공학박사 학위논문

Computational Approaches for Exploring the Relationships in High Dimensional Spaces of Multi-Omics Data Utilizing Biological Prior Knowledge

생물학적 사전 지식을 활용한 고차원의 다중 오믹스 관계를 찾는 컴퓨터 공학적 접근 방법

2021 년 8 월

서울대학교 대학원

컴퓨터 공학부

오 민 식

Computational Approaches for Exploring the Relationships in High Dimensional Spaces of Multi-Omics Data Utilizing Biological Prior Knowledge

생물학적 사전 지식을 활용한 고차원의 다중 오믹스 관계를 찾는 컴퓨터 공학적 접근 방법

지도교수 김 선

이 논문을 공학박사 학위논문으로 제출함

2021 년 5 월

서울대학교 대학원

컴퓨터 공학부

오 민 식

오민식의 공학박사 학위논문을 인준함

2021 년 6 월

위 원 장	박근수
부위원장	김선
위　원	황승원
위　원	김건희
위　원	정인욱

Abstract

Computational Approaches for Exploring the Relationships in High Dimensional Spaces of Multi-Omics Data Utilizing Biological Prior Knowledge

Minsik Oh
Department of Computer Science & Engineering
College of Engineering
Seoul National University

Understanding how cells function or respond to external stimuli is one of the most important questions in biology and medicine. Thanks to the advances in instrumental technologies, scientists can routinely measure events within cells in single biological experiments. Notable examples are multi-omics data: sequencing of genomes, quantifications of gene expression, and identification of epigenetic events that regulate expression of genes. In order to better understand cellular mechanisms, it is essential to identify regulatory relationships between multi-omics regulators and genes. However, regulatory relationships are very complex and it is infeasible to validate all condition-specific relationships experimentally. Thus, there is an urgent need for an efficient computational method to extract relationships from different types of high-dimensional omics data. One way to address these high-dimensional data is to incorporate external biological knowledge such as relationships between omics and functions of genes curated in various databases.

In my doctoral study, I developed three computational approaches to identify the regulatory relationships from multi-omics data utilizing biological prior knowledge.

The first study proposes a method to predict *one-to-m* relationships between miRNA and genes. The computational challenge of miRNA target prediction is that there are many miRNA target candidates, and the ratio of false positives to false negatives needs to be adjusted. This challenge is addressed by utilizing literature knowledge for determining the association between miRNA-gene and a given context. In this study, I developed ContextMMIA to predict miRNA-gene relationships from miRNA and gene expression data. ContextMMIA computes scores of miRNA-gene relationships based on statistical significance and literature relevance and prioritizes the relationships based on the scores. In experiments on breast cancer data with different prognosis, ContextMMIA predicted differentially activated miRNA-gene relationships in invasive breast cancer. The experimentally verified miRNA-gene relationships were predicted with high priority and those genes are known to be involved in breast cancer-related pathways.

The second study proposes a method to predict *n-to-one* relationships between regulators and gene on drug response. The computational challenge of drug response prediction is how to integrate multi-omics data of 20,000 genes for determining drug response mediator genes. This challenge is addressed by utilizing low-dimensional embedding methods, literature knowledge of drug-gene associations, and gene-gene interaction knowledge. For this problem, I developed DRIM to predict drug response relationships from the multi-omics data and drug-induced time-series gene expression data. DRIM uses autoencoder, tensor decomposition, and drug-gene association to determine *n-to-one* relationships from multi-omics data. Then, regulatory relationships of mediator genes are determined by gene-gene interaction knowledge and cross-correlation of drug-induced time-series gene expression data. In experiments on breast cancer cell line data, DRIM extracted mediator genes relevant to drug response and regulatory relationships of genes involved in the PI3K-Akt pathway targeted by lapatinib. In addition, DRIM revealed distinguished patterns of relationships in breast cancer cell lines with different lapatinib resistance.

The third study proposes a method to predict *n-to-m* relationships between regulators and genes. In order to predict *n-to-m* relationships, this study formulated an objective function that measures the deviation between observed gene expression values and estimated gene expression values derived from gene regulatory networks. The computational challenge of minimizing the objective function is to navigate the search space of relationships exponentially increasing according to the number of regulators and genes. This challenge is addressed by the iterative local optimization with regulator-gene interaction knowledge. In this study, I developed a two-step iterative RL-based method to predict *n-to-m* relationships from regulator and gene expression data. The first step is to explore the *n-to-one* gene-oriented step that selects regulators by reinforcement learning based heuristic to add edges to the network. The second step is to explore the *one-to-m* regulator-oriented step that stochastically selects genes to remove edges from the network. In experiments on breast cancer cell line data, the proposed method constructed breast cancer subtype-specific networks from the regulator and gene expression profiles with a more accurate gene expression estimation than previous combinatorial optimization methods. Moreover, regulatory relationships involved in the networks were associated with breast cancer subtypes.

In summary, in this thesis, I proposed computational methods for predicting *one-to-m*, *n-to-one*, and *n-to-m* relationships between multi-omics regulators and genes utilizing external domain knowledge. The proposed methods are expected to deepen our knowledge of cellular mechanisms by understanding gene regulatory interactions by analyzing the ever-increasing molecular biology data such as The Cancer Genome Atlas, Cancer Cell Line Encyclopedia.

Keywords: High dimensional data, Multi-omics, Gene expression, Machine learning, Biological prior knowledge, Search space

Student Number: 2015-22904

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cell state is characterized by the complex interactions of various genetic molecules such as the genome, transcriptome, epigenome, proteome, metabolome, and microbiome, called multi-omics. With recent sequencing technology development, each genetic events can be measured, and the integration of multi-omics data help interpretation of the cell state by identifying regulatory relationships of genes. However, identifying multi-omics relationships demands efficient computational approaches due to the complexity of relationships between high-dimensional omics data. In this dissertation, I developed algorithms for exploring multi-omics relationships to interpret cell states utilizing external biological prior knowledge as guidance for high-dimensional space.

1.1 Biological background

1.1.1 Multi-omics analysis

The recent advance in high-throughput sequencing technologies provides opportunities to measure the state of the cell represented by multi-omics. Depending on the type of omics data, various experimental protocols are used for measuring genetic events such as genome sequencing (Bentley *et al.*, 2008), RNA sequencing (Nagalakshmi *et al.*, 2008), chromatin immunoprecipitation sequencing (ChIP-seq) (Kharchenko *et al.*, 2008), bisulfite sequencing (Chatterjee *et al.*, 2012), and mass spectrometry (Cox and Mann, 2011). Each omics data represents unique characteristics and provides a particular view of biological systems at the molecular level. However, for the comprehensive understanding of biological systems, there is a need the integration of different types of omics data since genetic molecules are activated by interactions with inter- and intra-omics molecules. In order to understand the complex biological process, various computational methods have been proposed for the integration of multi-omics data (Subramanian *et al.*, 2020; Oh *et al.*, 2020b).

1.1.2 Multi-omics relationships indicating cell state

Cell state can be represented by gene expression that is the series of processes DNA into a functional product in the cell (Crick, 1970). The expression patterns of thousands of genes determine which functions are turned on or off in specific cells and determine cellular phenotypes such as disease development.

As shown in Figure 1.1, various genetic molecules regulate the expression levels of genes at different levels, such as interactions with other genes, microRNA (miRNA), DNA methylation, mutations, transcription factors (TFs), which is represented by the *n-to-m* relationship between regulators and genes. However, verifying the relationship between multi-omics regulators and gene expression through biological experiments is infeasible because there are too many relationship candidates. Therefore, a computational approach is necessary to investigate multi-omics relationships from multi-omics data to characterize condition-specific cellular states represented by complex relationships between multiple regulators and genes.

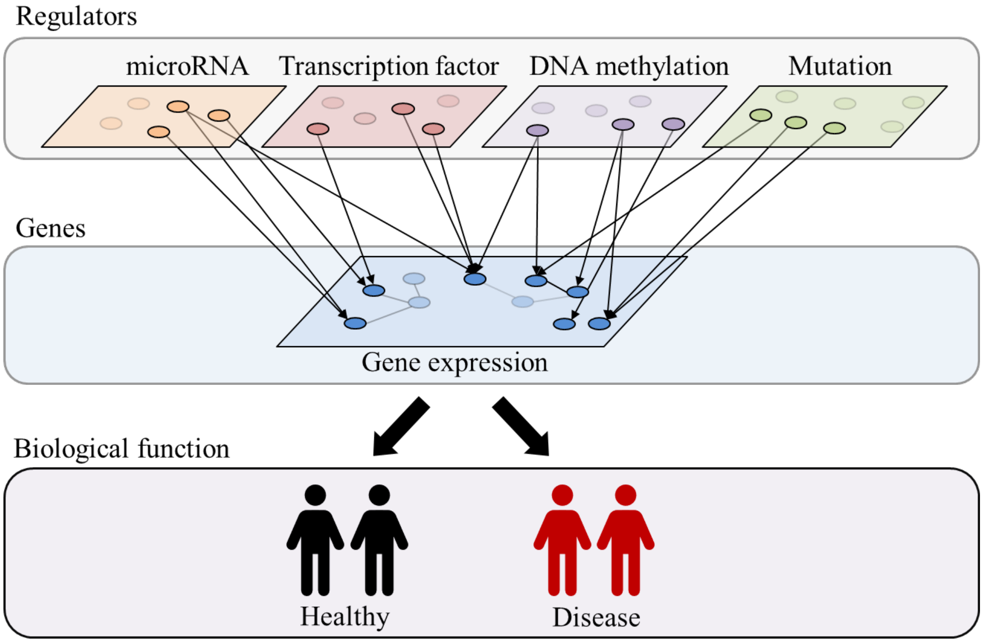
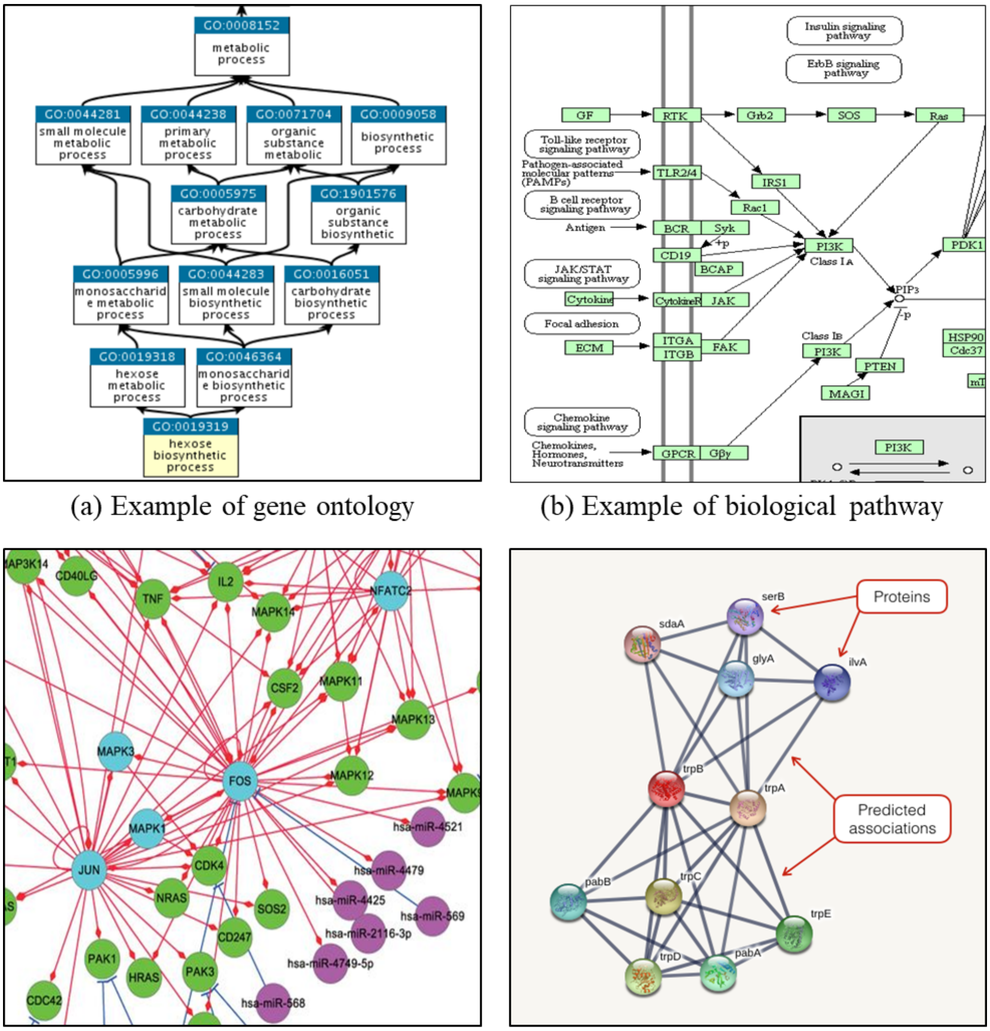


Figure 1.1: The example of *n-to-m* relationship between regulators and genes. The cell state determined by *n-to-m* relationship between regulators and genes. According to cell state, biological phenotype such as disease development is determined.

1.1.3 Biological prior knowledge

Biological prior knowledge provides information about biological functions of genes or interactions between different genetic molecules provided as computable structures. The prior knowledge can be acquired in various ways such as biological experiments, combined databases, computational predictions, and literature text mining (Figure 1.2). For instance, Gene ontology (GO) database provides the function of genes as a comprehensive and computational representation from molecules to the cellular level (Ashburner *et al.*, 2000). Regulatory relationship databases provide gene regulatory relationships such as TF-gene, miRNA-gene relation as a network structure (Liu *et al.*, 2015; Lewis *et al.*, 2003; Matys *et al.*, 2006). Biological pathway databases provide relationships of genes involved in biological pathways that lead to the product of changes in cell state (Kanehisa and Goto, 2000). STRING network provides protein-protein interaction (PPI) that two proteins jointly contribute to a specific biological function (Szklarczyk *et al.*, 2015). The external knowledge has been used for identifying multi-omics relationships by filtering disease-specific molecules or interactions between omics (Steele *et al.*, 2009; Ideker *et al.*, 2011).



(a) Example of gene ontology

(b) Example of biological pathway

(c) Example of regulatory network

(d) Example of PPI network

Figure 1.2: The example of biological prior knowledge

1.2 Research problems for the multi-omics relationship

The main purpose of my thesis is to develop computational approaches for identifying multi-omics relationships to understand condition-specific cellular states. In particular, the thesis focuses on the relationship between multi-omics regulators and gene expressions in multi-omics data since gene expression is one of the factors determining cellular states (Zare *et al.*, 2014). The multi-omics relationships can be represented as graph structure described in Figure 1.3. In a graph $G = (V, E)$, V is the union of the set of genes and regulators, each node in V has a weight that is quantitative measures of each omics such as miRNA expression level, DNA methylation level, gene expression level. E is a set of edges that represents regulatory relationships between regulators and genes. The research problem is to determine condition-specific *n-to-m* relationships between multi-omics regulators and genes represented by the graph G for given multi-omics profiles.

1.3 Computational challenges and approaches in the exploring multi-omics relationship

The main challenge of the problem is the high dimensionality of multi-omics relationships. Each omics data is high-dimensional with thousands to billions of dimensions depending on the omics type. Furthermore, relationships between multi-omics data have an exponential search space and the number of data is relatively small. To handle the high dimensional low sample multi-omics data, utilizing external biological prior knowledge is one way to explore the search space by reducing candidate relationships.

My doctoral study proposes three methods for exploring relationships of high dimensional multi-omics data incorporating biological prior knowledge

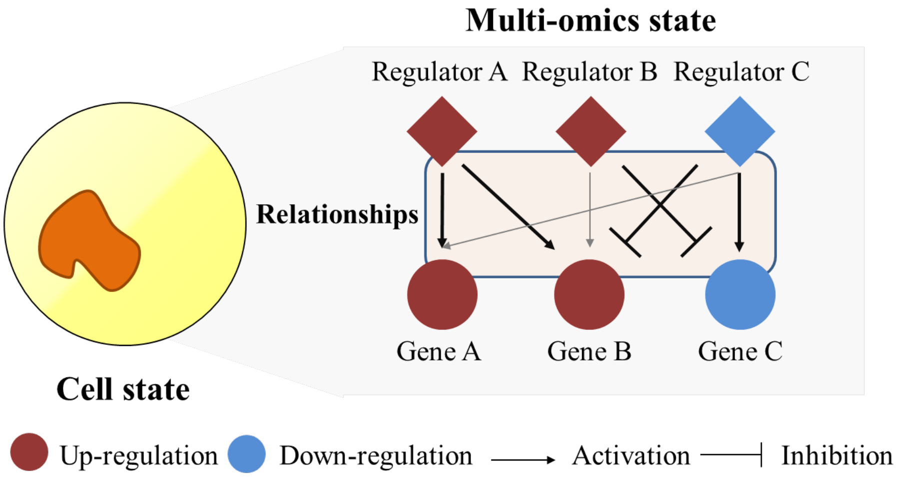


Figure 1.3: Graph representation of multi-omics relationships.

to understand complex cellular regulatory mechanisms. The first problem is miRNA target gene prediction that affects the difference in status between groups. In this study, the scoring method was used to balance false positive and false negative using expression value as well as literature knowledge.

The second problem is sample-specific drug response prediction combining multi-omics data and time-series gene expression data. In this study, drug response mediator genes are selected to reduce candidates using tensor decomposition, autoencoder, and literature knowledge. From the mediator genes with the time domain, upstream relationships are determined with TF gene target databases, and downstream relationships are determined with PPI networks and biological pathways.

The third problem is finding a sample-specific network of *n-to-m* relationships between regulators and genes. In this study, to find the optimal network estimating gene expression close to the measured gene expression, an iterative two-step search method was used for efficient network estimation in an *n-to-*

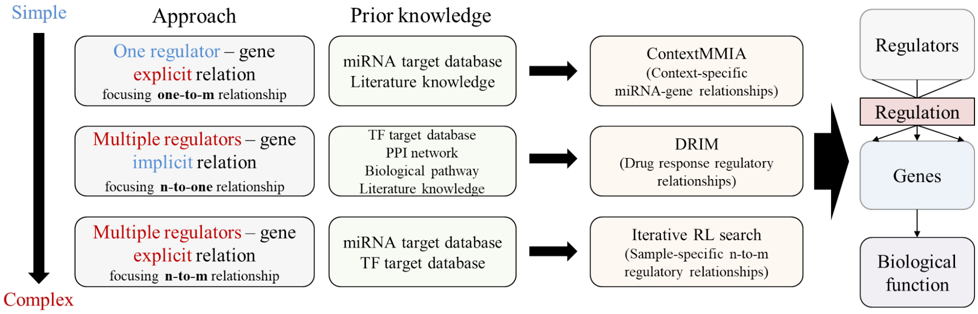


Figure 1.4: Approaches of each study for exploring relationships between regulators and genes.

m global search space. The detailed problem description with challenges and approaches are listed as follows (Figure 1.4).

- **Literature prior knowledge guided approach for exploring miRNA-gene relationships (ContextMMIA) (Oh *et al.*, 2017)**:

  Problem & Challenges: MiRNA can inhibit gene expression at the post-transcriptional level. In order to represent the cell state from miRNA expression data and gene expression data, it is necessary to predict the target gene of miRNA. The problem of this study is identifying condition-specific miRNA-gene relationships from miRNA expression and gene expression data. Previous miRNA-gene target prediction methods can be categorized into sequence-based prediction and expression-based prediction. There are many false positives since sequence-based prediction methods do not consider sample-specific information such as expression profiles. Expression-based prediction methods take advantage of statistical cutoffs, which have the false-negative problem of rejecting true-positive targets due to stringent cutoffs.

Approach: In this study, I focused on the *one-to-m* relationship between miRNA and genes. I proposed ContextMMIA, a data context-specific miRNA target prediction algorithm that combines miRNA, gene expression data, and biological prior knowledge. ContextMMIA utilizes two prior knowledge. One is the miRNA target databases curated miRNA-gene relationships from computational prediction and experimental validation to reduce candidate relationships. The other is literature knowledge that computes the association between miRNA-gene and a given data context such as disease and treated drug. The input of ContextMMIA is two groups of miRNA, gene expression data and context of data. ContextMMIA computes the statistical significance using expression data and literature relevance using literature knowledge. The miRNA-gene pairs in the target database are prioritized by scores of miRNA-gene relationships.

- **Drug response prediction with implicit *one-to-m* relationships by dimension reduction approach (DRIM) (Oh *et al.*, 2020a)**:

  Problem & Challenges: The individual variation of drug response depends on the cell state determined by the multi-omics state. It is essential to identify drug response relationships at the gene level to understand the variability of drug responses. The problem of this study is determining condition-specific drug response regulatory relationships from multi-omics data and drug-treated time-series gene expression data. The challenge of drug response prediction is integrating high-dimensional multi-omics data of 20,000 genes to determine drug response mediator genes, and determining gene-gene interactions from time-series gene expression data.

Approach: In this study, I focused on the *n-to-one* implicit relationship between multiple regulators and gene. I proposed DRIM, a drug response regulatory relationship prediction algorithm utilizing biological pathways, TF-target database, PPI network, and literature knowledge. DRIM takes multi-omics data and drug-treated time-series gene expression data as input. DRIM performs three steps for determining drug response. 1) It selects drug response mediator genes by multi-omics analysis. 2) It constructs a network representing upstream and downstream relationships of mediator genes using time-series gene expression data. 3) It computes the most probable drug-response relationships in the network using the influence-maximization algorithm. For the multi-omics analysis, tensor decomposition and autoencoder were used to address complex multi-omics relationships. The TF-target network was used to determine the upstream relationship of mediator genes. The PPI network and biological pathway were used to determine the downstream relationship of mediator genes.

- **Combinatorial modeling and optimization using iterative RL search for inferring sample-specific regulatory network**:

  Problem & Challenges: In order to identify the explicit *n-to-m* relationship between regulators and genes, I formulated the optimization problem to estimate observed gene expression from the network constructed by the *n-to-m* relationship. For the given regulator capacity and the amount of gene expression that changes due to the selected edges, the problem is to find the optimal network providing an accurate estimation of observed gene expression. The challenge of this problem is the search space growing exponentially according to the number of genes and regulators.

Approach: In this study, I focused on *n-to-m* relationships between multiple regulators and genes. Due to the size of the search space, it is challenging to explore a global *n-to-m* relationship. Thus, I proposed a search algorithm that iteratively explores *n-to-one* gene-oriented relationships and *one-to-m* regulator-oriented relationships. In the gene-oriented step (G-step), edges between regulators and a gene estimating the gene expression are added to the network determined by the reinforcement learning heuristics. In the regulator-oriented step (R-step), edges between regulator and genes violating the capacity constraint are removed from the network determined by a stochastic process. The G-step and R-step are iteratively computed until a terminate condition.

1.4 Outline of the thesis

Chapter 2, 3, and 4 introduce independent studies that are prior knowledge-guided methods for determining multi-omics relationships. In Chapter 2, a literature-based miRNA-gene target prediction method, ContextMMIA, is described. Chapter 3 describes a drug response prediction method DRIM is described. Chapter 4 proposes an iterative RL-based search method for determining the *n-to-m* relationship of regulators and genes.

Chapter 5 summarizes my contributions to the study of exploring the high dimensional space of multi-omics relationships. The thesis is concluded by an appendix of the bibliography of the cited references.

Chapter 2

Literature-based condition-specific miRNA-mRNA target prediction

MiRNAs are small non-coding RNAs that regulate gene expression by binding to the $3'$-UTR of genes. Many recent studies have reported that miRNAs play important biological roles by regulating specific mRNAs or genes. Many sequence-based target prediction algorithms have been developed to predict miRNA targets. However, these methods are not designed for condition-specific target predictions and produce many false positives; thus, expression-based target prediction algorithms have been developed for condition-specific target predictions. A typical strategy to utilize expression data is to leverage the negative control roles of miRNAs on genes. To control false positives, a stringent cutoff value is typically set, but in this case, these methods tend to reject many true target relationships, i.e., false negatives. To overcome these limitations, additional information should be utilized. The literature is probably the best resource that can be utilized. Recent literature mining systems compile millions of articles with experiments designed for specific biological questions, and the systems provide a function to search for specific information. To utilize the literature information, I used a literature mining system, BEST, that automatically extracts information from the literature in PubMed and that allows the user to perform searches of the literature with any English words. By integrating omics data analysis methods and BEST, I developed ContextMMIA, a miRNA-mRNA target prediction method that combines expression data analysis results and the literature information extracted based on the user-specified context. In the pathway enrichment analysis using genes included in the top 200 miRNA-targets, ContextMMIA outperformed the four existing target prediction methods that I tested. In another test on whether prediction methods can re-produce experimentally validated target relationships, ContextMMIA outperformed the four existing target prediction methods. In summary, ContextMMIA allows the user to specify a context of the experimental data to predict miRNA targets, and I believe that ContextMMIA is very useful for predicting condition-specific miRNA targets.

