Sensitivity analysis of drug repositioning study

B080359

August 18, 2016

MSc in High Performance Computing with Data Science
The University of Edinburgh
Year of Presentation: 2016
Abstract

A bottleneck in traditional drug discovery is the extremely long and expensive procedure from theoretical research to industrial development and finally to clinical application, so drug repositioning, which can do drug discovery with smaller investment and shorter development period, is increasingly widely used nowadays. One effective drug repositioning framework based on gene expression has been developed, which predicts similarity in drug mode of action (MoA) by exploiting the gene expression profiles of drug treatment and building a drug network. Then, the drug network is partitioned into some groups which significantly contains the compounds with similar MoA and can be used to discover novel application of existing drugs and predict the application of new drugs. The robustness and valid of the framework is critical because all later analysis studies are based on the result of the framework, so the dissertation focuses on the sensitivity analysis of the framework to verify whether it is robust or not.

In order to meet the aim, we re-implemented the framework and got a drug network with 1,302 nodes and 42,805 edges. Based on the drug network, we identified 87 drug communities. Then, we selected two parameters, the size of optimal signature and distance threshold, to do sensitivity analysis on both drug network and drug communities to see how the framework is affected by the two parameters, comparing to the result obtained from the framework with default parameters (the size of optimal signature is 250 and the distance threshold is 5% quantile). We chose confusion matrix as the quantity measurement to assess the similarity between networks of different parameters. We found that the drug network and drug communities are affected by both parameters. The drug network with smaller distance threshold is a sub-network of the drug network with larger threshold and the drug network nearly keeps unchanged when the size of optimal signature is larger than 200. As for drug communities, we come up with a method to build the cluster network based on drug communities and then compare the cluster networks of different parameters to see how the drug communities change. We found that the drug communities are very sensitive to both two parameters. Therefore, we concluded that the framework mentioned above is not a robust method to do drug reposition, because it is affected by parameters. This dissertation recommends that future study of the network can study the best parameter values for the framework to do drug repositioning.

Keywords: drug repositioning, drug network, sensitivity analysis
# Contents

<table>
<thead>
<tr>
<th>List of Tables</th>
<th>iii</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>v</td>
</tr>
</tbody>
</table>

## 1 Background
1.1 Introduction ................................................. 1
1.2 Outline ....................................................... 4

## 2 Literature Review
2.1 Introduction .................................................. 5
2.2 Biology: basic concepts ...................................... 5
2.2.1 Basic introduction of the cell and DNA .................. 5
2.2.2 Gene expression and regulation .......................... 8
2.3 Open-source dataset: CMap data ............................. 8
2.4 Network theory: basic principles ........................... 9
2.5 Sensitivity analysis: definition and basic techniques ... 10

## 3 Description of algorithms and Implementation
3.1 Introduction .................................................... 12
3.2 Merge the ranked gene probe lists of Each Drug .......... 14
3.2.1 Spearman’s Footrule ...................................... 15
3.2.2 Borda Merging Method .................................... 15
3.2.3 KRUBOR algorithm ......................................... 15
3.3 Drug distance .................................................. 16
3.3.1 Drug representative signature ............................ 16
3.3.2 Computation of distance between two drugs ............ 17
3.4 Cluster algorithm ............................................. 19
3.5 Implementation of the framework ............................ 21
3.5.1 Pre-processing of CMap data .............................. 21
3.5.2 Re-build the drug network ............................... 24
3.5.3 Implementation of cluster algorithm .................... 28

## 4 Sensitivity analysis
4.1 Introduction .................................................... 29
4.2 Measurement to assess the difference between networks .. 30
List of Tables

2.1 Confusion matrix ...................................................... 10
3.1 Some instances summary Information of CMap data .............. 23
3.2 Summary information of cell line .................................. 24
3.3 Statistical summary of AES distance and MES distance .......... 28
4.1 Drug network information of different thresholds ............... 32
4.2 Cluster network information of different thresholds .......... 32
4.3 Statistic information of distances of different $K$ ............... 35
4.4 Network information of different $K$ ............................ 35
4.5 Cluster network information of different $K$ .................... 35
4.6 Statistic information of comparison between networks versus threshold . 37
4.7 Statistic information of comparison between networks versus $K$ ........ 39
4.8 Cluster network information of different thresholds ............ 42
4.9 Statistic information of comparison between cluster networks versus threshold ....................................................... 42
4.10 Cluster network information of different thresholds in new method . 45
4.11 Statistic information of comparison between cluster networks versus threshold in new method ....................................................... 46
4.12 Cluster network information of different $K$ ...................... 47
4.13 Statistic information of comparison between cluster networks versus $K$ .... 47
4.14 Cluster network information of different $K$ in new method ..... 48
4.15 Statistic information of comparison between cluster networks versus $K$ in new method ....................................................... 49
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1</td>
<td>Confusion matrix of drug networks of different thresholds</td>
<td>56</td>
</tr>
<tr>
<td>B.2</td>
<td>Confusion matrix of drug networks of different $K$</td>
<td>56</td>
</tr>
<tr>
<td>B.3</td>
<td>Confusion matrix of cluster networks of different thresholds</td>
<td>57</td>
</tr>
<tr>
<td>B.4</td>
<td>Confusion matrix of cluster networks of different $K$</td>
<td>57</td>
</tr>
<tr>
<td>B.5</td>
<td>Confusion matrix of cluster networks of different thresholds in new method</td>
<td>57</td>
</tr>
<tr>
<td>B.6</td>
<td>Confusion matrix of cluster networks of different $K$ in new method</td>
<td>57</td>
</tr>
</tbody>
</table>
List of Figures

2.1 The cell structure .................................................. 6
2.2 The DNA double helix ............................................ 7
2.3 The relation between DNA, mRNA and Protein .......... 7
2.4 The undirected network (left) and directed network (right) 9
3.1 The process of building the DN ................................. 13
3.2 The example of the merge stage ............................... 17
3.3 The message exchange between nodes .................... 20
3.4 The flow of merge .................................................. 25
3.5 The empirical probability density function of AES .......... 27
3.6 The empirical probability density function of MES .......... 27
4.1 Network change versus threshold ............................. 34
4.2 The number of drugs and edges versus threshold .......... 34
4.3 Statistic analysis on the distances versus $K$ ............. 36
4.4 Comparison between networks versus threshold .......... 38
4.5 Comparison between networks versus $K$ ................. 40
4.6 Comparison between cluster networks versus threshold .... 42
4.7 The sub-network of the community in assumption ....... 44
4.8 The network of the community in reality .................. 44
4.9 The network of the community which contains 5 drugs using new method 45
4.10 Comparison between cluster networks versus threshold in new method 46
4.11 Comparison between cluster networks versus $K$ .......... 48
4.12 Comparison between cluster networks versus $K$ in new method 49
A.1 The empirical probability density function of AES . . . . . . . . . . . . 55
A.2 The empirical probability density function of MES . . . . . . . . . . . 55
Acknowledgements

On the completion of my thesis, I would like to express my gratitude to all those who helped me during the writing of this thesis.

First of all, I would like to express my deepest gratitude to my supervisors, professor Ally Hume and professor Menolascina Filippo, whose kindness and advice have made this work possible. They gave me valuable instructions and helped me when I faced some problems. Their constant encouragement, effective advice and shrewd comments kept the thesis in the right direction.

I also would like express my gratitude to professor Michael Bareford, who helped in the parallelization of my program to make the program practical and improve the performance of the program. His expert guidance and effective advice have made the parallelization possible.

Finally, I feel grateful to all the professors in EPCC who once offered me valuable courses and advice during my study. From their devoted teaching and enlightening lectures I have benefited a lot and learned a lot about the programming skills and parallel programming knowledge, which helps me a lot in my project.
Chapter 1

Background

1.1 Introduction

Recently, drug repositioning is increasingly used in drug discovery and concerned by government departments, pharmaceutical companies and academic institutions. It refers to the process of re-screening, combining and reforming the existing drugs with relevant technical methods to discover the unknown novel effect. It is important to verify that the relevant technical methods are valid and robust to do drug repositioning because all later studies to analyze the application of drug are based on it. Sensitivity analysis can be used to verify whether the method is robust or not by changing some parameters of the method. It compares the outcomes of the method with different parameters to see how the outcome varies with the change of parameters. If the outcomes of different parameters nearly keep unchanged, the method can be regarded as a robust method to do drug repositioning.

As for why drug reposition is widely used in drug discovery recently, the traditional drug discovery method has to be introduced. In the past, the discovery of a new drug was typically based on the knowledge of biological mechanisms of the diseases to be treated, which means that before the drug has been discovered, the gene targets of the drug have been hypothesized. Based on this, the chemical structure of the drug could be systematically analyzed so that the drug can cure the disease by working on the corresponding gene targets. In general, after a long period of biology and clinical medicine research in a specific field, experts can accumulate plenty of knowledge for a particular disease. They can find the gene targets based on the experience and design drug with corresponding chemical structure. Then, after a large-scale clinical trials, the drug is verified that it satisfies the criteria of safety and effectiveness in terms of the statistical level. Only then, the new drug is discovered. Recently, in order to use the experience of experts, based on the previous drug discovery method, the drug libraries have been development which records the experiments results of identification of molecules acting on specific gene targets [1].

It is clear that the traditional drug discovery is an extremely long and expensive pro-
procedure to discover a new drug, from theoretical research to industrial development and finally to clinical application. Besides, the research methods, technical conditions and project management involved in traditional drug discovery are very complex. It not only needs a large research team, but also depends on the effective corporation between any two parts of the project. Any error or accident may have a significant negative impact on the project[2], which makes the drug discovery with great uncertainty and risk [3]. In addition, it is because the long period and strict condition of the discovery of new drug that the cost is very high [4]. In general, a large number of candidate drugs can not be used ultimately, so the investment of previous research becomes the enormous cost for pharmaceutical industry. If the failure risk is taken into account, the average investment for every successful drug will be even more enormous. These negative factors have become a bottleneck in the recent research of the development of new drugs.

Drug repositioning can solve these problems. Compared to the development of new drug, drug repositioning based on re-development of the existing drugs can save lot of investment, because it may make a failed drug "revive" and even may further expand the use of a successful drug [5]. Besides, compared with the traditional drug discovery, drug repositioning can shorten the period of drug development from 12-17 years to 2-12 years. Therefore, the drug reposition is increasingly widely used in drug discovery.

Many of current successes of drug repositioning come through occasional clinical observation, which is not systematic and repeatable, so systematic data-driven approaches are now showing an increasing promise to do drug repositioning for multiple drugs and diseases simultaneously using a wide range of data sources. Existing computational drug repositioning approaches use similar strategies with different principles. The systematic drug repositioning based on the linkage map of genome expression uses the methods to compare gene expression of cellular response to a specific drug with gene expression of the cell in common condition. Drug repositioning based on genetics connects a known drug target with a genetically associated phenotype. Drug repositioning based on drug-protein interaction group to find novel targets of drugs builds a network which links drugs or diseases according to shared features. Drug repositioning based on side effects of drugs uses the similarity in side-effect to infer novel indications [1].

The dissertation mainly concerns the drug repositioning based on the gene expression. The principle of the method is to investigate the drug mode of action (MoA) to see how the cell responds to the drug, which means to detect which genes are affected by the drug.

In biology, gene is a locus of DNA which can encode a functional RNA and a protein product. Message RNA (mRNA) is a kind of RNA which conveys genetic information from DNA and instructs the production of protein. In recent research, mRNA level is used as the genomic signature to describe genes, because it is the intermediate molecule between gene and its products and due to the development of DNA microarray technique, it is practical to generate mRNA levels from a small number of cells at low cost and in high throughput.

DNA microarray technique is able to measure gene expression levels at a genome-wide
scale simultaneous, which means it can produce genome-wide gene expression profile (GEP). Based on this, it can be used to monitor the expression of all genes in a cell to measure how the cell responds to a specific drug at the transcriptional level. It is the logical next step of genome sequence: the sequence of genome reveals what the cell possibly do and the gene expression profile reveals what the cell actually do at a point of time [6].

There is an efficient drug repositioning method based on gene expression to study drug in mammalian system, which is done by Francesco Lorio and his colleagues [7]. They presented a novel and efficient framework to investigate the MoA of drugs by using transcriptional expression data only and to suggest novel therapeutic application of already known drugs by building the drug network, based on a publicly accessed database, CMap database. In order to do drug repositioning, first of all, they made an assumption on the similarity of MoA between drugs: whatever two drugs act on the same or different directly gene targets, if two drugs have the similar effects on transcriptional response of the cell, then the two drugs would be regarded as sharing the same MoA or therapeutic application. Then, based on the assumption, one similarity measure was chosen to analyze the similarity between drugs and the distances between drugs were calculated. The smaller the distance between two drugs is, the more similar the two drugs are. After they got the distances between drugs, they chose a threshold of distances. If the distance between two drugs is below the threshold, the two drugs are regarded as similar drugs. They used the drug network to present the similarity between drugs, in which each drug was referred as a vertex and the similarity between drugs was referred as a weighted edge. As the result, they constructed a drug network with 1,302 vertices (drugs) and 41,047 edges (similarity between drugs). Finally, they further studied the drug network and found the communities of drugs, which means that drugs in the same cluster are similar drugs and share similar drug MoA. The above drug communities not only can identify the drug MoA of already known drugs, but also can predict the novel therapeutic application of already known drugs. Using the framework, they successfully predicted the application pattern of nine anticancer drugs and found that one of the nine anticancer drugs, topoisomerase inhibitor, has the effect of inhibition of cell-cycle dependent protein kinase. The study and reality show that the systematic calculation of gene expression profile is one of the effective methods for the discovery of the novel therapeutic application of drugs and for drug repositioning.

