DEVELOPMENT OF NOVEL UNSUPERVISED AND SUPERVISED INFORMATICS METHODS FOR DRUG DISCOVERY APPLICATIONS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctorate of Philosophy in the

Graduate School of The Ohio State University

By

Syed Basha Mohiddin, M.A.S.

The Ohio State University
2006

Dissertation Committee
Professor James F. Rathman, Adviser
Professor Bhavik R. Bakshi
Professor Jeffrey J. Chalmers

Approved by

Adviser
Graduate Program in Chemical Engineering
ABSTRACT

As of 2002, the cost of discovering a new drug was nearly $802 million with a timeline of nearly 13.6 years. Despite the large investments in time and money, drugs that were successfully introduced in the market had to be withdrawn later due to efficacy (38%) and safety (20%) reasons. Improving the success rate in drug discovery is linked with two key steps in the process. First, in order to improve efficacy, there is a need for improved understanding of genetic biomarkers (targets for drug action) that are responsible for characterizing a given disease. Second, it is possible to improve drug safety, by predicting the activity/toxicity of potential drug candidates at an early stage prior to the initiation of expensive clinical trials. In this work, we develop a novel unsupervised informatics methodology that addresses characterization of both biological and chemical samples and identification of underlying key non-redundant features responsible for characterization. Biological samples are characterized into different groups (e.g. cancer types) based on gene expression profiling and the genetic biomarkers most responsible for characterization are identified. Similarly, chemical compounds are characterized into different groups with varying activity/toxicity based on structural, physical and chemical property data of the chemical compounds.
The methodology developed in this work relies largely on the multivariate aspects of principal component analysis and the application of \( k \) means clustering algorithm in a hierarchically recursive manner to achieve unsupervised multi-class classification. The principal components are replaced by the corresponding partial least square (PLS) components in the supervised scenario. Selection of influential components (principal components in unsupervised case and PLS components in supervised case) for the purpose of classification is demonstrated and is one of the key steps for the success of this methodology. Hierarchical \( k \)-means is applied recursively to achieve binary classification at each stage eventually resulting in multi-class classification. Identification of features responsible for classification is achieved by examining the appropriate loadings of the principal or PLS components along with their coefficient of correlation with influential components.
Dedicated to The One God, my family, teachers and friends.
ACKNOWLEDGMENTS

I would like to thank my adviser, Dr. James Rathman for his intellectual guidance and support throughout my graduate study and research, and for his unfailing encouragement.

I am grateful to Dr. Chihae Yang and Dr. Glen Myatt for introducing me to the world of drug discovery and for their stimulating lectures that sparked my interest and added to my understanding.

I also wish to thank my group members for their support and the entire faculty for their valuable lectures and discussions that made my education a truly rich experience.

This research was supported by the State of Ohio Technology Action Fund, Leadscope, Inc, Columbus, Ohio and the Department of Chemical & Biomolecular Engineering here at The Ohio State University.
VITA

March 23, 1978 ............................. Born — Vijayawada, AP, India

1996 – 2000 ............................... B.S. Chemical Engineering
Birla Institute of Technology & Sciences
Pilani, RJ, India

2000 – 2001 ............................... Programmer Analyst
Cognizant Corporation
Chennai, India

2001 – 2002............................... University Fellow
The Ohio State University

The Ohio State University

2002 – Present............................ Graduate Research Associate
Department of Chemical Engineering
The Ohio State University

FIELDS OF STUDY

Major Field: Chemical Engineering
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
</tbody>
</table>

## Chapters:

1. **Introduction** .................................................................................. 1  
   1.1 Drug discovery process overview ............................................. 1  
   1.2 Drug discovery informatics ...................................................... 4  
   1.3 Scope and objectives of research ............................................. 8  
   1.4 Outline of thesis document ..................................................... 10  

2. **Literature review** ........................................................................... 12  
   2.1 Molecular characterization of biological samples ........................ 12  
   2.1.1 Microarray and GeneChip technology ....................................... 14  
   2.1.2 Cancer classification ............................................................ 26  
   2.1.3 Existing techniques .............................................................. 28  
   2.2 Genetic biomarker identification .............................................. 32  
   2.3 Characterization of chemical compounds .................................... 35  
   2.3.1 Quantitative structure activity relationship ............................ 37  
   2.3.2 Chemical compound classification .......................................... 40  
   2.3.3 Current methodologies .......................................................... 42  
   2.4 Identification of active chemical features .................................. 44  
   2.5 Conclusions .................................................................................. 45
3. Systems and Methods..................................................................................47

3.1 Introduction..............................................................................................47
3.2 Unsupervised classification methodology.............................................49
  3.2.1 Principal component analysis and selection........................................50
  3.2.2 Hierarchical $K$-means clustering using influential principal
        components.........................................................................................58
  3.2.3 Identification of key features.............................................................63
3.3 Supervised classification methodology..................................................66
  3.3.1 PLS component analysis and selection.................................................67
  3.3.2 Hierarchical $K$-means clustering of PLS components.........................72
  3.3.3 Identification of key features.............................................................73
3.4 Conclusions.............................................................................................73

4. Classification of cancer..............................................................................74

4.1 Introduction..............................................................................................74
4.2 Unsupervised analysis: Leukemia data set..............................................74
  4.2.1 Principal component analysis and selection.........................................76
  4.2.2 Classification results and discussion.....................................................83
  4.2.3 Identification of differentiating genes................................................92
4.3 Unsupervised analysis: NCI 60 data set..................................................99
4.4 Supervised analysis: Leukemia data set................................................106
4.5 Conclusions.............................................................................................117

5. Classification of chemical compounds....................................................118

5.1 Introduction..............................................................................................118
5.2 Unsupervised analysis: Mutagenicity data set......................................120
5.3 Supervised analysis: Mutagenicity data set..........................................127
5.4 Conclusions.............................................................................................132

6. Prediction of octanol-water partition coefficient....................................134

6.1 Introduction..............................................................................................134
6.2 Existing Methods.....................................................................................135
6.3 Non-ideal solution activity coefficient approach..................................138

7. Conclusions and Recommendations.......................................................147

7.1 Conclusions.............................................................................................147
7.2 Impact of current research......................................................................149
7.3 Future work and recommendations.....................................................151
Appendices:

A. Matlab programs for gene expression studies........................................155
B. Matlab programs for classifying chemical compounds..........................159
C. Matlab programs for predicting $\log K_{OW}$............................................162

