Expression Profiling & Classification using Convolutional Neural Networks of Tumor Suppressor Genes Linked with Stress

Kaajal Nishandh, Sanjay Kumar P, P.K Krishnan Namboori

Abstract*: Tumor suppressor genes are always linked with stress, directly or indirectly which results in mutation. Therefore the probability of turning these mutations into cancer increases. Identification of major tumor suppressor genes and its presence among Indian population is analyzed. Due to great advancement in the field of deep learning, and wide variety of scopes in future, deep learning is incorporated in this project to perform the classification task .The requirement of large amount of data to perform classification task is one of the major drawback of deep learning. In order to solve this problem, one-shot learning algorithm is introduced which gave the accuracy of 70.2%. A secure data sharing platform has been developed using blockchain technique.

***Index Terms*: Block Chain Technique, Deep Learning, Tumor suppressor genes.**

I. INTRODUCTION

Stress is a typical natural response to dangerous situations of any kind of demand or threat, which is accompanied with a number of chemical and biological aberrations and resulted by the release of chemicals and hormones such as adrenaline and cortisol. In a biomolecular perspective, it is found to be extended even up to the cellular level. Cellular stresses is generally resulted out of the cells being exposed to external and internal stressors which damages the integrity of the cell and its genome such as cigarette smoke, hypoxia, ionizing radiations, oxidative stress, carcinogens, oncogene activation etc. [1]. Apart from mental and physical stresses, another main stress is the cellular stress. Cellular stresses can happen when human cell is exposed to external and internal stressors which damages the integrity of the cell and its genome such as cigarette smoke, hypoxia, ionizing radiations, oxidative stress, carcinogens, oncogene activation etc. [1]. This can lead to damage of DNA and also malignant transformation of cells. To ensure the survival of the organisms, cells have built up various methodologies to adjust to stressors i.e. the “Tumor Suppressor Genes” [2]. Due to this evolving nature of biological systems, it becomes difficult to carry out the big data analysis using the basic techniques. This gap could be bridged using emerging machine learning techniques such as deep learning which works with huge data sets. With the implementation of Siamese net or One- shot learning, one could carry out the classification problem with minimal data with better prediction accuracy [3].

Gene expression analysis is important to understand the association of tumor suppressor genes in the regulation of the disease. The ‘gene expression profiling’ can be considered as the ‘mutation signature’ of the disease. Although gene expression profiling is presently used as a primary research tool, many other potential clinical applications of this method are being instigated. The expression profile renders prognostic data about the aggressiveness nature of tumor, response to therapies, sensitivity and resistance to different chemotherapies [4]. The expression level varies from person to person, demanding a pharmacogenomic investigation in the subject [4] [5]. The knowledge of pharmacogenomic with genomic, epigenomic, metagenomic and environmental genomic components, helps in developing safe and effective medications according to the person’s genetic composition. The proneness of the disease as well as the susceptibility of drugs depends up on the individual variations. Epigenomics deals with the study of the entire set of epigenetic alterations on the genetic material of a cell (epigenome), Metagenomics is the study of genetic material taken from micro-organisms such as bacteria, yeast etc and environmental genomics deals with the prediction of how the organisms respond to external environmental factors [6]. Pharmacogenomics promises personalized treatment for patients suffering from common diseases, especially with multiple treatment modalities [7]. Pharmacogenomics provides the population specific marker frequency profiles and also drug efficiency. Various machine learning applications in healthcare helps in early diagnosis and prediction of suitable treatment strategy to the patient. Recently, deep learning algorithm has been identified as most suitable for identifying cancerous tumors from mammogram. At present, training the machine with minimum number of samples could be possible with modern learning experiences such as ‘one shot learning or Siamese net’. One-shot learning ranks the similarity of inputs in which the classification is done using K-nearest neighbor calculating the Euclidean distance between the test and train data, classifying the nearest ones as in the same class. Since health care deals with confidential patient

Revised Manuscript Received on January 25, 2019

Kaajal Nishandh, Department of Electronics and Communication Engineerining, Amrita School Of Engineerining, Amrita Vidhyapeetham, Coimbatore, India.