2.1 Computational Problem & Evaluation criterion

This chapter describes the method for solving the miRNA-gene target prediction problem. The input data is miRNA expression and gene expression data of samples from two different groups. The computational problem is to predict the miRNA-gene relationships that change the cellular state. Challenges of this problem are (1) there are too many miRNA-gene candidates, (2) false positive and false negative rates according to the cutoff value. Comparative experiments were performed to evaluate the proposed method in reproducibility of experimentally validated relationships and enrichment analysis of disease-related pathways with other methods.

2.2 Related works

MicroRNAs (miRNAs) are small non-coding RNAs that are 19-24 nucleotides in length. These RNAs regulate gene expression at the post-transcriptional level by binding to the $3'$-UTR of mRNAs (Ambros, 2004; Bartel, 2004); thus, miRNAs are functionally important. There are numerous scientific findings on the functional roles of miRNAs by regulating specific genes. For example, it is reported that miR-15 and miR-16-1 bind to BCL2 (Cimmino *et al.*, 2005) and that apoptosis is induced. Another example is that miR-125b, miR-145, miR-21 and miR-155 are dysregulated in breast cancer cells, and different expression levels of these miRNAs have significant correlations with breast cancer phenotypes, such as tumor stages and status of estrogen and progesterone receptors (Iorio *et al.*, 2005). Moreover, it is well known that miRNAs are related to proliferation, differentiation, and cell death (Hwang and Mendell, 2006).

The functional roles of miRNAs differ in different contexts. In other words, the relationship between miRNA and target genes is dynamic in different conditions. Thus, it is very important to identify which genes are targeted by miRNAs in a given context.

There are more than 1000 miRNAs, and approximately 60% of protein-coding genes are regulated by miRNAs (Friedman *et al.*, 2009). Since it is not possible to perform biological experiments for such a large number of miRNAs and genes, computational prediction is very important, and numerous computational methods have been developed for predicting targets of miRNAs. The first generation of computational tools leverage sequence complementary information and binding energy potentials. These prediction methods include TargetScan (Lewis *et al.*, 2005), PITA (Kertesz *et al.*, 2007), mirSVR (Betel *et al.*, 2010), miRanda (John *et al.*, 2004) and PicTar (Krek *et al.*, 2005). These tools generally come with corresponding databases that compile miRNA-target information. In addition to sequence complementary information, there are different approaches used in each of these methods. miRanda estimates the energy on sequence matching of miRNA and mRNA pairs to predict targets (John *et al.*, 2004). PicTar first finds candidate $3'$-UTR sites and uses a hidden Markov model (HMM) to filter out target sites (Krek *et al.*, 2005). TargetScan considers a conservation seed match and then considers regions outside seed matches (Lewis *et al.*, 2005). The mirSVR algorithm uses a support vector regression method to compute scores on candidate target sites that are identified by miRanda (Betel *et al.*, 2010). PITA uses the accessibility of target sites as a main feature to predict targets (Kertesz *et al.*, 2007).

Target prediction methods based on the sequence similarity score rely on the existence of target sites, and these methods are accompanied by target databases. However, such target information is not condition specific without considering which miRNAs and which genes are expressed; thus, there are many false positives even if the target information is accurate, which is not the case since many target databases do not agree on the miRNA-target relationship. To make the target information condition specific, many expression-based target prediction methods have been developed. These methods take miRNA-mRNA expression data and several sequence-based target databases as input data and filter out miRNA-mRNA targets using statistical significance or computational algorithms. I briefly summarize the previous expression-based algorithms. GenMiR++ used a Bayesian model and expectation maximization algorithm to predict the posterior probability of a miRNA target for mRNA (Huang *et al.*, 2007). MMIA employs a two-step method, where the first step is to select differentially expressed miRNA, and the second step is to select negatively correlated differentially expressed mRNA (Nam *et al.*, 2009) only for the differentially expressed miRNAs. MMIA also supports sequence data analysis on a cloud environment, which enables the user to utilize both microarray data and NGS data (Chae *et al.*, 2014). MAGIA2 is a web-based tool that considers the correlation among miRNA and mRNA and transcription factor (TF) regulation (Bisognin *et al.*, 2012). CoSMic extracts the significant target mRNA cluster for each miRNA (Ben-Moshe *et al.*, 2012). CoSMic employs methods similar to gene set enrichment analysis (GSEA) to identify miRNA targets (Subramanian *et al.*, 2005). miRNAmRNA is a target prediction algorithm based on the global test of a linear regression model (van Iterson *et al.*, 2013). To extract condition-specific miRNA activity, identifying causal relationships using intervention calculus when the DAG is absent was proposed (Zhang *et al.*, 2014). A recent tool, PlantMirnaT, was designed as a plant-specific miRNA-mRNA sequencing data analysis algorithm (Rhee *et al.*, 2015). The unique feature of PlantMirnaT is using the expression quantity information from sequencing data and employing a split ratio model to identify the relationship of target pairs.

2.3 Motivation

There are approximately 1,500 known miRNAs in the human genome. The number of possible miRNA-gene pairs exceeds 30 million when more than 20,000 protein-coding genes are considered. Among these pairs, only a fraction of the relationships are significant in terms of biological functions, e.g., phenotypes or cancer subtypes. Computational methods for predicting the miRNA target employ various techniques to identify phenotype-specific miRNA targets. Because this is a typical prediction problem, the challenges can be summarized in terms of false positives and false negatives.

- **Target databases have high false positive rates**: Sequence-based target prediction algorithms, such as TargetScan, mirSVR, and PITA, and their corresponding databases generally produce high false positives.

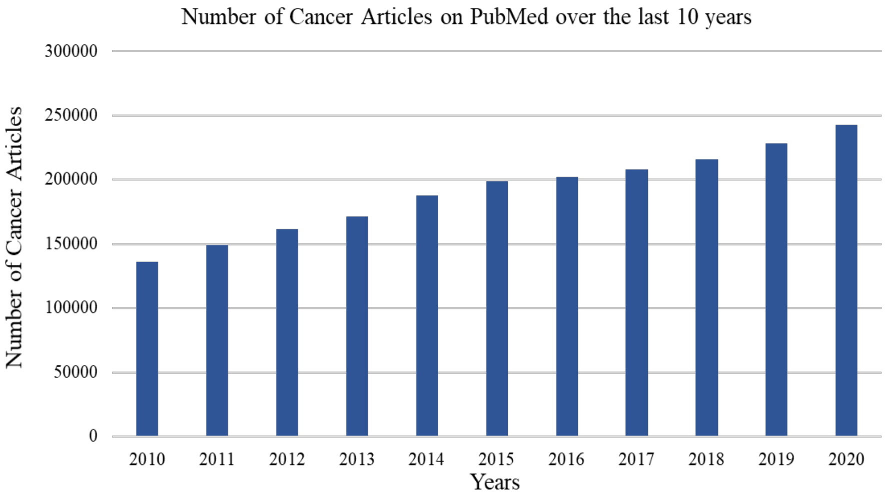


Figure 2.1: The number of published papers related to the keyword 'cancer' since 2010. More than 100,000 papers have been published every year.

There are two major reasons for these high false positives. First, these databases contain all known targets; thus, the target information is not condition specific. For this reason, when transcriptome data measured in a specific condition are analyzed, many targets are false positives. Second, sequence-based prediction methods do not consider the regulatory role of miRNA, which generally results in a negative correlation between miRNA and the target gene. In addition, sequence-based prediction methods do not consider sample-specific sequence information. For example, sequence variations in the target regions can affect the target relationship, but the current algorithms do not consider minor but subtle sequence variations.

- **Expression-based methods may have false negative rates**: Expression-based methods utilize negative correlation information between miRNA and targets or similar approaches. For these methods, there is always an issue of establishing a cutoff threshold value, e.g., for a negative correlation. If the cutoff value is not stringent, then there are too many miRNA-target relationships. Thus, in general, it is a common practice to set a quite stringent cutoff value. In this case, many true miRNA-target relationships can be rejected, i.e., the false negative issue.

Addressing the false positive and false negative issues is a very challenging problem. Using sequence pairing information and gene expression information is very useful because such methods have already produced many biologically meaningful results. However, one important information source, the literature, is not utilized in current methods. The scientific literature is currently growing exponentially. As shown in Fig 2.1, more than 100,000 papers related to 'cancer' are published every year. Thus, if I combine sequence pairing information and gene expression information with the literature information, I can certainly make a good improvement in predicting miRNA targets, reducing both false positives and false negatives. In particular, as with the use of gene expression information, the use of the literature information should be condition specific. The main issues are how to handle the vast amount of studies in the literature, how to allow the user to specify the experimental conditions, and finally, how to combine sequence pairing information, gene expression information and the literature information in a single computational framework.

Toward this goal, two research groups are working together to design and implement a novel human-specific miRNA-target prediction method.

First, I compute the **omics score** by utilizing sequence pairing information and gene expression information to produce candidate miRNA-target pairs. Then, I compute the literature-based **context score** to evaluate each candidate miRNA-target pair using the Biomedical Entity Search Tool (BEST) (Lee *et al.*, 2016). Using BEST, the user can specify the experimental condition using a set of any keywords, which will automatically be translated to a set of genes and related miRNAs. Subsequently, the two scores, the **omics score** and the **context score**, are combined into a single score in a conditional probabilistic form.

The remainder of this study is organized as follow. In the Methods section, I explain how to compute the **omics score** based on the expression data and miRNA-gene relationship and the **context score** from the literature according to user-provided keywords. In the Results section, I show how my proposed method performs compared with four existing methods in experiments with omics datasets in the public domain.

2.4 Methods

In this section, I explain how my method, ContextMMIA, predicts human miRNA targets by combining the literature information and gene expression data. ContextMMIA takes two-class (control vs. treated) human miRNA-mRNA expression data as input. Then, with user-specified keywords as the context of the experiment, it computes the probabilities of miRNA-gene pairs relevant to the phenotype differences by combining gene/miRNA expression data and the literature data. Fig 2.2 illustrates the workflow of ContextMMIA. First, differentially expressed miRNAs (DEmiRNAs) and differentially expressed mRNAs or genes (DEmRNAs) are determined with a cutoff value at the relaxed level such that most of the true positives can be retained in this step. Note that I use negative correlation information and the literature information to filter out and re-weight candidates for interaction pairs in the following steps. In the second step of processing omics data, human miRNA-mRNA pairs are predicted using miRNA target databases such as TargetScan, mirSVR, and PITA. These miRNA-mRNA pairs are further screened by negative correlation information between miRNA and mRNA. In the third step, for

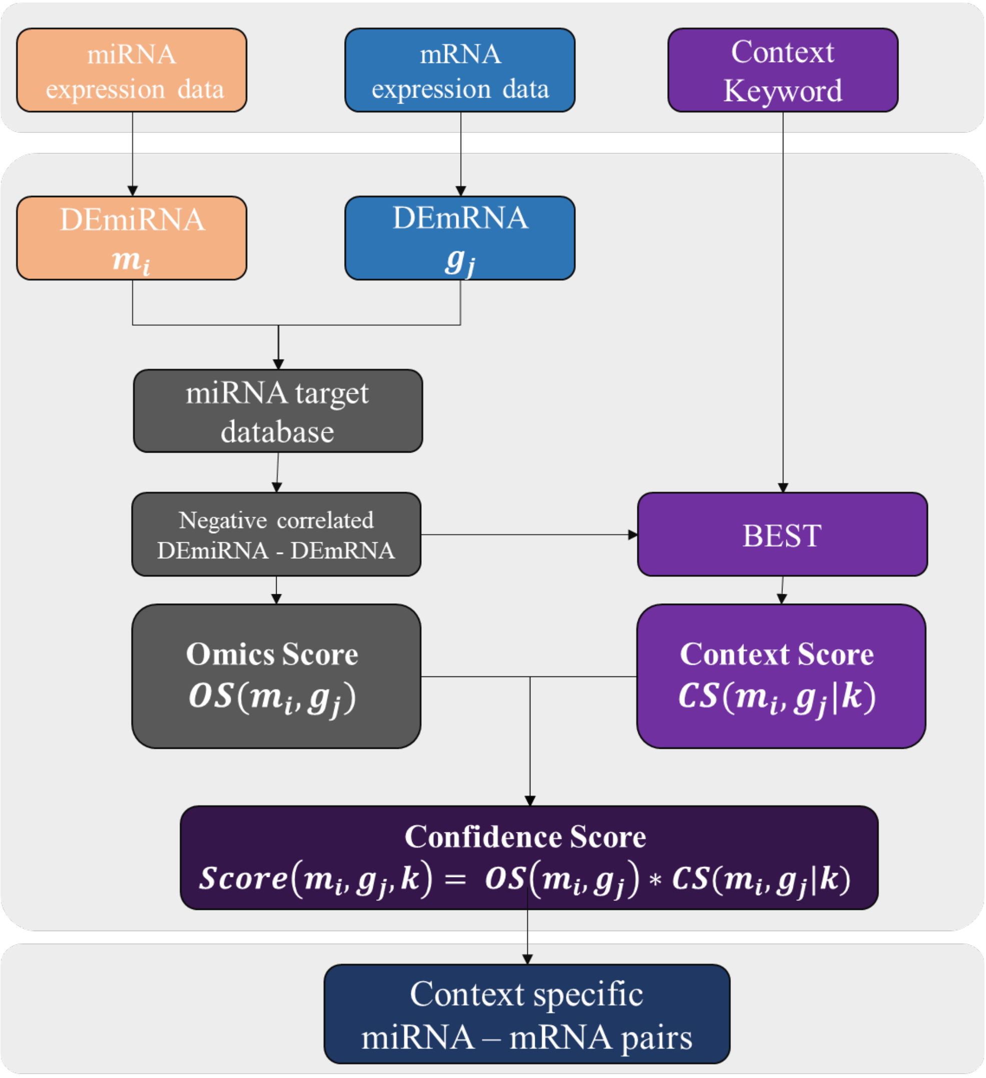

miRNA expression data
mRNA expression data
Context Keyword
DEmiRNA m_i
DEmRNA g_j
miRNA target database
Negative correlated DEmiRNA - DEmRNA
BEST
Omics Score $OS(m_i, g_j)$
Context Score $CS(m_i, g_j|k)$
Confidence Score $Score(m_i, g_j, k) = OS(m_i, g_j) * CS(m_i, g_j|k)$
Context specific miRNA – mRNA pairs

Figure 2.2: Schematic workflow for ContextMMIA. The system accepts expression information of miRNA data, mRNA data, and keyword as inputs. DEmiRNAs and DEmRNAs are extracted based on their expression level difference, and negative correlated target pair are extracted. Then, the system computes omics and context scores based on user-provided keywords by utilizing the BEST system. Finally, the system ranks context specific miRNA-mRNA pairs using the confidence scores.

each pair of miRNA and mRNA, ContextMMIA calculates the **omics score** based on expression data and the **context score** based on the literature information compiled based on the user-provided keywords. Finally, target pairs are ranked by combining the **omics score** and **context score**. For each miRNA-mRNA pair, ContextMMIA computes alignments of human miRNA and the $3'$-UTR of mRNA and generates the visualization of the miRNA-mRNA alignment on the website.

2.4.1 Identifying genes and miRNAs based on the user-provided context

ContextMMIA takes a set of keywords from the user to specify the context of the experiment. Currently, the most widely used biomedical literature database, PubMed, contains over 26 millions records. When I perform a search with the keyword 'cancer', over 3 million records are retrieved. Thus, I believe that this literature database contains enough articles to rank miRNA-gene pairs in terms of the user-provided context. However, there are two major issues in ranking miRNA-gene pairs: given the keywords, relevant papers should be identified and relevant gene names and miRNA names should also be identified. Since not all papers contain the user-provided keywords, it is necessary to infer the relevance of the words to extract genes and miRNAs in the relevant articles. To address this issue, I use BEST to identify relevant words and genes/miRNAs (Lee *et al.*, 2016). BEST has predefined biomedical entities for each category, such as drug, pathway, gene, and disease, and then it identifies relevant entities extracted from PubMed articles from the user query. For example, it returns entities such as 'ERBB2', 'wnt signaling pathway', and 'tamoxifen' with the keyword 'breast cancer' as an input. BEST has its own scoring system for entities, which is very useful in ranking gene-miRNA pairs with respect to the user-provided keywords. For example, there are keywords ‘breast cancer’ and entities ‘cell cycle’, ‘mir-200c’, ‘BRCA1’, and ‘ESR’. At the beginning, BEST compiles PubMed articles containing ‘breast cancer’ and the four entities in the abstract. Then, it measures the score and the rank for each entity and lists entities ordered by score. After compiling articles containing ‘BRCA1’ and ‘breast cancer’, BEST calculates a document score for each article and sums the score to measure the entity score, which is denoted as $BEST(BreastCancer, BRCA1)$. In this study, I use BEST to measure the relevance of each miRNA and mRNA for a given user query.

2.4.2 Omics Score

The **omics score** (OS) is the probability of a gene-miRNA contributing to the class difference when expression data are analyzed. The OS is based on the general principle that differentially expressed miRNA targets genes differentially, resulting in negative correlations between genes and miRNA; then, differentially expressed gene explains the phenotype differences. ContextMMIA computes the **omics score** based on a strategy similar to MMIA. It measures miRNA differential scores, mRNA differential scores, and then correlation scores. The DEmiRNAs and DEmRNAs can be determined by MMIA. After the DEmRNAs and DEmiRNAs are determined, the probability of miRNA-mRNA contributing to the class difference is calculated. Let the p-values of miRNA and mRNA be p_{m_i} and p_{g_j}, respectively. For miRNA m_i, m_i’s differential score $\mathit{diff}(m_i)$ is defined by Eq 2.1, and its normalization $\mathit{diff}_n(m_i)$ is defined by Eq 2.2.

$$\mathit{diff}(m_i) = -log_2(p_{m_i}) \tag{2.1}$$

$$\mathit{diff}_n(m_i) = \frac{\mathit{diff}(m_i) - \min(\mathit{diff})}{\max(\mathit{diff}) - \min(\mathit{diff})} \tag{2.2}$$

The calculation of diff_n for mRNA is similar to that of miRNA. The range of diff_n is between 0 and 1 by Eq 2.2. If miRNA is significantly differentially expressed in a given condition, then the value of diff_n will be close to 1.

Correlation score is defined by measuring the Pearson's correlation coefficient of the miRNA-mRNA pair's logarithmic expression as in (Mukherji *et al.*, 2011). ContextMMIA considers only negatively correlated miRNA-mRNA pairs; thus, a negative value of the coefficient is defined as the correlation score as in Eq 2.3.

$$corr(m_i, g_j) = -pearson\_correlation(m_i, g_j) \tag{2.3}$$

The **omics score** of miRNA-mRNA $OS(m_i, g_j)$ is defined in Eq 2.4.

$$OS(m_i, g_j) = \mathit{diff}_n(m_i) * corr(m_i, g_j) * \mathit{diff}_n(g_j) \tag{2.4}$$

By definition, $OS(m_i, g_j) \in$ [0,1]; thus, a value of OS close to 1 means that the miRNA and mRNA are both significantly differentially expressed and anticorrelated. Thus, I predict that the pair is related to the phenotype difference with a high confidence in terms of expression data.

2.4.3 Context Score

I defined the context score (CS) to measure the probability of a miRNA-mRNA pair contributing to the phenotype difference in terms of the literature information. As described in the previous section, BEST estimates a score between predefined entities and keywords. I denoted the user-input keyword as k, which is context specified by the user (e.g., disease, gene, pathway, and so forth). As shown in Eq 2.5, $CS(m_i, g_j|k)$ measures the significance of the m_i-g_j pair for k in terms of the literature information.

$$CS(m_i, g_j|k) = P(m_i|k) * P(g_j|k) \tag{2.5}$$

To compute $P(m_i|k)$, I used Bayes' rule and transformed $P(m_i|k)$ into Eq 2.6 because BEST only measures the score for predefined entities and does not support undefined keywords (e.g., broad keyword, new drug or pathway, and so on) (Lee *et al.*, 2016).

$$P(m_i|k) = \frac{P_n(k|m_i) * P_n(m_i)}{\sum\limits_{l=1}^{p} P_n(k|m_l) * P_n(m_l)} \tag{2.6}$$

By converting $P(m_i|k)$ using Bayes' rule, my method provides the user with a freeform keyword environment, which allows the user to easily utilize my system even when the user is not familiar with biological terms.

$$P(k|m_i) = log_2(BEST(k, m_i) + 1) \tag{2.7}$$

The literature significance of miRNA (m_i) for a given keyword k, $P(k|m_i)$, is computed as shown in Eq 2.7. $BEST(k, m_i)$ is the score of m_i for k computed by BEST, and I converted the scale of the score by taking the logarithm of the BEST score. For example, assume that the keyword 'immune system' and the miRNA 'miR-155' are used in an analysis. If the relation between 'miR-155' and 'immune system' is well studied, then $P(\,immune\ system\,|\,miR155)$ and $BEST(\,immune\,system\,,\,miR155)$ will have a high score.

$$P(m_i) = log_2(BEST(m_i, m_i) + 1) \tag{2.8}$$

Eq 2.8 describes how to compute $P(m_i)$, which denotes how much literature information exists for m_i; the more that papers report m_i, the higher the value it will have. After computing $P(m_i)$ and $P(k|m_i)$, normalization terms $P_n(m_i)$ and $P_n(k|m_i)$ are defined by the min-max normalization.

$P(m_i|k)$ is computed using Bayes' rule and specifies the significance of m_i given the literature domain k, and the value of $P(m_i|k)$ has a correlation with the amount of studies, i.e., the number of papers about m_i in domain k. For mRNA g_j, $P(g_j|k)$ is computed in a similar way, and I measured the significance of the m_i-g_j pair in k by computing $CS(m_i, g_j|k)$ using $P(m_i|k)$ and $P(g_j|k)$.

2.4.4 Confidence Score

The confidence score of m_i, g_j and k is denoted as $Score(m_i, g_j, k)$, which is a confidence value of target prediction in terms of both expression and literature data.

$$Score(m_i, g_j, k) = OS(m_i, g_j) \ * \ CS(m_i, g_j \mid k) \tag{2.9}$$

Eq 2.9 can be interpreted as a weighted **omics score**, where the weight is determined by a probability of a m_i, g_j pair being true in terms of the user-provided context given keywords k.

2.5 Results

To evaluate ContextMMIA, I performed three experiments in comparison with four existing tools: MMIA, MAGIA2, CoSMic and GenMiR++. The three experiments were pathway analysis, reproducibility of validated miRNA targets in human, and sensitivity tests when different keywords were used for specifying the experimental context. I used 2-class microarray datasets containing miRNA and mRNA expression profiles in humans. GSE21411 (Cho *et al.*, 2011), GSE40059 (Luo *et al.*, 2013), and GSE53482 (Norfo *et al.*, 2014) from human disease studies were used. Each study reports experimentally validated miRNA and the correlated target mRNA pair, which was used to evaluate the miRNA target prediction methods in this section. A detailed description of each dataset is listed in Table 2.1.

Table 2.1 summarizes the validated target pair and the domain of the experimental design in each dataset. In the interstitial lung diseases (ILD) study, it was reported that ZEB-1 affects the persistence of disease in ILD through suppression of NEDD4L by miR-23a. In the GSE40059 breast cancer study, the authors investigated differences between aggressive breast cancer cell lines and less-aggressive cell lines and reported that CFL2 was up-regulated by miR-

Table 2.1: Each GEO study comes with an experimentally validated miRNA-mRNA target (the second column) to affect their disease domain (the third column). Disease information was used to test performances when different contexts are specified.

Data	Experimentally validated target	Disease
GSE21411	hsa-miR-23a - NEDD4L	Interstitial Lung Diseases
GSE40059	hsa-miR-200c - CFL2	Breast Cancer
GSE53482	hsa-miR-155 - JARID2	Primary Myelofibrosis

200c. The authors also reported that CFL2 expression was correlated with tumor grade. In the primary myelofibrosis (PMF) study, the authors revealed that overexpressed miR-155-5p regulates JARID2, and they suggested that regulated JARID2 may be related to MK hyperplasia in PMF. Disease information was used to test performances when different contexts are specified for ContextMMIA. It is necessary to choose keywords to specify contexts. 'Interstitial lung disease' and 'primary myelofibrosis' are too specific to use literature data; thus, I used the more general words 'lung disease' and 'myelofibrosis' as the keywords for ContextMMIA.

2.5.1 Pathway analysis

To evaluate the effectiveness of the approach used in ContextMMIA, I compared it with four expression-based methods: MMIA, MAGIA2, GenMiR++, and CoSMic. GenMiR++ computes probabilities for target pairs using an EM algorithm. MMIA extracts DEmiRNA to reduce the search space by a user-defined cutoff and finds negatively expressed target DEmRNAs. MAGIA2 provides several methods for the integrated analysis, and I chose Pearson's correlation method from among these methods. After measuring the correlation, MAGIA2 calculates the false discovery rate (FDR) for each target. CoSMic extracts an mRNA cluster for each miRNA and computes the significance of a cluster using permutation tests. Likewise, each algorithm uses a different strategy to predict the miRNA target and to reduce the search space. I used these four algorithms to compare performances in terms of the predictive power. The methods compute confidence values for the predicted miRNA and mRNA targets, typically probability or p-value. I ranked the prediction results in terms of the confidence values. In the experiments, I used a p-value cutoff of 0.1 for ContextMMIA. For MMIA, a p-value of 0.05 was used for both DEmiRNA and DEmRNA selection.

