

*“Keep true to the dreams of thy youth. ”*

Friedrich von Schiller

*German dramatist & poet (1759 - 1805)*

# University of Alberta

## Pattern Recognition of Time-dependent Cellular Response of Chemicals Based on Profile Shape Similarity

by

Zhankun Xi

A thesis submitted to the Faculty of Graduate Studies and Research  
in partial fulfillment of the requirements for the degree of

Master of Science

in

Process Control

Department of Chemical and Materials Engineering

© Zhankun Xi  
Spring 2014  
Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the authors prior written permission.

To my beloved parents and friends

# Abstract

As a potential approach to interpret Mode of Action (MoA), the shape of cellular response profiles associated with chemicals has been a key consideration. In this thesis, statistical pattern recognition methods using multi-concentration time-dependent cellular response profiles (TCRPs) are explored. Cell Index (CI) values are used to reflect changes in cell population, morphology and the degree of cell attachment and are recorded dynamically as multiple time series data *via* the xCELLigence real-time cell analysis high-throughput (RTCA HT) system. Data processing techniques such as denoising and TCRP selection are applied to generate appropriate data for further analysis. These techniques also screen out the TCRPs which are not responsive enough and retain only those TCRPs which are the representative of action of the chemical compound based on the given cell population. Therefore, all the designed approaches are aimed at pattern recognition of TCRPs and classifying chemicals represented by different numbers of TCRPs. The results of these data-driven classification approaches show reasonable discrimination of chemicals based on profile shape similarity, which provides a potential guideline to determine Mode of Action of chemicals.

# Acknowledgements

I am thankful and grateful with so many memorable moments and days during my stay in Computer Process Control (CPC) group at the University of Alberta. I will always cherish this wonderful time I spend with so many nice people who supported me to pursue my M.Sc. degree, no matter where I am.

First, I would express my sincere appreciation to my supervisor, Dr. Biao Huang. His professional insights and meticulous guidance support me to get through the tough time. His diligence and rigorousness in work told me how to become a qualified researcher. It is my honor to become his student and I appreciated his trusting for offering me a challenging research project. I would also like to express my gratitude to all the members of Bioinformatics in CPC group: Swanand, Tianhong, Fadi, Yaojie, Chandy and Aaron. Their scrupulous attitude to work impressed me and drove me to become as well as they were. I am lucky to have them as my team members. Special thanks go to Swanand, Tianhong, Chandy and Aaron. Their courage and help made a better me. Financial support from the NSERC Industrial Research Chair in Control of Oil Sands Processes and Alberta Health and Wellness is gratefully acknowledged. During my study, I received so much help from my friends. I could not live a colorful life without them. Special thanks to Boon and every friend who offered me their help and care. I will treasure our friendship for the rest of my life.

Last but not least, I would like to express my deepest love and gratitude to my parents from the bottom of my heart. Thank you for your unconditional love, selfless support and clement understanding. This thesis is dedicated to you.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation and objective . . . . .	1
1.2	Main contribution . . . . .	3
1.3	Outline . . . . .	4
<b>2</b>	<b>Experiments, data collection and processing</b>	<b>6</b>
2.1	Experimental design . . . . .	6
2.1.1	Cell line and Chemicals . . . . .	6
2.1.2	xCELLigence RTCA HT system . . . . .	8
2.1.3	Experiment test procedures . . . . .	9
2.2	Data preprocessing . . . . .	11
2.2.1	Calculation of normalized cell index ( <i>NCI</i> ) . . . . .	11
2.2.2	Denoising . . . . .	11
2.2.3	Eligible TCRP selection . . . . .	14
2.3	Nonuniform dimensionality problem . . . . .	16
2.4	Conclusion . . . . .	16
<b>3</b>	<b>Majority-voting and <math>K</math>-means integrated classification</b>	<b>17</b>
3.1	Introduction . . . . .	17
3.2	Majority-voting based feature extraction . . . . .	18
3.2.1	Slope quantization . . . . .	18
3.2.2	Majority voting . . . . .	19

3.2.3	Slope representatives . . . . .	20
3.2.4	Example . . . . .	20
3.3	<i>K</i> -means clustering . . . . .	28
3.4	Results and discussion . . . . .	31
3.4.1	Results . . . . .	31
3.4.2	Discussion . . . . .	37
3.5	Conclusion . . . . .	39
<b>4</b>	<b>A hierarchical correlation based classification</b>	<b>41</b>
4.1	Introduction . . . . .	42
4.2	Step 1: Categorization . . . . .	43
4.3	Step 2: Curve by curve correlation-based clustering . . . . .	45
4.3.1	Correlation analysis between TCRPs . . . . .	46
4.3.2	Correlation analysis between first-order differences of TCRPs	48
4.3.3	Clustering algorithm based on curve by curve correlation analysis . . . . .	49
4.4	Results and discussions . . . . .	54
4.4.1	Results . . . . .	54
4.4.2	Discussion . . . . .	60
4.5	Conclusion . . . . .	61
<b>5</b>	<b>PCA &amp; FDA based hierarchical classification</b>	<b>62</b>
5.1	Introduction . . . . .	62
5.2	Principal component analysis (PCA) . . . . .	63
5.2.1	Overview . . . . .	63
5.2.2	Problem formulation . . . . .	64
5.3	Functional data analysis (FDA) . . . . .	66
5.3.1	Basis function selection . . . . .	67
5.3.2	Computing coefficients . . . . .	68
5.4	Model-based hierarchical clustering method . . . . .	72

5.4.1	Model-based clustering . . . . .	72
5.4.2	Hierarchical clustering . . . . .	73
5.4.3	Model-based hierarchical clustering . . . . .	74
5.5	Application . . . . .	75
5.5.1	Feature extraction . . . . .	76
5.5.2	Model-based hierarchical clustering analysis . . . . .	78
5.5.3	One-level clustering . . . . .	79
5.5.4	Two-level clustering . . . . .	86
5.5.5	Automatic determination of cluster number . . . . .	97
5.5.6	Chemical classification GUI module development . . . . .	104
5.6	Conclusion . . . . .	104
<b>6</b>	<b>Conclusions and future work</b>	<b>106</b>
6.1	Conclusion . . . . .	106
6.2	Future work . . . . .	108
	<b>Appendix</b>	<b>116</b>



# List of Tables

2.1	Summary of the Chemical Compounds . . . . .	7
3.1	Distribution of logical values (Cytochalasin D) . . . . .	26
3.2	Logical representation with conflict (Cytochalasin D) . . . . .	27
3.3	Improved logical representation (Cytochalasin D) . . . . .	27
3.4	Results of Majority-voting and $K$ -means integrated classification (cell line: HepG2. 47 of 65 chemicals are eligible.) . . . . .	32
3.5	Excerpted Logical representation (Vincristine Sulfate) . . . . .	38
4.1	Results of Hierarchical correlation based classification (cell line: HepG2. 47 of 65 chemicals are eligible.) . . . . .	55
5.1	Results of PCA & FDA based hierarchical classification with one-level dendrogram cutting (cell line: HepG2. 47 of 65 chem- icals are eligible.) . . . . .	81
5.2	Results of PCA & FDA based hierarchical classification with two-level dendrogram cutting (cell line: HepG2. 47 of 65 chem- icals are eligible.) . . . . .	92
5.3	Results of PCA & FDA based hierarchical classification with the number of clusters determined using BIC (cell line: HepG2. 47 of 65 chemicals are eligible.) . . . . .	99
6.1	Advantages and disadvantages of Majority-voting and $K$ -means clustering integrated classification . . . . .	107

6.2	Advantages and disadvantages of Hierarchical correlation based classification . . . . .	107
6.3	Advantages and disadvantages of PCA & FDA based hierarchical classification . . . . .	108

# List of Figures

2.1	Layout of the experiment plate (384x). In order to achieve robust and reliable results, the experiment for each chemical was repeated in quadruplicates. The concentrations were in a descending order from the 1 <sup>st</sup> column to the 11 <sup>th</sup> column in the left hand side and from the 24 <sup>th</sup> column to the 14 <sup>th</sup> column in the right hand side. The serial dilution ratio was 1:3. Negative control and DMSO (high: 0.2%, low: 0.067% or 0.04%) control panels were placed in the 12 <sup>th</sup> and 13 <sup>th</sup> columns in order to minimize the effect of chemicals under different concentrations. .	11
2.2	TCRPs of 5-FU and Paclitaxel . . . . .	13
2.3	TCRPs of 5-FU and Paclitaxel (relative) . . . . .	13
2.4	Relative TCRPs and selected relative TCRPs of 5-FU and Paclitaxel. Note that each chemical is now represented using TCRPs with different numbers of concentrations (profiles) retained after eligible profile selection. . . . .	15
3.1	Eligible TCRPs of 5-FU, Cytochalasin D and Paclitaxel . . . . .	22
3.2	Slopes of 5-FU, Cytochalasin D and Paclitaxel at the highest concentration level . . . . .	23
3.3	Logical representation . . . . .	24
3.4	Majority-voting logical representation . . . . .	25
3.5	Improvement of majority-voting logical representation . . . . .	27

3.6	Slope features . . . . .	28
3.7	Extracted feature vectors of eligible chemicals using majority voting methods . . . . .	31
3.8	Extracted feature vectors of chemicals in cluster 3 using majority voting methods . . . . .	38
3.9	TCRPs of Vincristine Sulfate and Paclitaxel . . . . .	39
4.1	Hierarchical correlation based classification schematic structure .	42
4.2	Negative pattern . . . . .	44
4.3	Positive pattern . . . . .	44
4.4	High correlation coefficient indicating similar trend . . . . .	47
4.5	High correlation in TCRP but not in the first-order differences of TCRPs . . . . .	49
4.6	TCRPs of 5-FU and Etoposide . . . . .	50
4.7	TCRPs of Monesin and CCCP . . . . .	52
5.1	Orthogonal projection of data in the original space onto the prin- cipal space . . . . .	64
5.2	Curve smoothness are determined by the penalized parameter $\lambda$	71
5.3	TCRPs of Gemcitabine HCl are normalized with mean 0 and standard deviation 1. . . . .	76
5.4	Basis functions: 15 cubic spline functions . . . . .	77
5.5	First three PC scores and the coefficients after FDA . . . . .	78
5.6	Dendrogram of chemicals . . . . .	79
5.7	Colored dendrogram of chemicals with the first two coefficient vectors aligned as features. 6 reasonable clusters are generated. .	80
5.8	Dendrogram of the chemicals in the 1 <sup>st</sup> level classification . . . .	87
5.9	Dendrogram with three reasonable subtrees marked with colors .	88
5.10	Dendrogram of chemicals in subtree 1 . . . . .	88

5.11	Dendrogram of chemicals in subtree 1 with three reasonable clusters marked in colors . . . . .	89
5.12	Dendrogram of chemicals in subtree 2 . . . . .	89
5.13	Dendrogram of chemicals in subtree 2 with four reasonable clusters marked in colors . . . . .	90
5.14	Dendrogram of chemicals in subtree 3 . . . . .	90
5.15	Dendrogram of chemicals in subtree 3 with one reasonable cluster marked in blue . . . . .	91
5.16	Model selection using BIC score . . . . .	98
6.1	Main graphic user interface (GUI) of Mode of Action Classification	116
6.2	First level dendrogram . . . . .	119
6.3	Second level dendrogram . . . . .	119
6.4	A prompted dialogue for the users to input appropriate cluster numbers . . . . .	120
6.5	Table saved via the GUI . . . . .	121
6.6	TCRP figures saved via the GUI . . . . .	121

# List of Abbreviations

MoA	Mode of Action
CI	Cell Index
TCRP(s)	Time-dependent Cellular Response Profile(s)
PCA	Principal Component Analysis
FDA	Functional Data Analysis
HTS	High-throughput Screening

# Chapter 1

## Introduction

### 1.1 Motivation and objective

With the expansion of industrialization, problems related to the environment and human health are drawing more attention. Chemicals which can lead to adverse effects on the human body are gradually detected and studied. Many chemicals may introduce diverse changes in cellular morphology and growth rate under sufficient doses or concentration levels. Some cause toxicity effects such as apoptosis and necrosis, while others can lead to uncontrolled cellular proliferation such as the case of cancer [1]. It is a challenge to effectively assess the toxicity of chemicals in pharmaceutical and biotechnological industries.

Experts in biology and medical science are studying several methods to assess and estimate the toxicity and infer effective solutions according to the biological phenomena of chemicals. Several analytical techniques have been used to characterize the toxicity of chemicals in toxicity testing of environment [2]. *In vivo* assays investigate the toxic effects of chemicals on living organisms and explain detailed mechanistic understanding of the molecular targets [3]. However, performing *in vivo* assays is time consuming and expensive, usually requiring a large number of samples to be observed. Furthermore, biological contamination and the ethics of causing distress and pain on living bodies are

potential concerns [4].

To effectively assess the toxicity of chemicals, cell-based *in vitro* assays play an integral role in today’s drug discovery [5]. They have become a key component of some research fields such as disease modelling, chemical screening, and safety assessment [6] and are now adopted for automated high-throughput screening (HTS) of large chemical libraries, providing *in situ* analysis for a variety of biological targets [7, 8]. Quantitative high-throughput assays have become a crucial and effective tool in drug discovery and development because they can expand the coverage of existing and new chemicals that need to be evaluated for human health risk assessment [9]. The main advantages of bioactivity profiling using high-throughput assays are the reduced cost and less time required for toxicological screening of environmental chemicals and also reduced the need for animal testing [10].

Mode of Action (MoA) is described as a set of key events and processes starting with the interaction of an agent with a cell, through physiological and tissue or organ changes [11]. The interaction decides if there is an adverse effect occurring between the organism and the chemicals. Therefore, MoA is a means of analysis based on physical, chemical, and biological information that is helpful in explaining key events in a chemical’s influence on organism [12]. As a necessary and crucial element in MoA analysis, a “*key event*” is an empirically observable precursor step that is itself a necessary element in MoA or a biological marker for the element [12]. In our study, we focus our attention on the cellular behaviors in the key events. Studying MoA is crucial in ecotoxicology not only to improve our understanding on the effects of pollutants on ecosystems, but also to build relevant and effective tools which can be applied in environmental risk assessment of chemicals and of polluted sites [13].

The main objective of this thesis is to design and develop three different pattern recognition strategies to classify chemicals based on time-dependent



cellular responses profiles (TCRPs), which record the cellular response continuously and dynamically, and hence assist biologists and experts in analyzing and discovering new drugs. These proposed methods are mainly based on profile shape similarity evaluated from a statistical machine learning and data mining perspective. Chemicals with similar response profiles are typically classified into one group. However, after data processing including denosing and eligible TCRPs selection, each chemical compound is represented by multiple time series; wherein, each time series represents a cellular response corresponding to a single dose of chemical compound. For some dose of the chemical compound, the cellular response is not significant and is screened out so that the subsequent analysis is not affected by non responsive curves. However, this data preprocessing adds to the challenge in pattern recognition that each chemical may be represented by different number of time series data. Thus, classifying chemicals with multiple doses is equivalent to classifying data points represented by different batches of time series. All the algorithms designed in this thesis are aimed at solving this problem.

The significance of this research lies in investigating various ways of handling huge data sets conveniently, increasing efficiency and saving human labor by means of bioinformatics. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge. Bioinformatics techniques have also been applied in many important areas such as finding homologues, rational drug design, large-scale censuses as well as in medical sciences [14]. By establishing a human machine interface (HMI), people can easily implement pattern recognition of different data sets (in our study, they are TCRPs of chemicals) effectively.

## 1.2 Main contribution

The main contributions of the thesis lie in the following aspects:

1. Propose three exploratory approaches to realize pattern recognition of multi-concentration TCRPs.
2. Formulate nonuniform data dimension reduction under the framework of majority-voting as well as Principal Component Analysis (PCA) & Functional Data Analysis (FDA) and extract valid feature vectors for clustering.
3. Test the strategies of classification successfully and implement them on chemicals based on dynamic information in TCRPs.
4. Design a user friendly GUI module and supply basic guidelines on the use of GUI module.

## 1.3 Outline

The thesis is organized as follows: Chapter 2 presents how chemicals and the target cell line are selected and how the experiments are carried out. Data processing methods including denoising and eligible TCRP & chemical selection are also introduced. Eligible TCRPs and chemicals are retained for further analysis. To realize the goal of discriminating chemicals, all the proposed approaches are aimed at pattern recognition of TCRPs with flexible numbers of profiles. In Chapter 3, a majority-voting and  $K$ -means clustering integrated classification method is elaborated. The proposed approach can effectively classify chemicals represented by different numbers of TCRPs into several groups according to the similarity of profiles. In Chapter 4, a hierarchical correlation-based classification method including categorization and curve by curve correlation-based clustering is proposed. In this approach, similarity between the TCRPs is measured using correlation coefficients and with this similarity metric, the chemicals are classified into different groups. In Chapter 5, a method incorporating Principal Component Analysis (PCA) and Functional Data Analysis

(FDA) as statistical feature extracting techniques, and a model-based hierarchical clustering approach to cluster the extracted features, is presented. All of these methods can effectively discriminate chemicals according to the similarity among shapes of TCRPs. In Chapter 6, conclusions based on the comparison about the advantages and disadvantages of three methods are reported. The recommendations for future work are also included.

# Chapter 2

## Experiments, data collection and processing

### 2.1 Experimental design

#### 2.1.1 Cell line and Chemicals

All the experiments were conducted in human hepatocellular carcinoma cells (HepG2) (Order # HB-8065, Cat.# 30-2003) which were obtained from ATCC (Manassas, VA, USA), and described in this section. HepG2 cells were routinely maintained in EMEM (Eagle's minimum essential medium) with 10% (v/v) fetal bovine serum (FBS). Cells were sub-cultured twice per week up to passage 20, and grown in an incubator with a set temperature at 37 °C, with 95% relative humidity and a CO<sub>2</sub> level of 5%. They were examined on workdays under a phase contrast microscope. Any changes to cell morphology or their adhesive properties were recorded.

Stock solutions of 65 chemicals obtained from Sigma-Aldrich (St. Louis, MO, USA) were prepared using water (H<sub>2</sub>O), dimethyl sulfoxide (DMSO) and ethanol (EtOH) separately and stored in amber vials at -80 °C. Solutions were diluted into 11 working concentrations in a serial ratio 1:3. The chemicals and

their solutions and concentration ranges are provided in Table 2.1.

Table 2.1: Summary of the Chemical Compounds

SN	Chemicals	Solvent	Concentration (1:3)
1	5-fluorouracil (5-FU)	DMSO	200 $\mu$ M – 3.39nM
2	Etoposide phosphate	DMSO	200 $\mu$ M – 3.39nM
3	Cordycepin	DMSO	200 $\mu$ M – 3.39nM
4	Cytochalasin D	DMSO	20 $\mu$ M – 0.339nM
5	Cytochalasin B	DMSO	20 $\mu$ M – 0.339nM
6	Latrunculin B	DMSO	20 $\mu$ M – 0.339nM
7	Emetine	H <sub>2</sub> O	50 $\mu$ M – 0.847nM
8	Paclitaxel	DMSO	20 $\mu$ M – 0.339nM
9	Actinomycin D	DMSO	2 $\mu$ M – 0.0339nM
10	Puromycin	H <sub>2</sub> O	1000 $\mu$ M – 17nM
11	Anisomycin	H <sub>2</sub> O	10 $\mu$ M – 0.17nM
12	Clofarabine (CLOF)	H <sub>2</sub> O	25 $\mu$ M – 0.42nM
13	Hydroxyurea (HU)	H <sub>2</sub> O	10mM – 169nM
14	Valproic Acid	H <sub>2</sub> O	50mM – 847nM
15	Vincristine Sulfate	H <sub>2</sub> O	250 $\mu$ M – 4.23nM
16	Doxorubicin (DOX)	H <sub>2</sub> O	100 $\mu$ M – 1.69nM
17	Brefeldin A (BEF)	DMSO	40 $\mu$ M – 0.68nM
18	Leptomycin B (LMB)	EtOH	20nM – 0.000339nM
19	Exo 1	DMSO	300 $\mu$ M – 5.08nM
20	Monensin	DMSO	4 $\mu$ M – 0.068nM
21	Concanamycin A (CMA)	DMSO	0.2 $\mu$ M – 0.003nM
22	Oligomycin	DMSO	20 $\mu$ M – 0.339nM
23	Antimycin A	EtOH	200 $\mu$ M – 3.387nM
24	Rotenone	DMSO	200 $\mu$ M – 3.387nM
25	Thapsigargin	DMSO	2 $\mu$ M – 0.0339nM
26	BHQ	DMSO	400 $\mu$ M – 7nM
27	Ochratoxin A	DMSO	10 $\mu$ M – 0.17nM
28	Cyclosporin A	DMSO	100 $\mu$ M – 1.69nM
29	FK-506 (tacrolimus)	DMSO	50 $\mu$ M – 1nM
30	BAPTA-am	DMSO	60 $\mu$ M – 1nM
31	Latrunculin A	EtOH	2 $\mu$ M – 0.03nM
32	CCCP	DMSO	100 $\mu$ M – 1.69nM
33	SAHA	DMSO	151 $\mu$ M – 2.56nM
34	(S)-HDAC-42	DMSO	128 $\mu$ M – 2.17nM
35	Mitoxantrone Dihydrochloride	DMSO	150 $\mu$ M – 2.54nM
36	Mitomycin C	DMSO	200 $\mu$ M – 3.39nM
37	NU7026	DMSO	20 $\mu$ M – 0.34nM
38	CRT0044876	DMSO	194 $\mu$ M – 3.29nM
39	Topotecan	DMSO	95 $\mu$ M – 1.61nM

*Continued on next page ...*

... continued from previous page

SN	Chemicals	Solvent	Concentration (1:3)
40	Gemcitabine HCl	H <sub>2</sub> O	1650 $\mu$ M – 27.94nM
41	Cisplatin	H <sub>2</sub> O	150 $\mu$ M – 2.54nM
42	Merbarone	DMSO	200 $\mu$ M – 3.39nM
43	Irinotecan (CPT-11)	DMSO	160 $\mu$ M – 2.71nM
44	Cytosine	H <sub>2</sub> O	8950 $\mu$ M – 151.57nM
45	ABT-888 (veliparib)	DMSO	308 $\mu$ M – 5.22nM
46	Benzo[a]pyrene	DMSO	100 $\mu$ M – 1.69nM
47	Gemcitabine	H <sub>2</sub> O	2 $\mu$ M – 0.03nM
48	Monastrol	DMSO	100 $\mu$ M – 1.69nM
49	S-trityl-Cysteine	DMSO	100 $\mu$ M – 1.69nM
50	Dimethylenastron	DMSO	40 $\mu$ M – 0.68nM
51	W7 HCl	DMSO	200 $\mu$ M – 3.39nM
52	Y-27632	DMSO	188 $\mu$ M – 3.18nM
53	Ro32-3555	DMSO	200 $\mu$ M – 3.39nM
54	Batimastat	DMSO	200 $\mu$ M – 3.39nM
55	FAKInhibitor14	H <sub>2</sub> O	2500 $\mu$ M – 42.34nM
56	MLCKInhibPep18	H <sub>2</sub> O	94.5 $\mu$ M – 1.6nM
57	PF573228	DMSO	40 $\mu$ M – 0.68nM
58	Blebbistatin	DMSO	100 $\mu$ M – 1.69nM
59	Docetaxel	DMSO	1 $\mu$ M – 0.02nM
60	SN-38	DMSO	200 $\mu$ M – 3.39nM
61	Vinblastine Sulfate	H <sub>2</sub> O	40 $\mu$ M – 0.68nM
62	Bafilomycin A1	DMSO	0.3212 $\mu$ M – 0.01nM
63	ML7 hydrochloride	DMSO	100 $\mu$ M – 1.69nM
64	HA1100 hydrochloride	H <sub>2</sub> O	1000 $\mu$ M – 16.94nM
65	PF431396	DMSO	5 $\mu$ M – 0.08nM

### 2.1.2 xCELLigence RTCA HT system

The experiments of chemical assays were performed using the xCELLigence real-time cell analysis high-throughput (RTCA HT) system with HepG2 cell line exposed to 65 chemicals. The RTCA HT system is developed by ACEA Biosciences Inc. (San Diego, USA) in the 96x or 384x well plate format. The system runs automatically and is equipped with the Biomek<sup>®</sup> FXP Dual Arm System with Multichannel Pipettor and Span-8 Pipettors. It can pinpoint the optimal time points for conducting endpoints assays and be used to monitor the

dynamics of cell viability continuously by measuring the electronic impedance of sensor electrodes integrated on the bottom of microtiter plates [1, 15]. The principles behind RTCA HT system are described by Abassi et al. [16], Slanina et al. [17], and Xing et al. [18]. Briefly, the system uses microelectronic plates (E-plates) integrated with gold micro-electrode arrays on glass substrate in the bottom of the wells to measure cellular status in real time. Cells are cultivated onto the surfaces of the microelectronic sensors. Under the control of RTCA software, the sensor analyzer automatically selects wells to be measured and continuously conducts measurements on wells. The electronic impedance is then transferred to a computer and recorded [19]. The extent of impedance change is related to the number of cells inside the wells and the inherent morphological and adhesive characteristics of the cells [16]. Cell index, often abbreviated as  $CI$ , is derived to provide quantitative information about the biological status of the cells such as cell number [18].  $CI$  is calculated as

$$CI = \max_{k=1,\dots,K} \left[ \frac{R_{cell}(f_k)}{R_b(f_k)} - 1 \right] \quad (2.1)$$

where  $k$  is the number of the frequency points at which the impedance is measured and  $R_b(f_k)$  and  $R_{cell}(f_k)$  are the frequency-dependent electrode impedance (resistance) without and with cells present in the wells, respectively.

### 2.1.3 Experiment test procedures

The layout of the plate is schematically represented in Figure 2.1. The experimental test procedures are as follows.

Before cell plating, the background cell index of each well was examined. 20  $\mu L$  of media was then added into each well of the 384x well E-plate using the VIAFLO 300  $\mu L$  multichannel pipet. The plate was spun for one minute in approximate 1,000 *rpms* to bring the media down to the bottom of the wells and loaded onto the stacker in the Cytomat 2C incubator. After the background cell

indexes were recorded, the plate was removed from the incubator and stored inside the TC hood.

After the mixed cell suspension was transferred from the T75 flask into 50 *mL* tube and mixed three times in the tube, the 500  $\mu L$  of cell suspension was transferred into microcentrifuge tube. 10  $\mu L$  of the cell suspension aliquot was transferred into fresh microcentrifuge tube with a 20  $\mu L$  Rainin Pipet and 10  $\mu L$  of Cedex Trypan Blue solution was added into the same tube (1:2 dilution) to mix well. The Cedex automated cell counter was used for *HepG2* cell line. The entire volume of the diluted sample was transferred into the Cedex counting slide and the cell density per *mL* was determined using the automated cell counter. 40  $\mu L$  of media with cells was added into each well of the E-plated to make 60  $\mu L$  per well. The E-plates were incubated with cells at room temperature for 30 minutes and then added into the Cyto2 incubator.

The cells were seeded (4,000 cells per well) and incubated for 20 ~ 24 hours. When *CI* reached to 10% or 20% of the maximum value of *CI*, solutions of each chemical with 11 concentration levels were applied onto the wells with automatic pipetting. Then, the cells were incubated for 89 hours including initial attachment/growth to ensure that the cells were in exponential growth state and the toxicity results were directly affected by the tested chemicals. The instrument recorded the time-dependent values of *CI* continuously. The readings of the *CI* were once per minute within the first 8 minutes, once every 15 minutes in the following 7 hours, and once every 2 hours afterwards until the experiment was completed. Both short-term and long-term responses were monitored. Short-term response referred to cellular reaction within the phase of the first 12 hours after treatment and long-term response referred to the phase from the 12<sup>th</sup> hour to the 89<sup>th</sup> hour.



Histamine					Rotenone					Brefeldin A					Antinomycin D					
1	SAHA									control negative	SLS							SAHA	1	
1																				1
2	(S)-HDAC-42																		(S)-HDAC-42	2
2																				2
3	Mitoxantrone																	Mitoxantrone	3	
3																			3	
4	Mitomycin C									DMSO 0.2 %	DMSO 0.04 %							Mitomycin C	4	
4										DMSO 0.04 %	DMSO 0.2 %								4	
5	NU 7026																	NU 7026	5	
5																			5	
6	CRT0044876									SLS	control negative							CRT0044876	6	
6																				6
7	Topotecan																		Topotecan	7
7																				7
Forskolin					Paclitaxel					Thapsigargin					5-FU					

Figure 2.1: Layout of the experiment plate (384x). In order to achieve robust and reliable results, the experiment for each chemical was repeated in quadruplicates. The concentrations were in a descending order from the 1<sup>st</sup> column to the 11<sup>th</sup> column in the left hand side and from the 24<sup>th</sup> column to the 14<sup>th</sup> column in the right hand side. The serial dilution ratio was 1:3. Negative control and DMSO (high: 0.2%, low: 0.067% or 0.04%) control panels were placed in the 12<sup>th</sup> and 13<sup>th</sup> columns in order to minimize the effect of chemicals under different concentrations.

