Bicluster Method to Predict Gene Patterns to Classify Differential Gene Expressions in Non-Small Cell Lung Cancer

Sumalatha Mani, Latha Parthiban

***Abstract*: In recent years, there are numerous efforts to overcome the constraints of data mining approaches to classify "BIG DATA". There are several types of data which has identical and similar expressions but there are dependent classification algorithms to predict these classes of expressions. Totally Different algorithms have been developed and enforced to research and differentiate the categories of data groups based on their functions. Zero-suppressed Binary Decision Diagram (ZBDD) algorithms help to classify the data with several categories. In the present study, lung cancer gene expression datasets 25 samples contain 10 mouth buccal cavity epithelial tissue samples and 15 nasal epithelial tissue samples from never smokers and current smokers were used to classify the genes and their expressions with various conditions. Using R and BioConductor software to normalize and predict differential expressed genes by Affy, Affycore tools and Limma packages to predict the gene expression with various functional properties. ZBDD algorithm and parallel coordination helps to predict the functional genes and the results shows 345 nasal epithelial genes were predict of which 54 genes were present in bicluster and 35 genes from mouth epithelial tissues show 14 were present in ZBDD bicluster. The results conclude that ZBDD algorithm has great advantage to classify big data and this algorithm can introduce in any large datasets for the accurate predict of large datasets*.**

***Keywords* : Clustering; bi-clustering; parallel co-ordination plots; R software implementation .**

I. INTRODUCTION

Data mining is the method of collection an outsized pool of knowledge and classifying it to grasp the functions [1]. Biological research includes sequences of genes and proteins, gene expressions, biological functions, pathways etc. There are several techniques that can generate large data of genes, proteins, and their expressions [2]. Different data mining methods are used to predict the gene expression data, and various algorithms are used to predict the genes and their functions [3-5]. Microarray data of both normal and disease conditions have several conditions of data that use cluster ways and microarray knowledge to classify the genes. Based on the literature and different statistical approaches, there are several clustering algorithms that are used to built gene expression data and classification, however little attention has paid to uncertainty within the results obtained[6-9].

In clustering, the patterns of expression of various genes across time, treatments, tissues, and intensity of color are classified into distinct clusters (perhaps organized hierarchically and k-means), in which genes in the same cluster are assumed to be probably functionally related or to be influenced by a standard upstream factor. Such a cluster structure is usually used to aid the elucidation of regulatory networks [10-14].

Agglomerative hierarchical clustering (HC) algorithm is one among the foremost frequently used techniques for the clustering of gene expression and therefore the classification of gene profiles [15-17]. However, HC ways have faith within the setting of some score threshold to distinguish members of a selected cluster from non-members, creating the determination of the number of clusters arbitrary and subjective [18]. The algorithm gives no guide to choose the "precise" range of clusters or the dimension at which to prune the tree. It is usually hard to grasp that distance metric to select, especially for structured data, such as gene expression profiles. Also, these methodologies do not provide a measure of ambiguity concerning the clustering, making it hard to calculate the prognosticative quality of clustering and to create comparisons between clustering on the basis of different models supported by totally different model assumptions (e.g., number of clusters, shape of clusters, etc.). We tried to handle these issues in a classical statistical framework and have targeted on the employment of bootstrap and permutation procedures to calculate local p-values for the significance of cluster development [19-21].

II. MATERIALS AND METHODS

The gene expression based on the intensity values of each gene can be predicted by using microarray data. Microarray data selected from the GEO database (GSE8987) is used in the current research. The dataset containing epithelial cells of respiratory tract is morphologically changed during smoking and passive smoking and undergoes genetic alterations to cause lung cancer [25-26]. The dataset contains 25 samples out of which 10 samples are mouth buccal cavity epithelium tissues mRNA samples and 15 nasal epithelial samples from never smokers and current smokers. The annotation platform of these two samples, namely, GPL96 (hgu133a) of mouth buccal cavity mRNA samples containing 22283 probes and GPL571 (hgu133a2) of nasal cavity mRNA samples containing 22277 probes, is used to predict differential gene expression analysis.

Revised Manuscript Received on November 06, 2019.

Sumalatha Mani, Research Scholar, Periyar University Salem, Tamilnadu, India. C.Kandaswami Naidu College for Women, Cuddalore, Tamilnadu, India-607001. Email:latha7suma@gmail.com

Latha Parthiban, Assistant Professor, Pondicherry University Community College, Kalapet, Pondicherry-605014
India.Email:lathap.ucc@pondiuni.edu.in.

The overall experiment is carried by five different steps: 1) Preprocessing and normalization, 2) Cluster analysis, 3) Differential gene expression analysis, 4) Functional analysis, and 5) Novel gene identification, and is potentially used for drug targets.

