Turkish Journal of Computer and Mathematics Education Vol.12 No. 7 (2021), 2585- 2593

Research Article

A Model Based Approach On Gene Expression Profiling Of Colorectal Cancer And Normal Mucosa Using Logistic Regression, Artificial Neural Network And Structural Equation Modelling

A Saranya[1], S Venkatesan[1]
[1]Sri Ramachandra faculty of Engineering & Technology - Bioinformatics,
Sri Ramachandra Institute of Higher Education & Research [DU],Chennai.

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: Colorectal cancer is one of the leading biomedical issues of concern. Our understanding of the disease has improved using microarray expression data analysis over last few decades. Therefore, it is of interest to design conceptual statistical models to analyze volumes of data to glean useful information. We used data from the open source Gene Expression Omnibus database containing information on160 normal mucosa tissues and 203 CRCs. The model described in this report selected 44 candidate genes exceeding 4-fold threshold expression. The reliability was determined using Cronbach's alpha measurement for further statistical analysis. Structural Equation Modelling and binary logistic regression and neural networks were used to incur the strength of association of genes with the disease outcome.

Keywords: CRC-Colorectal cancer; Microarray data, Logistic Regression, Multilayer Perceptron Neural Network, Structure Equation Modelling, Confirmatory Factor Analysis.

Introduction

High throughput profiling of biological systems in a cost-efficient manner aids in the generation of arduous volume in "Big Data" [1]. The complete analysis of Big Data provides an insight to the biology of disease that expedite the progress toward precision medicine which is, tailoring prevention, diagnosis, and treatment based on the molecular characteristics of a patient's disease. The urge for Bioinformatics research in cancer arena is due to emerging new technologies that increase the cancer data in exponential rate.

Colorectal cancer is one of the third mostly diagnosed cancer and leading causes of the death. A country wide study from 2012-2014 from various Population Based Cancer Registries and Hospital Based Registries report on significant increase in the mortality and incidence rates of Colorectal cancer in males at Chennai, Bangalore and Delhi and Females at Barshi and Bhopal [2]. The American cancer society and Centre for Disease Control and Prevention (CDC) estimated the incidence rates of colon and rectal cancer in males (70,820 cases) and in females (63,670 cases) and mortality rate in males (26,020) and females (23,170) by 2016 [3]. Due to 13% of demographic changes and 19% of cancer risk, the total cancer burden is anticipated to increase by 30% in Chennai by 2012-2016. On predicted incidence rates in the Chennai registry and current rates in Dindigul district, large bowel cancer likely to be surpassing Cervical Cancer in ranking in Chennai 2016 [4]. Colorectal cancer usually begins from the growth of the polyps in the inner layer of large intestine. The succession rate of genetic mutations and polymorphisms would progress the dynamics of polyps into subsequent unbound growth of Cancer. Therefore, studying molecular correlation is highly necessary to decipher the clinically important and useful biomarkers that can correlate with disease prognosis and treatment of the disease [5]. This provides clues to oncologist to predict the confounding factor associable to the disease. It is quintessential to construct a biological model with the selected genes and apprehend the biological phenomena of the disease. Recently, many efforts are made to solicit the genes responsible for the colorectal cancer. But there are only few reports about the statistical analyses on microarray gene signatures of colorectal cancer. The current research is to construct a statistical model to predict the most influencing factors highly responsible for the colorectal cancer from the recent dataset available on the GEO database. Though the studies are available pertaining to the genes involved in the colon and rectal cancer, the molecular classification of the disease is poorly understood. Although Structural Equation Modelling is widely used to study the physical and behavioural sciences to model human interactions, the technique is not fully exploited on analysing the microarray gene expression profiles to model the molecular profiles of the cancer. The primary task of the technique is to presumably define the cause and effect variables of CRC using SEM that encompasses a set of linear equations between predictors and unobserved variables.