## 2.2 Data preprocessing

### 2.2.1 Calculation of normalized cell index (*NCI*)

All TCRPs are normalized by dividing  $CI$  at each time instant by  $CI$  at a reference time instant  $CI(0)$  as in Eq. (2.2). Therefore, normalized cell index ( $NCI$ ) is 1 at the reference time point. Since the assays of chemicals come from different batches of experiments, only the initial 72 hours are considered.

$$NCI(k) = \frac{CI(k-1)}{CI(0)}, \quad k = 1, 2, \dots, K \quad (2.2)$$

### 2.2.2 Denoising

The proposed classification approach employed a denoising algorithm introduced by Pan and Huang [20]. The denoising algorithm overcomes the lim-

itation of various sensors and the disturbance, and increases the accuracy of measurement, feature extraction and classification analysis. The main principle behind the denoising algorithm is to detect and interpolate the abnormal difference in  $CI$ ,  $\Delta NCI(k)$ :

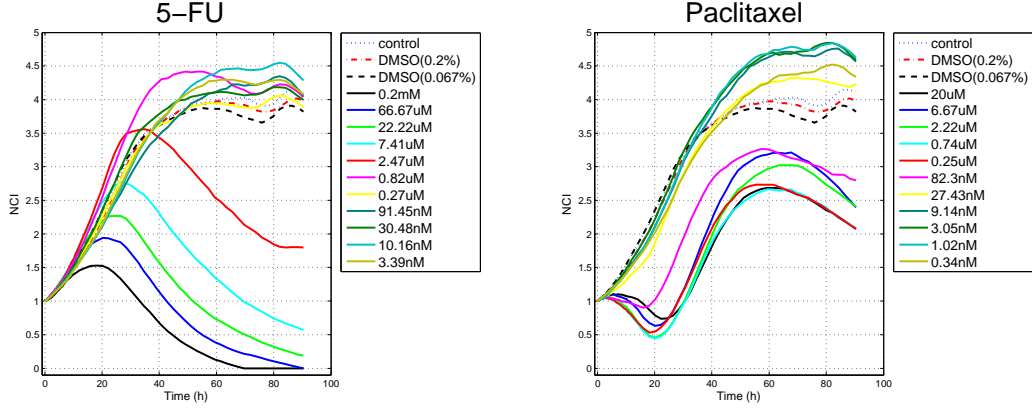
$$\Delta NCI(k) = NCI(k+1) - NCI(k), \quad k = 1, \dots, K-1 \quad (2.3)$$

The detection is similar to the empirical  $3\sigma$  rule. A magnitude coefficient  $\alpha$  is used to control the possibility of the abnormal cell deviation. According to Pan and Huang,  $\alpha = 2$ , which means a 95% confidence interval is used to screen out outliers [20]. If  $\Delta NCI(k)$  is located outside of the confidence interval, we are approximately 95% confident that the difference in  $CI$  is abnormal. Then a linear interpolation is used within the nearest two points of the abnormal index  $k$  to substitute the original  $NCI(k)$ .

In order to distinguish the TCRP patterns of the tested chemicals from the control patterns and to make a uniform comparison, the relative  $NCI$  ( $RNCI$ ) is calculated by dividing  $NCI$  of each chemical ( $NCI^d(k)$ ) by the  $NCI$  of the negative control or vehicle control ( $NCI^c(k)$ ) at the  $k^{th}$  time instant, as in Eq. (2.4). Specifically, the average representative of negative control was used to normalize  $CI$  for water soluble substances at 11 concentrations and non water soluble substances for the  $3^{rd} - 11^{th}$  concentrations. The average representative of 0.2% DMSO was used to normalize  $CI$  for non-water soluble substances at the  $1^{st}$  highest concentration; 0.067% and 0.04% for non water soluble substances at the  $2^{nd}$  high concentration.

$$RNCI(k) = \frac{NCI^d(k)}{NCI^c(k)} \quad k = 1, 2, \dots, K \quad (2.4)$$

The raw TCRPs of two chemicals collected from the experiments are illustrated in Figure 2.2.

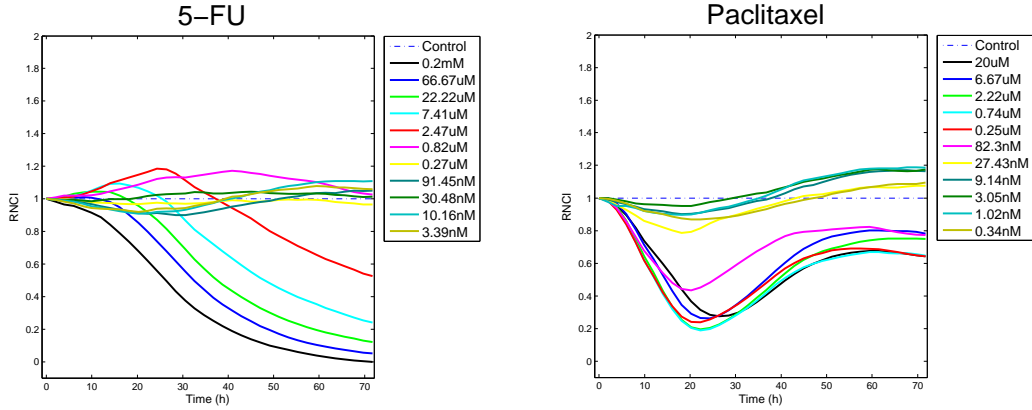


(a) TCRPs of 5-FU depict cellular reaction after cells HepG2 are exposed to different concentrations:  $C_e = 0$  (control line & DMSO),  $C_e = 0.2$  mM,  $C_e = 66.67 \mu\text{M}$ ,  $C_e = 22.22 \mu\text{M}$ ,  $C_e = 7.41 \mu\text{M}$ ,  $C_e = 2.47 \mu\text{M}$ ,  $C_e = 0.82 \mu\text{M}$ ,  $C_e = 0.27 \mu\text{M}$ ,  $C_e = 91.45 \text{ nM}$ ,  $C_e = 30.48 \text{ nM}$ ,  $C_e = 10.16 \text{ nM}$ ,  $C_e = 3.39 \text{ nM}$ . Cell indexes of all TCRPs increase consistently within the initial 20 hours while some show decreasing values after the 20<sup>th</sup> hour.

(b) TCRPs of Paclitaxel depict cellular reaction after cells HepG2 are exposed to different concentrations:  $C_e = 0$  (control line & DMSO),  $C_e = 20 \mu\text{M}$ ,  $C_e = 6.67 \mu\text{M}$ ,  $C_e = 2.22 \mu\text{M}$ ,  $C_e = 0.74 \mu\text{M}$ ,  $C_e = 0.25 \mu\text{M}$ ,  $C_e = 82.3 \text{ nM}$ ,  $C_e = 27.43 \text{ nM}$ ,  $C_e = 9.14 \text{ nM}$ ,  $C_e = 3.05 \text{ nM}$ ,  $C_e = 1.02 \text{ nM}$ ,  $C_e = 0.34 \text{ nM}$ . Cell indexes of some TCRPs decrease in the initial 20 hours, recover to increase gradually after the 20<sup>th</sup> hour and decrease after the 60<sup>th</sup> hour.

Figure 2.2: TCRPs of 5-FU and Paclitaxel

According to the above mentioned standards, TCRPs of chemicals are transferred into relative TCRPs for further analysis. Examples of relative TCRPs are shown in Figure 2.3.



(a) Relative TCRPs of 5-FU. TCRPs associated with low concentrations are close to the baseline 1.

(b) Relative TCRPs of Paclitaxel. TCRPs associated with low concentrations are close to baseline 1.

Figure 2.3: TCRPs of 5-FU and Paclitaxel (relative)

### 2.2.3 Eligible TCRP selection

An eligible TCRP is a cellular response to a toxicant at a specific concentration level that indicates distinguishable toxicity effect with a clearly different shape from the TCRP of the control line which is a toxic-free cellular response. An ineligible TCRP is then defined as a cellular response which is close to the TCRP of the control line. In our study, we use “Feature Extraction Step 1: Identification of Eligible TCRPs” introduced by Pan and Huang [20] to exclude ineligible TCRPs and retain eligible TCRPs. A screening index  $\Xi$  is defined as

$$\begin{cases} \Xi_m = \sum_{k=1}^K \xi_m(k) \\ \xi_m(k) = \begin{cases} 0 & 1 - \delta \leq RNCI_m(k) \leq 1 + \delta \\ 1 & \text{otherwise} \end{cases} \end{cases} \quad (2.5)$$

where  $m = 1, 2, \dots, M$  denotes the  $m^{\text{th}}$  concentration in  $M$  TCRPs (here,  $M = 11$ ),  $k = 1, 2, \dots, K$  is the sampling instant in a TCRP (here,  $K = 36$ , sampling interval of 2 hours), and  $\delta$  is an adjustable empirical threshold to screen eligible TCRPs. The  $\delta$  value in Eq. (2.5) determines the number of TCRPs for the classification algorithm. The larger  $\delta$  is, the smaller number of TCRPs is retained. To guarantee an adequate number of TCRPs for pattern recognition and reasonable features, the variation of negative control is suggested as reference. Coefficient of variation (CV%) is used to measure the inter/intra-plate reproducibility of the experiments. As calculated, the CV% of all intra-Plates are less than 17.9%; the  $\delta$  value should be greater than the maximum CV% of negative control. Therefore,  $\delta = 0.2$  is suggested. The  $m^{\text{th}}$  TCRP with  $\Xi_m > \epsilon$  (here,  $\epsilon = 10$ ) is eligible and included for further consideration.

The selection of eligible TCRPs also indicates the selection of specific concentration levels in the experiments. Meanwhile, the method can be used to screen chemicals with inconspicuous patterns. A chemical will be excluded

from being used in our proposed classification methods if at most two TCRPs are selected as eligible TCRPs. In this case, we consider the chemical illegible and without meaningful patterns, possibly because of experimental disturbance or design. For this reason, the inclusion of such chemicals may be misleading. By using the screening index  $\Xi$ , we successfully screen out ineligible TCRPs as well as ineligible chemicals. Ineligible TCRPs of chemicals will be deleted and ineligible chemicals will be categorized into “Unclassified group”.

According to the Eq. (2.5), relative TCRPs are selected and preserved. The process how TCRPs are selected is illustrated in Fig 2.4.

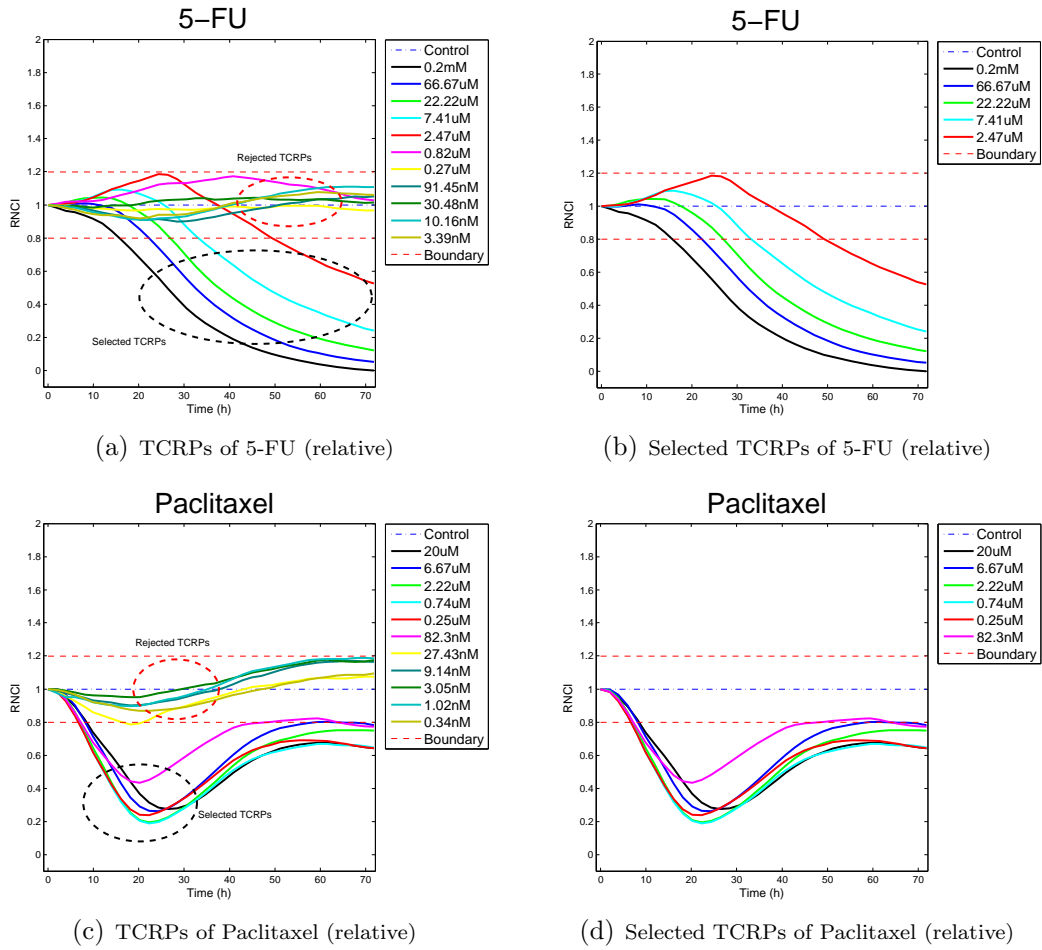


Figure 2.4: Relative TCRPs and selected relative TCRPs of 5-FU and Paclitaxel. Note that each chemical is now represented using TCRPs with different numbers of concentrations (profiles) retained after eligible profile selection.

## 2.3 Nonuniform dimensionality problem

The key problem after data processing lies in the existence of nonuniform dimensionality associated with each chemical data object. As indicated in 2.4, 5-FU has 5 eligible TCRPs while Paclitaxel has 6 eligible TCRPs. Therefore, 5-FU is denoted by a 36-by-5 matrix while Paclitaxel is denoted by a 36-by-6. Although both of them share the same time length which is 36 time intervals, the difference in column numbers makes the comparison and classification of chemicals challenging. In order to unify the classified chemicals for feasible comparison, feature extraction strategies are proposed and used to filter the data associated with chemicals and generate new data for manipulation.

## 2.4 Conclusion

This chapter introduced how the experiments were designed and carried out. In order to filter irrelevant information out, a data processing strategy including denoising and selection of eligible TCRP and chemicals was presented to smooth and screen TCRPs. Although the TCRPs associated with each chemical were denoised and cleaned, the nonuniform dimensionality which resulted from the data processing procedure made the comparison among chemical data difficult. Therefore, extracting valid features from such data becomes necessary and crucial. In the following chapters, three different approaches are proposed to handle this problem and then cluster data into proper groups.

## Chapter 3

# Majority-voting and $K$ -means integrated classification

In this chapter, a classification method including majority-voting slope feature extraction and  $K$ -means clustering algorithm based on TCRPs is explored. First, the majority-voting based feature extraction is elaborated in detail.  $K$ -means clustering algorithm is then introduced and implemented on the extracted feature vectors. The proposed classification approach addresses the problem of clustering chemicals into similar groups where TCRPs show similar tendency with a good performance.

### 3.1 Introduction

In order to classify chemicals which are denoted by different numbers of eligible TCRPs, the first problem to be solved is how to unify the dimension of data matrices. Data matrix associated with each chemical is composed of different numbers of TCRPs. Although the RNCI's in the TCRPs are recorded at the same time instants, the number of eligible TCRPs varies from one chemical to another. A feasible method to unify the data matrices is to tag each chemical with a feature vector. To achieve this goal, all the TCRPs of a chemical are

represented using a feature vector. Some basic statistical indices can be used, e.g. mean and median of TCRPs. However, important information about multiple TCRPs is lost inevitably as mean or median can merely reflect the average tendency of all the profiles of each chemical. The diversity among the tendencies of TCRPs is not taken into consideration. To address this problem, the majority-voting based feature extraction is utilized as a way to capture as much information about the major tendency of TCRPs as possible.

This chapter is organized as follows. First, the majority-voting based feature extraction is elaborated in Section 3.2. The proposed feature extraction scheme is illustrated using several examples. In Section 3.3,  $K$ -means clustering algorithm is considered. With its application on the extracted feature vectors, the classification of chemicals is addressed. In Section 3.4, classification results are summarized in a table and analysis of results is also presented. Section 3.5 concludes this chapter.

## 3.2 Majority-voting based feature extraction

### 3.2.1 Slope quantization

As ineligible TCRPs as well as the corresponding concentrations are screened out, feature extraction will be imposed on each chemical data represented by different numbers of TCRPs. All chemicals are then denoted as a cubic matrix of diverse dimensionality.

$$\mathbf{X} = \{RNCI_j^k(T)\} \quad (3.1)$$

where  $T$  is sampling instant,  $T = 1, \dots, 36$ . There are  $M$  concentration levels  $\{c_p\}_{p=1}^M$ .  $\{c_p\}_{p=1}^M$  denotes the concentration serial number after selection.  $k$  is the index vector for eligible chemicals.

The main challenge lies in the unification of data with non-uniform di-



mensionality. Due to the non-uniform dimensionality, a majority-voting based feature extraction is applied in order to fuse multiple curves into one representative profile based on the assumption that the shape displayed from TCRPs of a chemical is decided by the major tendency of profiles.

Due to uniform sampling, the slope of each profile can be denoted as  $\Delta X^{(k)} = \{\Delta RNCI_j^k(t)\}$ , where  $t$  is the sampling interval,  $t = 1, \dots, T - 1$ . Given some thresholds, a quantization for slope is as follows:

$$\Delta X^{(k)} = \{\Delta RNCI_j^k(t)\} \rightarrow \{B_j^k(t)\}_{j=c_1}^{c_M} = \begin{cases} 1 & \Delta RNCI_j^k(t) \geq \delta_1 \\ 0 & -\delta_1 \leq \Delta RNCI_j^k(t) < \delta_1 \\ -1 & -\delta_2 \leq \Delta RNCI_j^k(t) < -\delta_1 \\ -2 & \Delta RNCI_j^k(t) < -\delta_2 \end{cases} \quad (3.2)$$

where  $\delta_1, \delta_2$  are tuning parameters (here,  $\delta_1 = 0.001, \delta_2 = 0.1$ ) designed specifically for HepG2 cell line.  $\Omega = \{1, 0, -1, -2\}$  denote an increasing, constant, decreasing and quickly decreasing tendency in profiles accordingly. After all slopes of profiles in time intervals are quantified, a majority voting is carried out.

### 3.2.2 Majority voting

The representative profile is able to reflect the dominant profile tendency and variation from one time interval to another. The majority voting of the  $\{B_j(t)\}$  at individual sample interval  $t$  is carried out for all selected TCRPs. The represented feature vector  $\bar{B}(t)$  of chemical  $k$  is obtained using Eq. (3.3).

$$\bar{B}(t) = \arg \max_{\omega_i \in \Omega} \sum_{m=c_1}^{c_M} I(B_m(t) = \omega_i) \quad (3.3)$$

where  $I(\cdot)$  is an indicator function with  $\omega_i \in \Omega$  in Eq. (3.4).

$$I = \begin{cases} 1 & B_m(t) = \omega_i \\ 0 & otherwise \end{cases} \quad (3.4)$$

If  $B_m(t) = \omega_i$ , then  $I = 1$ , otherwise  $I = 0$ .  $M$  is the number of TCRPs of each chemical included.

### 3.2.3 Slope representatives

All the logical values of  $\Delta RNCI_j(t)$  which are equal to the dominant one are selected out. The corresponding concentration levels compose a new set  $\mathbf{c}(t)$ , which is indicated in Eq. (3.5).

$$\mathbf{c}(t) = \{j | B_j(t) = \bar{B}(t)\}, \quad j = c_1, \dots, c_M \quad (3.5)$$

where  $\{c_p\}_{p=1}^M$  are selected  $M$  concentration levels.

In each sampling interval, the median value of all  $\Delta RNCI_j(t)$  in correspondence with the selected concentrations then forms the feature vectors. The feature vectors describe the major tendency in each sampling interval. Concentrations are involved and contribute diversely in different sampling interval.

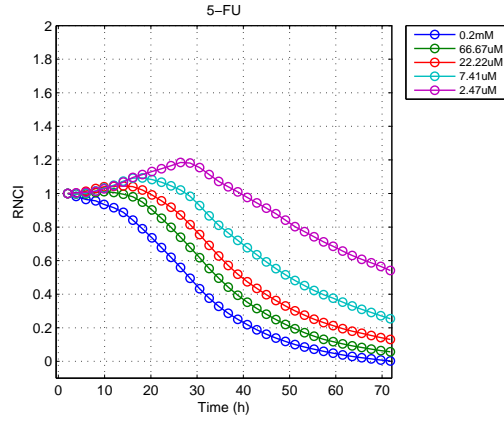
$$\bar{F}(t) = \text{median}_{j \in \mathbf{c}(t)} \{\Delta RNCI_j(t)\} \quad (3.6)$$

where  $\bar{F}(t)$  denotes the feature vectors.

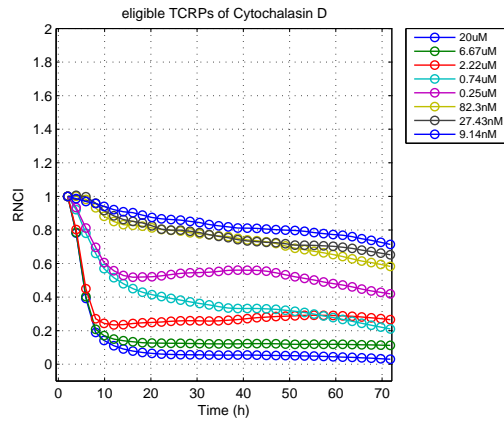
### 3.2.4 Example

A detailed example follows. TCRPs of 5-Fu, Cytochalasin D, Paclitaxel are taken as examples together to illustrate the feature extraction procedure.

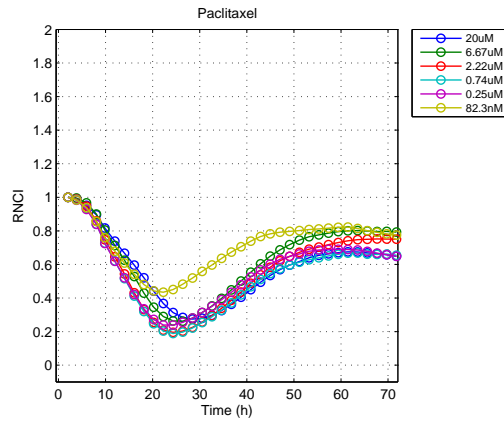
After eligible profile selection, 5, 8 and 6 TCRPs are retained accordingly. Figures 3.1(a), 3.1(b) and 3.1(c) show all the eligible TCRPs after selection. The highest level of concentration in each chemical considered is chosen as an example to show the corresponding slope tendency. Figures 3.2(a), 3.2(b) and 3.2(c) show the changing trend of  $\Delta RNCI$  with respect to cumulative time. From the figure, we observe that  $\Delta RNCI$ 's of Cytochalasin D in the first three sampling intervals is far smaller than a predefined threshold  $-\delta_2$  (here,  $\delta_2$  is 0.1). It discriminates itself from Figure 3.2(a) and Figure 3.2(c) where  $\Delta RNCI$ 's vary within the yellow range. Figures 3.3(a), 3.3(b), 3.3(c) show the logical representation of slopes after quantization. As the slopes are quite small, the logical values within the first three sampling intervals in TCRPs of Cytochalasin D are set as -2. Figures 3.4(a), 3.4(b) and 3.4(c) show the results of majority voting on logical variables accordingly.



(a) Eligible TCRPs of 5-FU

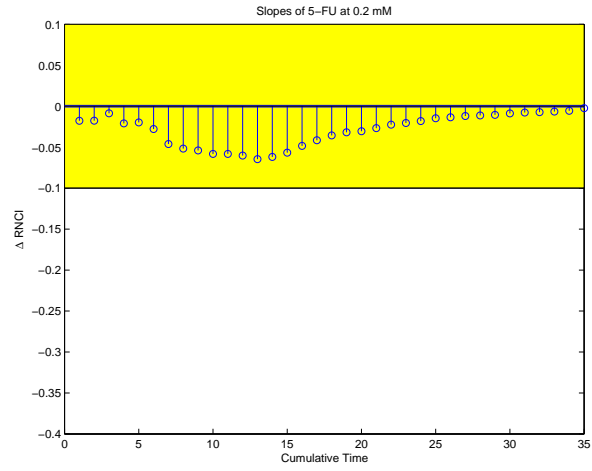


(b) Eligible TCRPs of Cytochalasin D

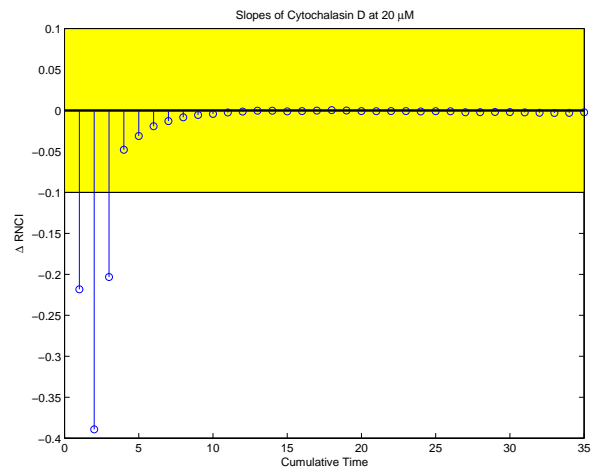


(c) Eligible TCRPs of Paclitaxel

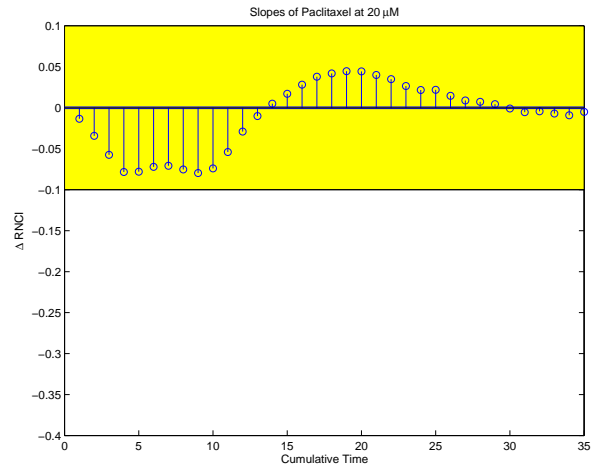
Figure 3.1: Eligible TCRPs of 5-FU, Cytochalasin D and Paclitaxel



(a) Slopes of 5-FU at 0.2 mM

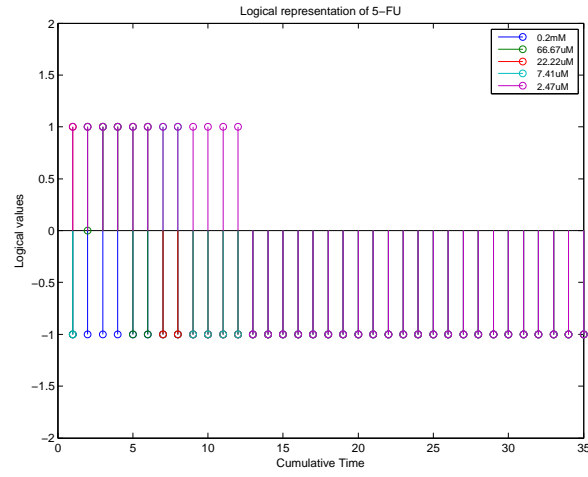


(b) Slopes of Cytochalasin D at 20  $\mu M$

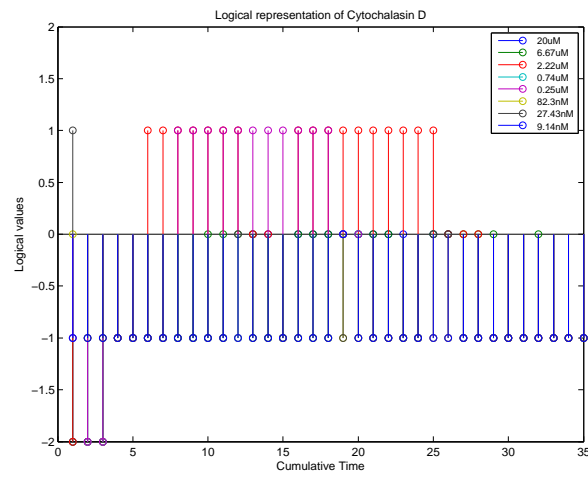


(c) Slopes of Paclitaxel at 20  $\mu M$

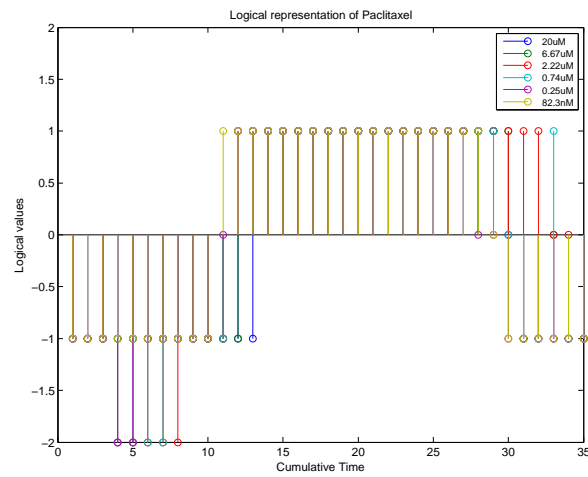
Figure 3.2: Slopes of 5-FU, Cytochalasin D and Paclitaxel at the highest concentration level



(a) Logical representation of 5-FU

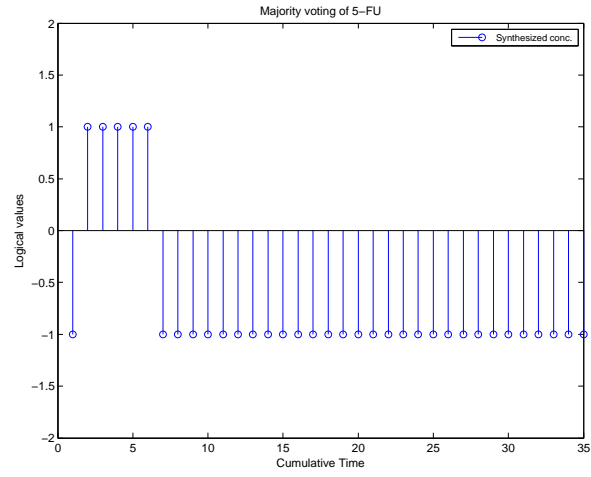


(b) Logical representation of Cytochalasin D

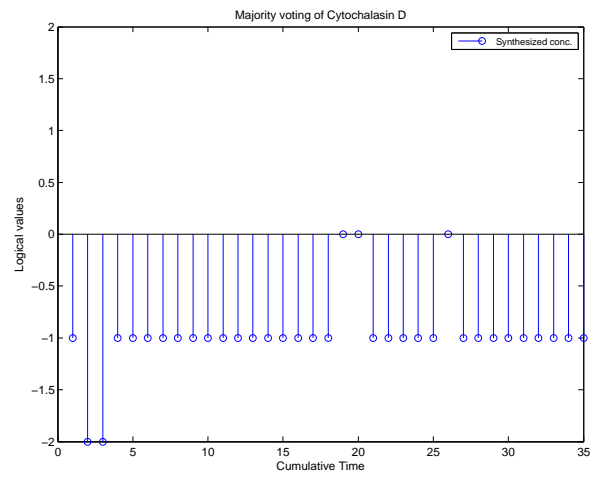


(c) Logical representation of Paclitaxel

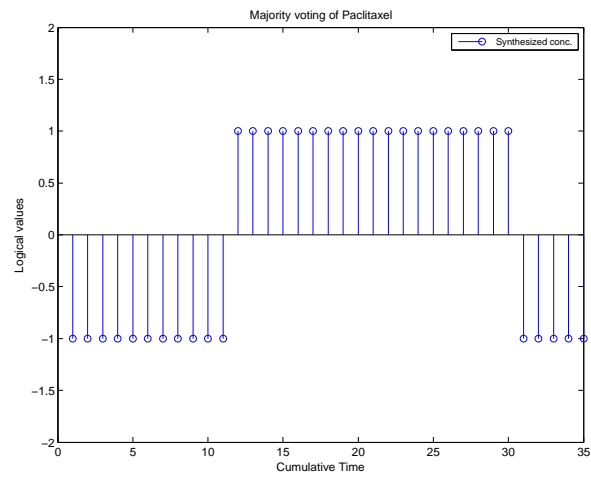
Figure 3.3: Logical representation



(a) Majority-voting logical representation of 5-FU



(b) Majority-voting logical representation of Cytochalasin D



(c) Majority-voting logical representation of Paclitaxel

Figure 3.4: Majority-voting logical representation

It should be noted that sometimes conflicts are observed in this majority-voting. Table 3.1 shows the distribution of logical values. Conflicts occur at the 1<sup>st</sup> and 26<sup>th</sup> entry where the amount of negative logical values (3) is equal to the amount of quick negative logical values (3) and the amount of constant logical values (4) is equal to the amount of negative logical values (4).