The dissertation will focus on the sensitivity analysis of the drug repositioning method mentioned above to verify whether the method is robust or not. In order to achieve the aim, the dissertation will re-implement the framework of Francesco Lorio to obtain the drug network and the drug communities. Then, based on network theory, the drug communities can be represented by a cluster network. Sensitivity analysis will be done on selected parameters of the framework to see how the drug network and cluster network responds to the change of parameters. According to a chosen comparison measurement between different networks, if the networks are nearly same regardless of change of parameters, the framework of drug repositioning will be regarded as robust.
1.2 Outline

The dissertation describes the process of re-implementation of the framework of drug repositioning. Then, it describes the sensitivity analysis on such framework to verify whether the drug repositioning method is robust or not.

In chapter 2, some basic biological concepts and principles as the background information is provided, which are needed to understand the principle of the framework. Moreover, in the same chapter, the publicly accessed dataset, CMap dataset, will be briefly introduced to help understand what data will be used in the paper. Finally, some concepts of network theory and sensitivity analysis are also covered in the chapter to help understand the results of the framework and what will be done later to verify the robustness.

Chapter 3 introduces the work flow of the framework and what algorithms used in the framework in details to help understand the framework more. Then, based on the algorithms, it introduces the details of how we deal with the dataset and re-implemented the network.

Based on the re-implementation results, which parameters will be chosen to do sensitivity analysis and how to do sensitivity analysis will be introduced in chapter 4.

Final conclusion chapter briefly introduces what we done in the project and what conclusions has been reached in the project. Besides, it also offers some recommendations for future study on the framework of drug repositioning.
Chapter 2

Literature Review

2.1 Introduction

This chapter introduces some basic biological concepts, which are necessary to understand what the experimental data are and what are analyzed in the project. (Section 2.2)

Then, Section 2.3 briefly introduces the publicly accessed dataset, CMap dataset, which is mostly used to investigate the drug Mode of Action (MoA) in the present research. It is helpful to understand what data will be processed in the project.

In Section 2.4, because the result of the framework mentioned above is represented by the network and some cluster algorithm are based on the principles of network, some basic definitions of network which is useful to understand the result of the framework and the cluster algorithm is introduced.

The final section introduces what is sensitivity analysis and one basic method, confusion matrix, in detail.

2.2 Biology: basic concepts

2.2.1 Basic introduction of the cell and DNA

The cell is the simplest and most elementary functional unit of life and it can be regarded as the most basic building blocks of human beings. According to the number of cells which are used to constructing the organisms, organisms can be divided into unicellular organism and multicellular organism. The unicellular organism, such as bacteria, consists of only one cell. The multicellular organism, such as humans, consists of more than one cell [8].
Using the animal cell as an example to briefly explain the cell structure (see Figure 2.1). Three important components of the cell are cell membrane, nucleus and cytoplasm [9]. The cell membrane is the boundary of the cell and keeps the cell together by containing all the organelles within it. It can be selectively-permitted to allow materials to move into and out of the cell. The cytoplasm is a thick and clear liquid between the cell membrane and nucleus, which maintains the cell’s chemical climate. The nucleus is the most important component of a cell, because it contains the genetic material, the Deoxyribonucleic acid (DNA).

A DNA is a polynucleotide, which is a chain of four kinds of nucleotides, adenine (A), guanine (G), cytosine (C) and thymine (T). The end of DNA is marked with 5’ left and 3’ right. The two polynucleotides are complementary, if one can be obtained from the other by exchanging A with T and C with G. Two complementary polynucleotides chains resembles a helix, which is knowns as a DNA double helix [10] (see Figure 2.2).

DNA can instruct all the development and function of nearly all living organisms and can long-term store information. It is always regarded as the cell’s program, because it can instruct the construct of other components of the cell, such as protein and Ribonucleic acid (RNA). Gene is a segments of DNA which carries the genetic information. The DNA sequences, except the genes, have structural purposes or can do the regulation in the expression of genetic information [9]. Both the change of these DNA sequences and the change of gene sequence may lead to abnormal performance of gene.

Another important component of the cell is ribosomes, which is critical to assemble the proteins. To make a protein, the DNA will be transcribed into a functional RNA, mRNA [9]. Then, the ribosomes can translate the mRNA into instructions and weld chemical building blocks one by one according to the instructions. The relation between DNA, mRNA and protein can be saw in Figure 2.3.
Figure 2.2: The DNA double helix

Figure 2.3: The relation between DNA, mRNA and Protein
2.2.2 Gene expression and regulation

As mentioned above, mRNA is the intermediate molecular that carries the information of the DNA sequence that represents a protein. The gene expression is a procedure that translates the nucleotide sequence of mRNA into a sequence of amino acids comprising a protein, which means transforming the information from gene to gene product [11].

Gene expression is a very complex process and different cellular components (DNA, mRNA and protein) build a complex hierarchical network of interaction. For DNA transcription, some proteins play an important role by regulating the expression of a gene in a positive or negative way. If they act in a positive sense, the level of mRNA transcribed for that gene will be increased, otherwise the level of mRNA will be decreased. For the RNA translation, recent research shows that RNA is the only regulation factor and act through RNA interference mechanism as "silencer" of gene expression [12]. It means that the products of some genes influence the transcriptional activity of other genes.

The gene expression can be digitally represented by the level of mRNA transcribed for that gene, which is actually the transcriptional gene expression and can be measured by the DNA microarray technique. DNA microarray technique is able to measure gene expression levels at a genome-wide scale simultaneous, which means it can produce genome-wide gene expression profile (GEP). It is the logical next step of genome sequence: the sequence of genome reveals what the cell possibly do and the gene expression profile reveals what the cell actually do at a point in time [6].

2.3 Open-source dataset: CMap data

The Connectivity Map (CMap) is an open-source dataset, which describes the functional connections between drugs, genes and diseases which are revealed by common gene-expression changes. It consists of 6,100 gene expression profiles as instances, which are created from five human cell lines treated with 1,309 drugs at different concentrations and recorded at different durations [13]. To be specific, one instance is a transcriptional gene expression of a cell under the drug treatment at a concentration after a period of duration.

The CMap consists of many batches (experiments). Each batch consists of a number of gene expressions of drug-treated cell lines and untreated cell lines and the number of gene expressions of each batch maybe different. The change of gene expression in a cell line treated with a specific drug can be calculated by considering the difference between the gene expression of the treated cell line and the gene expression of the same kind of cell which is not treated with the drug. The difference can reveal how the cell responses to the specific drug at a concentration after a period of duration, known as drug effect.
2.4 Network theory: basic principles

A network is simply a collection of connected objects. The objects are referred as nodes or vertices and denoted by the points in the graph, and the connections are referred as edges and denoted by lines between two points [14]. Formally, a network $G$ can be defined as the pair $(V, E)$, where the $V = \{v_1, v_2, ..., v_n\}$ is the set of vertices and $E = \{e_1, e_2, ..., e_m\}$ is the set of edges. A single edge $e \in E$ represents the connection between a pair of vertices $(x, y) \in V^2$ and in the case vertices $x$ and $y$ are said to be joined by edge $e$.

According to whether the order of the pair $e = (x, y)$ is relevant or not, network can be divided into two types, directed network and undirected network (see Figure 2.4). In directed network, the sorting order of the pair $e = (x, y)$ is relevant, which means that the edge $(x, y)$ is different from the edge $(y, x)$. In the case, the vertex $x$ is said to be the source node of edge $e$, denoted by $s(e) = x$, and the vertex $y$ is said to be the destination node of edge $e$, denoted by $d(e) = y$. In the undirected network, the order of the pair $(x, y)$ is irrelevant, which means that the edge $(x, y)$ is the same as the edge $(y, x)$. In the case, the vertices $x$ and $y$ joined by the edge can be regarded as source node and destination node interchangeably.

![Figure 2.4](image)

**Figure 2.4:** The undirected network (left) and directed network (right)

The network can be represented by a square (0,1)-matrix, adjacency matrix [14]. The element of the matrix indicates whether the pair of vertices is connected (1) or not (0) in the network. If the network is an undirected network, which means that the edge $(x, y)$ is the same as edge $(y, x)$, the adjacency matrix is symmetric. Following is the adjacent matrix of above undirected network.

\[
\begin{pmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
\end{pmatrix}
\]  

(2.1)

The network theory is the study of the network and it is a cross-disciplinary concept. In the network theory, one important topic is "weighted networks", which means that every edge in the network has a weight, representing the "loss" or "gain", such as the distance
between the two vertices. In the case, many problems become the problem to identify the sub-network to obtain the optimal value, the biggest "gain" or the smallest "loss". As for the cluster problem which is needed to be tackled in the paper, it is based on the "weighted network", so it is fundamental and critical to quantify the similarity between two objects. Then, based on the similarity measurement and the network topology, the sub-network will be identified as the communities to obtain the optimal value.

2.5 Sensitivity analysis: definition and basic techniques

Sensitivity analysis is to analyze how the result of a mathematical model or a system is affected by the change of parameters [15]. To be specific, it analyzes the difference of the results when the value of selected parameter is changed while the values of other parameters are held. If the results of the model or system are nearly the same when the selected parameters change, the model or the system can be considered as a robust system.

One basic technique of sensitivity analysis is confusion matrix, which is also known as \(2 \times 2\) contingency table. Using a two-class classification problem as an example to explain what is confusion matrix. In the binary classification problem, the outcomes are labeled as positive (\(p\)) or negative (\(n\)). There are four kinds of combinations of prediction value and actual value. If the prediction value is \(p\) and the actual value is \(p\), it is called a true positive (TP). If the prediction value is \(p\) while the actual value is \(n\), it is called a false positive (FP). If the prediction value is \(n\) while the actual value is \(p\), it is called a false negative (FN). If both the prediction value and actual value are \(n\), it is called a true negative (TN). It can be summarized in a \(2 \times 2\) matrix, confusion matrix [16] (see Table 2.1).

<table>
<thead>
<tr>
<th>Prediction Condition</th>
<th>True Condition</th>
<th>Predicted Positive</th>
<th>Predicted Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition Positive</td>
<td>TP</td>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>Condition Negative</td>
<td>FP</td>
<td>TN</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Confusion matrix

Based on the confusion matrix, four statistic variables, Accuracy, Precision, true positive rate and false positive rate can be calculated using the following equation.

\[
\text{Accuracy} = \frac{TP + TN}{\text{Total number of samples}} \tag{2.2}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{2.3}
\]

\[
TPR(\text{sensitivity}) = \frac{\sum \text{True Positive}(TP)}{\sum \text{Condition Positive}} \tag{2.4}
\]
\[ FPR(1 - \text{specificity}) = \frac{\sum \text{False Positive}(FP)}{\sum \text{Condition Negative}} \] 

(2.5)

Accuracy indicates how many objects are correctly classified, reflecting the general discrimination ability of the classifier. Precision indicates how many positive predicted objects are really positive. TPR is also known as sensitivity or recall in machine learning, which indicates how many correct positive results occur among all actual positive samples, reflecting discrimination ability of positive category. FPR is also known as 1-specificity, which indicates how many incorrect positive results occur among all actual negative samples, reflecting the discrimination ability of negative category. All four statistic variables range from 0 to 1.

According to the definition of the four statistic variables, it is obvious that the larger Accuracy, Precision and TPR are and smaller FPR is, the better the classifier is. Hence, the perfect prediction parameter would yield 100% Accuracy, 100% Precision, 100% TPR and 0 FPR, which means that all positive objects and negative objects are correctly identified.
Chapter 3

Description of algorithms and Implementation

3.1 Introduction

This chapter focuses on the principle of the framework and how we re-implemented the framework. In the CMap data, each drug may work on more than one cell line at different concentrations, recorded at different durations, so each drug may have more than one transcriptional response. In the situation, it is difficult to decide which transcriptional response to be chosen to compute the distance between drugs. Hence, in [17], it takes all the transcriptional responses of the same drug, across different cell lines, and/or at different concentrations and durations, into consideration. Each of the transcriptional response is presented by a ranked lists of Microarray Probe-set Identifiers (MPI) according to the differential gene expression. Then, a single ranked gene probe list for each drug is obtained, the Prototype Ranked List (PRL), by merging all the ranked list of the same drug. Section 3.2 introduces how to merge all the ranked gene probe lists of the same drug to generate a general single ranked list of the drug, regardless of cell line, duration and concentration.

After the single PRL is obtained for each drug, the [17] chooses the distance between two drugs as the similarity measure. Section 3.3 introduces how to calculate the distances between drugs. Once the distances between drugs are calculated, it chooses 5% quantile of distances as the distance threshold value. Then, the paper regards each drug as a node in a network. Each pair of nodes is connected if the distance of the pair is below the threshold value. If a drug is not connected with any other drug by at least one edge, the drug would be excluded from the network. The paper gets a drug network (DN) with 1,302 nodes and 41,047 edges.

In order to analyze and visualize the drug network, the communities of drugs are identified. Section 3.4 introduces the cluster algorithm which is used to cluster the drugs into communities. Until then, the framework is built.
Final section will introduce how we re-implemented the framework and what results we got from the re-implementation.

In the rest of the section, following notation will be used. The process of building the drug network can be seen in Figure 3.1

- $P$: a list of Microarray Probe-set Identifiers (MPI)
- $m$: the total number of MPIs, in the CMap dataset, $m = 22,283$.
- $D$: all 6100 ranked lists of the same set of $m$ MPI by sorting the differential gene expressions in decreasing order.
- $X$: a set of ranked lists of MPI of the same drug, $X \in D$
- $r$: $P \times D \rightarrow [1, ..., m]$: a function assigns a rank to each MPI of $P$ according to a ranked list $d \in D$.
- $\delta$: $D^2 \rightarrow N$: calculate the Spearman’s Footrule similarity between each pair of ranked list in $X$
- $B$: $D^2 \rightarrow D$: the Borda Merging function merges two ranked list in $X$ to one ranked list using Borda Merging method.

Figure 3.1: The process of building the DN
3.2 Merge the ranked gene probe lists of Each Drug

In order to equally weight the contribution of each instance of the drug to the single
PRL, regardless of cell line, concentration, duration, rank merging is achieved with
a procedure, which is based on a hierarchical majority-voting scheme, where gene
probes consistently over-expressed/down-regulated across the ranked lists are moved
to top/bottom [19].

