Spring 5-22-2021

# Using an integrative machine learning approach to study microRNA regulation networks in pancreatic cancer progression

Roland Madadjim
*University of Nebraska-Lincoln*, rmadadjim2@huskers.unl.edu

# USING AN INTEGRATIVE MACHINE LEARNING APPROACH TO STUDY MICRORNA REGULATION NETWORKS IN PANCREATIC CANCER PROGRESSION

by

Roland Madadjim

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Juan Cui

Lincoln, Nebraska

May, 2021

# USING AN INTEGRATIVE MACHINE LEARNING APPROACH TO STUDY MICRORNA REGULATION NETWORKS IN PANCREATIC CANCER PROGRESSION

Roland Madadjim, MS

University of Nebraska, 2021

Adviser: Juan Cui

With advances in genomic discovery tools, recent biomedical research has produced a massive amount of genomic data on post-transcriptional regulations related to various transcript factors, microRNAs, lncRNAs, epigenetic modifications, and genetic variations. In this direction, the field of gene regulation network inference is created and aims to understand the interactome regulations between these molecules (e.g., gene-gene, miRNA-gene) that take place to build models able to capture behavioral changes in biological systems. A question of interest arises in integrating such molecules to build a network while treating each specie in its uniqueness. Given the dynamic changes of interactome in chaotic systems (e.g., cancers) and the dramatic growth of heterogeneous data on this topic, building scalable models is crucial. Indeed, recovering a model that can capture the relationships within this data constitutes a major challenge. This thesis addresses this challenge by using an integrative network learning model based on gene expression data to elucidate miRNA – gene interactions in cancer progression. First, we present a preprocessing pipeline for miRNA-gene interactions based on De Novo approach. Second, we introduce a machine learning approach for data integration of multiple data types such microarray, RNA-seq and CLIP based miRNA-RNA

interactions along with graphical model fusion. Last, we show how the latter enabled transforming static interactions into semi-conditional ones. In a case study of human pancreatic cancer, we have identified gene regulatory networks distinctly associated with four progressive stages with a list of 12 miRNA-gene conditional interactions; The functional analysis with focus on microRNA-mediated dysregulation revealed significant changes in major cancer hallmarks. The identified novel pathological signaling and metabolic processes shed light on the regulatory roles that microRNAs play in pancreatic cancer progression. We believe this integrative model can be a robust and effective discovery tool to further the understanding of key regulatory characteristics in complex biological systems.

# DEDICATION

To God Almighty,

To Cheryl and Ariella.

# ACKNOWLEDGMENTS

I would like to express my appreciation and sincere thanks to my advisor Dr. Juan Cui for her guidance during the choice of my master thesis' subject, her precious advice that helped me accomplish this project.

I would like to thank Dr. Jitender Deogun and Dr. Massimiliano Pierobon for being on my master committee. I want acknowledge my colleagues in SBBI lab, especially Haluk Dogan and Zeynep Hakguder for their effort and support throughtout my master program. Thanks to all those who support me from far or near, Lizzy Isaacson, Dr. Bonodji Nako's family, Erika Hepburn, Dan Hutt, Veronica Riepe, Minal Khatri, Salome Maria. Last but not the less, to my family, my wife and my child for all their sacrifices and prayer while I was away from them.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Overview

In the post-genomic era in which we are, one of the challenging task for researchers is how to decipher genes regulations. Gene regulation presents a set of mechanisms used by cell to control gene expressions' products, these products are RNA or proteins. Understanding of such regulations enables discovery of new drug targets for disease treatment, which remains challenging as the dynamic nature of miRNA interactions, as well as the evolving functions, are largely undetermined. However determining those interactions under the context of a multi-layer regulation network leads to challenges in information fusion and network fusion. Information fusion, which refers to an effective integration of heterogeneous data analyses that reflect distinct regulatory mechanisms. For example, each type of high-throughput data such as microarray or RNA-seq based expression profiles, CLIP or CLASH based miRNA-RNA interactions, and ChIP-seq TF binding profiles, as well as methylation and genetic profiles from DNA sequencing analysis can be used to infer a certain type of molecular interaction. Molecular interactions such as microRNA-gene regulations are important to understand the pathological mechanism of human

cancers. Network fusion, the second identified challenge lies in the integration of heterogeneous interaction networks inferred from different models without losing proper causality inference among the interactions.

## 1.2 Motivation

In system biology, GRN reconstruction is of great interest as it enables the discovery of genetic mechanisms driving diverse diseases, such as cancer [51]. These mechanisms are defined by interactions among molecules that made up the organism. Hence, relationships between such molecules can be summarized as a network, where nodes are molecules and the links between the nodes are interactions of interest. Several network models have been built that captured biology interactions of genes, but very few have succeeded to capture miRNA-gene interactions in an integrative way. Many of the conventional methods used in genomic data analysis do not take advantage of the interactions between multiple factors, as in the case of miRNA and mRNA they regulate, failing thereby to uncover the cellular processes that are unique to specific tissues [29]. We have identified two major challenges, information fusion and network fusion. First, to address the information fusion challenge, we explored the De Novo approach for miRNA-gene interactions based on CLIP data and graphical lasso network method for building network from microarray and RNA-seq based expression profiles data. Effective information fusion can transform a static interaction analysis into a semi-conditional manner, leading to more practically useful results. Secondly, to tackle the network fusion challenge, we proposed a more generalized Bayesian Network framework for model integration while keeping the proper causality. The models used so far

to predict miRNA-gene interactions have proven to be limited. A network that capture gene-gene or miRNA-gene interactions based on a unique data source will not provide the full mechanism that underlies miRNA-gene interactions for a given tissue or specific disease. Those limitations vary from the unsupervised nature of the model to the model's assumptions that do not always hold [56]. With this in mind, it is crucial to develop computational model that will exploit the full potential of high-throughput data to discover the full potential of miRNA impact on cancer associated genes regulation. GRN is generally represented as a graph where the nodes are the genes and the edges their interactions. This graph's edges can be directed or undirected. A directed graph will provide causality in the network, hence the GRN inference. Our problem is how to capture these causality in a meaningful way, how to find the true causal links in the network, without biased (without limiting the types of available interactions i.e., gene-gene, miRNA-gene) based on the participating genes. Individual interactions do not provide a full picture of these biological entities (this has as consequence more false positive than true positive) where the need to explore additional interactions, specifically gene-gene interactions, will expands the benefits of a GRN with regards to functional analysis.

## 1.3 Contribution

The aim of the study is then to design a more generalized network framework for model integration while keeping the proper causality which will identify with high confidence a set of miRNA-gene interactions through an information fusion scheme that will infer regulatory mechanism underlying cancer progression. Knowledge of the functional miRNAs-gene interactions can help

find the driving factor (source or reason) of a genetic disease, which can provide more insights to biologist and researchers in their quest for targeted and efficient care and treatment. Taking advantage of the prior knowledge, we proposed a graphical model using Bayesian Network to predict interactions between miRNA-gene and then using those to update the GRN and discover more true causal structures. To demonstrate the proposed techniques, Pancreatic cancer will be used as case study where miRNA-gene will hep uncover cancer progression and functional analysis.

## 1.4 Thesis outline

The next chapter provides the background required for network reconstruction and related works using Regression, Bayesian and deep learning methods in gene network reconstruction. Chapter 3 describes the cancer data used in this study and provides an exploratory data analysis. Detailed explanation of the datasets is provided along with the preprocessing techniques we use to narrow down the number of genes and to identify the interactions of interest. Chapter 4 introduces a De Novo approach to identify miRNA-gene interactions and a Bayesian framework for GRN reconstruction along with a binding network for network fusion. Chapter 5 presents and discuss the results. Finally, Chapter 6 presents the conclusion and future of this thesis.