Table 3.1: Distribution of logical values (Cytochalasin D)

Sampling interval	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_{25}$	$t_{26}$	$\dots$	$t_{32}$	$t_{33}$	$t_{34}$	$t_{35}$
Positive	1	0	0	0	$\dots$	1	0	$\dots$	0	0	0	0
Constant	1	0	0	0	$\dots$	3	4	$\dots$	1	0	0	0
Negative	3	3	3	8	$\dots$	4	4	$\dots$	7	8	8	8
Quick Negative	3	5	5	0	$\dots$	0	0	$\dots$	0	0	0	0

To solve the conflict situations, rules of identifying logical vectors with conflicts are designed as follows:

- If the conflict does not appear in the first interval, the majority-voting result at an interval follows the result in its previous interval provided the curve tendency within two neighbouring intervals does not vary rapidly ( $l_t = l_{t-1}$ );
- If the conflict is observed in the 1<sup>st</sup> interval, the constant logical variables are not taken into consideration because they are close to the horizontal line and do not show obvious tendency. Instead,  $\arg \max_{\omega_i \in \Omega \setminus \{0\}} \sum_{m=c_1}^{c_M} I(B_m(t) = \omega_i)$  is calculated and becomes the logical feature if the conflict is solved. However, if the conflict is not solved, two situations are considered separately:

1. if there is still conflict on logical variables, and

$$\{-2\} \subset \arg \max_{\omega_i \in \Omega \setminus \{0\}} \sum_{m=c_1}^{c_M} I(B_m(t) = \omega_i) \quad (3.7)$$

then  $l_1 = -2$ ;



2. if there is still conflict on logical variables, and

$$\{-2\} \not\subset \arg \max_{\omega_i \in \Omega \setminus \{0\}} \sum_{m=c_1}^{c_M} I(B_m(t) = \omega_i) \quad (3.8)$$

then  $l_1 = 0$ .

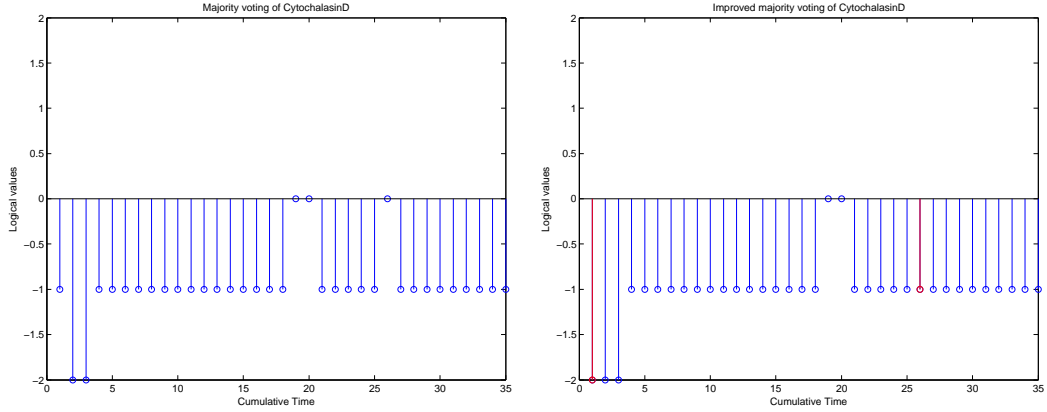
Tables 3.2 and 3.3 display the change of logical variables. Figures 3.5(a) and 3.5(b) are illustrated to compare the change of the logical feature of Cytochalasin D. The improved entries are marked with red in Figure 3.5(b).

Table 3.2: Logical representation with conflict (Cytochalasin D)

Sampling interval	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_{25}$	$t_{26}$	$\dots$	$t_{32}$	$t_{33}$	$t_{34}$	$t_{35}$
Logical feature	?	-2	-2	-1	$\dots$	-1	?	$\dots$	-1	-1	-1	-1

Table 3.3: Improved logical representation (Cytochalasin D)

Sampling interval	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_{25}$	$t_{26}$	$\dots$	$t_{32}$	$t_{33}$	$t_{34}$	$t_{35}$
Logical feature	<b>-2</b>	-2	-2	-1	$\dots$	-1	<b>-1</b>	$\dots$	-1	-1	-1	-1



(a) Original majority-voting logical representation of Cytochalasin D (b) Improved majority-voting logical representation of Cytochalasin D

Figure 3.5: Improvement of majority-voting logical representation

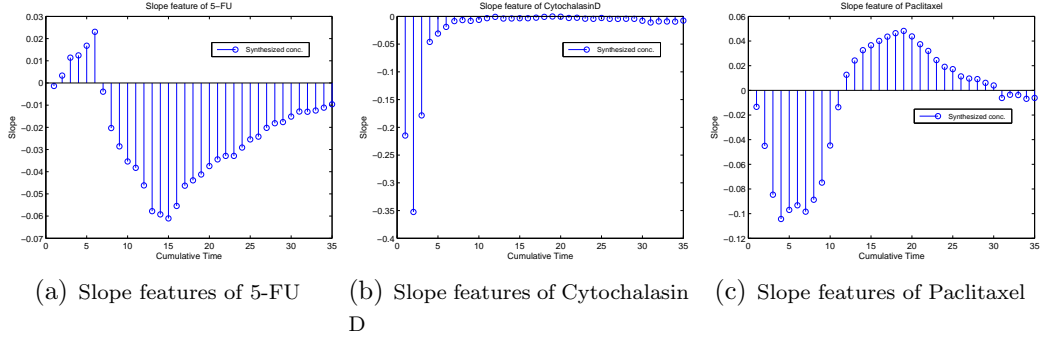


Figure 3.6: Slope features

The median of slopes whose logical representations are dominant in each sampling interval is taken as the slope feature for the corresponding entry. Figures 3.6(a), 3.6(b), 3.6(c) display the slope features of TCRPs of three chemicals. If there is a conflict in logical variables in an interval, the median of slopes whose logical representations are equal to  $l_{t-1}$  is taken as representatives. If

$$\mathbf{c}(t) = \{j | B_j(t) = \bar{B}(t-1), j = c_1, \dots, c_M\}$$

is an empty set, the median value of *all* slopes in the interval is taken as the slope feature for the entry. Thus, a chemical with diverse TCRP volumes is consistently represented by a 1-by-35 slope feature vector.

### 3.3 *K*-means clustering

*K*-means clustering is one of the unsupervised clustering algorithms and commonly used in many areas such as signal processing, market segmentation, computer vision etc. In cluster analysis, *K*-means method is aimed to partition the input data points into *K* clusters.

In this section, *K*-means clustering is elaborated with mathematical equations from [21]. Suppose a data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  is composed of *N* observations with *D* dimensions. A cluster is formulated as a group of data points whose inter-point distances are smaller than the distances to points outside the clus-

ter.  $D$ -dimensional vectors  $\boldsymbol{\mu}_k$  ( $k = 1, \dots, K$ ) are introduced to denote the prototype of each cluster. In  $K$ -means method, the prototypes are the centers of the clusters. Given a number  $K$ , the problem is formulated as assigning all data points to clusters such that the sum of distances of data points to its closest vector  $\boldsymbol{\mu}_k$  is minimized.

As an optimization problem, an objective function is defined as

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 \quad (3.9)$$

where  $\{\mathbf{x}_n\}_{n=1}^N$  are data points.  $r_{nk}$  is a binary indicator variable for  $\mathbf{x}_n$ ,  $n = 1, \dots, N$ ,  $k = 1, \dots, K$ . It is used to identify which cluster  $\mathbf{x}_n$  belongs to:

$$\begin{cases} r_{nk} = 1 & \text{if } \mathbf{x}_n \text{ is assigned to the } k^{th} \text{ cluster} \\ r_{nj} = 0 & \text{if } j \neq k. \end{cases} \quad (3.10)$$

Each data point is then coded into a 1-by- $K$  vector. To find values of  $r_{nk}$  and  $\{\boldsymbol{\mu}_k\}$  to optimize the objective function  $J$ , an iterative method including two successive steps is involved to realize a successive optimization. Some values for  $\boldsymbol{\mu}_k$  are initialized first. The objective function,  $J$ , is minimized with respect to  $r_{nk}$  with  $\boldsymbol{\mu}_k$  fixed. Then  $J$  is minimized with respect to  $\boldsymbol{\mu}_k$  with  $r_{nk}$  fixed. The new  $\boldsymbol{\mu}_k$  is then used as an input for optimization in successive iterations until  $\boldsymbol{\mu}_k$  and  $r_{nk}$  are convergent.

The binary indicator variable is determined in the following way.  $J$  in E 3.9 is a linear function of  $r_{nk}$ . Terms involving  $n$  are independent. Therefore, choosing  $r_{nk}$  to be 1 for the value of  $k$  that gives the minimum value of  $\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$  can optimize the objective function, which is indicated in Eq. (3.11).

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (3.11)$$

Once the  $r_{nk}$  is identified, the optimization is imposed on  $\boldsymbol{\mu}_k$ . Setting the derivative of  $J$  with respect to  $\boldsymbol{\mu}_k$  to zero, we get

$$2 \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \quad (3.12)$$

Therefore,  $\boldsymbol{\mu}_k$  is given as

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} \quad (3.13)$$

where the denominator is equal to the number of all of the data points assigned to cluster  $k$ . Eq. (3.13) is interpreted in a way that the prototype of each cluster, i.e.  $\boldsymbol{\mu}_k$ , is set as the mean value of the data points  $\mathbf{x}_n$  assigned in cluster  $k$ .

The iteration of assigning data points continues until the maximum number of iterations is reached or the estimation of parameters including binary indicator  $r_{nk}$  and cluster mean  $\boldsymbol{\mu}_k$  converges. However, some drawbacks of this algorithm should be mentioned. First, a local minimum value of  $J$  may be reached during the two-step iteration. Convergence properties of  $K$ -means method can be referred to [22]. Second,  $K$ -means algorithm performs a hard assignment of data points to clusters in which each data point belongs to one cluster uniquely. Data points which lie in the boundaries are sometimes misclassified. Third, the number of clusters is supposed to be specified in advance. For more details about the application of  $K$ -means algorithm and its connection to the mixture model, the readers are referred to Bishop and Nasrabadi [21].

## 3.4 Results and discussion

### 3.4.1 Results

Results of classification on all eligible chemicals are shown in this section. Eligible TCRPs of chemicals are extracted using majority voting and then input into  $K$ -means clustering elaborated above. Figure 3.7 displays all of the feature vectors of eligible chemicals using majority voting method.

Function *kmeans* in MATLAB<sup>®</sup> R2011b (version 7.13) is used to generate the results. The distance parameter set in the parameter list is cityblock distance. It is the sum of absolute differences, known as  $L1$  distance between two feature vectors. Because initial values to start  $K$ -means are chosen differently each time and  $K$ -means algorithm is an algorithm converging quickly to a local optimum, the replicates of clustering is set to 500 so as to obtain robust results.

According to the prior knowledge about cluster number, 6  $\sim$  10 is proper. 6 is chosen as the cluster number input to generate the results listed in Table 3.4.

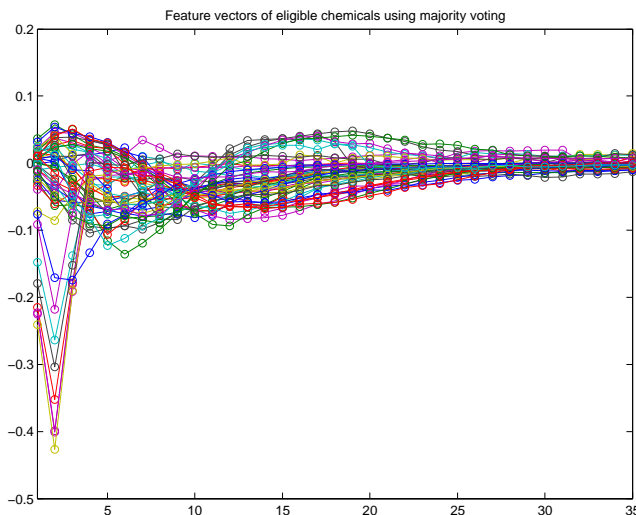
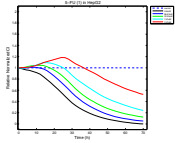
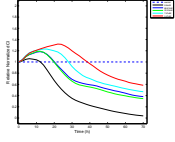
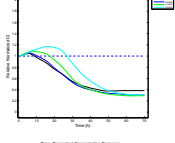
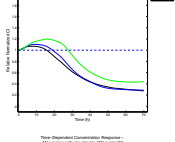
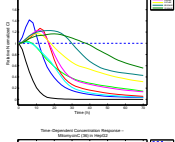
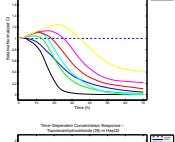
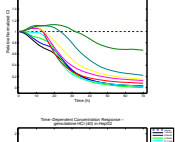
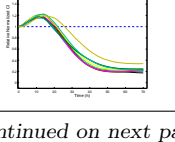


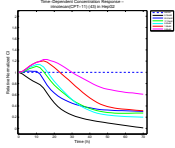
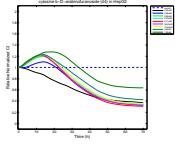
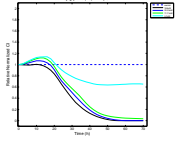
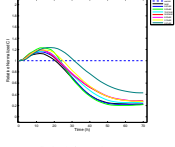
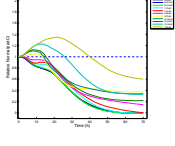
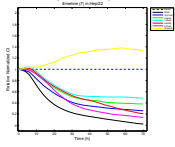
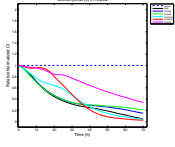
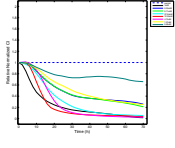
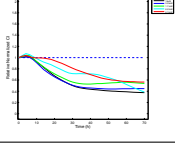
Figure 3.7: Extracted feature vectors of eligible chemicals using majority voting methods

Table 3.4: Results of Majority-voting and  $K$ -means integrated classification (cell line: HepG2. 47 of 65 chemicals are eligible.)

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
1	1	5-fluorouracil (5-FU)	DMSO	200 $\mu$ M – 3.39nM	
	2	Etoposide phosphate	DMSO	200 $\mu$ M – 3.39nM	
	12	Clofarabine (CLOF)	H <sub>2</sub> O	25 $\mu$ M – 0.42nM	
	13	Hydroxyurea (HU)	H <sub>2</sub> O	10mM – 169nM	
	35	Mitoxantrone	DMSO	150 $\mu$ M – 2.54nM	
	36	Mitomycin C	DMSO	200 $\mu$ M – 3.39nM	
	39	Topotecan	DMSO	95 $\mu$ M – 1.61nM	
	40	2'-Deoxy-2'	H <sub>2</sub> O	1650 $\mu$ M – 27.94nM	

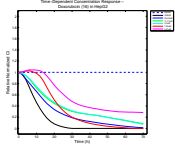
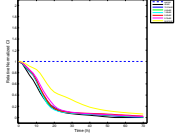
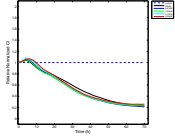
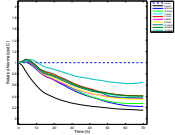
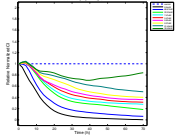
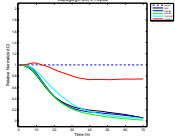
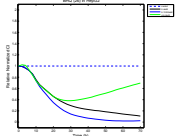
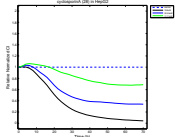
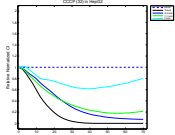
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	43	Irinotecan (CPT-11)	DMSO	160 $\mu$ M – 2.71nM	
	44	Cytosine	H <sub>2</sub> O	8950 $\mu$ M – 151.57nM	
	46	Benzo[a]pyrene	DMSO	100 $\mu$ M – 1.69nM	
	47	Gemcitabine	H <sub>2</sub> O	2 $\mu$ M – 0.03nM	
	60	SN-38	DMSO	200 $\mu$ M – 3.39nM	
2	7	Emetine	H <sub>2</sub> O	50 $\mu$ M – 0.847nM	
	9	Actinomycin D	DMSO	2 $\mu$ M – 0.0339nM	
	10	Puromycin	H <sub>2</sub> O	1000 $\mu$ M – 17nM	
	11	Anisomycin	H <sub>2</sub> O	10 $\mu$ M – 0.17nM	

continued on next page

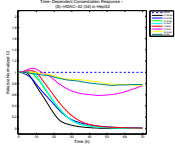
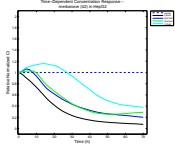
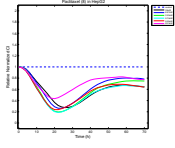
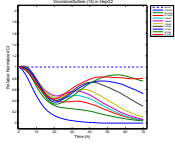
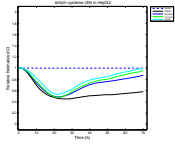
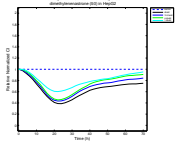
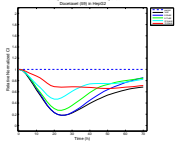
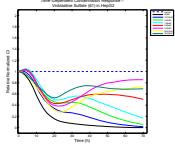
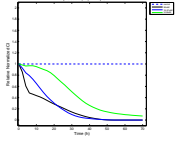
continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	16	Doxorubicin (DOX)	H <sub>2</sub> O	100 $\mu$ M – 1.69nM	
	17	Brefeldin A (BEF)	DMSO	40 $\mu$ M – 0.68nM	
	22	Oligomycin	DMSO	20 $\mu$ M – 0.339nM	
	23	Antimycin A	EtOH	200 $\mu$ M – 3.387nM	
	24	Rotenone	DMSO	200 $\mu$ M – 3.387nM	
	25	Thapsigargin	DMSO	2 $\mu$ M – 0.0339nM	
	26	BHQ	DMSO	400 $\mu$ M – 7nM	
	28	Cyclosporin A	DMSO	100 $\mu$ M – 1.69nM	
	32	CCCP	DMSO	100 $\mu$ M – 1.69nM	

continued on next page

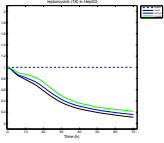
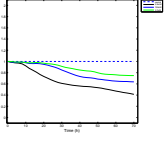
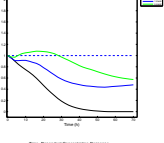
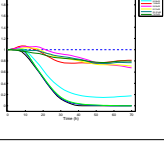
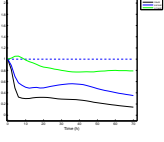
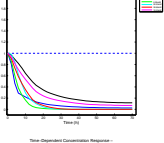
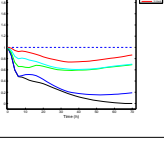
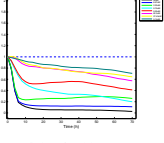
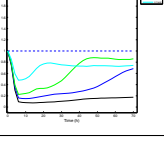


continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	34	(S)-HDAC-42	DMSO	128 $\mu$ M – 2.17nM	
	42	Merbarone	DMSO	200 $\mu$ M – 3.39nM	
3	8	Paclitaxel	DMSO	20 $\mu$ M – 0.339nM	
	15	Vincristine Sulfate	H <sub>2</sub> O	250 $\mu$ M – 4.23nM	
	49	S-trityl-Cysteine	DMSO	100 $\mu$ M – 1.69nM	
	50	Dimethylenastron	DMSO	40 $\mu$ M – 0.68nM	
	59	Docetaxel	DMSO	1 $\mu$ M – 0.02nM	
	61	Vinblastine Sulfate	H <sub>2</sub> O	40 $\mu$ M – 0.68nM	
4	14	Valproic acid	H <sub>2</sub> O	50mM – 847nM	

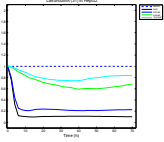
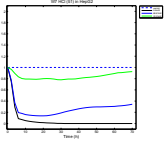
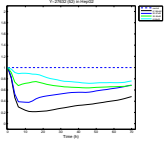
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	18	Leptomycin B (LMB)	EtOH	20nM – 0.000339nM	
	19	Exo 1	DMSO	300μM – 5.08nM	
	20	Monensin	DMSO	4μM – 0.068nM	
	33	SAHA	DMSO	151μM – 2.56nM	
5	5	Cytochalasin B	DMSO	20μM – 0.339nM	
	55	FAKInhibitor14	H <sub>2</sub> O	2500μM – 42.34nM	
	64	HA1100 hydrochloride	H <sub>2</sub> O	1000μM – 16.94nM	
6	4	Cytochalasin D	DMSO	20μM – 0.339nM	
	6	Latrunculin B	DMSO	20μM – 0.339nM	

continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	31	Latrunculin A	EtOH	$2\mu\text{M} - 0.03\text{nM}$	
	51	W7 HCl	DMSO	$200\mu\text{M} - 3.39\text{nM}$	
	52	Y-27632	DMSO	$188\mu\text{M} - 3.18\text{nM}$	

### 3.4.2 Discussion

The majority-voting feature extraction strategy is able to transfer multiple time series with different concentrations into one time series in order that TCRPs with various dimensionality can be integrated and the data points with different representations are therefore comparable with each other. Along the whole time intervals, different segments of TCRPs are involved in the calculation of the logical entry for each sampling interval. However, there is some inevitable loss of curve information during the fusion of multiple curves. For example, as the results of cluster 3 indicate, TCRPs of Vincristine Sulfate (#15) and Vinblastine Sulfate (#61) are mutually similar while TCRPs of Paclitaxel (#8), S-trityltl-Cysteine (#49), Dimethylenastron (#50) and Docetaxel (#59) are similar to each other. We check the feature vectors of chemicals in cluster 3. Figure 3.8 shows the extracted feature vectors of chemicals.

According to Figure 3.8, those feature vectors are close to each other. The pairwise distances are small enough so that  $K$ -means algorithm clusters them together. This is a limitation of the proposed majority-voting feature extraction to represent the several TCRPs into one logical vector. For each sampling interval, some TCRPs with low concentration levels are not involved in the

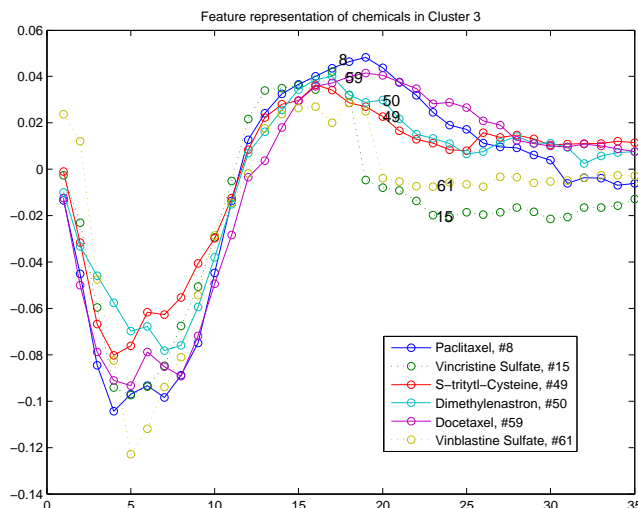


Figure 3.8: Extracted feature vectors of chemicals in cluster 3 using majority voting methods

calculation of feature slopes because the amount of the corresponding logical variable is subordinate. However, sometimes the majority is not absolutely dominant, indicating that the amount of subordinate TCRPs is close to that of dominant ones during slope quantization. For example, the TCRPs of Vincristine Sulfate (#15) and some excerpted logical entries are shown in Figure 3.9(a) and Table 3.5 accordingly. Values marked in red are considered as conflicting situations which can be improved using the methods introduced above. The underlined logical values in those columns are close to each other. By using majority-voting feature extraction, information about minor logical variables and their corresponding TCRPs has to be ignored.

Table 3.5: Excerpted Logical representation (Vincristine Sulfate)

Sampling interval	$t_1$	$t_2$	$t_3$	$t_4$	$t_5$	$t_6$	...	$t_{17}$	$t_{18}$	$t_{19}$	$t_{20}$	...
Positive	<u>4</u>	0	0	0	0	0	...	5	5	<u>4</u>	<u>4</u>	...
Constant	0	0	0	0	0	0	...	0	0	0	0	...
Negative	<u>6</u>	10	9	8	<u>6</u>	<u>6</u>	...	5	5	<u>6</u>	<u>6</u>	...
Quick Negative	0	0	1	2	<u>4</u>	<u>4</u>	...	0	0	0	0	...

Another possible reason for “misclassification” is that the values of  $RNCI$  are not fully utilized in feature extraction. As indicated in the TCRPs of

Vincristine Sulfate (#15) and Paclitaxel (#8) (Figure 3.9), after about 48 hours of the addition of chemicals, the majority of the TCRPs of both chemicals show a decreasing tendency. According to the calculation of slopes, the values for both #15 and #8 are rather close. However, the TCRPs of #15 disperse while those of #8 assemble. Slope values just consider the variability of cells by calculating  $\Delta RNCI$  from one sampling interval to another. Although slopes can indicate the tendency for each TCRP, the spread of TCRPs cannot be reflected in slope values, which leads to the difference in shape between #15 and #8 to some extent.

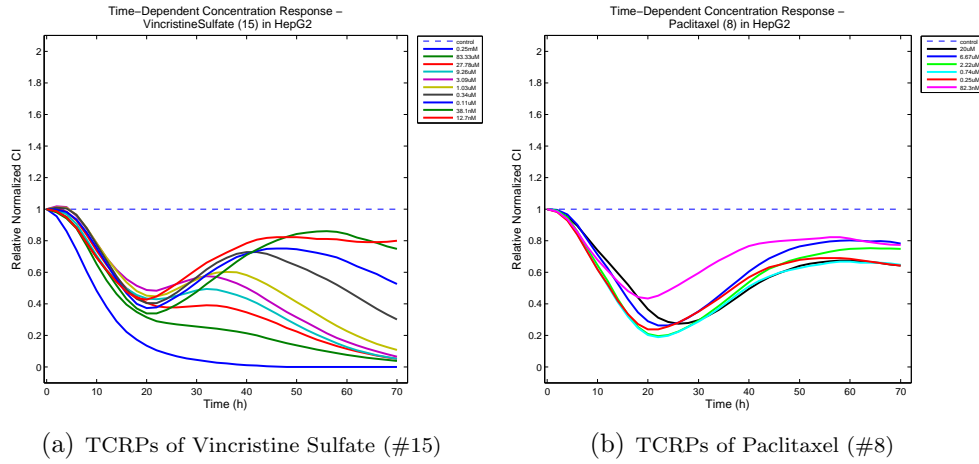


Figure 3.9: TCRPs of Vincristine Sulfate and Paclitaxel

Although there is some deficiency in the majority-voting feature extraction method, the overall clustering results show good discrimination from a curve shape similarity point of view. To avoid unnecessary loss of information, a curve by curve correlation-based clustering approach is proposed and introduced in next chapter.

### 3.5 Conclusion

After data processing, each chemical is depicted by multiple TCRPs with uneven dimensionality. To realize the goal of classifying chemicals using their

TCRPs, standardization about uneven data is imperative and crucial. A majority-voting based feature extraction strategy is proposed to fuse multiple TCRPs into consistent feature vectors with equal dimensionality to compare using slopes. The slope feature vectors represent the chemicals with different segments of TCRPs under diverse concentration levels involved in each sampling interval. All the feature vectors of eligible chemicals consist of a feature matrix. Then  $K$ -means clustering algorithm is briefly introduced and utilized as a key tool to classify the feature matrix. With the help of prior knowledge about the cardinality of clusters, chemicals are classified into six groups. The results indicate good differentiation on profile shapes and can be used for further analysis.