A famous method, Borda Merging method, is chosen to pair-wisely merge the ranked
gene probe lists of the same drug. The function is defined as $B: D^2 \rightarrow D$. The input
of the function is two ranked lists of the same set of MPIs of the same drug and the output
of the function is a new ranked lists of the same set of MPIs which summarizes the
ranking order of the both two inputs. The method realizes the voting scheme majorly
by computing the list of values $H = [h_1, h_2, ..., h_m]$, where $h_i = r(i, x) + r(i, y)$, $i = 1, 2, 3, ..., m$. Then, the merged ranked list of gene probes is generated by sorting $P$
according to there values in $H$ in increasing order.

As mentioned above, it is necessary to equally consider all the instances of the same
drug while only using Borda Merging method cannot realize this, because for a partic-
ular drug, the number of ranked gene probe lists for each cell line is different, which
will lead to that the responses of the cell line with smaller number of instances of the
drug are poorly represented. Therefore, a weight is needed to balance the inequality
mentioned above and Spearman’s Footrule measure is used. It calculates the distance
between each pair of ranked lists. The smaller the distance is, the more similar the two
ranked lists are.

In order to use the distance (Spearman’s Footrule) and Borda merging method to merge
all ranked lists to a new one, the idea of Kruskal Algorithm is used. It is a greedy
algorithm in network theory to find a minimum spanning tree for a connected weighted
network by repeating the two steps: finding the edge with minimum weight and adding
increasing costs arcs at each step [20]. It is clear the algorithm is based on the hierarchy
of the weight of edge.

The KRUBOR merging algorithm to generate a single PRL is built based on the similar-
ity measure between two ranked lists (Spearman’s Footrule measure), a method merge
two ranked lists (Borda Merging method) and an algorithm to create a single ranked list
from a set using a hierarchical way (Kruskal Algorithm) [21].

The merging stage firstly compares the similarity between each pair of ranked lists of
the same drug using Spearman’s Footrule measure. Then, it chooses the two lists with
the smallest distance, which means that the two lists are the most similar pair among
all pairs, and merges them using the Borda Merging method. The procedure repeats the
above two steps until a single PRL of the drug is obtained.
3.2.1 Spearman’s Footrule

Spearman’s Footrule is calculated using the following equation:

\[ \delta(x_p, x_q) = \sum |r(i, x) - r(i, y)|, i = 1, 2, 3, \ldots, m \]

Where \( r \) is the function mentioned above. For example, if \( a = (1, 2, 3, 4) \) and \( b = (5, 6, 7, 8) \), the Spearman’s Footrule distance between \( a \) and \( b \) is \( \delta(a, b) = 16 \).

3.2.2 Borda Merging Method

As mentioned above, the input of the method is two ranked gene probe lists of the same drug, denoted by \( a = \{a_1, a_2, \ldots, a_m\} \) and \( b = \{b_1, b_2, \ldots, b_m\} \). Based on the input, the method computes a list of values \( H = \{h_1, h_2, \ldots, h_m\} \), where \( h_i = r(i, a) + r(i, b) \). Finally, the new ranked list is generated by sorting the gene probes in \( P \) according to their value in \( H \) in an increasing order.

There is an example to explain the method. In the example, \( m = 5 \) and \( P \) is

\{gene_1, gene_2, gene_3, gene_4, gene_5\}

For drug \( A \), there are two ranked list

\( a = \{gene_1, gene_4, gene_5, gene_2, gene_3\} \)

and

\( b = \{gene_4, gene_5, gene_1, gene_3, gene_2\} \)

Based on \( P \), \( r(a) = \{1, 4, 5, 2, 3\} \) and \( r(b) = \{3, 4, 5, 1, 2\} \). Hence, \( H = \{4, 9, 9, 3, 5\} \).

Sorting the \( P \) according to corresponding value in \( H \) in an increasing order, a new ranked list is generated:

\{gene_4, gene_1, gene_5, gene_2, gene_3\}

As shown in the new ranked list, \( gene_4 \) ranked at the top of the new list, it is because that \( gene_4 \) consistently ranked nearly at the top of both \( a \) and \( b \) and it got the smallest number, 3, in \( H \). Due to the same reason that \( gene_2 \) and \( gene_3 \) consistently ranked at the bottom of \( a \) and \( b \), they also ranked at the bottom of the new list.

3.2.3 KRUBOR algorithm

The KRUBOR algorithm is described in algorithm 1. The input of the algorithm is \( X \), which has been defined in section 3.1. According to the KRUBOR algorithm, the algorithm first select two ranked lists with minimum distance (Spearman’s Footrule) of \( X \), which means to find the most similar two ranked lists [line 6]. Then, it uses Borda
Algorithm 1 KRUBOR algorithm

1: **procedure** MERGE(X)
2: **input:** X, all PRLs of the same drug
3: **output:** single PRL of the drug
4: \( n = |X| \)
5: **while** \( n > 1 \) **do**
6: \( \text{find } i, j : \delta(x_i, x_j) = \min \delta(x_p, x_q), p, q = 1, 2, \ldots, p \neq q \)
7: \( y = B(x_i, x_j) \)
8: \( X = (X/\{x_i, x_j\}) \cup y \)
9: \( n = |X| \)
10: **end while**
11: **end procedure**

merging method to merge the selected two ranked lists to create a new one \( y \) [line 7]. In the next step, it removes the selected two ranked lists from \( X \) and adds the new one \( y \) to \( X \) [line 8]. The procedure repeats the iteration until only one list left in \( X \), which means that the single PRL of the drug is obtained.

Figure 3.2 shows an example of the merge of all ranked lists obtained by treating MCF7, PC3 and HL60 cell lines with acetylsalicylic acid, also known as aspirin, which maybe an effective medicine at preventing certain types of cancer, particularly colorectal cancer [23]. In the example, it starts from the pair-wise calculation of Spearman’s Footrule distance between each pair of the ranked lists. In the figure, each node is a ranked list and the Spearman’s Footrule distance corresponds to the euclidean distance between nodes. Then, it selects two ranked lists with the smallest Spearman’s Footrule distance, list 1 and 2. The two lists are merged using Borda Merge method to generate a new ranked list, denoted by node 3 in the figure. The example continues by merging node 4 and 5, generating the new ranked list 6. The algorithm continues until the merge of node 23 and 24, generating a general PRL of acetylsalicylic acid.

3.3 Drug distance

3.3.1 Drug representative signature

In the PRL of a drug, the closer to the top of the ranked list, the larger the differential gene expression of the gene probe is compared with the control one. The closer to the bottom of the ranked list, the smaller the differential gene expression of the gene probe is compared with the control one. The gene probes locating at the middle of the ranked list maybe not affected by the drug. The gene probes whose gene expression change a lot due to the drug compared with the control one are what we really concerns. Optimal signature is a set of these gene probes which can be used to summarizing the general cellular response to the drug. It usually consists of top-ranked \( K \) gene probes and
bottom-ranked $K$ gene probes, where $K$ is the size of optimal signature. As PRL is the merge of all ranked lists of the drug, the optimal signature gene probes are those gene probes that are consistently over- or under-expressed in all the transcriptional response to the drug, regardless of duration, cell line and concentration. The paper chooses 250 as the size of the optimal signature, which means that the top-ranked 250 gene probes and bottom-regulated 250 gene probes as optimal signature of each PRL (denoted by $p$ and $q$ respectively) are selected as the synthetic descriptor summarizing the general cellular response to the drug.

### 3.3.2 Computation of distance between two drugs

The paper uses the Gene Set Enrichment Analysis (GSEA) to calculate enrichment score of each pair of 1,309 drugs by comparing the corresponding PRLs and uses the enrichment score to compute distances between drugs. The smaller the distance is, the more similar the two drugs are. Hence, the paper gets 856,086 pairwise distances. This section uses the calculation of distance between drug $d$ and drug $x$ as an example to explain how to calculate the distance between two drugs using GSEA.

According to [17], given the representative signature of drug $d$, with top-regulated 250 gene probes list

$$p = \{p_1, p_2, \ldots, p_{250}\}$$
and bottom-regulated 250 gene probes list

\[ q = \{q_1, q_2, \ldots q_{250}\} \]

the distance between drug \( d \) and drug \( x \) is defined as the Inverse Total Enrichment Score (TES) of drug \( d \) signature \( \{p, q\} \) with respect to the PRL of drug \( x \), as:

\[ TES_{d,x} = 1 - \frac{ES_{x}^{p} - ES_{x}^{q}}{2} \]

The \( ES_{x}^{y} \), \( y = p \) or \( q \) represents enrichment score of the signature \( y \) with respect to \( x \), ranging in [-1,1]. It quantifies how many gene probes in \( y \) are at the top of \( x \). The closer to 1 the \( ES_{x}^{y} \) is, the more gene probes in \( y \) are at the top of \( x \). The closer to -1 the \( ES_{x}^{y} \) is, the more gene probes in \( y \) are at the bottom of \( x \). When \( ES_{x}^{p} \) is 1 and \( ES_{x}^{q} \) is -1, it means all the top-regulated gene probes of \( d \) are at the top of \( x \) and all the bottom-regulated gene probes of \( d \) are at the bottom of \( x \), which is the best situation. In the situation, the \( TES_{d,x} \) is equal to 0. When \( ES_{x}^{p} \) is -1 and \( ES_{x}^{q} \) is 1, it is the worst situation and in the situation \( TES_{d,x} \) is equal to 2. Hence, the \( TES_{d,x} \) ranges in [0,2].

The method to calculate the Enrichment Score (ES) is called Gene Set Enrichment Analysis (GSEA), which is based on the Kolmogorov-Smirnov (KS) statistic. The paper [18] shows how the enrichment score is calculated (see algorithm 2). Given the priori defined set of gene probes \( p \) and a full ranked list of gene probes \( B \), ES is calculated by walking down the list \( B \), if a gene in \( p \) is encountered, a increment will be add to a running-sum statistic, otherwise a punishment will be done to the statistic [line 8]. Then, ES is the statistic value which is the maximum derivation from 0 in the walk.

**Algorithm 2 Gene Set Enrichment Analysis**

```plaintext
procedure ENRICHMENT_SCORE(p,B)
2:   input: \( p \), the representative gene list of the drug A; \( B \), the PRL of the drug B
3:   output: \( ES(p,B) \); the enrichment score of \( p \) of drug A in the PRL of drug B,
4:   ES(p,B)
5:   \( N_R = size(p \cap B) \)
6:   \( N_H = size(p) \)
7:   \( N = size(B) \)
8:   for all \( i = 1 : n \) do
9:       \( P_{hit}(i) = \sum(g_j \in p, j \leq i)(\frac{1}{N_R}) \)
10:      \( P_{miss}(i) = \sum(g_j / p, j \leq i)(\frac{1}{N-1-N_H}) \)
11:      \( distance_{i}(i) = P_{hit}(i) - P_{miss}(i) \)
12: end for
13:   \( ES(p,B) = the \ max \ derivation \ from \ 0 \ of \ distance, \)
14: end procedure
```

Following are some examples.

\( B = \{"a", "b", "c", "d", "e", "f", "g", "h", "i", "j"\} \)

When \( p = P(A) = \{"a"\} \), \( ES_{B}^{p} = 1 \)
When \( p = P(A) = \{"a", "c", "b"\} \), \( ES^B_B = 1 \)
When \( p = P(A) = \{"j", "d", "f"\} \), \( ES^B_B = -0.4285714 \)
When \( p = P(A) = \{"a", "b", "d", "f"\} \), \( ES^B_B = 0.6667 \)
When \( p = P(A) = \{"j", "h", "i"\} \), \( ES^B_B = -\frac{9}{11} = -1 \)

As shown in the principle of the calculation of the \( TES_{d,x} \), it is clear that the \( TES_{A,B} \) is a single direction distance of drug \( d \) to drug \( x \). Hence, in order to equally consider both direction, it would be better the calculate the distance between drug \( d \) and \( x \) by combining the both direction distance as the following equations.

\[
AES(d, x) = \frac{TES_{d,x} + TES_{x,d}}{2} \tag{3.1}
\]
\[
MES(d, x) = \frac{\min(TES_{d,x}, TES_{x,d})}{2} \tag{3.2}
\]

It is clear that the Average Enrichment Score (AES) ranges in [0,2], while the Maximum Enrichment Score (MES) ranges in [0,1]. The smaller the distance is, the similar the two drugs are. The distance between each pair of 1,309 drugs is calculated, for a total number of \( \binom{1309}{2} = 856086 \) distances using both distance definitions. The paper chooses the 5% quantile of the distances as the threshold. Then, each drug is regarded as a node in a network. Each pair of drugs is connected if the distance is below 5% quantile. If a drug is not connected with at least one drug, the drug would be excluded from the network. The paper gets a Drug Network (DN) with 1,302 nodes and 41,047 edges.

### 3.4 Cluster algorithm

Clustering data based on a measure of similarity is a critical step in data analysis, which means to identify the similar data as a community according to the chosen similarity measurement. The paper uses affinity propagation algorithm to cluster the drugs, which is chosen because the number of clusters does not need to be specific at beginning of the algorithm and it considers each data point as a node of the network and considers the similarity between points as the weighted edge of the network. For the algorithm, clusters are defined by the exemplars and other nodes are assigned to one of the exemplars. The algorithm considers all the data as potential exemplars at the beginning and iteratively transmits messages along edges of the network. With the iteration of the algorithm, the clusters gradually emerge and the exemplars are selected based on the messages [22].