For the performance evaluation, I used the top 200 predicted miRNA-mRNA pairs predicted by each method. Then, I mapped genes included in the interacting pairs to human pathways using DAVID (Huang *et al.*, 2009a,b) to determine which pathways were significantly enriched. Among these pathways, I carefully selected pathways that are most likely related to the disease through the literature study as shown in Table 2.1. I set evaluation criteria as how these literature-guided pathways were predicted by each method.

Table 2.2 shows the ratios of the number of genes that are mapped to significantly enriched pathways to the number of genes included in the top 200 miRNA-target edges. The number of genes is less than 200 because the same gene was multiply targeted, e.g., miR-200c-BRCA1 and miR-23a-BRCA1.

As shown in Table 2.2, the number of genes mapped to the significantly enriched pathways is quite different for each method even though the number of genes does not considerably differ for each method. In terms of the ratio of mapped genes to predicted genes, ContextMMIA outperforms the existing methods 2 to 4 times. A gene set in a pathway means that genes have similar biological functions in terms of regulating molecular processes. Thus, the ratios in Table 2.2 indicate that ContextMMIA produces more functionally coherent

Table 2.2: The ratio of the mapped genes and the number of the genes in the top 200 miRNA-target pairs. From each method, I extracted the top 200 target pairs using each method and performed pathway analysis using DAVID. The numerator is the number of genes mapped to the enriched pathways, and the denominator is the genes in the top 200 edges. The ratio of ContextMMIA is the largest for each dataset.

Methods	GSE21411	GSE40059	GSE53482
ContextMMIA	37 / 79	45 / 157	42 / 127
MMIA	12 / 157	20 / 179	11 / 124
GenMiR++	0 / 194	18 / 197	26 / 200
MAGIA2	18 / 182	12 / 191	19 / 193
CoSMic	24 / 196	9 / 195	X

gene sets.

Table 2.3 lists pathways related to 'breast cancer' and enriched pathways predicted by each method for the GSE40059 dataset. The circles in Table 2.3 mean an enriched pathway when DAVID pathway analysis was performed by using genes in the top 200 edges. For example, if the ECM-receptor interaction is enriched in the ContextMMIA and GenMiR++ results, circles are marked in the context column and the second column for the corresponding tools. As shown in Table 2.3, more pathways related to 'breast cancer' were enriched in the gene sets produced by ContextMMIA than in the gene sets produced by the competing methods. In addition, several important pathways were enriched only in ContextMMIA. For example, it is well known that approximately half of breast tumors have stronger MAP kinase activity than the surrounding benign tissues (Santen *et al.*, 2002). Inflammation plays a pivotal role in tumor initiation, promotion, angiogenesis and metastasis. Cytokines are important

Table 2.3: Enriched pathway analysis on GSE40059 breast cancer data. Breast-cancer-related pathways are selected by the literature search. A circle in a cell means that the pathway is enriched by the gene set predicted by each method (A: ContextMMIA, B: MMIA, C: GenMiR++, D: MAGIA2, and E: CoSMic). More pathways are enriched by the gene set in the ContextMMIA result.

Breast-Cancer-Related Pathway	A	B	C	D	E
Purine metabolism (Schramm *et al.*, 2010)		O			
Pyrimidine metabolism (Sigoillot *et al.*, 2004)		O			
ABC transporters (Fletcher *et al.*, 2010)			O		
MAPK signaling pathway (Santen *et al.*, 2002)	O				
Cytokine-cytokine receptor interaction (Esquivel-Velázquez *et al.*, 2015)	O				
Neuroactive ligand-receptor interaction (Finak *et al.*, 2008)			O		
p53 signaling pathway (Gasco *et al.*, 2002)	O	O			
Apoptosis (Lipponen, 1999)		O			
Notch signaling pathway (Reedijk, 2012)				O	
TGF-beta signaling pathway (Moses and Barcellos-Hoff, 2011)	O				
Axon guidance (Mehlen *et al.*, 2011)				O	
Focal adhesion (McLean *et al.*, 2005)	O	O		O	
ECM-receptor interaction (Lu *et al.*, 2012)		O			
Cell adhesion molecules (CAMs) (Saadatmand *et al.*, 2013)	O		O		
Adherens junction (Haidari *et al.*, 2013)	O				
Regulation of actin cytoskeleton (Flamini *et al.*, 2009)	O				
Glioma (Piccirilli *et al.*, 2005)	O				
Melanoma (Goggins *et al.*, 2004)	O				

in all the phenomena, and it has been reported that cytokines participate in regulating both induction and protection in breast cancer (Esquivel-Velázquez *et al.*, 2015). In addition, many studies have reported that TGF-beta signaling is critically important in the regulation of breast cancer (Moses and Barcellos-Hoff, 2011). High focal adhesion kinase expression is known to be related to aggressive breast cancer phenotypes (Lark *et al.*, 2005). Furthermore, cell adhesion molecules (CAMs) have a strong relationship with the process of metastasis, which is an important feature in predicting breast cancer prognosis (Saadatmand *et al.*, 2013). Moreover, a study revealed that activated leukocyte cell adhesion molecule (ALCAM) expression has a correlation with clinical outcomes such as grade, TNM stage, and NPI (King *et al.*, 2004).

2.5.2 Reproducibility of validated targets in humans

Table 2.4 shows the rankings of experimentally validated targets among the targets predicted by each method. Because ContextMMIA computes the context score using the literature data for given keywords, there is a possibility that the original papers of the datasets can affect the context score. Thus, I penalized the validated targets to compute $P(k|m_i)$ by excluding each paper when the BEST tool measures a score $BEST(k, m_i)$.

As shown in Table 2.4, ContextMMIA outperformed the other expression-based methods even though the penalized score is used. MMIA took the second place in reproducing the validated targets, but it ranked validated targets much lower than ContextMMIA. Although not rejecting the validated targets, GenmiR++ ranked validated targets very low. This result shows that GenmiR++ produced too many false positives for the three datasets. MAGIA2 failed to identify the validated targets as positive target pairs in any datasets because none of the validated target pairs satisfied the statistical cutoff. CoSMic also failed to identify the validated target pairs for two datasets, GSE21411 and

Table 2.4: This table contains the rankings of validated target pairs in three datasets. The validated targets are listed in the second column of Table I. ContextMMIA outperformed existing tools in predicting the validated targets. MAGIA2 and CoSMic failed to reproduce the validated targets.

Methods	GSE21411	GSE40059	GSE53482
Context-MMIA	**481**	**338**	**21**
MMIA	1411	387	1465
GenMiR++	8625	1673	95492
MAGIA2	X	X	X
CoSMic	X	X	X (Not Work)

GSE40059. In addition, CoSMic did not run successfully for dataset GSE53482 due to an input error issue. Many tools were not successful in reproducing validated targets, which can be an indication of false negatives.

To further confirm the reproducibility of my algorithm, we investigated how many experimentally verified targets in humans are detected in the top 200 miRNA-mRNA pairs by each of the methods. Experimentally validated human miRNA-mRNA pairs were extracted from miRTarBase (Hsu *et al.*, 2011), which curated experimentally validated miRNA-target interactions (MTI) by reporter assay, western blot, microarray, and next-generation sequencing experiments. I used human functional MTIs with strong evidence for functionality in humans as true interacting pairs. Table 2.5 summarizes the number of validated targets in the top 200 miRNA-mRNA pairs predicted by each method. As shown in table 2.5, ContextMMIA predicted two to five times more validated targets compared to the existing methods. ContextMMIA predicted more than 10% of the experimentally validated MTIs in humans, with is a considerably higher prediction accuracy than existing methods. It suggests

Table 2.5: This table contains the number of validated target pairs in three datasets. The validated targets are extracted from miRTarBase target pairs filtered by human functional miRNA target interaction (MTI).

Methods	GSE21411	GSE40059	GSE53482
ContextMMIA	**27**	**38**	**24**
MMIA	5	4	12
GenMiR++	3	4	3
MAGIA2	0	0	0
CoSMic	7	0	X (Not Work)

that ContextMMIA may provide good candidates for further experimental validation.

2.5.3 Sensitivity tests when different keywords are used

The performance of ContextMMIA depends on how the keywords to specify context are related to the goal of the experiment. In addition to disease-related keywords, I performed experiments using less-relevant keywords such as insulin resistance, influenzas, HIV and hepatocellular carcinoma. The results of ContextMMIA using less-relevant keywords are presented in Table 2.6.

The relevant keywords for the three datasets are listed in the third column of Table 2.1. As shown in Table 2.6, the rankings of the validated pairs were considerably higher when the keywords that reflect experimental designs were used. This result indicates that my method is able to reflect the degree of relevance to the experimental design and capture the different miRNA-mRNA pairs when different keywords were used. In summary, the experiments with irrelevant keywords showed that my method can capture the miRNA-mRNA pairs, reflecting the user-specified biological context.

Table 2.6: Sensitivity tests when different keywords are used. Rankings of validated targets are shown when different keywords are used. The validated targets had high ranks when disease-related keywords were used.

Keyword	GSE21411	GSE40059	GSE53482
Correct keyword	**481**	**338**	**21**
Insulin resistance	12479	2036	4250
Influenzas	6826	1169	1623
HIV	5865	4002	3238
Hepatocellular carcinoma	5278	3265	7180

2.6 Summary

I presented ContextMMIA, a human-specific miRNA-mRNA target pair prediction system that utilizes both expression profiles and the literature information from the user-specified experimental design goals. A major contribution of my system is that I handled the false positives and false negatives, which are an inherent issue in expression-based prediction tools, by incorporating the user-specified context information from the literature. Analyses on three independent human datasets showed that ContextMMIA can capture the true positive miRNA-mRNA target pairs that are specific to a biological context. ContextMMIA outperformed existing tools in a series of experiments, such as pathway analysis, validated target ranking, and irrelevant keyword experiments.

I emphasize that computational predictions of miRNA-mRNA target pairs should be further validated in biological experiments and that my system is intended to provide good candidates for experimental validation. ContextMMIA is available at http://biohealth.snu.ac.kr/software/contextMMIA

Chapter 3

DRIM: A web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration

Pharmacogenomics is the study of how genes affect a person's response to drugs. Thus, understanding the effect of drug at the molecular level can be helpful in both drug discovery and personalized medicine. Over the years, transcriptome data upon drug treatment has been collected and several databases compiled before drug treatment cancer cell multi-omics data with drug sensitivity (IC_{50}, AUC) or time-series transcriptomic data after drug treatment. However, analyzing transcriptome data upon drug treatment is challenging since more than 20,000 genes interact in complex ways. In addition, due to the difficulty of both time-series analysis and multi-omics integration, current methods can hardly perform analysis of databases with different data-characteristics. One effective way is to interpret transcriptome data in terms of well-characterized biological pathways. Another way is to leverage state-of-the-art methods for multi-omics data integration.

In this study, I developed Drug Response analysis Integrating Multi-omics and time-series data (DRIM), an integrative multi-omics and time-series data analysis framework that identifies perturbed sub-pathways and regulation mechanisms upon drug treatment. The system takes drug name and cell line identification numbers or user's drug control/treat time-series gene expression data as input. Then, analysis of multi-omics data upon drug treatment is performed in two perspectives. For the multi-omics perspective analysis, IC_{50}-related multi-omics potential mediator genes are determined by embedding multi-omics data to gene-centric vector space using a tensor decomposition method and an autoencoder deep learning model. Then, perturbed pathway analysis of potential mediator genes is performed. For the time-series perspective analysis, time-varying perturbed sub-pathways upon drug treatment are constructed. Additionally, a network involving transcription factors (TFs), multi-omics potential mediator genes, and perturbed sub-pathways is constructed and paths to perturbed pathways from TFs are determined by an influence maximization method.

To demonstrate the utility of my system, I provide analysis results of sub-pathway regulatory mechanisms in breast cancer cell lines of different drug sensitivity. DRIM is available at http://biohealth.snu.ac.kr/software/DRIM/.

3.1 Computational Problem & Evaluation criterion

This chapter describes the method for solving the drug response prediction problem. The input data is gene-centric multi-omics data of samples including mutation, gene expression, copy number variation, DNA methylation, and drug-treated time-series gene expression data measured at different doses and time points. The computational problem is to predict the gene interaction relationships representing cell line-specific drug response. Challenges of this problem are (1) the multi-omics relationship of 20,000 genes representing drug response states is too complex, (2) determining gene-gene interactions from time-series data is complex. To evaluate the proposed method, I investigated whether mediator genes are associated with drug response and whether distinct drug response pathways exist for each cell line.

3.2 Related works

The variability in drug responses among cells is a major challenge in cancer drug therapy, thus personalized drug response research is much needed (Sweeney, 1983). With the recent advances in instrument technologies, drug response analysis at the molecular level has become possible, thus there is an opportunity to investigate relationship between drug response phenotypes and corresponding molecular data, e.g., multi-omics data upon drug treatment. Large-scale drug-response genomics data help identify molecular markers related with therapeutic response (Garnett *et al.*, 2012). Furthermore, more than 100 FDA-approved drugs have been developed from rapidly growing pharmacogenomics studies. This shows that pharmacogenomics data could be used for drug development at various stages, from drug targets to patient therapeutics. Moreover, genomics data of the patient can be regarded as a predictive factor for drug response. It can be thought of as an early response signal before the phenotypic change of cells by drug (Surendiran *et al.*, 2008).

Current pharmacogenomics data analysis can be extended in two directions to broaden the understanding of drug response. The first direction is to perform a pathway-level analysis. Analyzing drug responses at the individual gene level is difficult to explain biological variability and also difficult to interpret gene-drug associations (Wang *et al.*, 2019). Thus, focus of pharmacogenomics research is changing to investigate multiple gene products at the biological pathway level (Weinshilboum and Wang, 2004). A recent study shows that analysis of transcriptome data can be effectively done at the pathway level, which facilitates clear biological interpretation (Lim *et al.*, 2020). The second direction is to perform multi-omics level analysis. Recently, precision medicine studies have been conducted at the multi-omics level, which is called "pharmaco-omics" beyond pharmacogenomics by integrating genomics, proteomics, epigenomics, and metabolomics data (Adam and Aliferis, 2019; Ginsburg *et al.*, 2019). Many studies have shown that multi-omics integration helps unravel complex biological mechanisms (Subramanian *et al.*, 2020). Integrative analysis of multi-omics data can help understand cell line-specific gene regulation mechanisms for pathway activation (Kim *et al.*, 2016; Oh *et al.*, 2020b) and it can be used as a signature for drug response sub-pathway identification (Xu *et al.*, 2019). Single omics analysis can detect only a smaller subset, but multi-omics analysis can detect more comprehensive pathways that respond to chemical exposure (Canzler *et al.*, 2020).

There are several pharmacogenomics databases such as Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio *et al.*, 2016), Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012), Patient-Derived Xenograft (PDX) mice models (Gao *et al.*, 2015) and NCI-60 Human Tumor Cell Lines Screen (Abaan *et al.*, 2013). These databases can be used for cell line-specific drug sensitivity analysis with multi-omics signature at the molecular level. In addition, data from after drug treatment time-series experiments can be used to capture time-varying cell line-specific drug response as signature of cell death, proliferation and drug resistance. The Library of Integrated Network-based Cellular Signatures (LINCS) L-1000 (Subramanian *et al.*, 2017) project measures cell viability upon genetic and chemical perturbations by 978 landmark genes. Another database compiled time-series transcriptome data using the NCI-60 cell line upon anti-cancer drug treatment. (Monks *et al.*, 2018).

There are several databases that enable computational pharmacogenomics study. GDSC measured the response of 988 cell-lines to 518 drug-compounds (Iorio *et al.*, 2016). It provides mutation, copy number variation, DNA methylation, and gene expression data of cell lines before drug treatment. CCLE (Barretina *et al.*, 2012) measured genomics profiles and response to 24 anticancer drugs in 947 cell lines. A recent study (Ghandi *et al.*, 2019) performed RNA sequencing (RNA-seq), whole-exome sequencing (WES), whole-genome sequencing (WGS), reverse-phase protein array (RPPA), reduced representation bisulfite sequencing (RRBS), microRNA expression profiling, histone modification profiling, metabolites profiling (Li *et al.*, 2019), and 1,448 drugs response (Corsello *et al.*, 2020) for CCLE cell lines. NCI-60 cell lines are the most widely studied cell lines in human cancer research. CellMiner (Reinhold *et al.*, 2012) is a website that provides 20,503 chemical compounds response of NCI-60 cells and also genomics data before drug treatment as mutation, DNA methylation, microRNA expression, gene expression, and protein data. The NCI Transcriptional Pharmacodynamics Workbench (NCI TPW) (Monks *et al.*, 2018) provides time-series pharmacogenomics data and a web page that allows data exploration. They measured gene-expression changes the NCI-60 cell line after drug exposure 2h, 6h, and 24h to 15 anticancer drugs. NCI-DREAM community (Bansal *et al.*, 2014) measured gene-expression changes the OCI-LY3 cell line after 14 anticancer drug treatment 6h, 12h, and 24h to predict the activity of pairs of compounds.

By utilizing pharmacogenomics data in various databases, a number of studies have been performed to analyze pharmacogenomics data in terms of IC50 prediction, drug response gene/pathway identification. Table 3.1 summarizes pharmacogenomics data analysis methods. Multi-Omics Late Integration (MOLI) (Sharifi-Noghabi *et al.*, 2019) is an end-to-end deep neural network-based drug response prediction method. MOLI takes mutation, copy number,

Table 3.1: Pharmacogenomics data analysis methods, their input, output, and algorithms.

Method	Input	Output	Algorithm
MOLI	Multi-omics data	Drug response (IC50)	Deep learning
DSPLMF	Multi-omics data, Chemical structures	Drug response (IC50)	Logistic Matrix Factorization
CancerDAP	Multi-omics data	Sub-pathway signatures	Random forest, Logistic regression
DryNetMC	Drug-treated time-series gene expression data	Clinically relevant genes	Differential network analysis

and gene expression as input, and predicts drug response using each omics type-specific encoder. Drug Sensitivity Prediction using a novel regularization approach in Logistic Matrix Factorization (DSPLMF) (Emdadi and Eslahchi, 2020) is a drug response prediction method based on recommender systems. DSPLMF takes cell line similarity matrix consisted of gene, copy number, mutation, and IC50 and drug similarity matrix as input, and predict drug response using matrix factorization and nearest neighbor algorithm. CancerDAP (Xu *et al.*, 2019) is a pipeline that integrates gene expression, copy number variation, and DNA methylation to identify sub-pathway signature of anticancer drug response. The user can browse drug active sub-pathway using CancerDAP webpage. Differential regulatory Network-based Modeling and Characterization (DryNetMC) (Zhang *et al.*, 2019) is a network-based algorithm to detect key cancer resistance genes based on time-series RNA-seq data. DryNetMC uses time-series RNA-seq data after drug treatment as input. From the data, it constructs drug-sensitive network and drug-resistant network utilizing ordinary differential equations and extracts differential network. Using differential network, a node importance is measured by topology, entropy, and gene expression changes to prioritize genes of clinical relevance.

Lv et al (Lv *et al.*, 2018), performed an analysis of differentially expressed genes on hepatocellular carcinoma (HCC) patients for drug discovery from gene expression data. They divided HCC patients into two groups: high/low-PKM2 to investigate the effect of pyruvate kinase isozymes M2 (PKM2) gene expression on HCC patients in terms of metabolic changes and prognosis. The study identified metabolic genes related to poor HCC patient survival and screened drugs that target metabolic enzymes associated with poor survival. Some of the screened drugs have been used in antitumor clinical studies. Another study proposed a tensor decomposition-based drug discovery method for neurological disorder from gene expression data (Taguchi and Turki, 2019). They selected genes through tensor decomposition-based feature extraction using mouse Alzheimer's single-cell RNA-seq data. These genes are significantly overlapped with the target genes of Alzheimer's disease drugs. Recently, a deep learning-based generative model (Méndez-Lucio *et al.*, 2020) proposed to design active-like molecules from gene expression signatures. The generative model takes the desired gene expression profile induced by drug-treatment or gene knock-out experiment as input. The study generates a molecular representation that is likely to have caused a change in gene expression.

3.3 Motivation

To utilize rapidly accumulating drug response omics data, many computational methods for drug response prediction have been developed. Machine learning methods are often used to process high-dimensional genomics data, such as matrix-factorization models (Wang *et al.*, 2017; Brouwer and Lió, 2017), network-based models (Zhang *et al.*, 2015, 2018) and deep learning models

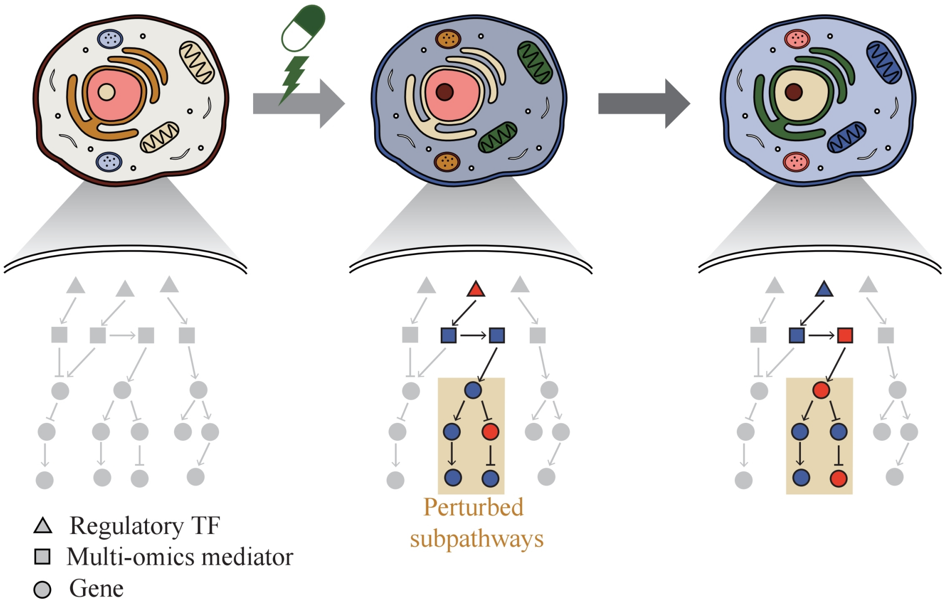


Figure 3.1: Phenotypic change of cell over time by drug. DRIM makes it possible to interpret drug response at molecular level by investigating perturbed sub-pathways.

(Sharifi-Noghabi *et al.*, 2019; Baptista *et al.*, 2020). Moreover, analysis methods for time series omics data have been developed (Jo *et al.*, 2016; Ahn *et al.*, 2019; Kim *et al.*, 2019; Kang *et al.*, 2019). However, utilizing these tools for the analysis of pharmacogenomics databases requires expert-level bioinformatics skill.

Thus, a web-based system called Drug Response analysis Integrating Multi-omics and time-series data (DRIM), was developed and presented in this study by integrating condition-specific multi-omics data to investigate temporal drug response at the molecular level. The condition of the sample can be defined as a combination of three variables that are cell-line type, drug type, and drug dose. DRIM aims to identify perturbed sub-pathways and regulatory mechanisms upon drug treatment using an integrative analysis framework on both multi-omics and time-series data. By simply taking drug name and cell lines or user's drug control/treat time-series gene expression data as input, DRIM performs the analysis in two perspectives. First, IC_{50}-related multi-omics potential mediator genes are chosen by embedding multi-omics data into gene-centric vector space using either a tensor decomposition or an autoencoder deep learning model. The tensor decomposition does not require pre-training to determine relationship among different omics components. Feature space from tensor decomposition is linear combination of input features, thus it is easy to interpret how the feature space combines input features. On the other hand, the autoencoder can learn non-linear relationship of multi-omics data. Autoencoder requires pre-training but it can generate a feature space dynamically for new incoming multi-omics data. In terms of computation time, tensor decomposition is faster that the autoencoder. Then, the potential mediator genes are extended to the identification of perturbed pathways upon drug treatment over time. This time-series analysis construct a network containing transcription factors (TFs), multi-omics mediator genes, and perturbed sub-pathways by an influence maximization based method.

To demonstrate the utility of my system, I provide analysis results of sub-pathway regulatory mechanisms in breast cancer cell lines of different breast cancer drug sensitivity.