## Chapter 4

# A hierarchical correlation based classification

In this chapter, a classification approach based on correlation coefficients is proposed. In order to classify chemicals associated with Mode of Action, Pan and Huang implemented correlation analysis on logical representations of chemicals to cluster them. If the correlation coefficient between logical feature vectors of any two chemicals was larger than  $\tau$ , a predefined tuning parameter, the two chemicals were grouped together [20].

The proposed curve by curve correlation-based clustering method deals with all the eligible TCRPs in order to keep valid information in each TCRP of a given chemical. Since each chemical is represented by multiple TCRPs, compressed data may not reveal all the information about the response patterns. Also, a challenge to compare the similarity between chemicals is that each chemical is represented by different numbers of eligible TCRPs. Some chemicals may have 3 eligible TCRPs while some others may have all 11 eligible TCRPs. The difference in dimensions makes the comparison of similarity among chemicals difficult. The proposed clustering method avoids data compression and the comparison of dimension and facilitates in quantifying the shape similarity among TCRPs of chemicals and classifying *similar* chemicals into one group.

## 4.1 Introduction

The proposed classification approach consists of two steps. The first step, termed as *categorization*, classifies chemicals according to apparent shapes of TCRPs. The main purpose of this step is to separate chemicals roughly using discernible curve features, which provides the foundation for the next step. The second step, termed as *curve by curve correlation-based clustering*, further improves classification results by grouping *similar* chemicals in each category. It is carried out by analyzing the correlation coefficients between TCRPs of chemicals within the same category. Using the correlation coefficients in a discerning way is a good measure of similarity in shape among TCRPs. It can group chemicals with similar shape patterns within one category. This two-step approach provides good classification results as demonstrated in the following sections.

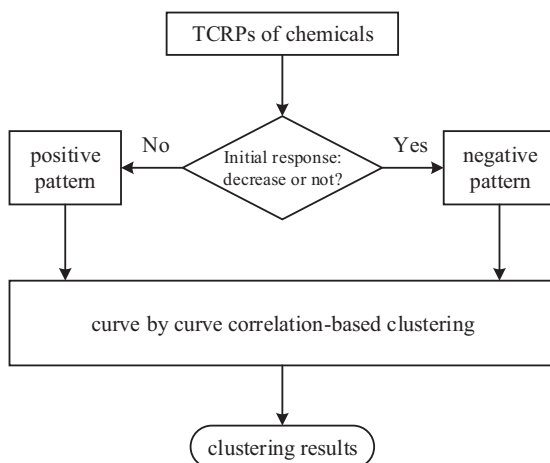


Figure 4.1: Hierarchical correlation based classification schematic structure

Figure 4.1 shows the schematic structure of the hierarchical correlation-based classification method. As seen from Figure 4.1, categorization is aimed at the initial 6 hours after the addition of chemicals. It discriminates chemicals with sharply decreasing responses from those that do not display such a ten-



dency. All chemicals are therefore divided into two categories: chemicals with sharply decreasing responses, termed as *negative pattern* group, and without sharply decreasing responses, termed as *positive pattern* group. The two categories are then subject to curve by curve correlation-based clustering under the hierarchical correlation-based classification framework.

This approach avoids the feature extraction procedure by fully utilizing all of the TCRPs of chemicals and can generate better results as well.

This chapter is organized as follows. First, Section 4.2 introduces how curve tendency is used as a means to discriminate chemicals. A correlation-based clustering method to classify the chemicals using correlation coefficient as a metric to measure similarity is introduced in Section 4.3. Examples about how correlation coefficients are formulated as well as the improvement are elaborated. Section 4.4 displays the results and lists some deficiency of this approach.

## 4.2 Step 1: Categorization

The first-order differences of TCRPs are a good indicator of decreasing and increasing tendencies of curves. A predefined threshold on the first-order differences can distinguish chemicals exhibiting negative pattern from positive pattern. In Figures 4.2 and 4.3, chemicals inducing negative pattern as well as positive pattern are shown respectively.

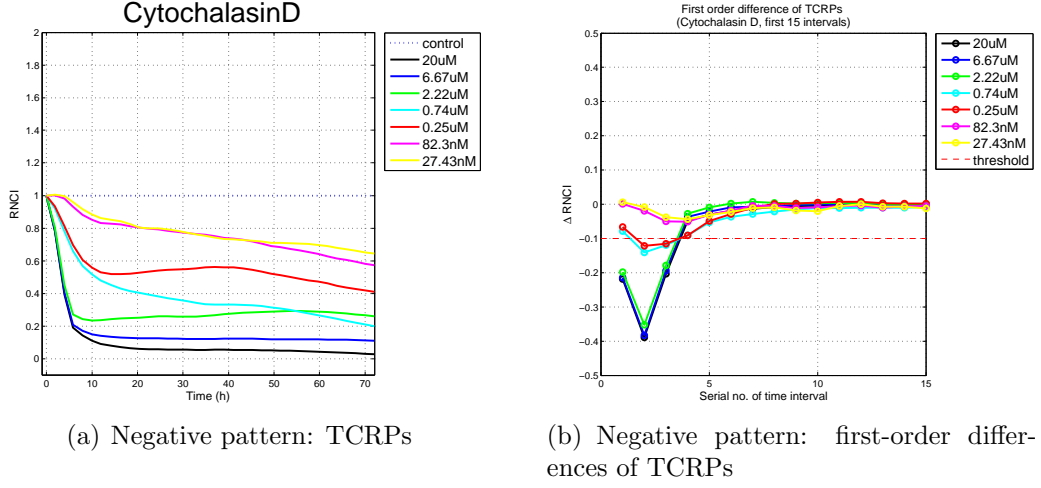


Figure 4.2: Negative pattern

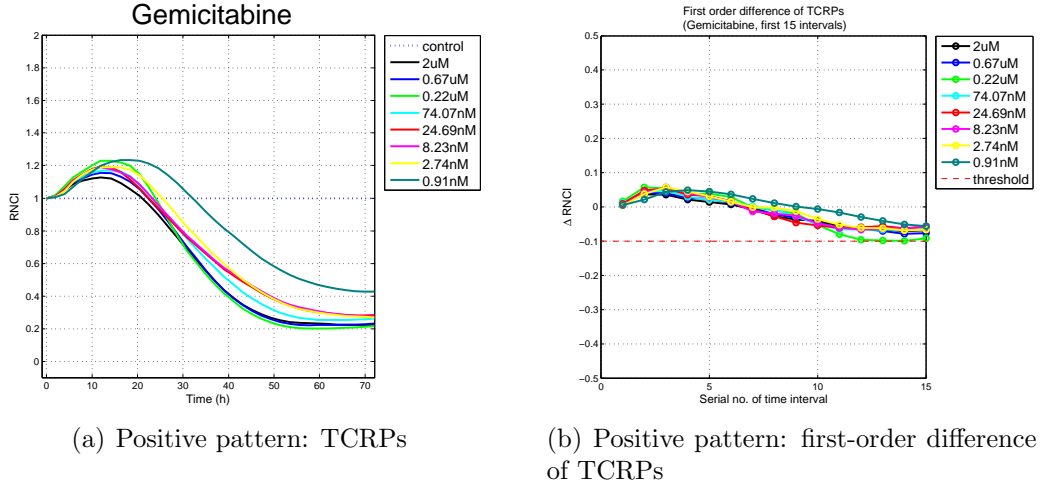


Figure 4.3: Positive pattern

The categorization is aimed at screening out chemicals with a sharply decreasing tendency in the initial 6 hours of cellular response. Some chemicals can kill cells in a very short time period. In order to separate those chemicals, we calculate the first-order differences of  $RNCI$ ,  $\Delta RNCI$ , in Eq. (4.1).

$$\Delta RNCI(k) = RNCI(k+1) - RNCI(k), k = 1, \dots, K-1 \quad (4.1)$$

We notice that the initial pattern of Figure 4.2 is different from Figure 4.3.

The corresponding  $\Delta RNCI$ 's in Figure 4.2 are evidently smaller than an adjustable empirical threshold -0.1 with some even falling down to -0.4. However, the first-order differences of the TCRPs in Figure 4.3 lie within the empirical thresholds as opposed to those in Figure 4.2. Figure 4.2 shows that in the initial phase, cells are killed rapidly by some chemicals under particular concentrations. Therefore, a chemical holds negative pattern if there are more than or equal to  $S$  TCRPs whose first-order differences are *all* less than an adjustably empirical threshold  $\mu$ .

$$i^{th} \text{ chemical} \begin{cases} \in \text{negative pattern} & \text{if } \sum_{k=1}^M I(\Delta RNCI(k) \leq \mu) \geq S \\ \in \text{positive pattern} & \text{otherwise} \end{cases} \quad (4.2)$$

where  $\mu = -0.1$ ,  $S = 2$  by default.  $I$  is an *indicator* function as defined in the following equation:

$$I(F) = \begin{cases} 1 & \text{if } F \text{ occurs} \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

Chemicals with evidently small first-order differences are screened out via the rules and form the negative pattern group. All chemicals are then separated into two groups, i.e. negative pattern group and positive pattern group.

### 4.3 Step 2: Curve by curve correlation-based clustering

In order to achieve the goal of exploring a pattern recognition strategy in which TCRPs can be grouped according to their similarity, the correlation coefficient is used as a similarity metric between TCRPs of chemicals.

The Pearson correlation coefficient is given in Eq. (4.4).

$$R(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.4)$$

where  $\mathbf{x} = [x_1, x_2, \dots, x_n]$  and  $\mathbf{y} = [y_1, y_2, \dots, y_n]$  are time series vectors, and  $\bar{x}$  and  $\bar{y}$  represent means of time series  $\mathbf{x}$  and  $\mathbf{y}$ , respectively.

### 4.3.1 Correlation analysis between TCRPs

The curve by curve correlation-based clustering is now explained in detail. Let chemical  $p$  and chemical  $q$  be classified in the same category based on previous categorization. An arbitrary TCRP  $u$  in chemical  $p$  is compared with all TCRPs in chemical  $q$  using the correlation coefficient. If one of the correlation coefficients is larger than the threshold set a priori, there exists a similar counterpart of TCRP  $u$ , denoted as  $v$ , in chemical  $q$ . The search for similar counterpart in chemical  $q$  for each TCRP in chemical  $p$  is carried out. Subsequently, starting from chemical  $q$ , the similar counterpart TCRPs in chemical  $p$  are found. The need to repeat the exercise of finding similar counterpart TCRPs for chemical  $q$  in chemical  $p$  is important for chemical  $p$ , as all TCRPs can find similar counterpart TCRPs in chemical  $q$  but the converse may not be true.

**Notation 1** Let  $C_{p \rightarrow q}^1(u, v)$  denote the correlation coefficient between TCRP  $u$  in chemical  $p$  and TCRP  $v$  in chemical  $q$ . Let  $u_1, u_2, \dots, u_m$  and  $v_1, v_2, \dots, v_n$  be eligible TCRPs in chemicals  $p$  and  $q$  respectively. Then the matrix of correlation coefficients is constructed as follows:

$$C_{p \rightarrow q}^1 = [C_{p \rightarrow q}^1(u_i, v_j)]_{i=1, \dots, m; j=1, \dots, n} \quad (4.5)$$

Here the size of  $C_{p \rightarrow q}^1$  is  $m \times n$ . Similarly, one can construct  $C_{q \rightarrow p}^1$  of size  $n \times m$ .

The correlation analysis described above cannot by itself guarantee that the TCRPs in both the chemicals are similar in shape and therefore the chemicals show the same pattern. This is because correlation coefficient between TCRPs alone cannot completely capture the similarity in shape of curves. Correlation is a measure to describe the linear relationship between two random variables. A high degree similarity can have a large correlation coefficient, but a large correlation coefficient cannot guarantee a high degree similarity in shape. This is illustrated in the following example.

**Example 1** In this example we illustrate the inadequacy of correlation coefficient to distinguish between *different* TCRPs. TCRPs of Etoposide and Mitoxantrone at concentrations levels 0.2mM and 5.56 $\mu$ M are presented in Figure 4.4(a). Further, TCRPs of the same chemicals at concentration levels 0.2mM and 68.59nM respectively are presented in Figure 4.4(b). For ease of notation,

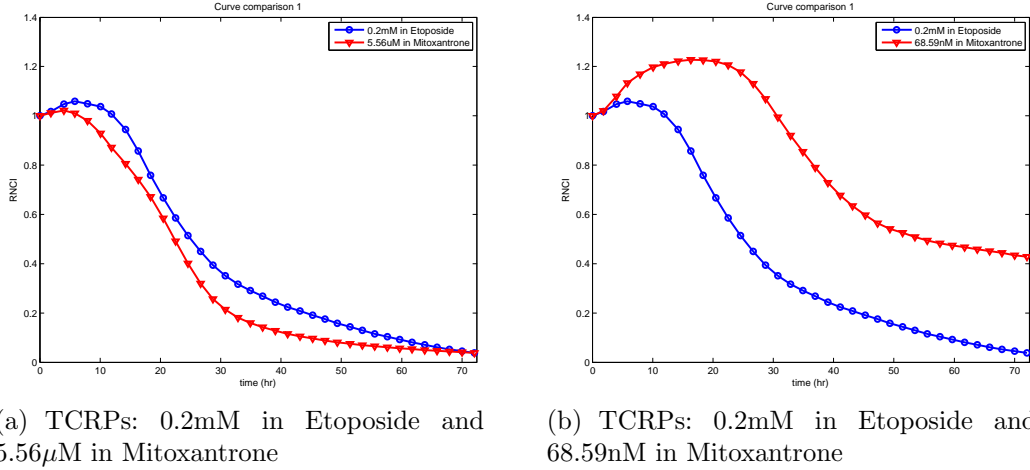


Figure 4.4: High correlation coefficient indicating similar trend

let  $p$  denote chemical Etoposide and  $q$  denote chemical Mitoxantrone. Let  $u_1$  and  $v_1$  denote TCRPs for chemicals  $p$  and  $q$  at concentration levels 0.2mM and 5.56 $\mu$ M, respectively. Let  $u_2$  and  $v_2$  denote TCRPs for chemicals  $p$  and  $q$  at concentration levels 0.2mM and 68.59nM, respectively. It can then be noted that  $C_{p \rightarrow q}^1(u_1, v_1) = 0.9920$  and  $C_{p \rightarrow q}^1(u_2, v_2) = 0.8689$  which indicates a high

correlation coefficient. However, the shapes of the TCRPs in Figure 4.4(b) are different because the peak times are different. The illustrative examples underline the need for supportive metric to correlation coefficient to measure similarity between TCRP shapes.

### 4.3.2 Correlation analysis between first-order differences of TCRPs

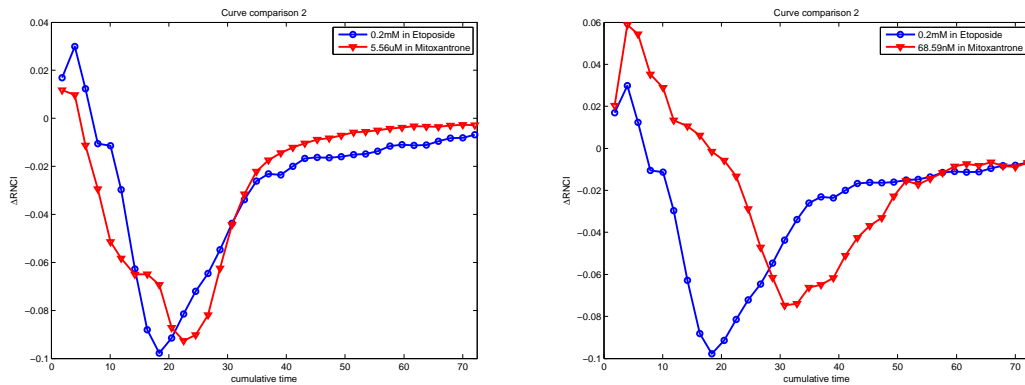
In order to circumvent the shortcoming of using correlation coefficient alone to measure similarity, a concept of correlation coefficient between the first-order differences of TCRPs is introduced (as defined in Eq. (4.1)). This analysis includes the information about the rate of change in TCRPs in addition to the trend information provided by correlation analysis of TCRPs. In other words, the correlation on TCRPs is not sensitive to time shift (see Figure 4.4) whereas the correlation between the first-order differences of TCRPs captures this phenomenon.

**Notation 2** For TCRPs  $u$  and  $v$  in chemicals  $p$  and  $q$ , respectively, define the first-order differences of TCRPs, denoted as  $\Delta u$  and  $\Delta v$ , respectively, according to Eq. (4.1). Let  $C_{p \rightarrow q}^2(u, v)$  denote the correlation coefficient between  $\Delta u$  and  $\Delta v$ . Let  $u_1, u_2, \dots, u_m$  and  $v_1, v_2, \dots, v_n$  be eligible TCRPs in chemicals  $p$  and  $q$ , respectively. The matrix of correlation coefficients among the first-order difference of TCRPs is then constructed as follows:

$$C_{p \rightarrow q}^2 = [C_{p \rightarrow q}^2(u_i, v_j)]_{i=1, \dots, m; j=1, \dots, n} \quad (4.6)$$

Here the size of  $C_{p \rightarrow q}^2$  is  $m \times n$ . Similarly, one can construct  $C_{q \rightarrow p}^2$  of size  $n \times m$ .

**Example 2** This example is a continuation of Example 1. Figure 4.5 presents the first-order differences of TCRPs of the TCRPs presented in Figure 4.4. In the same notation as in Example 1,  $C_{p \rightarrow q}^2(u_1, v_1) = 0.8943$  and  $C_{p \rightarrow q}^2(u_2, v_2) =$



(a) First-order differences of TCRPs: 0.2mM in Etoposide and 5.56 $\mu$ M in Mitoxantrone

(b) First-order differences of TCRPs: 0.2mM in Etoposide and 68.59nM in Mitoxantrone

Figure 4.5: High correlation in TCRP but not in the first-order differences of TCRPs

0.2819. A high correlation in the case of Figure 4.5(a) ( $C_{p \rightarrow q}^2(u_1, v_1)$ ) is an indication of the similarity not only in trend but also in time shifts. On the other hand, a low correlation in the case of Figure 4.5(b) ( $C_{p \rightarrow q}^2(u_2, v_2)$ ) indicates that though the trend in the TCRPs is similar, the trend in time shifts does not match. These examples show that using correlation coefficient in TCRPs as well as first-order differences of TCRPs serves as a more reasonable metric for similarity in shapes.

### 4.3.3 Clustering algorithm based on curve by curve correlation analysis

After finalizing the similarity metric (correlation coefficient in TCRPs and first-order differences of TCRPs) to identify chemicals with the same pattern, the clustering algorithm based on this similarity metric is now discussed.

1. Start with two chemicals  $p$  and  $q$  with eligible TCRPs  $u_1, \dots, u_m$  and  $v_1, \dots, v_n$ , respectively.
2. Construct the correlation matrix among TCRPs  $C_{p \rightarrow q}^1$  and the correlation

matrix among the first-order differences of TCRPs  $C_{p \rightarrow q}^2$  as introduced in Notation 1 and 2.

3. Compare correlation coefficients of TCRPs and first-order differences of TCRPs in two correlation matrices with predefined thresholds  $\tau_1$  and  $\tau_2$  respectively. If the values of two matrices at the same entry are larger than  $\tau_1$  and  $\tau_2$  respectively, a parameter called score for that entry is set to 1; otherwise, the score for that entry is 0.
4. Repeat steps 2 and 3 in the reverse direction. That is, compute  $C_{q \rightarrow p}^1$  and  $C_{q \rightarrow p}^2$ , compare the entries of two matrices with predefined thresholds and record the score.
5. Decide whether the two chemicals  $p$  and  $q$  exhibit similarity according to the score.

Before giving the detailed description of the algorithm, the importance of Step 4 is emphasised in the following example.

**Example 3** Consider two chemicals: 5-FU and Etoposide. For ease of notation, let  $p$  denote 5-FU and  $q$  denote Etoposide. Both  $p$  and  $q$  have 5 eligible TCRPs. The TCRPs are shown in Figure 4.6.

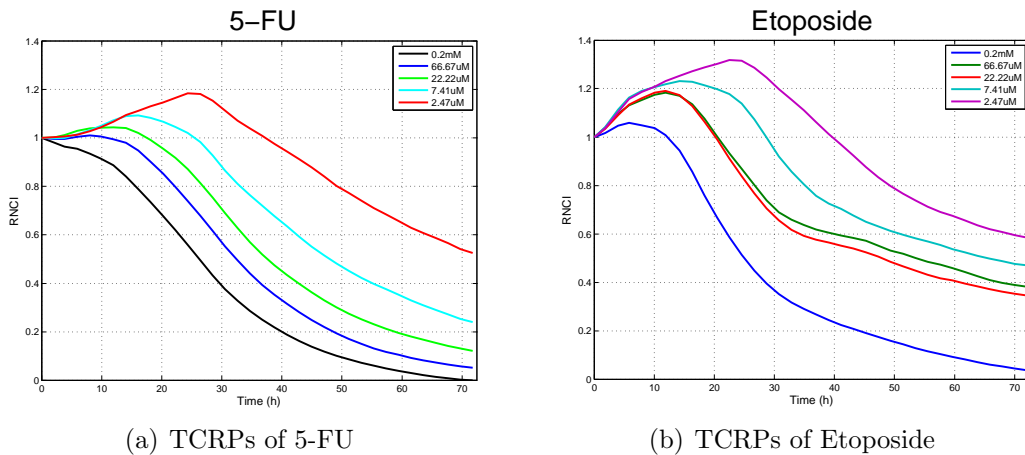


Figure 4.6: TCRPs of 5-FU and Etoposide



According to the clustering algorithm above, we calculate  $C_{p \rightarrow q}^1$  and  $C_{p \rightarrow q}^2$  as well as  $C_{q \rightarrow p}^1$  and  $C_{q \rightarrow p}^2$ .

$$C_{p \rightarrow q}^1 = \begin{pmatrix} 0.993 & 0.984 & 0.984 & 0.936 & 0.785 \\ 0.975 & 0.984 & 0.986 & 0.969 & 0.857 \\ 0.948 & 0.972 & 0.975 & 0.986 & 0.907 \\ 0.884 & 0.930 & 0.933 & 0.984 & 0.962 \\ \mathbf{0.681} & 0.758 & 0.763 & 0.885 & 0.983 \end{pmatrix} \quad C_{p \rightarrow q}^2 = \begin{pmatrix} 0.886 & 0.806 & 0.822 & 0.610 & \mathbf{0.167} \\ 0.740 & 0.829 & 0.849 & 0.811 & 0.499 \\ 0.526 & 0.718 & 0.754 & 0.933 & 0.759 \\ \mathbf{0.139} & 0.459 & 0.499 & 0.843 & 0.883 \\ \mathbf{-0.367} & \mathbf{-0.036} & \mathbf{-0.007} & 0.518 & 0.839 \end{pmatrix} \quad (4.7)$$

$$C_{q \rightarrow p}^1 = \begin{pmatrix} 0.993 & 0.975 & 0.948 & 0.884 & \mathbf{0.681} \\ 0.984 & 0.984 & 0.972 & 0.930 & 0.758 \\ 0.984 & 0.986 & 0.975 & 0.933 & 0.763 \\ 0.936 & 0.969 & 0.986 & 0.984 & 0.885 \\ 0.785 & 0.857 & 0.907 & 0.962 & 0.983 \end{pmatrix} \quad C_{q \rightarrow p}^2 = \begin{pmatrix} 0.886 & 0.740 & 0.526 & \mathbf{0.139} & \mathbf{-0.367} \\ 0.806 & 0.829 & 0.718 & 0.459 & \mathbf{-0.036} \\ 0.822 & 0.849 & 0.754 & 0.499 & \mathbf{-0.007} \\ 0.610 & 0.811 & 0.933 & 0.843 & 0.518 \\ \mathbf{0.167} & 0.499 & 0.759 & 0.883 & 0.839 \end{pmatrix} \quad (4.8)$$

From Figure 4.6, we notice that the TCRPs of both 5-FU and Etoposide are similar. In each row of  $C_{p \rightarrow q}^1$  and  $C_{p \rightarrow q}^2$ , there exist at least one entry whose values are larger than thresholds  $\tau_1, \tau_2$  (Here,  $\tau_1 = 0.7$ ,  $\tau_2 = 0.4$ ). It means that for each TCRP in 5-FU, there is at least one similar TCRP in Etoposide. The situation is the same in reverse. Therefore, these two chemicals are grouped because they are mutually similar. Each TCRP in chemical  $p$  has its counterpart in chemical  $q$  and vice versa.

**Example 4** Another example is as follows. Two chemicals considered are Monesin and CCCP from HepG2 cell line. For ease of notation, let  $p$  denote Monesin and  $q$  denote CCCP. There are 3 eligible TCRPs in Monesin and 4 eligible TCRPs in CCCP respectively. The TCRPs are shown in Figure 4.7.

According to the clustering algorithm discussed, we calculate  $C_{p \rightarrow q}^1$  and  $C_{p \rightarrow q}^2$  as well as  $C_{q \rightarrow p}^1$  and  $C_{q \rightarrow p}^2$ .

$$C_{p \rightarrow q}^1 = \begin{pmatrix} 0.953 & 0.999 & 0.992 & 0.838 \\ 0.896 & 0.983 & 0.963 & 0.790 \\ \mathbf{0.564} & 0.781 & 0.712 & \mathbf{0.323} \end{pmatrix} \quad C_{p \rightarrow q}^2 = \begin{pmatrix} 0.768 & 0.957 & 0.883 & 0.744 \\ \mathbf{0.266} & 0.633 & 0.462 & \mathbf{0.289} \\ \mathbf{-0.809} & \mathbf{-0.503} & \mathbf{-0.647} & \mathbf{-0.593} \end{pmatrix} \quad (4.9)$$

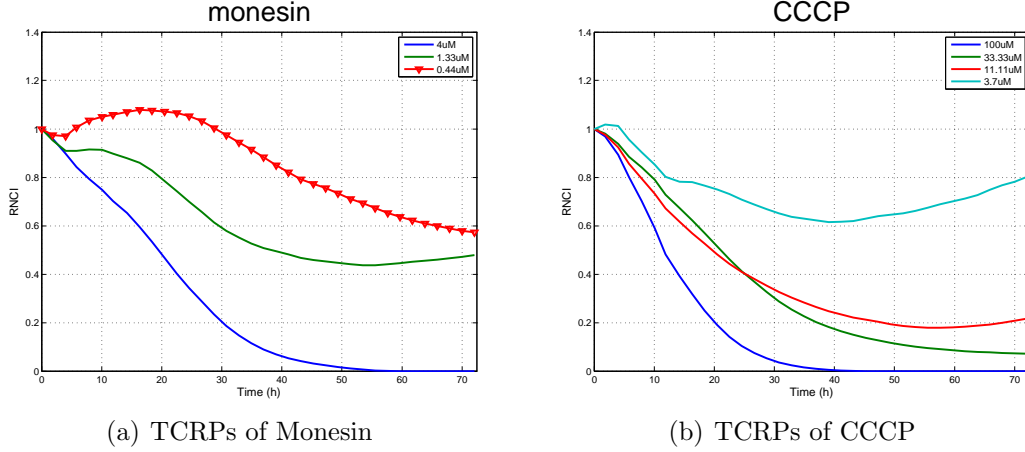


Figure 4.7: TCRPs of Monesin and CCCP

$$C_{q \rightarrow p}^1 = \begin{pmatrix} 0.953 & 0.896 & \mathbf{0.564} \\ 0.999 & 0.983 & 0.781 \\ 0.992 & 0.963 & 0.712 \\ 0.838 & 0.790 & \mathbf{0.323} \end{pmatrix} \quad C_{q \rightarrow p}^2 = \begin{pmatrix} 0.768 & \mathbf{0.266} & -\mathbf{0.809} \\ 0.957 & 0.633 & -\mathbf{0.503} \\ 0.883 & 0.462 & -\mathbf{0.647} \\ 0.744 & \mathbf{0.289} & -\mathbf{0.593} \end{pmatrix} \quad (4.10)$$

From Figure 4.7, we notice that except for the TCRP marked with triangles (down) in Monesin, the others can find their matching counterpart TCRPs in CCCP while all the TCRPs in CCCP have their counterpart TCRPs in Monesin. This phenomenon can be reflected from the correlation matrices. We can see that the first two rows of  $C_{p \rightarrow q}^1$  and  $C_{p \rightarrow q}^2$  are larger than the default thresholds  $\tau_1 = 0.7$ ,  $\tau_2 = 0.4$ , respectively while the highlighted entries in the third row do not satisfy the conditions. Specifically, the third row of  $C_{p \rightarrow q}^2$  are all negative. So, for the TCRP with the third highest concentration in Monesin, there is no counterpart TCRP in CCCP, which is denoted as a *mismatch*. Conversely, in each row of  $C_{q \rightarrow p}^1$  and  $C_{q \rightarrow p}^2$ , there exist at least one pair of elements which are larger than the thresholds defined. Therefore, CCCP is *similar* to Monesin while Monesin is not so *similar* to CCCP. In order to cluster chemicals in an acceptable scope, a tuning parameter  $M$  is set to control the number of mismatch. In this study, at most 1 mismatch is tolerated. If the number of mismatch is larger than 1, the two compared chemicals will not be

clustered.

The detailed algorithm is presented in Algorithm 1.