Materials and Methods

Data source

Microarray dataset was collected from GEO database with key terms: ("Colorectal neoplasm"[MeSH Terms] OR Colorectal cancer [All Fields]) AND "Homo sapiens"[porgn] AND "Expression profiling by array"[Filter]

The inclusion criteria to collect the dataset were as follows: (1) Sufficient sample size with matched normal mucosa and colorectal cancer tissues (2) the Microarray platform with whole Human genome Microarray chip from Agilent. (3) Data collected from DNA, extracted from peripheral blood lymphocytes and cancer tissues. (4) Data that contains metastatic colorectal cancer samples. Based on this criterion, 812 GEO datasets were listed of which the dataset (GSE87211) that contains a platform (GPL13497) with 363 samples of 203 colorectal cancers and 160 matched normal Mucosa were selected. These samples are retained with various characteristics like Colorectal cancer samples with wild type KRAS and mutated KRAS allele, preoperative radio chemotherapy and metastatic cancer examined by number of lymph nodes. The study dataset contains 15373 genes with least fold change of -6.37 and substantial fold change of 7.37. The genes with fold change of above 4.5 were mined from the samples and processed for further analysis.

Statistical Analysis

All the analysis described below was performed using statistical software SPSS version 16.0 and its package .

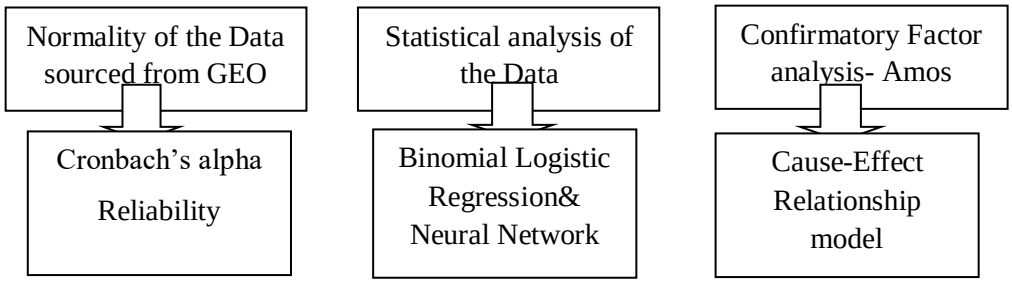

Figure1: ***Course of steps performed to generate a model that comprehend the causal model of colorectal cancer***

The data obtained from GEO2R was pre-processed in order to avoid the redundancy of the data. The data was checked for outliers which may highly present in microarray data samples caused by biological factors, mislabelled and contaminated samples [6]. Colorectal cancer is reported to be highly susceptible for presence of outliers in microarray data samples [7].

In order to index the reliability associated with the variation accounted for by the true score of the "underlying construct", Cronbach alpha was used to determine the internal consistency of the derived knowledge present in the data [8] which includes 44 items that measures the extent of association with colorectal cancer. The study Construct shows the good internal consistency in this sample (Cronbach's $\alpha = 0.99$). KMO (Kaiser-Meyer-Olkin) & the Bartlett's test of sphericity was used to validate if the data is suited for further factor analysis and also it measures the variance (common variance) among the predictors present in the data set [9]. The value that ranges between 0.8 and 1 and Bartlett's test of sphericity value <0.05 indicate the sampling is adequate [10], whereas the value obtained in the study dataset was 0.994 and Bartlett's test of sphericity value was 0.0001 that indicates adequacy of data for further analysis.

Logistic Regression Analysis

Binary logistic regression is a statistically fitting model and its use in medical diagnosis due to its visceral clinical interpretation by its entrenched approach [11][12]. The regression analysis is thus applied to study the relation between a set of predictors and a Dichotomous response variable. In this study, predictors are a set of highly expressed genes that can be used to generate a pattern to measure their strength of association with the colorectal cancer [13]**.** For example, the presence or the absence of the cancer for a specified period of time can be predicted from the knowledge of the fold changes of the genes that found to be differentially expressed. If X_1, X_2, X_3… denotes n predictor variables, Y denotes the presence(Y=1) or absence(Y=0) of disease and p determines the probability of presence or absence of the disease.

$\text{Log } (p/1\text{-}p) = ß_0 + ß_1X_1 + \ldots + ß_nX_n$

$ß_0$ is a constant and $ß_1$, $ß_2$, $ß_3$... represents regression coefficients of predictor variables which can be estimated from the available data. And thus the probability of the disease presence can be evaluated from the equation. The degree of contribution of predictor variables on the outcomes can be obtained from the regression coefficients of the predictors. Also the odds ratio can be estimated as follows. If $ß_1$ is the coefficient of the variable (INSL5- a gene with substantial fold change in this study), and p represents the probability of the presence of disease, Odds ratio is the exponential of coefficient of the variable. If the odds ratio value is 2, then the probability of the disease is twice as often in person with the presence of highly expressed gene (INSL5) than compared with a healthy person [14]. Wald Forward stepwise selection method was adopted, where the statistically significant variables that do meet the level of staying in the model are added in prior and the non significant variables are excluded based on Wald's Probability.

