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

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

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.

KEYWORDS

Cancer classification; lncRNA; RNA-seq; GDC-TCGA; CNN; DAE; LSTM; MLP

ACM Reference Format:
Abdullah Al Mamun and Ananda Mohan Mondal. 2019. Long Non-coding RNA Based Cancer Classification using Deep Neural Networks. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3307339.3343249

Mis-expression of lncRNAs can lead to change in expression profiles of various target genes related to cell homeostasis [1]. Cheetham et. al. identified different lncRNAs that are actively involved in different stages of cancer development [1].

Few studies developed models for classification of multiple cancer types. For example, Li et. al. developed GA/KNN to classify cancer and normal samples for 31 cancers using RNA-seq gene expression data [2].

In this paper, we aim to develop models for classifying multiple cancer types using lncRNA expression employing deep neural networks.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

ACM ISBN 978-1-4503-6666-3/19/09.
https://doi.org/10.1145/3307339.3343249

To validate the idea, we downloaded (April, 2019) RNA-seq FPKM normalized expression data for 8 cancers from UCSC xena [3].

The dataset has 12309 common lncRNA IDs with expression data for all selected cancers. The number of combined list of samples is 3656 after merging all the cancers.

For training 75% of each cancer type is selected randomly using seed 123 for random number generation. The remaining 25% is used for testing. This training and testing procedure has been repeated 10 times. The average of this 10 results are used as the performance of the model. Python is used for pre-processing and deeplearning4J, a java machine learning package is used for model development. All models are executed on a CPU Intel core i7 with 16GB RAM.

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.

Table 1: Performance Comparison

Model Name	Accuracy	Precision	Recall	F1
MLP	0.9371	0.9324	0.9290	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	0.9590	0.9600

Present study shows that lncRNA expression is a significant feature to differentiate multiple cancer types, which is evidenced by the performance of the proposed models that are able to correctly classify nearly 98% of the cancer samples.

ACKNOWLEDGMENT

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] 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.
[3] Mary Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Akhil Kamath, Fran McDade, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. The ucsc xena platform for cancer genomics data visualization and interpretation. *BioRxiv*, page 326470, 2019.