Preprocessing and Normalization

The raw data of selected gene expression dataset is carried by the Affymetrix method and the data analysis is carried by the R and BioConductor packages. There are several gene expression analysis packages that contain biological algorithms and can help to predict differential gene expressions. While comparing, the two totally different sample sets are hybridized to the identical array at systematic changes of intensity effects. The intensity values of background correct, log2 transformation and quality normalization are corrected by using RMA algorithm and summarization of corrected data with standard methods is implemented by RMA and MAS 5 method. The GCRMA method is used to filter the log2 intensity values on the origin of the intensity of mean, median, and standard deviation calculations. The results of normalized values are predicted using box plot, histogram, MA, PCA (principal component analysis) and quality plots. In order to research Gene Chip data with multiple arrays, the data preprocessing at the probe level is a critical step. The global background correction by signal and noise (background) convolution model is the one within which PM intensity distribution is modeled by an exponentially distributed signal element S with parameter λ, and a normally distributed background element B with mean μ and standard deviation σ.

PM=S+B

$$S \sim \exp(\lambda)$$

$$B \sim N(\mu, \sigma)$$

$$E(S \mid PM) = PM - \mu - \lambda\sigma^2 + \sigma \frac{\phi((PM - \mu - \lambda\sigma^2)/\sigma) - \phi((\mu + \lambda\sigma^2)/\sigma)}{\Phi((PM - \mu - \lambda\sigma^2)/\sigma) - \Phi((\mu + \lambda\sigma^2)/\sigma) - 1}$$

E (S|PM) represents background corrected value of each PM. φ and Φ is the normal density and cumulative density, respectively. Positive signal elements are estimated after the adjustment of background elements. This background correction is implemented in the robust multi-array average (RMA).

Clustering Analysis

Hierarchical clustering

The hierarchical clustering of these data can be calculated by using three techniques like Node Score, Level score and Tree score. The Node score is for calculating the node that specifies a cluster and therefore the enrichment p-values can be calculated to assign the given node with one amongst the categories within the data. The significant p-value of observing K instances assigned by the algorithm to a given category in a set of n instances is given by

$$P = \sum_{x=k}^{n} \binom{K}{x}\binom{N-K}{n-x} \Big/ \binom{N}{n}$$

where K is the total number of instances assigned to the class (the category) and N is the number of instances in the dataset. The p-values for all nodes and all categories also be viewed as dependent set estimations. Within the Level score, a level l of the tree contains all nodes that are separated by one edge from the root; every level specifies a partition of the data into clusters. Selecting for every node, the category for which it turned out to be has a significant node score, (J=tp/ (tp+fn+fp),

where tp is the number of true positive cases, fn is the number of false negative cases, and fp is the number of false positive cases. If the node in question is judged to be non-significant by the enrichment criterion, then its J-score is set to null. The level score is defined because the average of all J-scores at the given level.

The Tree score method is to define the weighted best J-Score

$$J^* = \frac{1}{N}\sum_{i}^{c} n_i J_i^*$$

where J*i is the best J-Score for class i in the tree, ni is the number of instances in categories i, c is the range of categories, and N is the number of instances within the dataset.

The k-means clustering can be used for calculating data to find the means of noise data

$$J(K, m) = \sum_{k=1}^{K}\sum_{i=1}^{N} (u_{ki})^m d^2(x_i, c_k)$$

K and N are the number of clusters and genes in the data sets, m is a parameter which relates to 'fuzziness' of resulting clusters, uki is the degree of membership of gene xi in cluster k, d2(xi; ck) is the distance from gene xi to centroid ck.

Bi-Clustering

The ZBDD algorithm is used to identify the bi-clustering of binary data using 0s and 1s as columns and rows[27] Zero-suppressed BDDs (ZBDDs) are a variant of ROBDDs and represent a set of combinations. A combination of n elements is an n-bit vector (x1; x2; . . .; xn)Є Bn where B = {0,1}. The i-th bit reports whether the i-th element is contained in the combination. Thus, a group of combinations can be described by a Boolean function f : Bn→ B. A combination given by the input vector (x1; x2; . . . ; xn) is contained in the set, if and only if f(x1; x2; . . . ; xn) = 1.

$$ARV(I, J) = \frac{1}{|I||J|}\sum_{i \in I}\sum_{j \in J}(a_{ij} - a_{i\bullet})^2$$

Parallel Co-ordination

The parallel coordinate (PC) plot is used to visualize the calculated information from R Software by using a way to visualize the high dimensional data. All axes are organized in parallel to every alternative on a 1-D plane. The additive-related bicluster shows a number of lines with the same slope across the conditions. Thus, columns {C2-C1, C3-C1} with rows R1, R3, R5, R9 and R11 can be visualized by this type of arrangement in PC plots.

Analysis of gene expression

The significant quality filtered probe sets of all datasets are used to predict differential expression analysis by LIMMA package [28]. Limma is specifically designed to predict linear models of datasets assigned with differential expressions.

