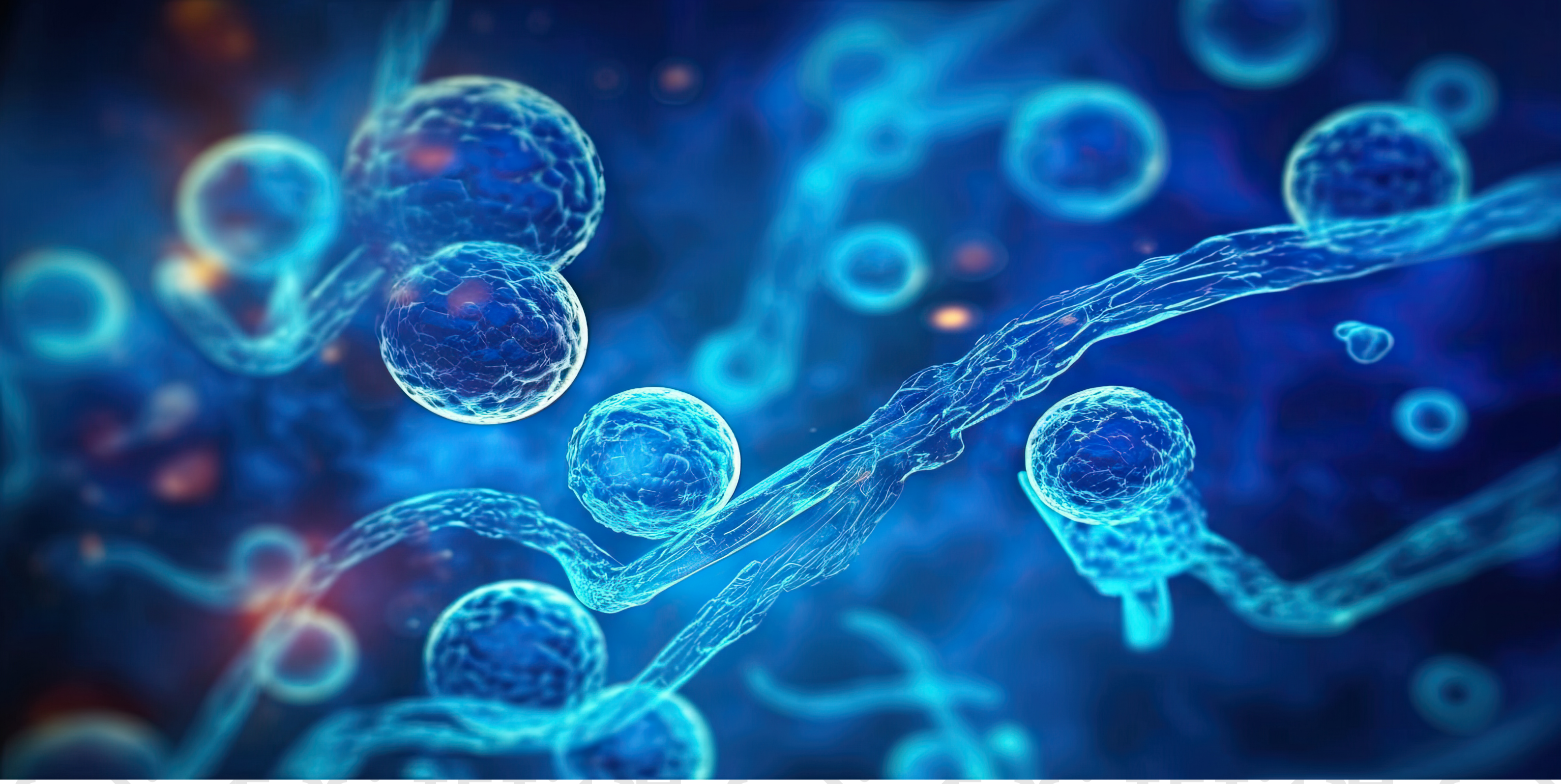

miDGD: Prediction of miRNA activity levels based on gene expression data using Deep Generative Decoder model

Master's in Bioinformatics Thesis

Farhad Zamani

Supervisor:

Jakob Skou Pedersen

Spring 2024

AARHUS UNIVERSITY

miDGD: Prediction of miRNA activity levels based on gene expression data using Deep Generative Decoder model

Master's in Bioinformatics Thesis

Farhad Zamani
Supervisor: Jakob Skou Pedersen

Spring 2024

AARHUS UNIVERSITY

Aarhus University
http://www.au.dk

Thesis title:
miDGD: Prediction of miRNA activity levels based on gene expression data using Deep Generative Decoder model