3.4 Methods

The system workflow is illustrated in Figure 3.2. In Step 1, The user selects a drug and cell lines to be analyzed for perturbed pathway analysis or uploads their drug control/treat time-series gene expression data. In Step 2, through time-series gene expression data analysis after drug treatment, perturbed sub-pathways are identified. In Step 3, multi-omics potential mediator genes are selected by multi-omics integration methods. In Step 4, a time-bounded network is constructed and the most regulatory path is identified by influence maximization. In Step 5, the system visualizes networks involving TF, mediator genes, and perturbed sub-pathways that change over time upon drug treatment. A detailed description of each step in the workflow is below.

3.4.1 Step 1: Input

The user selects a drug and cell lines to be analyzed for perturbed pathway analysis or uploads their own drug control/treat time-series gene expression data. The system uses two time-series gene expression after drug treatment databases LINCS L-1000 and NCI-60. In both databases, there are control and treated data for drugs per cell line. For each condition, the gene expression was measured at each time point. These databases are available as GSE70138 and GSE116438 in GEO.

3.4.2 Step 2: Identifying perturbed sub-pathway with time-series

Step 2 is for identifying perturbed sub-pathways of differentially expressed genes that are defined using a time-series data analysis tool, TimeTP (Jo *et al.*, 2016). First, each pathway is represented as a directed graph from the KEGG pathway database. For each node in the pathway, the system assigns a time vector $\vec{v}$ of 1 (overexpressed) or -1 (underexpressed) and 0 (unchanged) that are defined by comparing gene expression levels, treated vs. control. `Limma` (Smyth, 2005; Ritchie *et al.*, 2015) was used to define differentially expressed genes (DEGs) at each time point with Robust Multiarray Average (RMA) normalization (Kupfer *et al.*, 2012). When there is no control sample, differential expression genes are defined by comparing either to the expression level of the previous time point or to the expression level of initial time point. Second, a

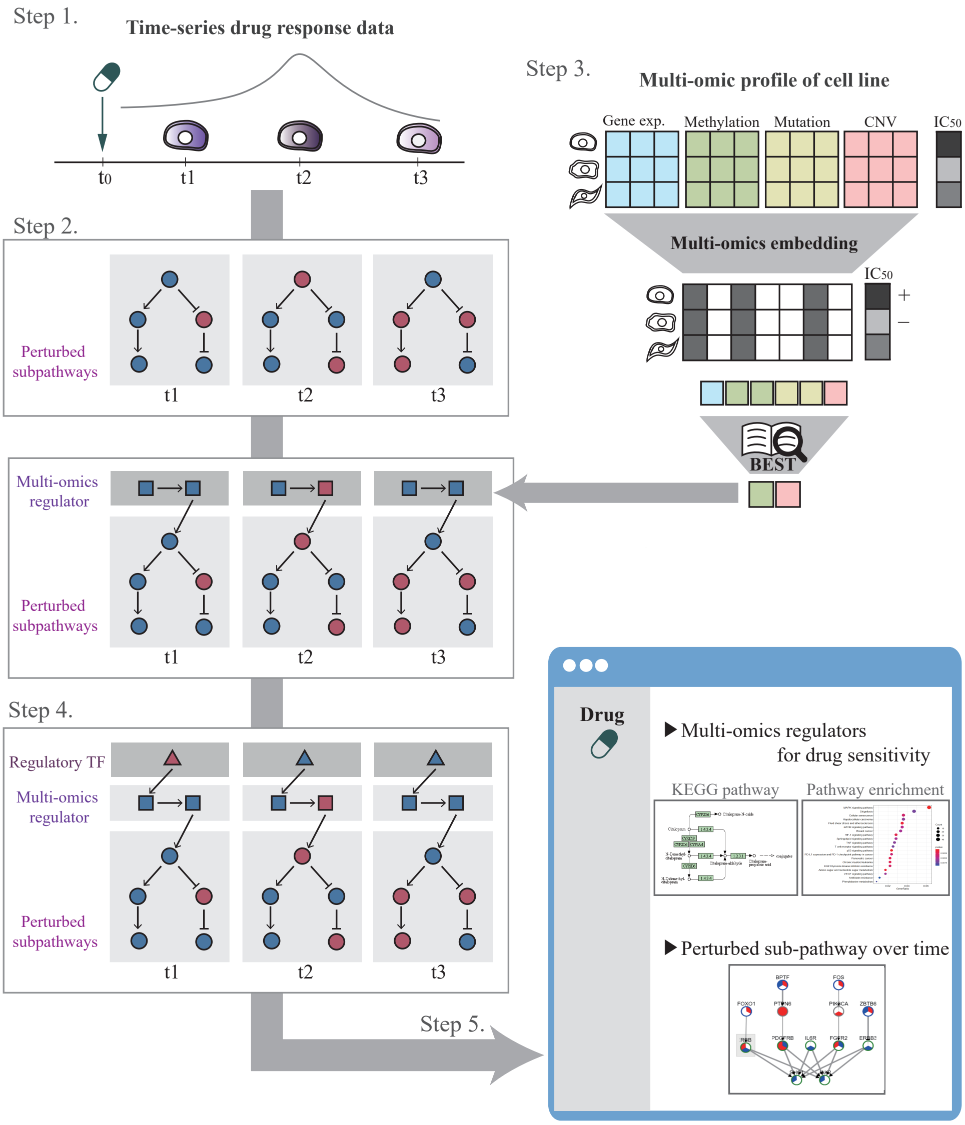


Figure 3.2: The systematic workflow of DRIM. Step 1 is for drug and cell line selection. Step 2 is for perturbed sub-pathway identification using expression propagation. Step 3 is for selecting multi-omics potential mediator genes by multi-omics embedding methods. Step 4 is for constructing time-bound network and determining regulatory path by influence maximization. Step 5 is to visualize the analysis result.

perturbed sub-pathway is determined by choosing valid edges in the pathway graph. Assume that there is an edge $N1 \rightarrow N2$ between two genes, $N1$ and $N2$, that have differential time vectors $\vec{v_1}$ and $\vec{v_2}$. To measure the direction of propagation and the number of delayed time points between two vectors, cross-correlation is defined as

$$(\vec{v_1} \star \vec{v_2})(n) = \sum_{t=-\infty}^{\infty} \vec{v_1}(t)\vec{v_2}(t+n) \tag{3.1}$$

where $\vec{v}(t) = 0$ for $t \leq 0$ or $t > T$ (This happens at the preceding or trailing entries of two vectors). Cross-correlation is maximized when the two vectors overlap most with n delay.

$$d(\vec{v_1}, \vec{v_2}) =_n (\vec{v_1} \star \vec{v_2})(n) \tag{3.2}$$

If $d(\vec{v_1}, \vec{v_2})$ is negative, it means that the propagation direction is opposite to the given direction. The opposite edge is considered as invalid and excluded from the perturbed sub-pathway. When delay n is larger than a threshold value, the edge is filtered out. After choosing valid edges, a sub-graph that has more than two valid edges is determined as a perturbed sub-pathway. P-value of a perturbed sub-pathway is determined by permutation test. The null distribution is generated by randomly re-assigning differential expression vector for each gene in the sub-pathways. A sum of cross-correlations of edges is used as a pathway-level statistics and p-value for a perturbed sub-pathway is calculated from the null distribution.

3.4.3 Step 3: Embedding multi-omics for selecting potential mediator genes

Step 3 determines potential mediator genes related to drug sensitivity from the multi-omics regulation perspective. The system integrates four multi-omics data such as gene-expression, copy number variation, DNA methylation, and

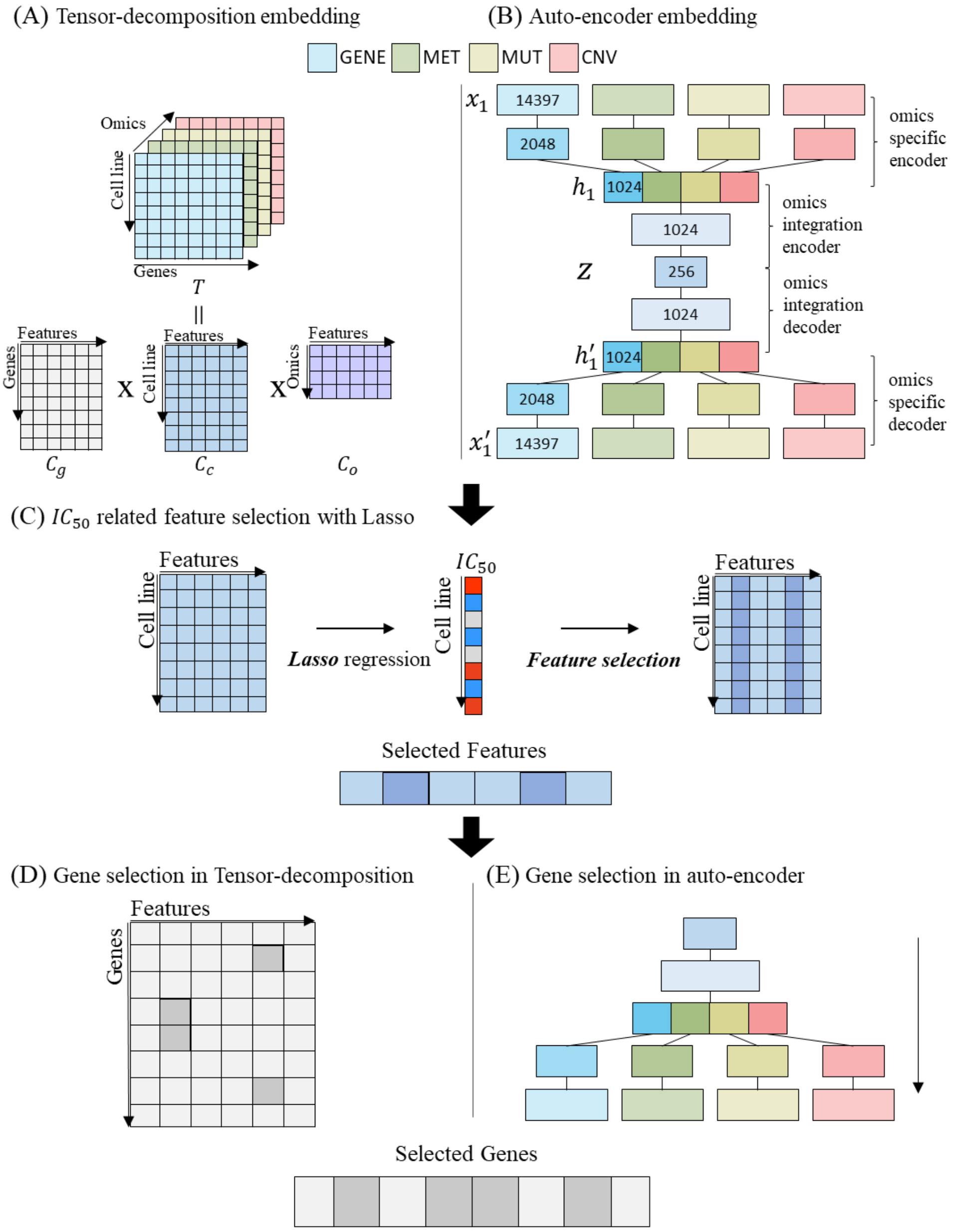


Figure 3.3: Multi-omics potential mediator gene selection. (A) multi-omics integration by tensor decomposition. (B) multi-omics integration by autoencoder. (C) IC_{50} related feature selection using Lasso regression with embedded feature matrix. (D) gene selection of tensor decomposition from selected features. (E) is gene selection of autoencoder from selected features.

mutation from the CCLE database. Each omics data processed to a gene-centric (*cell line* $\times$ *gene*) matrix to discover potential mediator genes from the perspective of multi-omics regulation. The gene expression and copy number variation values were normalized by min-max normalization. The mutation data was binarized to 1 if mutations exist in the gene or 0 otherwise. To process methylation data, methylation levels of probes located within the transcription start site and 1KB upstream of promoter regions were averaged per gene. The IC_{50} value measured for each cell line is used as the drug response phenotype.

The system uses two machine learning algorithms, a tensor decomposition method and an autoencoder method, to embed high dimensional multi-omics data to low dimensional feature space. The embedding of the multi-omics data is to create a "gene-centric" feature space, which means that regulation information, such as copy number variation, DNA methylation, and mutation, is tied to a gene while embedding multi-omics data.

Figure 3.3.(A) and 3.3.(B) illustrate the process of embedding gene-centric multi-omics data with two algorithms. For tensor decomposition, I used the PARAFAC model that decomposes a tensor into three two-dimensional matrices (Rabanser *et al.*, 2017). As shown in Figure 3.3.(A), tensor T with elements x_{ijk} composed of *cell line* $\times$ *gene* $\times$ *omics* matrix and is factorized three matrix C_g,C_c and C_o with g_{if}, c_{jf} and o_{kf}. C_g, C_c and C_o are defined as gene, cell line and omics components, respectively. $f = 1, ..., R$, R is the number of features.

$$x_{ijk} = \sum_{f=1}^{R} g_{if} c_{jf} o_{kf} + e_{ijk} \tag{3.3}$$

I used C_c matrix that embeds cell line-specific multi-omics relationship.

Figure 3.3.(B) describes the process of autoencoder embedding that is unsupervised artificial neural network to learn efficient encoded representation of data (Kramer, 1991). I constructed a late-integration autoencoder that encodes gene-centric multi-omics data. An input vector is represented as $x = (x_1, ..., x_n, x_{n+1}, ..., x_{2n}, x_{2n+1}, ..., x_{3n}, x_{3n+1}, ..., x_{4n})$ that is a concatenation of four multi-omics values and n is the number of genes. An autoencoder is to reconstruct x' as output for an input vector x. For each layer l, I used $relu$ as activation function between input layer x and output layer y.

$$y = f_l(x) = relu(W_l x + b_l) \tag{3.4}$$

The autoencoder consists of four system components: an omics-specific encoder, an omics-integration encoder, an omics-integration decoder, and an omics-specific decoder. In the omics-specific encoder, features are learned individually for each omics data. For each omics data of x_i with $i = (1, 2, 3, 4)$, x_i is encoded to h_i.

$$h_i = F^k(x) = f_k \circ ... \circ f_1(x) \tag{3.5}$$

Where k is the number of layer, $f_k \circ f_{k-1}(x) = f_k(f_{k-1}(x))$ is the composition function of f. The omics-integration encoder learns relationship among multi-omics data using concatenated omics features $h = (h_1, h_2, h_3, h_4)$ and encodes h to z in a similar way to Eq 3.5. z is an embedding vector that learns the regulation of multi-omics relationship. The omics-integration decoder decodes z to h'. The omics-specific decoder decodes omics specific h'_i to x'_i and reconstruct input $x' = (x'_1, x'_2, x'_3, x'_4)$ in the opposite way to the encoder. For each encoder and decoder, I used 2 layers and 2048, 1024 hidden neurons in the omics specific layers, 1024, 256 hidden neurons in the omics integration layers. I used $Mean\ Squared\ error\ (MSE)\ loss$ as a loss function with $L2$ regularization on the weight vector such as Eq 3.6.

$$Loss = \sum_{i=1}^{N} \frac{1}{N}(x_i - x'_i)^2 + \lambda * \sum_{i=1}^{P} |w_i| \tag{3.6}$$

N is the number of data, P is the number of layer, w_i is the weight of ith layer. Figure 3.3.(C) illustrates the feature selection process, using C_c matrix by tensor decomposition or z vector by autoencoder multi-omics embedding matrix. Least Absolute Shrinkage and Selection Operator (LASSO)-regression model (Tibshirani, 1996) is constructed using IC_{50} as a target value. Features with non-zero coefficients in regression are considered as features that are significantly associated with the IC_{50} value.

Figure 3.3.(D) and 3.3.(E) depict the gene selection step related to associated features from Figure 3.3.(C). In Figure 3.3.(D), tensor decomposition method using C_g matrix is for gene selection. For each gene, the row-wise *argmax* operation can be used to obtain the feature most related to the gene, and if the feature is among the IC_{50} related features obtained in the previous step (features whose coefficients are large in Lasso regression), the gene is selected. The product of $C_g(g, f)$ and $coef(f)$ is defined as the omics score of the gene, where $coef(f)$ is the coefficient of $f'th$ feature in Lasso regression.

The autoencoder method uses decoder part for gene selection in Figure 3.3.(E). To evaluate features of a gene in terms of multi-omics, process selected feature in the decoder is activated and propagated to the omics data layer. Activation of the final layer is measured through the gene-wise summation and the omics score is computed. The significant genes related with the features are selected.

Selection of multi-omics potential mediator genes is done by combining the two scores, a condition-specific omics score and a literature-based score using BEST, a Biomedical Entity Search Tool (Lee *et al.*, 2016). When a drug name is submitted to the BEST system, genes that are known to be related to the drug are selected in a ranked list in the order of relevance to the drug. Combining the two scores is done by a method that was developed for microRNA and target gene interaction (Oh *et al.*, 2017).

3.4.4 Step 4: Construct TF-regulatory time-bounded network and identify regulatory path

Step 4 is for constructing TF-regulatory time-bounded network and determining regulatory paths. First, two networks are constructed to search upstream regulators of perturbed sub-pathway. A Gene Regulatory Network (GRN) is constructed from HTRIdb (Bovolenta *et al.*, 2012) for interaction information between TF and multi-omics potential mediator. A Protein-Interaction Network (PIN) is instantiated from STRING (Szklarczyk *et al.*, 2015) database for gene-gene interaction. To combine GRN, PIN, and perturbed sub-pathways as TF-regulatory time-bounded networks, I used the method described in Step 2.

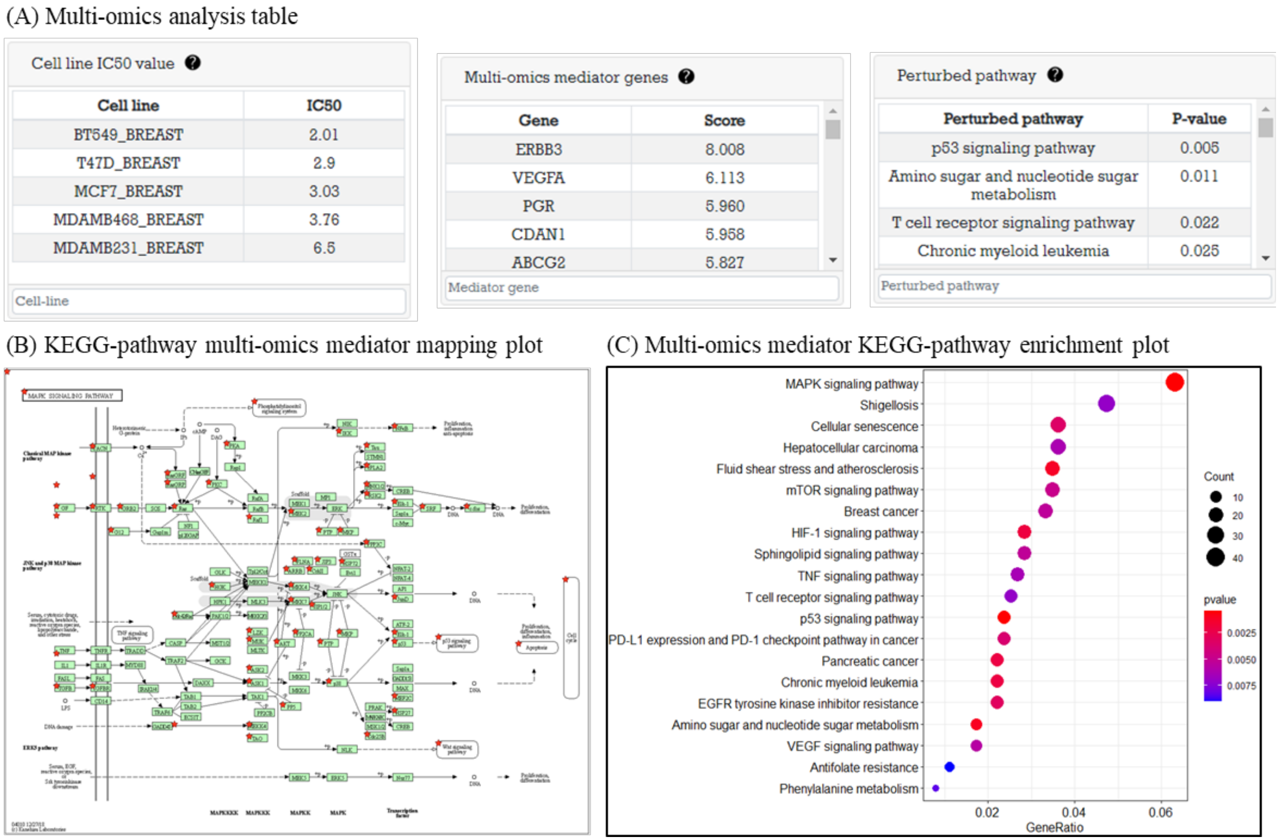
Next, the most likely regulatory paths are identified by the influence maximization method that has been widely used to select marketing targets in the social network to maximize the spread of influence (Kempe *et al.*, 2003). My system uses a labeled influence maximization algorithm (Li, 2011) to the time bounded network to identify most influential regulatory path from TF to perturbed sub-pathway (Jo *et al.*, 2016).

3.4.5 Step 5: Analysis result on the web

The system provides analysis results on the web from two perspectives: multi-omics data before drug treatment and time-series gene expression data after drug treatment.

Multi-omics analysis result before drug treatment

In this part, system provides analysis results of multi-omics data before drug treatment. As an example, in Figure 3.4.(A), there are tables representing cell line IC_{50}, multi-omics potential mediator genes related to IC_{50} value, and perturbed pathways that are enriched by potential mediator genes. In Figure



(A) Multi-omics analysis table

Cell line IC50 value

Cell line	IC50
BT549_BREAST	2.01
T47D_BREAST	2.9
MCF7_BREAST	3.03
MDAMB468_BREAST	3.76
MDAMB231_BREAST	6.5

Cell-line

Multi-omics mediator genes

Gene	Score
ERBB3	8.008
VEGFA	6.113
PGR	5.960
CDAN1	5.958
ABCG2	5.827

Mediator gene

Perturbed pathway

Perturbed pathway	P-value
p53 signaling pathway	0.005
Amino sugar and nucleotide sugar metabolism	0.011
T cell receptor signaling pathway	0.022
Chronic myeloid leukemia	0.025

Perturbed pathway

(B) KEGG-pathway multi-omics mediator mapping plot

(C) Multi-omics mediator KEGG-pathway enrichment plot

Figure 3.4: Multi-omics data analysis result before drug treatment. (A) consists of three tables. Cell line with IC_{50} table, multi-omics potential mediator genes with score table, perturbed pathway with p-value table (B) is a perturbed pathway mapping to KEGG pathway. (C) is an enriched pathway dot plot.

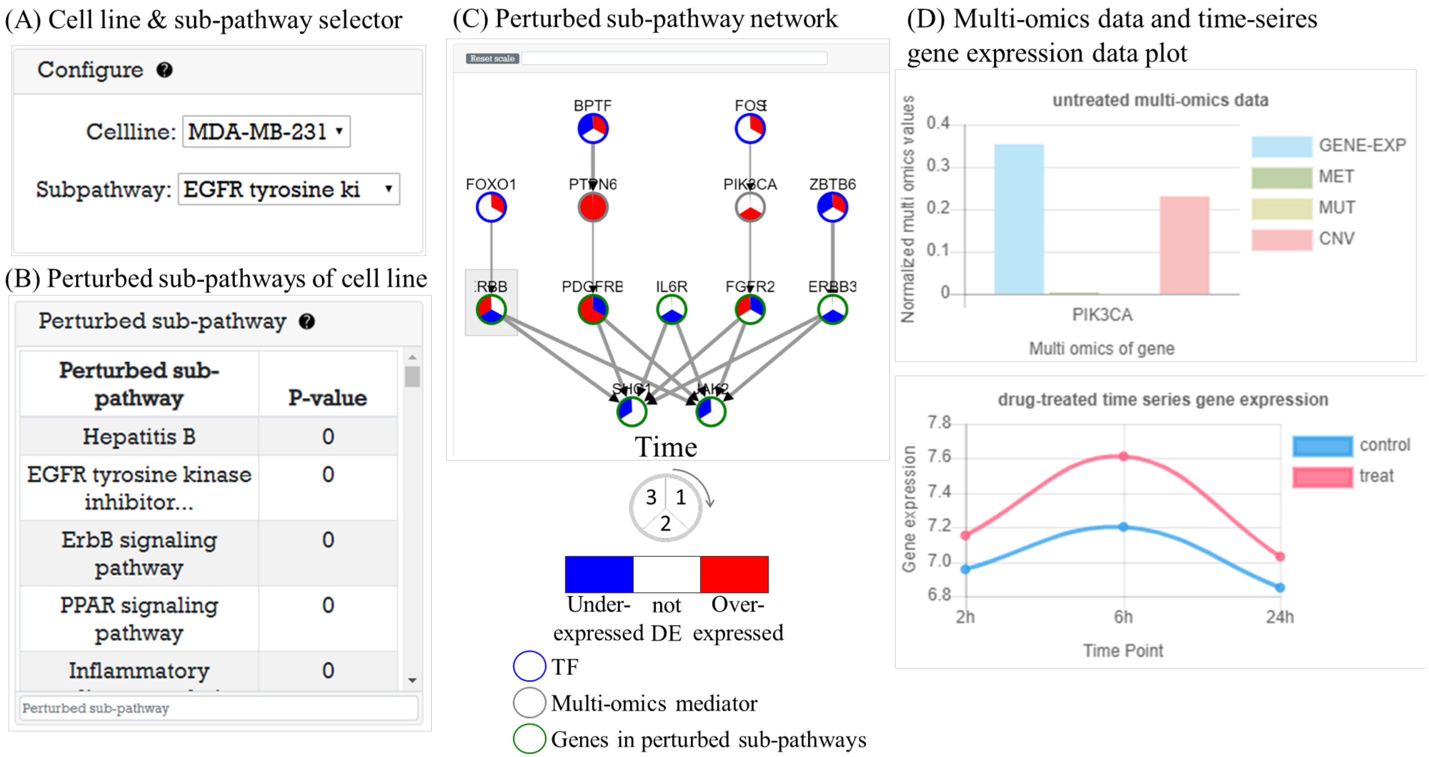
3.4.(B), when the user clicks on the pathway in the pathway table, a KEGG pathway plot is created. Figure 3.4.(C) is GO enrichment analysis plot of potential mediator genes to show the biological functions of the multi-omics potential mediator gene set in relation to drug sensitivity.