Conditions of samples are aligned to extract the datasets based on sample sets and differential expressions are calculated by empirical Bayes shrinkage of the standard errors towards a common value by

calculating the sensible t-statistics, moderated F-statics, and log-odds. The creation of topTable of n number of differentially expressed probes for any contrasts is imposed by fold change cutoff and one can see the number of gens returned by using ifc modifier for topTable. Then, the list of genes with

adj.P.Val ≤ 0.05 and fold change ≥ 2 is created for the first contrasts that can make a heat map of the expressions. In order to annotate the probe sets into gene symbols, there is a need to load associated database package and annotate package that can extract the probe sets' IDs from the topTable results and match the symbols.

Functional Annotation and Enrichment analysis

The differentially expressed genes are annotated with reference genome database to predict the gene names on the basis of Affymetrix probe ids. The DAVID functional annotation database is used to predict the gene names and functions based on clusters and to convert the gene names based on the annotation platform [29-30]. The gene ontology database is also employed to detect the functions of each gene and their associated networks that can be predicted in large scale genomics and where FDR = 0.05 is used as a cut-off criterion. The functional enrichment analysis is predicted on the origin of GO functional terms and is compared with process, function, and component of different co-expressions that can be predicted by using Gorilla functional enrichment analysis tool [31] ToppFun, and FunRich Tools [32].

Protein-protein interaction network analysis

String database is used to predict the functionally enriched genes and their associated protein functions are predicted online. The information of human PPI network relationship genes and associated proteins related to the disease is collected to construct the PPI network and to classify the most novel genes that are significantly associated to carcinoma. The differential expression of genes was mapped to the String information and also the known and predicted associations were then scored and integrated. The combinations of genes with the threshold were scored >0.4 of the median confidence to visualize the maximum score and the results were observed in Cytoscape software.

III. RESULT AND DISCUSSION

There are several algorithms that will help to identify gene expressions and also to construct cluster and bi-clusters. The current research focuses on the raw fact of mouth buccal cavity and nasal cavity and epithelial tissue samples were used for preprocessing and normalization by using Affy package. The signal extraction methods, such as MAS5, RMA, and GCRMA methods, were used for the preprocessing, normalization, and quality analysis of different intensity values screened by log2 intensity calculation. The differential expression of mouth epithelial dataset has 22283 probe sets that are filtered by using normalization and show that 13411 probes are accepted with logP intensity ranges. The redundant log2 intensity values that are finally filtered show that 8804 genes were predicted in differential expressions (Table: 1a, 1b).

The information of human PPI network relationship genes and associated proteins associated with the malady is collected to construct the PPI network and to classify the foremost novel genes that are significantly related to carcinoma. The differential expression of genes was mapped to the String information and also the famous and foreseen associations were then scored and integrated. The nasal datasets had the lowest number variables (1108) that were screened based on the log2 transformation, filtered by 11107 genes, and used as differential expressions. We have designed the matrix of every individual samples by contrast matrix to design the coefficients (Never smoker-Current smoker). We selected the sample sets of condition column that represent four coefficients only and the coordinates of each top Table were calculated by empirical Bayes method (Table: 2a, 2b). Here, 82 up-regulated and 3747 down-regulated genes were predicted in mouth epithelial tissues.

Table: Ia Contrast matrix and coefficient of mouth buccal cavity epithelial samples to predict differential expressions

Coefficients (Coef)	Fold Change (IFC)	No of genes	Up	Down
Coef=1 (Mouth1-Mouth2) Never smoker- Current smoker	1	853	1	852
Coef=2 (Mouth1 – Mouth3) Never smoker – Current smoker	1	1265	26	1239
Coef= 3(Mouth1 – Mouth4) Never Smoker – Never Smoker	1	1597	65	1532
Coef=4 (Mouth2 – Mouth3) Current Smoker – Current Smoker	1	6	0	6
Coef=5 (Mouth2 – Mouth4) Current Smoker – Never Smoker	1	76	14	62
Coef=6 (Mouth3 – Mouth4) Current Smoker – Never Smoker	1	72	16	56

Coefficients (Coef) represents the selection of matrix based on type of sample comparison, Fold Change (IFC) the fold change represents the probability of samples selection based on differential expression, Further, we have predicted the functional enriched genes of both tissues, which show that only 35 transcriptional genes are expressed in transcriptional regulation out of which only 14 genes are most commonly involved in different signaling pathways in mouth epithelial tissues. The nasal epithelial tissues have 1243 up-regulated and 2979 down-regulated genes out of which 345 genes are involved in different functional mechanisms, biological processes, and cellular components. Out of these, only 249 genes are involved in transcriptional regulation in different signaling pathways, which shows that only 69 gene expressions are involved in lung cancer associated signaling pathways.

Based on the results of various differential gene identification methods, we have used the ZBDD bi-clustering algorithm to compare the differentially expressed genes and observe the gene expression on the origin of the sample's rigidity (Fig: 1a, 1b).

