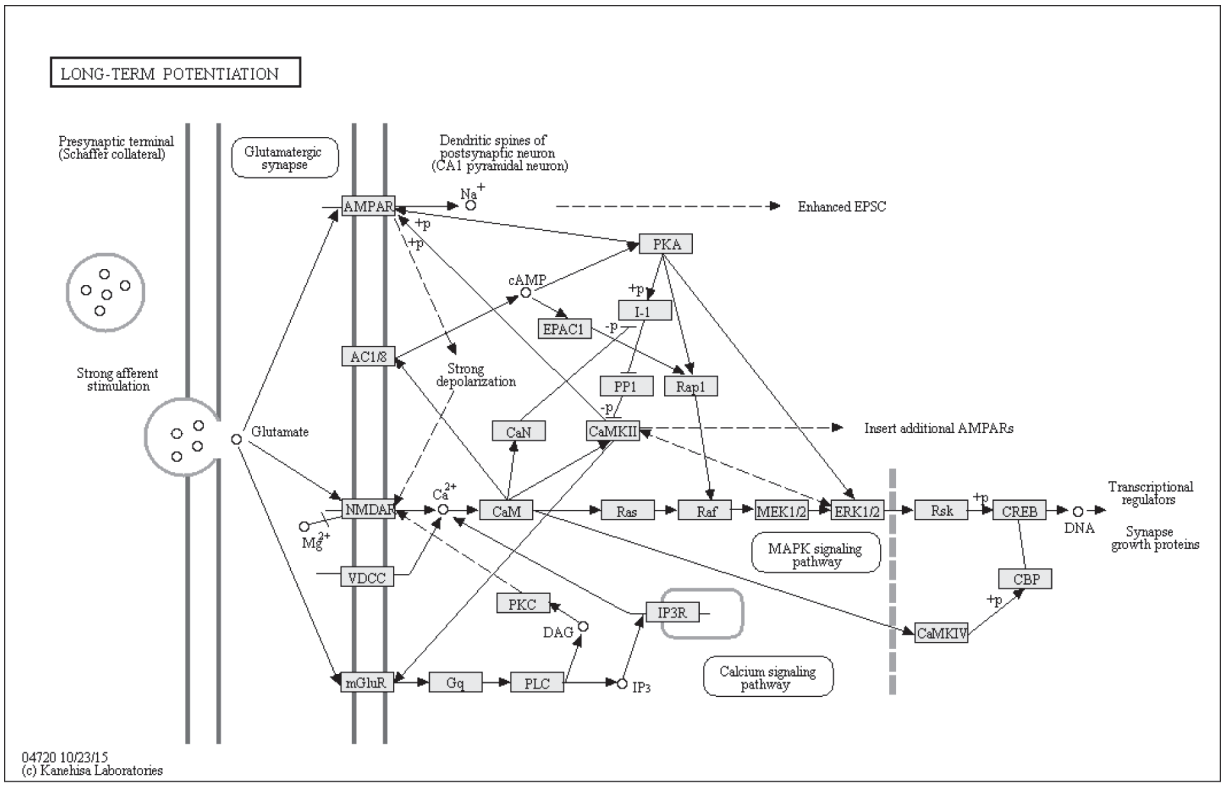


Fig. 6. KEGG pathway of 11 DEGs which are significant in long term potentiation, identified from sample group 2.