List of References.........................................................................................172
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Percent variability accounted by the corresponding PC and the proportion of negative coefficients for the first 10 eigenvectors in the case of ALL/AML distinction</td>
</tr>
<tr>
<td>4.2</td>
<td>Prediction results for ALL/AML classification using unsupervised method</td>
</tr>
<tr>
<td>4.3</td>
<td>Leave-one-out cross validation prediction results for ALL/AML classification using unsupervised method</td>
</tr>
<tr>
<td>4.4</td>
<td>Prediction results for B-cell ALL/T-cell ALL class distinction within the 47 ALL samples using unsupervised method The five misclassified samples are shown in diamonds</td>
</tr>
<tr>
<td>4.5</td>
<td>Prediction results for ALL/AML classification after the inclusion of $\textbf{PC}_1$ in the predictive subset of PCs</td>
</tr>
<tr>
<td>4.6</td>
<td>Prediction results for B-cell ALL/T-cell ALL classification after the inclusion of $\textbf{PC}_1$ in the predictive subset of PCs</td>
</tr>
<tr>
<td>4.7</td>
<td>A partial list of 20 most differentiating genes identified through overall correlation rank method for ALL/AML classification</td>
</tr>
<tr>
<td>4.8</td>
<td>A partial list of 20 most differentiating genes identified through feature loading method for ALL/AML classification</td>
</tr>
<tr>
<td>4.9</td>
<td>A list of first 10 PLS components and their corresponding percent variation accounted in predictor ($\textbf{X}$) and response ($\textbf{Y}$) variables for the acute leukemia data set</td>
</tr>
</tbody>
</table>
4.10 Prediction results for ALL/AML classification using all the first 10 PLS components as predictors in the supervised method……………………………111

4.11 Prediction results for ALL/AML classification with PLS components #3 and #4 excluded from the first 10 PLS components as predictors in the supervised method…………………………………………………………111

4.12 A list of 9 PLS components and their corresponding percent variation accounted in predictor (X) and response (Y) variables for the 38 samples in the training set of acute leukemia data………………………………………………………….113

4.13 Supervised prediction results of ALL/AML classification for a test set of 34 samples using PLS components derived from a training set of 38 samples……116

4.14 Supervised prediction results of ALL/AML classification for 20 random runs on a test set of 34 samples using PLS components derived from a training set of 38 samples………………………………………………………………………….116

5.1 (a) Percent variability accounted by the first 10 PCs; (b) Cumulative percent variability accounted by first 150 PCs…………………………………………………………122

5.2 Unsupervised prediction results for mutagenicity prediction using FRAG descriptors with the inclusion of PC₁ and PC₂ in the predictive subset of PCs……………………………………………………………………………124

5.3 Supervised prediction results for mutagenicity prediction using Leadscope® descriptors………………………………………………………………………………130
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The central dogma of molecular biology</td>
</tr>
<tr>
<td>2.2</td>
<td>Example schematic of nucleotide hybridization</td>
</tr>
<tr>
<td>2.3</td>
<td>Example schematic of Affymetrix oligonucleotide GeneChip</td>
</tr>
<tr>
<td>2.4</td>
<td>Schematic of hybridization step in an Affymetrix GeneChip</td>
</tr>
<tr>
<td>2.5</td>
<td>Schematic of gene expression measurement in an Affymetrix chip</td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic of eigenvector distributions: (a) good predictor for classification purpose (b) poor predictor for classification purpose</td>
</tr>
<tr>
<td>3.2</td>
<td>Cartoon illustration of $k$-means clustering process: (a) 15 data points in 2-dimensional space; (b) selection of random seeds; (c) assignment of data points to one of the two clusters; (d) re-assignment of data points</td>
</tr>
<tr>
<td>3.3</td>
<td>Schematic of hierarchical $K$-means clustering</td>
</tr>
<tr>
<td>4.1</td>
<td>A plot of cumulative percent variability accounted vs. the number of principal components in the leukemia data set for ALL/AML distinction</td>
</tr>
<tr>
<td>4.2</td>
<td>A plot showing proportion of negative coefficients for the first 10 eigenvectors vs. the percent variability accounted by their corresponding principal components</td>
</tr>
<tr>
<td>4.3</td>
<td>A plot showing the distribution of eigenvector coefficients for eigenvectors 1 and 2 in the ALL/AML case study</td>
</tr>
</tbody>
</table>
4.4 A plot showing prediction results for unsupervised classification of ALL/AML tumor samples in the dimensional space of $\text{PC}_2$ and $\text{PC}_4$.................................84

4.5 A plot showing proportion of negative coefficients for the first 10 eigenvector vs. the percent variability accounted by their corresponding principal components for the subset of 47 ALL samples.............................................................88

4.6 A plot showing proportion of negative coefficients for the first 10 eigenvector vs. the percent variability accounted by their corresponding principal components for the NCI 60 data set.................................................................101

4.7 Partitioning of the NCI 60 cell lines using recursive $k$-means clustering process........................................................................................................103

4.8 Flowchart depicting our unsupervised methodology of recursive $k$-means clustering using predictive subset of PCs..........................................................105

4.9 A plot showing proportion of negative coefficients for the first 10 weight vectors corresponding to the PLS components in the supervised analysis of leukemia data...........................................................................109

4.10 A plot showing proportion of negative coefficients for the first 9 weight vectors corresponding to the PLS components of the 38 samples in the training set of acute leukemia data.................................................................114

5.1 A plot showing proportion of negative coefficients for the first 10 eigenvector vs. the percent variability accounted by their corresponding principal components for the mutagenicity data set using FRAG descriptors..................................................123

5.2 A plot showing proportion of negative coefficients for the first 25 eigenvector vs. the percent variability accounted by their corresponding principal components for the mutagenicity data set using Leadscope® descriptors..........................126

5.3 A plot showing proportion of negative coefficients for the first 10 weight vector vs. the percent variability accounted by their corresponding PLS components for the mutagenicity data set using FRAG descriptors.........................................128

5.4 A plot showing proportion of negative coefficients for the first 15 weight vector vs. the percent variability accounted by their corresponding PLS components for the mutagenicity data set using Leadscope® descriptors.................................131
CHAPTER 1

INTRODUCTION

1.1 Drug Discovery Process Overview

Traditional drug discovery process consists of several steps including disease selection, biomarker targeting, lead compound identification and optimization, preclinical and clinical phases and the final step of pharmacogenomic optimization. Most diseases are caused by imbalance of particular biochemicals in the body, invasion of foreign matter or aberrant cell growth. The selection of a disease with the aim of pursuing therapeutic compounds is followed by the search for biological targets. A biological target (biomarker) is a molecule that is believed to be the cause of the disease state. Usually these molecules are enzymes or receptors, on which pharmaceutical researchers focus when designing a drug. Often, the target molecule is from an external entity such as a virus or a bacterium. Sometimes, a target may be an abnormal human protein inside the body. The main objective of traditional drug discovery is to design a small molecule, a drug, which can interact with the target molecule in a manner such that it rectifies the disease state. In other words, a drug is any substance presented for treating or preventing
disease in human beings or in animals. A drug may also be used for medical diagnosis or for modifying physiological functions.