Time-series gene expression analysis result after drug treatment

This part provides time-series gene expression data after drug treatment analysis results with perturbed sub-pathways. As an example, in Figure 3.5.(A), user can select cell line and perturbed sub-pathway to explore. When the user select a cell line, a perturbed sub-pathway table (Figure 3.5.(B)) is generated with p-value. Figure 3.5.(C) is a TF-pathway network in time clock. When user clicks the gene node, a popup window appears to display multi-omics measurement of gene and expression plot of gene over time. Furthermore, the user can search genes in the network. The user can control the network size by choosing a cut-off value for DEGs to identify perturbed pathway. If the cutoff is low, the number of nodes edges increases, which may cause false positive problems. In the opposite case, there may be a false negative problem. In either case, predicted perturbed pathways are computationally predicted, thus the user may need to further investigate perturbed pathways.

3.5 Case study: Comparative analysis of breast cancer cell lines that have different sensitivity with lapatinib

To demonstrate the usefulness of DRIM, I conducted an analysis on breast cancer cell lines in response to `lapatinib` administration. The `lapatinib` is a dual inhibitor on both targets epidermal growth factor receptor (EGFR) and human epidermal growth factor receptor 2 (HER2) tyrosine kinases (Med-



(A) Cell line & sub-pathway selector

Configure

Cellline: MDA-MB-231

Subpathway: EGFR tyrosine ki

(B) Perturbed sub-pathways of cell line

Perturbed sub-pathway

Perturbed sub-pathway	P-value
Hepatitis B	0
EGFR tyrosine kinase inhibitor...	0
ErbB signaling pathway	0
PPAR signaling pathway	0
Inflammatory	0

Perturbed sub-pathway

(C) Perturbed sub-pathway network

(D) Multi-omics data and time-seires gene expression data plot

Figure 3.5: Time-series gene expression data analysis result after drug treatment. (A) is selector to visualize network of cell line. (B) is a perturbed sub-pathway table of cell line. (C) is visualized time varying network TF to perturbed sub-pathway. (D) is gene information window that contains time series gene expression plot and multi-omics data before drug treatment .

Table 3.2: Five breast cancer cell lines that are available multi-omics data before drug treatment with `lapatinib` sensitivity and time-series gene expression data after drug treatment.

Cell Line	Molecular Sub-subtype	$IC_{50}(\mu M)$
BT-549	Basal B	2.02
T-47D	Luminal	2.90
MCF7	Luminal	3.04
MDA-MB-468	Basal A	3.77
MDA-MB-231	Basal B	6.50

ina and Goodin, 2008). It was approved by US Food and Drug Administration (FDA) in combination therapy for HER2-positive/overexpressed breast cancer patients. I chose five representative breast cancer cell lines that have distinct sensitivity/resistance on `lapatinib` (Table 3.2). These cell lines are all available on both multi-omics and time-series data to fully utilize the nature of DRIM.

3.5.1 Multi-omics analysis result before drug treatment

For multi-omics analysis for before drug treatment cells, DRIM selected IC_{50}-related multi-omics potential mediator gene sets that are obtained by multi-omics integration analysis as shown in Figure 3.4. I carefully examined the set of candidate multi-omics potential mediator genes predictive of `lapatinib` sensitivity.

The top 15 multi-omics potential mediator genes `lapatinib` are shown in Table 3.3 sorted by their relevance score with respect to `lapatinib`. Among the genes, ERBB3 (HER3) was previously known for its critical role in HER2-amplified breast cancer cells (Lee-Hoeflich *et al.*, 2008). It is strongly associated

Table 3.3: Top 15 multi-omics potential mediator genes that related to lapatinib sensitivity.

ERBB3	CDAN1	CASP8	CNTN4	NF2
VEGFA	ABCG2	TP53	DCTN6	CBL
PGR	ESR1	MAP2K7	CD274	E2F1

with lapatinib sensitivity in coexpression with neuregulin-1 (NRG1) (Wilson *et al.*, 2011). Genetic perturbations on other genes such as ABCG2, TP53, and HSF1 were also well known for lapatinib resistance (Dai *et al.*, 2008; Rahko *et al.*, 2003; Yallowitz *et al.*, 2018).

3.5.2 Time-series gene expression analysis after drug treatment

For the temporal pharmacogenomic analysis, I investigated cell line-specific perturbed sub-pathways that may be related to different lapatinib response. The lapatinib mainly targets PI3K signaling pathway, which plays a critical role in cell growth, survival, and proliferation (Fruman *et al.*, 2017). Conceivably, aberrant activation of PI3K signaling is known to confer resistance to drugs targeting various receptor tyrosine kinases (Eichhorn *et al.*, 2008; Wang *et al.*, 2011). As expected, I collectively observed a significant time-course perturbation of PI3K signaling in each of the five cell lines in Table 3.4.

I further examined in detail if there are differential sub-pathway-level regulations among cell lines that mediate the response to the drug. Specifically, I asked whether each cell line harbors a distinct time-course regulatory path that governs the expression of a shared protein at the terminus of a pathway. To systematically identify such examples, I seeked for the regulatory paths with shared terminator protein for at least two cell lines using the “overview” network generated by DRIM. To simplify the analysis, I defined the terminator

Table 3.4: The p-value of PI3K-Akt signaling pathway in each cell line.

Cell Line	P-value
BT-549	1.6e-05
T-47D	6.07e-03
MCF7	7.11e-04
MDA-MB-468	4.94e-05
MDA-MB-231	1.08e-03

proteins as the nodes without outgoing edges in the network. Moreover, for biological interpretability, I only considered the paths starting from the transcription factors, and also enforced the paths to contain at least one multi-omics mediator. Different cell lines responded to `lapatinib`, accompanying distinct molecular perturbations, and shared the same terminal protein at the end of the paths (Figure 3.6).

Interestingly, I observed that many proteins involved in PI3K signaling pathway were regulated by different signaling pathways in a cell line-specific manner. For example, vascular endothelial growth factor A (VEGFA), a well-known effector molecule induced by PI3K signaling pathway (Karar and Maity, 2011), was shown to be activated through different signaling cascades, as shown in Figure 3.6.(A). In MDA-MB-468, VEGFA seemed to be induced by aryl hydrocarbon receptor (AhR) and aryl hydrocarbon receptor nuclear translocator (ARNT) signaling, presumably by the increased level of AhR/ARNT heterodimer as shown in Figure 3.6.(A). In BT-549 and T-47D cell lines, activation of JNK and NF-κB signaling were shown to be associated with increased level of VEGFA, respectively. Intriguingly, the time-bounded network allows the interpretation of the temporal difference of VEGFA induction between `lapatinib`-treated BT-549 and T-47D cell lines, as it can be deduced that the

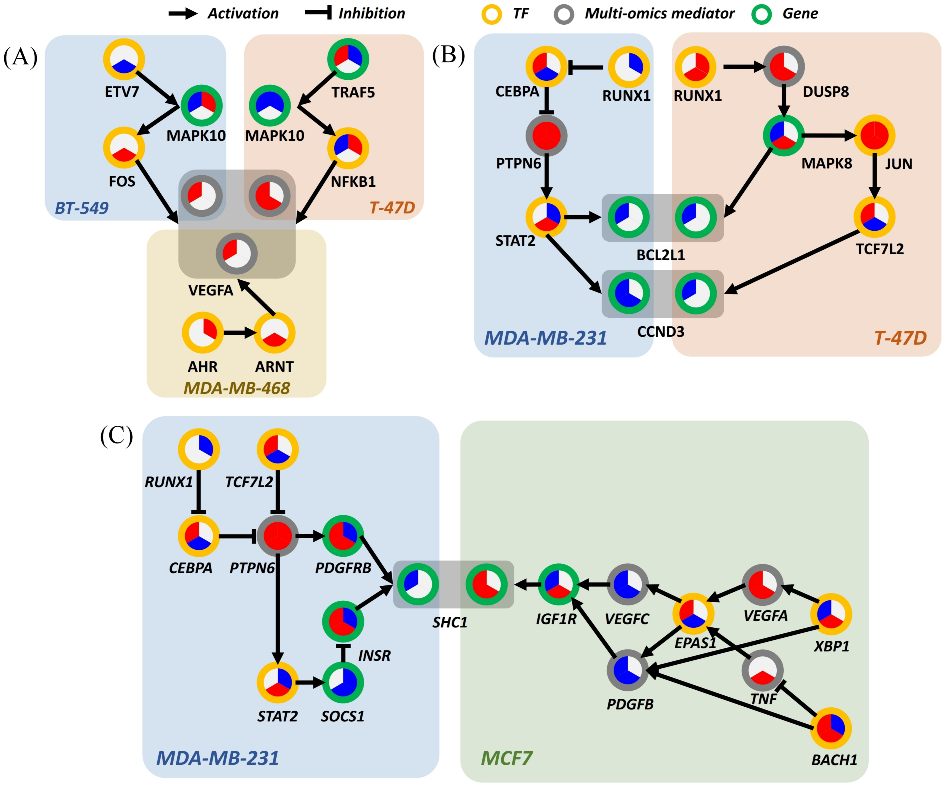


Figure 3.6: Differentially perturbed sub-pathway networks. (A) Regulatory path sharing VEGFA in BT-549, T-47D, and MDA-MB-468. (B) Regulatory path sharing CCND3, BCL2L1 in MDA-MB-231, T-47D. (C) Regulatory path sharing SHC1 in MDA-MB-231, MCF7.

earlier response of T-47D was due to the more rapid induction of NF-κB than that of FOS in BT-549.

Bcl-2-like protein 1 (BCL2L1) and cyclin D3 (CCND3), overexpressed in human breast cancer, are anti-apoptotic proteins that delay cell death and increases cell survival (Chi *et al.*, 2015; Simonian *et al.*, 1997). In Figure 3.6.(B) T-47D and MDA-MB-231 cell line, BCL2L1 and CCND3 are downregulated in response to `lapatinib` that leads to cell death. In T-47D, overexpression of JUN throughout whole phases is a prominent characteristic. JUN is a well-known transcription regulator that induces apoptotic cell death (Bossy-Wetzel *et al.*, 1997). It can be hypothesized that promoted cell death in response to `lapatinib` is attributed to the increased c-Jun.

Another interesting characteristic is that expression of downstream molecule of JUN—transcription factor 7-like 2 (TCF7L2)—increases over time, while T-47D cell line retained a high expression level of JUN. Since activity of c-Jun is predominantly regulated through phosphorylation, expression of molecules in regulatory relations should not be necessarily correlated. In MDA-MB-231, downregulation of BCL2L1 and CCND3 is induced by signal transducer and activator of transcription 2 (STAT2) (Furth, 2014), which involved in the JAK-STAT signaling pathway that leads to oncogenesis (Thomas *et al.*, 2015). Although temporal relations between molecules are not clear, it still gives insight into which pathways are involved in elevated cell death.

SHC-transforming protein 1 (SHC1), a core regulator of receptor tyrosine kinase signalling, is an essential gene for promoting immune suppression. Downstream effects of SHC1 perturbation lead to STAT3/STAT1-related immune impairment. As previously mentioned, SHC1 can respond to EGF stimulation using multiple paths of protein phosphorylations and interactions (Zheng *et al.*, 2013). There also exists PTPN12 as a turning point of SHC1 pTyr/Grb2 signaling that regulates cell invasion and morphogenesis. Reflecting the previous findings, my results on the perturbed sub-pathways can also show multiple regulatory mechanisms that each of the breast cancer cell lines can potentially utilize favorable/possible sub-paths on the temporal flow such as in Figure 3.6.(C). This implies that upstream stimuli including EGF regulation directs multiple paths of temporal information among breast cancer cell lines (Zheng *et al.*, 2013).

Even though all the five breast cancer cell lines were treated with the same drug, `lapatinib`, targeting receptors at cell surface with extreme specificity, each cell line showed different sensitivity to the drug. This heterogeneity may occur due to the complex crosstalk between various signaling pathways, which makes the inactivation of single signaling pathway by drug treatment not enough to cause systemic dysregulation of cellular machineries. Our system allows us to dissect this phenomena by differentially characterizing the fragments of regulatory cascades towards various effector molecules for each individual cell line as shown in Figure 3.6. Furthermore, DRIM provides intracellular mechanistic portraits of drug response for each of the cell lines, it may allow us to devise novel combination therapeutic strategies targeting additional molecules that cells depend on after the primary drug is applied.

3.6 Summary

For understanding the cell variability in drug response, personalized drug response analysis is demanded. In spite of increasing drug response genomics data, the interaction of high dimension multi-omics and time-series analysis are challenges for pharmacogenomics analysis. Pathway level analysis and multi-omics integration can be effective ways to interpret drug response data.

I developed an integrative multi-omics and time-series data analysis framework DRIM that finds perturbed sub-pathways and regulatory mechanisms in drug response. DRIM identifies the most likely regulatory path involving

TF, multi-omics mediator gene, and perturbed sub-pathway for each cell line. DRIM provides analysis results in two perspectives. As a demonstration, I conducted an analysis of breast cancer cell lines that have different `lapatinib` sensitivity. In the multi-omics perspective result, DRIM selected multi-omics potential mediator genes that are related to `lapatinib` resistance in previous studies. In the temporal pharmacogenomic analysis result, I showed that DRIM can be used to discover distinct temporal regulatory mechanisms governing the induction of several common downstream proteins across cell lines.

Chapter 4

Combinatorial modeling and optimization using iterative RL search for inferring sample-specific regulatory network

Understanding the cell state from the multi-omics data requires identifying the global gene regulation network described by *n-to-m* relationships between regulators and genes. In this study, I set the capacity of target genes for each regulator and the edge weight that is the contribution value of the regulator to the gene expression level from known regulator-gene interaction knowledge. From the capacity and edge weight, I formulated an objective function to measure the deviation between observed gene expression and estimated gene expression from the network represented by the *n-to-m* relationship. This optimization problem is a combinatorial optimization problem of which search space is $2^{n \times m}$. For exploring the search space, I leveraged the reinforcement learning (RL) framework to learn heuristics from data for complicated combinatorial problems. However, current RL frameworks are hard to handle the multi-omics relationship since the search space is too large. Thus, I proposed the iterative search method computing gene step and regulator step. In the gene-oriented step, the edges are added by selecting regulators using RL attention model. In the regulator-oriented step, the edges are removed by selecting genes using stochastic selection.

4.1 Computational Problem & Evaluation criterion

This chapter describes the method for *n-to-m* relationships between regulators and genes that estimate observed gene expression. The input data is miRNA, TF, and gene expression data of a single sample with regulator capacity and the value that edge changes the target gene expression level. The computational problem is to find the optimal network that estimates observed gene expression from the network. Challenges of this problem are (1) it is necessary to navigate a combinatorial search space of *n-to-m* relationships between the regulators and genes, (2) local optimization of each gene cannot finds a reasonable solution due to the constraint of regulator capacity. In order to evaluate the proposed method, a quantitative comparison experiment was performed on how accurately estimating gene expression and a qualitative comparison experiment was performed whether the network detects biologically meaningful relationships.

4.2 Related works

Biological processes, such as disease development and phenotype changes, are associated with gene expression patterns determined by the complex relationship between genetic molecules such as mutation, DNA methylation, miRNA, and histone modification (Oh *et al.*, 2020b). Recent technological developments have enabled scientists to measure different types of omics data representing cell state and help understand the complex biological mechanisms (Subramanian *et al.*, 2020). In order to interpret cell state from the multi-omics data, identification of global gene regulatory relationship helps to understand the current state of the cell. However, constructing a condition-specific regulatory network is to explore the search space of *n-to-m* relationships, which requires solving a combinatorial optimization problem that is often NP-complete or NP-hard. Traditional approaches rely on heuristics to solve combinatorial optimization problems using domain-specific expert knowledge that is inappropriate for other combinatorial problems.

With recent advances in deep neural networks, two approaches have been proposed to handle combinatorial optimization problems. One is neural directed acyclic graph (DAG) inference, (Zheng *et al.*, 2018) convert combinatorial DAG inference problem to continuous optimization problem by using a continuous function that represents acyclicity constraint. Various neural DAG inference methods were proposed and used on the different data domains (Vowels *et al.*, 2021). In addition, there is a study for constructing a gene regulation network from gene expression data using neural DAG inference. (Lee *et al.*, 2019). However, neural DAG inference methods are difficult to handle prior knowledge such as TF target database and microRNA target database. The other is reinforcement learning (RL) for learning heuristic, which can be interpreted as policy, to solve combinatorial optimization problems (Bello *et al.*, 2016). Recent RL-based methods for combinatorial problems generate reasonable solutions without handcrafted heuristics required expert domain knowledge (Mazyavkina *et al.*, 2020; Bengio *et al.*, 2020).

In this study, I formulated a combinatorial optimization problem to model a sample-specific regulator-gene network to determine the *n-to-m* relationship. Finding an optimal subset of the prior target network is hard to solve because of the enormous search space. Thus I proposed an iterative search method to determine the *n-to-m* relationship between regulators and genes that iteratively explores *n-to-one* gene-oriented relationship and *one-to-m* regulator-oriented relationship.

4.3 Motivation

In order to determine the explicit *n-to-m* relationship between regulators and genes, TF and miRNA were used as regulators that have been well studied its regulatory mechanisms at the transcriptional level and the post-transcriptional level (Chen and Rajewsky, 2007). I constructed a model of a regulator-gene network by assuming that the sum of regulator influences determines the deviation of gene expression. The regulator influence is determined by the deviation of regulator expression and edge weight between a regulator and a gene. Each regulator has the capacity of the number of target genes.

Figure 4.1 illustrates the example of modeling idea that G_1 is targeted by TF_1, miR_1, and miR_2 and the sum of influences determines upper deviation of G_1 expression. In order to reduce the search space and determine values for formulating an objective function, I utilized a regulator-gene regulatory network in the curated database that collects potential relationships between regulators and target genes such as transcription factor binding sites (TFBS) miRNA target database. I used RegNetwork (Liu *et al.*, 2015) as a prior network that collects 17 databases providing regulatory relationships between miRNA, TF and gene. Details of problem formulation and algorithm are described in the method section.

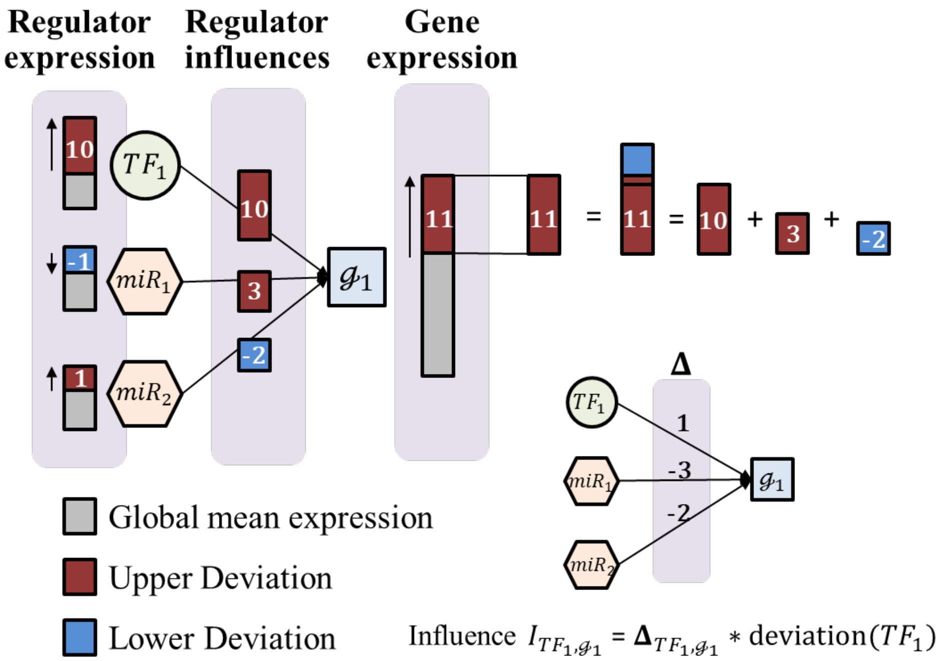


Figure 4.1: Illustration of modeling regulator-gene network. The red bar denotes an upper deviation of expression. The blue bar denotes a lower deviation of expression. Influence I is computed by multiplying the deviation of the regulator and $\boldsymbol{\delta}$ that is the amount of expression change of gene per regulator. G_1 is targeted by TF_1, miR_1, and miR_2 and its upper deviation is 11 that is sum of three regulator influences.

4.4 Methods

4.4.1 Formulating an objective function

In this section, an objective function is described to estimate gene regulatory network and the frequently used notations in this section are represented in Table 4.1.

I focused on constructing the regulator-gene network that estimates observed gene expression deviation induced by the sum of regulator influences. Let $\mathscr{R}$ is the set of regulators, $X_{\mathscr{R}}$ is the deviation of regulator expression, $\mathscr{G}$ is the set of genes, $X_{\mathscr{G}}$ is the deviation of gene expression, $\boldsymbol{I}$ is the influence matrix, and G_{prior} is the prior network. The deviation value is defined by the difference between the expression value and the mean value of expression.

The influence matrix is estimated by multiplication of the deviation of regulator expression the average rate of change between a regulator and a gene from expression profiles. For given $\mathscr{R}$, $X_{\mathscr{R}}$, $\mathscr{G}$, $X_{\mathscr{G}}$, $\boldsymbol{I}$, and G_{prior}, an objective function $F(G)$ is formulated based on Eq (4.1), Eq (4.2). A function $f(G)$ denotes an error term that is the sum of absolute error between observed gene expression and estimated gene expression from G, and $h(G)$ denotes a penalty term that is the regulator capacity constraint calculated by KL-divergence of regulator degree distribution between prior network and estimated network. The objective function $F(G)$ consists of the error term $f(G)$ and the penalty term $h(G)$.

$$\begin{aligned}
f(G) &= \frac{1}{|\mathscr{G}|}\sum_{\mathscr{g}\in\mathscr{G}} |x_{\mathscr{g}} - \sum_{\mathscr{r}\in\mathscr{N}^{-}(\mathscr{g})} \boldsymbol{I}_{\mathscr{r},\mathscr{g}}|) \\
h(G) &= \sum_{\mathscr{r}\in\mathscr{R}} \mathscr{P}_{prior}(\mathscr{r}) \log \frac{\mathscr{P}_{prior}(\mathscr{r})}{\mathscr{P}(\mathscr{r})} \\
\mathscr{P}(\mathscr{r}) &= \frac{|\mathscr{N}^{+}(\mathscr{r})|}{|E|}, \mathscr{P}_{prior}(\mathscr{r}) = \frac{|\mathscr{N}^{+}_{prior}(\mathscr{r})|}{|E_{prior}|}
\end{aligned} \tag{4.1}$$

Table 4.1: Frequently used notations.