---

**Algorithm 1** Curve by curve correlation-based clustering

---

**Initialization:**

Set  $\text{Index} \neq \emptyset$ ,  $k_2 = 1$ ;  $t_1, t_2$ ,  $M$ ,  $\tau_1, \tau_2$  (Default values:  $\tau_1 = 0.7, \tau_2 = 0.4$ )

**Iteration:**

```

1: while  $\text{Index} \neq \emptyset$  do
2:    $p \leftarrow \text{Index}(1)$ ;  $C_k \leftarrow \text{Index}(1)$ ; removing  $p^{\text{th}}$  chemical's label;
3:   for each  $l \in [2, \dots, \text{length}(\text{Index})]$  do
4:      $q \leftarrow \text{Index}(l)$ ;
5:     Calculating the correlation matrix  $C_{p \rightarrow q}^1 = [C_{p \rightarrow q}^1(u_i, v_j)]_{i=1, \dots, m; j=1, \dots, n}$ 
       for  $RNCI$  and the correlation matrix  $C_{p \rightarrow q}^2 = [C_{p \rightarrow q}^2(\Delta u_i, \Delta v_j)]_{i=1, \dots, m; j=1, \dots, n}$ 
       for the deviation of  $RNCI$  from  $p^{\text{th}}$  to  $q^{\text{th}}$  chemical between  $t_1$  and  $t_2$  separately. Similarly,  $C_{q \rightarrow p}^1$  and
        $C_{q \rightarrow p}^2$  are generated from  $q^{\text{th}}$  to  $p^{\text{th}}$  chemical. ( $m, n$ : # of selected
       curves in  $p^{\text{th}}$  and  $q^{\text{th}}$  chemical);  $m \times 1$  vector  $flag1 = \mathbf{0}$ ,  $n \times 1$  vector
        $flag2 = \mathbf{0}$ .
6:     for  $i \in [1, \dots, m]$  do
7:       for  $j \in [1, \dots, n]$  do
8:         if  $C_{p \rightarrow q}^1(u_i, v_j) > \tau_1, C_{p \rightarrow q}^2(\Delta u_i, \Delta v_j) > \tau_2$  then
9:            $flag1(i) = 1$  (TCRP  $v_j$  is matched with TCRP  $u_i$ .)
10:        end if
11:      end for
12:    end for
13:    for  $j \in [1, \dots, n]$  do
14:      for  $i \in [1, \dots, m]$  do
15:        if  $C_{q \rightarrow p}^1(v_j, u_i) > \tau_1, C_{q \rightarrow p}^2(\Delta v_j, \Delta u_i) > \tau_2$  then
16:           $flag2(j) = 1$  (TCRP  $u_i$  is matched with TCRP  $v_j$ .)
17:        end if
18:      end for
19:    end for
20:    if  $m - \sum_{i=1}^m flag1(i) \leq M, n - \sum_{j=1}^n flag2(j) \leq M$  then
21:       $C_k \leftarrow C_k \cup \text{Index}(l)$ ;
22:    end if
23:  end for
24:  Cluster  $\#k_2 \leftarrow C_k$ ;  $k_2 \leftarrow k_2 + 1$ ;
25: end while

```

**Output:**

clusters and cluster #;

---

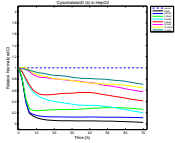
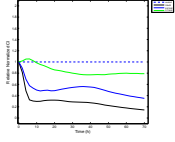
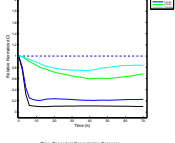
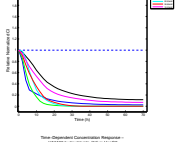
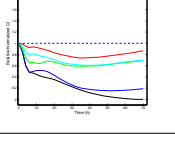
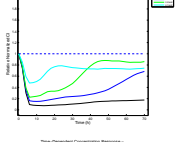
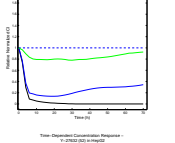
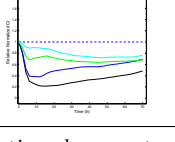
## 4.4 Results and discussions

### 4.4.1 Results

The proposed algorithm is demonstrated using HepG2 cell line with seven batches of experiments. All chemicals whose TCRPs are similar to the negative control lines are screened out, which is introduced in Data Preprocessing. Those chemicals are categorized into “Unclassified Group” because they do not show apparent patterns. All eligible chemicals are then fed into the proposed categorization step. Following the categorization, chemicals in each category are analyzed via curve by curve correlation analysis in order to specify the similarity within each category. The curve by curve correlation analysis considers the cell reactions within the whole time range (all 72 hours from the time when chemicals are added).

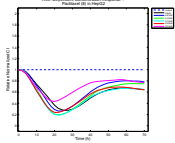
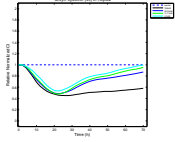
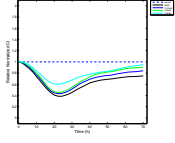
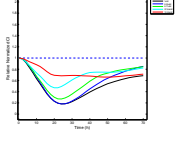
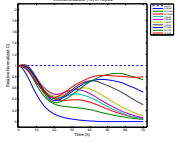
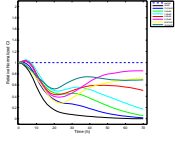
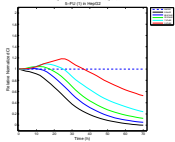
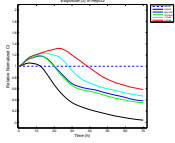
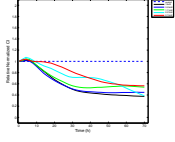
The classification results on HepG2 cell line are displayed in Table 5.1. Specifically, the classification results are illustrated in both table and the eligible relative TCRP figures. Chemicals with similar shapes of TCRPs are clustered together into one group based on the results. Some empirical parameters, e.g. correlation coefficient thresholds  $\tau_1, \tau_2$ , can be fine tuned because  $\tau_1$  and  $\tau_2$  determine the number of clusters in each category. Larger  $\tau_1, \tau_2$  will result in a larger cluster number.

Table 4.1: Results of Hierarchical correlation based classification (cell line: HepG2. 47 of 65 chemicals are eligible.)

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
1	4	Cytochalasin D	DMSO	20 $\mu$ M – 0.339nM	
	5	Cytochalasin B	DMSO	20 $\mu$ M – 0.339nM	
	31	Latrunculin A	EtOH	2 $\mu$ M – 0.03nM	
	55	FAKInhibitor14	H <sub>2</sub> O	2500 $\mu$ M – 42.34nM	
	64	HA1100 hydrochloride	H <sub>2</sub> O	1000 $\mu$ M – 16.94nM	
2	6	Latrunculin B	DMSO	20 $\mu$ M – 0.339nM	
	51	W7 HCl	DMSO	200 $\mu$ M – 3.39nM	
	52	Y-27632	DMSO	188 $\mu$ M – 3.18nM	

continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
3	8	Paclitaxel	DMSO	20 $\mu$ M – 0.339nM	
	49	S-tritytl-Cysteine	DMSO	100 $\mu$ M – 1.69nM	
	50	Dimethylenastron	DMSO	40 $\mu$ M – 0.68nM	
	59	Docetaxel	DMSO	1 $\mu$ M – 0.02nM	
4	15	Vincristine Sulfate	H <sub>2</sub> O	250 $\mu$ M – 4.23nM	
	61	Vinblastine Sulfate	H <sub>2</sub> O	40 $\mu$ M – 0.68nM	
5	1	5-fluorouracil (5-FU)	DMSO	200 $\mu$ M – 3.39nM	
	2	Etoposide phosphate	DMSO	200 $\mu$ M – 3.39nM	
	11	Anisomycin	H <sub>2</sub> O	10 $\mu$ M – 0.17nM	

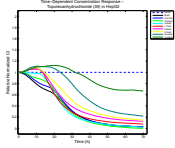
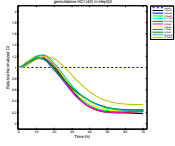
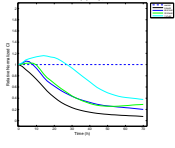
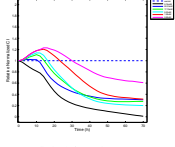
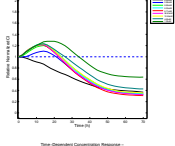
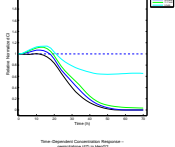
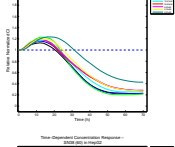
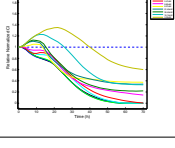
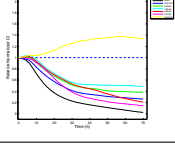
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	12	Clofarabine (CLOF)	H <sub>2</sub> O	25μM – 0.42nM	
	13	Hydroxyurea (HU)	H <sub>2</sub> O	10mM – 169nM	
	19	Exo 1	DMSO	300μM – 5.08nM	
	20	Monensin	DMSO	4μM – 0.068nM	
	25	Thapsigargin	DMSO	2μM – 0.0339nM	
	28	Cyclosporin A	DMSO	100μM – 1.69nM	
	33	SAHA	DMSO	151μM – 2.56nM	
	34	(S)-HDAC-42	DMSO	128μM – 2.17nM	
	36	Mitomycin C	DMSO	200μM – 3.39nM	

continued on next page

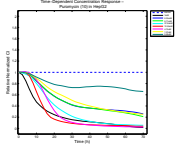
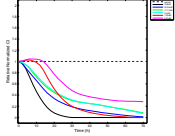
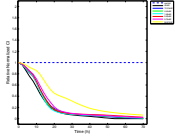
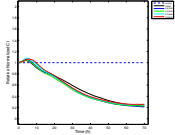
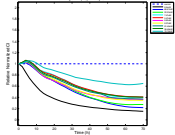
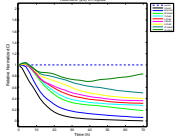
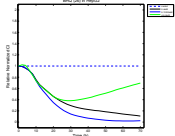
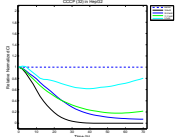
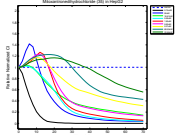
continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	39	Topotecan	DMSO	95 $\mu$ M – 1.61nM	
	40	2'-Deoxy-2'	H <sub>2</sub> O	1650 $\mu$ M – 27.94nM	
	42	Merbarone	DMSO	200 $\mu$ M – 3.39nM	
	43	Irinotecan (CPT-11)	DMSO	160 $\mu$ M – 2.71nM	
	44	Cytosine	H <sub>2</sub> O	8950 $\mu$ M – 151.57nM	
	46	Benzo[a]pyrene	DMSO	100 $\mu$ M – 1.69nM	
	47	Gemcitabine	H <sub>2</sub> O	2 $\mu$ M – 0.03nM	
	60	SN-38	DMSO	200 $\mu$ M – 3.39nM	
6	7	Emetine	H <sub>2</sub> O	50 $\mu$ M – 0.847nM	

continued on next page

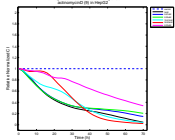
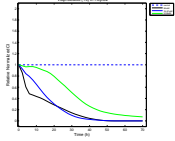
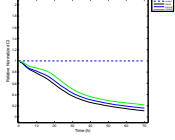


continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	10	Puromycin	H <sub>2</sub> O	1000μM – 17nM	
	16	Doxorubicin (DOX)	H <sub>2</sub> O	100μM – 1.69nM	
	17	Brefeldin A (BEF)	DMSO	40μM – 0.68nM	
	22	Oligomycin	DMSO	20μM – 0.339nM	
	23	Antimycin A	EtOH	200μM – 3.387nM	
	24	Rotenone	DMSO	200μM – 3.387nM	
	26	BHQ	DMSO	400μM – 7nM	
	32	CCCP	DMSO	100μM – 1.69nM	
	35	Mitoxantrone	DMSO	150μM – 2.54nM	

continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
7	9	Actinomycin D	DMSO	2 $\mu$ M – 0.0339nM	
	14	Valproic acid	H <sub>2</sub> O	50mM – 847nM	
	18	Leptomycin B (LMB)	EtOH	20nM – 0.000339nM	

#### 4.4.2 Discussion

The proposed approach can distinguish the TCRPs of chemicals and give a separation of chemicals within each category, but there are additional improvements that need to be done.

An example is used here to indicate why the transitivity is not applicable to correlation coefficients as a measure of similarity. Specifically, three chemicals A, B and C are compared among each other. Because the correlation test is pairwise, based on a particular chemical A, (A,B) and (A,C) pairs are subject to the proposed method to decide if B and C can be absorbed into the same cluster as A. However, all the pairwise tests are carried out with A involved. The high similarity between A and B, A and C cannot guarantee a high similarity between B and C.

Also, this correlation-based clustering method does not update the centers iteratively. Each cluster is constructed based on the particular element in each round of correlation test. When a new data object associated with its chemical is updated, the non-uniform dimensionality of data makes the calculation of centers infeasible and difficult. Centers can be updated if a “distance” measure between chemicals is defined properly. So even though the correlation based

approach has shown promising results, further improvement through additional approaches can be considered as none single approach can be perfect.

## 4.5 Conclusion

In this chapter, a hierarchical correlation based classification approach including categorization and curve by curve correlation-based clustering was developed and implemented on TCRPs of chemicals. The method uses the kinetic information, e.g. growth and killing rate, in the profiles to distinguish chemicals. By setting proper parameter values and thresholds, the designed classification method was able to distinguish chemicals based on the similarity of TCRPs via correlation coefficient. However, empirical tuning on parameters is inevitable in this approach. In order to solve this problem and make the classification more objective, a new approach using PCA & FDA will be established in the next chapter.

## Chapter 5

# PCA & FDA based hierarchical classification

In this chapter, a systematic model-based hierarchical clustering approach using principal component analysis (PCA) and functional data analysis (FDA) is proposed and elaborated in order to reduce the number of tuning parameters. First, we present two statistical techniques, PCA and FDA, including their theories and application in feature extraction and large data dimension reduction. A model-based hierarchical clustering algorithm is then introduced. Our classification problem is finally addressed by the proposed approach and the results are presented in tables.

### 5.1 Introduction

As introduced in previous chapters, the data we encounter are of different dimensions after irrelevant and uninformative response profiles of chemicals are deleted. Chemicals represented by different numbers of TCRPs make the comparison of shape similarity difficult. The strategy of majority-voting within each time interval indicated in Chapter 3 can fuse the information of eligible TCRPs into one synthesized representative. However, the key deficiency of this

approach lies in that the number of profiles for some chemicals is largely reduced to 3 after the selection of TCRPs. The voting results are not precise for such a small sample number, which leads to a conflict in the identification of feature vectors. Therefore, to address this issue, PCA is utilized as a more reliable tool to extract valid feature vectors for further analysis. The experiments using different levels of concentration are considered as a realization of multiple observations. The advantage of PCA lies in its capability of transforming data with nonuniform dimensionality into uniform volumes. In other words, the problem of identifying the profile similarity of chemicals which are denoted with flexibly dimensionalities is then solved with principal component (PC) scores in the principal component space, while a set of observations of possibly correlated variables are converted into a set of values of linearly uncorrelated variables simultaneously.

This chapter is organized as follows. First, the principles of how PCA works in feature extraction and dimension reduction is introduced in Section 5.2. In Section 5.3, the principles of FDA are exhibited in order to show its advantage in smoothing, denoising and compression. In Section 5.4, an agglomerative model-based hierarchical clustering algorithm is explained in detail. By applying it to the *CI* data in Section 5.5, this proposed classification approach shows a good discrimination based on profile similarity.

## 5.2 Principal component analysis (PCA)

### 5.2.1 Overview

As an effective means of compressing data with high dimension, principal component analysis (PCA) is a mathematical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables using an orthogonal transformation [21]. Those uncorrelated variables are called principal components. The main advantage of PCA lies in

the fact that it can retain most of the variation in the principal component space. This technique is widely used for applications such as dimensionality reduction, feature extraction, prediction for incomplete data and overview of any data table etc. [23]

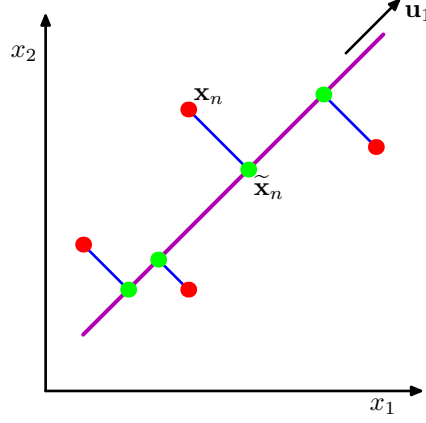


Figure 5.1: Orthogonal projection of data in the original space onto the principal space

The significance of PCA is to project the data orthogonally onto a lower dimensional linear space, termed as the *principal subspace*, such that the variance of the projected data is maximized. In Figure 5.1, we can observe the process of projection from data (red dots) onto the principal component space with a lower dimension (green dots in line marked in magenta). The orthogonal projection of the red dots maximizes the variance of the green dots.

### 5.2.2 Problem formulation

In this subsection, PCA is elaborated with mathematical equations from [21]. Consider a data set of observations  $\{\mathbf{x}_i\}$ , where  $\mathbf{x}_i$  is a vector whose dimension is  $d$  and  $i = 1, \dots, N$ . The goal is to project all  $\mathbf{x}_i$ 's onto a smaller dimensional space  $m$  ( $m < d$ ) so as to maximize the variance of the projected data.

To simplify this problem, we first consider the projection onto a one-dimensional space ( $m = 1$ ). The direction of the space is then defined using a  $d$ -dimensional vector  $\mathbf{u}_1$  which is chosen to be a unit vector ( $\mathbf{u}_1^T \mathbf{u}_1 = 1$ ). Each data point  $\mathbf{x}_i$

is projected to a scalar value using a linear transformation  $\mathbf{u}_1^T \mathbf{x}_i$ . The mean of the projected data is  $\mathbf{u}_1^T \bar{\mathbf{x}}$  where  $\bar{\mathbf{x}}$  is the sample mean of the data.

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \quad (5.1)$$

The variance of the projected data is

$$\frac{1}{N} \sum_{i=1}^N \{\mathbf{u}_1^T \mathbf{x}_i - \mathbf{u}_1^T \bar{\mathbf{x}}\}^2 = \mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 \quad (5.2)$$

where  $\mathbf{S}$  is the data covariance matrix.

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \quad (5.3)$$

To maximize the projected variance  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$  with respect to  $\mathbf{u}_1$ , we introduce the Lagrange multiplier which is denoted by  $\lambda$ . As a meaningful optimization problem, the constrained optimization problem is then transferred into an unconstrained problem.

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 + \lambda(1 - \mathbf{u}_1^T \mathbf{u}_1) \quad (5.4)$$

To maximize the objective function, we set the derivative with respect to  $\mathbf{u}_1$  equal to zero. A stationary point exists when

$$\mathbf{S} \mathbf{u}_1 = \lambda \mathbf{u}_1 \quad (5.5)$$

Eq. (5.5) indicates that  $\mathbf{u}_1$  is an eigenvector of  $\mathbf{S}$ . By left multiplying  $\mathbf{u}_1^T$  and using  $\mathbf{u}_1^T \mathbf{u}_1 = 1$ , the variance of the projected data is

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda \quad (5.6)$$

Therefore, the variance of the projected data will be maximized when  $\mathbf{u}_1$  is equal to the eigenvector with the largest eigenvalue  $\lambda$ . The eigenvector  $\mathbf{u}_1$  is

the first principal component.

Similarly, for an  $m$  dimensional space, the projection for which the variance of the projected data is maximized can now be defined via a matrix  $\mathbf{U}$  whose rows are the transposed  $m$  eigenvectors of the data covariance matrix  $\mathbf{S}$ . The  $m$  eigenvectors are associated with the  $m$  largest eigenvalues, i.e.  $\lambda_1, \dots, \lambda_m$ .

$$\tilde{\mathbf{x}}_i = \mathbf{U}\mathbf{x}_i = \begin{pmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \dots \\ \mathbf{u}_m^T \end{pmatrix} \mathbf{x}_i \quad (5.7)$$

where  $\tilde{\mathbf{x}}_i$  is an  $m$  dimensional projected variable. The data dimension is compressed while most of the variance in the projected data is retained.

PCA includes evaluating the mean,  $\bar{\mathbf{x}}$  and the data covariance matrix,  $\mathbf{S}$  and subsequently finding the  $m$  eigenvectors of  $\mathbf{S}$  corresponding to the  $m$  largest eigenvalues that become the key component in PCA. There are different algorithms that can find the eigenvalues of the data covariance matrix as well as the corresponding eigenvectors. Details about the efficiency of the eigenvalue algorithm can be found in Golub and Van Loan [24]. Details about how PCA is formulated in minimum-error scenario and its applications can be accessed via Bishop and Nasrabadi [21].

### 5.3 Functional data analysis (FDA)

Functional data analysis (FDA) is a statistical tool to analyze data from a curve perspective. It assumes the functionality behind time series data. Even though each measurement is a finite set of numbers, their values can reflect a smooth variation that can be assessed. Besides, this data analysis technique can also be used for data sets which are not functional [25]. It is proved that functional data analysis has been applied in many areas such as modeling of



gene expression data [26], life course data in criminology [27] etc.

To build functions, two steps are required. First, a set of *basis functions* are specified. Second, a vector, or array of coefficients are generated to define the functions as a linear combination of these basis functions. As a way to compress data dimensionality, FDA can describe a curve successfully with fewer parameters and coefficients while retaining as much information about the profile shapes as possible.

Instead of treating observed numbers or values in each sampling interval as the units of data in traditional statistical methods [26], functions defined on some interval are considered as the data units in FDA.

We first build basis functions. The main reason to build basis functions is to make the description of curves flexible. We cannot specify the characteristics of curves certainly. Meanwhile, by building up basis functions we can save the burden of computation.

### 5.3.1 Basis function selection

A set of functional building blocks  $\phi_k, k = 1, \dots, K$  which can be combined linearly are called basis functions. Therefore, a function  $x(t)$  can be defined as a linear combination of basis functions written as

$$x(t) = \sum_{k=1}^K c_k \phi_k(t) = \mathbf{c}' \boldsymbol{\phi}(t) \quad (5.8)$$

Eq. (5.8) is called basis function expansion. Parameters  $c_1, c_2, \dots, c_K$  are coefficients of the expansion.  $\mathbf{c}$  is used to denote the vector of all coefficients  $c_k$  and  $\boldsymbol{\phi}$  denotes the vector of all basis functions  $\phi_k(t)$ .

To choose proper basis functions, some prior knowledge about the profile tendency is required. Since the Cell Index data considered are non-periodic, B-spline basis functions are chosen. Spline basis functions are piecewise polynomial and constructed over subintervals divided by boundary points called

*break points*. *Degree* and *order* define the power and the argument number of the polynomials within the subinterval. The degree of the spline functions is fixed within any subinterval while the order can be different from one interval to another. In our study, we assume the order of polynomials is the same for each subinterval. Every *knot* has the same value as a break point, but multiple knots can be located at certain break points.

Therefore, spline systems are defined by three main elements:

1. The break points which define the subintervals,
2. The degree or order of the polynomials,
3. The sequence of knots.

The number of basis functions  $K$  is determined by

$$n_{basis} = n_{order} + n_{interior\ knots} \quad (5.9)$$

where  $n_{order}$  denotes the order and  $n_{interior\ knots}$  denotes the number of interior knots except for the beginning and the end boundary points.

In MATLAB<sup>®</sup>, the equation can be expressed as

$$n_{basis} = n_{order} + length(params) - 2 \quad (5.10)$$

where  $length(params)$  indicates the number of breaks. Order 4 ( $n_{order} = 4$ ) is frequently chosen, which implies piece-wise cubic polynomials. So the equation is written as

$$n_{basis} = length(params) + 2 \quad (5.11)$$

### 5.3.2 Computing coefficients

In this section, we discuss the method to compute the coefficients with the basis functions to obtain an optimal fit to data.

Two strategies can be used to compute the coefficients. The simpler one uses regression analysis to estimate the coefficients. Another one introduces the penalization on the “roughness” of functions.

### Regression analysis

Regression analysis can effectively estimate the coefficients given the basis functions. It fits the data by minimizing the sum of squared errors ( $SSE$ ):

$$SSE = \sum_j^n [y_j - x(t_j)]^2 \quad (5.12)$$

where  $y_j$ 's are the real measurements, and  $x(t_j)$ 's are values of the fitted function at  $t_j$ . According to the basis function expansion in Eq. (5.8), Eq. (5.12) can be written as

$$SSE(\mathbf{y}, \mathbf{c}) = \sum_j^n [y_j - \boldsymbol{\phi}(t_j)' \mathbf{c}]^2 \quad (5.13)$$

where  $\mathbf{y} = [y_1, \dots, y_n]^T$ . The model is then formulated with the error:

$$y_j = x(t_j) + \epsilon_j = \boldsymbol{\phi}(t_j)' \mathbf{c} + \epsilon_j \quad (5.14)$$

where the errors  $\epsilon_j$  are statistically independent and have a Gaussian distribution with mean 0 and a constant variance. Eq. (5.14) is a standard regression model. If  $n$  values are fit and  $\boldsymbol{\epsilon}$  is the residual vector,  $\Phi$  is an  $n - by - k$  matrix of regressors and the model is written in a vector form as

$$\mathbf{y} = \Phi \mathbf{c} + \boldsymbol{\epsilon} \quad (5.15)$$

The least square estimate of the coefficients is

$$\hat{\mathbf{c}} = (\Phi' \Phi)^{-1} \Phi' \mathbf{y} \quad (5.16)$$

To make regression analysis work in the smoothed data, the number of basis functions  $K$  is smaller than the number of sampling points  $n$  in order to avoid overfitting.

### **Penalization of roughness**

This method defines a measure of the roughness of the fitted curve. The way to further smooth the function by attaching an additional term that controls the roughness of some derivative is called *regularization*.

$$PENSSE = SSE(\mathbf{y}, \mathbf{c}) + \lambda PEN(x) \quad (5.17)$$

where  $PENSSE$  denotes the penalized SSE;  $SSE(\mathbf{y}, \mathbf{c})$  is sum of squared errors. The second term on the right hand side  $\lambda PEN(x)$  penalizes the roughness of fitted function  $x(t)$ .

$$PEN(x) = \int [D^2x(t)]^2 dt \quad (5.18)$$

The penalization term uses the second derivative  $D^2(x)$  of the square of  $x$ 's. This second derivative is called the total curvature and is used to smooth the curve. The term will be smaller if  $x$  is close to a linear function structure. The smoothing parameter  $\lambda$  controls the smoothness of curves.

### **Example: smoothing TCRPs**

The effect of  $\lambda$  on the smoothness of curves is illustrated in this part. Figure 5.2 shows four situations with different  $\lambda$  values in an ascending manner. It can be observed that the value of  $\lambda$  determines the smoothness of curves. The larger  $\lambda$  is, the less complex the fitted function is. In other words, a larger  $\lambda$  constrains the order of the fitted function because of the integration in Eq. (5.18). If the  $\lambda$  approaches infinity, the functional form  $x(t)$  will approximately approach a linear function structure. On the contrary, as  $\lambda$  tends to zero, the

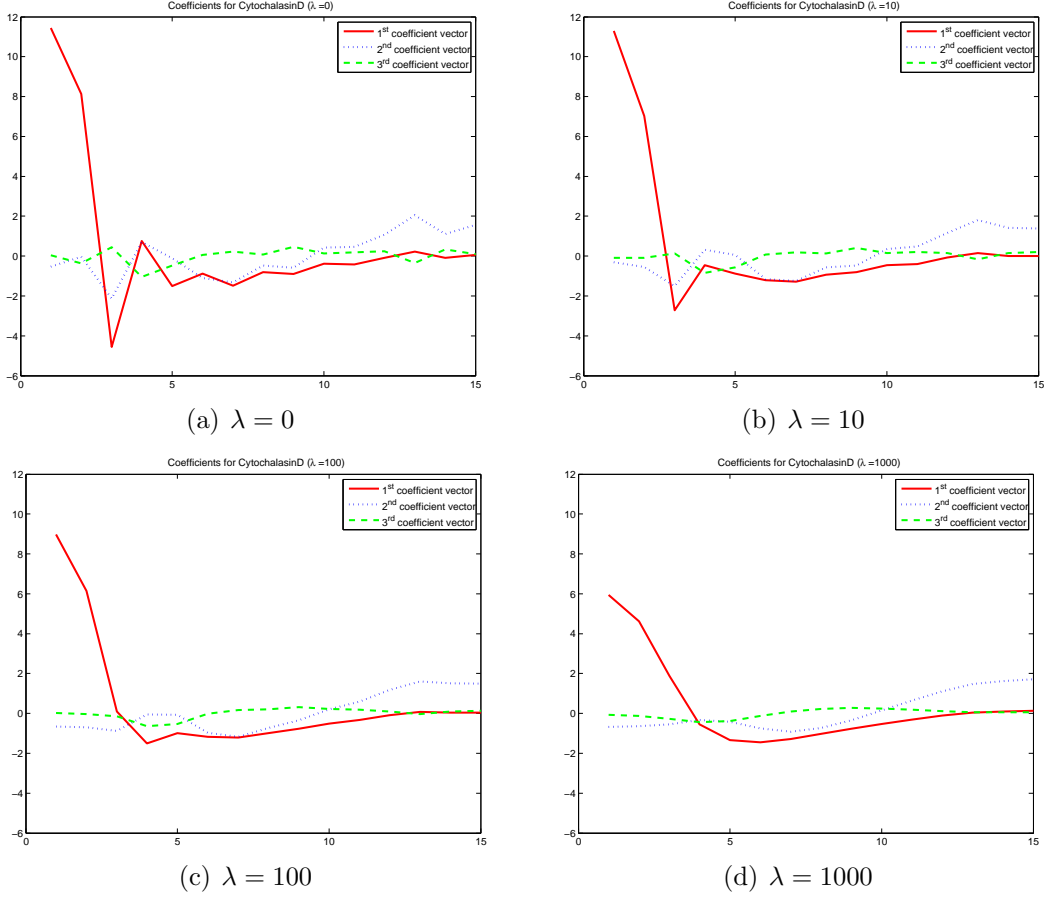


Figure 5.2: Curve smoothness are determined by the penalized parameter  $\lambda$

functional complexity is not constrained because the second term approaches zero.  $x(t)$  can be as rough as  $\mathbf{y}$  and it will pass the data points exactly.