A drug molecule interacts with the target and modulates its biological actions in different ways. Three most common types of interactions involve inhibitors, agonists and antagonists. A molecule that inhibits or blocks the biological action of another molecule is called an inhibitor. Inhibitors can be competitive and non-competitive. A molecule that binds to a receptor of a cell and triggers a response by the cell is called an agonist. An agonist often mimics the action of a naturally occurring substrate. A full agonist results in maximal cellular response while a partial agonist is unable to induce the maximal action. A molecule that binds to the agonist receptor and inhibits the action is called an antagonist. Antagonists are further classified as competitive, reversible and non-competitive agents. It is unlikely to find molecules that fit in the exact descriptions above that can be used as drug molecules readily. Most of the molecules need to be modified and transformed into a clinically useful drug. The chemical entities that serve as the starting point for such transformations are called leads or lead compounds.

Leads are new chemical entities which possess the desired biological activity, but in a non-optimized fashion. Discovery of such new molecules is called lead identification. The systematic modification of a biologically active compound (lead) to fulfill all the stereo-electronic, physico-chemical, pharmacokinetic and toxicological properties required for clinical usefulness is called lead optimization. The goal of lead optimization is to increase the efficacy and decrease the toxicity of the drug at the same time. More specifically, efficacy and toxicity are measured in terms of absorption, distribution, metabolism, elimination and toxicity (ADMET) abilities expressed by the
drug. Absorption is passage of the molecule through the physiological membranes via either active or passive transport. Distribution of drugs deals with pharmacokinetic issues after dosing the drug into the targeted tissue. Metabolism comprises the entire physical and chemical processes involved in breaking down of drugs to give simpler molecules (catabolism) which by themselves may be used to form more complex molecules (anabolism). Elimination is the process of achieving the reduction of concentration of the drug in the human body. Toxicity is the degree to which a chemical substance elicits a deleterious effect upon a biological system over a specified time period.

Pre-clinical and clinical trials are research studies to answer specific questions about the safety and efficacy of the new drugs in development. Phase 1 trials include the initial introduction of an investigational drug into humans. These studies are carried out with healthy volunteers and are designed to determine the metabolic and pharmalogical actions of the drug in humans, the side effects associated with increasing doses in order to establish a safe dose range and to gain early evidence of effectiveness. The total number of subjects involved in phase 1 trials is generally in the range of 20-80. Phase 2 trials are controlled clinical studies to evaluate the drug’s effectiveness for a particular indication in patients with the disease or condition under investigation. At this stage, the common short-term side effects and risks associated with the use of new drugs are assessed within a population of no more than several hundred subjects. Phase 3 trials involve the administration of a new drug to a large number of patients, usually several hundred to thousand subjects in different clinical settings in order to establish safety, efficacy and appropriate dosage. In this stage, the drug is administered the same way it would be administered when marketed. Phase 4 trials run concurrent with marketing approval and
include studying the use of drug in other patient populations or other stages of the disease or use of drug over longer periods.\textsuperscript{5}

It is evident that the process of drug discovery is not only a long and expensive process involving human trials but also a process that incorporates a high level of risk and associated cost of failure. In early times, the drug discovery procedure was to synthesize compounds more or less at random and to see if they showed the desired properties.\textsuperscript{6} However, in view of the ever-increasing number of chemical compounds and stringent demands to be met by these new chemicals, the random and empirical selection method is no longer effective. The cost of drug discovery has sharply risen from $54 million in the year 1976 to $802 million in the year 2002 with an average time span of 13.6 years to discover a new drug.\textsuperscript{7} Nevertheless, the problem of drug safety and efficacy still persists with 20% of the drugs in the past being terminated for safety reasons while 38% being terminated for lack of efficacy. Thus, the fundamental objective of drug discovery is to develop methodologies that find efficacious compounds with maximal safety in shortest time and minimal cost.

1.2 Drug Discovery Informatics

The development of new technologies during the first half of 20th century has had a profound effect in understanding and adopting rational approaches to drug design and development. Chemistry, pharmacology, microbiology and biochemistry played the key roles in changing the shape of drug discovery from a random selection of compounds to a more rational approach that involves understanding of biological structure, biochemical
mechanisms and underlying functions to create novel chemical entities. Molecular biology especially influenced drug discovery by quantifying the concept of genetic information and making it more concrete in biochemical and chemical terms. The main contribution of molecular biology towards drug discovery lies in the potential to understand disease processes at the molecular (genetic) level and to determine the optimal molecular targets for therapeutic intervention. It is estimated that the current drug therapy is based on less than 500 molecular targets while the potential drug targets may lie between 5000 and 10,000. Therefore, there are at least 10 times as many molecular targets to be exploited for future therapies than are being used as of today.

The advances in genomic studies coupled with high throughput experimentation have produced huge amounts of data that requires to be transformed into useful information. A large amount of research effort is directed towards developing informatics methodologies that glean knowledge from the large raw data sets either produced as a result of high throughput experimentation or available historically over a long period of studies. In recent literature, it has been shown that molecular targets identified using such informatics methods can be associated with certain diseases – for example, leptin or the leptin receptor with obesity, the low density lipoprotein receptor with atherosclerosis, complement receptors with inflammation, interleukin-4 (IL-4) with allergic diseases and so on. The credibility of identified targets has been demonstrated by using animal models in which target modifications led to the desired phenotypic changes. This reversal of disease symptoms in a clinical situation by the modification of a target (e.g., the blocking of a receptor or the inhibition of an enzyme) provides the highest degree of validation.
Microarray technologies shifted the scale of genomic research as they facilitated the study of several thousand genes in a single experiment. DNA microarrays measure the relative amounts of mRNA in isolated cells or biopsied tissues from patients. As transcriptional changes accurately reflect the status of many diseased states including cancer, it is believed that gene expression profiles can be used to classify different types of cancer. These experiments generated large amounts of data from few experimental runs and therefore posed a challenge for understanding and analyzing the information hidden underneath the flood of data. Under the premise that gene expression patterns act as unique molecular signatures characterizing a biological state, systematic methods to classify cancer types have been developed by various groups. Some of the techniques include Golub’s neighborhood analysis, support vector machines and self-organizing maps. After classifying the samples based on differences in genetic profiles, those genes that are most responsible for the observed classifications are identified using a variety of informatics methods. The genes identified in this manner provide insight into the possible causes of diseased states and are indicative of potential biological targets that can be explored for therapeutic intervention. The above methodologies fall under the broad area of bioinformatics and are primarily used at the early stage of target identification in the long process of drug discovery.