Machine Learning Technique

Artificial Neural Network is an inspiring machine learning model that represents biological neurons interconnecting between the (nodes) biological variables and dynamic in discovering the complex hidden interactions between the variables. But their robust performance in pledging any kinds of data [15][16] is broadly applicable to estimate the risk in variety of cancers [17][18].

Structural Equation Modelling

The tool encompasses various multivariate techniques such as multiple regression, path analysis and confirmatory factor analysis [19]. SEM is not a commonly used technique in mainstream statistics, it provides the researchers flexibility (lacks a strict formalization) in modelling relationships among multiple predictors and criterion variables, modelling errors in measurement for observed variables, and standardized terminology. And thus, SEM model is found to be highly potential, powerful and more flexible for developing new modelling approaches in Microarray data analysis [20]

Confirmatory Factor analysis (CFA)

This process is mainly adopted to reduce the dimensions of the data and to validate the construct of the data using SPSS package- Amos (Analysis of Moment of Structure). Confirmatory factor analysis is performed to construct a factor structure of the dataset based on the output of EFA. If the proposed model has maximum correlations inherent in the dataset, then the model is found to be good fit.

The metrics that ought to be reported to measure the fitness of the model are listed (Table 2) with their specific thresholds. Maximum Likelihood method was adopted to perform CFA as the method handles well even if the data has any missing values [21]. After conducting Confirmatory Factor Analysis using Maximum Likelihood Estimation, the fitness of the model was evaluated to assist degree of consistency of the proposed model accounting for correlations among the variables present in the dataset using Goodness of Fit Index (GFI) [22], Root Mean Square Deviation (RMSEA) and Chi- Square Fit Index. GFI evaluates the fitness between estimated and observed Covariance Matrix. The RMSEA evaluates the fitness by assessing the degree of unknown but optimally chosen parameter fitness with population covariance matrix. The RMSEA value <0.05 found to be a good model, value that ranges between 0.05-0.1 found to be moderate and value above 0.1 indicates the model fit is not good [23]**.** The Chi- Square Fit Index (CFI) range from 0 to 1 reflects the improvement in fit of hypothesized model over an independence model among the measured variables. The fitness of the model may also be affected not only by sample size and number of variables but also due to the discrepancies might present in the model. The modification indices offer remedies to remove the discrepancies between the proposed and estimated model and recommended to covary the error terms among the observed variables. The larger modification indices are given precedence before addressing to the minor ones. Standard Residual Covariances are similar to the modification indices as they even locate the discrepancies between the models. A standard significant residual covariance with absolute value lesser than 2.5 determines the good model fit.

Discriminant analysis

It is absolutely prerequisite to establish convergent and Discriminant validity as well as reliability of the model. Without validating the model, it is unfit on moving to causal interference analysis. The measures that describe validity and reliability are Composite Reliability (CR), Average Variance Extracted (AVE), Maximum Shared Variance (MSV), and Average Shared Variance (ASV). The Convergent Reliability should be greater than 0.7 and Average Variance Extracted should be greater than 0.5 [24]. Whereas, Discriminant validity was evaluated with Maximum Shared Variance (MSV) and Average Shared Variance (ASV) where both should be lower than the Average Variance Extracted (AVE) for all the constructs.

Results

Logistic Regression

From the Binomial Logistic Regression Analysis, 5 genes (ATP1A2, CA7, CLCA1, ABCA8, and SI) were found to be significant predictors.

From the Table 1 the fitted model is:

Logit (p) = (0.238) + (-1.687 X CA7) + (-1.643 X ATP1A2) + (.851 X ABCA8) +
(-1.034 X CLCA1) + (.567 X SI)

The regression analysis predicts that 2 genes (ABCA8, SI) were predicted with a significant odds ratio, indicating higher risk of disease studied, that the diseased subjects have a higher risk of mutational load compared to non-diseased one.