Sanjay Kumar P, Amrita Molecular modeling and synthesis (AMMAS) Research Lab, Computational Engineerining and Networking, Amrita School Of Engineerining, Amrita Vidhyapeetham, Coimbatore, India.

P.K Krishnan Namboori, Molecular modeling and synthesis (AMMAS) Research Lab, Computational Engineerining and Networking, Amrita School Of Engineerining, Amrita Vidhyapeetham, Coimbatore, India.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

information, security plays a very important role in dealing with pharmacogenomics. Biometric authorization is suggested in order to reduce the security threats. There are several algorithms such as AES, DES etc. that are used to improve the security levels of details provided by the person. Secure sharing of data could be made possible via block chain technology [8]. Health security becomes difficult because all the client architectures are for central administrator. Block chain technology eliminates the requirement of a central administrator by the use of cryptography and also all the users can control and manage their information. Thus block chain offers access security, data privacy and scalability on a single go [9].

II. MATERIALS AND METHODS

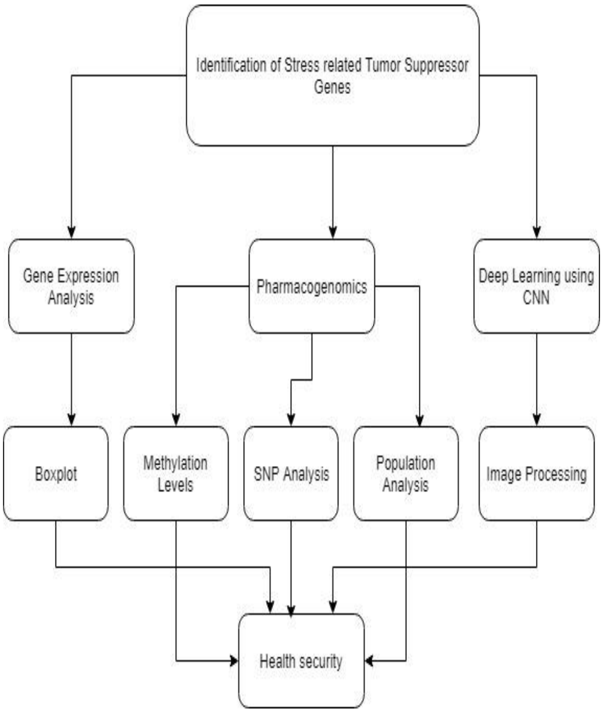

Fig.1 Workflow

Pharmacogenomics study was carried out by using various data base tools. Most relevant ten tumor suppressor genes was acquired using National Centre for Biotechnology Information (NCBI) database. [10][11] Instead of taking single gene, stress associated genes were also considered and expression profile was taken. Gene expression analysis was carried out using Expression Atlas and then a gene profile was taken. Gene expression was analyzed [12] using deep learning. Using R-Studio, a Boxplot consisting of all the 10 genes identified in the expression profile, was plotted. The SNPs were identified using 'SNP Database' of NCBI, which helps in finding the genetic signature of each gene mutations. Using 'Indian Genome Variation Database (IGVDB) a population analysis was carried out [13]. Methylation level of each gene in promoter and gene body was identified using 'Meth bank' [14].

Convolutional Neural Networking which is the latest approach to deep learning – one shot algorithm also known as Siamese network was used for microscopic pathological images of protein expression analysis, which was acquired from Human Protein Atlas database[15][16]. The training was done by dividing the pathological images into 80% training and 20% evaluation.

The patient specific information has been secured by using health security norms. A platform is set up for encrypting the details of the patients by storing the information in the block chain, requesting an entry to already created block chain. Each block in the block chain consists of hash value that depends on the previous block. The information and patient id are stored which can be used by authorized organization under specific requirement. All the details in block chains are encrypted using cryptography. Personal digital signatures will be given to all the participants to access the provided data. The signature will be invalid if a single data is altered which will alert the participants about the modification made. Further, damage of data can be prevented by early notification about the attacks. The requirement of large amount of computing power to access and make corresponding changes acts as a resistance to modifications of data.