Symbol	Description
$\mathscr{G}$	Set of genes $\{\mathscr{g}_1, \mathscr{g}_2, ..., \mathscr{g}_{\lvert\mathscr{G}\rvert}\}$
$\mathscr{R}$	Set of regulators $\{\mathscr{r}_1, \mathscr{r}_2, ..., \mathscr{r}_{\lvert\mathscr{R}\rvert}\}$
$X_{\mathscr{G}}$	Set of deviation of gene expressions $\{x_{\mathscr{g}_1}, x_{\mathscr{g}_2}, ..., x_{\mathscr{g}_{\lvert\mathscr{G}\rvert}}\}$
$X_{\mathscr{R}}$	Set of deviation of regulator expressions $\{x_{\mathscr{r}_1}, x_{\mathscr{r}_2}, ..., x_{\mathscr{r}_{\lvert\mathscr{R}\rvert}}\}$
$\boldsymbol{\Delta}$	Amount of gene expression change of per regulator expression
$\boldsymbol{I}$	Amount of gene expression change by regulator
s	A problem instance for RLRegSearchPerGene
$\boldsymbol{\theta}$	Model parameters of RLRegSearchPerGene.
$\boldsymbol{\pi}$	Set of regulators generated by RLRegSearchPerGene.
$L(\boldsymbol{\pi} \mid s)$	Error of the regulator set $\boldsymbol{\pi}$, given s.
$p(\boldsymbol{\pi} \mid s)$	A stochastic policy for selecting regulator set $\boldsymbol{\pi}$, given s.
$F(G)$	Fitness function value of graph G.
$G = (V,\, E)$	An unweighted directed graph
$\mathscr{N}^+(v)$	Set of out-neighborhood of a node v (departing in v)
$\mathscr{N}^-(v)$	Set of in-neighborhood of a node v (arriving in v)
$\mathscr{L}(\mathscr{g} \mid G)$	Error of gene $\mathscr{g}$, given G

$$\begin{aligned} &\min F(G) \\ &F(G) = f(G) + \lambda h(G) \ (\lambda > 0). \end{aligned} \tag{4.2}$$

Finding an optimal solution of the objective function $F(G)$ means finding a edge set between regulators and gene, which is challenging since the search space is combinatorial. In addition, gene-wise local minimization cannot find a reasonable solution. Thus, I proposed an iterative search algorithm to find a sample-specific regulatory network minimizing (4.2).

4.4.2 Overview of an iterative search method

In order to minimize the objective function $F(G)$, it is necessary to explore the search space of the *n-to-m* relationship between regulators and target genes. However, since the search space of the problem is $O(2^{|\mathscr{R}|\times|\mathscr{G}|})$, an exhaustive search cannot solve the problem. Thus, I proposed a two-step iterative edge search framework to find an appropriate solution within an acceptable time. Figure 4.2 and algorithm 1 shows the overview of the proposed search algorithm.

The first step is the gene-oriented step (G-step) to explore *n-to-one* relationships between regulators and each target gene. In G-step, edges are added to the network G using RL-based heuristics to reduce the value of $f(G)$ that is the absolute error between observed expression and estimated expression. The second step is the regulator-oriented step (R-step) to explore *one-to-m* relationships between each regulator and target genes. In R-step, edges are removed from the network G using stochastic selection to reduce the value of $h(G)$ that is the penalty term of capacity constraint. A detailed explanation of each step is described in the following section.

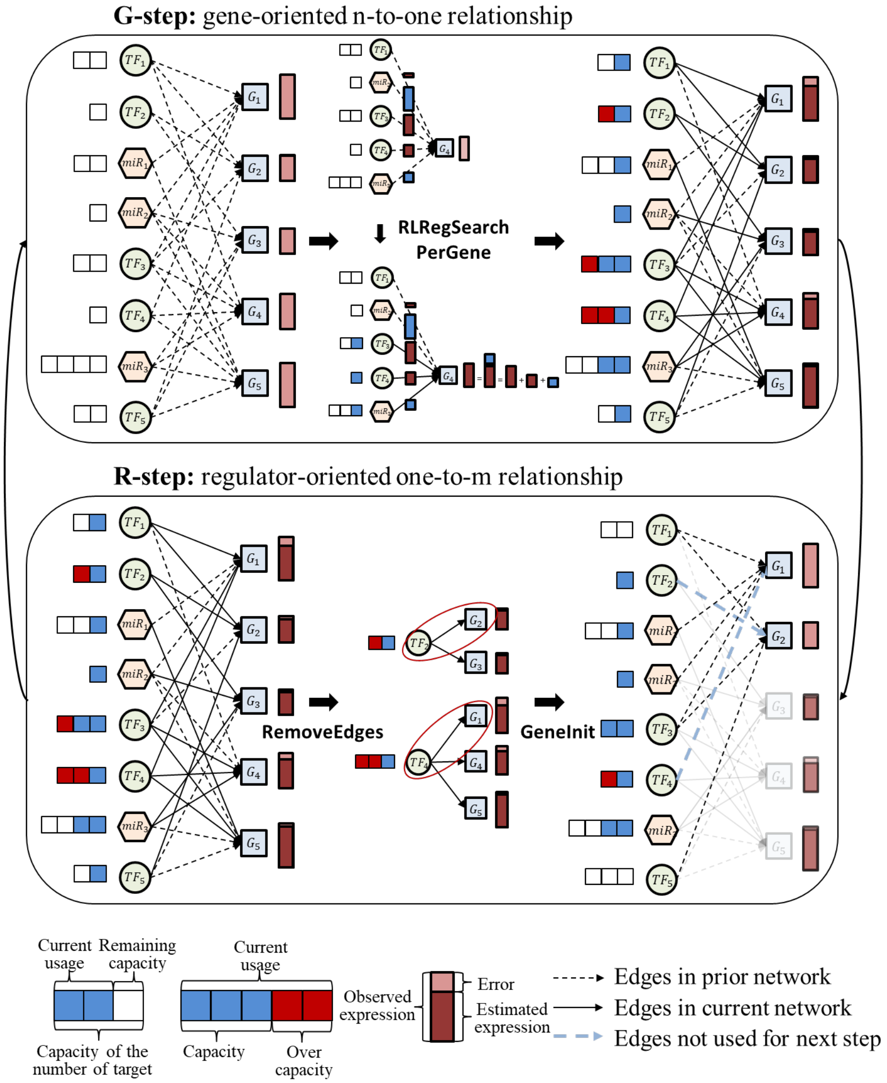


Figure 4.2: The overview of proposed search method that iteratively compute G-step and R-step. In G-step, exploring *n-to-one* relationship between regulators and target gene is performed by **RLRegSearchPerGene** for g_1, ..., g_5. In R-step, exploring *one-to-m* relationship between regulator and target genes is performed. **RemoveEdges** samples TF_2, g_2 and TF_4, g_1 that excess the capacity and removes. **GeneInit** initializes edges targeting g_1, g_2 to re-find regulators in next G-step with masking $TF_2 - g_2$, $TF_4 - g_1$ edges that are not used for next G-step.

Algorithm 1 Iterative search methods

```
**Input**: $\mathscr{R}$, $X_{\mathscr{R}}$, $\mathscr{G}$, $X_{\mathscr{G}}$, a pre-trained RL model parameter $\boldsymbol{\theta}$, a prior graph $G_{prior}$.
**Output**: Regulatory network $G$
$G = (V, E)$, $V = \mathscr{G} \cup \mathscr{R}$, $E = \emptyset$
**G-step** : Explore *n-to-one* gene-oriented relationship.
    **for** $\mathscr{g} \in \mathscr{G}$ **do**
        $G =$ **RLRegSearchPerGene**$(\mathscr{R}, \mathscr{g}, G_{prior}, \boldsymbol{\theta}, G)$
    **end for**

**while** not stop condition **do**
    $G_{cand} = G$
    **R-step** : Explore *one-to-m* regulator-oriented relationship.
        $E^{-} =$ **RemoveEdges**$(G_{prior}, G_{cand}, \mathscr{R})$
        // Remove edges exceeding the degree constraint.
        $\mathscr{G}_{init} =$ **GeneInit**$(E^{-}, G_{cand})$
        // Initialize edges targeting gene set in $E^{-}$.
    **G-step** : Explore *n-to-one* gene-oriented relationship.
        **for** $\mathscr{g} \in \mathscr{G}_{init}$ **do**
            $G_{cand} =$ **RLRegSearchPerGene**$(\mathscr{R}, \mathscr{g}, G_{prior}, \boldsymbol{\theta}, G_{cand})$
        **end for**
        //Re-compute G-step for $\mathscr{G}_{init}$ without $E^{-}$.
    **if** $f(G_{cand}) < f(G)$ **then**
        $G = G_{cand}$
    **end if**
**end while**
**return** best network $G$
```

4.4.3 G-step for exploring *n-to-one* gene-oriented relationship

Problem formulation of gene-oriented relationship

In this section, I describe G-step to explore the *n-to-one* relationship for minimizing $f(G)$. Given a problem instance s determined by states of $\mathscr{g}$, $\mathscr{R}$, $X_{\mathscr{R}}$, $x_{\mathscr{g}}$, and G_{prior}, the G-step finds regulator set $\boldsymbol{\pi}$, which minimizes a cost function $L(\boldsymbol{\pi} \mid s)$ that measures the difference between the deviation of observed expression of $\mathscr{g}$ and estimated expression determined by the sum of influences of regulators targeting $\mathscr{g}$. The $L(\boldsymbol{\pi} \mid s)$ is defined as Eq (4.3).

$$L(\boldsymbol{\pi} \mid s) = |x_g - \sum_{r \in \boldsymbol{\pi}} \boldsymbol{I}_{r,g}| \tag{4.3}$$

To minimize (4.3), I used RL model inspired by previous work (Kool *et al.*, 2018) that improved optimization performance of TSP using multi-head attention-based model. I leveraged the model architecture of the previous work (Kool *et al.*, 2018) for learning an optimal stochastic policy determining the optimal regulator set for the given problem instance s. The policy can be factorized with parameters $\boldsymbol{\theta}$ as Eq (4.4).

$$p_{\boldsymbol{\theta}}(\boldsymbol{\pi} \mid s) = \prod_{t=1}^{K} p_{\boldsymbol{\theta}}(\boldsymbol{\pi}_t \mid s, \boldsymbol{\pi}_{1:t-1}) \tag{4.4}$$

A encoder generates regulator embeddings from the influences of regulators. A decoder generates the sequence of regulators $\boldsymbol{\pi}$. The input of the decoder is regulator embeddings and current state representation that is the difference of observed gene expression and currently estimated gene expression at each time step. Figure 4.3 describes the model architecture and details of the model are described in the following section.

Encoder

A transformer-based encoder without positional encoding is used as an encoder architecture (Vaswani *et al.*, 2017). The encoder takes regulator influences as

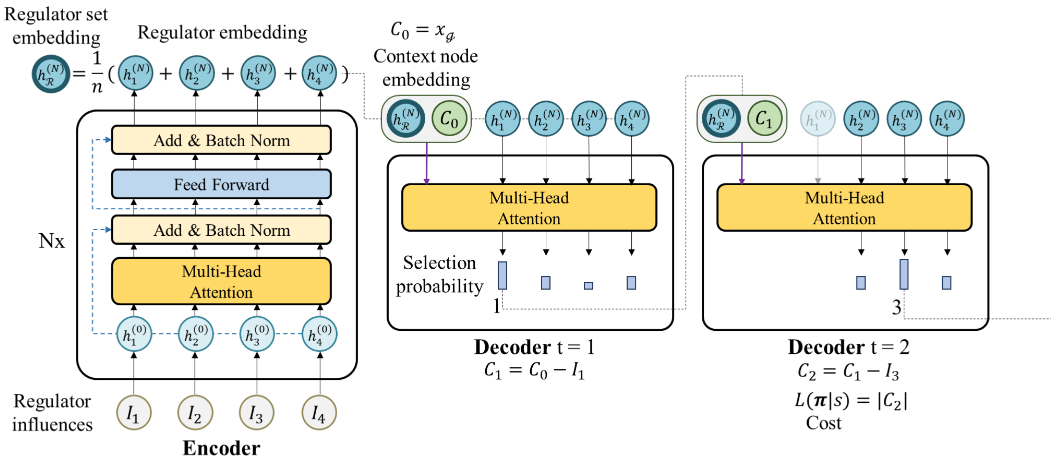


Figure 4.3: Attention-based encoder and decoder for regulator selection. The encoder takes regulator influences I_i as input to generate regulator embeddings $h_i^{(N)}$ using N times multi-head attention and feed-forward network. The regulator set embedding $h_{\mathscr{R}}^{(N)}$ is computed by the average of regulator embeddings $h_i^{(N)}$. The decoder takes the context node embedding and regulator embedding $h_i^{(N)}$ to generate the selection probability of regulator among the possible regulators for each step. The context node embedding is defined by concatenating $h_{\mathscr{R}}^{(N)}$ and C_t that is the cost in current t step. In the example, from the initial step $C_0 = x_{\mathscr{g}}$, decoder randomly selects the first regulator with selection probability and the cost change to C_1. If the step t reaches to the K that is the maximum number of selections (K=2 in this example), the regulator set $\boldsymbol{\pi} = \{1, 3\}$ is determined. The cost value $L(\boldsymbol{\pi} \mid s)$ is the absolute value of C_2.

input, $I_{\mathscr{r}_i,\mathscr{g}} = \mathbf{\Delta}_{\mathscr{r}_i,\mathscr{g}} * x_{\mathscr{r}_i}$, that is the amount of expression changes of $\mathscr{g}$ when regulator $\mathscr{r}_i$ is selected. From the one-dimensional regulator influences vector, the encoder computes regulator embeddings using N multi-head attention (MHA) layers and feed-forward (FF) layers with skip-connection and batch normalization (BN), which is a similar architecture to previous work (Kool *et al.*, 2018). Regulator embeddings h_i^l generated by layer $l \in \{1, .., N\}$, h_i^l is computed as Eq (4.7) and definition of MHA layer is descirebed as Eq (4.5) and Eq (4.6).

$$\begin{gathered} \mathbf{q}_i = W^Q\mathbf{h}_i,\ \mathbf{k}_i = W^K\mathbf{h}_i,\ \mathbf{v}_i = W^V\mathbf{h}_i \\ u_{ij} = \frac{\mathbf{q}_i^T\mathbf{k}_j}{\sqrt{d_k}},\ a_{ij} = \frac{e^{u_{ij}}}{\sum_{j'} e^{u_{ij'}}},\ \mathbf{h'}_i = \sum_j a_{ij}\mathbf{v}_j \end{gathered} \tag{4.5}$$

Eq (4.5) represents the attention mechanism. With dimensions d_k and d_v, the key $\mathbf{k}_i \in \mathbb{R}_k^d$, value $\mathbf{v}_i \in \mathbb{R}_v^d$, and query $\mathbf{q}_i \in \mathbb{R}_k^d$ are computed by the projection of $\mathbf{h}_i$ for each regulator. W^Q, W^K, and W^V are learning parameters. The compatibility $u_{ij} \in \mathbb{R}$ of the query $\mathbf{q}_i$ of regulator i with the key $\mathbf{k}_j$ of regulator j is computed as dot-product. The attention weights $a_{ij} \in \{0, 1\}$ is computed by softmax of compatibilities. The $\mathbf{h}'_i$ can be computed by the convex combination of value $\mathbf{v}_j$.

$$\mathrm{MHA}_i(\mathbf{h}_1, ..., \mathbf{h}_n) = \sum_{m=1}^{M} W_m^O \mathbf{h'}_{im} \tag{4.6}$$

The multi-head attention value of regulator i is a function of regulator embeddings $\mathbf{h}_1, \ldots, \mathbf{h}_n$ as Eq (4.6) using $\mathbf{h}'_{im}$ that is the output vectors of each attention with $M = 8$ times attention with different parameters and $d_k = d_v = \frac{d_h}{M} = 16$. Let $\mathbf{h}'_{im}$ the output vectors of each attention.

$$\begin{gathered} \mathbf{h}_i = \mathrm{BN}^l(\mathbf{h}_i^{(l-1)} + \mathrm{MHA}_i^l(\mathbf{h}_1^{(l-1)}, ..., \mathbf{h}_n^{(l-1)})) \\ \mathbf{h}_i^l = \mathrm{BN}^l(\mathbf{h}_i + \mathrm{FF}^l(\mathbf{h}_i)) \end{gathered} \tag{4.7}$$

The regulator embedding $\mathbf{h}_i^l$ in l-th layer can be computed with previous layer embedding $\mathbf{h}_i^{l-1}$, the MHA layer, the FF layer with dimension $d_{ff} = 512$, the rectified linear unit (ReLU) layer, and the BN layer.

Decoder

A decoder sequentially generates a regulator π_t at each time step $t \in \{1, .., K\}$ based on regulator embeddings and the previous selected regulators $\pi_{t'}$ at the time step $t' < t$. The decoder uses a context embedding vector $\mathbf{h}_c^N$ representing a current state. The context embedding $\mathbf{h}_c^N$ is defined by concatenating regulator set embedding $\bar{\mathbf{h}}^N$ and difference between observed gene expression $x_{\mathscr{g}}$ and currently estimated gene expression $\hat{x}_{\mathscr{g};(t-1)}$ at the time step t like (4.8).

$$\mathbf{h}_c^N = \begin{cases} [\bar{\mathbf{h}}^N, x_{\mathscr{g}} - \hat{x}_{\mathscr{g};(t-1)}] & t > 1 \\ [\bar{\mathbf{h}}^N, x_{\mathscr{g}}] & t = 1. \end{cases} \tag{4.8}$$

The decoder generates $p_{\boldsymbol{\theta}}(\boldsymbol{\pi}_t \mid s, \boldsymbol{\pi}_{1:t-1})$ that is the selection probabilities of regulator at the time step t form the context embedding $\mathbf{h}_c^N$ and regulator embeddings $\mathbf{h}_i^N$. The decoder uses a multi-head attention layer similar to encoder but only computes a single query $\mathbf{q}_c$ from the context embedding like Eq (4.9).

$$\mathbf{q}_c = W^Q \mathbf{h}_c^N,\ \mathbf{k}_i = W^K \mathbf{h}_i^N,\ \mathbf{v}_i = W^V \mathbf{h}_i^N \tag{4.9}$$

The compatibility u_{cj} of the query can be computed by Eq (4.10) with clipping to make the result within $[-C, C](C = 10)$ using *tanh*. If a regulator was selected in previous time step, it is masked ($u_{cj} = -\infty$).

$$u_{cj} = \begin{cases} C \cdot \tanh(\frac{\mathbf{q}_c^T \mathbf{k}_j}{\sqrt{d_k}}) & \text{if } j \neq \pi_{t'}, \forall t' < t \\ -\infty & \text{otherwise} \end{cases} \tag{4.10}$$

The compatibility u_{cj} can be interpreted as unnormalized log-probabilities. The output probability $\mathbf{p}$ can be computed by softmax as Eq (4.11).

$$p_i = p_{\boldsymbol{\theta}}(\pi_t = i|s, \boldsymbol{\pi}_{1:t-1}) = \frac{e^{u_{cj}}}{\sum_j e^{u_{cj}}}. \tag{4.11}$$

REINFORCE with baseline

In the previous section, the attention-based encoder-decoder model is presented, which generates a regulator set $\boldsymbol{\pi}$ for a given instance s. The policy gradient method is used to learn model parameter set $\boldsymbol{\theta}$ of regulator selection policy $p_{\boldsymbol{\theta}}(\boldsymbol{\pi}|s)$. A loss function is defined as $\mathscr{L}(\boldsymbol{\theta} \mid s) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{\pi}|s)}[L(\boldsymbol{\pi})]$ that is the expectation of the cost function $L(\boldsymbol{\pi})$. I optimized $\mathscr{L}$ using gradient descent with REINFORCE algorithm with baseline $b(s)$ as Eq (4.12) (Williams, 1992).

$$\nabla \mathscr{L}(\boldsymbol{\theta} \mid s) = \mathbb{E}_{p_{\boldsymbol{\theta}}(\boldsymbol{\pi}|s)}[(L(\boldsymbol{\pi}) - b(s))\nabla \mathrm{log} p_{\boldsymbol{\theta}}(\boldsymbol{\pi}|s)]. \tag{4.12}$$

A baseline $b(s)$ is used to reduce the variance of gradient and increases learning speed. In this study, a deterministic greedy rollout of the policy $p_{\boldsymbol{\theta}^{\mathrm{BL}}}$ that is the best model parameter on during training is used as a baseline policy (Kool *et al.*, 2018). At the end of every epoch, parameters of baseline policy $\boldsymbol{\theta}^{\mathrm{BL}}$ are updated to parameters of training policy if there is a significant improvement according to a paired t-test, on separate evaluation instances. I trained my model similar to (Kool *et al.*, 2018), but added a simple greedy selection policy $\boldsymbol{\pi}^{\mathrm{G}}$ that selects a regulator minimizing the cost at each step to provide reasonable baseline in the initial training step.

Adam optimizer (Kingma and Ba, 2014) was used for optimization. Algorithm 2 is the description of REINFORCE with baseline for model training.

The algorithm 3 describes G-step for given pre-trained model parameter $\boldsymbol{\theta}$ and a gene $\mathscr{g}$, which is called **RLRegSearchPerGene** that finds the best

Algorithm 2 REINFORCE with Baseline

```
Input number of epochs $E$, steps per epoch $T$, batch size $B$, significance $\alpha$
for epoch = $1, \ldots, E$ do
    for step = $1, \ldots, T$ do
        $s_i = \text{MakeInstance}(x_{\mathscr{g}i})\ \forall i \in \{1, \ldots, B\}$
        $\boldsymbol{\pi}_i = \text{SampleRollout}(s_i,\, p_{\boldsymbol{\theta}})\ \forall i \in \{1, \ldots, B\}$
        $\boldsymbol{\pi}_i^{\text{BL}} = \text{GreedyRollout}(s_i,\, p_{\boldsymbol{\theta}^{\text{BL}}})\ \forall i \in \{1, \ldots, B\}$
        $\boldsymbol{\pi}_i^{\text{G}} = \text{GreedySelection}(s_i)\ \forall i \in \{1, \ldots, B\}$
        $\nabla \mathscr{L} = \sum_{i=1}^{B} (L(\boldsymbol{\pi}_i) - \min(L(\boldsymbol{\pi}_i^{\text{BL}}), L(\boldsymbol{\pi}_i^{\text{G}})))\nabla_{\boldsymbol{\theta}} \text{log} p_{\boldsymbol{\theta}}(\boldsymbol{\pi}_i)$
        $\boldsymbol{\theta} = \text{Adam}(\boldsymbol{\theta}, \nabla \mathscr{L})$
    end for
    if OneSidedPairedTTest($p_{\boldsymbol{\theta}}, p_{\boldsymbol{\theta}^{\text{BL}}}$) $< \alpha$ then
        $\boldsymbol{\theta}^{\text{BL}} = \boldsymbol{\theta}$
    end if
end for
```

Algorithm 3 RLRegSearchPerGene

```
Input: set of regulators $\mathscr{R}$, a gene $\mathscr{g}$ to find regulators, a prior network $G_{prior}$, a pre-trained RL model parameter $\boldsymbol{\theta}$, a current network $G$
Output: Gene-centric regulator search network $G$
$s = \text{EncodeStateRL}(\boldsymbol{I}_{:,\mathscr{g}},\, X_{\mathscr{R}},\, x_{\mathscr{g}},\, G_{prior})$ // Encode input for RL
$\boldsymbol{\pi}_i = \text{SampleSolution}(p_{\boldsymbol{\theta}}(. \mid s))$; $i \in \{1 \text{ to } 5120\}$ // Sampling regulator sets
$\boldsymbol{\pi}_{min} = \text{argmin}(L(\pi_i \mid s))$ // Choose best solution
$E = E \cup \{(\mathscr{r}, \mathscr{g}) \mid \mathscr{r} \in \boldsymbol{\pi}_{min}\}$
return $G = (V,\, E)$
```

regulator set $\boldsymbol{\pi}$ among the sampling results and adds edges to current network G.

4.4.4 R-step for exploring *one-to-m* regulator-oriented relationship

In this section, I describe R-step to explore the *one-to-m* relationship for minimizing $h(G)$ by re-distributing the degree of regulators. In the R-step, edges of which regulator exceeds the capacity constraint are randomly removed from the current network G with entropy-based probability, which consists of two steps **RemoveEdges** and **GeneInit**.

Algorithm 4 RemoveEdges

```
Input: $G_{prior}$, a current network $G$, $\mathscr{R}$
Output: Removed edges $E^-$ that exceeds the regulator degree constraint.
for $\mathscr{r}$ in $\mathscr{R}$ do
    $\mathscr{P}(\mathscr{r}) = \frac{|\mathscr{N}^+(\mathscr{r})|}{|E|}, \mathscr{P}_{prior}(\mathscr{r}) = \frac{|\mathscr{N}^+_{prior}(\mathscr{r})|}{|E_{prior}|}$, $p$ = rand(0,1)
    //Fraction of out-degree and random prob $p$
    if $\mathscr{P}(\mathscr{r}) > \mathscr{P}_{prior}(\mathscr{r})$ and $p >$ entropy($\mathscr{P}(\mathscr{r}), \mathscr{P}_{prior}(\mathscr{r})$) then
        //Entropy-based selection of regulators that exceed capacity
        $\mathscr{g}$ = a sample with softmax$(\boldsymbol{f})_g$, $\boldsymbol{f} = \{f(g \mid G) \mid g \in \mathscr{N}^+(\mathscr{r})\}$
        //A sample $\mathscr{g}$ according to error among the target genes of $\mathscr{r}$
        $(\mathscr{r}, \mathscr{g})$ append to $E^-$
    end if
end for
$E = E \setminus E^-$ // Remove edges exceeding degree constraint.
return $E^-$
```

The **RemoveEdges** determines the edges E^- to remove from the current G by stochastic process as algorithm 4. For each regulator $\mathscr{r}$ with exceeding degree constraint, $\mathscr{r}$ is randomly sampled by probability based on the entropy between prior regulator degree and current regulator degree. For each sampled $\mathscr{r}$, a target gene $\mathscr{g}$ of $\mathscr{r}$ is also randomly selected among the current network G by softmax probability of target gene error $f(\mathscr{g} \mid G)$ that is the absolute error of a gene $\mathscr{g}$ in current network G.