In the long drug discovery cycle, there also comes a stage where large amounts of chemical data are generated from novel technologies like automated synthesis, combinatorial chemistry and high-throughput screening (HTS). The objective of automated synthesis and combinatorial chemistry is to design compound libraries for HTS. HTS identifies lead compounds that exhibit activity or potency against a biological
target of interest. The resulting data is used to establish a predictive relationship between the properties (structural, physical and chemical) of lead compounds and their activity. The data sets generated at this stage are often similar to those that are presented at the stage of target identification (gene expression studies). In combinatorial chemistry, the objective is to enable the synthesis of large molecular libraries that are used as input for screening of potential lead compounds in HTS experiments. It is obvious that design of such large libraries should address the question of compound diversity. Ideally, the compound libraries are expected to span a vast chemical space and yet include a good number of entries from each chemically diverse sub system.\textsuperscript{20, 21} It would be prudent to screen compounds coming out of diverse sub systems than to screen compounds that have a high amount of redundancy in the information they carry. The task of classifying large number of chemical compounds into diverse sub systems based on their structural, physical and chemical properties is strikingly similar to the problem of classifying biological samples into different cancer types based on genetic profiles.

Similarly, in HTS experiments, large amounts of data are generated about the degree of activity or potency exhibited by each compound. At this stage, informatics methods are again needed to classify screened compounds into varying levels of activity ranging from highly inactive compounds to highly active compounds. After classification of screened compounds, the differentiating profiles of physico-chemical and structural properties that characterize a compound for its level of activity are to be identified. The informatics techniques developed to achieve the above tasks include structure representation and fingerprinting, defining key descriptors that represent various physical and chemical properties of a compound vital to its activity such as polar surface area,
number of hydrogen bond acceptors and donors, measure of solubility (logP), molecular weight and so on. The techniques that correlate the descriptors with activity include classical least squares, principal component regression, partial least squares, artificial neural networks, discriminant analysis, cluster analysis, nonlinear regression analysis. These techniques, commonly referred to as chemoinformatics, share a good deal of similarity with the bioinformatics techniques useful in early stages of drug discovery.

1.3 Scope and Objectives of Research

In the current study, most of our research effort was focused on developing informatics methodology that is applicable at two critical stages in drug discovery process – (i) target identification and (ii) lead compound identification and optimization. At the target identification stage, a methodology was developed to classify biological samples based on genetic fingerprints and to identify those biological targets that were most responsible for observed classes. More specifically, we focused on developing a methodology to analyze microarray data for classification of cancers into different subtypes when no prior knowledge of differences in classification is available i.e., unsupervised classification. Our approach was extended to the supervised scenario, where prior knowledge of differences could be used to build the predictive system. We demonstrated our methodology with improved and comparable accuracies to other existing techniques for both unsupervised and supervised classifications. The genes identified via our approach as potential targets were validated with supporting studies already published in literature. Finally, the methodology was adapted and extended to the
chemoinformatics framework at the lead identification and optimization stage of drug discovery, where the objectives were to identify and classify chemical compounds into varying levels of activity against a given biological target and to identify structural features most responsible for the observed classification of compounds.

Although the scope of the current work is demonstrated in the context of drug discovery process, it is easy to adapt it to diverse scenarios where the objectives are to classify samples into various groups and identify underlying features most responsible for the classification. Some potential applications that readily come to mind include the development of catalysts based on physico-chemical and structural properties of compounds for achieving specific catalytic activity, applications that broadly come under bioinformatics and chemoinformatics but have different objectives than those that are considered in this work – for example, gene clustering and design of diverse chemical libraries. In this work however, the specific objectives were two fold – in the cancer classification stage, the objectives were (i) to classify biological samples into different types of cancer and (ii) to identify genes that were most responsible for observed cancer types. Whereas, in the chemoinformatics framework, the objectives were (i) to classify chemical compounds into different classes based on varying level of activity or potency and (ii) to identify those structural features most responsible for the observed classes.

The methodology developed in this work relies largely on the multivariate aspects of principal component analysis and the application of $k$ means clustering algorithm in a hierarchical manner to achieve unsupervised multi-class classification. The principal components are replaced by the corresponding partial least square (PLS) components in the supervised scenario. Selection of influential components (principal components in
unsupervised case and PLS components in supervised case) for the purpose of classification is demonstrated and is one of the key steps for the success of this methodology. Hierarchical $k$-means is applied recursively to achieve binary classification at each stage eventually resulting in multi-class classification. Identification of features responsible for classification is achieved by examining the appropriate loadings of the principal or PLS components along with their coefficient of correlation with influential components. The multivariate nature of principal components ensures that the relative importance of underlying features is considered simultaneously where a feature’s relative importance as indicated by its component loading is determined in the presence of other features. This overcomes the problem of selecting one feature at a time, which is inherent in some of the prior techniques developed in this field. In addition, the mutual independence and orthogonality of principal components ensures that selection of redundant features at the expense of key non-redundant features is minimized. The details are further discussed in Chapter 3 (Systems and Methods).

1.4 Outline

This dissertation includes six chapters. Chapter 1 (introduction) and Chapter 2 (literature review) present an overview of this study and the most recent investigations in the field of bio and chemoinformatics as it applies to the process of drug discovery. Chapter 3 covers the development of a novel methodology in this work to address specific research questions raised at two critical stages of drug discovery. This chapter first presents the unsupervised methodology involving principal component analysis and
hierarchical $k$-means clustering. Next, the methodology is extended to a supervised scenario using PLS components in place of principal components. The chapter also discusses the identification of key features responsible for classification. Chapter 4 demonstrates the methodology in the context of cancer classification. Cancers are characterized at the molecular (genetic) level based on DNA microarray gene expression studies and the genetic biomarkers are identified as potential targets. Results and discussion on the application of this work in the context of cancer classification are presented. In Chapter 5, the methodology is applied to the lead identification and optimization stage within chemoinformatics framework where lead compounds are classified into different classes based on their level of activity/potency and the underlying structural, physical and chemical features are identified that are most responsible for observed activity of the chemical compounds. Results and discussion pertaining to the application of this work in chemoinformatics framework are presented. Chapter 6 discusses stand-alone work done on predicting octanol-water partition coefficients. Finally, Chapter 7 lists the major conclusions and makes recommendations on some future work that is worthy to pursue.
CHAPTER 2