III. RESULTS AND DISCUSSIONS

Top 10 Tumor suppressor genes such as TP53, AKT1, VEGFA, CDKN2A, ESR1, BCL11B, STAT3, BRCA1, BRCA2, PTEN were considered during this study. The expression of the genes were tested in 56 different cell lines. As per the gene expression analysis, it was found that STAT3 and AKT1 are the driver genes and the rest of genes namely VEGFA, CDKN2A, ESR1, BCL11B, BRCA1, BRCA2, PTEN, TP53 are identified as the passenger genes. With the help of the TPM values of each gene, gene expression profiling is done (Fig 2).

A. Gene Expression Profile

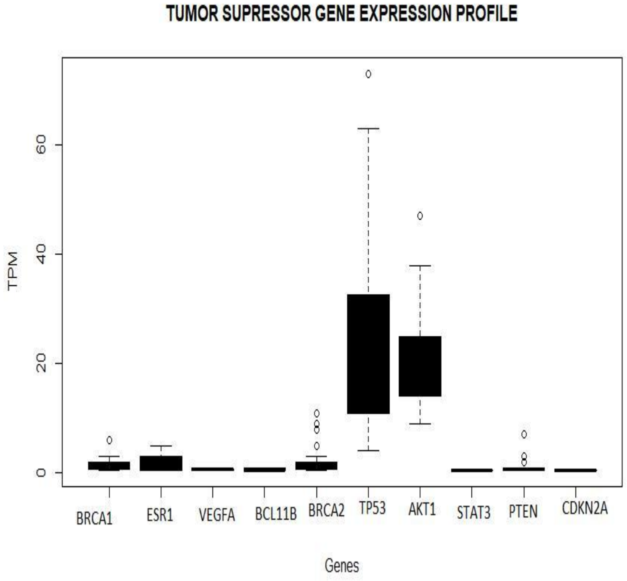

Fig.2. Gene Expression Profile

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Fig.2 indicates boxplot which includes all the 10 tumor suppressor genes. From the gene expression profile, we can predict that TP53 and AKT1 is expressed 25% more out of 56 cell lines compared to other genes and the least expressed gene is CDKN2A, VEGFA, STAT3.

B. Pharmacogenomics

Table .1 SNPs present in Indian Population

Sl. no	Genes	No: of SNPs Pathogenic/ Synonymous	SNPs Specific To Indian Population
1	BRCA1	527/18	rs8176265 , rs8176257, rs2070833, rs1060915
2	TP53	97/3	rs1642785, rs1042522, rs8064946,rs12951053 rs1641549 , rs1625895, rs2078486
3	PTEN	124/5	rs17107012, rs12572106,rs2299939, rs2735343,rs532678, rs1234225, rs1234223
4	BCL11B	1	NIL
5	ESR1	6	NIL
6	CDKN2A	16/2	NIL
7	AKT1	3	rs2230506, rs2498799, rs2498796 rs2494731, rs2494732, rs1130214 , rs3730358 ,rs2494743
8	STAT3	21/1	NIL
9	VEGFA	NIL	rs3025010, rs25648
10	BRCA2	609/20	rs11571579, rs542551, rs206146

From the Table.1, we could identify that TP53 has got more number of SNPs present in Indian population. The missense pathogenic SNPs were included in the analysis.

C. Methylation

Methylation levels (%) of the promoter and gene body is included in Table.2. It can be observed that, in TP53, AKT1, VEGFA, STAT3, CDKN2A, BCL11B, BRCA2, PTEN the chances of methylation in the promoter side is much lesser compared to that of the percentage methylation levels in gene body. But in the case of BRCA1 and ESR1, the chances of methylation in promoter is much higher than that of the gene body.