# Chapter 2

# Background and Related Work

## 2.1 Background

### 2.1.1 Genes and microRNA

**Genes** are regions of DNA that encodes functional RNAs or proteins and they are the molecular units of heredity. RNA molecule folds into a "unique" tertiary structure to carry a particular set of biochemical functions. The Human Genome Project estimated that humans have about 21,000 genes. Messenger RNA (mRNA) is copy of RNA obtained after transcription. mRNA is a single-stranded RNA molecule of RNA that is complementary to one of the DNA strands of a gene and is read by a ribosome in the process of synthesizing a protein. Changes in mRNA expression are key in cancer onset and progression. Whereas **MicroRNAs** are a class of single-stranded endogenous non-coding RNAs that are about 19 to 25 nucleotides (nt) in length [6]. They are used by cell to control gene expression mainly by binding with mRNA. One miRNA can simultaneously bind to various target mRNAs, long non-coding RNAs, and circular RNAs and meanwhile, one gene can be regulated by multiple miRNAs. MicroRNAs do not bind to mRNAs on their own but rather function as a component of the RNA-induced silencing complexes (RISCs). Mature miRNAs are

mostly known to repress gene expression at post-transcriptional levels by binding to the 30 untranslated region (UTR) of the target mRNA transcripts [6]. There are evidences that miRNA expression is dysregulated in cancer through various mechanisms, including amplification or deletion of miRNA genes, abnormal transcriptional control of miRNAs, dysregulated epigenetic changes and defects in the miRNA biogenesis machinery. miRNAs influence numerous cancer-related cellular processes such as cell proliferation, cell cycle control, apoptosis and metabolism [13] via its own dysregulation. Their dysregulation confer malignant cells their tumorigenic potential. Understanding the impact of such regulation is of highest importance but miRNA interactions and functions are still largely unknown. A miRNA can regulate one to many genes while several miRNA can target and regulate the same gene.

### 2.1.2 Gene regulation networks (GRNs)

Gene regulation network is a directed graph representing all the molecules interacting jointly to control genes expression. It is made up of regulators such Transcription factors (TFs), microRNA (miRNA) which bind to the promotor regions of their target genes and act on them as activators or inhibitors. These directed edges of the GRN enable clear identification of regulators and regulated nodes which explain the direction of causality in the network. Reconstruction and understanding of such network, GRN, has great potential in understanding diseases initiation and progression. Reconstructing GRN therefore is required to understand how gene expression dysregulation contributes to cancer and other complex heritable diseases [5].

## 2.2 Literature review on network inference

Many algorithms have been developed to infer the GRNs from unsupervised to supervised methods, from model-based to probabilistic methods. Here we present regression based, probabilistic-Bayesian Network and deep learning approaches. Note that not all of these studies that will follow are directly related to miRNA-gene interaction. However, the proposed methods are in line with the approaches under scrutiny in this study.

### 2.2.1 Data fusion and integration

In the high-throughput biomolecular data context, data integration is typically performed in four different manners. One is to analyze each data type separately first and then integrate the final findings. Another manner is to preprocess each type of data independently, then perform cross-platform normalization across the data types, then combine the normalized figures and finally perform an overall analysis. The third type of integration consists of performing a statistical integration. The fourth approach is to integrate the data by modeling the data types based on the biological meaning of the molecules and their interactions [21] As reported in [21, 50], data integration is done in vertical and horizontal direction. In this work, we will follow the vertical direction, ie the vertical multi-omics analysis which is performed within the cohort as opposed to horizontal, out of scope, which is a cross-cohort data integration. In line with data integration, Glass et al. proposed **PANDA** [16] an integrative method based on information sharing between different data sources by aggregating such information to build a coherent regulatory network. Their method is based on similarity. They defined the notion of agreement and availability

which enable reconstruction of genome-wide, condition-specific regulatory networks by weighing and integrating such data in a manner which first checks the availability of a target gene to be regulated by a TF against its likely the responsibility a TF is measured to have in regulating that gene.

### 2.2.2 Probabilistic methods

Bayesian based methods [34] when dealing with small sample size, Bayesian networks are preferred as they proved to be less influenced by biological noise, BN is based on probability theory. They proved to be make accurate prediction in face incomplete data. BN is one of the first established methods for integrating prior knowledge which encodes biological information as a prior distribution over graph structures [24] (Imoto et al., 2003). Using a heuristic greedy optimization, both the hyperparameters and network structure are inferred by maximizing the joint posterior distribution. This framework has been applied to a wide variety of priors in conjunction with gene expression data.

### 2.2.3 Machine and deep learning based methods

Lyu et al. [39] in their work on tumor type classification using GE data, presented a DL approach that used Convolutional Neural Networks (CNN) at its core. They proposed an embedding of high dimensional RNA-seq data into a 2-D image. To learn tumor-specific gene, they combined a large set of genes from various tumors from the PANCAN for training the model, using this knowledge from multiple tumor types they were able identified tumor-specific gene. They achieved an accuracy of more than 95% on the PANCAN datasets. One major take-away in their work was the the image embedding which tackled

perfectly the challenges imposed by high dimensional genomic data. Yuan et al. [56] also offered an embedding of GE into 2-D image following a NEPDF as input to a CNN model. Their CNN allowed concatenation of additional data type at FC layer. In [40] the authors proposed 3 different implementations of their CNN models based on 1D (vector input), 2D (matrix input) and 2D and again high accuracy was achieved here as well (93-95%). The authors took a completely novel/different approach on their CNN architecture by implementing only 1 convolution layer, suggesting shallower models suitable for cancer type prediction as such models help addressed the curse of dimensionality problem. **CNNC** offered a Deep Learning approach where expression data are transformed in an image-like input and fed into a CNN architecture to learn interaction between pair of gene.

### 2.2.4 Regression based methods

Another important category of GRN inference methods is based on regression methods, which are used to predict one target gene based on one or more input genes. These methods are amongst the most popular and scalable approaches for reconstructing directed networks [23]. Regression based methods enable learning high-order conditional dependencies between genes expressions. TargetScan [2] based on multiple linear regression, predicts biological targets of miRNAs by focusing on canonical binding sites within 3 -UTR regions and performs species-specific prediction. It uses stepwise variable selection based on Akaike information criterion. Lu et al. [38] proposed a Lasso regression based on Targetscan [2] for target prediction, and their proposed method offered reliable miRNA-mrna interactions based on combined sequence-based prediction, co-regulation and RISC availability and expression data. Another example of

method using linear regression is TIGRESS [19], GRN inference is formulated as a sparse linear regression problem where regulators of each target gene are learned by combining feature selection with least-angle regression with stability selection [48]. Jacobsen et al. [25] use multivariate linear regression model in miRNA-mRNA association to account and deal with noise introduced as a result of gene alterations. miRanda [7] on the other hand uses support vector regression and assesses the thermodynamic folding energy mirna:utr duplex.

## 2.3 Graph theory basics for GRN

Graph theory is the branch of mathematics that studies graphs, and networks are often referred to as graphs. In this section we will present a brief review of the most important terminologies needed to understand GRN. Mathematically speaking a graph, network, is an ordered pair $G = (V, E)$ where V is a finite set of vertices or nodes and E a set edges. Edges in a graph are said to be adjacent if they have a common node, in the same way vertices are said to be adjacent if they shared common edges. A particular vertex can be connected with more than one vertex, the number of edges that end (or start) at this particular edge is called the degree of the vertex. One major property of graph network, essential in GRN is the degree distribution. First, the degree of a node is defined as the number of edges it has with other nodes. In the case of a causal relationship i.e. a directed graph therefore we identify two types of node's degree: the in-degree (number of incoming edges) and the out-degree (the number of outgoing edges). Knowing the degree of each nodes give the degree distribution, which is defined as the probability of these degrees across all nodes in the network, denoted $P_{deg(k)}$. It is the fraction of nodes in the

graph with in-degree and out-degree. This property, the degree distribution of a graph enables us to learn network's structure (such as the presence of hubs) and to have a clear distinction between the types of network. It decays as a power law for many real networks [5], which is the defining property of scale-free networks. A scale-free network is a signature for a GRN. The topology of scale-free networks is dominated by a few highly connected nodes (hubs) which link the rest of the less connected nodes to the system. This property gives the network a great tolerance against errors, the nodes are able to communicate even with very high failure rates.

## 2.4 GRN performance evaluation

Several prediction programs and algorithms exist to predict miRNA-target interactions. Since they do not have the same approaches and use different principles in their implementations, they can not get the exact same results and conclusions. For example, one program will predict a miRNA-target interaction to be functional whereas a second program will conclude that this same interaction is not. Hence, how can the inferred network be assessed? Networks are compared by means of their edges, presence or absence of edge in both network: inferred and actual regulatory network. GRN inference can therefore be classified as a binary classification problem, i.e. each inferred interaction will either be classified as true or not, and it can be actually true or not. From those inferred interactions we can derive a confusion matrix. As opposed to standard interpretation, the meaning of the confusion matrix's metrics (TP, TN, FP and FN) are as follow:

- TP: True Positive is the number of predicted miRNA-target interactions

that do actually exist. These are are edges occurring in the reconstructed network, and that also occur in the gold standard network.