Model Assessment

R^2 statistic (coefficient of determination) is used to measure the suitability of predictor variables in predicting the outcome. Cox & Snell R Square are two such statistics that should attain maximum value less than 1. The

adjusted version of Cox & Snell R Square is Nagelkerke R^2 that should range between 0 to 1 [25]. The value ranging from 0.64 - 0.70 for Cox & Snell R Square and the values ranging from 0.85 – 0.94 for Nagelkerke R^2 obtained for the studied predictor variables indicated that the model is useful in predicting the colorectal cancer.

Table 1: Coefficients and Wald's test for logistic regression on colorectal Cancer data

Predictors (Genes)	Coefficients	Wald	df	P	OR	95%CI for Odds Ratio	
						LB	UB
CA7	-1.687	21.00	1	0.00	0.185	0.090	0.381
ATP1A2	-1.643	18.36	1	0.00	0.193	0.091	0.410
ABCA8	.851	4.317	1	0.03	2.341	1.049	5.222
CLCA1	-1.034	10.76	1	0.01	0.356	0.192	0.660
SI	.567	4.177	1	0.04	1.76	1.024	3.034

Artificial Neural Network

Using IBM SPSS, colorectal cancer ANN model was generated with a three layer feed forward network. The network consists of three layers with 44 independent variables in input layer, a hidden layer with 4 nodes and a single output layer with dichotomous response (0 and 1) that represents the risk of colorectal cancer. The obtained network was trained using Scaled Conjugate Algorithm (a supervised learning algorithm) which is reported as a benchmark against Standardised backpropogation algorithm and faster than other algorithms [26]. The ROC (Receiver Operating Characteristic) curve is a graph to plot the specificity (true positive rate) and sensitivity (false positive rate) for a different cut points of the predictive test studied. The ROC curve was used to validate the diagnostic performance of the test or its accuracy to distinguish the diseased cases from normal cases [27]. The area under the curve measures discrimination, that is, comprehension of the test to classify the diseased and non-diseased pairs accurately. The value of AUC (Area under Curve) should vary between 0.5(random guess) to 1(perfect fit). The area obtained for this study in classifying the groups is 0.99 which reveals the excellence and perfectness of the test.

Structural Equation Modelling

The fitness of the model was assessed with GFI (GFI = 0.731) and RMSEA (RMSEA = 0.1). The raw chi-square statistic χ^2 is 3292.668 and χ^2/df is 4.631 with p-value <0.0001. The raw and scaled χ^2 may be influenced by the sample size [28].

Table 2: Metrics for Good Model Fitness

Measure	Threshold	Model Outcome
Chi-Square/df (CMIN/DF)	<5	4.63
p-Value	<0.05	0.0001
CFI	>0.90	0.924
GFI	>0.90	0.731
AGFI	>0.80	0.609
NFI	>0.90	0.906
IFI	>0.90	0.925
TLI	>0.90	0.894
RFI	>0.90	0.869
RMSEA	<0.5-good; 0.5-1.0- moderate	0.100

CFI- Confirmatory Fit Index; GFI- Goodness of Fit Index; AGFI- Adjusted Goodness of Fit Index; NFI-Normed Fit Index; IFI- Incremental Fit Index; TLI- Tucker-Lewis Index; RFI- Relative Fit Index; RMSEA- Root Mean Square Deviation

Discriminant Analysis

Convergent validity is measured by the Reliability and Average Variance Extracted (AVE), Factor loadings of all the variables implemented to construct the model using CFA. The Factor loadings are represented as Regression weights in Amos. On addressing to the Factor loadings, the Standardised Regression weights should be greater than 0.5 or higher, ideally the range above 0.7 is found to be highly significant. The Factor loadings of the variables implied in this study found to be highly significant (Table 4). In addition to Factor loadings, Cronbach's alpha reliability *(Cronbach, 1951)* scale indicates that the indicators present in the model are consistent with their construct (latent variable).

A best approach to elucidate the Discriminant validity is to compare the squared correlation between the construct with its individual Average Variance Extracted (AVE) estimates. Discriminant validity issues are due to very high cross loadings in the model [29].

