Generative AI for Predicting Gene Expression from cfDNA Fragmentation Patterns

Christina Espeseth Grand

Supervisor: Jakob Skou Pedersen

Department of Bioinformatics
Aarhus University

This thesis is submitted for the degree of
Master of Science in Bioinformatics

June 2025

Acknowledgements

I would like to express my gratitude to Professor Jakob Skou Pedersen for his invaluable guidance and support throughout this project. His expertise, insightful feedback, and encouragement have played a big role in shaping the direction of my research and in helping me navigate the various challenges and frustrations encountered along the way.

I am also deeply thankful to my co-supervisors, Farhad Zamani and Mathilde Diekema, for their support. Farhad's assistance, especially his insights and the use of code from his master's thesis, has been important in the development of this project. Mathilde's thoughtful advice, support, and willingness to help whenever needed have been essential to my progress.

Finally, I would like to thank everyone in the research group for good discussions and conversations. Their support has made my time here a truly positive and enriching experience.

Abstract

Motivation

Nucleosomes regulate gene expression by blocking DNA access to proteins important for transcription, requiring the displacement of nucleosomes at promoter regions for transcription to occur. Their positioning, especially around transcription start sites (TSSs), reflects gene activity. This positioning can be inferred from cell-free DNA (cfDNA) in blood, as nucleosomes protect DNA from cleavage during apoptosis. This protection results in nucleosome-bound DNA being more likely to appear in cfDNA, whereas DNA from actively transcribed genes, particularly around the TSS, are more exposed and thus underrepresented due to degradation. In cfDNA sequencing, this creates an inverse relationship between gene expression and the read depth around the TSS of the corresponding gene. This relationship suggests the possibility of modeling gene expression from cfDNA sequencing, offering potential for noninvasive disease detection based on expression changes.

This study primarily explores deep generative models to predict gene expression from cfDNA sequencing data. XGBoost is used as a baseline model for comparison. Two variants of a Deep Generative Decoder (DGD) are applied: a multimodal DGD and a newly developed factor multimodal DGD. The latter model explicitly separates the contributions of genes and samples to both cfDNA read depths and gene expression. These models aim to find the patterns linking the cfDNA signal to transcriptional activity. This approach can potentially enable noninvasive disease detection by leveraging differences in gene expression between diseased and normal cells.

Results

Analysis of the data suggested a potential relationship between cfDNA read depth patterns around the TSS and gene expression levels. However, the models struggled to effectively learn this pattern. One of the reasons for this could be significant noise within the data. A key factor contributing to the poor performance of the models is that the cfDNA sequencing data and gene expression data were obtained from different individuals, which means that the gene expression associated with the cfDNA pattern may not always be correct, due to variance between patients.

Table of contents

Abbreviations

cfDNA cell-free DNA.

ctDNA circulating tumor DNA.

CV Coefficient of Variation.

DGD Deep Generative Decoder.

FFT Fast Fourier Transformation.

FPKM Fragments Per Kilobase of transcript per Million mapped reads.

FPM Fragments Per Million.

GMM Gaussian Mixture Model.

mRNA messenger RNA.

NDR nucleosome-depleted region.

NFR nucleosome-free region.

PBMC Peripheral Blood Mononuclear Cell.

PFE promoter fragmentation entropy.

PIC pre-initiation complex.

RNA-seq RNA-sequencing.

rRNA ribosomal RNA.

SGD stochastic gradient descent.

SVM support vector machine.

TER Transcription End Region.

TF transcription factor.

TSS Transcription Start Site.

VAE Variational Autoencoder.

WGS Whole Genome Sequencing.

WPS Window Protection Score.

XGBoost Extreme Gradient Boosting.