Following is the briefly description of the algorithm. \( \{x_1, x_2, \ldots x_n\} \) is the dataset needed to be classified and is assumed that there is no structure between data. The input of the algorithm is the real-valued similarity matrix \( S \). It has \( S(i, j) > S(i, k) \), if and only if the similarity between \( x_i \) and \( x_j \) is larger than that between \( x_i \) and \( x_k \). Affinity propagation algorithm carries out the exchange of two messages to update the following two matrix (see Figure 3.3):
• responsibility matrix $R$: $r(i, k)$ is sent from point $i$ to point $k$, indicating how well-suited the point $k$ is as the exemplar for point $i$

• availability matrix $A$: $a(i, k)$ is sent from point $k$ to point $i$, indicating how appropriate it is for point $i$ to choose the point $k$ as the exemplar

FIGURE 3.3: The message exchange between nodes

At the beginning, the two matrix are initialized to be 0 and the two messages are updated at each iteration using the following rules. Because the similarity matrix always use negative distances between data, the matrix $R$ and $A$ are negative and iteratively become more close to 0 or larger than 0.

\[
R(i, k) \leftarrow S(i, k) - \max_{k' \neq k} \{ A(i, k') + S(i, k) \}
\]

When $i \neq k$, $A(i, k) \leftarrow \min \{ 0, R(k, k) + \sum_{i' \neq i, i' \in \{i, k\}} \max \{ 0, R(i', k) \} \}$

When $i = k$, $A(i, k) \leftarrow \sum_{i' \neq i, i' \in \{i, k\}} \max \{ 0, R(i', k) \}$

Besides, in order to avoid the oscillations, affinity propagation algorithm uses damping factor $\lambda$, which ranges in $(0,1)$. The larger the damping factor, the faster the algorithm converges. Using the damping factor, the update of two messages at $t + 1$ iteration become:

\[
R(i, k) \leftarrow (1 - \lambda)R_{t+1}(i, k) + \lambda R_t(i, k)
\]

\[
A(i, k) \leftarrow (1 - \lambda)A_{t+1}(i, k) + \lambda A_t(i, k)
\]

At any point of iteration, responsibility and availability can be combined to identify the exemplar. To be specific, the algorithm find the data index whose diagonal element of the $E = R + S$ is larger than 0 and regards the data as one exemplar. For example, the $i$th diagonal element of $E$ is larger than 0, the algorithm chooses data $i$ as one exemplar.
Then, after the exemplars are identified, each data is assigned to one exemplar according to the similarity matrix. For example, there are 5 data and the first and second data are exemplars. For the third data, the similarity between the third data and other data is $S[3, \cdot] = \{-1, -5, 0, -7, -4\}$ and the third data chooses data 1 as exemplar, because the similarity between the third data and first exemplar is larger than that between third data and second exemplar.

The Affinity Propagation algorithm is described in algorithm 3. The input of the algorithm is the real-valued similarity matrix $S$ and damping factor $\lambda$. At the beginning, the two messages of all pair of nodes are initialized to be 0. Then, the algorithm updates the responsibility $R(i, j)$ [line 13] and availability $A(i, j)$ [line 20] for each pair of nodes which have connection in the network. The process will continue until the iteration number satisfies the maximum iteration number, which can be changed. After the update, responsibility matrix $R$ and availability matrix $A$ are combined to determine the exemplars [line 27]. Based on the sum of $R$ and $A$, the algorithm finds all the data indices whose diagonal elements of the sum matrix is larger than 0 [line 30]. Finally, based on the exemplars selected above and similarity matrix $S$, each drug is assigned to the corresponding cluster [line 33].

As mentioned above, the drug network is built by regarding each drug as a vertex and regarding the connection between two drugs as an weighted edge. In the problem, 1,309 drugs are needed to be classified. When applying the affinity propagation algorithm on the problem, the input of the algorithm is distance matrix calculated above and the damping factor. Using the clustering algorithm, the paper identifies 106 drug communities.

### 3.5 Implementation of the framework

This section focuses on the implementation of the framework, including which language is used to implement the framework and the performance of the code.

Before re-implementing the framework, what we should do first is to obtain the CMap data and do some necessary pre-processing work on the data, which is introduced in section 3.5.1. Then, based on the results of the pre-processing work, section 3.5.2 introduces how we re-built the drug network. Finally, section 3.5.3 describes how we implemented the affinity propagation algorithm on the problem.

#### 3.5.1 Pre-processing of CMap data

Originally, we downloaded the CMap data and metadata at http://portals.broadinstitute.org/cmap/, which are raw and machinery data. The pre-processing work of the data costed us several weeks and finally it was not work. The detail of the preprocessing work based on raw CMap data can be seen in Appendix A. Therefore, we
Algorithm 3 Affinity Propagation Algorithm

procedure AP(S, λ)
input: S, the similarity between data. It has \( S(i, j) > S(i, k) \), if and only if the similarity between \( i \) and \( j \) is larger than that between \( i \) and \( k \)
output: exemplars, an array indicating which data are the exemplars
n_clusters_, the number of clusters
labels, an array indicating which cluster the data belongs to

1. \( N = 10000 \)
drug_number = |S|

2. for all iter = 1 : N do
3. \( Rold = R \)
4. for all \( i = 1 : drug_number \) do
5. for all \( k = 1 : drug_number \) do
6. \( R(i, k) \leftarrow S(i, k) - \max_{k', s.t. k' \neq k} \{ A(i, k') + S(i, k') \} \)
7. end for
8. end for
9. \( R \leftarrow (1 - \lambda) R + \lambda Rold \)
Aold = A

10. for all \( i = 1 : drug_number \) do
11. for all \( k = 1 : drug_number \) do
12. if (\( k! = i \))
13. \( A(i, k) \leftarrow \min \{ 0, R(k, k) + \sum_{i', s.t. i' \neq i, k} \max \{ 0, R(i', k) \} \} \)
14. else
15. \( A(i, i) \leftarrow \sum_{i', s.t. i' \neq k} \max \{ 0, R(i', k) \} \)
16. end if
17. end for
18. end for
19. \( A \leftarrow (1 - \lambda) A + \lambda Aold \)
20. end for
21. \( E = R + A \)
22. exemplars = find(diag(E) > 0)
n_cluster_ = |exemplars|
23. for all \( i = 1 : drug_number \) do
24. \( k = \max_{j \in \text{exemplars}} S \)
25. labels[i] \leftarrow k
26. end for
27. end procedure
changed to use the data package in R called "ConnectivityMap". The ranked differential gene expression lists of CMap data can be obtained directly using the data package. It contains two data objects:

- \textit{rankMatrix}: 22,283 rows \times 6,100 "instances"
- \textit{instances}: 14 columns of metadata describing each of the 6,100 instances.

The \textit{rankMatrix} object contains 6,100 ranked gene probe lists with the same order of 22,283 gene probes, which means that \textit{rankMatrix}[i,j] represents the rank of \textit{i}th gene probe in the \textit{j}th instance and that \textit{rankMatrix}[i,] is a vector representing the rank of \textit{i}th gene probe in all 6,100 instances. The \textit{instances} object contains the metadata of CMap data, which is a 14 \times 6100 matrix and followings are some important and useful columns of the matrix:

- \textit{instance_id}
- cmap\_name: drug name
- concentration(M)
- duration(h)
- cell2: cell name
- perturbation\_scan\_id: data file name of the instance.
- vehicle\_scan\_id4: vehicle data file name of the instance.

No other pre-processing work should be done other than some statistics work based on the metadata data to know more about CMap data, because the data package has got the ranked lists which can be used directly in later work. Table 3.1 shows the metadata of some instances in CMap.

<table>
<thead>
<tr>
<th>instance_id</th>
<th>cell2</th>
<th>cmap_name</th>
<th>concentration(M)</th>
<th>duration(h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MCF7</td>
<td>metformin</td>
<td>0.00001</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>MCF7</td>
<td>metformin</td>
<td>0.00001</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>MCF7</td>
<td>metformin</td>
<td>0.001</td>
<td>6</td>
</tr>
<tr>
<td>21</td>
<td>MCF7</td>
<td>phenformin</td>
<td>0.00001</td>
<td>6</td>
</tr>
<tr>
<td>460</td>
<td>PC3</td>
<td>deferoxamine</td>
<td>0.0001</td>
<td>6</td>
</tr>
<tr>
<td>485</td>
<td>MCF7</td>
<td>deferoxamine</td>
<td>0.0001</td>
<td>6</td>
</tr>
<tr>
<td>573</td>
<td>MCF7</td>
<td>deferoxamine</td>
<td>0.0001</td>
<td>12</td>
</tr>
<tr>
<td>825</td>
<td>MCF7</td>
<td>rottlerin</td>
<td>0.00001</td>
<td>6</td>
</tr>
</tbody>
</table>

From the Table 3.1, it describes the following situations:

- One drug may work on the same cell line with same duration and different concentrations, such as instance 2 and 3.
• One drug may work on different cell lines with the same dose and duration, such as instance 460 and 485.

• One drug may work on the same cell line with same concentration and different durations, such as instance 485 and 573.

• The cell line may be treated with different drugs with same or different durations and concentrations, such as instance 21 and 852 (with different doses and different drugs), and instance 1 and 852 (with different drugs).

It is very clear that one instance is decided by four factors, cell line, drug, dose and duration. In the case, the first instance and the second instance in the table are repeated experiments which create two instances in the dataset. Hence, the number of unique instances was counted and there were 3,742 unique instances in CMap data, which means that if one instance which is repeated several times will be counted just once.

Besides, as mentioned in section 3.2, the difference between instance number for each cell line of the same drug may lead to unequal consideration of all instances of the same drug, so based on the distinct instances, we summarized how many distinct instances of each cell line and the result shows that the number of instances of MCF7, PC3 and HL60 nearly occupies 90% of the distinct instances, see Table 3.2.

<table>
<thead>
<tr>
<th>cell line</th>
<th>number of distinct instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssMCF7</td>
<td>16</td>
</tr>
<tr>
<td>HL60</td>
<td>1078</td>
</tr>
<tr>
<td>SKMEL5</td>
<td>17</td>
</tr>
<tr>
<td>MCF7</td>
<td>1294</td>
</tr>
<tr>
<td>PC3</td>
<td>1182</td>
</tr>
</tbody>
</table>

Finally, as shown in section 3.2, the merge of all instances of the same drug plays an critical role, so we did statistics work on how many drugs in CMap data are tested on at least two cell lines to verify the merge is valuable, which means for these drugs they have at least two ranked gene probe lists which are suitable for being merged to generate single PRL. We got the result that nearly 94% of the 1,309 drugs in the CMap dataset are tested on at least two cell lines and only 6% of 1,309 drugs are tested on just one cell line.

### 3.5.2 Re-build the drug network

**Merge the ranked lists of each drug**

In order to merge the ranked lists of each drug, firstly we read all the ranked lists from *rankMatrix* object and metadata from *instances* object in "ConnectivityMap"
package of R. Then, based on the metadata, we split the 6,100 ranked gene lists into 1309 .csv files and each file contains all the ranked lists for one drug (see code pre_work.R). Finally, based on the 1,309 files, we implemented the KRUBOR algorithm. As mentioned in section 3.2.3, the input of the algorithm is a set of ranked lists of Microarray Probe-set Identifiers (MPI) of the same drug and then using r function to work on it to get the ranked lists with the same order of 22,283 gene probe list P to do further merge. The output of each merge is a new ranked gene probe list. Because the ranked lists in rankMatrix is the same as the result of r function, when we implemented the algorithm, the input of the algorithm is all the ranked lists with the same order of 22,283 gene probe list P of the same drug, other than X, and the output of each merge is a new ranked list with the same order of 22,283 gene probe list P. In order to get the ranked gene probe list at the end of merge, we sorted the gene probes in P according to their value in the last ranked lists in an increasing order (see code aggregation.R). Besides, according to P, each gene probe has a number which is the location of the gene probe in P and we used this number to represent the gene probe in final ranked gene probe list. Hence, the final merged ranked gene probe list is a numeric vector and each number represent the corresponding gene probe in P. Using the same example as section 3.2.2, the output of the example using our program is \{4, 1, 5, 2, 3\} rather than \{gene_4, gene_1, gene_5, gene_2, gene_3\}. The output of the merge program is a .csv file which contains 1,309 single PRL for 1,309 drugs. The flow of the merge can be seen in Figure 3.4.

As for the performance of the above two program, pre_work.R costs just a few seconds and aggregation.R costs around half an hour, which is acceptable.

![Figure 3.4: The flow of merge](image-url)
Calculation of the distance

As mentioned in section 3.3, we selected the 250 top-regulated gene probes and 250 bottom-regulated gene probes as optimal signature of each PRL. Then, we implemented the GSEA algorithm to calculate the enrichment score and Inverse Total Enrichment Score (TES). Finally, based on the TES, the distances of drugs were calculated using both Average enrichment score (AES) and Maximum enrichment score (MES). The input of the algorithm is the 1,309 single PRL of drugs which had been obtained above.

Originally, we implemented the algorithm using R language. Because R is not good at dealing with for loop and the algorithm has a large for loop, the performance of the R code is undesirable. Even thought it sounds good that it costs nearly several second to calculate one TES, there are \( \binom{1309}{2} = 856086 \) TES needed to be calculated and the time cost of the R code is predicted to at least 40 days, which is not practical.

Therefore, we re-implemented the algorithm using C language and paralleled it using Message-passing interface. The program is run on Morar and each process calculates the TES between one specific drug with all other drugs and outputs an 1309 × 1 vector. As we had 1,309 drugs and needed 1,309 processes, which exceeds the maximum number of processes of Morar, so the program uses 200 processes to calculate TES of 200 drugs at once and it is run 7 times to calculate the TES of all drugs, which totally costs one and half an hour. The output of the program is 1,309 1309 × 1 vectors (see code ES.c). Then, the 1,309 vectors are combined to form a 1309 × 1309 matrix, denoted by \( E \), which contains all pairwise TES. Finally, based on the matrix, the AES distance matrix and MES distance matrix can be calculated, denoted by \( A \) and \( M \) respectively. The output is two .csv files, one for AES distance matrix and the other for MES distance matrix (see code distance.R). The performance of distance.R is great and it costs around two minutes.

\[
A = \frac{E + \text{transpose}(E)}{2}
\]

\[
M[i,j] = \frac{\min\{E[i,j], E[j,i]\}}{2}
\]

As mentioned above, the number of pairwise distance values (856,086) is huge, so the empirical probability density function estimated from the data is assumed to be a good approximation of the real distance values. The empirical Probability Density Function (pdf) of AES and MES can be seen in Figure 3.5 and Figure 3.6 respectively. The red line in both two figures is the 5% quantile line.