The advantage of FDA lies in its application on the smoothness of curves so as to remove noise from the measurements. Further, as long as basis functions are given, data can be expressed with fewer coefficients in coefficient space. Besides, as a technique to treat all the measurements as a functional entity, this statistical method can deal with missing values, and nonuniform sampling problems to some extent [28]. Specific details about FDA can be accessed via Ramsay et al. [25].

## 5.4 Model-based hierarchical clustering method

After feature extraction via *PCA* and *FDA*, all the processed PC score vectors will be input to a model-based hierarchical clustering algorithm developed via Gaussian finite mixture models and the *MCLUST* algorithm by Fraley and Raftery. For more details about the algorithm, the readers are referred to Fraley and Raftery [29].

Agglomerative model-based clustering is a separate function in Model-based Clustering Toolbox in MATLAB® developed by A. Martinez and W. Martinez [30]. Different from traditional hierarchical clustering which merges two closest clusters in terms of some distance metric (e.g. Euclidean distance, City Block distance, Mahalanobis distance), the model-based hierarchical clustering algorithm merges clusters such that a likelihood function is maximized given a model structure. The model-based clustering algorithm is presented briefly using the equations introduced by Fraley [31].

### 5.4.1 Model-based clustering

For model based clustering, each observation is generated by

$$\mathcal{L}(\theta_1, \dots, \theta_G; \gamma | \mathbf{x}) = \prod_{i=1}^n f_{\gamma_i}(\mathbf{x}_i | \theta_{\gamma_i}) \quad (5.19)$$

where  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  is the observation vector,  $\gamma = (\gamma_1, \dots, \gamma_n)^T$  is the label vector and  $\theta_1, \dots, \theta_G$  are the parameters.

When  $f_k(\mathbf{x} | \theta_k)$  follows a multivariate normal distribution, the likelihood from Eq. (5.19) is

$$\begin{aligned} \mathcal{L}(\mu_1, \dots, \mu_G; \Sigma_1, \dots, \Sigma_G; \gamma | \mathbf{x}) \\ = \prod_{k=1}^G \prod_{i \in \mathcal{I}_k} (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)\right\} \end{aligned} \quad (5.20)$$

where  $\mathcal{I}_k = \{i : \gamma_i = k\}$  is the set of indices corresponding to observations belonging to the  $k^{th}$  group. The estimation of  $\mu_k, \hat{\mu}_k$  is  $\bar{x}_k = s_k/n_k$  where  $s_k$  and  $n_k$  are the sum and number of observations in the  $k^{th}$  group, respectively. The likelihood of  $\Sigma = (\Sigma_1, \dots, \Sigma_G)$  is

$$\begin{aligned} l(\Sigma_1, \dots, \Sigma_G; \gamma | \mathbf{x}; \hat{\mu}_1, \dots, \hat{\mu}_G) \\ = -\frac{pn \log(2\pi)}{2} - \frac{1}{2} \sum_{k=1}^G \{ \text{tr}(W_k \Sigma_k^{-1}) + n_k \log |\Sigma_k| \} \quad (5.21) \end{aligned}$$

where  $W_k = \sum_{i \in \mathcal{I}_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T$  is the sample cross product matrix for the  $k^{th}$  group.

According to four different structures of  $\Sigma_k$ , there are four criteria to maximize the likelihood. If  $\Sigma_k = \sigma^2 I$ , maximizing the log-likelihood is equivalent to minimizing  $\text{Tr}(\sum_{k=1}^G W_k)$ ; if each group has a different variance, i.e.  $\Sigma_k = \sigma_k^2 I$ ,  $\gamma$  is chosen to minimize  $\sum_{k=1}^G n_k \log \text{Tr}(\frac{W_k}{n_k})$ ; if  $\Sigma_k$  is the same for all groups without structural constraints, then the  $\gamma$  which minimizes  $|\sum_{k=1}^G W_k|$  maximizes the log-likelihood; if  $\Sigma_k$  varies substantially between groups, then  $\gamma$  minimizes  $\sum_{k=1}^G n_k \log |\frac{W_k}{n_k}|$ . The toolbox implements the general fourth situation.

### 5.4.2 Hierarchical clustering

Agglomerative hierarchical clustering is a bottom-up clustering method. Initialized in their own clusters, pairs of objects are successively connected and merged moving up the hierarchy to form clusters until finally only one cluster remains.

#### Merging cost update

Classical agglomerative methods merge a pair of clusters according to a metric or “cost” which measures how much the sum of squares will increase when we merge them. We consider Ward’s method as an example. The cost of merging

clusters  $A$  and  $B$  is

$$\Delta(A, B) = \sum_{i \in A \cup B} \|\mathbf{x}_i - \mathbf{c}_{A \cup B}\|^2 - \sum_{i \in A} \|\mathbf{x}_i - \mathbf{c}_A\|^2 - \sum_{i \in B} \|\mathbf{x}_i - \mathbf{c}_B\|^2 \quad (5.22)$$

$$= \frac{n_A n_B}{n_A + n_B} \|\mathbf{c}_A - \mathbf{c}_B\|^2 \quad (5.23)$$

where  $\mathbf{x}_i$  is an observation vector in its cluster,  $\mathbf{c}$  is the center of a cluster and  $n$  is the number of observations in a cluster. The center of the merged cluster is

$$\mathbf{c}_{A \cup B} = \frac{n_A \mathbf{c}_A + n_B \mathbf{c}_B}{n_{A \cup B}} = \frac{n_A \mathbf{c}_A + n_B \mathbf{c}_B}{n_A + n_B} \quad (5.24)$$

If there is a group  $C$  which will be merged into  $A \cup B$ , the cost will be

$$\Delta(\langle A, B \rangle, C) = \frac{(n_A + n_C)\Delta(A, C) + (n_B + n_C)\Delta(B, C) - n_C \Delta(A, B)}{n_A + n_B + n_C} \quad (5.25)$$

As long as neither of the clusters in that pair is involved in a merge, the cost will remain fixed [31].

### 5.4.3 Model-based hierarchical clustering

The relationship between successive stages exists in terms of sample cross-product matrix. It is expressed as

$$W_{\langle i, j \rangle} = W_i + W_j + w_{i, j} w_{i, j}^T \quad (5.26)$$

where  $W$  is the sample cross-product matrix and

$$w_{i, j} = \eta_{ji} s_i - \eta_{ij} s_j \quad (5.27)$$

$$\eta_{ij} = \sqrt{\frac{n_i}{n_j(n_i + n_j)}} \quad (5.28)$$

The derivation of the relationship is described in more detail in [31]. When each group has an unconstrained covariance structure  $\Sigma_k$ , the objective function



$\sum_{k=1}^G n_k \log \left| \frac{W_k}{n_k} \right|$  is to be minimized and can be updated as

$$\sum_{k=1}^G n_k \log \left[ \left| \frac{W_k}{n_k} \right| + \beta \left\{ \frac{\text{Tr}(W_k) + \alpha \frac{\text{Tr}(\mathcal{W})}{np}}{n_k} \right\} \right] \quad (5.29)$$

If  $L$  denotes the Cholesky factor of  $W$ , Eq. (5.26) becomes

$$L_{\langle i,j \rangle} L_{\langle i,j \rangle}^T = L_i L_i^T + L_j L_j^T + w_{ij} w_{ij}^T = (L_i \quad L_j \quad w_{ij})(L_i \quad L_j \quad w_{ij})^T \quad (5.30)$$

By applying Given's rotation [32] on a composite matrix  $\begin{pmatrix} L_i^T \\ L_j^T \\ w_{ij}^T \end{pmatrix}$ , another

composite matrix  $\begin{pmatrix} L_{\langle i,j \rangle}^T \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}$  can be generated. Hence,  $L_{\langle i,j \rangle}$  is calculated effi-

ciently. The pair of clusters will be chosen to merge if the updated objective function is minimized. Specific details about the time efficacy can be accessed via [31].

In model-based hierarchical clustering strategy, the pair of clusters are merged at each stage with the maximum likelihood. The advantage of this strategy is that model-based clustering can be associated with Bayesian criterion which can determine the best partition according to the model defined in the hierarchy automatically.

## 5.5 Application

By synthesizing the methods above, including feature extraction based on PCA & FDA and hierarchical clustering approaches, this section will illustrate the pattern recognition application on TCRP data denoted with nonuniform dimensionality and show its efficiency. PCA is used here to unify and format the dimensionality of data by extracting the first three principle component

scores (PC scores) consistently. FDA is then used to indicate the PC scores with fewer coefficients and remove the noise at the same time. Agglomerative model-based hierarchical clustering will cluster the extracted coefficient vectors by merging the pairs with the maximum likelihood.

### 5.5.1 Feature extraction

Gemcitabine HCl is taken as an example to illustrate how features are extracted.

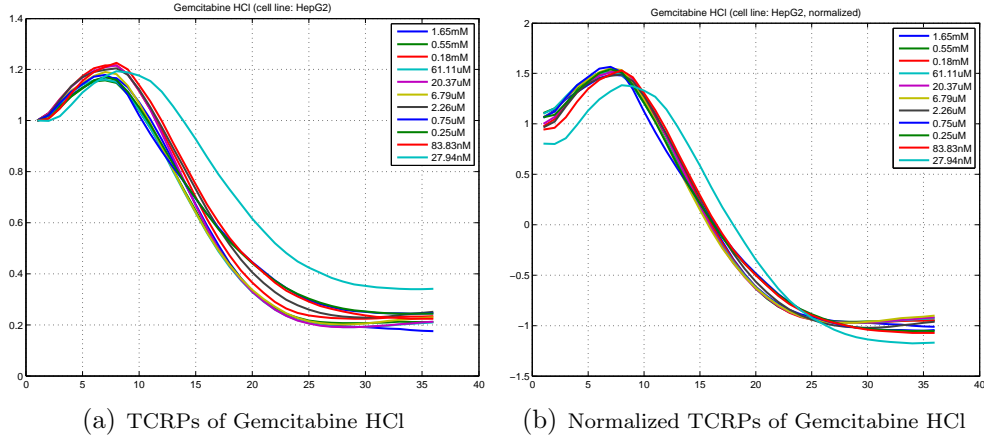


Figure 5.3: TCRPs of Gemcitabine HCl are normalized with mean 0 and standard deviation 1.

Figure 5.3(a) shows the selected TCRPs. To perform PCA with standardized variables based on correlations, a normalization is applied to the data matrix. The eligible *RNCI* data matrix is denoted as  $X$ . Each element of  $X$  is transformed so that columns of  $X$  are normalized with mean 0 and standard deviation 1. The normalized data is denoted as  $X_N$ . Figure 5.3(b) shows all the normalized TCRPs of Gemcitabine HCl. PCA is then used to transform  $X_N$  into the representation, termed as PC scores, in the principal component space. The first three PC scores are retained regardless of how much variation exists in the projected data. Figure 5.5(a) shows the first three PC scores of  $X$ . The first PC score (red, solid) shares the similar shapes with the TCRPs

before PCA while the second and third PC scores are close to zero.

FDA uses coefficients of basis functions to denote the PC scores. Fifteen cubic spline functions construct the basis functions. Figure 5.4 illustrates the structures and shapes of basis functions. In that figure, the red dot vertical lines indicate horizontal coordinates of interior knots. The basis spline functions in the center reach the peak point at knots. The first basis function rises to the peak value of one at the left boundary point and decreases to a value of zero when reaching the first interior knot from the left. The last basis function on the right hand side is similar.

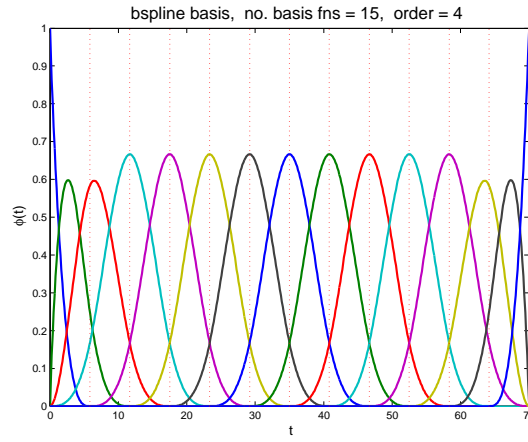


Figure 5.4: Basis functions: 15 cubic spline functions

FDA is applied to the PC scores using the basis function defined above. Figure 5.5(b) shows the extracted coefficients with  $\lambda = 10$ . From the figure, we see that the shapes and dynamics are retained while the  $x$ -axis is effectively compressed. Also, the noise in PC scores level is removed in coefficient level to some extent.

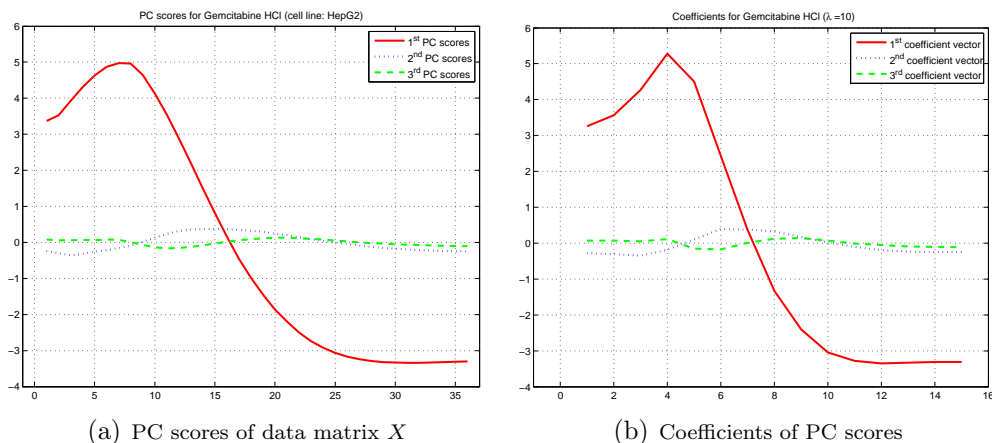


Figure 5.5: First three PC scores and the coefficients after FDA

All chemicals with data matrices denoted by non-uniform dimensionality are formulated in the same way, which generates the input feature matrix for model-based hierarchical clustering.

### 5.5.2 Model-based hierarchical clustering analysis

This subsection is aimed at clustering the feature representation matrix under a model-based hierarchical clustering scenario. The feature representation matrix is input into the algorithm and a dendrogram is generated as an output for users to analyze. One challenging problem here is the determination of the number of clusters using dendrogram. The use of hierarchical clustering can avoid this problem. However, the results of this connectivity based clustering are not easy to use, because it cannot produce a unique partitioning of the data set. Therefore, no general conclusion can be made about comparison among groups.

We adopt two strategies to determine the numbers of clusters. First, we rely on users' interpretation as well as background knowledge and use it as a guideline in dendrogram cutting. It is relatively subjective; however, it can be smoothly interpreted by experts such as biological scientists according to their satisfaction and requirement. Second, to let the data decide how many clusters are supposed to be, we use Bayesian information criterion (BIC) to estimate

the proper number. It is justified and operates as an efficient way in statistical model selection considering likelihood [33].

### 5.5.3 One-level clustering

One-level clustering means cutting the dendrogram in *one* time. Under the one-level clustering framework, to ensure more information is involved in classification, the first two coefficient vectors of the PC score is chosen. These vectors are concatenated as the feature representation. Figure 5.6 shows the dendrogram of chemicals whose features are the first two coefficient vectors.

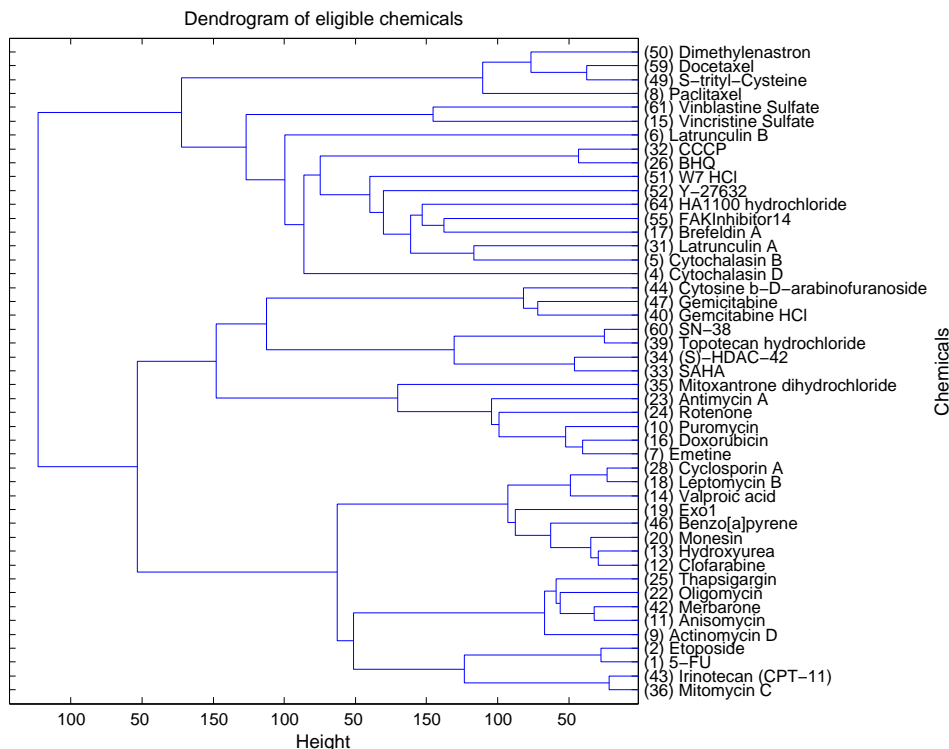


Figure 5.6: Dendrogram of chemicals

Prior knowledge about the number of clusters indicates that 6 to 12 is appropriate and accepted according to the mechanisms of action. However, it is not fully applicable because we intend to classify chemicals based on profile shape similarity so as to classify them into mode of action categories. Although

mechanism of action is associated with mode of action, they are essentially different concepts in discriminating and evaluating chemicals from different levels and non-proportionality between them weakens the identical distribution of chemicals [13, 34].

As referential information, we first use 6 to generate the classification results.

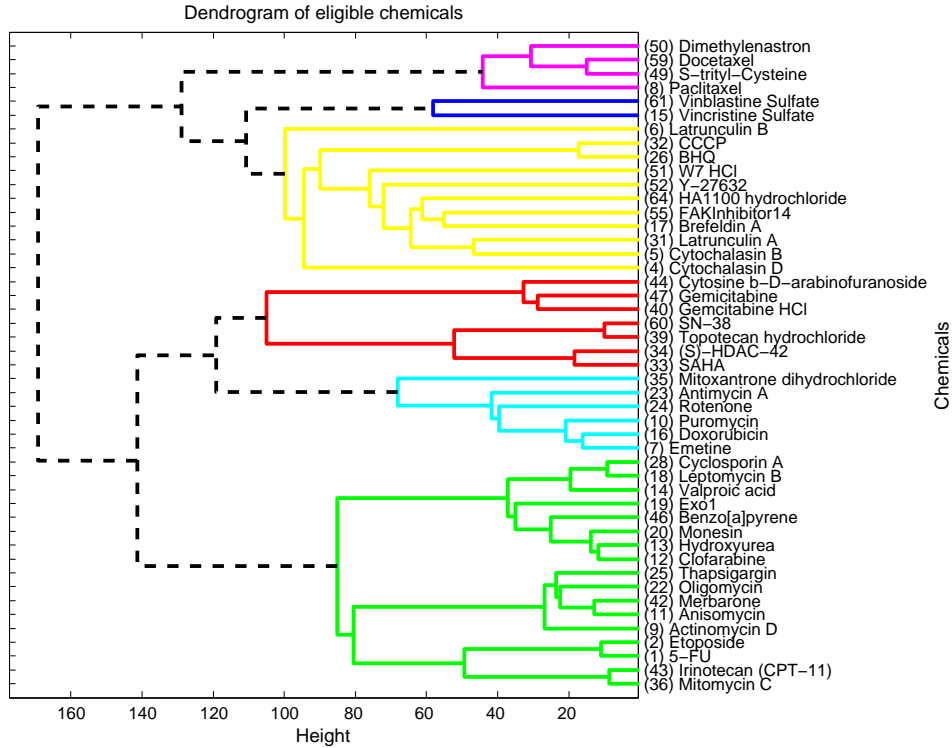
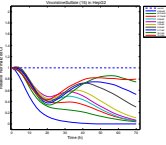
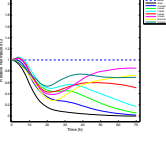
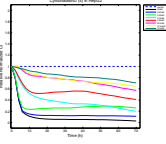
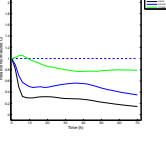
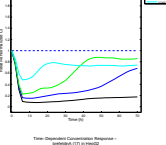
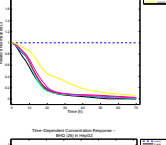
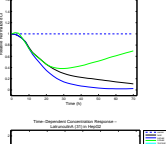
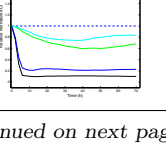


Figure 5.7: Colored dendrogram of chemicals with the first two coefficient vectors aligned as features. 6 reasonable clusters are generated.

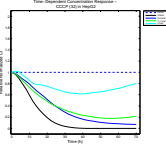
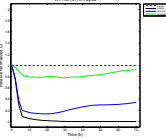
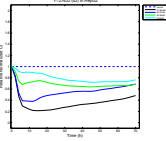
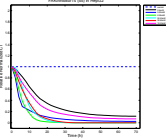
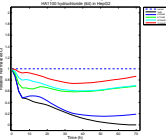
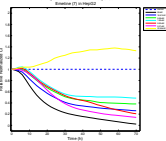
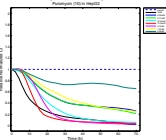
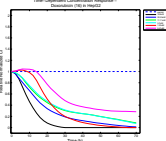
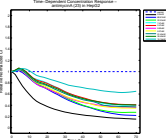
Figure 5.7 indicates the distribution of chemicals in each cluster when the dendrogram is cut into 6 separate clusters. The TCRPs are then illustrated to give specification on the results.

Table 5.1: Results of PCA & FDA based hierarchical classification with one-level dendrogram cutting (cell line: HepG2. 47 of 65 chemicals are eligible.)

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
1	15	Vincristine Sulfate	H <sub>2</sub> O	250 $\mu$ M – 4.23nM	
	61	Vinblastine sulfate	H <sub>2</sub> O	40 $\mu$ M – 0.68nM	
2	4	Cytochalasin D	DMSO	20 $\mu$ M – 0.339nM	
	5	Cytochalasin B	DMSO	20 $\mu$ M – 0.339nM	
	6	Latrunculin B	DMSO	20 $\mu$ M – 0.339nM	
	17	Brefeldin A (BEF)	DMSO	40 $\mu$ M – 0.68nM	
	26	BHQ	DMSO	400 $\mu$ M – 7nM	
	31	Latrunculin A	EtOH	2 $\mu$ M – 0.03nM	

continued on next page

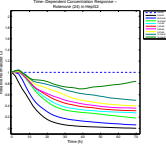
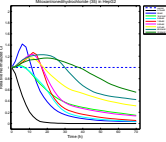
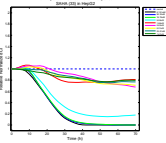
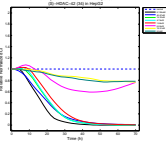
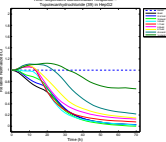
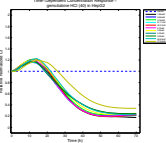
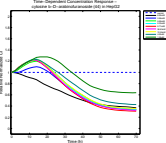
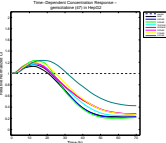
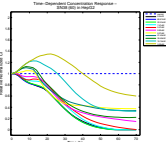
continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	32	CCCP	DMSO	100 $\mu$ M – 1.69nM	
	51	W7 HCl	DMSO	200 $\mu$ M – 3.39nM	
	52	Y-27632	DMSO	188 $\mu$ M – 3.18nM	
	55	FAKInhibitor14	H <sub>2</sub> O	2500 $\mu$ M – 42.34nM	
	64	HA1100 hydrochloride	H <sub>2</sub> O	1000 $\mu$ M – 16.94nM	
3	7	Emetine	H <sub>2</sub> O	50 $\mu$ M – 0.847nM	
	10	Puromycin	H <sub>2</sub> O	1000 $\mu$ M – 17nM	
	16	Doxorubicin (DOX)	H <sub>2</sub> O	100 $\mu$ M – 1.69nM	
	23	Antimycin A	EtOH	200 $\mu$ M – 3.387nM	

continued on next page

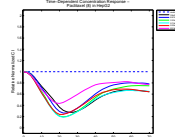
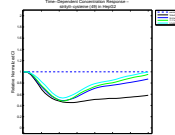
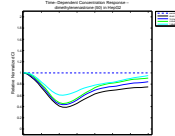
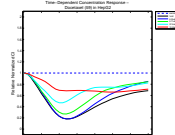
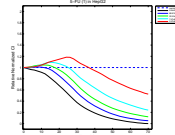
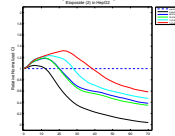
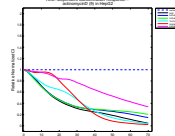
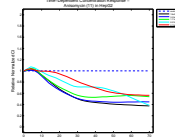
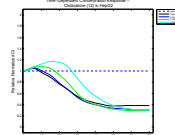


continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	24	Rotenone	DMSO	200 $\mu$ M – 3.387nM	
	35	Mitoxantrone	DMSO	150 $\mu$ M – 2.54nM	
4	33	SAHA	DMSO	151 $\mu$ M – 2.56nM	
	34	(S)-HDAC-42	DMSO	128 $\mu$ M – 2.17nM	
	39	Topotecan	DMSO	95 $\mu$ M – 1.61nM	
	40	Gemcitabine HCl	H <sub>2</sub> O	1650 $\mu$ M – 27.94nM	
	44	Cytosine	H <sub>2</sub> O	8950 $\mu$ M – 151.57nM	
	47	Gemcitabine	H <sub>2</sub> O	2 $\mu$ M – 0.03nM	
	60	SN-38	DMSO	200 $\mu$ M – 3.39nM	

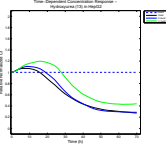
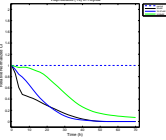
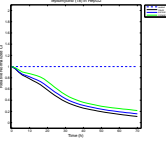
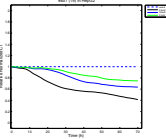
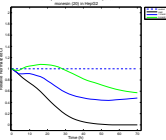
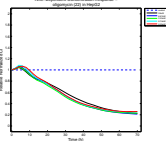
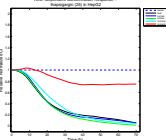
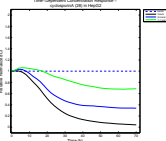
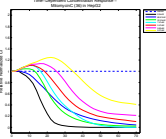
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
5	8	Paclitaxel	DMSO	20 $\mu$ M – 0.339nM	
	49	S-trityl-Cysteine	DMSO	100 $\mu$ M – 1.69nM	
	50	Dimethylnastron	DMSO	40 $\mu$ M – 0.68nM	
	59	Docetaxel	DMSO	1 $\mu$ M – 0.02nM	
6	1	5-fluorouracil (5-FU)	DMSO	200 $\mu$ M – 3.39nM	
	2	Etoposide phosphate	DMSO	200 $\mu$ M – 3.39nM	
	9	Actinomycin D	DMSO	2 $\mu$ M – 0.0339nM	
	11	Anisomycin	H <sub>2</sub> O	10 $\mu$ M – 0.17nM	
	12	Clofarabine (CLOF)	H <sub>2</sub> O	25 $\mu$ M – 0.42nM	

continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	13	Hydroxyurea (HU)	H <sub>2</sub> O	10mM – 169nM	
	14	Valproic acid	H <sub>2</sub> O	50mM – 847nM	
	18	Leptomycin B (LMB)	EtOH	20nM – 0.000339nM	
	19	Exo 1	DMSO	300μM – 5.08nM	
	20	Monensin	DMSO	4μM – 0.068nM	
	22	Oligomycin	DMSO	20μM – 0.339nM	
	25	Thapsigargin	DMSO	2μM – 0.0339nM	
	28	Cyclosporin A	DMSO	100μM – 1.69nM	
	36	Mitomycin C	DMSO	200μM – 3.39nM	

continued on next page

continued from previous page

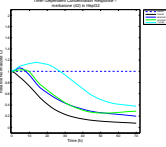
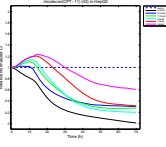
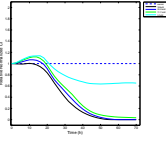
Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	42	Merbarone	DMSO	200 $\mu$ M – 3.39nM	
	43	Irinotecan (CPT-11)	DMSO	160 $\mu$ M – 2.71nM	
	46	Benzo[a]pyrene	DMSO	100 $\mu$ M – 1.69nM	

Table 5.1 indicates that except for some chemicals, most of the chemicals are classified properly according to shapes. The overall classification quality is acceptable. However, the main deficiency of one-level clustering lies in the fact that it may generate a singleton cluster. Because the dendrogram is cut consistently, if a singleton is connected with other chemicals at a large height, the singleton may be possibly separated. Since in the application we expect to avoid many singleton clusters, based on this idea, a two-level hierarchical clustering scenario is designed by cutting the dendrogram of each level at a reasonable height. Therefore, the structures of dendrogram are not destroyed by an improper choice of referential clustering number. Dendrograms are cut at a height where mistakes occur with a small probability.