Table 3: Standardized Regression Weights (Factor loadings)

	Indicators		Estimate
LOC157503	<---	F1	.863
CA7	<---	F1	.889
SCGN	<---	F1	.936
LOC283454	<---	F1	.846
CPB1	<---	F1	.910
SST	<---	F1	.921
SCN7A	<---	F1	.864
CA2	<---	F1	.966
GUCA2B	<---	F1	.932
NOL4	<---	F1	.872
GCG	<---	F1	.965
ATP1A2	<---	F1	.814
OR8D4	<---	F1	.781
RFX6	<---	F1	.879
ABCA8	<---	F1	.910
INSM1	<---	F1	.882
NEUROD1	<---	F1	.872
SLC4A4	<---	F1	.972
CA1	<---	F1	.976
INSL5	<---	F1	.967
ADH1C	<---	F1	.932
CCL23	<---	F1	.774
GUCA2A	<---	F1	.945
ADH1A	<---	F1	.931
TMIGD1	<---	F1	.967
C21orf88	<---	F1	.942
TTR	<---	F1	.874
SLC30A10	<---	F1	.936
HEPACAM2	<---	F1	.917
CHGA	<---	F1	.929
CWH43	<---	F1	.854
CLDN8	<---	F1	.969
SLC26A3	<---	F1	.867
CDKN2BAS	<---	F1	.921
CLCA1	<---	F1	.870
DPP10	<---	F1	.785
MS4A12	<---	F1	.926
CD177	<---	F1	.876
CA4	<---	F1	.932
LOC389023	<---	F1	.814
CLCA4	<---	F1	.896
SI	<---	F1	.880
ITLN1	<---	F1	.886
ZG16	<---	F1	.931

Note: F1 represents the dichotomous variable (0(case) &1(control))

Table 4: Convergent and Discriminant Validities computation of CFA model generated

INDICATORS	CR	AVE	MSV	ASV

INSM1	**0.78**	**0.78**	**0.16**	**0.04**
LOC157503	**0.74**	**0.74**	**0.32**	**0.07**
LOC283454	**0.72**	**0.72**	**0.06**	**0.03**
ATP1A2	**0.66**	**0.66**	**0.50**	**0.12**
SCN7A	**0.75**	**0.75**	**0.09**	**0.03**
GUCA2B	**0.76**	**0.76**	**0.12**	**0.03**
ZG16	**0.86**	**0.86**	**0.09**	**0.04**

Discussion

The current study investigated whether a new microarray dataset of colorectal cancer from GEO database is suitable for the better understanding of molecular risk factors of the disease devised through factor analysis. The 44 genes that are expressed in a substantial fold change of greater than 4.5, was taken to study their association for the cause of disease. Data mining was performed in 363 samples which include 203 colorectal cancer and 160 matched normal healthy samples with respect to the genes taken for the study. The statistical investigation was further performed with internal consistency and reliability of the data. The binomial logistic regression was executed in the study using Forward stepwise Wald technique in order to locate the most significant genes associated with the outcome. The genes ATP1A2, CA7, CLCA1, ABCA8 and SI were predicted to be more significant and two genes ABCA8 and SI were found to have significant odds ratio (>95%CI). Further, Area under curve obtained from Multilayer Perceptron Neural Network was investigated and the significant value (0.99) of Area under the curve predicted that the model generated from the data studied was good and more efficient in classifying the diseased and non diseased data.

Structural Equation Modelling analysis on the gene markers predicted the inter correlations among the endogenous variables and latent factor. For that Confirmatory Factor Analysis was conducted and fitness of the model was inspected. Therefore the CFI value of 0.92 shows the study model is relatively a good fit. The RMSEA value of 1.00 in this sample indicates a moderate fit. The GFI and AGFI value of this sample 0.73, 0.60 are below 0.9, but both the Fit indices are known to depend on sample size [30]**.** The other fit index NFI, IFI, TLI, RFI of this sample is almost close to 0.9 indicates an acceptable model fit.

The Factor loadings of all the items are greater than 0.7 (Table 3) which shows significant factors in association with the disease [31]. After the validating the fitness using Excel Stats Package, the genes INSMI, LOC157503, LOC283454, ATP1A2, SCN7A, GUCA2B, ZG16 were predicted to be highly correlated with one another and do not measure well separated latent factor. However many items were identified with larger factor loadings than the corresponding reliability and validity metrics (Table 4). As shown in Table 4, all the items except ATP1A2 were relatively good convergent and well-associated with one another.

ATP1A2

ATP1A2 encodes the protein alpha-2 isoform of the Na (+), K (+)-ATPase (EC 3.6.1.9) responsible for maintaining electrochemical gradient of NA^+ and K^+ ions across the plasma membrane. Mutations in this gene reported to result in rare alternating hemiplegia of childhood [32]. No studies reported on significant mutation of this gene in association to colorectal cancer.