Then, the 5% quantile of the empirical pdf is selected to be the distance significance threshold value, which is 0.8337 of AES distance and 0.4033 of MES distance (see Table 3.3), comparing to 0.8327 of AES distance and 0.4032 of MES distance in [7]. Table 3.3 also shows some statistic information of both two distances.

Then, each drug was presented by a vertex in the network. If the distance between two drugs is below the threshold, the two corresponding vertices will be connected by
Figure 3.5: The empirical probability density function of AES

Figure 3.6: The empirical probability density function of MES
Table 3.3: Statistical summary of AES distance and MES distance

<table>
<thead>
<tr>
<th></th>
<th>5% quantile</th>
<th>mean</th>
<th>variance</th>
<th>median</th>
<th>maximum</th>
<th>minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>AES distance</td>
<td>0.8337</td>
<td>0.9653</td>
<td>0.0066</td>
<td>0.9653</td>
<td>1.3364</td>
<td>0.1182</td>
</tr>
<tr>
<td>MES distance</td>
<td>0.4033</td>
<td>0.4686</td>
<td>0.0017</td>
<td>0.4659</td>
<td>0.6512</td>
<td>0.0613</td>
</tr>
</tbody>
</table>

a weighted edge. If the node is not connected by at least one edge with other nodes, the node will be removed from the network. Hence, we created a drug network with 1,305 vertices and 42,805 based on the AES distance and a drug network with 1,302 vertices and 42,805 edges based on MES distance, respectively. However, our drug network based on MES distance is different from the drug network with 1,302 vertices and 41,048 edges based on MES distance in [7].

3.5.3 Implementation of cluster algorithm

As mentioned in section 3.4, the affinity propagation algorithm is chosen to identify the drug communities. In section 3.5.2, the drug network built based on MES distance has 1,302 vertices, so we just took the 1,302 drugs into consideration to do the classification. Hence, first of all, we remove the drugs with no edge with other drugs from the MES distance matrix obtained above and got a $1302 \times 1302$ distance matrix (see code create_network.R). Then, we used the $1302 \times 1302$ distance matrix as the input of the algorithm. We implemented the affinity propagation algorithm using the AffinityPropagation function in the package sklearn.cluster of PYTHON (see code cluster.py). The output of the program is two .csv file, one for the exemplars and the other for drugs with its corresponding exemplar. After the cluster procedure, we identified the 87 clusters, which is different from the cluster result, 106 communities, in [7]. Until then, the framework has been re-implemented. The cluster.py costs no more than one minutes.
Chapter 4

Sensitivity analysis
4.1 Introduction

This chapter mainly focuses on the sensitivity analysis of the framework of drug repositioning based on MES distance. As mentioned in section 2.5, the sensitivity analysis is to analyze how the drug repositioning of the framework changes with the change of parameters. If the drug repositioning is nearly not affected by the change of parameters, the framework can be regarded as a robust method to do drug repositioning. In order to test whether the framework is robust or not, what we should do is to test whether the drug communities are changed or not with the change of parameters.

Before doing the sensitivity analysis, what we should do first is to choose a measurement to compare the similarity between networks, which is described in section 4.2.

Then, Two parameters are chosen to re-implement the framework. The first one is the distance threshold. Once the distances between drugs are calculated, the threshold distance value is chosen to build the DN. In the framework, the threshold value is assigned to the 5% quantile of all 856,086 distance values. If the distance between two drugs is below the distance threshold value, then the two drugs will be connected by an edge in the drug network and the element of the two drugs will be 1 in the adjacency matrix of the drug network. If the drug is not connected by at least one edge, the drug will be removed from the network. The threshold of the distance is changed to 2% quantile, 3% quantile, 4% quantile, 6% quantile, 7% quantile, 10% quantile, 20% quantile and 30% quantile to re-implement the framework, which is described in detail in section 4.3. The results of the framework with different distance thresholds are also shown in the section 4.3.

Another parameter is the size of optimal signatures, $K$. In the original framework, $K$ is assigned to 250, which affects the calculation of the distances between drugs. In section 4.4, the size of optimal signature is changed to 1, 2, 5, 10, 20, 50, 75, 100, 200, 300, 350 and 450 to re-implement the framework. The results of the framework with different size of optimal signature are also shown in the section 4.4.

In section 4.5, based on the measurement in section 4.2 and the re-implementation results in the above two section, we did sensitivity analysis on DN of the framework to see how the DN is affected by parameters.

Then, in section 4.6, we come up with a method to compare drug communities of different parameters. Based on the measurement mentioned in 4.2, the re-implemented results in section 4.3 and section 4.4 and the method which we come up with, we did sensitivity analysis of drug communities to see how the drug repositioning is affected by the two parameters.

Finally, based on the analysis in the section 4.5 and section 4.6, whether the framework is robust or not is analyzed in final section.
4.2 Measurement to assess the difference between networks

As mentioned above, we need to compare the drug networks of different parameters, so we need a measurement to describe how similar two networks are.

Based on the network theory mentioned in section 2.4, each network can be represented by a square (0,1)-matrix, adjacency matrix. In the matrix, the element of the matrix indicates whether the pair of nodes is connected or not. If the element is 1 between node \(i\) and node \(j\), there is an edge between node \(i\) and node \(j\) in the network, otherwise there is no edge between the two nodes. Therefore, the network can be regarded as a binary classification, each pair of nodes is classified to 1 or 0. The network obtained above is the control one and the networks built from different parameters can be regarded as different classifiers. The closer to the control one, the better the network is.

In a binary classification problem, the task is to assign each object to one categories: the positive or negative one. Based on the confusion matrix mentioned in section 2.5, four statistical variables, Accuracy, Precision (Positive Predicted Value), True positive rate (TPR or Sensitivity) and False positive rate (FPR or 1-Specificity), can be calculated to assess the classifier. Accuracy indicates how many objects are correctly classified, reflecting the general discrimination ability of the classifier. Precision indicates how many positive predicted objects are really positive. TPR indicates how many correct positive results occur among all actual positive samples, reflecting discrimination ability of positive category, and FPR indicates how many incorrect positive results occur among all actual negative samples, reflecting the discrimination ability of negative category. All four statistic variables range from 0 to 1. The better the classifier is, the larger the Accuracy, the Precision and the TPR are and the smaller the FPR is.

\[
\text{Accuracy} = \frac{TP + TN}{\text{Total number of samples}}
\]  
(4.1)

\[
\text{Precision} = \frac{TP}{TP + FP}
\]  
(4.2)

\[
\text{TPR}(\text{sensitivity}) = \frac{\sum \text{True Positive} (TP)}{\sum \text{Condition Positive}}
\]  
(4.3)

\[
\text{FPR}(1 - \text{specificity}) = \frac{\sum \text{False Positive} (FP)}{\sum \text{Condition Negative}}
\]  
(4.4)

Therefore, based on the measurement mentioned above, the closer to the control network means the larger Accuracy, Precision and TPR and the smaller the FPR.

4.3 Distance threshold

As common knowledge, applying the same data with different thresholds, the smaller the threshold is, the stricter condition it is. Hence, the smaller the threshold is, the less
edge of the network will be kept. Considering whether there is enough connections between nodes to be clustered, the choose of distance threshold is very critical.

The distance threshold value is originally assigned to the 5% quantile of all 856,086 distances. It will be changed to 2% quantile, 3% quantile, 4% quantile, 6% quantile, 7% quantile, 10% quantile, 20% quantile and 30% quantile, which are simply represented by threshold later and the value of threshold is one of the set

\[(0.02, 0.03, 0.04, 0.06, 0.07, 0.1, 0.2, 0.3)\]

Keeping the size of optimal signature \(K\) unchanged, for each threshold, based on the same MES distance, if the distance between two drugs is below the corresponding quantile distance value, there will be an edge between two drugs. Based on these, DNs of different thresholds were built. The details of the DNs of different thresholds can be seen in Table 4.1. Then, the DNs were applied with the affinity propagation algorithm to identify the communities. The number of drug communities of different thresholds can be seen in Table 4.2.

Table 4.1: Drug network information of different thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>node</td>
<td>1219</td>
<td>1273</td>
<td>1294</td>
<td>1306</td>
<td>1306</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
</tr>
<tr>
<td>edge</td>
<td>17122</td>
<td>25683</td>
<td>34245</td>
<td>51366</td>
<td>59926</td>
<td>85611</td>
<td>17122</td>
<td>256828</td>
</tr>
</tbody>
</table>

Table 4.2: Cluster network information of different thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>community</td>
<td>165</td>
<td>140</td>
<td>111</td>
<td>74</td>
<td>64</td>
<td>38</td>
<td>16</td>
<td>10</td>
</tr>
</tbody>
</table>

Based on the re-implementation results, before comparing the drug networks built from these threshold values with the control one, we could see how the threshold value affect the topology of the drug network first.

When the threshold value is 0.02, the drug network is built by linking two drugs with an edge if their MES distance value is below the 2% quantile threshold. If the drug is not connected by at least one edge, it will be removed. The drug network consists of 1,219 drugs and 17,122 connections. Figure 4.1 (a) present the sub-network of the whole drug network. The sub-network contains one connected components, which consists of the following drugs: cloxacillin, pentoxifylline, iproniazid and meclofenamic acid. All of the drugs are mainly used as the inhibitor of bacterial or cell to avoid some diseases.

If the drug network is built by linking two drugs with an edge if their MES distance value is below the 3% quantile threshold, the whole drug network consists of 1,273 drugs and 25,683 connections. Figure 4.1 (b) presents the sub-network of the whole drug network. With respect to the previous network, the sub-network contains two components. The new component consists of two new drugs, midodrine (the ACE inhibitor)
and gentamicin (an antibiotic). Besides, two new connections, the connection between mefexamide and pentoxifylline and the connection between cloxacillin and pentoxifylline, appear in the first component. The new drug, mefexamide, is the inhibitor of adrenal cortical.

Further increasing the distance threshold to 4% quantile yields a drug network consisting of 1,294 drugs and 34,245 connections. The sub-network is presented in Figure 4.1 (c). Some new connections appear, such as the connection between tranexamic acid and midodrine and the connection between DL-PPMP and pentoxifylline.

The drug networks when threshold is equal to 0.06 and 0.07 consist of 1,306 drugs with 59,926 connections and 1,306 drugs with 59,926 connections, respectively. If the threshold value is increasing to 0.1, all the drugs will be concluded in the drug network. The change of the number of drugs and the change of the number of connections can be seen in Table 4.1 and Figure 4.2. Until then, it is clear that when the distance threshold value increases, the number of drug nodes and connections between drugs increases, which means that the drug network with smaller threshold is a sub-network of the drug network with larger threshold.

### 4.4 The size of optimal signature K

$K$ is the size of optimal signature. It determines how many gene probes are used to calculate the distance. For example, if the size of optimal signature $K$ is 20, it means that the 20 top-regulated gene probes and 20 bottom-regulated gene probes are selected as optimal signature to be used to calculate the distance. The larger the size of optimal signature is, the more gene probes are used to calculate the distances between drugs and the more real the distances are. Therefore, in order to get more real distances, it would be better to use larger $K$. However, the Gene Set Enrichment Analysis aims to use the up-regulated gene probe lists and bottom-regulated gene probe lists to calculate the enrichment score and calculate the distance between drugs. If the $K$ is too large, some irrelevant gene probes, whose differential gene expression maybe very close to 0, locating at the medium part of the ranked gene probe list, are considered. In addition to this, the increase of $K$ may increase the work load of the calculation of distance and increase the time consumption. Therefore, the choose of the size of optimal signature is very critical.

The size of optimal signature is originally assigned to 250 in the above framework. In this section, it will be changed to 1, 2, 5, 10, 20, 50, 75, 100, 200, 300, 350 and 450 and the MES distances of these $K$ are calculated. The statistic information of the distances of different $K$ can be seen in Table 4.3. Based on the distances, with the same 5% quantile distance threshold, the DNs can be built. For each $K$, if the distance between two drugs is below the 5% quantile, there will be an edge between the two drugs. The details of the drug networks of different $K$ can be seen in Table 4.4. Then, these drug networks are applied with affinity propagation algorithm to identify drug communities.
Figure 4.1: Network change versus threshold

Figure 4.2: The number of drugs and edges versus threshold
The number of drug communities of different $K$ can be seen in Table 4.12.