LITERATURE REVIEW

2.1 Molecular Characterization of Biological Samples

Undeniably, one of the most influential advances of the 19th century is the unveiling of the digital code for the basis of life – the discovery of double helical structure of DNA by James Watson and Francis Crick in their landmark paper.27 This discovery had a profound effect on the understanding of biology and transformed it from a traditional branch of knowledge to an information science characterized by precise quantitative knowledge. It was soon realized that there are two features of DNA structure that account for much of its remarkable impact on science: its digital nature and its complimentarity.28 The digital nature of DNA is twofold – the ability of genes to encode for proteins, which are the smallest units of functionality in human body and the ability to regulate gene networks that code for the collective behavior of genes. The complimentarity of DNA arose out of the four types of bases that constitute the DNA – adenine (A), cytosine (C), guanine (G) and thymine (T). Any strand of DNA is made up of a combination of the sequence of the four bases A, C, G and T. The specific sequence of these four bases and its length differentiates one gene fragment from another. After a
decade of the discovery of DNA structure, it was discovered that each gene (a specific sequence of the 4 bases – A, C, G and T) encodes a complimentary RNA transcript, called messenger RNA or mRNA. Messenger RNA is made up of combination of the four bases – A, C, G and uracil (U), instead of T. Proteins, which are the smallest units of functionality, are also made up of sequences of building blocks known as amino acids. There are 20 different types of amino acids. Each protein is different from the other by the specific sequence of amino acids and the length of the sequence. Every bodily activity can be tracked down to the functions carried out by one or more proteins. Therefore, malfunction in any bodily organ can be characterized by identifying which of the different types of protein functions that are involved in that organ are disrupted. Furthermore, it was also discovered that there is a specific way by which the four bases of DNA or RNA are related to the 20 amino acids of the proteins. This is called the triplet code, in which three consecutive bases known as codons encode for a specific amino acid. Thus, it was finally understood that relative changes in the amounts of production of proteins are related and more importantly guided by the changes in expression of DNA. This is often referred to as the central dogma of molecular biology (Figure 2.1). The premise that held tremendous promise was that it would be possible to characterize the biological state of a sample at the genetic level by studying and patterning the expression level of DNA or RNA or proteins. It would be possible to understand the genetic causes of disease states and pursue therapeutics that specifically target the rectification of malfunctionality at the genetic level.
It was essential to study the behavior of thousands of genes simultaneously in order to compare the genetic profiles of disease states with normal states and to determine the set of genes most responsible for abnormal behavior. As a result of the human genome project, nearly 40,000 different genes have been identified and their functions characterized. More are being identified and characterized functionally every day. Armed with the growing annotations of genes with respect to their function, biological process and cellular component, the logical next step was to conduct high throughput experiments to study the expression of thousands of genes. Microarrays and GeneChip® brand arrays evolved as the standard high throughput technologies used for expression analysis, DNA re-sequencing, polymorphism detection and other related genome scale studies.
2.1.1 Microarray and GeneChip® Technology

Any organism can be viewed to be made up of various organs (e.g., lungs, heart, liver, kidneys etc.) that function in harmony with each other, while carrying out each of their specialized functions. Each of these organs is made up of their respective tissues that comprise of cells containing a complete copy of the organism’s genome (entire set of genes). It is known that cells are of different types (e.g., cells from liver tissue, kidney, blood cells, nerve cells, skin cells and so on) and are present in different states (normal cells, dividing cells, cancerous cells and so on). The essential difference in these cells arises from the differences in functionality which in turn are directly related to differential gene expression i.e., when, where and how much of each gene is expressed. Microarray experiments are a means to measure this differential gene expression. Analyzing data from microarray experiments allows researchers to assign a unique molecular signature (gene expression profile) to a unique state of the cells and thereby to the samples from where the cells have been derived.

Microarrays and GeneChip® brand arrays are microscopic arrays that are based on the principle of complimentarity in DNA double helical structure. The DNA bases exhibit a unique complimentarity whereby each base on one strand of DNA pairs with its complimentary base on the partner strand (A is complimentary to T; C is complimentary to G). This unique property allows for copying of genetic material (replication) and the generation of mRNA strands (transcription). The microarray technologies exploit the complimentarity at the nucleotide level to hybridize two complimentary strands of nucleotides and detect florescent signals that indicate whether hybridization took place or not and if it did, to what extent. Hybridization refers to the annealing of two
complimentary nucleic acid strands following the base-pairing rules (Figure 2.2). The nucleic acids in a duplex can be separated (denatured) by heating to destroy the hydrogen bonds that exist between the base pairs. Denaturing and annealing are mutually reversible processes with the proper control of temperature.

Figure 2.2: Example schematic of nucleotide hybridization. (a) complimentary DNA strands in un-hybridized state; (b) complimentary DNA strands in hybridized state.
Hybridization is also the principle behind the discovery of another key technology, polymerase chain reaction (PCR), whereby DNA sequences could be amplified exponentially.\textsuperscript{35} PCR has become so widespread and deep-rooted in molecular biology applications that it is difficult to imagine a technique that does not utilize PCR. PCR was also the technique that enabled high throughput microscopic arrays by overcoming the obstacle of insufficient nucleic acid derived from samples.

There is a marked difference between the two most popular and widely used microscopic arrays – the cDNA microarrays and the oligonucleotide GeneChip\textsuperscript{®} brand arrays. The cDNA microarrays are prepared either by spotting or printing complimentary DNA molecules on the glass surface. The length of sequence immobilized on the cDNA microarray surface is usually between 500 and 5000 bases long. Oligonucleotide GeneChip\textsuperscript{®} brand arrays, on the other hand, are synthesized one base after another using photolithographic techniques and the length of the immobilized sequence is usually between 20 to 80 bases. In either case, the sequence of nucleotides immobilized on the surface is called the probe sequence and the strand of nucleotides from the sample tissue that will complimentarily hybridize to the probe is called the target sequence. In the case of microarrays, a pair of mRNA samples (e.g. a control and an experimental) is separately replicated as complimentary DNA strands using reverse transcriptase, tagged with distinguishable fluorochromes, mixed and then hybridized on to the immobilized probe strands of the same spotted microarray.\textsuperscript{36} In the case of oligonucleotide GeneChip\textsuperscript{®} brand arrays, each experiment requires two chips – the mRNA derived from the control sample and experimental sample is tagged with a single fluorochrome and each labeled cDNA is hybridized to separate chips.\textsuperscript{37} In other words, there are two target
samples per array in the case of spotted microarrays while there is only one target sample per array in the case of oligonucleotide GeneChip® brand arrays. Another key difference between the two is that, there is only one probe sequence per gene in spotted microarrays while there are 11-20 probe-pairs per gene in oligonucleotide GeneChip® brand arrays. In both the cases, fluorescence intensities are calculated for each position on the chip or the array, which are then translated into the relative estimates of expression of genes.