#### 5.5.4 Two-level clustering

As a hierarchical clustering scenario, this method is designed to use partial information of the feature representation at different level so as to classify chemicals. In the first level, the coefficients of the first PC scores with respect to the slopes of TCRPs are used, while in the second level, the coefficients of the first PC scores with respect to the TCRPs are used. The first coefficients

denote the main tendency of TCRPs.

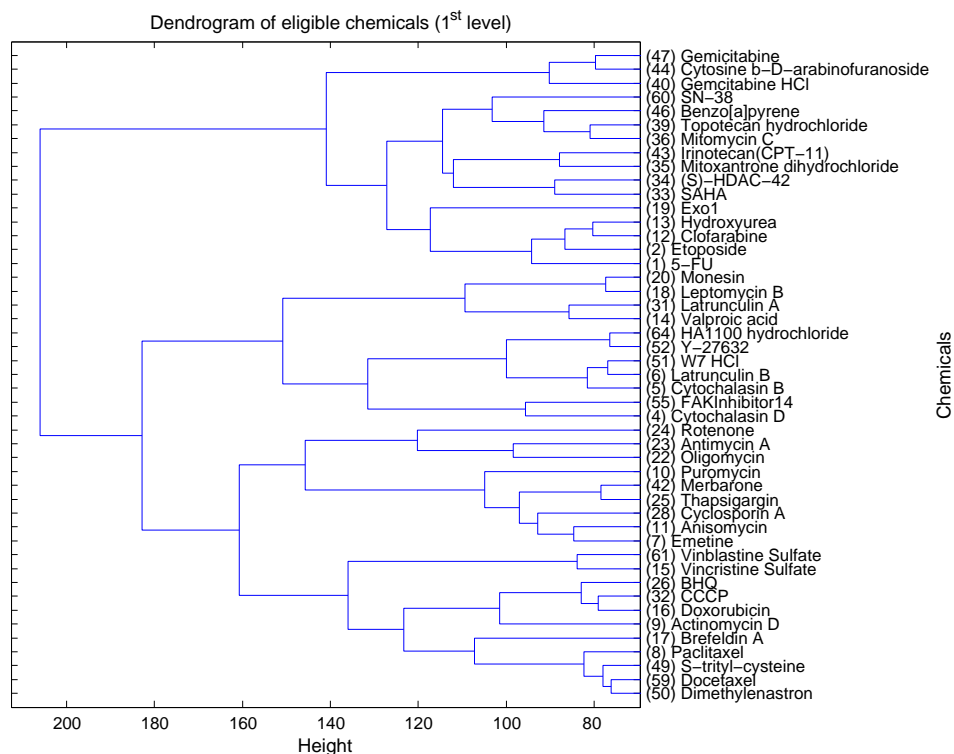


Figure 5.8: Dendrogram of the chemicals in the 1<sup>st</sup> level classification

Figure 5.8 shows the dendrogram generated via model-based hierarchical clustering based on coefficients of the first PC scores with respect to the slopes of TCRPs. Slope is an important indicator to reflect the degree of incline in TCRPs. It measures the variation of *RNCI* within each time interval. The coefficients of the first PC scores on slopes of TCRPs then reflect the main variation of curve trends among all eligible TCRPs and can effectively discriminate chemicals with apparent changes in slopes. Figure 5.9 shows the dendrogram cut at a proper height with three reasonable subtrees generated. They are marked with different colors.

Each subtree generated from 1<sup>st</sup> level is then input into the model-based hierarchical clustering algorithm again by using the coefficients of the first PC scores from the TCRPs. The information about the values of *RNCI* is used in this level to discriminate chemicals further. Figure 5.10 shows the dendrogram

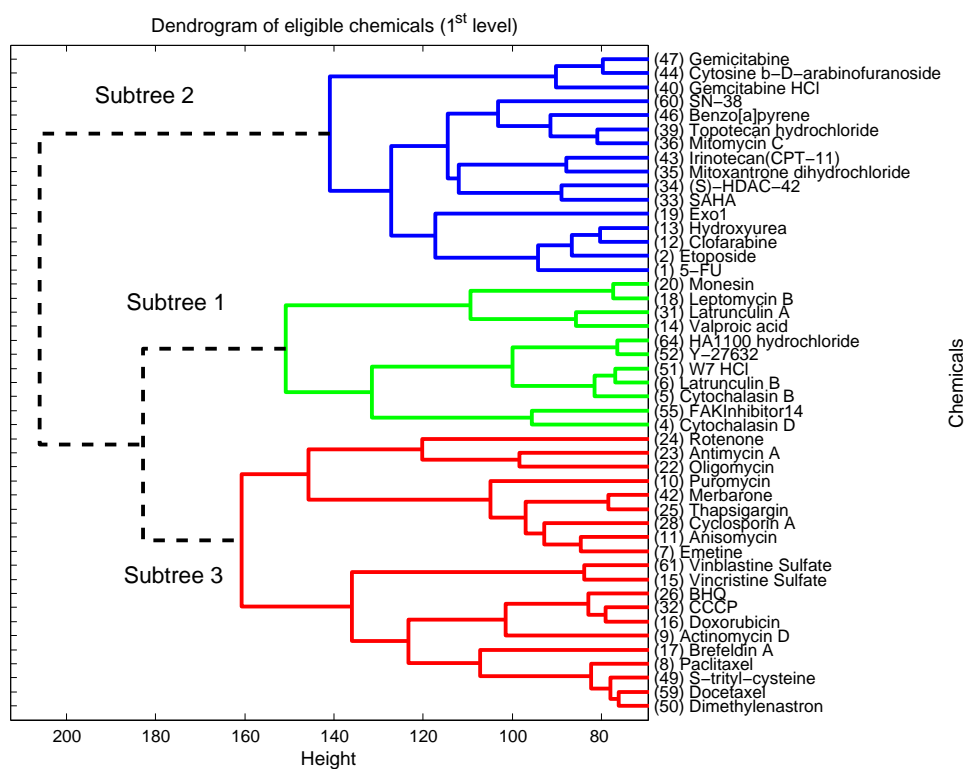


Figure 5.9: Dendrogram with three reasonable subtrees marked with colors

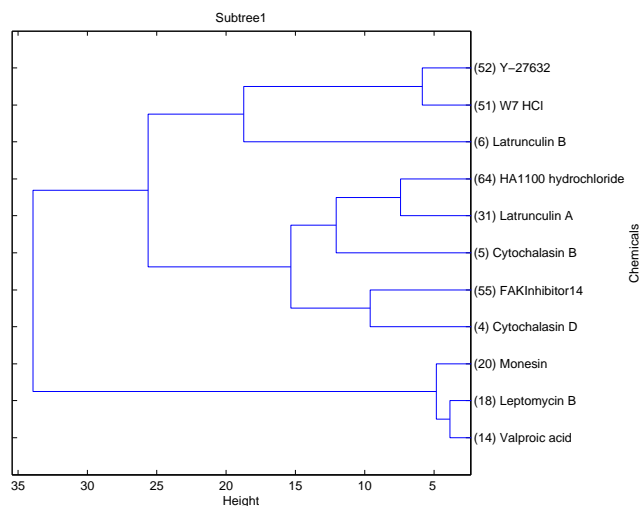


Figure 5.10: Dendrogram of chemicals in subtree 1

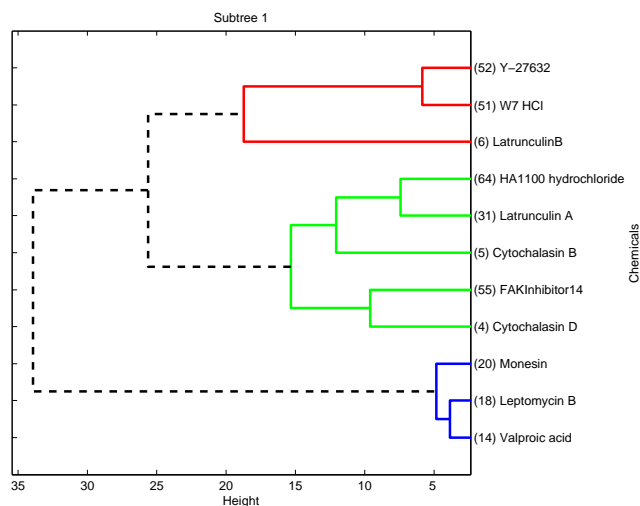


Figure 5.11: Dendrogram of chemicals in subtree 1 with three reasonable clusters marked in colors

of chemicals in subtree 1. Figure 5.11 shows the cutting results in Figure 5.10 marked with different colors. Other dendrograms in the  $2^{nd}$  level are cut at proper heights similarly. Figures 5.12, 5.13 and 5.14 show the dendrograms and the marked dendrograms for other subtrees in  $2^{nd}$  level.

Clustering results of eligible chemicals are listed in Table 5.2.

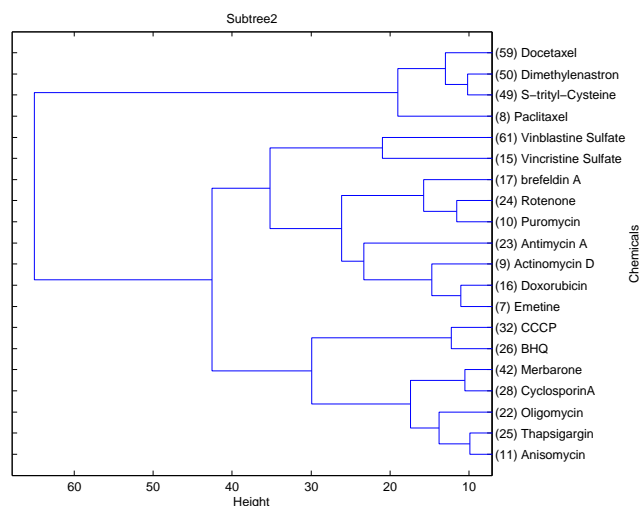


Figure 5.12: Dendrogram of chemicals in subtree 2

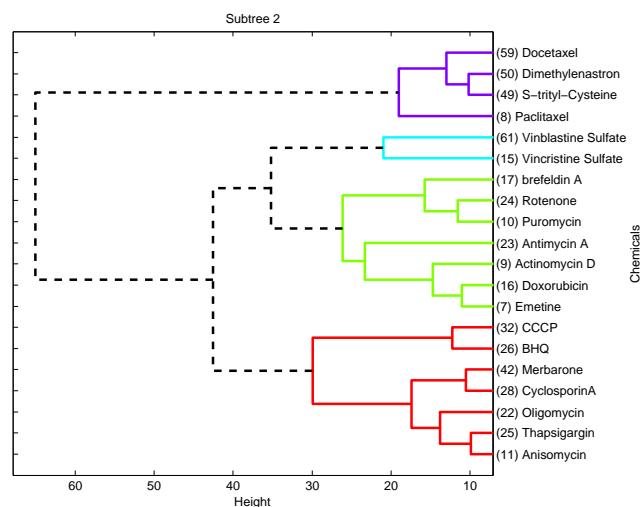


Figure 5.13: Dendrogram of chemicals in subtree 2 with four reasonable clusters marked in colors

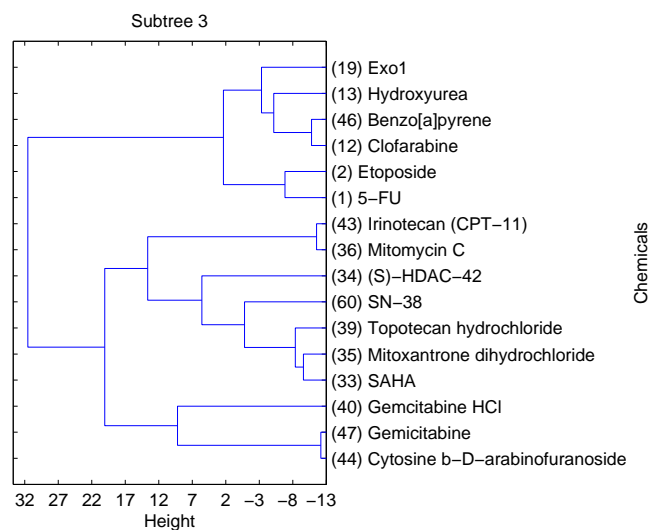


Figure 5.14: Dendrogram of chemicals in subtree 3



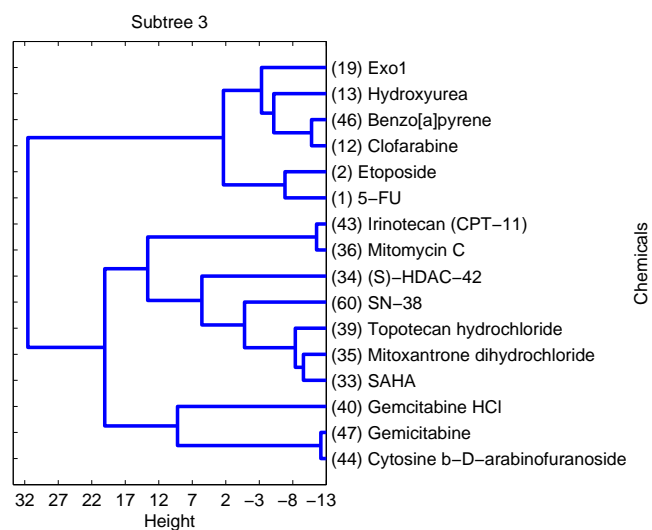
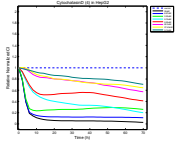
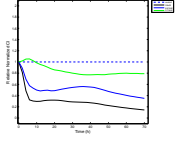
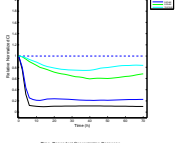
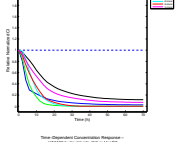
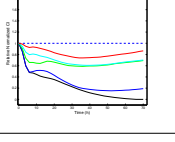
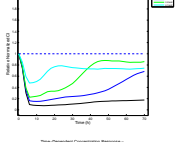
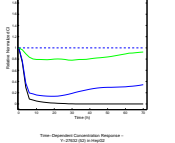
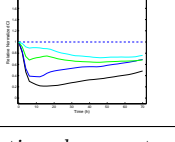


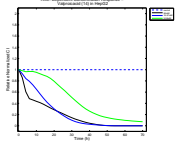
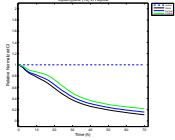
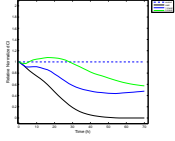
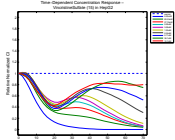
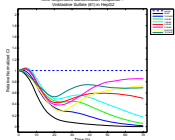
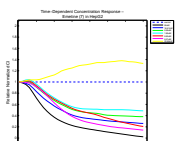
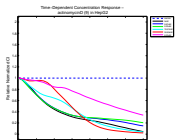
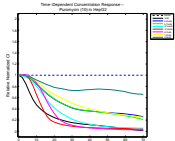
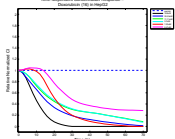
Figure 5.15: Dendrogram of chemicals in subtree 3 with one reasonable cluster marked in blue

Table 5.2: Results of PCA & FDA based hierarchical classification with two-level dendrogram cutting (cell line: HepG2. 47 of 65 chemicals are eligible.)

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
1	4	Cytochalasin D	DMSO	20 $\mu$ M – 0.339nM	
	5	Cytochalasin B	DMSO	20 $\mu$ M – 0.339nM	
	31	Latrunculin A	EtOH	2 $\mu$ M – 0.03nM	
	55	FAKInhibitor14	H <sub>2</sub> O	2500 $\mu$ M – 42.34nM	
	64	HA1100 hydrochloride	H <sub>2</sub> O	1000 $\mu$ M – 16.94nM	
2	6	Latrunculin B	DMSO	20 $\mu$ M – 0.339nM	
	51	W7 HCl	DMSO	200 $\mu$ M – 3.39nM	
	52	Y-27632	DMSO	188 $\mu$ M – 3.18nM	

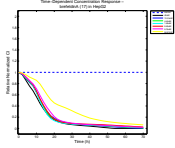
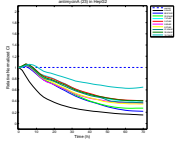
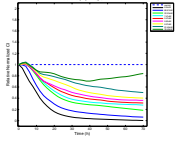
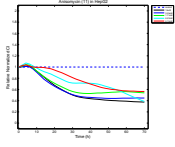
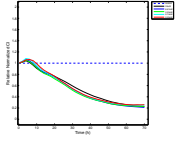
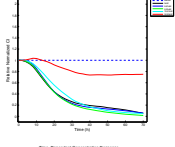
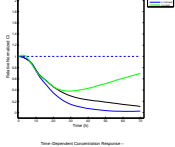
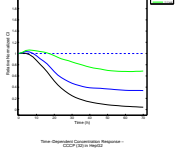
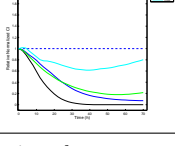
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
3	14	Valproic acid	H <sub>2</sub> O	50mM – 847nM	
	18	Leptomycin B (LMB)	EtOH	20nM – 0.000339nM	
	20	Monensin	DMSO	4μM – 0.068nM	
4	15	Vincristine Sulfate	H <sub>2</sub> O	250μM – 4.23nM	
	61	Vinblastine Sulfate	H <sub>2</sub> O	40μM – 0.68nM	
5	7	Emetine	H <sub>2</sub> O	50μM – 0.847nM	
	9	Actinomycin D	DMSO	2μM – 0.0339nM	
	10	Puromycin	H <sub>2</sub> O	1000μM – 17nM	
	16	Doxorubicin (DOX)	H <sub>2</sub> O	100μM – 1.69nM	

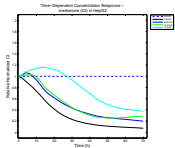
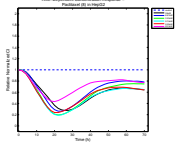
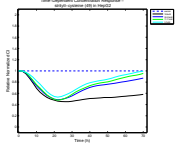
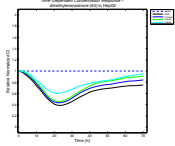
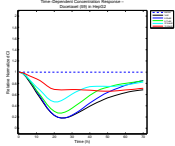
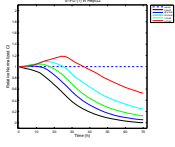
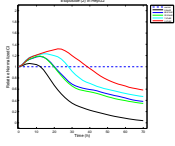
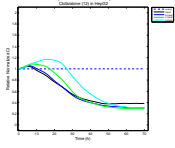
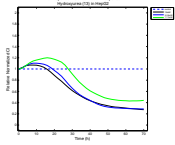
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	17	Brefeldin A (BEF)	DMSO	40 $\mu$ M – 0.68nM	
	23	Antimycin A	EtOH	200 $\mu$ M – 3.387nM	
	24	Rotenone	DMSO	200 $\mu$ M – 3.387nM	
6	11	Anisomycin	H <sub>2</sub> O	10 $\mu$ M – 0.17nM	
	22	Oligomycin	DMSO	20 $\mu$ M – 0.339nM	
	25	Thapsigargin	DMSO	2 $\mu$ M – 0.0339nM	
	26	BHQ	DMSO	400 $\mu$ M – 7nM	
	28	Cyclosporin A	DMSO	100 $\mu$ M – 1.69nM	
	32	CCCP	DMSO	100 $\mu$ M – 1.69nM	

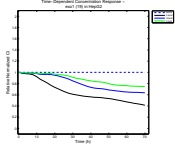
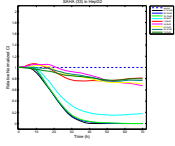
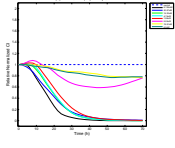
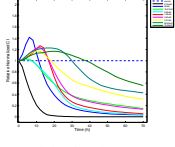
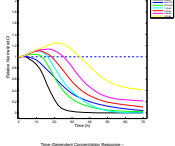
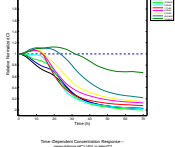
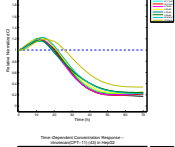
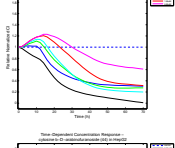
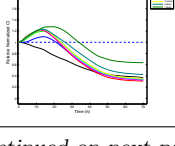
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	42	Merbarone	DMSO	200 $\mu$ M – 3.39nM	
7	8	Paclitaxel	DMSO	20 $\mu$ M – 0.339nM	
	49	S-tritytl-Cysteine	DMSO	100 $\mu$ M – 1.69nM	
	50	Dimethylenastron	DMSO	40 $\mu$ M – 0.68nM	
	59	Docetaxel	DMSO	1 $\mu$ M – 0.02nM	
8	1	5-fluorouracil (5-FU)	DMSO	200 $\mu$ M – 3.39nM	
	2	Etoposide phosphate	DMSO	200 $\mu$ M – 3.39nM	
	12	Clofarabine (CLOF)	H <sub>2</sub> O	25 $\mu$ M – 0.42nM	
	13	Hydroxyurea (HU)	H <sub>2</sub> O	10mM – 169nM	

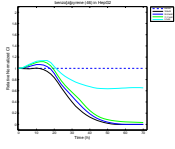
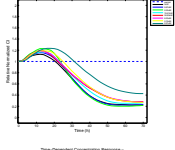
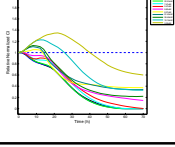
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	19	Exo 1	DMSO	300 $\mu$ M – 5.08nM	
	33	SAHA	DMSO	151 $\mu$ M – 2.56nM	
	34	(S)-HDAC-42	DMSO	128 $\mu$ M – 2.17nM	
	35	Mitoxantrone	DMSO	150 $\mu$ M – 2.54nM	
	36	Mitomycin C	DMSO	200 $\mu$ M – 3.39nM	
	39	Topotecan	DMSO	95 $\mu$ M – 1.61nM	
	40	Gemcitabine HCl	H <sub>2</sub> O	1650 $\mu$ M – 27.94nM	
	43	Irinotecan (CPT-11)	DMSO	160 $\mu$ M – 2.71nM	
	44	Cytosine	H <sub>2</sub> O	8950 $\mu$ M – 151.57nM	

continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	46	Benzo[a]pyrene	DMSO	100 $\mu$ M – 1.69nM	
	47	Gemcitabine	H <sub>2</sub> O	2 $\mu$ M – 0.03nM	
	60	SN-38	DMSO	200 $\mu$ M – 3.39nM	

### 5.5.5 Automatic determination of cluster number

Cluster number is an important parameter in unsupervised clustering algorithms. The determination of cluster number is considered as a process of model selection. Most clustering algorithms require parameters that either directly or indirectly specify the number of clusters. Setting these parameters requires either the existing knowledge of the data or time-consuming trial and error. The latter case still requires that the user has sufficient domain knowledge to gauge good clustering results [35].

Except for users' interpretation and specification on cluster number, an appropriate number can be obtained via Bayesian Information Criterion (BIC) which is based on penalized likelihood estimation. BIC is defined as follows.

$$BIC = 2L_M(\mathbf{x}, \hat{\theta}) - m_M \log(n) \quad (5.31)$$

where  $m_M$  is the number of parameters in model M,  $L_M$  is the log likelihood, and  $n$  is the number of observations. The selected model will be the one with the highest BIC value defined in Eq. (5.31).

Model-based Clustering Toolbox in MATLAB<sup>®</sup> developed by A. Martinez

and W. Martinez [30] includes functions which apply Expectation Maximization (EM) algorithm to obtain the final estimates of model parameters as well as those which calculate BIC scores. Details about how EM algorithm works can be accessed via [36]. Different ways of model parametrization are discussed in varying degrees in [37], [31] and [38]. Specifically, [37] analyzed the assumptions for [39] and [40] in model structure parametrization and its applicability and proposed general parametrization criteria based on eigenvalue decomposition as a proper extension. Since different constraints are imposed on covariance matrices of models, different model types are involved in the calculation of BIC scores. To make the application more understandable, the selected scenario is based on the parametrization introduced above.

Figure 5.16 shows the BIC scores calculated based on the coefficients of first PC scores on slopes of TCRPs.

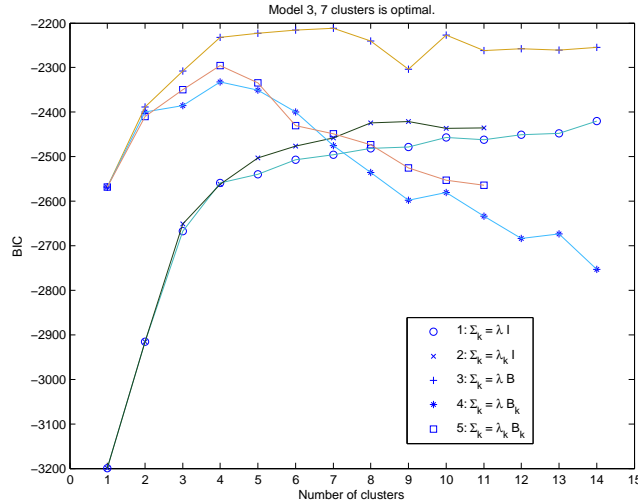
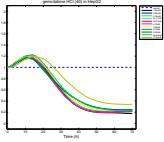
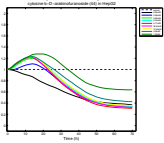
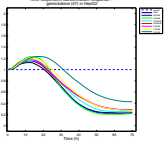
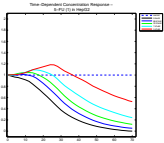
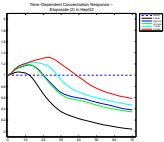
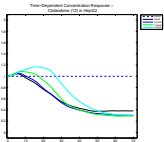
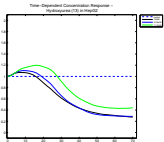
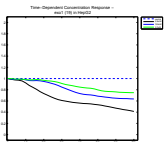


Figure 5.16: Model selection using BIC score

The figure indicates that the third model structure ( $\Sigma_k = \lambda B$ ) gives the largest BIC score. 7 clusters are optimal in this situation. The clustering number is reasonably consistent with prior information. The results are listed in Table 5.3.

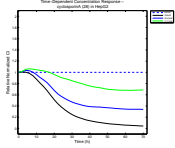
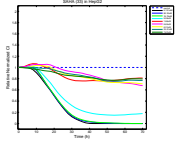
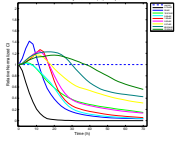
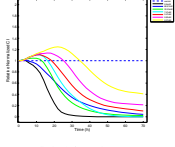
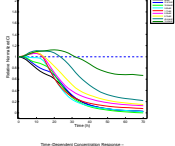
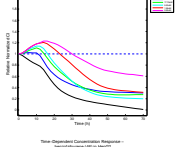
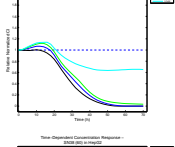
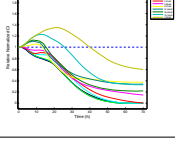
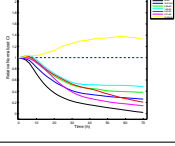


Table 5.3: Results of PCA & FDA based hierarchical classification with the number of clusters determined using BIC (cell line: HepG2. 47 of 65 chemicals are eligible.)