<table>
<thead>
<tr>
<th>$K$</th>
<th>5%</th>
<th>mean</th>
<th>variance</th>
<th>median</th>
<th>max</th>
<th>min</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0315</td>
<td>0.3151</td>
<td>0.0516</td>
<td>0.3936</td>
<td>0.9966</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.1217</td>
<td>0.3536</td>
<td>0.0286</td>
<td>0.3333</td>
<td>0.9703</td>
<td>0.0003</td>
</tr>
<tr>
<td>5</td>
<td>0.2307</td>
<td>0.3936</td>
<td>0.0136</td>
<td>0.3707</td>
<td>0.8353</td>
<td>0.0003</td>
</tr>
<tr>
<td>10</td>
<td>0.2915</td>
<td>0.4163</td>
<td>0.0078</td>
<td>0.3982</td>
<td>0.7533</td>
<td>0.0008</td>
</tr>
<tr>
<td>20</td>
<td>0.3359</td>
<td>0.4344</td>
<td>0.0047</td>
<td>0.4219</td>
<td>0.7317</td>
<td>0.0027</td>
</tr>
<tr>
<td>50</td>
<td>0.3733</td>
<td>0.4515</td>
<td>0.0028</td>
<td>0.4446</td>
<td>0.6806</td>
<td>0.0196</td>
</tr>
<tr>
<td>75</td>
<td>0.3841</td>
<td>0.4571</td>
<td>0.0024</td>
<td>0.4519</td>
<td>0.6733</td>
<td>0.0223</td>
</tr>
<tr>
<td>100</td>
<td>0.3902</td>
<td>0.4604</td>
<td>0.0021</td>
<td>0.4563</td>
<td>0.6666</td>
<td>0.0318</td>
</tr>
<tr>
<td>200</td>
<td>0.4009</td>
<td>0.4669</td>
<td>0.0018</td>
<td>0.4642</td>
<td>0.6508</td>
<td>0.0465</td>
</tr>
<tr>
<td>250</td>
<td>0.4033</td>
<td>0.4686</td>
<td>0.0017</td>
<td>0.4659</td>
<td>0.6521</td>
<td>0.0513</td>
</tr>
<tr>
<td>300</td>
<td>0.4049</td>
<td>0.4698</td>
<td>0.0017</td>
<td>0.4673</td>
<td>0.6573</td>
<td>0.0536</td>
</tr>
<tr>
<td>350</td>
<td>0.4063</td>
<td>0.4707</td>
<td>0.0016</td>
<td>0.4681</td>
<td>0.6605</td>
<td>0.0628</td>
</tr>
<tr>
<td>450</td>
<td>0.4078</td>
<td>0.4722</td>
<td>0.0016</td>
<td>0.4697</td>
<td>0.6677</td>
<td>0.0747</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$K$</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>node</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
</tr>
<tr>
<td>edge</td>
<td>42817</td>
<td>42817</td>
<td>42805</td>
<td>42810</td>
<td>42806</td>
<td>42807</td>
</tr>
<tr>
<td>$K$</td>
<td>75</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>350</td>
<td>450</td>
</tr>
<tr>
<td>node</td>
<td>1309</td>
<td>1309</td>
<td>1304</td>
<td>1301</td>
<td>1300</td>
<td>1297</td>
</tr>
<tr>
<td>edge</td>
<td>42809</td>
<td>42805</td>
<td>42805</td>
<td>42807</td>
<td>42805</td>
<td>42805</td>
</tr>
</tbody>
</table>

Table 4.4: Network information of different $K$

Before doing sensitivity analysis, some simple statistic analysis can be done on the information of these distances. Figure 4.3 (a) shows that with the increase of the $K$, the 5% quantile value increases rapidly from $K = 1$ to $K = 100$. When the $K$ is larger than 200, the 5% quantile keeps nearly constant. The mean distance value of these distances shows the similar situation with the 5% quantile value. It increases rapidly from 0 to 100 and nearly keeps constant after $K$ is larger than 200, see Figure 4.3 (b). However, the maximum distance of these distances shows an opposite trend with above two statistic variables. It decreases rapidly from 0 to 100 and nearly keeps stable after $K$ is larger than 100, see Figure 4.3 (c). The minimum distance shows a nearly linear increase with the increase of $K$, see Figure 4.3 (d).
If $K$ is equal to 1, which means that only the first-regulated and last-regulated gene probes are used to calculate the distance, the distances of drugs will be discrete, which may lead to very large distance if both first-regulated and last-regulated gene probes are different, regardless of the situation that the first-regulated gene of one ranked list may located at the second place of the other ranked list, and no distance if both first-regulated and last-regulated gene probes are the same, regardless of the situation that the second-regulated gene probes are different. It can be seen in Figure 4.3 (c) and (d) that the maximum distance of the network when $K = 1$ is the largest among all $K$ and the minimum distance of the network when $K = 1$ is the smallest with 0 among all $K$. With the increase of $K$, the difference between similar drugs becomes larger, because more gene probes are used to calculate the distance and it increases the probability of the difference between two ranked gene lists, which leads to the increase of the minimum distance. Due to the nearly same principle, the increase of $K$ decreases the probability of the difference between two dissimilar lists, which leads to the decrease of the maximum distance. Moreover, because the maximum distance gets smaller and the minimum distance gets larger with the increase of $K$, the distances become more dense, which leads to the increase of mean distance and 5% quantile value.

Figure 4.3: Statistic analysis on the distances versus $K$
4.5 Sensitivity analysis on drug network

Based on the measurement mentioned in section 4.2 and the implementation results of above two section, this section focuses on sensitivity analysis of drug networks on distance threshold (section 4.5.1) and the size of optimal signature (section 4.5.2).

4.5.1 Distance threshold

Based on the DNs of different distance thresholds in section 4.3, each drug network was represented by its adjacency matrix. The drug network built from the default parameter was also represented by its adjacency matrix as the control one. Then, using the measurement in section 4.2 to compare the adjacency matrix of each threshold with the control one, the confusion matrix of each threshold was calculated, see Appendix B. For each threshold, based on its confusion matrix, the Accuracy, Precision, TPR and FPR were calculated, see Table 4.6 and Figure 4.4.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>1-Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.9699995</td>
<td>1</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>0.03</td>
<td>0.9799997</td>
<td>1</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>0.04</td>
<td>0.990001</td>
<td>1</td>
<td>0.8000234</td>
<td>0</td>
</tr>
<tr>
<td>0.06</td>
<td>0.9899998</td>
<td>0.8333333</td>
<td>1</td>
<td>0.0105265</td>
</tr>
<tr>
<td>0.07</td>
<td>0.9800008</td>
<td>0.7142967</td>
<td>1</td>
<td>0.02105176</td>
</tr>
<tr>
<td>0.1</td>
<td>0.949998</td>
<td>0.4999943</td>
<td>1</td>
<td>0.05263371</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8499964</td>
<td>0.2499985</td>
<td>1</td>
<td>0.15789868</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7499982</td>
<td>0.1666668</td>
<td>1</td>
<td>0.26315997</td>
</tr>
</tbody>
</table>

Table 4.6: Statistic information of comparison between networks versus threshold

As for the Accuracy, in Figure 4.4 (a), it is clear that it nearly keeps constant when the threshold is smaller than 0.07 and then it decreases with the increase of threshold. Because the dominator of Accuracy is the total number of edges which is unchanged for all thresholds, it can indicate that the number of edges that are correctly predicted, 1 or 0, nearly keeps unchanged when the threshold is smaller than 0.07 and then decreases with the increase of threshold. Table 4.6 shows that when the threshold is equal to 0.06, the Accuracy is the largest (0.99) and when the threshold is equal to 0.03, the Accuracy is the smallest (0.75).

The Precision has a similar situation. Figure 4.4 (b) shows that it keeps unchanged at 1 when the threshold is smaller than and equal to 0.04 and then decreases to only 0.16 when the threshold is equal to 0.3. From the confusion matrix in Appendix B, it is clear that the decrease of the Precision when the threshold is larger than 0.04 is because the increase of FP while the TP is unchanged.

As for Sensitivity, Figure 4.4 (c) shows that the Sensitivity increases until the threshold is equal to 0.06 and then keeps unchanged at 1 until threshold value is equal to 0.3.
Because the dominator of the Sensitivity is the total number of real positive edges which is unchanged for all thresholds, it means that the number of positive edges that are correctly identified increases first when the threshold value is smaller than 0.06 and then all the positive edges are correctly identified when the threshold is larger than 0.06.

Figure 4.4 (d) shows that 1-Specificity increases with the increase of the threshold value. It means that the Specificity decreases with the increase of the threshold value. Because the dominator of all TPR is the total number of real negative edges which is unchanged for all thresholds, it indicates that less negative edges are correctly identified with the increase of the threshold.

As for why the 1-Specificity for all drug networks is near 0, it is because that the distance quantile, such 5% quantile, just corresponds to the minority of 856,086 edges and the majority of the elements of the adjacency matrix of drug network are 0.

The result is obvious, because in section 4.3 we have analyzed that the network with smaller threshold value is the sub-network of the network with larger threshold value, which is the reason why compared with the drug network with 0.05 threshold, the Precision of the network with less than 0.05 threshold value is 1 and the Sensitivity of the network with larger than 0.05 threshold value is 1.
4.5.2 The size of optimal signature

Based on the DNs of different $K$ in section 4.4, each drug network was represented by its adjacency matrix. The drug network built from the default parameter was also represented by its adjacency matrix as the control one. Then, using the measurement mentioned in section 4.2 to compare the adjacency matrix of each $K$ with the control one, the confusion matrix of each $K$ was calculated, see Appendix B. For each $K$, based on its confusion matrix, the Accuracy, Precision, TPR and FPR were calculated, see Table 4.7 and Figure 4.5.

<table>
<thead>
<tr>
<th>K</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>1-Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9139759</td>
<td>0.1398743</td>
<td>0.1399136</td>
<td>0.045283242</td>
</tr>
<tr>
<td>2</td>
<td>0.9168728</td>
<td>0.1688348</td>
<td>0.1688821</td>
<td>0.043758553</td>
</tr>
<tr>
<td>5</td>
<td>0.9261313</td>
<td>0.2613246</td>
<td>0.2613246</td>
<td>0.038878321</td>
</tr>
<tr>
<td>10</td>
<td>0.9348255</td>
<td>0.3482831</td>
<td>0.3483238</td>
<td>0.034305486</td>
</tr>
<tr>
<td>20</td>
<td>0.9452637</td>
<td>0.4526468</td>
<td>0.4526574</td>
<td>0.028809231</td>
</tr>
<tr>
<td>50</td>
<td>0.961088</td>
<td>0.6108814</td>
<td>0.6109099</td>
<td>0.020481236</td>
</tr>
<tr>
<td>75</td>
<td>0.968377</td>
<td>0.6837581</td>
<td>0.683822</td>
<td>0.016646153</td>
</tr>
<tr>
<td>100</td>
<td>0.9739068</td>
<td>0.7390725</td>
<td>0.7390725</td>
<td>0.013733261</td>
</tr>
<tr>
<td>200</td>
<td>0.9891857</td>
<td>0.8918584</td>
<td>0.8918584</td>
<td>0.00569176</td>
</tr>
<tr>
<td>300</td>
<td>0.9911457</td>
<td>0.9114397</td>
<td>0.9114823</td>
<td>0.004661366</td>
</tr>
<tr>
<td>350</td>
<td>0.9878003</td>
<td>0.8780049</td>
<td>0.8780049</td>
<td>0.006420905</td>
</tr>
<tr>
<td>450</td>
<td>0.9833358</td>
<td>0.8333606</td>
<td>0.8333606</td>
<td>0.008770646</td>
</tr>
</tbody>
</table>

As for the Accuracy, in Figure 4.5 (a), it is clear that it has a small increase when $K$ is smaller than 200 and then it is nearly constant to be close to 1 with the increase of $K$. Because the dominator of Accuracy is the total number of edges which is unchanged for all $K$, it can indicate that the number of edges that are correctly predicted has a small increment and then nearly keeps unchanged to be extremely close to the total number of edges.

As for the Precision, Figure 4.5 (b) shows that with the increase of $K$, it increases first from 0 to 300 and then slowly decreases. Table 4.7 shows that when $K$ is equal to 300, the precision is the largest (0.9114) and when $K$ is equal to 1, the precision is the smallest (0.1398). From the confusion matrix in Appendix B, it is clear that the increase of the Precision when $K$ is smaller than 300 is because the increase of TN and the decrease of FP and the decrease of the Precision when $K$ is larger than 300 is because the decrease of TN and the increase of FP.

Figure 4.5 (c) shows that the Sensitivity has the same trend with Precision. Because the dominator of the Sensitivity is the total number of real positive edges which is the same for all $K$, it means that the number of positive edges that are correctly identified increases first when $K$ is smaller than 300 and then slowly decreases with the increase
of $K$. Table 4.7 shows that when $K$ is equal to 300, the Sensitivity is the largest (0.9114) and when $K$ is equal to 1, the Sensitivity is the smallest (0.1399).

In Figure 4.5 (d), the 1-Specificity experiences a small decrease when $K$ is smaller than 200 and then nearly keeps unchanged to be extremely close to 0 until $K$ is equal to 450. It means that the Specificity undergoes a small increase when $K$ is smaller than 200 and then keeps unchanged to be close to 1 with the increase of $K$. Because the dominator of TPR is the total number of real negative edges which is the same for all $K$, it indicates that nearly all negative edges are correctly identified. As for 1-Specificity for all drug networks is close to 0, it is because the majority elements of the adjacent matrix of drug network are 0.

From the four statistic variables, it is clear that when $K$ is smaller than 200, the drug network built from such $K$ is quiet different with the control drug network. When the $K$ is larger than 200, the difference between the drug network built from such $K$ with the control drug network is very small. Therefore, we could conclude that the drug network is affected by the size of signature and when the size of signature is larger than 200, the change of the drug network is small, which means when $K$ is larger than 200, the drug network nearly keeps unchanged.

---

**Figure 4.5: Comparison between networks versus $K$**

---
4.6 Sensitivity analysis on drug communities

This section focuses on the sensitivity analysis of drug communities. First of all, in order to compare drug communities of different parameters, we come up with a method to build a cluster network to represent drug communities, which is introduced in section 4.6.1. Then, based on the measurement of comparing different networks mentioned in section 4.2, we did sensitivity analysis of drug communities on distance threshold (section 4.6.2) and the size of optimal signature (section 4.6.3).

However, the result of the sensitivity analysis is not desirable based on the method, so later we come up with a new method to build the cluster network and re-did the sensitivity analysis on both parameters. The details of the later method and results of the sensitivity analysis on both parameters can be seen in corresponding sections.

4.6.1 Method to compare drug communities

In order to do sensitivity on drug communities of the framework in this section, we come up with a method to build the cluster network based on the drug network and communities. We made an assumption that the connections between drugs and its exemplars exist in drug network. Based on the assumption, combining the drug network and the corresponding cluster result, we just keep the connection between drugs with its exemplar, the connection between drugs of the same exemplar and the connection between exemplars of the network. After the cluster network is built, it can be represented by its adjacency matrix and we can use the measurement mentioned in section 4.2 to compare the cluster networks of different parameters.

Based on the method and control drug network ($K$ is 250 and threshold is 0.05), we got a cluster network with 87 communities, 1,294 nodes and 11,522 connections (see code `cluster_later.R`).

4.6.2 Distance threshold

Based on the method mentioned in section 4.6.1, drug communities of each distance threshold can be represented by a cluster network. The details of the cluster networks of different thresholds can be seen in Table 4.8. Each cluster network was represented by its adjacency matrix. Comparing the adjacency matrix of cluster network of each threshold with the control one, the confusion matrix of each threshold was calculated, see Appendix B. Based on the confusion matrix, the Accuracy, Precision, TPR (sensitivity) and FPR (1-specificity) were calculated for each threshold, see Table 4.9 and Figure 4.6.

