Long Non-coding RNA Based Cancer Classification using Deep Neural Networks

Abdullah Al Mamun and Ananda Mohan Mondal
Florida International University, Miami, FL
Email: {mmamu009, amondal}@fiu.edu

Abstract

Recent studies indicate that lncRNA plays key roles in tumorigenesis and misexpression of lncRNAs can lead to change in expression profiles of various target genes involved in different aspects of cancer progression. However, research on classifying multiple cancer types using only lncRNA is rarely found. In this paper, we explored the capability of lncRNA in classifying cancer types by employing four deep neural networks - multi-layer perceptron (MLP), long-short-term memory (LSTM), convolutional neural network (CNN) and deep autoencoder (DAE). For experiment, RNA-seq expression values from TCGA for 8 cancers - BLCA, CESC, COAD, HNSC, KIRP, LGG, LIHC, and LUAD - are used. The combined dataset consists of 3656 patients with expression values for 12309 lncRNAs. The performance of the models in terms of accuracy ranges from 94% to 98%, which shows lncRNA expression profiles as the better signature compared to the mRNA expression profiles in classifying cancer types.

Introduction

Several cancer risk loci are transcribed into lncRNAs and these transcripts play key roles in tumorigenesis and misexpression of lncRNAs can lead to change expression profiles of various target gene involved in different aspect of cell homeostasis [1–3]. They identified important lncRNAs those are actively involve in different stages of cancer development [1]. For example, ANRIL PCR1-mediated repression of INK4A-ARF- INK4b tumor suppressor locus, up-regulated in prostate cancer [4]. Similarly, H19 plays significant role in transcriptional co-activation and repression stages as up-regulated in gastric cancer [5]. In addition, Adam et. al. discussed the impact of lncRNA in cancer pathway. More specifically, authors describe the involvement of lncRNAs in different cancer stages such as proliferation, growth suppression, motify etc [6].

Integrated analysis of twelve cancer types based on gene, microRNA, protein expression, copy number variation, and DNA methylation revealed that many tumor types has unique features for a particular cancer type [7]. Consequently, cancer classification using RNA-seq expression have revealed distinct cancer subtypes and uncover various patterns which were supported by many clinical outcomes [8,9]. Thus, building an intelligent approach that will learn knowledge from multiple tumor types for classification can be a remarkable contribution to an efficient cancer diagnosis system. However, very few studies analyzed the multiple cancer classification problem using deep neural networks. Li et. al. developed GA/KNN to classify cancer and normal samples for 31 cancers using RNA-seq gene expression data [10]. Having similar goal, Boyu et. al. approached a different method to develop the cancer classifier for the same dataset [11]. They converted RNA-seq data of individual sample to a heat-map image, later the image has been used to train the CNN model which requires a lot of computational power. In this paper, we aim to classify multiple cancer types based on lncRNA expression values using deep neural networks. Contribution of this research is to develop an intelligent classifier that will be able to classify cancer types automatically.

Data and Preprocessing

To validate the idea, we downloaded (April, 2019) RNA-seq FPKM normalized expression data for 33 cancers from UCSC xena. Combined dataset consists of 11057 patients with respect to 60483 RNAs representing 33 tumor types. The row represents the RNA Ids while column represents the sample Ids. Values of each cell represents the normalized read counts of a sample for that specific RNA sequence. The number of combined list of samples is 3656 after merging all the cancers. In this study, we used lncRNA expression as classification features. The database TANRIC provides 12727 lncRNA Ids [34] that has been used to extract expression of lncRNAs from the combined RNA expression data. Now we have 12309 common lncRNA Ids with expression data for all selected cancers.

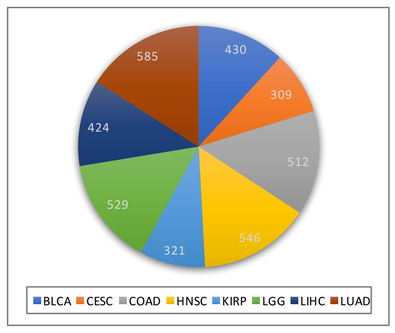

Figure 1: Sample Distribution of Selected Tumor Types

Methodology

We used four different deep learning algorithms - Multilayer Perceptron (MLP), Gated Recurrent Unit Long-Short Term Memory (GRU-LSTM), Convolutional Neural Network (CNN) and Deep Auto Encoder (DAE) - to develop the classification models.