Thesis Period:
Spring Semester 2024

Author:
Farhad Zamani

Supervisor:
Jakob Skou Pedersen

Page Numbers: 64

Date of Completion:
June 15, 2024

Master's of Science in Bioinformatics
Bioinformatics Research Center (BiRC)
Aarhus University
Denmark

In collaboration with:
Department of Molecular Medicine (MOMA)
Aarhus University Hospital
Denmark

Abstract

Motivation

microRNAs play an important role in regulating gene expression at a posttranscriptional level. In cancer cells, miRNAs are often dysregulated, acting as either oncomiRs or tumor suppressors. Therefore, understanding miRNA regulation is important in cancer research and potentially elsewhere. However, there are limitations to studying miRNAs in single-cell RNA sequencing (scRNA-seq) settings. The state-of-the-art to infer miRNA expression levels involves several computational approaches that leverage the relationship between miRNAs and their target mRNAs, including motif enrichment analysis and machine learning models like XGBoost. This thesis aims to explore and evaluate the ability of generative AI approaches to improve the prediction of miRNA expression levels from gene expression data in bulk RNA-seq and scRNA-seq settings.

Results

We present the miDGD, a Deep Generative Decoder (DGD) model that can infer miRNA activity levels based on only gene expression data. The miDGD model learns the shared representation of gene and miRNA expression and handles complex parameterized latent distributions. The result shows that miDGD model can be used to predict miRNA expression in bulk RNA-seq and sparse data equivalent to scRNA-seq experiments.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to everyone who has supported me throughout the completion of this thesis.

I am deeply grateful to my supervisor, Professor Jakob Skou Pedersen, for your invaluable guidance, continuous support, and patience during my research. Your immense knowledge and experience have been influential in shaping this thesis. Your enthusiasm for the research is truly inspiring.

I would also like to thank the members of the Skou Pedersen Group for the insightful discussions and help through the thesis, especially Mathilde Diekema for your valuable input and for reviewing my thesis, and Asta Mannstaedt Rasmussen for the assistance in navigating and working with the dataset. Your contributions have significantly improved the quality of my work.

Special thanks to the Bioinformatics Research Center (BiRC) and the Molecular Medicine Department (MOMA) at Aarhus University for providing a stimulating environment. I am particularly thankful to my fellow Master's in Bioinformatics students for their moral support and the fun times we shared.

I would like to extend my heartfelt thanks to my family for their unwavering support and encouragement. To my parents, thank you for your endless love and for believing in me. I am also grateful to my friends in PPI Denmark for their support and for making this journey enjoyable.

Abbreviations

ACC Adrenocortical Carcinoma.

BLCA Bladder Urothelial Carcinoma.

BRCA Breast Invasive Carcinoma.

CESC Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma.

CHOL Cholangiocarcinoma.

COAD Colon Adenocarcinoma.

DGD Deep Generative Decoder.

DLBC Lymphoid Neoplasm Diffuse Large B-cell Lymphoma.

ESCA Esophageal Carcinoma.

GMM Gaussian Mixture Model.

HNSC Head and Neck Squamous Cell Carcinoma.

KICH Kidney Chromophobe.

KIRC Kidney Renal Clear Cell Carcinoma.

KIRP Kidney Renal Papillary Cell Carcinoma.

LAML Acute Myeloid Leukemia.

LGG Brain Lower Grade Glioma.

LIHC Liver Hepatocellular Carcinoma.

LUAD Lung Adenocarcinoma.

LUSC Lung Squamous Cell Carcinoma.

MESO Mesothelioma.

OV Ovarian Serous Cystadenocarcinoma.

PAAD Pancreatic Adenocarcinoma.

PCPG Pheochromocytoma and Paraganglioma.

PRAD Prostate Adenocarcinoma.

READ Rectum Adenocarcinoma.

SARC Sarcoma.

scRNA-seq Single-cell RNA sequencing.

SKCM Skin Cutaneous Melanoma.

STAD Stomach Adenocarcinoma.

TCGA The Cancer Genome Atlas.

TGCT Testicular Germ Cell Tumors.

THCA Thyroid Carcinoma.

THYM Thymoma.

UCEC Uterine Corpus Endometrial Carcinoma.

UCS Uterine Carcinosarcoma.

UTR Untranslated Region.

UVM Uveal Melanoma.

Contents