In this work, we will focus on experiments with the Affymetrix® GeneChip® brand arrays. These arrays are more specific in terms of their target detection as they are fabricated carefully one base after another using photolithographic technique (Figure 2.3). These arrays contain pairs of probes, where one of the probes in the pair is a perfect match sequence of nucleotides while the other is a single base mismatch sequence at the center of the sequence. The purpose of mismatch probe is to obtain a measure of non-specific binding. A probe pair set is usually a collection of 11-20 probe pairs related to a common gene or gene fragment. These probe pairs are randomly located over the entire chip to enhance the accuracy of the measurement. The absence or presence of a gene is detected by the difference in the intensities of perfect match and mismatch signals of all probe pairs collectively belonging to a gene. A gene is present in a sample if consistently high intensities are detected for the perfect match probes and low intensities for the mismatch probes. Difference in these signal intensities corresponds to the strength of the presence of the gene.
Figure 2.3: Cartoon depicting a single feature on an Affymetrix GeneChip® microarray. Image courtesy of Affymetrix.
For example, one Affymetrix® brand array, HG U133A has the dimensions of 1.28cm by 1.28cm and the entire array is divided into 506,944 cells with 712 rows and 712 columns (Figure 2.3). With 11 probe pairs representing a gene fragment, the maximum possible gene fragments that can be studied on this array are \( \frac{506,944}{11*2} \), which equal 23,043 gene fragments. The HG U133A array actually contains 22,283 sequences that require \( (22,283*11*2) = 490,226 \) cells on the chip. The rest of the \( (506,944 - 490,226) = 16,718 \) cells are used as quality control sites. Each cell is allocated to a probe or a quality control sequence. The physical area of one cell is obtained by dividing the 1.28cm by 1.28cm array into 506,944 cells, which comes to 18µm by 18µm area for each cell. This allows for \( 10^6 - 10^7 \) identical probe molecules in each cell that are built up using photolithography. Sample preparation involves extracting the cells from tissue sample under investigation and lysing them in order to obtain the mRNA. This mRNA is tagged with fluorescent dye and allowed to hybridize on to the surface of the array (Figure 2.4). After allowing sufficient time for hybridization of the target sequence with the probe, fluorescent signal intensity for each cell on the array is measured and translated into an estimate for the expression of the gene (Figure 2.5). Separate arrays are required for control sample as well as the sample under investigation.
Figure 2.4: Cartoon depicting hybridization of tagged probes to Affymetrix GeneChip® microarray. Image courtesy of Affymetrix.
Figure 2.5: Cartoon depicting scanning of tagged and un-tagged probes on an Affymetrix GeneChip® microarray. Image courtesy of Affymetrix.
Microarrays and GeneChip® brand arrays have been used effectively in several studies. Although the experiment itself is straightforward, the task of pre-processing the raw data to obtain reliable gene-expression values involves several steps. These can be broadly categorized into four major steps – image segmentation, background adjustment, normalization and summarization of gene expression values. After allowing the target sample to hybridize with the probes, the array goes through a series of automated staining and washing steps using the Affymetrix® fluidics station. Subsequently, the array is scanned using Affymetrix® confocal laser scanner, which generates an image of the array by exciting each feature with its laser. The image now needs to be fragmented to measure the intensities corresponding to different cells on the array. There are several gridding algorithms to fragment the image and calculate individual cell intensities. The Affymetrix® software uses a dynamic gridding algorithm to segment the image and to compute the cell intensities based on a percentile algorithm. Several groups also describe other image processing algorithms. The process of accounting for unwanted variations like background noise, processing effect and cross-hybridization i.e., binding of non-specific DNA is called background adjustment. A GeneChip® experiment involves the use of two or more arrays and hence it is imperative to bring the different arrays in the experiment to the same scale (normalize) before meaningful comparisons between the arrays could be made. Normalization removes the effect of systematic variation across arrays except for differential gene expression. There are different methods of normalizing the arrays and one of the most frequently used method is called rank-invariant normalization. In this normalization procedure, it is assumed that an invariant set of features exist that do not change significantly between two experimental
conditions. Some of these features include control sequences while others include probe
sequences corresponding to house keeping genes that do not vary significantly from one
condition to another. All fluorescence intensities on each of the arrays are sorted (ranked)
according to increasing/decreasing intensity and items with similar ranks between two
arrays are identified as unchanged and hence constitute the invariant set.\textsuperscript{41} Normalization
is done based on these invariant set of features. When the normalized plot of base array
vs. any other array is made, it has the data points shifted such that the invariants lie on a
line whose slope equals one and this forms the basis for scaling other data points. After
normalizing all the arrays based on a baseline array, the next major step is
summarization. The purpose of summarization is to reduce the 11-20 individual probe
pair intensities to a single expression measure corresponding to a gene. One of the most
popular methods is the multi-chip statistical method called model based expression
index.\textsuperscript{42, 43} This method also allows the computation of standard errors for gene
expression values. According to this method, the model for calculating expression value
of a gene is given by:

\[
\text{MM}_{ij} = \nu_j + \theta_i \alpha_j + \epsilon \\
\text{PM}_{ij} = \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \epsilon
\]  
[eqn. 2.1]  
[eqn. 2.2]

where, $\text{PM}_{ij}$ and $\text{MM}_{ij}$ denote the perfect match (PM) and mismatch (MM) intensity
values for $i^{th}$ array and the $j^{th}$ probe pair for this gene, $\nu_j$ is the baseline response of the $j^{th}$
probe pair due to nonspecific hybridization. $\theta_i$ denotes the expression index for the gene
in $i^{\text{th}}$ sample array. $\alpha_j$ is the rate of increase of the MM response for the $j^{\text{th}}$ probe pair while $\phi_j$ is the additional rate of increase in the corresponding PM response, $\varepsilon$ is the random error and $i$ varies from 1 to I (the number of samples arrays in the experiment). The differential expression of a gene is measured by the difference in its PM and MM intensities:

\[
Y_{ij} = \text{PM}_{ij} - \text{MM}_{ij} = \theta_i \phi_j + \varepsilon_{ij} \quad \text{[eqn. 2.3]}
\]

In order for the parameters to be identifiable, the model places a constraint that the sum of squares of $\phi$s to be $J$, the number of probes. The model assumes that the errors, $\varepsilon_{ij}$ follow a normal distribution with mean zero and a constant variance, $\sigma^2$. Least square estimates of the model parameters are obtained by iteratively fitting the set of $\theta$s and $\phi$s, while assuming the other set to be known. Given $\phi$s, the least square estimate of $\theta$ is given by:

\[
\hat{\theta}_i = \frac{\sum_{j} Y_{ij} \phi_j}{\sum_{j} \phi_j^2} \quad \text{[eqn. 2.4]}
\]

such that,

\[
E(\hat{\theta}_i) = \theta_i \quad \text{and} \quad \text{Var}(\hat{\theta}_i) = \sigma^2/J \quad \text{[eqn. 2.5]}
\]
2.1.2 Cancer Classification