The down-regulated genes are chosen as 0th level and the up-regulated genes expression are chosen as 1th level (Figure: 2a, 2b). The expression of these genes show (a) The response time spent by each method to find all the embedded biclusters from the synthetic data expressions of various sizes (b) The quantity of biclusters found by each method within the same time spent as per our method.The parallel co-ordination plot is used to visualize the different clustering results (Figure: 3a-e). Further, a study conducted to predict the potential drug targets of these two epithelial tissues shows that only 14 genes in mouth epithelial tissues and 54 genes in nasal epithelial tissues are used as potential drug targets.

Table- Ib. Contrast matrix and coefficient of Nasal epithelial samples to predict differential expressions

Coefficients (Coef)	Fold Change (IFC)	No of genes	Up	Down
Coef=1 (Nose1-Nose2) Never smoker – Current Smoker	1	1	0	1
Coef=2 (Nose1 – Nose3) Never Smoker – Never Smoker	1	2255	414	949
Coef= 3(Nose1 – Nose4) Never Smoker – Current smoker	1	1608	193	162
Coef=4 (Nose2 – Nose3) Current Smoker – Never Smoker	1	2783	398	997
Coef=5 (Nose2 – Nose4) Current Smoker – Current Smoker	1	2259	238	205
Coef=6 (Nose3 – Nose4) Never Smoker – Current Smoker	1	665	0	665

Coefficients (Coef) represents the selection of matrix based on type of sample comparison, Fold Change (IFC) the fold change represents the probability of samples selection based on differential expression

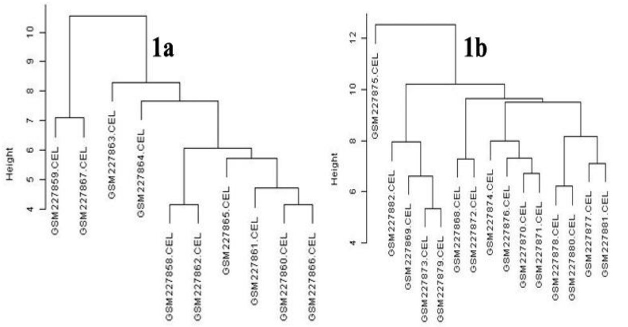

Fig. 1a & b. Cluster analysis of microarray predicted using hierarchical clustering methods of both smokers' and non-smokers' buccal and nasal cavities

A regulatory study shows that there are 10 genes that are most commonly present in lung adenocarcinoma, namely, CDKN1B, MTUS1, MID1, EMP1, VDR, IGF2BP3, NEDD4L, HPGD, TMPRSS2, and ZFAND5, which are highly expressed in mouth epithelial current smoker and never smoker datasets.

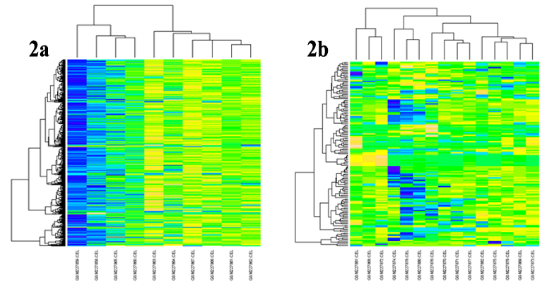

Fig. 2a. Hierarchical clustering of novel buccal cavity of lung cancer predicted with smoker and non-smoker samples (Fig. 2b.) Hierarchical clustering of nasal cavity of lung cancer predicted with smoker and non-smoker samples

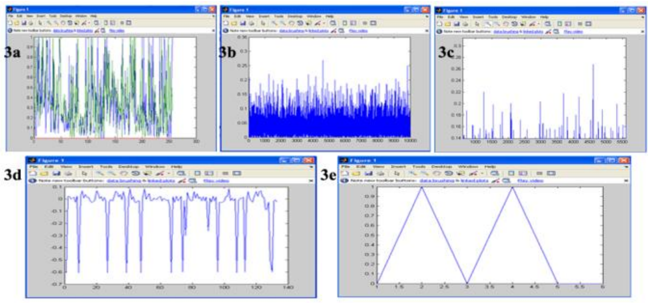

Fig. 3a showing the ZBDD algorithm used to construct the clustering of data. (Fig. 3b.) ZBDD algorithm to separate the clusters supported the p-values of differentially expressed data. (Fig. 3c) ZBDD algorithm to predict the clusters based on noise removal. (Fig. 3d) ZBDD algorithm used to construct up-regulated genes supported the bi-clustering technique. (Fig. 3e) ZBDD algorithm used to construct down-regulated genes based on the bi-clustering method of Parallel co-ordination method to construct bi-clusters

Further, we have predicted that the pathways show that there are 14 genes involved in different signaling pathways, such as S1P1 pathway, GMCSF-mediated signaling events, IGF1 pathway, EGFR-dependent Endothelia signaling events, Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met), CDC42 signaling events, Plasma membrane estrogen receptor signaling, Thrombin/protease-activated receptor (PAR) pathway, Syndecan-1-mediated signaling events, and Glypican 1 network. VDR, ASAP2, FOXO4, KAT2B, CTNNA1, TMPRSS2, CDKN1B, MDF1C, NEDD4L, ACTR2, MYO6, ZFAND5, ABI1, and EZR, and are the most common genes found in non-small-cell adenocarcinoma in both smoker and never smoker samples.