- TN:True Negative refer to edges that neither belong to the inferred network nor the actual network.
- FP: False Positive is the number of miRNA-target interactions occurring in the inferred network, but that do not appear in the actual network.
- FN: False Negative is the number of miRNA-target interactions that exist in the actual network but missing in the predicted network.

From these metrics, we can derive three statistical metrics essential in assessing GRN performance and these are:

- Precision is the proportion of interactions predicted as positives that are actual positives.

$$Precision = \frac{TP}{TP+FP} \tag{2.1}$$

- Sensitiviy, Recall or True Positive Rate (TPR): the proportion of actual positive interactions found

$$Sensitivity = \frac{TP}{TP+FN} \tag{2.2}$$

- Specificity, which can be seen as the ratio between the correctly non-predicted interactions and the number of total non-existing interactions (correctly non-predicted and incorrectly predicted). Thus defined as:

$$Specificity = \frac{TN}{TN+FP} \tag{2.3}$$

With the above statistical metrics, we can graphically explore the trade-off between these metrics and derive two important curves: AUPRC and AUROC. AUPRC, the area under the Precision-Recall curve plots the $Precision$ against the $Recall$. This measure is more suitable for imbalanced classes such in cancer data and it is commonly used in GRN inference. AUROC, on the other side display $Sensitivity$ as a function of $1 - Specificity$ also called False Positive Rate. The above described metrics and measurements will be used in the this study for performances analysis.

# Chapter 3

# Genomics Data on Pancreatic Cancer

In this chapter, data sources and data preprocessing is explored and data exploratory analysis is performed with outputs needed for downstream analysis. Figure 3.1 presents the workflow of the data preprocessing. In line with one of the challenges reported in this study, effective information fusion.

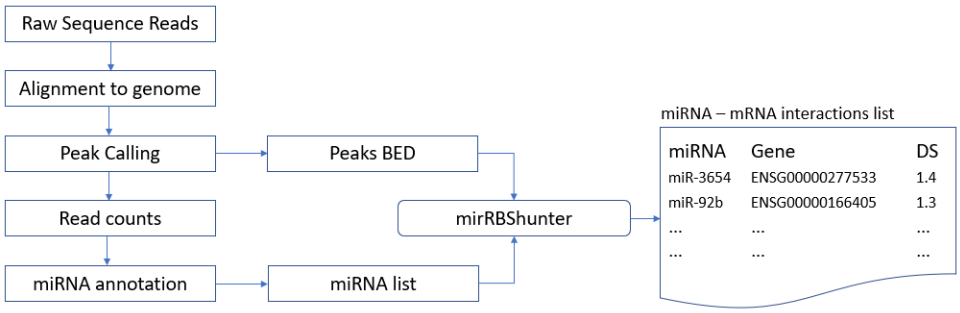



Figure 3.1: Workflow for miRNA-gene interaction data preprocessing

## 3.1 miRNA-mRNA interations detected through sequencing

MicroRNA targetome is known to use a seed sequence of 6–8 nt in their 5′ end (positions 2 to 7) to predominantly bind to either the 3′ UTR or the coding sequence (CDS) of mRNAs [41] [8]. This binding approach is defined as canonical

miRNA-binding site. However, it has been experimentally found that about 60% of miRNA binding activity is non-canonical, which involves other portions of miRNA sequence outside the seed or with seed-like motifs including mismatches or bulges [8]. Therefore, to effectively capture miRNA target mRNA interaction we use CLIP-seq data. CLIP-seq is a relatively new experimental technique to study the specificity of the binding activity for RNA-Binding Proteins (RBP). This technique provides a comprehensive and genome-wide map of the direct RNA binding sites for RBP. The application of CLIP-seq to Ago2 has been used to identify the miRNA-binding sites. The datasets are obtained from GEO under accession numbers SRP034075, these data contained two pancreatic cell lines (MIAPACA, HPNE) as reported in [10]. Raw sequence data were downloaded using SRAtoolit, an efficient module to download high-volume data from NCBI. The preprocessing was carried out following the workflow in Figure 3.1 above. To align the reads to a genome, Bowtie2 [30] aligner, an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences was used along with human genemone hg19 to locate the reads locations in the genome. The output of this aligner is a SAM (Sequence Alignment/Map) file. SAMtools [35] was used to convert SAM file into its compressed binary version BAM (Binary Alignment/Mapping), Samtools is a set of powerful utilities for interacting with and post-processing short DNA sequence read alignments in the SAM formats. An important step in this miRNA-RNA data interaction preprocessing is the peak calling task. This is a computational method used to identify areas in the genome that have been enriched with aligned reads as a result of performing ChIP-sequencing, CLIP-seq experiments among many other. Pyicoclip from Pyicoteo [3] library, a suite of tools for the analysis of high-throughput sequencing data, was used because it

| **Interaction List (SRX893318/SRX893321 data from GEO)** | | | | |
|---|---|---|---|---|
| hTERT-HPNE CLIP cell line (Normal) | | | | |
| #miRNA: | 20 | | #Genes: | 350 |
| Total Interaction #: | | | 436 | |
| MIA PACA-2 CLIP cell line (Carcinoma) | | | | |
| #miRNA: | 17 | | #Genes: | 285 |
| Total Interaction #: | | | 414 | |

Table 3.1: Cell line data summary

is suitably designed for CLIP-seq peaks. Pyicoclip was performed with p-value of 0.001, value used in [8]. The output was formatted into BED6 format. To get the reads count, featureCounts [37] from Subreads package was used. Reads counts provide an overall summary of the coverage for the genomics features of interest. In the output obtained from the previous steps, genomics coordinates were available, these were used to perform miRNA annotations with the help of intersect function from bedtools [42]. The annotation and miRNA sequence files were retrieved from miRBase [18, 27] and liftover [28] was performed on the annotatin file to match the assembly version in use (hg19). The resulting files were fed in miRBShunter [8] pipeline to identify interactions from both canonical and non-canonical binding site. This pipeline follows the de novo's approach for the identification of enriched motifs from Ago2 CLIP-seq peaks and it computes the calculate the miRNA::RNA heteroduplex score for the identified interactions. The output at this stage constitutes one input to our pipeline. Table 3.1 provides a summary of the data.

## 3.2 Gene expression data

The normal dataset (Normal solid tissues genes expression) for our study was obtained from GEO ( `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi`) under GSE28735 [57], a study of microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues. After a log transformation of the expressions, limma [43] was used to identify differentially expressed genes.

TCGA PAAD cohort data, RNA-seq and miRNA-seq were downloaded using GDCRNATools [36], a package that enable the preprocessing of the data and analysis of genes differential expression. For this cohort 183 samples were downloaded for RNA-seq along with their miRNA-seq, common samples data were extracted and to make up the tumor samples set 174 samples were kept which involves 21 samples in stage 1, 146 samples in stage 2, 3 samples in stage 3 and 4 samples in stage 4. To complete the preprocessing, TMM [45] normalization, an essential step in an RNA-seq analysis, in which the read count matrix is transformed to allow for meaningful comparison of counts across samples, it is required in that the proportion of mRNA corresponding to a given gene may change across biological conditions [14] and Voom [32] transformation needed to scale our data from raw counts onto a scale that accounts for library size difference as libraries sequenced at a greater depth will result in higher counts were performed. Processed data were fed into DEAnalysis method of GDCRNATools for Differential expression analysis using edgeR [44] method from Robinson. A list of 448 differentially expressed genes (number of up and down reg) were retained to match common genes set in both normal and tumor dataset. A cut off threshold of $|log2\ Fold\ Change\ (FC)| > 1$ and

*adjusted* $P - value < 0.05$ were considered for differential gene identification. The results of the preprocessing is summarized in table 3.2.

| **Normal Tissue (GSE28735 from GEO)** | | | | |
|---|---|---|---|---|
| #Samples: | 45 | | #Genes: | 448 |
| **Tumor Tissue (PAAD from TCGA)** | | | | |
| #Samples: | 174 | | #Genes: | 448 |
| Stage 1:<br>Stage 2:<br>Stage 3:<br>Stage 4: | | 21 Samples<br>146 Samples<br>3 Samples<br>4 Samples | | |
| **DE-miRNA (PAAD from TCGA)** | | | | |
| #Samples (T+N): | 178 | | #miRNA: | 26 |

Table 3.2: TCGA PAAD data summary

# Chapter 4

# Methodologies

To reconstruct this network, a graphical lasso approach was used to recover the structure of the network from the data.