The number of hidden layer is four. The number of node in input layer is equal to the number of features (lncRNAs), which is 12309. All the hidden layers also have the same number of nodes as in input layer and the layers are fully connected. The output layer has 8 nodes corresponding to 8 different cancer types. After tuning hyper-parameters and optimizing model parameters, we found a good convergence with learning rate 0.01, epoch 500 and seed 123. The hyper-parameter seed is used to randomly select the samples for testing and training. These parameters adjust the network for appropriate weights to prevent over-fitting. We used XAVIER as weight initializer in the model, which is a Gaussian distribution with mean 0. The function that learns the weight vector is called the optimizer function which is the stochastic gradient descent (SGD) in this experiment. The activation function allows the model to learn the complex data set. The activation function ReLU is used in all layers and negative log likelihood is used as the loss function.

Python is used for pre-processing and deeplearning4J, a java ma- chine learning package is used for model development. All models are executed on a CPU Intel core i7 with 16GB RAM. For training 75% of each cancer type is selected and the remaining 25% is used

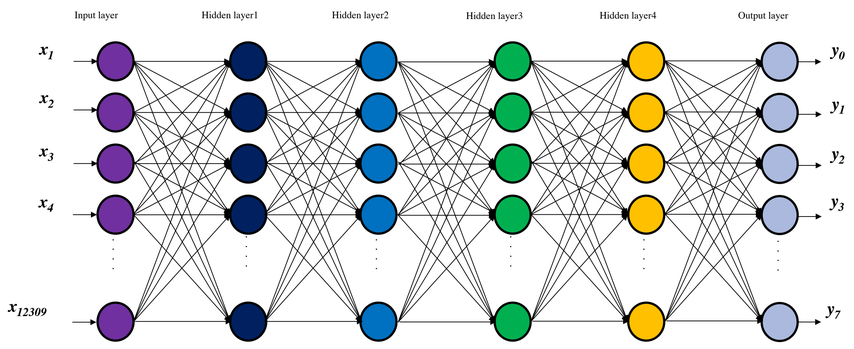

Figure 2: Deep Neural Architecture with Multiple Hidden Layers

Results

In this study, we developed deep neural network based classification models to classify 8 cancers - BLCA, CESC, COAD, HNSC, KIRP, LGG, LIHC, and LUAD - using four different algorithms - MLP, LSTM, CNN, and DAE. Fig. 3 shows one of the confusion matrices. Column labels represent the actual labels and row labels represent the predicted labels. For example, number of test samples for BLCA is 106 (430×25%). Out of 106, 101 samples are correctly identified as BLCA whereas 1, 2, 1, and 1 sample(s) is(are) incorrectly identified as HNSC, KIRP, LGG, and LUAD respectively.

Four different performance metrics - accuracy, precision, recall, and F1 score - are measured to compare the model performance. Table 1 shows the values of performance metrics for MLP, LSTM, CNN and DAE models. It is clear from the table that MLP has the lowest accuracy of 94% and CNN has the highest accuracy of 98%. Similarly, MLP has the lowest precision, recall, and F1 scores nearly 93% while CNN has the highest score, 98%, for all metrics. Clearly, CNN outperforms for this classification task. It is well-known that CNN or version of CNN beats other deep learning models in terms of image classification because CNN uses pixels of the image as features.

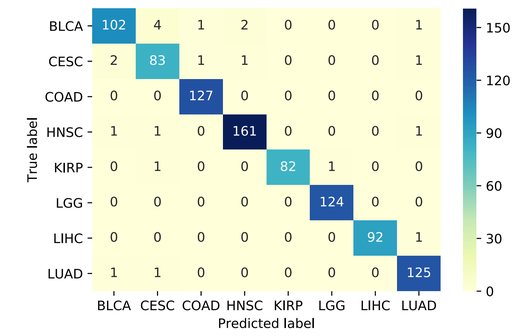

Figure 3: Confusion Matrix of CNN Model

In this experiment, lncRNA expression is nothing but an array and each cell contains the expression value of a lncRNA similar to the pixel of an image that leads to better performance for CNN model. Each execution has been repeated 10 times with an independent training/testing partition to get a stable result