The overall results of functional enrichment show that there are 3241 genes that are differentially expressed on the origin of molecular and biological functions. Out of these genes, only 345 genes are the top regulated genes in nasal epithelial tissues of both never smoker and current smoker tissues. Further, the prediction of the transcriptional regulatory genes among the 345 genes showed that only 249 genes were involved in transcriptional factors that can regulate the expression of different signaling pathways, such as mTOR signaling pathway, TRAIL signaling pathway, Arf6 signaling events, Arf6 downstream pathway, PAR1-mediated thrombin signaling events, IFN-gamma pathway,PDGFR-beta signaling pathway, GMCSF-mediated signaling events, Internalization of ErbB1, and Nectin adhesion pathway. 69 genes were differentially expressed and were further used to predict drug sensitiveness to non-small cell lung cancer. 54 genes were used as impending drug targets for nasal epithelial tissues in lung cancer.

Table: IIa. Functional enrichment analysis of topTable unregulated genes on never smoker – current smoker samples comparison of Mouth buccal cavity epithelial tissues.

Term	%	P-Value	Genes	Fold Enrichment	FDR
negative regulation of cell proliferation	14.285	0.00126	VDR, CDKN1B, KAT2B, NUPR1, SKAP2, FOXO4	7.026315	1.7962
cellular response to insulin stimulus	7.1428	0.01053	KAT2B, FOXO4, LPIN1	18.65073	14.100
steroid metabolic process	9.5238	0.01084	OSBPL2, CYP1B1, CYP1A1, FDFT1	8.371287	14.476
toxin metabolic process	4.7619	0.01140	CYP1B1, CYP1A1	169.1	15.176
response to insulin stimulus	7.1428	0.02188	KAT2B, FOXO4, LPIN1	12.682	27.201
response to organic substance	14.285	0.02276	CYP1B1, KAT2B, CYP1A1, DUOX2, FOXO4, LPIN1	3.51803	28.138
cell cycle arrest	7.1428	0.02312	CDKN1B, KAT2B, FOXO4	12.3131	28.515
hormone metabolic process	7.1428	0.02439	CYP1B1, CYP1A1, DUOX2	11.9646	29.837
regulation of cell proliferation	14.285	0.03175	VDR, CDKN1B, KAT2B, NUPR1, SKAP2, FOXO4	3.22299	37.059
cellular response to hormone stimulus	7.1428	0.03704	KAT2B, FOXO4, LPIN1	9.535714	41.818
peptidyl-lysine modification	4.7619	0.03827	KAT2B, DOHH	49.73529	42.872
isoprenoid biosynthetic process	4.7619	0.04487	CYP1A1, FDFT1	42.275	48.250
regulation of hormone levels	7.1428	0.04660	CYP1B1, CYP1A1, DUOX2	8.399006	49.574
response to peptide hormone stimulus	7.1428	0.04827	KAT2B, FOXO4, LPIN1	8.23538	50.829
hydrogen peroxide metabolic process	4.7619	0.05578	CYP1A1, DUOX2	33.82	56.115
oxidation reduction	11.904	0.05650	CYP1B1, CYP1A1, DOHH, DUOX2, FDFT1	3.307902	56.589
regulation of microtubule polymerization or depolymerization	4.7619	0.06872	CDKN1B, MID1	27.27419	63.994
regulation of cell size	7.1428	0.08046	CDKN1B, NUPR1, CDKN2AIP	6.156553	69.984
regulation of microtubule cytoskeleton organization	4.7619	0.09199	CDKN1B, MID1	20.13095	74.955
isoprenoid metabolic process	4.7619	0.09616	CYP1A1, FDFT1	19.21590	76.555
heme binding	7.1428	0.03359	CYP1B1, CYP1A1, DUOX2	10.05914	33.315
iron ion binding	9.5238	0.03642	CYP1B1, CYP1A1, DOHH, DUOX2	5.269074	35.592

tetrapyrrole binding	7.1428	0.03775	CYP1B1, CYP1A1, DUOX2	9.435319	36.643
aromatase activity	4.7619	0.05806	CYP1B1, CYP1A1	32.4575	50.801
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	4.7619	0.06928	CYP1B1, CYP1A1	27.04791	57.315
cell adhesion molecule binding	4.7619	0.06928	EZR, CTNNA1	27.0479	57.315
electron carrier activity	7.1428	0.09725	CYP1B1, CYP1A1, DUOX2	5.50749	70.273
electron carrier activity	7.1428	0.09725	CYP1B1, CYP1A1, DUOX2	5.50749	70.273
oxygen binding	4.7619	0.09783	CYP1B1, CYP1A1	18.8706	70.502

Table: II b. Functional enrichment analysis of topTable upregulated genes on never smoker – never smoker samples comparison of Nasal cavity epithelial tissues