Cancer in general can be defined as uncontrolled growth of cells. This overgrowth can sometimes be benign (no apparent harmful effects) but most often it is malign (harmful). Cancers are of various types – oral, digestive, respiratory, breast, genital, nervous system, lymphoma, leukemia etc. Leukemia is a type of cancer that affects the white blood cells in the body. White blood cells are the primary defense mechanism of the human body (host) against invasion of foreign entities (microbes, viruses and any other foreign matter). Whenever a host is attacked by foreign entities, white blood cells respond in different ways depending on the specific functionality required to counter the attack. This process is often referred to as immune response. White blood cells originate in bone marrow along with red blood cells and platelets. White blood cells can be classified into two main categories – myeloid and lymphoid types. Myeloid cells are further subdivided into neutrophils, basophils and eosinophils. Lymphoid cells are further subdivided into B-cell lymphocytes and T-cell lymphocytes. Each of these sub categories has a distinct role in mediating immune response.44

In leukemia, which is the cancer of white blood cells, there is an unchecked growth or proliferation of white blood cells in the bone marrow. This affects the production of red blood cells and platelets. Often, the highly proliferating white blood cells themselves are immature and lose their function. Loss of function leads to poor immune response and the host becomes susceptible to a wide variety of infections in the wake of poorly functional defense mechanism. If immature lymphocytes are involved in the overgrowth, it is referred to as acute lymphoid leukemia (ALL) and if immature myeloid cells are involved, it is termed as acute myeloid leukemia (AML). Furthermore,
if the involved lymphocytes are B-cell lymphocytes, the disease is called B-cell ALL and if T-cell lymphocytes are involved, the disease is called T-cell ALL.\textsuperscript{45, 46}

Leukemia is a systemic (whole body) disease and therefore treatment is primarily based on chemotherapy (treatment with drugs). Although, all of the above diseased conditions are referred to as leukemia in general, it has been observed that different forms of leukemia require different approaches for optimal therapeutic results. For example, ALL does not respond well to chemotherapy traditionally used for AML while B-cell ALL does not respond well to chemotherapy traditionally used for T-cell ALL and so on.\textsuperscript{47,48} Thus, one of the key challenges of cancer treatment is its accurate diagnosis, which facilitates targeting of specific therapies to distinct tumor types, thereby maximizing the efficacy of the treatment and minimizing the unwanted side effects.

The conventional diagnosis of cancer into its subtype has been primarily based on the morphological appearance of the sample tumor tissue extracted from a patient. An experienced pathologist examines the morphology of a sample tumor tissue and classifies it into ALL, AML etc. and thereafter, appropriate therapy is assigned for treatment. This approach can have serious limitations when morphological appearance of a tumor is atypical and the sample cannot be assigned to a particular subtype unambiguously or sometimes the appearance of two samples is similar despite their distinct biological origin.\textsuperscript{46} In such cases, it would be informative to look at the overall gene expression profile (gene expression patterns of all the genes) of the sample. Expression of a gene can be thought of as the amount of that gene produced in the cell. The basic idea is that any particular state (normal, cancerous etc) of a cell is directly related to its genetic condition since the molecular basis of life is at the genetic level. Therefore, by studying and
comparing the expression patterns of the samples, their biological origin could be
established and the sample could be classified into its appropriate tumor type. It was
believed that high throughput gene experiments would easily solve the problem of
characterizing disease conditions at the molecular level and identifying potential drug
targets. However, the enormous amount of data generated by these high throughput
experiments makes the conventional statistical analysis techniques difficult to apply
directly. There is a need to adapt conventional analysis techniques such that they can be
applied to the data in the context of high throughput experimentation. Several such
techniques have been developed and some of the existing techniques are discussed in the
next section.

2.1.3 Existing Techniques

Microarrays and GeneChip® experiments allow for simultaneous measurement of
thousands of gene expression levels in the sample of interest (e.g. tumor tissue vs. normal
tissue). The tumor tissue extracted from a patient could be examined at a genetic level
with the aim of assigning a unique genetic signature that is representative of its cancerous
state. Gene expression data obtained as a result of GeneChip® experiments is unusual in
the sense that 60,000 or more measurements are made to estimate expression levels of
thousands of genes. Such data exists for each tumor sample analyzed via a GeneChip®
experiment. The task of gleaning meaningful information from this enormous amount of
data is formidable. Several techniques have been developed to address the task of
analyzing gene expression data. Currently, gene expression based analyses for molecular
characterization of cancer can be broadly grouped in two categories – supervised and unsupervised. In a supervised method, the samples to be classified are divided into two groups – a training data set and a test data set. Each sample in the training set is pre-assigned to a tumor subtype based on prior knowledge of class distinction. This assignment of the classes to different samples in a supervised method builds critical knowledge into the system to predict unknown classes for each of the samples in the test set and thus supervises the analysis. Such supervised methods extract patterns from the training data and build a model to make predictions on new data. Examples of supervised classification methods include logistic regression, discriminant analysis, support vector machines, artificial neural networks and a technique based on neighborhood analysis.

In logistic regression, the classification variable is random and is predicted by the continuous variables by way of predicting the parameters of a predictor function, usually a log linear model. In discriminant analysis, the classifications are fixed and the procedure uses labeled examples from each set of classified samples to estimate a probability distribution for the values of features in that set. When a new sample is to be classified, the procedure determines the most probable distribution and assigns the sample to the set closest to that distribution. In both logistic regression and discriminant analysis, the categorical response value is predicted based on the continuous predictor variables. Neural networks are machine-learning methods, wherein a set of prior samples are used to train a multi-layered computational network that makes a prediction on the class membership of a new sample. Neighborhood analysis is another technique that uses a pre-selected subset of variables known as ‘informative genes’. In this approach, when a
new sample is presented for classification, each of these informative genes casts a
weighted vote for one of the classes and the votes are summed to determine the winning
class. The vote is cast based on the expression level of the selective genes in the new
sample and the weight attached to a vote is determined by the relative significance
(measured by degree of correlation between the gene and class distinction in a training set
of samples) of the gene in the informative subset of genes that is participating in the vote.
Several other supervised techniques have been developed specifically for microarray data
analysis. Most of the supervised techniques can be summarized as a combination of two
important steps – first, reducing the number of variables (expression of genes) to obtain a
small yet informative subset and second, building the predictive model using samples in
the training set. Therefore, the success of supervised techniques is dependent upon two
critical questions – first, how to select a good informative subset of genes and second,
how to assign representative samples to the training set from which distinct gene
expression patterns are extracted for future predictions.