## 4.1 Gene regulation network reconstruction

GRN is an abstraction to explain regulatory mechanisms behind gene expression. Our workflow is depicted in Figure 4.1. First, a Gaussian graphical

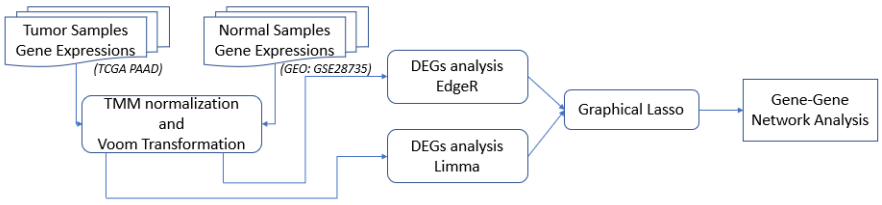



Figure 4.1: Worklow for gene-gene network construction

model (GGM) was explored to explain the dependency relationship between genes, the variables in a continuous multivariate system. In order to learn the underlying relationships embedded under complex GRN, we assume that our data is sampled from the following multivariate Gaussian distribution:

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right) \tag{4.1}$$

where $\mu$ is the mean vector and $\Sigma$ is the covariance matrix. $\Sigma$ is a square positive definite matrix and $\Omega = \Sigma^{-1}$ is called precision matrix. We can rewrite the formula in Equation 4.1 for $\mu = 0$ and $\Omega$ as shown in Equation 4.2.

$$p\left(x_1, x_2, \ldots, x_n \mid \mu = 0, \Omega\right) = \frac{|\Omega|^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_i \omega_{ii}\left(x_i\right)^2 - \sum_{i<j} \omega_{ij} x_i x_j\right) \tag{4.2}$$

Equation 4.2 can be considered as a continuous Markov Random Field (MRF) with potentials defined on every node and edge where $\omega_{ii}(x_i)^2$ is a node potential denoted as $\phi(x_i)$, and $\omega_{ij}x_i x_j$ is an edge potential denoted as $\phi(x_i, x_j)$. Given $n \times p$ data matrix $X$ where $n$ is the number samples, $p$ is the number of genes, and observations $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d) and sampled from $\mathcal{N}(\mu, \Sigma)$ where $\Sigma$ is $p \times p$ positive definite matrix. Two variables $p_i$ and $p_j$ are conditionally independent if and only if $\Omega[i, j] = 0$ [31]. The problem for learning the conditional independence relationship between the variables with the given data becomes now estimating the coefficients $\omega_{ii}$ and $\omega_{ij}$ shown in Equation 4.2. The scaled log-likelihood of a sample $x \in \mathbb{R}^p$ in a Gaussian graphical model with mean $\mu$ and precision matrix $\Omega$ is, up to a constant given by:

$$\mathcal{L}(\Omega, x) \equiv \log\det(\Omega) - (x - \mu)^\top \Omega (x - \mu) \tag{4.3}$$

We define the average scaled log-likelihood of $N$ samples $x^{(1)}, \ldots, x^{(n)}$ which depends only on sample covariance matrix $\widehat{\Sigma}$ by:

$$\begin{aligned} \mathcal{L}(\Omega, \widehat{\Sigma}) &\equiv \frac{1}{N}\sum_n \mathcal{L}\left(\Omega, \mathrm{x}^{(n)}\right) \\ &= \log\det(\Omega) - \mathrm{tr}(\widehat{\Sigma}\Omega) \end{aligned} \tag{4.4}$$

We impose some sparsity assumptions in our learning problem as sparse network are common in real-work applications. The reasons for such assumptions are: (1) biological networks are often sparse [20]; (2) computations on dense graphs require huge amount of resources; (3) dense graphs are difficult to interpret. Banerjee et al., [4] showed that finding the sparse precision matrix which fits most to a dataset is an NP-hard problem. Additionally, $p \times p$ covariance matrix $\Sigma$ requires $O(p^2)$ parameters for accurate estimation, however, we often have $n \ll p$. Therefore, some form of regularization can be used to make the computation tractable. Structured sparsity can be obtained by regularizing with $\ell_1$-norm. Our goal is to solve the following regularized maximum likelihood problem by minimizing regularized minus log-likelihood:

$$\min_{\Omega > 0} \mathcal{L}(\Omega) := \text{tr}(\widehat{\Sigma}\Omega) - \log\det(\Omega) + \lambda\|\Omega\|_1 \tag{4.5}$$

Equation 4.5 is a convex optimization problem where regularization parameter $\lambda > 0$, and linear term $(\text{tr}(\widehat{\Sigma}\Omega))$, the negative log determinant function $(\log\det(\Omega))$, the $\ell_1$ penalty, and the set of all positive definite matrices are convex. The solution to the convex optimization problem in Equation 4.5 is known as the graphical lasso [15]. Learning the structures using the observations in different groups separately does not take into consideration the similarities between their structures. In fact, the structure of a graphical model on a single sample group shouldn't deviate much from the rest. Since differences between the graphical models are of interest, Danaher et al., [12] proposed a technique for jointly estimating multiple graphical models. They solved the following optimization problem subject to constraint that $\Omega^1, \ldots, \Omega^{(K)}$ are positive def-

inite.

$$\min_{\{\Omega>0\}} \mathcal{L}(\{\Omega\}) := \sum_{k=1}^{K} \operatorname{tr}(\widehat{\Sigma}^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)}) + P(\{\Omega\}) \tag{4.6}$$

where $P(\{\Omega\})$ denotes a convex penalty function. They defined two regularization functions to foster the precision matrices to share certain characteristics. Their first proposed regularization function, fused graphical lasso as shown in Equation 4.7, applies $\ell_1$ regularization for sparsity constraint, and the fused lasso [22] penalty regularization function to the differences between corresponding elements of each pair of precision matrices to encourage similar edge values.

$$P(\{\Omega\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i\neq j} |\omega_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\omega_{ij}^{(k)} - \omega_{ij}^{(k')}| \tag{4.7}$$

where $\lambda_1$ and $\lambda_2$ are nonnegative tuning parameters. Second regularization function they proposed, group graphical lasso as shown in Equation 4.8, also applies $\ell_1$ regularization for sparsity constraint, and the group lasso penalty [55] to the $(i, j)$ element across all $K$ precision matrices in order to have an identical pattern of non-zero elements in the precision matrices.

$$P(\{\Omega\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i\neq j} |\omega_{ij}^{(k)}| + \lambda_2 \sum_{i\neq j} \sqrt{\sum_{k=1}^{K} \omega_{ij}^{(k)^2}} \tag{4.8}$$

## 4.2 miRNA-gene binding network inference

Next, we learned a BN to represent the binding relationship between miRNAs and genes. Interactions among miRNA and genes were obtained as per the steps described in section one of the Chapter 2. These interactions were derived from two Pancreactic cell lines: a total of 436 interactions involving 20 unique

miRNAs and 350 uniques genes for hTERT-HPNE and 414 interactions among 17 unique miRNAs and 285 unique genes. We used these interactions to build the evidence matrix $E$, where $E[i, j] = 1$ if there is a reported interaction between gene $i$ and miRNA $j$. We first used the Greedy Hill Climbing (GHC) method to find initial DAGs, and then applied the Tabu search algorithm starting with those DAGs with tabu size 100 and a maximum of 2 changes that decreases the score of model. The Bayesian Dirichlet equivalence uniform (BDeu) [9] scoring implemented in aGrUM package [17] was used for this score-based learning process. Figure 4.2 depicts the workflow of this stage. Once we obtained a DAG for binding network, we converted the DAG to its Markov equivalent undirected graphical model (moralized graph). In the equivalent undirected model, there is an undirected edge between two nodes if they share a directed edge in the original graph or they are parents of the same node. In the pancreatic cancer case, the binding network in the DAG and undirected graphical model have 1,278 and 5,484 edges, respectively. In order to explain the impact of miRNA-mediated regulation in the gene interaction networks, we used the entropic Gromov-Wasserstein distance [46] to assess the similarity between two phenotypes by including only expressions of genes involved in direct interactions of miRNAs as well as the interactions of their dependencies. As shown in Figure 4.3, calculating the distance of normal and cancer expression profiles to understand the impact of miRNA M1 involves both the expressions of direct interaction with G1 and all interactions of its dependencies G3. We only kept the distances if they are greater than the threshold, 0.0127, which is the distance between the normal and cancer profiles based on the entire DEGs. We then ranked the miRNA-mRNA interactions based on the distance with the reasoning that existence of top-ranked binding

cases differentiates the regulatory mechanisms in cancer compared to normal more than other binding sites.