Term	%	PValue	Genes	Fold Enrichment	FDR
protein catabolic process	8.4812	1.28E-06	RAD23B, SPG7, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ARIH1, CUL5, TPP1, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, ADAM9, TBL1XR1, UBE2J1, ERLIN2, PCNP, SOCS5, AFG3L2, HLTF, CLPX, UBE2N, PJA2, UBE2E3, HSP90B1, UBR5, FBXL5, ACE2, RNF138, USP46, CAND1, SIAH2, CUL4B, RNF111, UBE2E1	2.2481	0.0022
macromolecule catabolic process	9.8619	1.37E-06	RAD23B, SPG7, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ISG20, ZFP36L1, ARIH1, CUL5, TPP1, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, ADAM9, AGA, ABCE1, TBL1XR1, GUSB, UBE2J1, GTF2H3, ERLIN2, PCNP, SOCS5, AFG3L2, HLTF, CLPX, UBE2N, PJA2, DNASE2, UBE2E3, HSP90B1, UBR5, FBXL5, ACE2, RNF138, USP46, CAND1, SIAH2, CUL4B, UBE2E1, RNF111	2.0818	0.0023
protein transport	9.6646	1.59E-06	SEC24B, CHMP5, AP1AR, CCDC91, SNX4, PEX3, PDIA4, NXT2, CHMP2B, CEP57, ZFYVE16, AAGAB, NECAP1, TLK1, VPS16, SNX24, RAB6A, RANBP2, TNPO2, SAR1B, RAB27B, SEC24D, SEC61A1, RAB2A, SCAMP1, SEC23A, MCM3AP, STX4, ATG9A, LIN7C, NUP85, STXBP3, CLPX, EPS15, COG4, HSP90B1, ERBB2IP, IPO7, RAB22A, IPO5, YWHAQ, SDCBP, JAK2, SRP72, GGA1, SNX13, KPNA1, SERP1, F2R	2.0911	0.0027

establishment of protein localization	9.6646	2.06E-06	SEC24B, CHMP5, AP1AR, CCDC91, SNX4, PEX3, PDIA4, NXT2, CHMP2B, CEP57, ZFYVE16, AAGAB, NECAP1, TLK1, VPS16, SNX24, RAB6A, RANBP2, TNPO2, SAR1B, RAB27B, SEC24D, SEC61A1, RAB2A, SCAMP1, SEC23A, MCM3AP, STX4, ATG9A, LIN7C, NUP85, STXBP3, CLPX, EPS15, COG4, HSP90B1, ERBB2IP, IPO7, RAB22A, IPO5, YWHAQ, SDCBP, JAK2, SRP72, GGA1, SNX13, KPNA1, SERP1, F2R	2.0720	0.0035
protein localization	10.059	1.79E-05	SEC24B, CHMP5, AP1AR, PEX3, PDIA4, ZFYVE16, AAGAB, NECAP1, TLK1, RANBP2, RAB6A, VPS16, SAR1B, RAB27B, SEC24D, SEC23A, SCAMP1, MCM3AP, STX4, ATG9A, G3BP2, STXBP3, NUP85, CLPX, IPO7, IPO5, SDCBP, SRP72, SNX13, KPNA1, SERP1, SNX4, RDX, CCDC91, CHMP2B, NXT2, CEP57, SNX24, TNPO2, SEC61A1, RAB2A, LIN7C, EPS15, COG4, HSP90B1, ERBB2IP, RAB22A, YWHAQ, JAK2, GGA1, F2R	1.8803	0.0310
proteolysis involved in cellular protein catabolic process	7.6923	1.83E-05	RAD23B, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ARIH1, CUL5, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, ADAM9, TBL1XR1, UBE2J1, ERLIN2, PCNP, SOCS5, HLTF, CLPX, UBE2N, PJA2, UBE2E3, HSP90B1, UBR5, FBXL5, RNF138, USP46, CAND1, SIAH2, CUL4B, RNF111, UBE2E1	2.1137	0.0318
cellular protein catabolic process	7.692	2.05E-05	RAD23B, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ARIH1, CUL5, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, ADAM9, TBL1XR1, UBE2J1, ERLIN2, PCNP, SOCS5, HLTF, CLPX, UBE2N, PJA2, UBE2E3, HSP90B1, UBR5, FBXL5, RNF138, USP46, CAND1, SIAH2, CUL4B, RNF111, UBE2E1	2.1032	0.0356
cellular macromolecule catabolic process	8.6785	2.56E-05	RAD23B, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ISG20, ZFP36L1, ARIH1, CUL5, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, ADAM9, ABCE1, TBL1XR1, UBE2J1, GTF2H3, ERLIN2, PCNP, SOCS5, HLTF, CLPX, UBE2N, PJA2, DNASE2, UBE2E3, HSP90B1, UBR5, FBXL5, RNF138, USP46, CAND1, SIAH2, CUL4B, RNF111, UBE2E1	1.9735	0.0443