Algorithm 5 GeneInit

```
Input: Removed edges $E^-$, a current graph $G$
Output: A gene set $\mathscr{G}_{init}$ for next G-step.
for $\mathscr{r}$, $\mathscr{g}$ in $E^-$ do
    $E = E \setminus \{(\mathscr{r}^*, \mathscr{g}) \mid \mathscr{r}^* \in \mathscr{N}^-(\mathscr{g})\}$
    //Initialize the edges targeting $\mathscr{g}$
    masking ($\mathscr{r}$, $\mathscr{g}$) edge
    //The masked edges are not used in next G-step.
    $\mathscr{g}$ append to $\mathscr{G}_{init}$
end for
return $\mathscr{G}_{init}$
```

The **GeneInit** removes selected edges E^- from the current G that is determined by **RemoveEdges**. Then, it revmoes edges including genes involved in E^- to re-determine of regulators of genes with masking E^- to exclude edges in the next G-step. Details of **GeneInit** is described in algorithm 5.

4.5 Results

4.5.1 Cancer cell line data

A multi-omics dataset of breast cancer cell lines was used for experiments from the Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012). There are 47 breast cancer cell lines in the CCLE database with miRNA, TF, and gene expression data. There are 895 regulators (miRNAs and TFs) and 15,141 genes for each cell line. For training the RL model in G-step, I used 35 cell lines for training and 6 cell lines for baseline update, and 6 cell lines for validation. Since there is a limitation in training time and GPU memory, it is not easy to use all regulators as encoder input. Therefore, for each gene, regulators targeting the gene determined by the prior network were used as input to the encoder to improve training time and GPU memory efficiency.

4.5.2 Hyperparameters

In order to determine K the maximum number of regulators for each gene, I trained the regulator selection model and measured the average error for different $K \in \{5, 10, 15, 20\}$. Table 4.2 shows the average error of estimated gene expression, the maximum number of selected regulators, and the average number of selected regulators for each gene. The error slightly increases as K increases, but it does not seem to be a significant difference. However, it seems that the average number of selected regulators increases as K increases until K=15. Thus I chose K=15 as the maximum number of regulators since there is no significant difference between K=15 and K=20.

Table 4.2: The experiment result of G-step for different hyperparameter K

Model	Average error	The largest regulator number	Average regulator number
RL model K=5	0.190	5	2.711
RL model K=10	0.193	10	3.044
RL model K=15	0.194	14	3.232
RL model K=20	0.195	18	3.226

A weight value λ is an important hyper-parameter for finding an optimal network. There is a trade-off that $h(G)$ decreases as λ increases, but $f(G)$ increases. Thus, it is crucial to determine the appropriate λ to find a reasonable

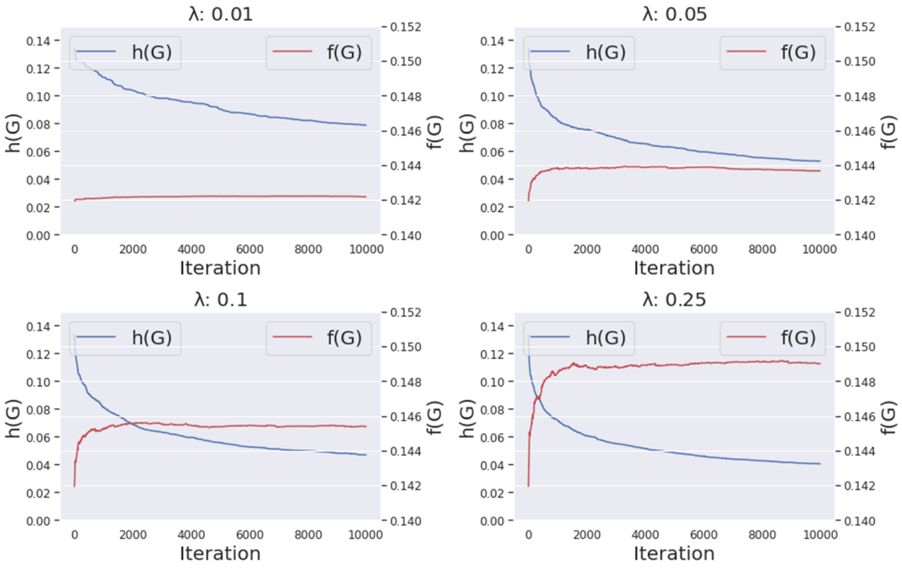


Figure 4.4: The value of error and penalty function for different λ

solution. To determine the λ, I measured the value of $f(G)$ and $h(G)$ for 10,000 iteration with different $\lambda \in \{0.01, 0.05, 0.1, 0.25\}$. Figure 4.4 shows the value of $f(G)$ and $h(G)$ for different λ. I determined $\lambda = 0.05$ is a reasonable trade-off between $f(G)$ and $h(G)$.

4.5.3 Quantitative evaluation

For the quantitative evaluation, I compared the performance of the iterative RL method with Gurobi (Gurobi Optimization, LLC, 2021) and Genetic Algorithm (GA), which are widely used to solve combinatorial optimization problems. Gurobi is a state-of-the-art commercial optimization solver that can handle integer programming (IP). Gurobi IP solver was used to optimize the objective function $F(G)$. GA is a population-based meta-heuristics widely used to solve various optimization problems. GA can be used to minimize objective function $F(G)$. I used the Python geneticalgorithm2 package with default parameters for evaluation. The miRNA, TF, and gene expression data of the MDA-MB-231 cell line were used for evaluation. Table 4.3 shows the objective function value of each method for various number of genes in MDA-MB-231 cell line data.

Table 4.3: The quantitative evaluation results for different problem size.

Number of	Number of	Iterative RL		Gurobi		GA	
genes	edges	Obj.	Time	Obj.	Time	Obj.	Time
100	917	0.187	0.5h	**0.175**	2h	0.656	1h
300	2755	0.267	0.5h	**0.258**	2h	0.748	2h
500	4587	0.225	0.5h	**0.217**	2h	0.717	3.5h
1000	9871	0.198	0.5h	**0.190**	3h	0.723	8h
2000	18987	**0.193**	1h	0.205	6h	0.701	16h
3000	28929	**0.193**	1h	0.554	12h	0.724	24h
4000	39307	0.184	2h	**0.181**	24h	-	-
5000	49925	**0.180**	2h	0.190	24h	-	-
6000	60060	**0.183**	2h	0.679	24h	-	-

Gurobi showed the best performance for problems with less than 1000 genes. However, as the complexity of the problem increases, the iterative RL method provided a more accurate gene expression estimation than Gurobi and GA. The iterative RL method provided a promising solution with a reasonable running time. The quantitative evaluation suggests that the iterative RL method is more suitable for exploring search spaces of gene regulatory networks than previous Gurobi and GA.

4.5.4 Qualitative evaluation

To demonstrate the biological usefulness of iterative RL network search method, I analyzed network consists of cancer hallmark genes in breast cancer cell line.

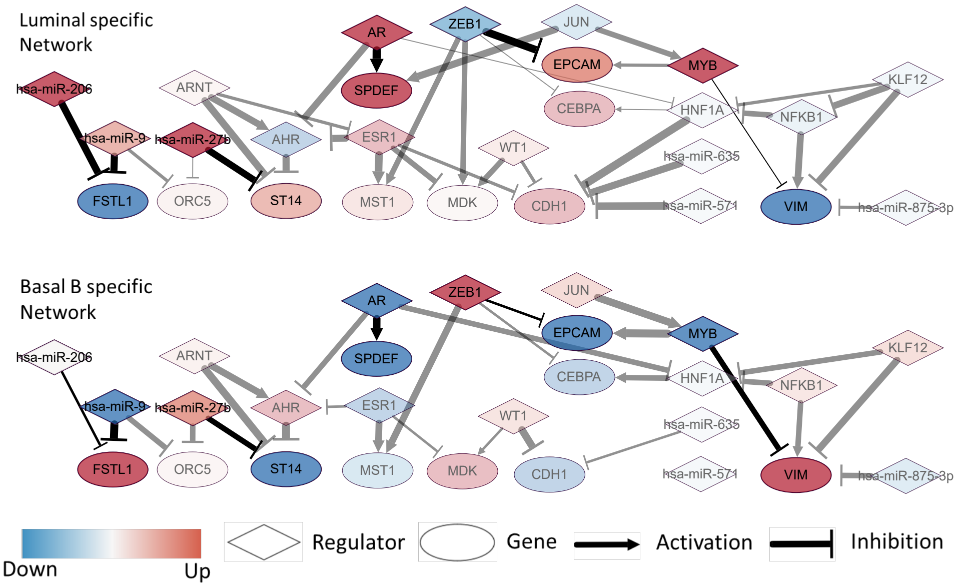


Figure 4.5: Breast cancer subtype-specific sub-network estimated by iterative RL method. The thickness of the network edges is proportional to the frequency of edge occurrences in the subtype. The bolded edges in the network mean that the regulatory relationships have been experimentally verified in previous studies.

The two breast cancer subtype samples were used for evaluation luminal that is less aggressive breast cancer subtype and basal-B that is more aggressive breast cancer subtype. For each cell line, the network is estimated by iterative RL method and aggregated by subtypes.

Figure 4.5 shows subtype-specific sub-networks estimated by the iterative RL method. It seems that some edges in each subtype-specific network represent subtype-specific patterns. Several regulatory relationships in estimated networks have been experimentally verified in previous studies. For instance, Androgen receptor (AR) directly upregulates expression of SAM pointed domain-containing Ets transcription factor (SPDEF) that promotes

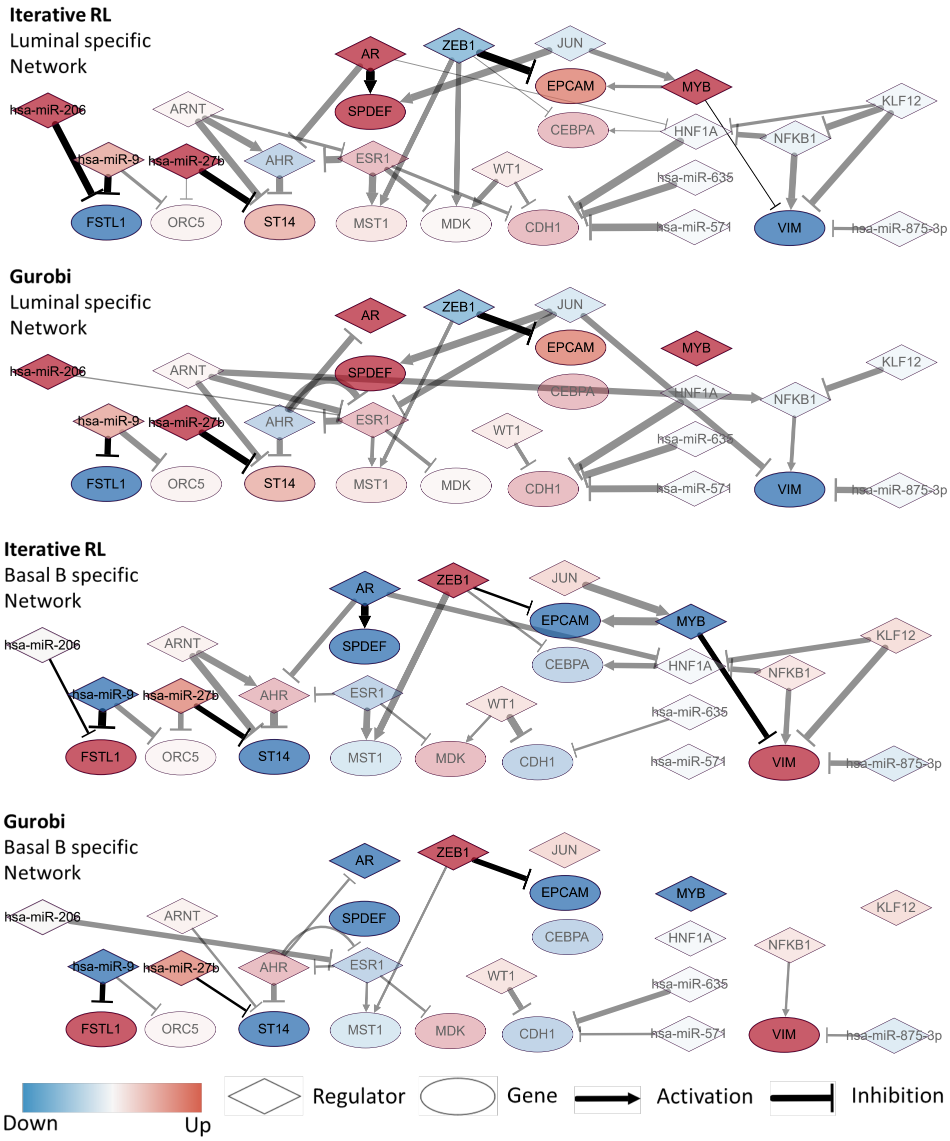


Figure 4.6: Breast cancer subtype-specific sub-network estimated by iterative RL method and Gurobi. The bolded edges in the network mean that the regulatory relationships have been experimentally verified in previous studies.

the proliferation and invasion of breast cancer cell lines (Cao *et al.*, 2018). Suppression of Tumorigenicity 14 (ST14) is a target gene of miR-27b, which increases cancer progression by decreasing ST14 expression (Wang *et al.*, 2009). Vimentin (VIM) is known to be over-expressed in basal breast cancer cell lines. There is an experimental result that the expression of VIM increased when MYB was knocked down in breast cancer cells (Hugo *et al.*, 2013). Zinc finger E-box-binding homeobox 1 (ZEB1) is known as a repressor of Epithelial cell adhesion molecule (EPCAM) in breast cancer (Vannier *et al.*, 2013). Follistatin-related protein 1 (FSTL1) downregulated by miR-9 and miR-206 leads to cell migration (Nowek *et al.*, 2018; Rosenberg *et al.*, 2006).

Figure 4.6 shows the subtype-specific network estimated with iterative RL and the network estimated with Gurobi. Some edges estimated by iterative RL and Gurobi seem to be different, and Gurobi cannot detect some experimentally validated edges.

4.6 Summary

In this study, I proposed combinatorial modeling for the *n-to-m* regulator-gene relationship from the multi-omics data. In order to navigate the search space, I suggested a two-step iterative search method. For G-step, the attention-based encoder-decoder model was used to determine the regulator set to minimize gene expression's absolute error. For R-step, the random selection of regulator with entropy-based probability and the random selection of its target gene with error-based probability were proposed to minimize the degree constraint of regulators. G-step and R-step are iteratively computed until a terminate condition to find a reasonable solution. In the network analysis from breast cancer cell line data, iterative RL reproduced the previously known regulatory relationships relevant to breast cancer subtypes. In summary, my method can determine the promising *n-to-m* relationship within a reasonable time, which may generate a hypothesis for experimental validation relationships related to sample-specific biological functions.

Chapter 5

Conclusions

The integration of multi-omics data provides an opportunity for understanding the cellular state by identifying relationships between omics. However, the multi-omics data analysis is challenging computational task due to the complexity of relationships between high-dimensional omics data. This thesis proposes three computational methods for determining condition-specific relationships from multi-omics data incorporating external knowledge about multi-omics relationships.

1. a literature knowledge guided miRNA-gene relationship prediction method that focuses on the *one-to-m* explicit relationship between miRNA and genes.

2. a method to predict drug response regulatory relationship using the *one-to-m* implicit relationship between multiple regulators and gene.

3. a method to predict sample-specific regulatory network by exploring the *n-to-m* explicit relationship between multiple regulators and genes.

In the first study, a literature-guided miRNA target prediction method, ContextMMIA, was proposed for analyzing two groups of miRNA expression and gene expression data. ContextMMIA used miRNA target databases and literature text mining knowledge as biological prior knowledge. ContextMMIA computes omics score from expression profiles using the statistical difference and negative correlation of miRNA-gene pair in target databases and computes context score for measuring literature relevance between miRNA-gene and data context. ContextMMIA was able to reproduce experimentally validated miRNA-gene relationships.

In the second study, a drug response prediction method, DRIM, was proposed for analyzing multi-omics data with drug sensitivity and drug-induced time-series gene expression data. DRIM used literature knowledge of drug-gene association, PPI-network, TF target database, and biological pathway as gene-gene interaction knowledge. DRIM determines drug response mediator genes by multi-omics integration using low dimensional embedding methods: tensor decomposition and autoencoder. The upstream and downstream relationships of mediator genes are determined by time-series gene expression analysis with gene-gene interaction knowledge. DRIM identified the distinct regulatory paths of PI3K pathway genes in the breast cancer cell line with different lapatinib responses.

In the final study, a sample-specific *n-to-m* relationship prediction method was proposed for analyzing miRNA, TF, and gene expression data. This study used the TF target database and miRNA target database as biological prior knowledge. This study addressed a combinatorial optimization problem to find *n-to-m* relationships estimating observed gene expression. In order to estimate an optimal network, an iterative search method was proposed that iteratively adds and removes edges using RL-based and stochastic-based heuristics. The proposed method constructed breast cancer subtype-specific networks that are estimating a more accurate gene expression than other methods for combinatorial optimzation and involving biologically meaningful edges relevant to breast cancer subtypes.

In conclusion, my doctoral study proposes three computational approaches to identify relationships between regulators and genes from multi-omics data for the comprehensive investigation of cell state. These methods are expected to contribute to the analysis and interpretation of large-scale multi-omics data from various databases such as TCGA and CCLE, which help expand our knowledge of how biological organisms function and respond to external stimulation such as diseases, drugs, and environmental factors.

Bibliography

Abaan, O. D., Polley, E. C., Davis, S. R., Zhu, Y. J., Bilke, S., Walker, R. L., Pineda, M., Gindin, Y., Jiang, Y., Reinhold, W. C., *et al.* (2013). The exomes of the nci-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer research*, **73**(14), 4372–4382.

Adam, T. and Aliferis, C. (2019). *Personalized and Precision Medicine Informatics: A Workflow-Based View*. Springer Nature.

Ahn, H., Jung, I., Chae, H., Kang, D., Jung, W., and Kim, S. (2019). Htrgene: a computational method to perform the integrated analysis of multiple heterogeneous time-series data: case analysis of cold and heat stress response signaling genes in arabidopsis. *BMC bioinformatics*, **20**(16), 588.

Ambros, V. (2004). The functions of animal micrornas. *Nature*, **431**(7006), 350–355.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, **25**(1), 25–29.

Bansal, M., Yang, J., Karan, C., Menden, M. P., Costello, J. C., Tang, H., Xiao, G., Li, Y., Allen, J., Zhong, R., *et al.* (2014). A community computational challenge to predict the activity of pairs of compounds. *Nature biotechnology*, **32**(12), 1213–1222.

Baptista, D., Ferreira, P. G., and Rocha, M. (2020). Deep learning for drug response prediction in cancer. *Briefings in Bioinformatics*.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., *et al.* (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**(7391), 603–607.

Bartel, D. P. (2004). Micrornas: genomics, biogenesis, mechanism, and function. *cell*, **116**(2), 281–297.

Bello, I., Pham, H., Le, Q. V., Norouzi, M., and Bengio, S. (2016). Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*.

Ben-Moshe, N. B., Avraham, R., Kedmi, M., Zeisel, A., Yitzhaky, A., Yarden, Y., and Domany, E. (2012). Context-specific microrna analysis: identification of functional micrornas and their mrna targets. *Nucleic acids research*, **40**(21), 10614–10627.

Bengio, Y., Lodi, A., and Prouvost, A. (2020). Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, **456**(7218), 53–59.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microrna targets predicts functional non-conserved and non-canonical sites. *Genome biology*, **11**(8), R90.

Bisognin, A., Sales, G., Coppe, A., Bortoluzzi, S., and Romualdi, C. (2012). Magia2: from mirna and genes expression data integrative analysis to microrna–transcription factor mixed regulatory circuits (2012 update). *Nucleic acids research*, page gks460.

Bossy-Wetzel, E., Bakiri, L., and Yaniv, M. (1997). Induction of apoptosis by the transcription factor c-jun. *The EMBO journal*, **16**(7), 1695–1709.

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). Htridb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC genomics*, **13**(1), 405.

Brouwer, T. and Lió, P. (2017). Bayesian hybrid matrix factorisation for data integration. *arXiv preprint arXiv:1704.04962*.

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., and Hackermüller, J. (2020). Prospects and challenges of multi-omics data integration in toxicology. *Archives of Toxicology*, pages 1–18.

Cao, L., Xu, C., Xiang, G., Liu, F., Liu, X., Li, C., Liu, J., Meng, Q., Jiao, J., and Niu, Y. (2018). Ar–pdef pathway promotes tumour proliferation and upregulates myc-mediated gene transcription by promoting mad1 degradation in er-negative breast cancer. *Molecular cancer*, **17**(1), 1–15.

Chae, H., Rhee, S., Nephew, K. P., and Kim, S. (2014). Biovlab-mmia-ngs: microrna–mrna integrated analysis using high-throughput sequencing data. *Bioinformatics*, page btu614.

Chatterjee, A., Stockwell, P. A., Rodger, E. J., and Morison, I. M. (2012). Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic acids research*, **40**(10), e79–e79.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and micrornas. *Nature Reviews Genetics*, **8**(2), 93–103.

Chi, Y., Huang, S., Liu, M., Guo, L., Shen, X., and Wu, J. (2015). Cyclin d3 predicts disease-free survival in breast cancer. *Cancer cell international*, **15**(1), 89.

Cho, J.-H., Gelinas, R., Wang, K., Etheridge, A., Piper, M. G., Batte, K., Dakhlallah, D., Price, J., Bornman, D., Zhang, S., *et al.* (2011). Systems biology of interstitial lung diseases: integration of mrna and microrna expression changes. *BMC medical genomics*, **4**(1), 1.

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., Wojcik, S. E., Aqeilan, R. I., Zupo, S., Dono, M., *et al.* (2005). mir-15 and mir-16 induce apoptosis by targeting bcl2. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(39), 13944–13949.

Corsello, S. M., Nagari, R. T., Spangler, R. D., Rossen, J., Kocak, M., Bryan, J. G., Humeidi, R., Peck, D., Wu, X., Tang, A. A., *et al.* (2020). Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, **1**(2), 235–248.

Cox, J. and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. *Annual review of biochemistry*, **80**, 273–299.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563.

Dai, C.-l., Tiwari, A. K., Wu, C.-P., Su, X.-d., Wang, S.-R., Liu, D.-g., Ashby, C. R., Huang, Y., Robey, R. W., Liang, Y.-j., *et al.* (2008). Lapatinib (tykerb, gw572016) reverses multidrug resistance in cancer cells by inhibiting the activity of atp-binding cassette subfamily b member 1 and g member 2. *Cancer research*, **68**(19), 7905–7914.

Eichhorn, P. J., Gili, M., Scaltriti, M., Serra, V., Guzman, M., Nijkamp, W., Beijersbergen, R. L., Valero, V., Seoane, J., Bernards, R., *et al.* (2008). Phosphatidylinositol 3-kinase hyperactivation results in lapatinib resistance that is reversed by the mtor/phosphatidylinositol 3-kinase inhibitor nvp-bez235. *Cancer research*, **68**(22), 9221–9230.

Emdadi, A. and Eslahchi, C. (2020). Dsplmf: A method for cancer drug sensitivity prediction using a novel regularization approach in logistic matrix factorization. *Frontiers in Genetics*, **11**, 75.

Esquivel-Velázquez, M., Ostoa-Saloma, P., Palacios-Arreola, M. I., Nava-Castro, K. E., Castro, J. I., and Morales-Montor, J. (2015). The role of cytokines in breast cancer development and progression. *Journal of Interferon & Cytokine Research*, **35**(1), 1–16.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., *et al.* (2008). Stromal gene expression predicts clinical outcome in breast cancer. *Nature medicine*, **14**(5), 518–527.

Flamini, M., Sanchez, A., Goglia, L., Tosi, V., Genazzani, A., and Simoncini, T. (2009). Differential actions of estrogen and serms in regulation of the actin cytoskeleton of endometrial cells. *Molecular human reproduction*, **15**(10), 675–685.

Fletcher, J. I., Haber, M., Henderson, M. J., and Norris, M. D. (2010). Abc transporters in cancer: more than just drug efflux pumps. *Nature Reviews Cancer*, **10**(2), 147–156.