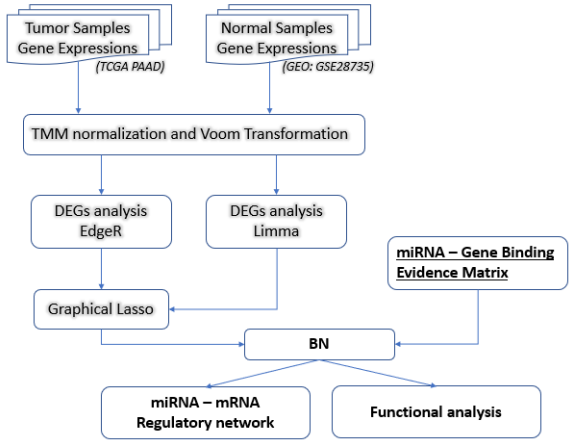



Figure 4.2: Worklow for miRNA-gene binding network inference

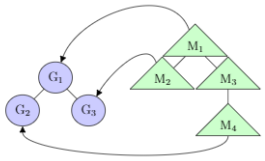



Figure 4.3: Illustration of connecting GRN and miRNA co-binding models

## 4.3 Functional analysis

We built networks for each progressive stage following the aforementioned steps and then investigated the structural differences of the learned models. Specifically, we focus on gene-gene or miRNA-gene interactions that are newly introduced to each stage and absent in the preceding stage, which are defined as stage-specific interactions. For instance, if there exists an interaction between gene A and gene B in stage 1, and interaction between gene B and gene C in stage 2,then we only consider gene A for stage 1, and gene C for stage 2

for functional analysis. We excluded the common gene B in these interactions from set enrichment analysis because including such a common gene may undermine differences between pairing patterns specific to stages. Based on this information, functional roles enriched in each stage was analyzed through the following approaches. Genes were clustered based on their projection at a specific level of the gene ontology corpus. Enrichment test were computed using clusterProfiler [52] for gene ontology terms and were followed by a KEGG [26] Pathways analysis. This was performed to assess biological significance of the gene sets obtained by testing over-representation of gene ontology terms.

## 4.4 Conditional interactions identified using supervised neural network model

We explore a deep learning solution based on Convolutional Neural Network (CNNs) [33] to elucidate the conditional miRNA-mRNA interaction based on gene and miRNA expression profiles in this study. CNN is a special kind of deep neural network, designed to be spatially invariant and to recognize patterns directly from an input. It is composed of multiple building blocks such as convolution layers, pooling layers, and fully connected layers. Early layers of CNN models learn low-level features, while deep layers learn high-level features which are composed of low-level features. We designed multiple layers of two-input 1D CNN model to discover spatial gene and miRNA features. In our CNN architecture, we add max pooling layers after convolution layers to reduce spatial dimensions which also helps to control overfitting. Additionally, we add dropout layers after dense layers to control overfitting. We truncate the CNN architecture at concatenation layer and saved the weights in the last

dense layers before concatenation layers after training is complete.

miRNA-gene binding interactions obtained from our De Novo preprocessing were considered to serve as a positive label for classifiers. One branch of an architecture gets gene expression values, and the other branch gets miRNA expression values and they are concatenated down the line for performing binary classification task to predict binding relationships from expression values.

# Chapter 5

# Results and Discussion

We used AIC criterion for model selection, and we found $\ell_1 = 0.7$ and $\ell_2 = 0.025$ gives the best fitted model to our data. We investigated the structural differences of the learned models for each cancer stage. The GRN models for each cancer stage are learned together with normal samples. We hypothesized that GRN model for different cancer stages should not deviate much from the GRN for normal condition. Figure 5.1 shows the numbers of edges in each subset and their intersections. The large number of common edges in the intersection of all GRNs verifies our hypothesis.

To analyze the result of this network, we used Cytoscape [47] for visualization and analysis. In gene regulation networks, including miRNA–mRNA interaction networks, one of the most relevant metrics is centrality. This is a measure of the degree, i.e., the number of edges connected to a vertex; the assumption is that vertices with the highest degrees (with the most connections) play important roles in the functioning of the system, making the degree of centrality a useful guide for focusing attention on the system's most crucial elements [11, 1]. We consider the betweenness centrality measurement. Table 5.1 report some statistics about the Gene-Gene network. A visualization of this network with Cytoscape is shown in Figure 5.2. The degree distribution

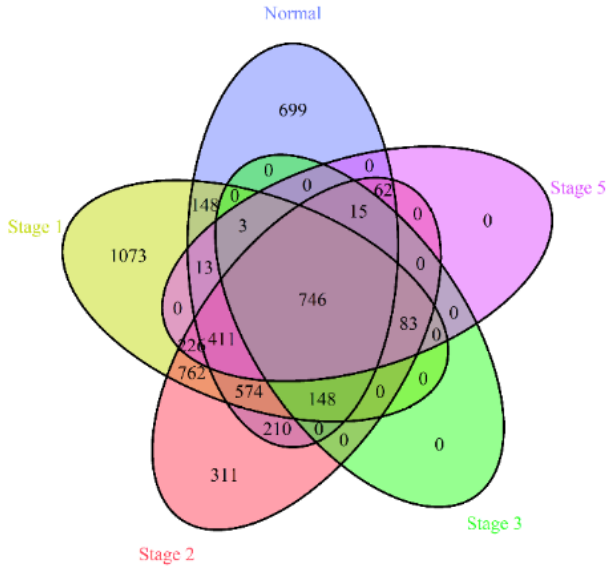



Figure 5.1: Venn diagram for gene interactions on the identified networks

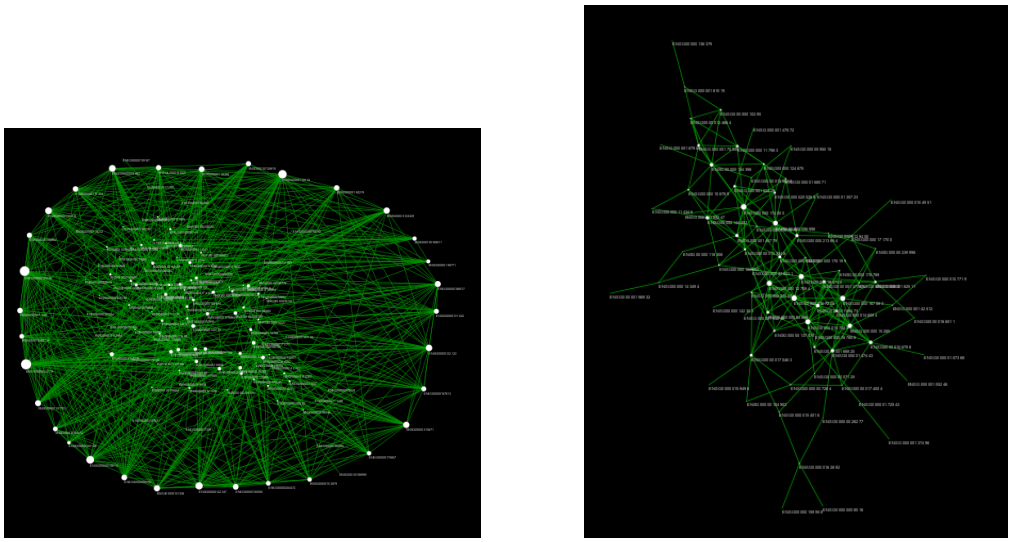
of each network inferred (stage-specific) fits the power law, which is a defining property of a scale-free network. This property gives the network a great tolerance against errors. Figure 5.3 reports the plot of each network.

| | Normal | Stage 1 | Stage 2 | Stage 3 | Stage 4 |
|---|---|---|---|---|---|
| Number of Nodes | 247 | 307 | 263 | 185 | 224 |
| Number of Edges | 3029 | 4187 | 3548 | 995 | 1559 |
| Min Degree of a Node | 1 | 1 | 1 | 1 | 1 |
| Max Degree of a Node | 89 | 97 | 96 | 46 | 62 |
| Average Degree | 26 | 28 | 27 | 12 | 14 |
| Betweenness Centrality | 0.6 | 1 | 1 | 0.1 | 0.1 |
| Clustering Coefficient | 0.511 | 0.455 | 0.551 | 0.453 | 0.465 |

Table 5.1: Results on the predicted gene-gene interactions

There are in total 100 new genes that appeared in the Stage 1 and not present in the solid normal tissue of Pancreatic cancer while from Stage 1 to Stage 2, 15 new genes make their appearance. From stage 2 to stage 3, 6 new genes introduced in the network and 44 new genes from Stage 3 to Stage 4.