From Table 4.8, it is clear that with the increase of threshold, the number of communities decreases, while the number of nodes and edges in the cluster network increase.
Table 4.8: Cluster network information of different thresholds

<table>
<thead>
<tr>
<th>Threshold</th>
<th>community</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>node</td>
<td>165</td>
<td>140</td>
<td>111</td>
<td>74</td>
<td>64</td>
<td>38</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>edge</td>
<td>1188</td>
<td>1257</td>
<td>1279</td>
<td>1303</td>
<td>1301</td>
<td>1304</td>
<td>1309</td>
<td>1309</td>
</tr>
</tbody>
</table>

Table 4.9: Statistic information of comparison between cluster networks versus threshold

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>1-Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.9875702</td>
<td>0.59910011</td>
<td>0.2311231</td>
<td>0.002109964</td>
</tr>
<tr>
<td>0.03</td>
<td>0.9867093</td>
<td>0.51249133</td>
<td>0.2563791</td>
<td>0.003327161</td>
</tr>
<tr>
<td>0.04</td>
<td>0.9867969</td>
<td>0.51486762</td>
<td>0.3291095</td>
<td>0.004230585</td>
</tr>
<tr>
<td>0.06</td>
<td>0.9881928</td>
<td>0.55646063</td>
<td>0.6047561</td>
<td>0.006576174</td>
</tr>
<tr>
<td>0.07</td>
<td>0.9864137</td>
<td>0.4961102</td>
<td>0.6032807</td>
<td>0.008359343</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9742888</td>
<td>0.22246918</td>
<td>0.3648672</td>
<td>0.017397142</td>
</tr>
<tr>
<td>0.2</td>
<td>0.9340113</td>
<td>0.13519696</td>
<td>0.7232251</td>
<td>0.063113038</td>
</tr>
<tr>
<td>0.3</td>
<td>0.9041907</td>
<td>0.07995805</td>
<td>0.5823642</td>
<td>0.091418768</td>
</tr>
</tbody>
</table>

Figure 4.6: Comparison between cluster networks versus threshold
The increase of the number of nodes and edges satisfies the meaning of threshold, because when the threshold increases, it is a more loose condition to the distances and more edges and nodes are kept in the cluster network. However, we do not know the exact reason why the number of communities decreases with the increase of threshold, but we guess that it is because when the threshold increases, more edges are kept in the cluster network and some communities are connected by the new edges to become one larger community.

As for Figure 4.6, even though the data are fluctuated, there is a general trend, which can generally conclude that the cluster network is affected by the threshold because if the cluster network is not affected, at least the Accuracy of different thresholds should be extremely close to 1. With the increase of threshold, the Accuracy and Precision decrease generally and the Sensitivity and Specificity increase generally.

Besides, the Precision and Sensitivity of all thresholds are under 0.6 and 0.7, respectively, which shows that the difference between cluster networks of different parameters with the control one is quiet large.

Based on the sensitivity analysis result, there are two possible conclusions, one is that the drug communities is very sensitive to the distance threshold, the other is that the method to build the cluster network is not valid. Therefore, in order to find the reason and get a more valid result, we checked the cluster networks and its communities. We found that our assumption that the connections between drugs and its exemplar exist in drug network is wrong, so when we built the cluster network, some drugs which did not have connection with its exemplar in drug network and did not have connection with other drugs in the community would be removed from the cluster network. It is the reason why the number of nodes of cluster network is different from that of drug network. There is one example which can verify that the assumption is wrong and explain why the number of cluster network is different from that of drug network.

A community of the control cluster network with 5% quantile distance threshold contains ten drugs, the exemplar of which is drug aminophenazone. If our assumption is correct, the sub-network of the community should show the situation as Figure 4.7. However, in reality (see Figure 4.8), the drug rilmenidine did not show in the cluster network, because it did not have connection with aminophenazone in drug network and it did not have connection with other drugs in the community. Besides, it also shows that the drug hydroflumethiazide did not have connection with the exemplar in cluster network and it is not removed from the cluster network because it had a connection with drug benzonatate in the community.

Therefore, our method to build a cluster network is wrong and we come up with a new method. The new method is not based on the drug network and it is just based on the drug communities. In the new method, all the drugs in one community are connected with each other and the weight of the connections are 1 (see code new_cluster_later.R). Using the method to build the cluster network, if one community contains five drugs, the network of the community should be shown in the way, see Figure 4.9.

Based on the new method, cluster networks of different thresholds and default threshold
Figure 4.7: The sub-network of the community in assumption

Figure 4.8: The network of the community in reality
Figure 4.9: The network of the community which contains 5 drugs using new method were built. Using the method, we got a control cluster network with 87 communities, 1302 nodes and 33448 edges. The details of the cluster networks of different thresholds can be seen in Table 4.10. Each cluster network was represented by its adjacency matrix. Comparing the adjacency matrix of each threshold with the control one, the confusion matrix of each threshold was calculated. Based on the confusion matrix, the Accuracy, Precision, TPR (sensitivity) and FPR (1-specificity) were calculated for each threshold, see Table 4.11 and Figure 4.10.

Table 4.10: Cluster network information of different thresholds in new method

<table>
<thead>
<tr>
<th>Threshold</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>community</td>
<td>165</td>
<td>140</td>
<td>111</td>
<td>74</td>
<td>64</td>
<td>38</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>node</td>
<td>1217</td>
<td>1271</td>
<td>1294</td>
<td>1306</td>
<td>1305</td>
<td>1309</td>
<td>1309</td>
<td></td>
</tr>
<tr>
<td>edge</td>
<td>13472</td>
<td>17506</td>
<td>22124</td>
<td>37158</td>
<td>40152</td>
<td>53977</td>
<td>121690</td>
<td>141961</td>
</tr>
</tbody>
</table>

From Figure 4.6 and Figure 4.10, it is clear that the result obtained from the new method is similar to the result obtained from original method. Besides, the Precision and Sensitivity of all thresholds are under 0.6 and 0.7, respectively, which means that the cluster networks built from these thresholds are quiet different from the control one.

Only then, we could say that the drug communities is affected by distance threshold and the framework is affected by distance threshold.

4.6.3 The size of optimal signature

Based on the method mentioned in section 4.6.1, drug communities of each $K$ can be represented by a cluster network. The details of the cluster networks can be seen in Table 4.12. Each cluster was represented by its adjacency matrix. Comparing the adjacency matrix of cluster networks of each $K$ with the control one, the confusion matrix
Table 4.11: Statistic information of comparison between cluster networks versus threshold in new method

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>1-Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.9632303</td>
<td>0.5731146</td>
<td>0.2308359</td>
<td>0.006990924</td>
</tr>
<tr>
<td>0.03</td>
<td>0.9606698</td>
<td>0.4936593</td>
<td>0.2583712</td>
<td>0.010775092</td>
</tr>
<tr>
<td>0.04</td>
<td>0.9586607</td>
<td>0.456111</td>
<td>0.3016922</td>
<td>0.01462733</td>
</tr>
<tr>
<td>0.06</td>
<td>0.9630808</td>
<td>0.524786</td>
<td>0.5829945</td>
<td>0.021465092</td>
</tr>
<tr>
<td>0.07</td>
<td>0.9572239</td>
<td>0.4605001</td>
<td>0.5527984</td>
<td>0.02633236</td>
</tr>
<tr>
<td>0.1</td>
<td>0.9236981</td>
<td>0.2047539</td>
<td>0.3304233</td>
<td>0.052179695</td>
</tr>
<tr>
<td>0.2</td>
<td>0.868595</td>
<td>0.1752157</td>
<td>0.6374671</td>
<td>0.122007493</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8343402</td>
<td>0.1183071</td>
<td>0.5021227</td>
<td>0.152151979</td>
</tr>
</tbody>
</table>

Figure 4.10: Comparison between cluster networks versus threshold in new method
of each $K$ was calculated, see Appendix B. For each $K$, based on the confusion matrix, the Accuracy, Precision, TPR (sensitivity) and FPR (1-specificity) were calculated for each $K$, see Table 4.13 and Figure 4.11.

Table 4.12: Cluster network information of different $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>Community</th>
<th>Node</th>
<th>Edge</th>
<th>Community</th>
<th>Node</th>
<th>Edge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>38</td>
<td>39</td>
<td>42</td>
<td>57</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>1306</td>
<td>1291</td>
<td>1300</td>
<td>1278</td>
<td>1299</td>
<td>1301</td>
</tr>
<tr>
<td>5</td>
<td>2530</td>
<td>2932</td>
<td>4331</td>
<td>4799</td>
<td>5092</td>
<td>7817</td>
</tr>
<tr>
<td>10</td>
<td>75</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>350</td>
<td>450</td>
</tr>
<tr>
<td>20</td>
<td>1302</td>
<td>1295</td>
<td>1300</td>
<td>1293</td>
<td>1296</td>
<td>1283</td>
</tr>
<tr>
<td>50</td>
<td>7429</td>
<td>8843</td>
<td>8983</td>
<td>8494</td>
<td>13001</td>
<td>11809</td>
</tr>
</tbody>
</table>

Table 4.13: Statistic information of comparison between cluster networks versus $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>1-Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9840904</td>
<td>0.08537549</td>
<td>0.01874675</td>
<td>0.002739875</td>
</tr>
<tr>
<td>2</td>
<td>0.9839105</td>
<td>0.1159618</td>
<td>0.02950877</td>
<td>0.003069039</td>
</tr>
<tr>
<td>5</td>
<td>0.9832388</td>
<td>0.17363196</td>
<td>0.06526645</td>
<td>0.00423769</td>
</tr>
<tr>
<td>10</td>
<td>0.98365</td>
<td>0.24213378</td>
<td>0.10085055</td>
<td>0.004306364</td>
</tr>
<tr>
<td>20</td>
<td>0.9826162</td>
<td>0.21535073</td>
<td>0.11031071</td>
<td>0.005483303</td>
</tr>
<tr>
<td>50</td>
<td>0.9840962</td>
<td>0.36612511</td>
<td>0.24839438</td>
<td>0.005866933</td>
</tr>
<tr>
<td>75</td>
<td>0.9849256</td>
<td>0.40593117</td>
<td>0.25898282</td>
<td>0.005170715</td>
</tr>
<tr>
<td>100</td>
<td>0.9842411</td>
<td>0.388669</td>
<td>0.29829891</td>
<td>0.006400936</td>
</tr>
<tr>
<td>200</td>
<td>0.9861638</td>
<td>0.4820216</td>
<td>0.37580281</td>
<td>0.005509352</td>
</tr>
<tr>
<td>300</td>
<td>0.986513</td>
<td>0.49857824</td>
<td>0.36755772</td>
<td>0.005042839</td>
</tr>
<tr>
<td>350</td>
<td>0.9858063</td>
<td>0.47580955</td>
<td>0.53688596</td>
<td>0.008069252</td>
</tr>
<tr>
<td>450</td>
<td>0.9848532</td>
<td>0.43881785</td>
<td>0.44974831</td>
<td>0.007846652</td>
</tr>
</tbody>
</table>

From Table 4.12, it is clear that with the increase of $K$, the number of communities and the number of edges of the cluster networks show an general increase trend, while the number of nodes nearly keeps unchanged.

As for Figure 4.11, even though the data are fluctuated, there is a general trend, which can generally conclude that the cluster network is affected by the size of signature, because if the cluster network is not affected, at least the Accuracy and Sensitivity should be very close to 1 all the time. Besides, the Precision and Sensitivity of all cluster networks are under 0.6, which shows that the cluster networks built from these $K$ are quiet different from the control one.

As analyzed in section 4.6.2, there are two possible conclusions can be made from the result, one is that the drug communities is very sensitive to the size of signature, the other is that the method we used to build cluster network is wrong. we have analyzed in
section 4.6.2 that the method we used to build the cluster network is wrong, so we used the new method, that all the drugs in one community are connected with each other and the weight of the connections are 1, to re-build the cluster network for each \( K \). The details of the cluster networks of different \( K \) based on new method can be seen in Table 4.14. Each cluster network was represented by its adjacency matrix. Comparing the adjacency matrix of each threshold with the control one, the confusion matrix of each threshold was calculated. For each \( K \), based on the confusion matrix, the Accuracy, Precision, TPR (sensitivity) and FPR (1-specificity) were calculated for each threshold, see Table 4.15 and Figure 4.12.