CA7

Carbonic anhydrase 7 is part of anhydrase family belongs to large family of Zinc metalloenzymes that involves in the reversible hydration of carbon dioxide. Their extensive diversity in tissues participates in various biological processes like acid base balance, bone resorption, calcification, and Cerebrospinal fluid formation. They also involve in positive regulation of pH in cells. The up regulated expression of this gene acts as a good indicator of disease prognosis in colorectal cancer [33]**,** but acts a poor prognostic indicator in patient with astrocytomas [34].

ABCA8:

ABCA8 is ATP-binding cassette, sub-family A (ABC1) member 8 encodes, membrane-associated protein that may regulate in lipid metabolism, formation and maintenance of myelin. The gene is found to be under expressed in Colorectal cancer patients in the early stage of cancer even before the patient is administered with 5-fluorouracil which is given in the first line treatment of disease [35]**.** But this study predicts that, the gene ABCA8 is found to be highly expressed in the advanced stages of cancer and may act as a good indicator for the disease prognostication. To defend this study, the gene is found to be manifested, in metastasis of liver and colon cancer (Stage II and III) [36] and considered as one of the top hub genes in the Cancer Network Galaxy

SI:

This protein encoding gene encodes Sucrose-Isomaltase enzyme expressed in Intestinal brush border, responsible for digestion of dietary carbohydrates. Various research has been carried out only on the under expression of gene/ deficiency of the protein, which would result in Irritable Bowel syndrome. But only few studies are reported on the highly expressed activity of SI gene that is identified in 80% of colon adenocarcinomas. This type of study would exploit in the clinical management of the disease [37].

CLCA1:

Chloride Channel, Calcium Activated, Family Member 1 (CLCA1) is protein encoding gene involve in calcium activated chloride conductance. This gene is mainly responsible for tumour suppressing activity. Many studies published on the critical roles of mutated CLCA1, in differentiation and proliferation of colorectal cancer and acts as a potential predictor of disease prognosis in human colorectal cancer [38]. The prognosis of disease is well predicted when the expression of the gene is high. But in advanced stages of CRC, the gene expression is studied to be low [39].

INSM1:

Insulinoma Associated protein-1 is an intronless gene that encodes a protein containing both Zinc Finger DNA binding domain and Putative Prohormone domain. The protein plays a crucial role in neuroendocrine differentiation in lung tumour and acts as a diagnostic marker. But there are no studies on this protein relating to colorectal cancer.

The items (**LOC157503, LOC283454)** are reputed as uncharacterised protein encoding gene found to be highly correlated with other genes in causing risk to colorectal cancer. From the GEO profiles, this protein encoding gene is found to be crucial for many diseases.

SCN7A:

Sodium Voltage-Gated Channel Alpha Subunit (SCN7A) gene encodes voltage gated sodium channels proteins responsible for membrane depolarization. Very few studies were published on SCN7A gene as a colorectal cancer risk factor. But this gene was studied to be down regulated in colorectal cancer [40]**.** In contrast, this study reported that the gene is found to be up regulated in the progression of colorectal cancer and highly correlated with other genes. Also the gene was studied to play a crucial role in oligodontia-colorectal cancer syndrome [41].

GUCA2B:

This gene encodes a preproprotein which cleaves into multiple proteins of guanylin family of peptides and also acts as an endogenous ligand of Guanylate cyclase- receptor to regulate salt and water homeostasis in intestines and kidneys. This protein encoding gene is highly found as a novel gene to be associated with colorectal cancer and acts a potential biomarker in the early detection of cancer [42].

ZG16:

Zymogen granule membrane protein plays an important role in protein trafficking, enzyme foldings. In contrast to this study, the expression of Zymogen granule protein 16 was reported to be completely lost in colon cancer [43].

Conclusion

We describe a combined statistical model using Logistic Regression, Artificial Neural Network and Structural Equation Modeling to infer information from Gene Expression Data analysis in the context of colorectal cancer. The model described in this report selected 44 candidate genes exceeding 4-fold threshold expression.

Acknowledgment

We would like to show our sincere gratitude to the management of Sri Ramachandra Institute of Higher Education & Research [DU], for providing us constant support and continuous encouragement during the course of this research.