Once miRNA and genes are bound together, We consider the miRNAs with



(a) Normal-All Stages (b) Normal-Stage 1

Figure 5.2: Network visualization
(a) represent the intersection between all the stage and normal whereas
(b) shows the edges only present in Stage 1.

smaller degrees (less than 20) to be particularly important. This is because when a miRNA with a high degree is excluded from the dataset, many potentially important dependents are also taken off from the analysis which leads to a greater difference from the baseline. miRNAs with degrees in top-20 percent are:hsa-miR-139-5p, hsa-miR-451a, hsa-miR-194-5p, hsa-miR-150-5p. Table 5.2 reports the mirna genes binding obtained. Four of which are experimentally validated microRNAs.

Using these 12 identified miRNA that have binding information to our DEGs list. we assessed their functional analysis using KEGG [26] and GO enrichment analysis. With the help of dotplot functionality of clusterProfile package [53] in R, Figure 5.4 showed significant functions enriched in more advanced stage (a) and detailed the genes associated with one or more terms (b) during cancer progression.

Gene Set Enrichment Analysis (GSEA) [49] on gene sets from KEGG path-

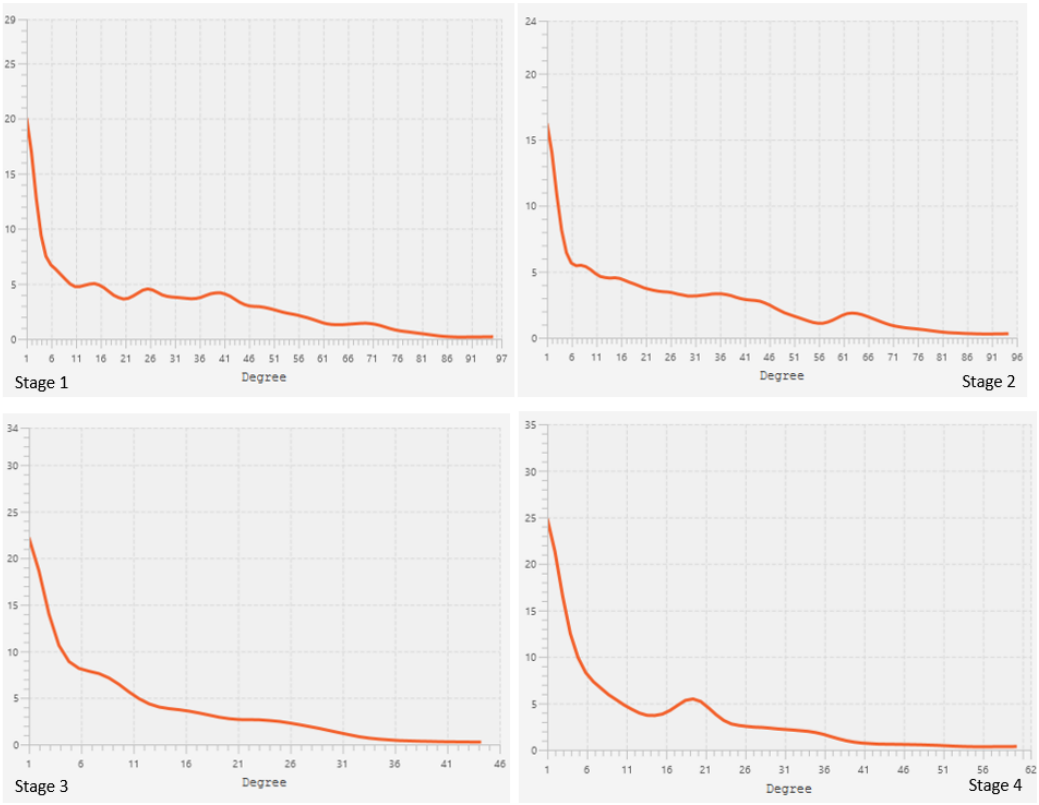


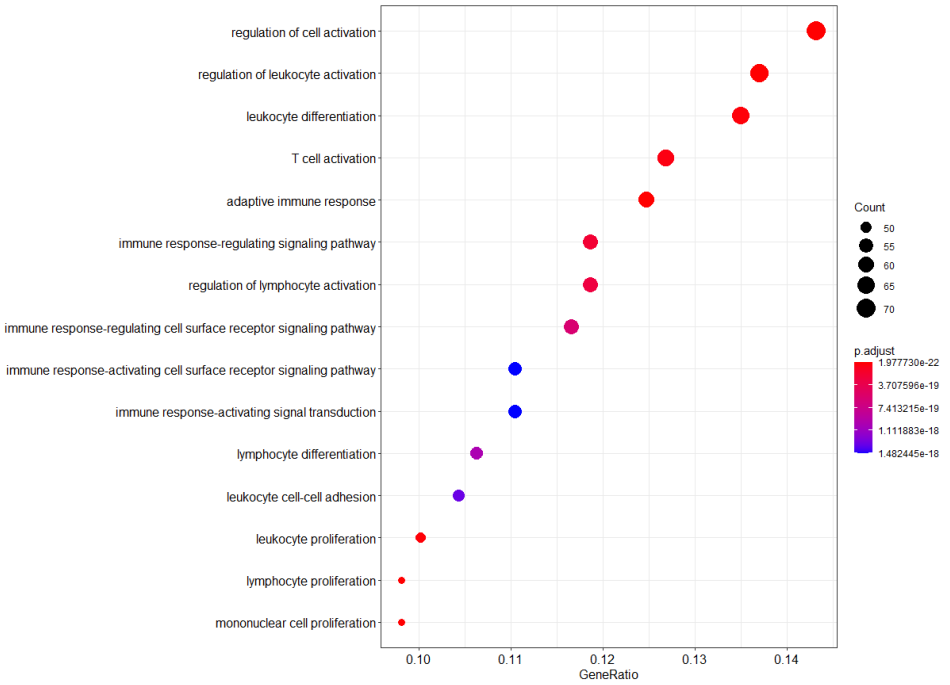

Figure 5.3: Degree distribution of network inferred for each stage

ways was performed and 74 enriched pathway were found and the resulting table can be found in Appendix **A1**. Most of the resulting pathways are also reported in [54].

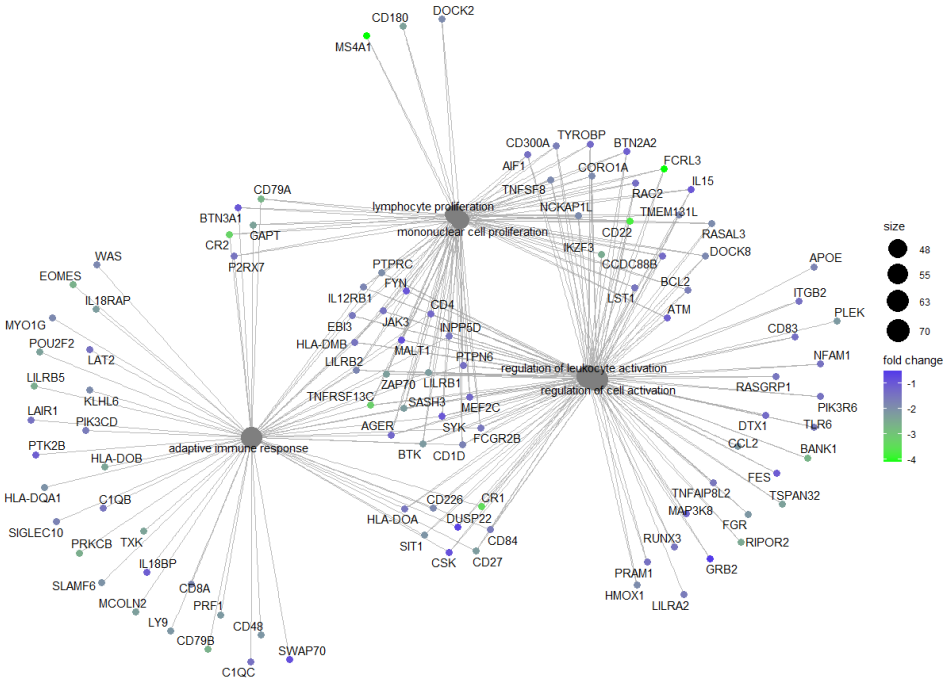
In the absence of benchmarking data, a 10-fold cross-validation was performed to analyze the model's performance. Our model has demonstrated promising prediction power on identifying conditional miRNA-Gene interactions. It achieved an average accuracy of about 97% and Figure 5.5 plots the receiver operating characteristic.