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
1	40	Gemcitabine HCl	H <sub>2</sub> O	1650 $\mu$ M – 27.94nM	
	44	Cytosine	H <sub>2</sub> O	8950 $\mu$ M – 151.57nM	
	47	Gemcitabine	H <sub>2</sub> O	2 $\mu$ M – 0.03nM	
2	1	5-fluorouracil (5-FU)	DMSO	200 $\mu$ M – 3.39nM	
	2	Etoposide phosphate	DMSO	200 $\mu$ M – 3.39nM	
	12	Clofarabine (CLOF)	H <sub>2</sub> O	25 $\mu$ M – 0.42nM	
	13	Hydroxyurea (HU)	H <sub>2</sub> O	10mM – 169nM	
	19	Exo 1	DMSO	300 $\mu$ M – 5.08nM	

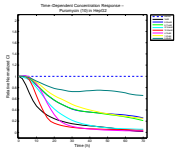
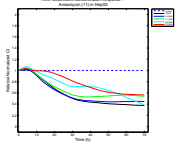
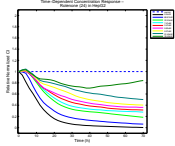
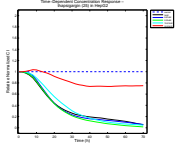
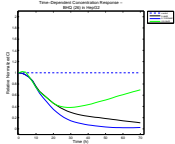
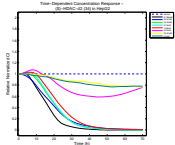
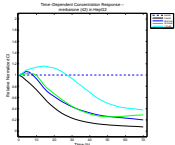
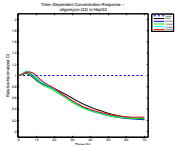
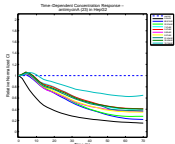
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	28	Cyclosporin A	DMSO	100 $\mu$ M – 1.69nM	
	33	SAHA	DMSO	151 $\mu$ M – 2.56nM	
	35	Mitoxantrone	DMSO	150 $\mu$ M – 2.54nM	
	36	Mitomycin C	DMSO	200 $\mu$ M – 3.39nM	
	39	Topotecan	DMSO	95 $\mu$ M – 1.61nM	
	43	Irinotecan (CPT-11)	DMSO	160 $\mu$ M – 2.71nM	
	46	Benzo[a]pyrene	DMSO	100 $\mu$ M – 1.69nM	
	60	SN-38	DMSO	200 $\mu$ M – 3.39nM	
3	7	Emetine	H <sub>2</sub> O	50 $\mu$ M – 0.847nM	

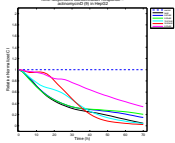
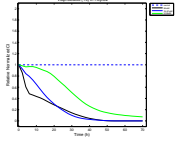
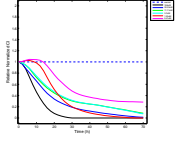
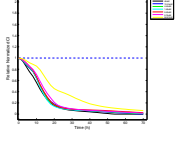
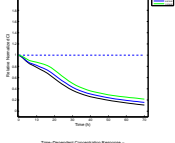
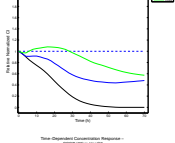
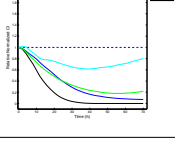
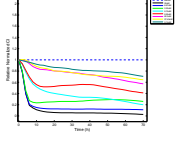
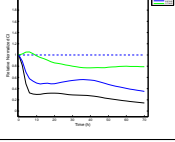
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	10	Puromycin	H <sub>2</sub> O	1000μM – 17nM	
	11	Anisomycin	H <sub>2</sub> O	10μM – 0.17nM	
	24	Rotenone	DMSO	200μM – 3.387nM	
	25	Thapsigargin	DMSO	2μM – 0.0339nM	
	26	BHQ	DMSO	400μM – 7nM	
	34	(S)-HDAC-42	DMSO	128μM – 2.17nM	
	42	Merbarone	DMSO	200μM – 3.39nM	
4	22	Oligomycin	DMSO	20μM – 0.339nM	
	23	Antimycin A	EtOH	200μM – 3.387nM	

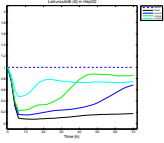
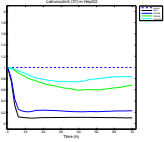
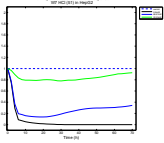
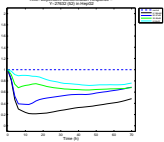
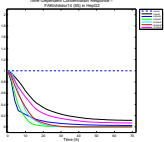
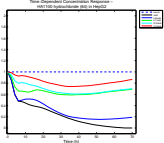
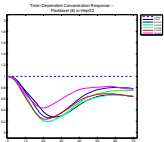
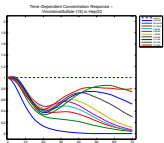
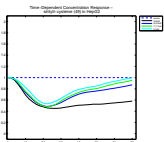
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
5	9	Actinomycin D	DMSO	$2\mu\text{M} - 0.0339\text{nM}$	
	14	Valproic acid	H <sub>2</sub> O	$50\text{mM} - 847\text{nM}$	
	16	Doxorubicin (DOX)	H <sub>2</sub> O	$100\mu\text{M} - 1.69\text{nM}$	
	17	Brefeldin A (BEF)	DMSO	$40\mu\text{M} - 0.68\text{nM}$	
	18	Leptomycin B (LMB)	EtOH	$20\text{nM} - 0.000339\text{nM}$	
	20	Monensin	DMSO	$4\mu\text{M} - 0.068\text{nM}$	
	32	CCCP	DMSO	$100\mu\text{M} - 1.69\text{nM}$	
6	4	Cytochalasin D	DMSO	$20\mu\text{M} - 0.339\text{nM}$	
	5	Cytochalasin B	DMSO	$20\mu\text{M} - 0.339\text{nM}$	

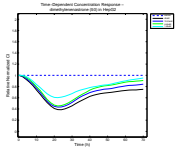
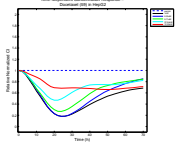
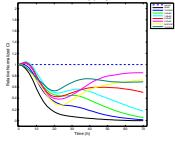
continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
	6	Latrunculin B	DMSO	20 $\mu$ M – 0.339nM	
	31	Latrunculin A	EtOH	2 $\mu$ M – 0.03nM	
	51	W7 HCl	DMSO	200 $\mu$ M – 3.39nM	
	52	Y-27632	DMSO	188 $\mu$ M – 3.18nM	
	55	FAKInhibitor14	H <sub>2</sub> O	2500 $\mu$ M – 42.34nM	
	64	HA1100 hydrochloride	H <sub>2</sub> O	1000 $\mu$ M – 16.94nM	
7	8	Paclitaxel	DMSO	20 $\mu$ M – 0.339nM	
	15	Vincristine Sulfate	H <sub>2</sub> O	250 $\mu$ M – 4.23nM	
	49	S-trityl-Cysteine	DMSO	100 $\mu$ M – 1.69nM	

continued on next page

continued from previous page

Cluster	SN	Chemical	Solution	Concentration (1:3)	TCRP
50		Dimethylenastron	DMSO	40 $\mu$ M – 0.68nM	
59		Docetaxel	DMSO	1 $\mu$ M – 0.02nM	
61		Vinblastine sulfate	H <sub>2</sub> O	40 $\mu$ M – 0.68nM	

The results shown in Table 5.3 are overall satisfactory, although some chemicals are misclassified according to the visualization of shapes. The main advantage in determining the cluster number using BIC lies in the fact that the structure of all the clusters including the clustering number is fully determined by the data. It can save human efforts when datasets are large. However, some expert knowledge is still required to verify the reliability of classification via biological interpretation.

### 5.5.6 Chemical classification GUI module development

A Graphic User Interface (GUI) module is designed according to the clustering approach proposed above. A user's guide that introduces the layouts and functions of the GUI is given in the Appendix.

## 5.6 Conclusion

This chapter introduced the proposed PCA and FDA based hierarchical classification. As effective and efficient strategies for feature extraction, PCA projected the data with uneven dimensionality into a lower-dimensional PC space

with the major variation of data retained. FDA used fewer coefficients to represent the extracted PC scores and removed noise to some extent simultaneously. The coefficients of PC scores were taken as valid features. A model-based hierarchical clustering algorithm was introduced and applied on the extracted feature matrices.

The application of the proposed approach on cell index data was elaborated in detail. To divide the dendrogram into some parallel and analyzable categories, a tree cutting was imposed. Based on users' interpretation, two tree cutting strategies, one-level and two-level dendrogram cutting, were introduced separately. One-level dendrogram cutting utilized two coefficient vectors concatenated together, while two-level dendrogram clustering processed the first coefficient of PC scores based on the TCRPs and the slopes of TCRPs in different levels. Although the clustering results indicated the success of discriminating chemicals from shape similarity point of views, the determination of clusters involved human interactions and was subjective. Also, the interpretation by users required expert knowledge. Hence, to solve this problem, model selection using Bayesian Information Criterion was introduced and applied to determine the cluster numbers automatically. The results indicated that the estimated clustering number is close to that generated via users' interpretation. A MATLAB<sup>®</sup>-based GUI was designed based on the approach proposed.

## Chapter 6

# Conclusions and future work

### 6.1 Conclusion

In this thesis, the aim is to classify chemicals into groups such that each group represents the same mode of action for all the chemical compounds. The shape of TCRPs is a good indication of Mode of Action. Several data based classification techniques were introduced to classify chemicals with similar TCRP shapes into the same group.

The main contribution of this thesis is to design and implement three different strategies to classify chemicals which are denoted by different numbers of TCRPs. TCRPs of chemicals were visualized and depicted with multiple time series. Therefore, the classification problem was interpreted by using dynamic information in profiles.

*Majority-voting and K-means clustering integrated classification* aimed at extracting first-order difference (slope) of profiles within each sampling interval to depict the dynamic tendency and shape of curves. With the help of majority voting, the feature vector of each chemical represents the major tendency of the curves. The advantages and disadvantages of this approach are concluded in Table 6.1:

*Hierarchical correlation based classification* focuses on measuring similarity



Table 6.1: Advantages and disadvantages of Majority-voting and  $K$ -means clustering integrated classification

Advantage	Disadvantage
1. evaluated chemicals using major tendency of TCRPs	1. conflicts in identifying feature vector
2. fully utilized slopes of TCRPs ( $\Delta RNCI$ )	2. not fully utilized information about $RNCI$
3. computationally efficient	3. predefined cluster number

between chemicals with different numbers of TCRPs using correlation coefficients. All the eligible chemicals are first divided into *positive* and *negative* groups. In order to classify chemicals from a curve shape similarity point of view, each TCRP is considered as a variable and each chemical is denoted by a vector of TCRPs. The significance of curve by curve correlation analysis is aimed to search for the pairs of curves from the compared chemicals with the maximum value  $R$  and thus decide whether two chemicals are combined or not.

Table 6.2: Advantages and disadvantages of Hierarchical correlation based classification

Advantage	Disadvantage
1. avoided unnecessary loss about profile information	1. empirical in tuning parameters
2. computationally efficient	
3. determined cluster number flexibly and automatically	

*PCA & FDA based hierarchical classification* aims at classifying chemicals denoted with different numbers of TCRPs from a relatively objective way. PCA projects data with nonuniform dimensionality into a lower-dimensional and unified PC space by selecting the principal components. FDA uses fewer coefficients to represent the extracted PC scores and remove noise at the same time. A model-based hierarchical clustering algorithm is applied on the extracted feature matrices. The number of clusters can be determined either by using prior knowledge manually or by using BIC criterion automatically.

Table 6.3: Advantages and disadvantages of PCA & FDA based hierarchical classification

Advantage	Disadvantage
1. synthesized profiles in a reasonable way	1. computationally time-consuming
2. classified chemicals in a maximum-likelihood framework	
3. determined cluster number manually and automatically	

## 6.2 Future work

Classification of chemicals denoted by different numbers of TCRPs is a challenging problem. The problem is considered as a classification of data objects. Each data object is defined as a matrix with a different number of columns. Usually, each data object is depicted by a variable or vector in  $\mathbb{R}^N$ . When data objects become complicated and multi-dimensional, they can be represented with matrices. However, classification on data objects denoted by matrices itself is not easy because the target feature matrix will become cubic. It is even harder to handle data objects denoted by matrices with nonuniform dimensionality. So, the key component to realize pattern recognition of data with nonuniform dimensionality is to extract valid and functional feature vectors with the information in the original data object compressed and integrated as much as possible.

Some works worthy of further investigation are listed as follows:

### 1. Improvement on feature extraction

Feature extraction is crucial for classification. This procedure integrates important information about original data and decreases the calculation time at the same time. However, fusing the information and rescaling the data object remains a challenge. Apart from majority-voting and *PCA* and *FDA* methods, some other ways to extract features can be considered such as adding different weights on profiles with different concentrations.

### 2. Similarity measurement

Calculating the degree of similarity between time series is also crucial

for classification especially for curve shape based classification. Although distance-based similarity is commonly used, it is constrained by vector dimensionality. Distance functions may lose their usefulness in high dimensionality [41]. A better way to quantify the similarity between vector data is needed.

### 3. Robust classification of chemicals using dose-response relationship

Dose-response curves are considered as a good way to realize a more robust classification of chemicals as dose-response curves are aimed to establish the relationship between toxicity index (e.g.  $LC_{50}$  [42]) and concentrations [43]. Therefore, for each concentration, dynamic information in the TCRPs is compressed and integrated using the toxicity index. Experimental noise which affects the dynamics of the TCRPs can be alleviated by introducing the toxicity index. Classification of chemicals denoted by TCRPs is thus addressed by switching to the classification of chemicals based on dose-response profiles.

# Bibliography

- [1] S. Khatibisepehr, B. Huang, F. Ibrahim, J.Z. Xing, and W. Roa. Data-based modeling and prediction of cytotoxicity induced by contaminants in water resources. *Computational Biology and Chemistry*, 35(2):69–80, 2011.
- [2] T. Pan, B. Huang, W. Zhang, S. Gabos, D.Y. Huang, and V. Devendran. Cytotoxicity assessment based on the  $AUC_{50}$  using multi-concentration time-dependent cellular response curves. *Analytica Chimica Acta*, 764:44–52, 2013.
- [3] K.A. Houck and R.J. Kavlock. Understanding mechanisms of toxicity: Insights from drug discovery research. *Toxicology and Applied Pharmacology*, 227(2):163–178, 2008.
- [4] National Research Council Committee on Methods of Producing Monoclonal Antibodies, Institute for Laboratory Animal Research. *Monoclonal Antibody Production*. The National Academies Press, 1999.
- [5] H. Hosseinkhani, M. Hosseinkhani, and A. Khademhosseini. Emerging applications of hydrogels and microscale technologies in drug discovery. *Drug Discovery*, 1:32–34, 2006.
- [6] S. Przyborski. Supporting cells with scaffold technology. *Genetic Engineering and Biotechnology News*, 31(16):38–39, 2011.

- [7] J.E. González, K. Oades, Y. Leychkis, A. Harootunian, and P.A. Negulescu. Cell-based assays and instrumentation for screening ion-channel targets. *Drug Discovery Today*, 4(9):431–439, 1999.
- [8] L. Silverman, R. Campbell, and J.R. Broach. New assay technologies for high-throughput screening. *Current Opinion in Chemical Biology*, 2(3):397–403, 1998.
- [9] National Research Council (US), Committee on Toxicity Testing and Assessment of Environmental Agents. *Toxicity Testing In the 21st Century: a Vision and a Strategy*. National Academies Press, 2007.
- [10] R. Judson, F. Elloumi, R.W. Setzer, Z. Li, and I. Shah. A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinformatics*, 9(1):241, 2008.
- [11] V. Dellarco and P.A. Fenner-Crisp. Mode of action: moving toward a more relevant and efficient assessment paradigm. *Journal of Nutrition*, 142(12):2192S–2198S, 2012.
- [12] U.S. EPA. Guidelines for carcinogen risk assessment. U.S. Environmental Protection Agency, Washington, DC, 2005. EPA/630/P-03/001F, 2005.
- [13] B.I. Escher and J.L. Hermens. Modes of action in ecotoxicology: their role in body burdens, species sensitivity, qsars, and mixture effects. *Environmental Science and Technology*, 36(20):4201–4217, 2002.
- [14] N.M. Luscombe, D. Greenbaum, and M. Gerstein. What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*, 1:83–99, 2001.
- [15] N. Ke, X. Wang, X. Xu, and Y.A. Abassi. The xcelligence system for real-time and label-free monitoring of cell viability. In *Mammalian Cell*

- Viability: Methods and Protocols*, volume 740 of *Methods in Molecular Biology*, pages 33–43. May 2011.
- [16] Y.A. Abassi, B. Xi, W. Zhang, P. Ye, S.L. Kirstein, M.R. Gaylord, S.C. Feinstein, X. Wang, and X. Xu. Kinetic cell-based morphological screening: prediction of mechanism of compound action and off-target effects. *Chemical Biology*, 16(7):712–723, 2009.
  - [17] H. Slanina, A. Knig, H. Claus, M. Frosch, and A. Schubert-Unkmeir. Real-time impedance analysis of host cell response to meningococcal infection. *Journal of Microbiological Methods*, 84(1):101 – 108, 2011.
  - [18] J.Z. Xing, L. Zhu, J.A. Jackson, S. Gabos, X.J. Sun, X. Wang, and X. Xu. Dynamic monitoring of cytotoxicity on microelectronic sensors. *Chemical Research in Toxicology*, 18(2):154–161, 2005.
  - [19] B. Huang and J.Z. Xing. Dynamic modelling and prediction of cytotoxicity on microelectronic cell sensor array. *The Canadian Journal of Chemical Engineering*, 84(4):393–405, 2006.
  - [20] T. Pan and B. Huang. MOA classification based on the integer variables. November 2012. Technical Report for Alberta Health, Edmonton, AB, Canada.
  - [21] C.M. Bishop and N.M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.
  - [22] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*. Citeseer, 1995.
  - [23] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1):37–52, 1987.

- [24] G.H. Golub and C.F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.
- [25] J.O. Ramsay, G. Hooker, and S. Graves. *Functional data analysis with R and MATLAB*. Springer, 2009.
- [26] J.J. Song, H.J. Lee, J.S. Morris, and S. Kang. Clustering of time-course gene expression data using functional data analysis. *Computational biology and chemistry*, 31(4):265–274, 2007.
- [27] J.O. Ramsay and B.W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer New York, 2002.
- [28] N. Coffey and J. Hinde. Analyzing time-course microarray data using functional data analysis-a review. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.
- [29] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.
- [30] A.R. Martinez and W.L. Martinez. Model-based clustering toolbox for MATLAB. *Naval Surface Warfare Center, Dahlgren Division, Tech. Rep*, 2004.
- [31] C. Fraley. Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281, 1998.
- [32] A. Dittrich and H. Schmeck. *Given’s rotation on an instruction systolic array*. Springer, 1989.
- [33] K.P. Burnham and D.R. Anderson. Multimodel inference understanding AIC and BIC in model selection. *Sociological methods and research*, 33(2):261–304, 2004.

- [34] B.I. Escher, R. Ashauer, S. Dyer, J.L. Hermens, J.H. Lee, H.A. Leslie, P. Mayer, J.P. Meador, and M.S. Warne. Crucial role of mechanisms and modes of toxic action for understanding tissue residue toxicity and internal effect concentrations of organic chemicals. *Integrated environmental assessment and management*, 7(1):28–49, 2011.
- [35] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584. IEEE, 2004.
- [36] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- [37] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.
- [38] C. Fraley and A.E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.
- [39] H.P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178, 1967.
- [40] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397, 1971.
- [41] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Database Theory – ICDT99*, pages 217–235. Springer, 1999.
- [42] C.E. Stephan. Methods for calculating an LC50. *Aquatic Toxicology and Hazard Evaluation*, 1:65–84, 1977.



- [43] K.S. Crump, D.G. Hoel, C.H. Langley, and R. Peto. Fundamental carcinogenic processes and their implications for low dose risk assessment. *Cancer Research*, 36(9 Part 1):2973–2979, 1976.

# Appendix

This user's guide document will illustrate how to use the GUI module designed for Mode of Action Classification.

## Interface Introduction

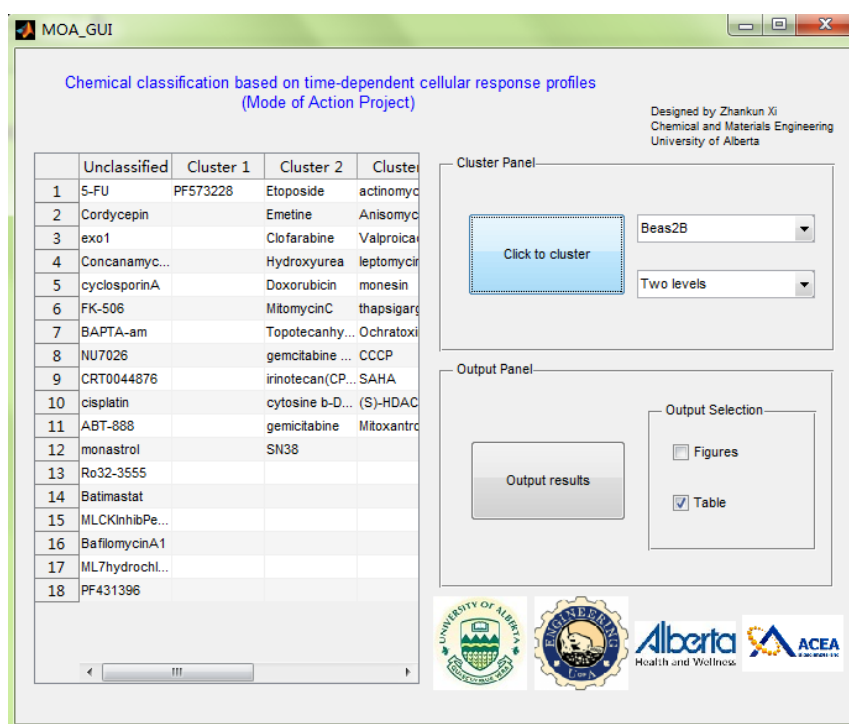


Figure 6.1: Main graphic user interface (GUI) of Mode of Action Classification

Figure 6.1 shows the main interface of Mode of Action Classification (MoA Classification). It is designed by Zhankun Xi, a master candidate in Com-

puter Process Control, Department of Chemical and Materials Engineering, University of Alberta. The GUI module is developed tentatively and more functions are under consideration. If there is any suggestion, please contact [zhankun@ualberta.ca](mailto:zhankun@ualberta.ca).

## Result Table

In the left hand side, a **uitable** displays the pattern recognition results for the data selected to be clustered. Chemicals are arranged cluster by cluster. The headers represent the category which chemicals belong to. Chemicals in *Unclassified* category are not eligible for classification due to their TCRPs are close to control lines, which do not show apparent pattern for analysis.

## Cluster Panel

**Cluster Panel** is used to control the input parameters to the GUI.

### Cell line selection

**Cell line selection** is the upper pop-up menu at the top right corner. It is used to select the cell line based on which the users classify the chemicals. One of five default cell lines in the database, ACHN (Tissue: kidney), ARPE19 (Tissue: retinal pigmented epithelium; retina), HepG2 (Tissue: liver), Beas-2B (Tissue: lung/bronchus) and H4 (Tissue: brain) can be selected at one time. Other new cell lines will be added into the database later if available. All the data files are formatted into *.mat* which can be directly recognized by MATLAB<sup>®</sup>.

### Clustering mode selection

**Clustering mode selection** is the lower pop-up menu at the top right corner. It is used to choose the clustering mode for the selected cell line. There are

two default modes available, *one level* and *two levels*. Although the designed algorithm belongs to hierarchical clustering, the difference between two modes lie in whether there is a uniform tree cutting or not in hierarchical clustering. The dendrogram generated from the algorithm is cut at one time in a one-level scenario, while it is cut level by level in a two-level one. In other words, the dendrogram is firstly cut into several subtrees whose volumes are relatively big. Then, each subtree is cut again according to its actual distribution. The main advantage of two-level hierarchical clustering scenario is that it can manipulate the number of clusters easily under each subtree and hence the overall number of clusters for all the input data.

#### **Pushbutton: Click to cluster**

Other than the two pop-up menus in the **Cluster Panel**, there is a push button, **Click to cluster**, to implement an approach utilizing multivariate statistical techniques including principal component analysis (PCA) and functional data analysis (FDA) and a model-based agglomerative hierarchical clustering algorithm to cluster the input chemicals using TCRPs. Via the push button, one or two dendrograms are generated according to users' choice in the lower pop-up menu. Users can cut the trees according to their satisfaction. Usually, for classification based on TCRPs in our study, a suitable number of clusters is around 10. The clustering results may vary due to different interpretation and understanding about the clusters. However, the goal of this project is to categorize chemicals whose TCRPs show similar profile shapes and tendency. Chemicals that show similar profile shapes and tendency may show similar Mode of Action if the experiment is done under appropriate doses or concentration levels.

## Example: Cluster Panel

We consider a two-level hierarchical clustering scenario as an example. When “Beas-2B” cell line and two-level mode are chosen separately, a dendrogram together with a dialogue to prompt the users to type in a number is generated after the button, **click to cluster**, is clicked. Users can cut the dendrogram according to the distribution under itself, which is indicated as in Figure 6.2. After users enter  $c$  as a number, another figure will pop out to ask the users to input the cluster numbers under each of  $c$  subtrees. The hierarchical distributions are illustrated in Figure 6.3. The prompted dialogue for the users to input appropriate cluster numbers is shown in Figure 6.4.

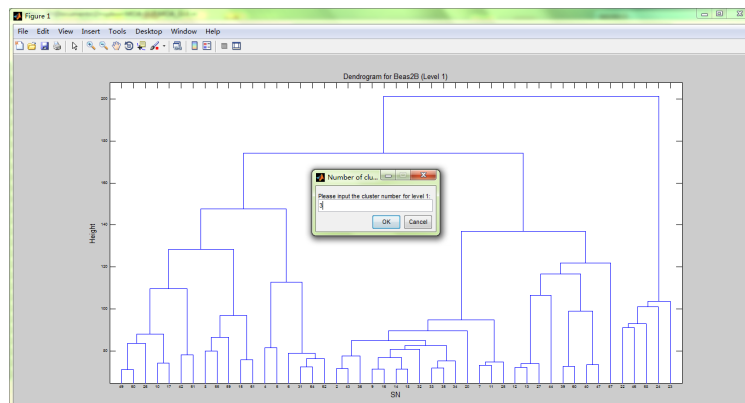


Figure 6.2: First level dendrogram

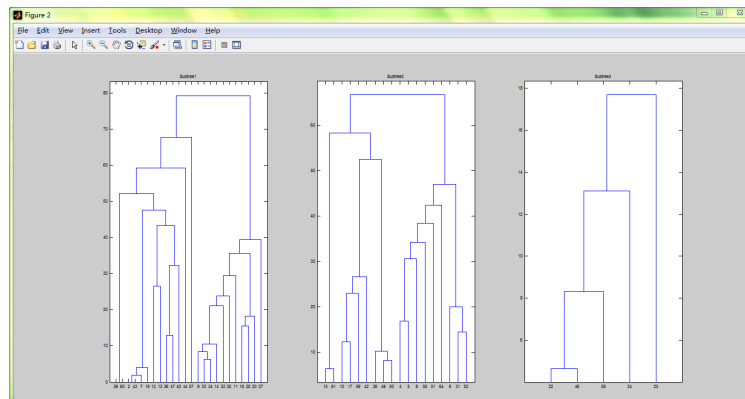


Figure 6.3: Second level dendrogram

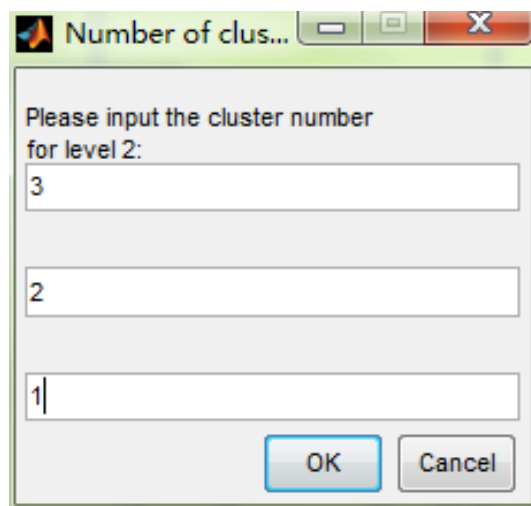


Figure 6.4: A prompted dialogue for the users to input appropriate cluster numbers

## Output Panel

**Output Panel** is used to control the outputs from the GUI.

### Pushbutton: Output results

In **Output panel**, **Output results** is a push button which can output the clustering results once the execution of clustering is finished.

### Output selection

The subpanel, **Output selection** ask users to choose the types of outputs: Table, Figures or both. When Table is ticked, an EXCEL file whose contents are the same as those indicated in the uitable module in MoA GUI can be generated and saved in “\Excel Result” folder when pushbutton **Output results** is clicked. Users can copy and paste the contents for their further study. When Figures is ticked, all eligible TCRPs and their corresponding concentration information will be saved into *.bmp* format in “\Figures\_Cellline\_65Chemicals” folder, where Cellline is a variable according to the cell line which the users choose in **Cluster Panel**.

Example: Table

1	A	B	C	D	E	F	G
2	unclassified	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
3	1-5-FU	57-PF573228	2-Etoposide	9-actinomycinD	4-CytochalasinD	10-Puromycin	22-oligomycin
4	3-Cordycepin		7-Emetine	11-Anisomycin	5-CytochalasinB	15-VincristineSulfate	23-antimycinA
5	19-exo1		12-Clofarabine	14-Valproicacid	6-LatrunculinB	17-brefeldinA	24-Rotenone
6	21-ConcanamycinA		13-Hydroxyurea	18-leptomycinb	8-Paclitaxel	26-BHQ	46-benzo[a]pyrene
7	28-cyclosporinA		16-Doxorubicin	20-monesin	31-LatrunculinA	42-merbarone	58-Blebbistatin
8	29-FK-506		36-MitomycinC	25-thapsigargin	51-W7 HCl	49-strityltl-cysteine	
9	30-BAPTA-am		39-Topotecanhydrochloride	27-OchratoxinA	52-Y-27632	50-dimethylenenastrone	
10	37-NU7026		40-gemcitabine HCl	32-CCCP	55-FAKInhibitor14	59-Docetaxel	
11	38-CRT0044876		43-irinotecan(CPT-11)	33-SAHA	64-HA1100hydrochloride	61-Vinblastinesulfate	
12	41-cisplatin		44-cytosine b-D-arabinofuranoside	34-(S)-HDAC-42			
13	45-ABT-888		47-gemcitabine	35-Mitoxantronedihydrochloride			
14	48-monastral		60-SN38				
15	53-Ro32-3555						
16	54-Batimastat						
17	56-MLCKInhibPep18						
18	62-BafilomycinA1						
19	63-ML7hydrochloride						
20	65-PF431396						
21							
22							
23							

Figure 6.5: Table saved via the GUI

Example: Figures

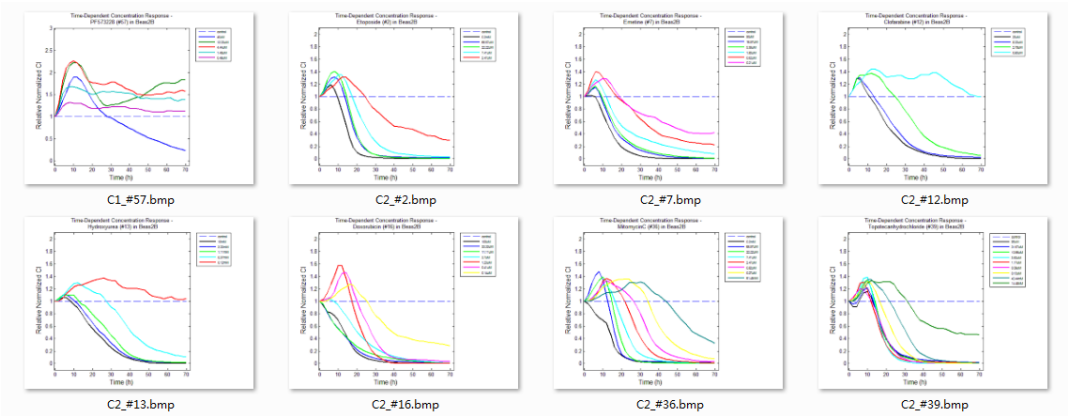


Figure 6.6: TCRP figures saved via the GUI