Table 4.14: Cluster network information of different \( K \) in new method

<table>
<thead>
<tr>
<th>( K )</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>community</td>
<td>63</td>
<td>38</td>
<td>39</td>
<td>42</td>
<td>57</td>
<td>67</td>
</tr>
<tr>
<td>node</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
<td>1309</td>
</tr>
<tr>
<td>edge</td>
<td>14646</td>
<td>24363</td>
<td>29071</td>
<td>27659</td>
<td>24692</td>
<td>27587</td>
</tr>
<tr>
<td>( K )</td>
<td>75</td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>350</td>
<td>450</td>
</tr>
<tr>
<td>community</td>
<td>75</td>
<td>75</td>
<td>88</td>
<td>94</td>
<td>101</td>
<td>104</td>
</tr>
<tr>
<td>node</td>
<td>1309</td>
<td>1307</td>
<td>1304</td>
<td>1301</td>
<td>1300</td>
<td>1296</td>
</tr>
<tr>
<td>edge</td>
<td>24765</td>
<td>30944</td>
<td>25555</td>
<td>25449</td>
<td>34282</td>
<td>34471</td>
</tr>
</tbody>
</table>

From Figure 4.11 and Figure 4.12, it is clear that the result obtained from new method
Table 4.15: Statistic information of comparison between cluster networks versus $K$ in new method

<table>
<thead>
<tr>
<th>$K$</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>1-Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9460451</td>
<td>0.06500068</td>
<td>0.02846209</td>
<td>0.01664645</td>
</tr>
<tr>
<td>2</td>
<td>0.9363393</td>
<td>0.06797192</td>
<td>0.04950969</td>
<td>0.02760266</td>
</tr>
<tr>
<td>5</td>
<td>0.9335125</td>
<td>0.09631592</td>
<td>0.08371203</td>
<td>0.03193507</td>
</tr>
<tr>
<td>10</td>
<td>0.9368708</td>
<td>0.1276665</td>
<td>0.10556685</td>
<td>0.02932882</td>
</tr>
<tr>
<td>20</td>
<td>0.9397841</td>
<td>0.13344403</td>
<td>0.09851112</td>
<td>0.02601023</td>
</tr>
<tr>
<td>50</td>
<td>0.9452333</td>
<td>0.25646138</td>
<td>0.21152236</td>
<td>0.02493442</td>
</tr>
<tr>
<td>75</td>
<td>0.9500716</td>
<td>0.31233596</td>
<td>0.23125448</td>
<td>0.02070169</td>
</tr>
<tr>
<td>100</td>
<td>0.9456176</td>
<td>0.28819804</td>
<td>0.26662282</td>
<td>0.02677484</td>
</tr>
<tr>
<td>200</td>
<td>0.9556668</td>
<td>0.41185678</td>
<td>0.31466754</td>
<td>0.01827049</td>
</tr>
<tr>
<td>300</td>
<td>0.954744</td>
<td>0.39596841</td>
<td>0.30127362</td>
<td>0.01868623</td>
</tr>
<tr>
<td>350</td>
<td>0.9557615</td>
<td>0.43547634</td>
<td>0.44633461</td>
<td>0.02352554</td>
</tr>
<tr>
<td>450</td>
<td>0.9530877</td>
<td>0.4026283</td>
<td>0.4149426</td>
<td>0.02503167</td>
</tr>
</tbody>
</table>

Figure 4.12: Comparison between cluster networks versus $K$ in new method
is similar to the result obtained from original method. Besides, the Precision and Sen-
sitivity are under 0.5, which is obvious that the drug communities of these $K$ are quiet
different from the control one.

Only then, we could conclude that the drug communities are affected by the size of
signature and the framework is affected by the size of signature.

4.7 Result of the sensitivity analysis

From section 4.5, it is clear that the drug network with larger threshold is the sub-
network of the drug network with large threshold. Besides, the drug network is affected
by the size of signature when the size is smaller than 200. When the size of signature
is larger than 200, the drug network nearly keeps unchanged. The analysis of how the
drug network is affected by the threshold and the size of signature is the same as the
analysis result in [7].

As for the drug communities, from section 4.6, it is clear that based on both method,
the originally wrong method or the new method, to build the cluster network, the results
are similar that the drug communities are affected by both parameters and sensitive to
both parameters.

Based on the above two section, it is clear that the framework is affected by the two
parameters, so the framework is not a robust method to do drug reposition.
Chapter 5

Conclusion
5.1 Introduction

The overall aim of the dissertation was to do sensitivity analysis on the drug repositioning method mentioned in [7] to verify whether the method is robust or not. This chapter will summarize what we did and introduce the findings and conclusions of our work in section 5.2. Then, some recommendations for future research in drug discovery will be made in section 5.3.

5.2 Summary of findings and conclusions

In order to achieve the aim, first of all, we re-implemented the framework of the paper using CMap data. Using KRUBOR algorithm mentioned in algorithm 1, we merged all the ranked gene probe lists of the same drug to generate one general single Prototype Ranked List (PRL) of each drug as the representative ranked gene probe list. Once the single PRL for each drug is generated, based on the Gene Set Enrichment Analysis, we calculated the distances between drugs using two methods, Average enrichment score distance (AES) and Maximum enrichment score distance (MES), and we got 856,086 distances for each method. We chose the 5% quantile of distances as threshold, 0.8337 of AES distance and 0.4033 of MES distance. Then, each drug is represented by a vertex in the network. If the distance between two drugs is below the threshold, the two drugs would be connected by a weighted edge. If a drug is not connected by at least one drug, the drug would be removed from the network. Based on this, we got a drug network with 1305 nodes and 42,085 edges based on AES distance and a drug network with 1,302 nodes and 42,085 edges based on MES distance. Finally, we implemented the affinity propagation cluster algorithm on the drug network based on MES distance and identified 87 drug communities. Until then, we re-implemented the framework.

Then, we chose two parameters, the size of optimal signature and distance threshold, to test whether the framework is affected by the two parameters or not by re-implementing the framework using the new parameter values. We found that both drug network and drug communities are affected by the two parameters. The drug network with smaller threshold is a sub-network of the drug network with larger threshold. Besides, even though the drug network is affected by the size of optimal signature, when the size is larger than 200, the drug network nearly keeps unchanged. As for the drug communities, we originally come up with a method to build a cluster network to represent the drug communities and found that the method is wrong. Then, we come up with a new method to build the cluster network based on drug communities and found that drug communities are very sensitive to both two parameters.

Therefore, we can make a conclusion that the framework is affected by the two parameters, so the framework is not a robust method to do drug repositioning.
5.3 Recommendations

Based on the result of the project that the framework is not robust, we have three recommendations for future study on the framework of drug reposition. The first one is that one more effective method to compare different drug communities can be studied to verify the result which is obtained in the project. Then, we recommend that the future study of the framework can study the best parameter values for the framework to do drug repositioning. Finally, after the best parameters are chosen and the drug communities has been identified, the drug MoA can be elucidated. Based on the study of drug MoA and the gene expression of patients, one more thing can be done is to make medical suggestion for doctors. The medical suggestion can be single drug or combination of drugs from the drug network.
Appendix A

The pre-processing work based on raw CMap data

As mentioned in section 3.3.1, we dealt with the raw CMap data at the beginning of the project. As what we got are raw and machinery data, the function `rma` in the `affy` package of R language was used to do normalization of the raw data and transformed it to gene expression data which can be used. Then, based on the metadata, we matched the transcriptional gene expression of each instance and its corresponding control transcriptional gene expression and got differential gene expression for each instance by calculating the difference between the original data and its control data. When we calculated the differential gene expression, we found that for one instance, it may have more than one control data. In the situation, if the instance has more than one control, we calculated the average gene expression of these controls first and then calculate the difference between the original data and the average control data to get the differential gene expression of the instance. Besides, when we calculated the differential gene expression, we also found that many instances may have same control data, so in order to accelerate the program, we put the instances with the same control together so that each control data is read just once. Based on the differential gene expression, we ranked the genes according to their differential gene expression in an decreasing order. Until then, we got the ranked gene list of each instance.

However, when we processes the merge stage ,we found that the dimension of instances of the same drug maybe different and found that there are three dimensions of 6,100 instances, 22,277, 22,283 and 22,944. It is impossible to merge two lists with different dimensions, so we got the intersection of the three gene lists and obtained one gene list whose dimension is 22,268. According to the intersection gene lists, we changed the dimension of ranked gene list of each instance to 22,268 by removing some genes. Based on this, we merged all the ranked lists of the same drug and got a single ranked list for each drug.

It is clear that the dimension of our ranked list is 22,268, which is different as the dimension of ranked list in "ConnectivityMap" package, 22,283. We calculated the
distances based on the 22,268 genes and the empirical PDF can be seen in Figure A.1 and Figure A.2. In both figure, the red line is the 5% quantile, 0.948979 for AES distance and 0.4642685 for MES distance. Compared with 0.8327 for AES distance and 0.4016 for MES distance in [7], it is clear that what distances we got is quiet different with [7]. Hence, we changed to use the data package in R.

Figure A.1: The empirical probability density function of AES

Figure A.2: The empirical probability density function of MES
Appendix B

Confusion matrix

This chapter supplements the confusion matrix of different drug networks, cluster networks of different size of optimal signature, $K$, and of different size of distance threshold.

Table B.1: Confusion matrix of drug networks of different thresholds

<table>
<thead>
<tr>
<th>threshold</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>17122</td>
<td>25683</td>
<td>34245</td>
<td>42805</td>
<td>42805</td>
<td>42805</td>
<td>42805</td>
<td>42805</td>
</tr>
<tr>
<td>FN</td>
<td>25683</td>
<td>17122</td>
<td>8560</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FP</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8561</td>
<td>17121</td>
<td>42806</td>
<td>128416</td>
<td>214023</td>
</tr>
<tr>
<td>TN</td>
<td>813281</td>
<td>813281</td>
<td>813281</td>
<td>804720</td>
<td>796160</td>
<td>770475</td>
<td>684865</td>
<td>599258</td>
</tr>
</tbody>
</table>

Table B.2: Confusion matrix of drug networks of different $K$

<table>
<thead>
<tr>
<th>$K$</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>5989</td>
<td>7229</td>
<td>11186</td>
<td>14910</td>
<td>19376</td>
<td>26150</td>
</tr>
<tr>
<td>FN</td>
<td>36816</td>
<td>35576</td>
<td>31619</td>
<td>27895</td>
<td>23429</td>
<td>16655</td>
</tr>
<tr>
<td>FP</td>
<td>36828</td>
<td>35588</td>
<td>31619</td>
<td>27900</td>
<td>23430</td>
<td>16657</td>
</tr>
<tr>
<td>TN</td>
<td>776453</td>
<td>777693</td>
<td>781662</td>
<td>785381</td>
<td>789851</td>
<td>796624</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$K$</th>
<th>75</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>350</th>
<th>450</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>29271</td>
<td>31636</td>
<td>38176</td>
<td>39016</td>
<td>37583</td>
<td>35672</td>
</tr>
<tr>
<td>FN</td>
<td>13534</td>
<td>11169</td>
<td>4629</td>
<td>3789</td>
<td>5222</td>
<td>7133</td>
</tr>
<tr>
<td>FP</td>
<td>13538</td>
<td>11169</td>
<td>4629</td>
<td>3791</td>
<td>5222</td>
<td>7133</td>
</tr>
<tr>
<td>TN</td>
<td>799743</td>
<td>802112</td>
<td>808652</td>
<td>809490</td>
<td>808059</td>
<td>806148</td>
</tr>
</tbody>
</table>
### Table B.3: Confusion matrix of cluster networks of different thresholds

<table>
<thead>
<tr>
<th>threshold</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>2663</td>
<td>2954</td>
<td>3792</td>
<td>6968</td>
<td>6951</td>
<td>4204</td>
<td>8333</td>
<td>6710</td>
</tr>
<tr>
<td>FN</td>
<td>8859</td>
<td>8568</td>
<td>8560</td>
<td>4554</td>
<td>4571</td>
<td>7318</td>
<td>3189</td>
<td>4812</td>
</tr>
<tr>
<td>FP</td>
<td>1782</td>
<td>2810</td>
<td>3573</td>
<td>5554</td>
<td>7060</td>
<td>14693</td>
<td>53303</td>
<td>77209</td>
</tr>
<tr>
<td>TN</td>
<td>842782</td>
<td>841754</td>
<td>840991</td>
<td>839010</td>
<td>837504</td>
<td>829871</td>
<td>791261</td>
<td>767355</td>
</tr>
</tbody>
</table>

### Table B.4: Confusion matrix of cluster networks of different K

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>216</td>
<td>340</td>
<td>752</td>
<td>1162</td>
<td>1271</td>
<td>2862</td>
</tr>
<tr>
<td>FN</td>
<td>11306</td>
<td>11182</td>
<td>10770</td>
<td>10360</td>
<td>10251</td>
<td>8660</td>
</tr>
<tr>
<td>FP</td>
<td>2314</td>
<td>2592</td>
<td>3579</td>
<td>3637</td>
<td>4631</td>
<td>4955</td>
</tr>
<tr>
<td>TN</td>
<td>842250</td>
<td>841972</td>
<td>840985</td>
<td>840927</td>
<td>839933</td>
<td>839609</td>
</tr>
</tbody>
</table>

### Table B.5: Confusion matrix of cluster networks of different thresholds in new method

<table>
<thead>
<tr>
<th>threshold</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.06</th>
<th>0.07</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>7721</td>
<td>8642</td>
<td>10091</td>
<td>19500</td>
<td>18490</td>
<td>11052</td>
<td>21322</td>
<td>16795</td>
</tr>
<tr>
<td>FN</td>
<td>25727</td>
<td>24806</td>
<td>23357</td>
<td>13948</td>
<td>14958</td>
<td>22396</td>
<td>12126</td>
<td>16653</td>
</tr>
<tr>
<td>FP</td>
<td>5751</td>
<td>8864</td>
<td>12033</td>
<td>17658</td>
<td>21662</td>
<td>42925</td>
<td>100368</td>
<td>125166</td>
</tr>
<tr>
<td>TN</td>
<td>816887</td>
<td>813774</td>
<td>810605</td>
<td>804980</td>
<td>800976</td>
<td>779713</td>
<td>722270</td>
<td>697472</td>
</tr>
</tbody>
</table>

### Table B.6: Confusion matrix of cluster networks of different K in new method

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>952</td>
<td>1656</td>
<td>2800</td>
<td>3531</td>
<td>3295</td>
<td>7075</td>
</tr>
<tr>
<td>FN</td>
<td>32496</td>
<td>31792</td>
<td>30648</td>
<td>29917</td>
<td>30153</td>
<td>26373</td>
</tr>
<tr>
<td>FP</td>
<td>13694</td>
<td>22707</td>
<td>26271</td>
<td>24127</td>
<td>21397</td>
<td>20512</td>
</tr>
<tr>
<td>TN</td>
<td>808944</td>
<td>799931</td>
<td>796367</td>
<td>798511</td>
<td>801241</td>
<td>802126</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>7735</td>
<td>8918</td>
<td>10525</td>
<td>10077</td>
<td>14929</td>
<td>13879</td>
</tr>
<tr>
<td>FN</td>
<td>25713</td>
<td>24530</td>
<td>22923</td>
<td>23371</td>
<td>18519</td>
<td>19569</td>
</tr>
<tr>
<td>FP</td>
<td>17030</td>
<td>22026</td>
<td>15030</td>
<td>15372</td>
<td>19353</td>
<td>20920</td>
</tr>
<tr>
<td>TN</td>
<td>805608</td>
<td>800612</td>
<td>807608</td>
<td>807266</td>
<td>803285</td>
<td>802046</td>
</tr>
</tbody>
</table>
Bibliography