| miRNA | Gene |
|---|---|
| hsa-miR-139-5p | ENSG00000186642 |
| hsa-miR-139-3p | ENSG00000182253 |
| hsa-miR-451a | ENSG00000188536 |
| hsa-miR-194-5p | ENSG00000006611 |
| hsa-miR-192-5p | ENSG00000170608 |
| hsa-miR-196a-5p | ENSG00000272763<br>ENSG00000123388 |
| hsa-miR-196b-5p | ENSG00000253293<br>ENSG00000106031 |
| hsa-miR-210-3p | ENSG00000130821 |
| hsa-miR-135b-5p | ENSG00000281406 |
| hsa-miR-1224-5p | ENSG00000089199 |
| hsa-miR-375 | ENSG00000167964<br>ENSG00000116299 |
| hsa-miR-155-5p | ENSG00000234883 |

Table 5.2: miRNA-Gene Network Result




(a) Enrichment analysis for GO Biological Process




(b) Top 5 significant terms GO Enrichment

Figure 5.4: Functional Analysis

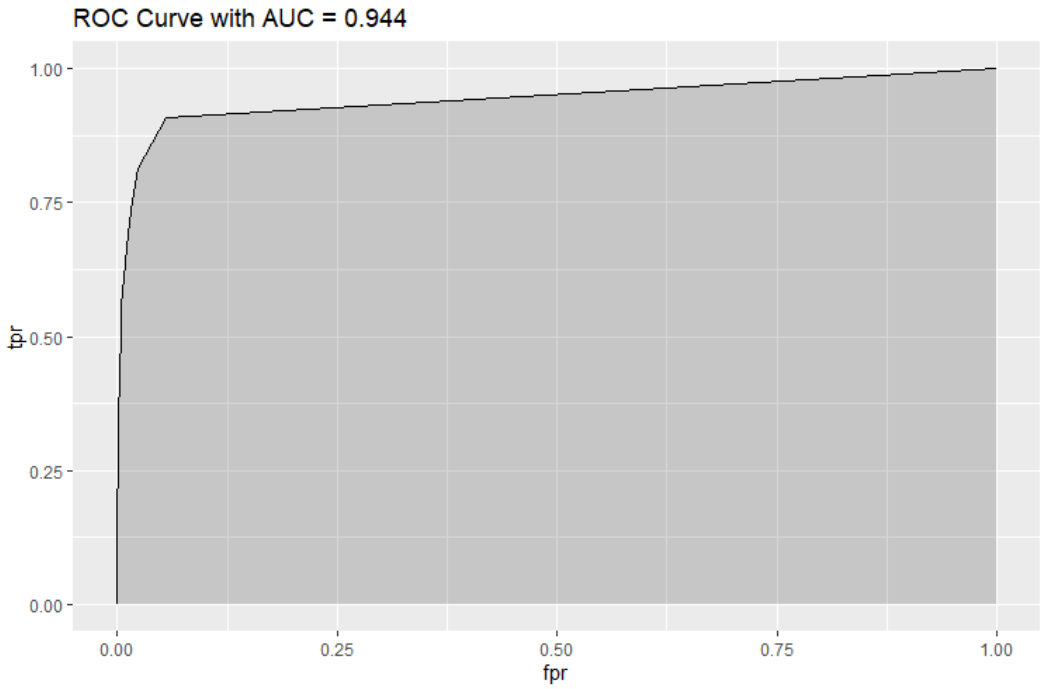



Figure 5.5: Performance evaluation based on ROC (Receiver Operating Characteristic)

# Chapter 6

# Conclusions and Future work

This thesis presented an integrative methodology to construct multifaceted gene regulatory networks among different types of bio-molecules in cancers. Gene-gene and miRNA binding networks were inferred based on sequencing-derived expression data and interaction information. The major contribution of this study is the presentation of a network learner that can merge continuous and discrete data models, and supports queries on variables of interest for interaction predictions, which can be generalized for similar applications in biomedical research.

Two important directions are identified and can be pursued in the future to improve this work. First, a deep investigation of gene expression conversion into image to unlock the full potential of a CNN-based model. An attempt to use expression value directly into a 1-D vector has been made and this will be pursued along with how to encode in a meaningful way expression value into image for a CNN model interactions inference. Second, how to address data augmentation within real biological data as limited samples size poses serious challenges while inadequate data augmentation has serious repercussions.

# Appendix A

# Appendix: Selected KEGG pathways

| ID | Description |
|---|---|
| hsa05034 | Alcoholism |
| hsa04974 | Protein digestion and absorption |
| hsa04972 | Pancreatic secretion |
| hsa00830 | Retinol metabolism |
| hsa04911 | Insulin secretion |
| hsa01230 | Biosynthesis of amino acids |
| hsa05204 | Chemical carcinogenesis |
| hsa00980 | Metabolism of xenobiotics by cytochrome P450 |
| hsa00982 | Drug metabolism - cytochrome P450 |
| hsa04950 | Maturity onset diabetes of the young |
| hsa05143 | African trypanosomiasis |
| hsa05320 | Autoimmune thyroid disease |
| hsa05330 | Allograft rejection |
| hsa04672 | Intestinal immune network for IgA production |
| hsa05332 | Graft-versus-host disease |
| hsa05340 | Primary immunodeficiency |
| hsa04080 | Neuroactive ligand-receptor interaction |
| hsa03010 | Ribosome |
| hsa05144 | Malaria |
| hsa04623 | Cytosolic DNA-sensing pathway |
| hsa05321 | Inflammatory bowel disease |
| hsa04260 | Cardiac muscle contraction |
| hsa05416 | Viral myocarditis |
| hsa05150 | Staphylococcus aureus infection |
| hsa04612 | Antigen processing and presentation |
| hsa00512 | Mucin type O-glycan biosynthesis |
| hsa05140 | Leishmaniasis |
| hsa00190 | Oxidative phosphorylation |
| hsa04061 | Viral protein interaction with cytokine and cytokine receptor |
| hsa04662 | B cell receptor signaling pathway |
| hsa04640 | Hematopoietic cell lineage |
| hsa04658 | Th1 and Th2 cell differentiation |
| hsa05323 | Rheumatoid arthritis |
| hsa04970 | Salivary secretion |

| | |
|---|---|
| hsa04660 | T cell receptor signaling pathway |
| hsa04620 | Toll-like receptor signaling pathway |
| hsa04650 | Natural killer cell mediated cytotoxicity |
| hsa05235 | PD-L1 expression and PD-1 checkpoint pathway in cancer |
| hsa05142 | Chagas disease |
| hsa04625 | C-type lectin receptor signaling pathway |
| hsa04666 | Fc gamma R-mediated phagocytosis |
| hsa04659 | Th17 cell differentiation |
| hsa04670 | Leukocyte transendothelial migration |
| hsa04064 | NF-kappa B signaling pathway |
| hsa04933 | AGE-RAGE signaling pathway in diabetic complications |
| hsa05145 | Toxoplasmosis |
| hsa04630 | JAK-STAT signaling pathway |
| hsa04611 | Platelet activation |
| hsa00010 | Glycolysis / Gluconeogenesis |
| hsa04613 | Neutrophil extracellular trap formation |
| hsa04380 | Osteoclast differentiation |
| hsa05162 | Measles |
| hsa04514 | Cell adhesion molecules |
| hsa05135 | Yersinia infection |
| hsa00601 | Glycosphingolipid biosynthesis - lacto and neolacto series |
| hsa04142 | Lysosome |
| hsa04145 | Phagosome |
| hsa04621 | NOD-like receptor signaling pathway |
| hsa05161 | Hepatitis B |
| hsa05164 | Influenza A |
| hsa05152 | Tuberculosis |
| hsa04664 | Fc epsilon RI signaling pathway |
| hsa04062 | Chemokine signaling pathway |
| hsa05133 | Pertussis |
| hsa05167 | Kaposi sarcoma-associated herpesvirus infection |
| hsa04060 | Cytokine-cytokine receptor interaction |
| hsa05169 | Epstein-Barr virus infection |
| hsa05170 | Human immunodeficiency virus 1 infection |
| hsa05417 | Lipid and atherosclerosis |
| hsa05146 | Amoebiasis |
| hsa05163 | Human cytomegalovirus infection |
| hsa04668 | TNF signaling pathway |
| hsa04915 | Estrogen signaling pathway |
| hsa05166 | Human T-cell leukemia virus 1 infection |

## Bibliography

[1] Networks: An introduction. 2010: Oxford university press. *Artificial life.*, 18, 2012.