References

1. Greene CS1, Tan J, Ung M, Moore JH, Cheng C, "Big data bioinformatics". *J Cell Physiology*, 2014
2. intoday.in. [Online].; 2016 [cited 2016 May 18. Available from: http://indiatoday.intoday.in/story/over-17-lakh-new-cancer-cases-in-india-by-2020-icmr/1/671461.html].
3. Rebecca L. Siegel MPH,Kimberly D. Miller MPH, "Cancer statistics". CA: *A Cancer Journal for Clinicians.* p. 7-30, 2016.
4. Swaminathan R1, Shanta V, Ferlay J, Balasubramanian S, Bray F, Sankaranarayanan R, "Trends in cancer incidence in Chennai city (1982-2006) and statewide predictions of future burden in Tamil Nadu (2007-16)". *PubMed.gov.* p. 72-7, 2011.
5. Shuji Ogino and Ajay Goel, "Molecular Classification and Correlates in Colorectal Cancer". *The Journal of Molecular Diagnostics.* p. 13-27, 2008.
6. Li, L., Darden,T. A., Wei nberg, C. R., Levine, A. J., and Pedersen, L. G, "Gene assessment and sample classification for gene expression data using a genetic. Combinatorial Chemistry & High.";: p. 727–739, 2001.
7. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine. Broad patterns of gene expression revealed by clustering of tumor and normal. Proceedings of the. p. 6745–50, 1999.
8. Mohsen Tavakol, Reg Dennick, "Making sense of Cronbach's alpha". *International Journal of Medical Education*. p. 2:53-55, 2011.

9. Snedecor, George W. and Cochran, William G. Statistical Methods Iowa: Eighth Edition, Iowa State University Press; 1989.
10. Cerny CA,&KHF, "A study of a measure of sampling adequacy for factor-analytic correlation matrices". *Multivariate Behavioral Research*. p. 43-47, 1977.
11. Bartfay E, Mackillop WJ, Pater JL, "Comparing the predictive value of neural network models to logistic regression models on the risk of death for small-cell lung cancer patients". *Eur J Cancer Care*. p. 115–124, 2006.
12. DJ S. Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer.* p. 1636–1642, 2001.
13. Jr. FEH, Binary Logistic Regression. In HarrellJr. FE. Regression Modeling Strategies. Nashville, TN, USA : *Springer International Publishing;* p. 219-274, 2015.
14. Turgay Ayer, MS, Jagpreet Chhatwal, PhD, Oguzhan Alagoz, PhD, Charles E. Kahn, Jr, MD, MS, Ryan W. Woods, MD, MPH, and Elizabeth S. Burnside, MD, MPH, MS, "Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. RadioGraphics".;: p. Volume 30, Issue 1, 2010.
15. Dvorchik I, Demetris AJ, Geller DA, Carr BI, Fontes P, et al. "Prognostic models in hepatocellular carcinoma (HCC) and statistical methodologies behind them". *Curr Pharm Des.* 2007;: p. 1527–1532.
16. Ho WH CC, "Genetic-algorithm-based artificial neural network modeling for platelet transfusion requirements on acute myeloblastic leukemia patients. Expert Systems With Applications". p. 6319–6323, 2011.
17. WG B, "Application of artificial neural networks to clinical medicine". *Lancet.*: p. 1135–1138, 1995.
18. Lisboa PJ, Taktak AF, "The use of artificial neural networks in decision support in cancer": *a systematic review. Neural Netw.* p. 408–415, 2006.
19. Rabe-Hesketh S and Skrondal A, "Classical latent variable models for medical research". *Stat Meth Med Res*. p. 5–32, 2008.
20. F Martella1 and JK Vermunt2, "Model-based approaches to synthesize microarray data: a unifying review using mixture of SEMs". *Statistical Methods in Medical Research*.;: p. 567-82, 2013.
21. Bentler, "EQS 6 Structural Equations Program Manual". Multivariate Software. ; 2006.
22. Tabachnick B, Fidell L. Using Multivariate Statistics. In Bacon Aa. Using Multivariate Statistics. New York; 2013.
23. Hu LT, Bentler PM, "Cutoff criteria for fit indexes in covariance structure analysis. Conventional criteria versus new alternatives". Structural Equation Modeling. p. 1-55, 1999.
24. Hair, J., Black, W., Babin, B., and Anderson, R. Multivariate data analysis (7th ed.) Upper Saddle River, NJ, USA.: Prentice-Hall, Inc.; 2010.
25. Viv Bewick1, Liz Cheek1 and Jonathan Ball2. Statistic Review: Logistic regression. Critical Care. p. Vol9 No1, 2005.
26. Møller MF, "A scaled conjugate gradient algorithm for fast supervised learning". *Science Direct ELSEVIER*. p. 525–533, 1993.
27. Zweig MH, Campbell G. Clinical Chemistry, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine". p. 561-577, 1993.
28. Hooper D, Coughlan J, Mullen M, "Structural equation modelling: guidelines for determining model fit". *The Electronic Journal of Business Research Methods.* p. 53-60, 2008.
29. U. K. Jayasinghe-Mudalige , J. M. M. Udugama and S. M. M. Ikram, "Use of Structural Equation Modeling Techniques to Overcome the Empirical Issues Associated With Quantification of Attitudes and Perceptions". *Sri Lankan Journal of Applied Statistics*. p. 15-37, 2012.
30. S. A. Mulaik, L. R. James, J. Van Alstine, N. Bennett, S. Lind, and C. D. Stilwell, "Evaluation of goodness-of-fit indices for structural equation models". *Psychological Bulletin*. p. 430–445, 1989.
31. D. Gefen, D. Straub, and M.-C. Boudreau, "Structural equation modeling and regression: guidelines for research practice". Communications of the Association for Information Systems. p. vol. 4, no. 1, article 7, 2000.
32. Colorectal Cancer Atlas. [Online]. [cited 2017 March 27. Available from: http://colonatlas.org/gene_summary?gene=ATP1A2.