modification-dependent macromolecule catabolic process	7.2978	3.76E-05	RAD23B, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ARIH1, CUL5, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, TBL1XR1, UBE2J1, ERLIN2, PCNP, SOCS5, HLTF, UBE2N, PJA2, UBE2E3, HSP90B1, UBR5, FBXL5, RNF138, USP46, CAND1, SIAH2, CUL4B, RNF111, UBE2E1	2.0961	0.0652
modification-dependent protein catabolic process	7.2978	3.76E-05	RAD23B, UBE3A, UBE2G1, UBA5, EDEM3, FEM1B, CD2AP, ARIH1, CUL5, PSMB3, WWP1, FBXO28, PSMD2, RNF11, ZMPSTE24, FBXW12, RANBP2, FBXO3, TBL1XR1, UBE2J1, ERLIN2, PCNP, SOCS5, HLTF, UBE2N, PJA2, UBE2E3, HSP90B1, UBR5, FBXL5, RNF138, USP46, CAND1, SIAH2, CUL4B, RNF111, UBE2E1	2.0961	0.0652
regulation of cellular protein metabolic process	6.1143	1.47E-04	GCLC, IL6ST, FEM1B, TIMP1, ZFP36L1, PRKAR2B, APP, PRKAR2A, MDFIC, PSMB3, RB1CC1, ITGAV, PSMD2, PUM2, INSR, ADAM9, EIF2B5, IBTK, PAIP1, SMAD4, NDFIP1, UBE2N, MAP4K5, EIF4E, EP300, HIPK3, PRKAR1A, JAK2, EIF2AK3, BMPR1A, UBE2E1	2.1267	0.2543

The overall analysis of mouth epithelial tissue and nasal epithelial tissue respiratory tract non-small cell lung cancer is predicted to identify the differentially expressed genes on the origin of molecular mechanism, biological process, and cellular component. Here, we predicted the 82 up-regulated and 3747 down-regulated genes in mouth epithelial tissues. The identification of different gene expression levels was observed by using clustering techniques followed by bi-clustering methods. A set of observations were allocated into subsets (called clusters) in order to that observations within the same cluster were similar in some sense.

Clustering is a method of unsupervised learning and is a general method for statistical data analysis. The outcomes of the clustering algorithmic method show the up-regulated and the down-regulated genes to be highly overlapped, whereas the bi-clustering ZBDD algorithmic method shows a clear interpretation of neural network clusters. Further, we have predicted that the functional enriched genes of both tissues show that only 35 transcriptional genes are expressed in transcriptional regulation, out of which only 14 genes are most commonly involved in different signaling pathways in mouth epithelial tissues. The nasal epithelial tissues have 1243 up-regulated and 2979 down-regulated genes out of which 345 gene data are implicated in different functional mechanisms, biological processes, and cellular components. Out of these, only 249 gene data are implicated in transcriptional regulation in different signaling pathways, which shows that only 69 gene data are implicated in carcinoma associated signaling pathways.

IV. CONCLUSION

In this study conducted to predict the potential drug targets of these two epithelial tissues shows that only 14 genes in mouth epithelial tissues and 54 genes in nasal epithelial tissues are used as potential drug targets. This study demonstrates that computational empirical Bayes method can be used to discover both tumor and control tissue expressions in lung cancer patients. This research helps clinicians and a molecular diagnostic professional to understand the genes involved in both never smoker and current smoker in NSCLC and is more invasive to diagnostic procedures for the patients. We expect that our results can facilitate researchers to grasp the tissue types and predict the genes and ensure that this knowledge be used for public. The models will be further improved as more knowledge is vailable and future discoveries could be made for different types of cancers.

Conflicts of interest

All authors declare that there are no conflicts of interest.

REFERENCES

1. Zhang Y., Guo S. L., Han L. N., Li T, L.: Application and exploration of big data mining in clinical medicine. Chin. Med. J. 129 731-738 (2016)
2. Anand Mariadoss A.V., Krishnan Dhanabalan A., Munusamy H., Gunasekaran K., David, E.: In silico studies towards enhancing the anticancer activity of phytochemical phloretin against cancer drug targets. Curr. Drug Targets 13 174-188 (2018).
3. Jin D., Lee, H.: FGMD: A novel approach for functional gene module detection in cancer. PloS One 12 e0188900. (2017)

