Deep Learning for Predicting Differentially Expressed Genes in Prostate Cancer

Healthcare, Generative Modeling

Zhongyi (James) Guo (SUID: guozy), Tiankai Yan (SUID: ytk2304), Luna Lyu (SUID: lunalyu)

Introduction

Prostate cancer (PCa) is the most commonly diagnosed cancer among men in 118 countries in 2022 (Bray *et al.*, 2024). Deep learning has revolutionized computational genomics by achieving exceptional performance in unraveling intricate patterns between gene expression and *cis*-regulatory elements, such as enhancers and promoters (Liu *et al.*, 2020; Zou *et al.*, 2019). How can we leverage deep learning to predict whether a given gene will be differentially expressed (DE) in primary prostate cancer?

In this study, we will: (1) identify DE genes as the ground truth; (2) train a binary Convolutional Neural Network (CNN) to predict the likelihood of a gene being DE in prostate cancer, based on the associated enhancer and promoter sequences; and (3) develop a Generative Adversarial Network (GAN) to generate synthetic enhancer and promoter sequences to augment the data, improving CNN performance.

Previous Work

Previously, many deep learning models have been developed to study prostate cancer, but most have focused on clinical diagnosis and disease progression (Wei *et al.*, 2022; Elmarakeby *et al.*, 2021; Ramírez-Mena *et al.*, 2023). One study developed an Artificial Neural Network to distinguish between prostate cancer and control samples using DE genes themselves instead of regulatory elements (Xie & Xie, 2024) that are the root cause of gene transcription.

Materials and Methods

Dataset

We obtained an RNA-seq count matrix from a Gene Expression Omnibus (GEO) case-control study GSE120741 (Stelloo *et al.*, 2018), which includes 39,376 genes from 48 primary PCa and 43 control samples collected from prostatectomy. For each gene, enhancer and promoter sequences will be matched using R packages *Ensembl* and *biomaRt*. We will then split the data into training, validation, and test sets.

Differential Gene Expression

We will use *PyDESeq2*, a Python package (Muzellec *et al.*, 2023), to identify DE genes among all genes.

Convolutional Neural Networks (CNN)

We will encode the training set to a matrix using one-hot encoding, then apply convolution with ReLU activation, followed by pooling, and finish with a fully connected layer using sigmoid activation, while using back propagation iteratively.

Generative Adversarial Network (GAN)

The goal of training the GAN is to generate synthetic enhancer and promoter sequences to enhance the performance of the CNN. During the training stage, we will feed the training set into the generator, and create dummy RNA-seq data for the discriminator using ChatGPT API. Once they have reached convergence, we will input generator-produced synthetic sequences into CNN, with the goal of improving CNN performance. The combination of CNN-GAN is the innovative point in our project.

Model Evaluation

To evaluate the CNN, we will use the accuracy and AUC-ROC by comparing predicted labels to ground truth labels in the test set.

To evaluate the GAN, we will compare the accuracy and AUC-ROC of CNN with and without GAN-generated sequences. An improvement in these metrics would suggest that the GAN performs well.

Possible Challenges

One potential challenge is that one gene can be regulated by multiple enhancers and promoters, each of which may vary in length. This variability could introduce complexity and confounding during the model training process.

References

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *74*(3), 229-263.

Liu, J., Li, J., Wang, H., & Yan, J. (2020). Application of deep learning in genomics. *Science China Life Sciences*, *63*, 1860-1878.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature genetics*, *51*(1), 12-18.

Wei, Z., Han, D., Zhang, C., Wang, S., Liu, J., Chao, F., ... & Chen, G. (2022). Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer. *Frontiers in oncology*, *12*, 893424.

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., ... & Van Allen, E. M. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature*, *598*(7880), 348-352.

Ramírez-Mena, A., Andrés-León, E., Alvarez-Cubero, M. J., Anguita-Ruiz, A., Martinez-Gonzalez, L. J., & Alcala-Fdez, J. (2023). Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression. *Computer Methods and Programs in Biomedicine*, *240*, 107719.

Xie, Y., & Xie, J. (2024). Integrates Differential Gene Expression analysis and deep learning for accurate and robust prostate cancer diagnosis. *Applied and Computational Engineering*, 57, 66-74.

Stelloo, S., Nevedomskaya, E., Kim, Y., Schuurman, K., Valle-Encinas, E., Lobo, J., ... & Zwart, W. (2018). Integrative epigenetic taxonomy of primary prostate cancer. *Nature communications*, *9*(1), 4900.

Muzellec, B., Teleńczuk, M., Cabeli, V., & Andreux, M. (2023). PyDESeq2: a python package for bulk RNA-seq differential expression analysis. *Bioinformatics*, *39*(9), btad547.