[2] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel. Predicting effective microrna target sites in mammalian mrnas. *eLife*, 4:e05005, aug 2015.

[3] S. Althammer, J. González-Vallinas, C. Ballaré, M. Beato, and E. Eyras. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics*, 27(24):3333–3340, Dec. 2011.

[4] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[5] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12(1):56–68, Jan. 2011.

[6] D. P. Bartel. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2):215–233, 2009.

[7] D. Betel, M. Wilson, A. Gabow, D. S. Marks, and C. Sander. The microRNA.org resource: targets and expression. *Nucleic Acids Research*, 36(suppl_1):D149–D153, 01 2008.

[8] S. Bottini, N. Hamouda-Tekaya, B. Tanasa, L.-E. Zaragosi, V. Grandjean, E. Repetto, and M. Trabucchi. From benchmarking HITS-CLIP peak detection programs to a new method for identification of miRNA-binding sites from Ago2-CLIP data. *Nucleic Acids Research*, 45(9):e71–e71, 01 2017.

[9] W. L. Buntine. Theory Refinement on Bayesian Networks. In B. D'Ambrosio and P. Smets, editors, *UAI '91: Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence, University of California at Los Angeles, Los Angeles, CA, USA, July 13-15, 1991*, pages 52–60. Morgan Kaufmann, 1991.

[10] P. M. Clark, P. Loher, K. Quann, J. Brody, E. R. Londin, and I. Rigoutsos. Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. *Scientific Reports*, 4(1):5947, Aug. 2014.

[11] W. A. da Silveira, L. Renaud, J. Simpson, W. B. Glen, Jr, E. S. Hazard, D. Chung, and G. Hardiman. miRmapper: A Tool for Interpretation of miRNAmRNA Interaction Networks. *Genes*, 9(9):458, Sept. 2018. Publisher: MDPI.

[12] P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 76:373–397, Mar. 2014.

[13] C. Dimitrakopoulos. *Computational Studies in Cancer Multi-Omic Data Integration.* Doctoral Thesis, ETH Zurich, 2018. Accepted: 2019-01-22T09:08:11Z.

[14] C. Evans, J. Hardin, and D. M. Stoebel. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792, 02 2017.

[15] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, July 2008.

[16] K. Glass, C. Huttenhower, J. Quackenbush, and G.-C. Yuan. Passing messages between biological networks to refine predicted interactions. *PLOS ONE*, 8(5):1–14, 05 2013.

[17] C. Gonzales, L. Torti, and P.-H. Wuillemin. aGrUM: A graphical universal model framework. In *Advances in Artificial Intelligence: From Theory to Practice*, pages 171–177. Springer International Publishing, 2017.

[18] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl_1):D154–D158, 11 2007.

[19] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Systems Biology*, 6(1):145, Nov. 2012.

[20] D. P. Hayden, Y. H. Chang, J. Goncalves, and C. J. Tomlin. Sparse network identifiability via Compressed Sensing. *Automatica*, 68:9–17, 2016.

[21] D. M. D. Herrera. Integrative Pathway Analysis Pipeline For Mirna And Mrna Data. page 109.

[22] H. Hoefling. A Path Algorithm for the Fused Lasso Signal Approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[23] V. A. Huynh-Thu and G. Sanguinetti. Gene regulatory network inference: an introductory survey, 2018.

[24] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 104–113, 2003.

[25] A. Jacobsen, J. Silber, G. Harinath, J. T. Huse, N. Schultz, and C. Sander. Analysis of microRNA-target interactions across diverse cancer types. *Nature Structural & Molecular Biology*, 20(11):1325–1332, Nov. 2013.

[26] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, Jan. 2000.

[27] A. Kozomara and S. Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):D68–D73, 11 2013.

[28] R. M. Kuhn, D. Haussler, and W. J. Kent. The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, 14(2):144–161, 08 2012.

[29] M. L. Kuijjer, M. Fagny, A. Marin, J. Quackenbush, and K. Glass. Puma: Panda using microrna associations. *bioRxiv*, 2019.

[30] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Apr. 2012. Number: 4 Publisher: Nature Publishing Group.

[31] S. L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

[32] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, Feb. 2014.

[33] Y. LeCun and Y. Bengio. *Convolutional Networks for Images, Speech, and Time Series*, page 255–258. MIT Press, Cambridge, MA, USA, 1998.

[34] S. Lee and X. Jiang. Modeling mirna-mrna interactions that cause phenotypic abnormality in breast cancer patients. *PLOS ONE*, 12(8):1–22, 08 2017.

[35] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 06 2009.

[36] R. Li, H. Qu, S. Wang, J. Wei, L. Zhang, R. Ma, J. Lu, J. Zhu, W.-D. Zhong, and Z. Jia. GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics*, 34(14):2515–2517, 03 2018.

[37] Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 11 2013.

[38] Y. Lu, Y. Zhou, W. Qu, M. Deng, and C. Zhang. A Lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, 27(17):2406–2413, 07 2011.

[39] B. Lyu and A. Haque. Deep learning based tumor type classification using gene expression data. *bioRxiv*, 2018.

[40] M. Mostavi, Y.-C. Chiu, Y. Huang, and Y. Chen. Convolutional neural network models for cancer type prediction based on gene expression, 2019.

[41] A. E. Pasquinelli. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, 13(4):271–282, Apr. 2012.

[42] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 01 2010.

[43] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015.

[44] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009.

[45] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, Mar. 2010.

[46] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[47] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks.

*Genome research*, 13(11):2498–2504, Nov. 2003. Publisher: Cold Spring Harbor Laboratory Press.

[48] C. Smetz and F. Université de Liège > Master ingé. civ. info. Gene regulatory network inference from observational and interventional expression data. June 2017. Accepted: 2017-07-01T02:07:02Z Publisher: Université de Liège, Liège, Belgique.

[49] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[50] G. Tseng, D. Ghosh, and X. J. Zhou. *Integrating Omics Data*. Cambridge University Press, 2015.

[51] D. Weighill, M. B. Guebila, C. Lopes-Ramos, K. Glass, J. Quackenbush, J. Platig, and R. Burkholz. Gene regulatory network inference as relaxed graph matching. *bioRxiv*, 2020.

[52] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology*, 16(5):284–287, May 2012. Edition: 2012/03/28 Publisher: Mary Ann Liebert, Inc.

[53] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 5 2012.

[54] S. Yu, Y. Wu, C. Li, Z. Qu, G. Lou, X. Guo, J. Ji, N. Li, M. Guo, M. Zhang, L. Lei, and S. Tai. Comprehensive analysis of the SLC16A gene family in pancreatic cancer via integrated bioinformatics. *Scientific Reports*, 10(1):7315, Apr. 2020.

[55] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 68(1):49–67, 2006.

[56] Y. Yuan and Z. Bar-Joseph. Deep learning for inferring gene relationships from single-cell expression data. *Proceedings of the National Academy of Sciences*, 116(52):27151–27158, 2019.

[57] G. Zhang, A. Schetter, P. He, N. Funamizu, J. Gaedcke, B. M. Ghadimi, T. Ried, R. Hassan, H. G. Yfantis, D. H. Lee, C. Lacy, A. Maitra, N. Hanna, H. R. Alexander, and S. P. Hussain. Dpep1 inhibits tumor cell invasiveness, enhances chemosensitivity and predicts clinical outcome in pancreatic ductal adenocarcinoma. *PLOS ONE*, 7(2):1–9, 02 2012.