4. Chiu C. C., Chan S Y., Wang C. C., Wu W,S: Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Systems Biology 6 S12 (2013).
5. Dhawan A., et al.: Mathematical modelling of phenotypic plasticity and conversion to a stem-cell state under hypoxia. Sci. Rep. 6 18074 (2016) https://doi.org/10.1038/srep/18074
6. Choi Y. H., J et al.: Mathematical modelling long-term effects of replacing Prevnar7 with Prevnar13 on invasive pneumococcal diseases in England and Wales. PloS one 7 e 39927 (2012).
7. Ferguson A. C., Pearce S., Band L. R., Yang C., Ferjentsikova I., King J., Yuan Z., Zhang D., Wilson ZA. Biphasic regulation of the transcription factor ABORTED MICROSPORES (AMS) is essential for tapetum and pollen development in Arabidopsis. New Phytologist 213(2) 778-90 (2017)
8. Place A. E., Huh S, J., Polyak, K.: The microenvironment in breast cancer progression: biology and implications for treatment. Breast Cancer Research 6 227 .(2011)
9. Chaffer C. L., et al.: Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. Proceedings of the National Academy x of Sciences 108 7950 (2011)
10. Kern M., Lex A., Gehlenborg N., Johnson C, R.: Interactive visual exploration and refinement of cluster assignments. BMC Bioinformatics 18 406 (2017).
11. Hristoskova A., Boeva V., Tsiporkova, E.: A formal concept analysis approach to consensus clustering ofmulti-experiment expression data. BMC Bioinformatics 15 151 (2014).
12. Zhao W., Zou W., Chen J, J.: Topic modeling for cluster analysis of large biological and medical datasets. BMC Bioinformatics 15 S11 (2014).
13. Lerato L., Niesler, T.: Clustering acoustic segments using multi-stage agglomerative hierarchical clustering. PloS One 10 e0141756 (2015)
14. Freyhult E., et al.: Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. BMC Bioinformatics 11 503 (2010).
15. Gaynor S., Bair, E.: Identification of relevant subtypes via preweighted sparse clustering. Comput. Stat. Data An. 116 139-154 (2017).
16. Zhou S., Xu Z., Liu, F.: Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. IEEE T. Neur. Net. Lear. 28 3007-17 (2017).
17. Kruse C., Eiken P., Vestergaard, P.: Clinical fracture risk evaluated by hierarchical agglomerative clustering. Osteoporosis Int. 28 819-32 (2017).
18. Ghosh S., Townsend J, P.: H-CLAP: Hierarchical clustering within a linear array with an application in genetics. Stat. Appl. Genet. Mol. 14 125-41 (2015) https://doi.org/10.1515/sagmb-2013-0076
19. Scrucca L., Raftery A, E.: Improved initialization of model-based clustering using Gaussian hierarchical partitions. Adv Data Anal Classi. 9 447-60 (2015)
20. McLachlan G.J., Bean R.W., Ng S, K.: Clustering. In: Keith J. (eds) Bioinformatics. Methods Mol. Biol. 1526 (2017)
21. Demir A., Cetingul H, E.: Sequential hierarchical agglomerative clustering of white matter fiber pathways. IEEE Trans. Biomed. Eng. 62 1478-89 (2015)
22. Banfield J. D., Raftery A, E.: Model-based Gaussian and non-Gaussian clustering. Biometrics. 49 803-821 (1993)
23. Dharan S., Nair A, S.: Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. BMC Bioinformatics 10(1) S27 (2009).
24. Madeira S. C., Oliveira A, L.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biol. Bioinform. 1 24-45 (2004).
25. Sridhar S., et al.: Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. BMC Genomics 9 259 (2008).
26. Wistuba II., et al.: Molecular damage in the bronchial epithelium of current and former smokers. J. Natl. Cancer Inst. 89 1366-73 (1997)
27. Yoon S., Nardini C., Benini L., De Michel,i G.: Discovering coherent biclusters from gene expression data using zero-suppressed binary decision diagrams. IEEE/ACM Trans. Comput. Biol. Bioinform. 2 339-54 (2005).
28. Diboun I., Wernisch L., Orengo C. A., Koltzenburg, M.: Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC Genomics 7 252 (2006)
29. Huang D.W., et al.: Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35 169-175 (2007)
30. Wen Z., et al.: Expression profiling and functional annotation of noncoding genes across 11 distinct organs in rat development. Sci. Rep. 6 38575 (2016) Kichaev G., Pasaniuc, B.: Leveraging functional-annotation data in trans-ethnic fine-mapping studies. Am. J. Hum. Genet. 97 260-271 (2015)
31. Erinjeri N. J.,et al.: Whole-exome sequencing identifies two discrete druggable signaling pathways in follicular thyroid cancer. J. Am. Coll. Surg. 226 950-959 (2018)
32. Vastrad C., Vastrad, B.: Bioinformatics analysis of gene expression profiles to diagnose crucial and novel genes in glioblastoma multiform. Pathology-Research and Practice 214 1395-461 (2018).

AUTHORS PROFILE

Mrs. Sumalatha Mani is a Phd Research Scholar in Periyar University, Salem India, cum Asst professor in the department of computer Science, C. Kandaswami Naidu College for Women, Cuddalore. As an budding researcher in computer science her search deals with the analysis of oxidative stress related gene expression using biclustering method

Dr.Lathaparthiban, is an Asst. Professor in the Pondicherry University: Pondicherry, Tamil Nadu, India, Her research activities mainly focus on Datamining , image processing, computer-aided diagnosis in biomedical science, in her research credited she published more than 30 journal in Peer reviewed journal