Table 1: Performance Comparison

Model Name	Accuracy	Precision	Recall	F1
MLP	0.9371	0.9324	0.929	0.9394
LSTM	0.9562	0.9508	0.9521	0.9514
CNN	**0.9781**	**0.9765**	**0.9764**	**0.9764**
DAE	0.9639	0.9613	0.959	0.96

In comparison with more recent supervised classifier, GA/KNN model achieved upto 90% accuracy [10]. In contrast, Boyu et. al. converted cancer and normal samples into heatmap images and trained a CNN model [11]. As a result, the model is able to separate disease and normal samples with accuracy upto 95.59% whereas proposed models archived 97.81% accuracy in classification of multiple cancer types. The improved performance indicates that this classifier might assist in cancer diagnosis and identification system. Worth mentioning that models are not able to differentiate between case and normal samples because models are trained with diseases samples only. Therefore, if a normal sample will be tested, model might forcefully classify it to one of the cancer types which is a pitfall of supervised learning.

Conclusion

In conclusion, using lncRNA expression alone, proposed models are able to correctly classify 98% of samples from 8 different cancer types. The achieved accuracy is remarkable given the total volume of samples and the number of cancer types involved. Moreover, the accuracy of the proposed approach exhibits improvement compared to the recent studies on multiple cancer types classification. One of the significant application of this classifier is to assist in cancer diagnosis and identification system where unknown cancer sample needs to be identified to a specific cancer type. In the present study, expression of all lncRNAs are used. In future, we would like to identify most significant lncRNAs which can be considered as cancer biomarkers. Those lncRNAs will be used in next study to develop the cancer classification model. Also, normal samples will be trained in the next study to overcome the limitation of current study.

Acknowledgements

This research is funded by NSF CAREER award #1651917 (transferred to #1901628) to AMM.

References

1. SW Cheetham, F Gruhl, JS Mattick, and ME Dinger. Long noncoding rnas and the genetics of cancer. British journal of cancer, 108(12):2419, 2013.
2. Hui Tao, Jing-Jing Yang, Xiao Zhou, Zi-Yu Deng, Kai-Hu Shi, and Jun Li. Emerging role of long noncoding rnas in lung cancer: Current status and future prospects. Respiratory medicine, 110:12–19, 2016.
3. Yiwen Fang and Melissa J Fullwood. Roles, functions, and mechanisms of long non-coding rnas in cancer. Genomics, proteomics & bioinformatics, 14(1):42–54, 2016.
4. Y Kotake, T Nakagawa, K Kitagawa, S Suzuki, N Liu, M Kitagawa, and Y Xiong. Long non-coding rna anril is required for the prc2 recruitment to and silencing of p15p15 ink4b tumor suppressor gene. Oncogene, 30(16):1956, 2011.
5. Zhe Yang, Lin Zhou, Li-Ming Wu, Ming-Chun Lai, Hai-Yang Xie, Feng Zhang, and Shu-Sen Zheng. Overexpression of long non-coding rna hotair predicts tumor recurrence in hepatocellular carcinoma patients following liver transplantation. Annals of surgical oncology, 18(5):1243–1250, 2011.
6. Adam M Schmitt and Howard Y Chang. Long noncoding rnas in cancer pathways. Cancer cell, 29(4):452–463, 2016.
7. Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158(4):929–944, 2014.
8. Jorge S Reis-Filho and Lajos Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. The Lancet, 378(9805):1812–1823, 2011.
9. ArindamBhattacharjee,WilliamGRichards,JaneStaunton,ChengLi,Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences, 98(24):13790–13795, 2001.
10. Yuanyuan Li, Kai Kang, Juno M Krahn, Nicole Croutwater, Kevin Lee, David M Umbach, and Leping Li. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC genomics, 18(1):508,2017.
11. Boyu Lyu and Anamul Haque. Deep learning based tumor type classification using gene expression data. In Proceedings of the 2018 ACM International Con- ference on Bioinformatics, Computational Biology, and Health Informatics, pages 89–96. ACM, 2018.