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mrnas are conserved targets of micrornas. *Genome research*, **19**(1), 92–105.

Fruman, D. A., Chiu, H., Hopkins, B. D., Bagrodia, S., Cantley, L. C., and Abraham, R. T. (2017). The pi3k pathway in human disease. *Cell*, **170**(4), 605–635.

Furth, P. A. (2014). Stat signaling in different breast cancer sub-types. *Molecular and cellular endocrinology*, **382**(1), 612–615.

Gao, H., Korn, J. M., Ferretti, S., Monahan, J. E., Wang, Y., Singh, M., Zhang, C., Schnell, C., Yang, G., Zhang, Y., *et al.* (2015). High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nature medicine*, **21**(11), 1318.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., *et al.* (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**(7391), 570–575.

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. *Breast Cancer Research*, **4**(2), 70.

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., *et al.* (2019). Next-generation characterization of the cancer cell line encyclopedia. *Nature*, **569**(7757), 503–508.

Ginsburg, G. S., Willard, H. F., Woods, C. W., and Tsalik, E. L. (2019). *Genomic and Precision Medicine: Infectious and Inflammatory Disease*. Academic Press.

Goggins, W., Gao, W., and Tsao, H. (2004). Association between female breast cancer and cutaneous melanoma. *International journal of cancer*, **111**(5), 792–794.

Gurobi Optimization, LLC (2021). Gurobi Optimizer Reference Manual.

Haidari, M., Zhang, W., and Wakame, K. (2013). Disruption of endothelial adherens junction by invasive breast cancer cells is mediated by reactive oxygen species and is attenuated by ahcc. *Life sciences*, **93**(25), 994–1003.

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., *et al.* (2011). mirtarbase: a database curates experimentally validated microrna–target interactions. *Nucleic acids research*, **39**(suppl_1), D163–D169.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**(1), 1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, **4**(1), 44–57.

Huang, J. C., Babak, T., Corson, T. W., Chua, G., Khan, S., Gallie, B. L., Hughes, T. R., Blencowe, B. J., Frey, B. J., and Morris, Q. D. (2007). Using expression profiling data to identify human microrna targets. *Nature methods*, **4**(12), 1045–1049.

Hugo, H. J., Pereira, L., Suryadinata, R., Drabsch, Y., Gonda, T. J., Gunasinghe, N. D., Pinto, C., Soo, E. T., van Denderen, B. J., Hill, P., *et al.* (2013). Direct repression of myb by zeb1 suppresses proliferation and epithelial gene expression during epithelial-to-mesenchymal transition of breast cancer cells. *Breast cancer research*, **15**(6), 1–19.

Hwang, H. and Mendell, J. (2006). Micrornas in cell proliferation, cell death, and tumorigenesis. *British journal of cancer*, **94**(6), 776–780.

Ideker, T., Dutkowski, J., and Hood, L. (2011). Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*, **144**(6), 860–863.

Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., *et al.* (2016). A landscape of pharmacogenomic interactions in cancer. *Cell*, **166**(3), 740–754.

Iorio, M. V., Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., *et al.* (2005). Microrna gene expression deregulation in human breast cancer. *Cancer research*, **65**(16), 7065–7070.

Jo, K., Jung, I., Moon, J. H., and Kim, S. (2016). Influence maximization in time bounded network identifies transcription factors regulating perturbed pathways. *Bioinformatics*, **32**(12), i128–i136.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human microrna targets. *PLoS Biol*, **2**(11), e363.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Kang, H., Ahn, H., Jo, K., Oh, M., and Kim, S. (2019). mirtime: identifying condition-specific targets of microrna in time-series transcript data using gaussian process model and spherical vector clustering. *Bioinformatics*.

Karar, J. and Maity, A. (2011). Pi3k/akt/mtor pathway in angiogenesis. *Frontiers in molecular neuroscience*, **4**, 51.

Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network in: Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining, 137–146. *ACM, New York*.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microrna target recognition. *Nature genetics*, **39**(10), 1278–1284.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of chip-seq experiments for dna-binding proteins. *Nature biotechnology*, **26**(12), 1351–1359.

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nature communications*, **7**(1), 1–12.

Kim, S., Jung, W., Ahn, H., Jo, K., Jung, D., Park, M., and Hur, J. (2019). Propanet: Time-varying condition-specific transcriptional network construction by network propagation. *Frontiers in plant science*, **10**, 698.

King, J. A., Ofori-Acquah, S. F., Stevens, T., Al-Mehdi, A.-B., Fodstad, O., and Jiang, W. G. (2004). Activated leukocyte cell adhesion molecule in breast cancer: prognostic indicator. *Breast Cancer Research*, **6**(5), 1.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kool, W., Van Hoof, H., and Welling, M. (2018). Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, **37**(2), 233–243.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., Da Piedade, I., Gunsalus, K. C., Stoffel, M., *et al.* (2005). Combinatorial microrna target predictions. *Nature genetics*, **37**(5), 495–500.

Kupfer, P., Guthke, R., Pohlers, D., Huber, R., Koczan, D., and Kinne, R. W. (2012). Batch correction of microarray data substantially improves the identification of genes differentially expressed in rheumatoid arthritis and osteoarthritis. *BMC medical genomics*, **5**(1), 23.

Lark, A. L., Livasy, C. A., Dressler, L., Moore, D. T., Millikan, R. C., Geradts, J., Iacocca, M., Cowan, D., Little, D., Craven, R. J., *et al.* (2005). High focal adhesion kinase expression in invasive breast carcinomas is associated with an aggressive phenotype. *Modern Pathology*, **18**(10), 1289–1294.

Lee, H.-C., Danieletto, M., Miotto, R., Cherng, S. T., and Dudley, J. T. (2019). Scaling structural learning with no-bears to infer causal transcriptome networks. In *Pac Symp Biocomput*. World Scientific.

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., Lim, S., Choi, D., Kim, S., Tan, A.-C., *et al.* (2016). Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, **11**(10), e0164680.

Lee-Hoeflich, S. T., Crocker, L., Yao, E., Pham, T., Munroe, X., Hoeflich, K. P., Sliwkowski, M. X., and Stern, H. M. (2008). A central role for her3 in her2-amplified breast cancer: implications for targeted therapy. *Cancer research*, **68**(14), 5878–5887.

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microrna targets. *Cell*, **115**(7), 787–798.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *cell*, **120**(1), 15–20.

Li, F. (2011). Privacy, security, risk and trust (passat) and 2011 ieee third inernational conference on social computing (socialcom), 2011 ieee third international conference on.

Li, H., Ning, S., Ghandi, M., Kryukov, G. V., Gopal, S., Deik, A., Souza, A., Pierce, K., Keskula, P., Hernandez, D., *et al.* (2019). The landscape of cancer cell line metabolism. *Nature medicine*, **25**(5), 850.

Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2020). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Briefings in bioinformatics*, **21**(1), 36–46.

Lipponen, P. (1999). Apoptosis in breast cancer: relationship with other pathological parameters. *Endocrine-related cancer*, **6**(1), 13–16.

Liu, Z.-P., Wu, C., Miao, H., and Wu, H. (2015). Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, **2015**.

Lu, P., Weaver, V. M., and Werb, Z. (2012). The extracellular matrix: a dynamic niche in cancer progression. *The Journal of cell biology*, **196**(4), 395–406.

Luo, D., Wilson, J. M., Harvel, N., Liu, J., Pei, L., Huang, S., Hawthorn, L., and Shi, H. (2013). A systematic evaluation of mirna: mrna interactions involved in the migration and invasion of breast cancer cells. *Journal of translational medicine*, **11**(1), 1.

Lv, W.-W., Liu, D., Liu, X.-C., Feng, T.-N., Li, L., Qian, B.-Y., and Li, W.-X. (2018). Effects of pkm2 on global metabolic changes and prognosis in hepatocellular carcinoma: from gene expression to drug discovery. *BMC cancer*, **18**(1), 1–13.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., *et al.* (2006). Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**(suppl_1), D108–D110.

Mazyavkina, N., Sviridov, S., Ivanov, S., and Burnaev, E. (2020). Reinforcement learning for combinatorial optimization: A survey. *arXiv preprint arXiv:2003.03600*.

McLean, G. W., Carragher, N. O., Avizienyte, E., Evans, J., Brunton, V. G., and Frame, M. C. (2005). The role of focal-adhesion kinase in cancer—a new therapeutic opportunity. *Nature Reviews Cancer*, **5**(7), 505–515.

Medina, P. J. and Goodin, S. (2008). Lapatinib: a dual inhibitor of human epidermal growth factor receptor tyrosine kinases. *Clinical therapeutics*, **30**(8), 1426–1447.

Mehlen, P., Delloye-Bourgeois, C., and Chédotal, A. (2011). Novel roles for slits and netrins: axon guidance cues as anticancer targets? *Nature reviews Cancer*, **11**(3), 188–197.

Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D., and Wichard, J. (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature communications*, **11**(1), 1–10.

Monks, A., Zhao, Y., Hose, C., Hamed, H., Krushkal, J., Fang, J., Sonkin, D., Palmisano, A., Polley, E. C., Fogli, L. K., *et al.* (2018). The nci transcriptional pharmacodynamics workbench: a tool to examine dynamic expression profiling of therapeutic response in the nci-60 cell line panel. *Cancer research*, **78**(24), 6807–6817.

Moses, H. and Barcellos-Hoff, M. H. (2011). Tgf-β biology in mammary development and breast cancer. *Cold Spring Harbor perspectives in biology*, **3**(1), a003277.

Mukherji, S., Ebert, M. S., Zheng, G. X., Tsang, J. S., Sharp, P. A., and van Oudenaarden, A. (2011). Micrornas can generate thresholds in target gene expression. *Nature genetics*, **43**(9), 854–859.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, **320**(5881), 1344–1349.

Nam, S., Li, M., Choi, K., Balch, C., Kim, S., and Nephew, K. P. (2009). Microrna and mrna integrated analysis (mmia): a web tool for examining biological functions of microrna expression. *Nucleic acids research*, **37**(suppl 2), W356–W362.

Norfo, R., Zini, R., Pennucci, V., Bianchi, E., Salati, S., Guglielmelli, P., Bogani, C., Fanelli, T., Mannarelli, C., Rosti, V., *et al.* (2014). mirna-mrna integrative analysis in primary myelofibrosis cd34+ cells: role of mir-155/jarid2 axis in abnormal megakaryopoiesis. *Blood*, **124**(13), e21–e32.

Nowek, K., Wiemer, E. A., and Jongen-Lavrencic, M. (2018). The versatile nature of mir-9/9* in human cancer. *Oncotarget*, **9**(29), 20838.

Oh, M., Rhee, S., Moon, J. H., Chae, H., Lee, S., Kang, J., and Kim, S. (2017). Literature-based condition-specific mirna-mrna target prediction. *PloS one*, **12**(3).

Oh, M., Park, S., Lee, S., Lee, D., Lim, S., Jeong, D., Jo, K., Jung, I., and Kim, S. (2020a). Drim: A web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration. *Frontiers in Genetics*, **11**.

Oh, M., Park, S., Kim, S., and Chae, H. (2020b). Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Briefings in bioinformatics*.

Piccirilli, M., Salvati, M., Bistazzoni, S., Frati, A., Brogna, C., Giangaspero, F., Frati, R., and Santoro, A. (2005). Glioblastoma multiforme and breast cancer: report on 11 cases and clinico-pathological remarks. *Tumori*, **91**(3), 256.

Rabanser, S., Shchur, O., and Günnemann, S. (2017). Introduction to tensor decompositions and their applications in machine learning. *arXiv preprint arXiv:1711.10781*.

Rahko, E., Blanco, G., Soini, Y., Bloigu, R., and Jukkola, A. (2003). A mutant tp53 gene status is associated with a poor prognosis and anthracycline-resistance in breast cancer patients. *European journal of cancer*, **39**(4), 447–453.

Reedijk, M. (2012). Notch signaling and breast cancer. In *Notch Signaling in Embryology and Cancer*, pages 241–257. Springer.

Reinhold, W. C., Sunshine, M., Liu, H., Varma, S., Kohn, K. W., Morris, J., Doroshow, J., and Pommier, Y. (2012). Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer research*, **72**(14), 3499–3511.

Rhee, S., Chae, H., and Kim, S. (2015). Plantmirnat: mirna and mrna integrated analysis fully utilizing characteristics of plant sequencing data. *Methods*, **83**, 80–87.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47–e47.

Rosenberg, M. I., Georges, S. A., Asawachaicharn, A., Analau, E., and Tapscott, S. J. (2006). Myod inhibits fstl1 and utrn expression by inducing transcription of mir-206. *The Journal of cell biology*, **175**(1), 77–85.

Saadatmand, S., De Kruijf, E., Sajet, A., Dekker-Ensink, N., van Nes, J., Putter, H., Smit, V., van de Velde, C., Liefers, G., and Kuppen, P. (2013). Expression of cell adhesion molecules and prognosis in breast cancer. *British Journal of Surgery*, **100**(2), 252–260.

Santen, R. J., Song, R. X., McPherson, R., Kumar, R., Adam, L., Jeng, M.-H., and Yue, W. (2002). The role of mitogen-activated protein (map) kinase in breast cancer. *The Journal of steroid biochemistry and molecular biology*, **80**(2), 239–256.

Schramm, G., Surmann, E.-M., Wiesberg, S., Oswald, M., Reinelt, G., Eils, R., and König, R. (2010). Analyzing the regulation of metabolic pathways in human breast cancer. *BMC medical genomics*, **3**(1), 1.

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). Moli: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, **35**(14), i501–i509.

Sigoillot, F. D., Sigoillot, S. M., and Guy, H. I. (2004). Breakdown of the regulatory control of pyrimidine biosynthesis in human breast cancer cells. *International journal of cancer*, **109**(4), 491–498.

Simonian, P. L., Grillot, D. A., and Nuñez, G. (1997). Bak can accelerate chemotherapy-induced cell death independently of its heterodimerization with bcl-x l and bcl-2. *Oncogene*, **15**(15), 1871–1875.

Smyth, G. (2005). Limma: linear models for microarray data. gentleman rcarey vdudoit sirizarry rhuber w bioinformatics and computational biology solutions using r and bioconductor.

Steele, E., Tucker, A., Hoen, P., and Schuemie, M. J. (2009). Literature-based priors for gene regulatory networks. *Bioinformatics*, **25**(14), 1768–1774.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). 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.

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., *et al.* (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**(6), 1437–1452.

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, **14**, 1177932219899051.

Surendiran, A., Pradhan, S., and Adithan, C. (2008). Role of pharmacogenomics in drug discovery and development. *Indian journal of pharmacology*, **40**(4), 137.

Sweeney, G. (1983). Variability in the human drug response. *Thrombosis Research*, **29**, 3–15.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., *et al.* (2015). String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, **43**(D1), D447–D452.

Taguchi, Y. and Turki, T. (2019). Neurological disorder drug discovery from gene expression with tensor decomposition. *Current Pharmaceutical Design*, **25**(43), 4589–4599.

Thomas, S., Snowden, J., Zeidler, M., and Danson, S. (2015). The role of jak/stat signalling in the pathogenesis, prognosis and treatment of solid tumours. *British journal of cancer*, **113**(3), 365.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

van Iterson, M., Bervoets, S., de Meijer, E. J., Buermans, H. P., AC't Hoen, P., Menezes, R. X., and Boer, J. M. (2013). Integrated analysis of microrna and mrna expression: adding biological significance to microrna target predictions. *Nucleic acids research*, **41**(15), e146–e146.

Vannier, C., Mock, K., Brabletz, T., and Driever, W. (2013). Zeb1 regulates e-cadherin and epcam (epithelial cell adhesion molecule) expression to control cell behavior in early zebrafish development. *Journal of Biological Chemistry*, **288**(26), 18643–18659.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Vowels, M. J., Camgoz, N. C., and Bowden, R. (2021). D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*.

Wang, L., Zhang, Q., Zhang, J., Sun, S., Guo, H., Jia, Z., Wang, B., Shao, Z., Wang, Z., and Hu, X. (2011). Pi3k pathway activation results in low efficacy of both trastuzumab and lapatinib. *BMC cancer*, **11**(1), 248.

Wang, L., Li, X., Zhang, L., and Gao, Q. (2017). Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC cancer*, **17**(1), 513.

Wang, X., Sun, Z., Zimmermann, M. T., Bugrim, A., and Kocher, J.-P. (2019). Predict drug sensitivity of cancer cells with pathway activity inference. *BMC medical genomics*, **12**(1), 15.

Wang, Y., Rathinam, R., Walch, A., and Alahari, S. K. (2009). St14 (suppression of tumorigenicity 14) gene is a target for mir-27b, and the inhibitory effect of st14 on cell growth is independent of mir-27b regulation. *Journal of Biological Chemistry*, **284**(34), 23094–23106.

Weinshilboum, R. and Wang, L. (2004). Pharmacogenomics: bench to bedside. *Nature reviews Drug discovery*, **3**(9), 739–748.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, **8**(3-4), 229–256.

Wilson, T. R., Lee, D. Y., Berry, L., Shames, D. S., and Settleman, J. (2011). Neuregulin-1-mediated autocrine signaling underlies sensitivity to her2 kinase inhibitors in a subset of human cancers. *Cancer cell*, **20**(2), 158–172.

Xu, Y., Dong, Q., Li, F., Xu, Y., Hu, C., Wang, J., Shang, D., Zheng, X., Yang, H., Zhang, C., *et al.* (2019). Identifying subpathway signatures for individualized anticancer drug response by integrating multi-omics data. *Journal of translational medicine*, **17**(1), 255.

Yallowitz, A., Ghaleb, A., Garcia, L., Alexandrova, E. M., and Marchenko, N. (2018). Heat shock factor 1 confers resistance to lapatinib in erbb2-positive breast cancer cells. *Cell death & disease*, **9**(6), 1–13.

Zare, H., Khodursky, A., and Sartorelli, V. (2014). An evolutionarily biased distribution of mirna sites toward regulatory genes with high promoter-driven intrinsic transcriptional noise. *BMC evolutionary biology*, **14**(1), 1–10.

Zhang, F., Wang, M., Xi, J., Yang, J., and Li, A. (2018). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Scientific reports*, **8**(1), 1–9.

Zhang, J., Le, T. D., Liu, L., Liu, B., He, J., Goodall, G. J., and Li, J. (2014). Inferring condition-specific mirna activity from matched mirna and mrna expression data. *Bioinformatics*, page btu489.

Zhang, J., Zhu, W., Wang, Q., Gu, J., Huang, L. F., and Sun, X. (2019). Differential regulatory network-based quantification and prioritization of key genes underlying cancer drug resistance based on time-course rna-seq data. *PLoS computational biology*, **15**(11), e1007435.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS computational biology*, **11**(9).

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, **31**, 9472–9483.

Zheng, Y., Zhang, C., Croucher, D. R., Soliman, M. A., St-Denis, N., Pasculescu, A., Taylor, L., Tate, S. A., Hardy, W. R., Colwill, K., *et al.* (2013). Temporal regulation of egf signalling networks by the scaffold protein shc1. *Nature*, **499**(7457), 166–171.

국문초록

세포가 어떻게 기능하고 외부 자극에 반응하는지 이해하는 것은 생물학, 의학에서 가장 중요한 관심사 중 하나이다. 기술의 발전으로 과학자들은 단일 생물학적 실험으로 세포의 변화요인들을 쉽게 측정할 수 있게 되었다. 주목할만한 예시로 게놈 시퀀싱, 유전자 발현량 측정, 유전자 발현을 조절하는 후성 유전체 측정 같은 다중 오믹스 데이터가 있다. 세포의 상태를 더 자세히 이해하기 위해서 다중 오믹스 조절자와 유전자 사이의 조절 관계를 알아내는 것이 중요하다. 하지만 다중 오믹스 조절 관계는 매우 복잡하고 모든 세포 상태 특이적인 관계를 실험적으로 검증하는 것은 불가능하다. 따라서, 서로 다른 유형의 고차원 오믹스 데이터로부터 관계를 예측하기 위한 효율적인 컴퓨터 공학적 접근방법이 요구된다. 이러한 고차원 데이터를 처리하는 한 가지 방법은 다양한 데이터베이스에서 선별된 유전자의 기능과 오믹스 간의 관계와 같은 외부 생물학적 지식을 통합하여 활용하는 것이다.

본 박사학위 논문은 생물학적 사전 지식을 활용하여 다중 오믹스 데이터로부터 유전자의 발현을 조절하는 관계를 예측하기 위한 세 가지 컴퓨터 공학적인 접근법을 제안하였다.

첫 번째는 마이크로 알엔에이와 유전자의 일대다 관계를 예측하기 위한 기법이다. 마이크로 알엔에이 표적 예측 문제는 가능한 표적 유전자의 개수가 너무 많으며 거짓 양성과 거짓 음성의 비율을 조절해야 하는 문제가 있다. 이러한 문제를 해결하기 위해 마이크로 알엔에이-유전자와 데이터의 맥락 사이의 연관성을 문헌 지식을 활용하여 결정하고 마이크로 알엔에이-유전자 관계를 예측하기 위한 ContextMMIA를 개발하였다. ContextMMIA는 통계적 유의성과 문헌 관련성을 기반으로 마이크로 알엔에이-유전자 관계의 점수를 계산하여 관계의 우선순위를 결정한다. 예후가 다른 유방암 데이터에 대한 실험에서 ContextMMIA는 예후가 나쁜 유방암에서 활성화된 마이크로 알엔에이-유전자 관계를 예측하였고 기존 실

험적으로 검증된 관계가 높은 우선순위로 예측되었으며 해당 유전자들이 유방암 관련 경로에 관여하는 것으로 알려졌다.

두 번째는 약물 반응을 일으키는 유전자의 다대일 조절 관계를 예측하기 위한 기법이다. 약물 반응 예측을 위해서 약물 반응 매개 유전자를 결정해야 하며 이를 위해 20,000개 유전자의 다중 오믹스 데이터를 통합 분석하는 방법이 필요하다. 이 문제를 해결하기 위해 저차원 임베딩 방법, 약물-유전자 연관성에 대한 문헌 지식 및 유전자-유전자 상호 작용 지식을 활용하여 약물 반응을 예측하기 위한 DRIM을 개발하였다. DRIM은 오토인코더, 텐서 분해, 약물-유전자 연관성을 이용하여 다중 오믹스 데이터에서 다대일 관계를 결정한다. 결정된 매개 유전자의 조절 관계를 유전자-유전자 상호 작용 지식과 약물 반응 시계열 유전자 발현 데이터의 상호 상관관계를 이용하여 결정한다. 유방암 세포주 데이터에 대한 실험에서 DRIM은 라파티닙이 표적으로 하는 PI3K-Akt 패스웨이에 관여하는 유전자들의 약물 반응 조절 관계를 예측하였고 라파티닙 반응성과 관련된 매개 유전자를 예측하였다. 그리고 예측된 조절 관계가 세포주 특이적인 패턴을 보이는 것을 확인하였다.

세 번째는 세포의 상태를 설명하는 조절자와 유전자의 다대다 조절 관계를 예측하기 위한 기법이다. 다대다 관계 예측을 위해 관찰된 유전자 발현 값과 유전자 조절 네트워크로부터 추정된 유전자 발현 값 사이의 차이를 측정하는 목적 함수를 만들었다. 목적 함수를 최소화하기 위하여 조절인자와 유전자의 수에 따라 기하급수적으로 증가하는 검색 공간을 탐색해야 한다. 이 문제를 해결하기 위해 조절자-유전자 상호 작용 지식을 활용하여 두 가지 연산을 반복하여 조절 관계를 찾는 최적화 기법을 개발하였다. 첫 번째 단계는 네트워크에 간선을 추가하기 위해 강화 학습 기반 휴리스틱을 통해 조절자를 선택하는 다대일 유전자 중심 관계를 탐색하는 단계이다. 두 번째 단계는 네트워크에서 간선을 제거하기 위해 유전자를 확률적으로 선택하는 일대다 조절자 중심 관계를 탐색하는 단계이다. 유방암 세포주 데이터에 대한 실험에서 제안된 방법은 이전의 최적화 방법보다 더 정확한 유전자 발현량 추정을 하였고 조절자 및 유전자 발현 데이터로 유방암 아형 특이적 네트워크를 구성하였다. 또한, 유방암 아형 관련 실험 검증된 조절

관계를 예측하였다.

요약하면, 본 박사학위 논문은 다중 오믹스 조절자와 유전자의 사이의 일대다, 다대일, 다대다 관계를 예측하기 위하여 생물학적 지식을 활용한 컴퓨터 공학적 접근법을 제안하였다. 제안된 방법은 증가하고 있는 분자 생물학 데이터를 분석하여 유전자 조절 상호 작용을 이해함으로써 세포 기능에 대한 심층적인 이해를 도와줄 수 있을 것으로 기대된다.

주요어: 다중 오믹스, 고차원 데이터, 생물학적 사전지식, 유전자 발현량, 기계학습, 탐색 공간

학번: 2015-22904