33. Yang GZ, Hu L, Cai J, Chen HY, Zhang Y, Feng D, Qi CY, Zhai YX, Gong H, "Prognostic value of carbonic anhydrase VII expression in colorectal carcinoma". BMC Cancer. p. 15:209, 2015.
34. Bootorabi F, Haapasalo J, Smith E, Haapasalo H, Parkkila S, "Carbonic anhydrase VII–a potential prognostic marker in gliomas". Health. p. 6-12, 2011.
35. I. Hlavata B. Mohelnikova-Duchonova R. Vaclavikova V. Liska P. Pitule P. Novak J. Bruha O. Vycital L. Holubec V. Treska, "The role of ABC transporters in progression and clinical outcome of colorectal cancer". *Mutagenesis.* p. Volume 27, issue 2, 2012.
36. the Cancer Network Galaxy 0.14. [Online].; 2013 [cited 2017 march 27. Available from: http://sign.hgc.jp.
37. Othon Wiltz. Carl J. O'Hara, Glenn D. Steele Jr, Arthur M. Mercurio, "Expression of enzymatically active sucrase-isomaltase is a ubiquitous property of colon adenocarcinomas". Science Direct, p. 1266-1278, 1991.
38. Bo Yang, Lin Cao, Bin Liu, Colin D. McCaig, Jin Pu, "The Transition from Proliferation to Differentiation in Colorectal Cancer Is Regulated by the Calcium Activated Chloride Channel A1". *PLOS one*, 2013.
39. Yang Yu, Vijay Walia, Randolph C. Elble., "Loss of CLCA4 Promotes Epithelial-to-Mesenchymal Transition in Breast Cancer Cells". PLOS one, 2013.
40. Beata Ostasiewicz, Paweł Ostasiewicz, Kamila Duś-Szachniewicz, Katarzyna Ostasiewicz, and Piotr Ziółkowski, "Quantitative analysis of gene expression in fixed colorectal carcinoma samples as a method for biomarker validation". *Molecular Medicine Reports.* p. 5084–5092, 2016.
41. Life map sciences. [Online].; 2017 [cited 2017 March 28. Available from: http://www.malacards.org/card/normokalemic_periodic_paralysis.
42. Shivashankar H Nagaraj, Antonio Reverter, "A Boolean based systems Biology approach to predict novel genes associated with cancer: Application to colorectal cancer". *BMC Systems Biology*. p. 5-35, 2011.
43. Liang Wang et al., "Loss of expression of zymogen granule protein 16 in colorectal cancer". *Cancer Research*, p. vol 26, 2005.