Table.2 Methylation levels

Sl. No	Genes	Promoter Methylation level (%)		Gene body Methylation level (%)	
		Female	Male	Male	Female
1	TP53	3.43	3.4	40.9	40.908
2	AKT1	24.38	24.2	84.97	85.12
3	VEGFA	40.4	40.3	48.47	48.49
4	STAT3	1.9	1.88	39.55	39.58
5	CDKN2A	4.688	4.61	9.039	9.19
6	BCL11B	6.701	6.54	46.25	46.12
7	BRCA1	60.66	59.7	33.47	33.51
8	BRCA2	8.86	8.53	34.9	34.96
9	ESR1	73.014	73.3	48.89	49.02
10	PTEN	3.56	3.5	26.74	26.73

D. Deep Learning

Two pathological images for ten tumor suppressor genes and for 17 different cancers have been taken for the analysis. The class prediction has been done using the one –shot algorithm with an accuracy of 70.2%.

E. Blockchain Technique

Patient specific data has been secured by the security system, which uses a block chain technology. Block chains help in secure sharing of data among the hospitals, research institutions, medical practitioners etc. The addition of health security norms into pharmacogenomics serves as a potential strategy for pharmacogenomics.

IV. CONCLUSION

Deep learning using convolutional neural networking is found to be an effective way for pre-detection of cancers caused due tumor suppressor genes. Moreover, this methodology helps in the prediction of classes from pathological images. This proposal model is presumed to make large scale and in -depth recognition of image on primary pathological descriptions. This method is an additional platform for pharmacogenomics predictions on proclivity towards cancer. The proneness to cancer could be predicted through SNP test analysis. Data analytic techniques of Indian population helps in the generation of genetic signatures. The use of convolutional neural network for pathological image processing serves as an ingenious and efficacious technique for more evaluation of suspected cases.

ACKNOWLEDGMENT

The authors express their sincere gratitude towards the officials and members of Amrita Molecular Modeling and Synthesis (AMMAS) Research Lab at Amrita Vishwa Vidyapeetham, Coimbatore for their support.

REFERENCES

1. S. Chen, B. Mulgrew, and J. Pflaum, S. Schlosser and M. MÃuller, "p53 Family and Cellular Stress Responses in Cancer", 2018. .
2. M. Chircop and D. Speidel, "Cellular Stress Responses in Cancer and Cancer Therapy", 2018.
3. Gregory Koch et al (2015) Siamese Neural Networks for One-shot Image Recogni-tion.Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.37
4. C. HimaVyshnavi A M, Lakshmi Anand C, Deepak O M, P K Krishnan Namboori.Evaluation of Colorectal Cancer (CRC)
5. Epidemiology A Pharmacogenomic Approach.J Young Pharm, 2017; 9(1): 36-39.
6. https://www.nature.com/subjects/pharmacogenomics
7. Annualreviews.org,2018.[Online].Available:https://www.annualreviews .org/doi/abs/10.1146/annurev-cancerbio-050216-121919.
8. Preethi M. Iyer et al (2016) Brca1 responsiveness towards breast cancer-a population-wise pharmacogenomic analysis. IJPPS 8(9):267
9. M. Zhang and Y. Ji, "Blockchain for healthcare records: A data perspective", 2018.
10. Capgemini (2017) Blockchain: A Healthcare Industry View
11. Sherry ST et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308-11
12. https://www.ncbi.nlm.nih.gov/gene accessed on November 2018.
13. https://www.ebi.ac.uk accessed on November 2018.
14. Indian Genome Variation Consortium (2005) The Indian Genome Variation database (IGVdb): a project overview. Hum Genet 118(1):1-11
15. R. Li, F. Liang, M. Li, D. Zou, S. Sun, Y. Zhao, W. Zhao, Y. Bao, J. Xiao and Z. Zhang, "MethBank 3.0: a database of DNA methylomes across a variety of species", 2018.
16. Iyer, Akshay & Vyshnavi A M, Hima & Namboori P K, Krishnan. (2018). Deep Convolution Network Based Prediction Model For Medical Diagnosis Of Lung Cancer - A Deep Pharmacogenomic Approach : deep diagnosis for lung cancer. 1-4. 10.1109/ICAECC.2018.8479499.
17. https://www.proteinatlas.org/ accessed on November 2018.
18. Tarca, R. Romero and S. Draghici, "Analysis of microarray experiments of gene expression profiling", American Journal of Obstetrics and Gynecology, vol. 195, no. 2, pp. 373-388, 2006.