When samples have to be grouped into different categories with no prior
knowledge of the class distinction whatsoever, the grouping is called unsupervised. There
is no training data with already known classification available to extract patterns and
build a predictive model. As a result, the task of achieving unsupervised classification
presents a more difficult problem than achieving supervised classification. Although
supervised methods are superior to unsupervised methods, success of supervised methods
is critically dependent upon the quality of data in training set and the ability of the
method to select an informative subset of genes. Some of the current unsupervised
techniques include hierarchical clustering, $k$-means clustering and self-organizing maps.\textsuperscript{58}
In hierarchical clustering, the distance between each pair of samples is calculated and the two samples closest to each other are grouped as one unit. The process is repeated by treating the previously grouped samples as a single new entity. At the end of the grouping procedure, the samples are organized into a hierarchical tree based on their distance from each other. There are several options for the distance metric used in the calculation, which include simple euclidean, standardized euclidean, mahalanobis, correlation coefficient and several others. The number of genes used in the analysis corresponds to the number of dimensions in the distance metric space. There is no provision to identify the actual number of clusters in the hierarchical clustering. $K$-means clustering requires the pre-specification of the expected number of clusters, $k$. The $k$-means algorithm works by initially partitioning the data points randomly into $k$ clusters. In each subsequent iteration, all of the data points are assigned to the cluster whose centroid they are nearest to, and after the assignment of samples, the cluster centroids are recalculated. This process is iteratively repeated until the sum of point-to-cluster centroid distances is minimized in all the clusters.\textsuperscript{58} Self-organized maps organize the cluster of points into a map where similar clusters are close to each other. The partitions are in a pre-specified geometrical configuration. The computational details in a self-organized map are similar to $k$-means algorithm except that the cluster centroids are recalculated at each iteration using both the samples within a cluster and those that are in neighboring clusters.\textsuperscript{59}

There is a subtle but significant difference between supervised and unsupervised techniques. A supervised technique, owing to its advantage of learning patterns from a training set, places differential weights on the predictor variables on the basis of their relevance to class prediction. An unsupervised technique however does not have such an
advantage and hence places equal weights on the predictor variables. This could be a cause of concern as less relevant features can potentially mask the effects of relevant features. Therefore, dimension reduction and predictor selection assume a role of greater importance in unsupervised techniques than in supervised techniques. The successful application of supervised or unsupervised technique to classify samples in a microarray data is typically followed by the task of identifying and validating the genes that were differentially expressed across samples and hence were most responsible for the classification obtained.

2.2 Genetic Biomarker Identification

Molecular characterization of biological states (e.g., tumor subtypes) is a significant achievement of microarray technology. However, the true contribution of this technology lies in identifying the genetic biomarkers that characterize the biological state and translating this information to develop potential drugs that specifically target the biomarkers to restore biological function. Microarray experiments involve the measurement of thousands of gene expressions, not all of which may be relevant to the condition under investigation. Presence of more variables than needed in the analysis may result in increasing the noise and cause over-fitting. It has been observed that reducing the number of genes used in microarray data analysis improves classification accuracy. Several techniques are applied to either pre-select genes for classification purposes or post-select differential genes after classification. Some of these techniques include principal component analysis (PCA), partial least squares (PLS), a correlation
metric that measures the relative class separation produced by expression values of a
gene, selecting genes based on maximum difference in average expression value that is
similar to t-test statistic, independent component analysis (ICA) and soft independent
modeling of class analogy (SIMCA).

Principal component analysis is a popular multivariate dimension reduction
technique in which the original variables are transformed into fewer new variables called
principal components. Principal components are linear combinations of the original
variables satisfying two conditions. Firstly, a small number of components capture
maximal variability in the data set and secondly, the principal components are orthogonal
to each other. \(^{61,62}\) This technique was primarily developed for the case where the original
predictor variables are very large in number and when they have a high degree of
correlation among them. Thus, it suits well for applications in gene expression analysis,
where thousands of genes are studied simultaneously and it is known that genes exhibit a
high degree of correlation and group behavior. The disadvantage of linear principal
component analysis is that it may not be as powerful if the dependence structure among
the original variables is non-linear. PCA is based on maximizing the variance of the
linear combination of the predictor genes (principal components) sequentially so that first
few of the principal components capture a reasonable amount of variability in the original
predictor variables. Partial least squares (PLS) is somewhat similar in principle to
principal component analysis except that the PLS components are extracted by
maximizing the covariance between the response variable and a linear combination of the
original predictors. \(^{53, 54}\) This difference restricts the use of PLS to supervised scenarios in
which the response data is available. In another supervised gene selection technique
called neighborhood analysis, the class distinction is represented by an idealized expression pattern, in which the expression level is consistently high in one class and low in another class. Each gene is represented by an expression vector, consisting of expression levels from each of the tumor samples. Subsequently, genes are identified based on their correlation to the idealized expression pattern as compared to equivalent random patterns.\textsuperscript{15} Several other supervised techniques rely on pre-selecting genes based on a t-test statistic. For each gene, a t-test value representing the maximum difference in average expression levels is computed based on sample responses in training data set. Genes are sorted (ranked) by their t-test values and a subset of genes is selected by taking the genes with the most positive t-test values and genes with the most negative t-test values.\textsuperscript{53,54} Independent component analysis assumes that gene expression values vary under the influence of a smaller number of transcription factors and aims at identifying these latent variables.\textsuperscript{63} Soft independent modeling of class analogy (SIMCA) is built upon principal component analysis. SIMCA works by considering each class separately i.e., for each class, a principal component analysis is performed resulting in a different PCA model for each category. When a new unclassified sample is encountered, it is assigned to the category with which its model has the nearest best fit. Selection of genes is achieved by maximizing within-class modeling power and between-class separability.\textsuperscript{64} Thus, a large number of informatics techniques have been proposed to achieve the task of biomarker identification. Identification of all possible biomarkers involved in a disease condition allows thorough exploration of biomarker space for development of successful therapeutics.
2.3 Characterization of Chemical Compounds

The need to minimize time and costs associated with discovering new drugs has given the greatest impetus to the task of characterizing chemical compounds on a structural level. The physical, biological and chemical properties of a compound that render it useful in any kind of application are directly related to the compound’s structure. For example, when a drug is administered, the ability of the drug to partition itself between non-aqueous tissue phase and aqueous blood phase depends to a large extent on the structural features of that compound. Presence of more hydrophilic features than hydrophobic ones makes the drug partition itself more into the aqueous blood phase than the tissue phase and vice-versa. The development of a model that can accurately predict solubility in any two given phases (e.g., octanol-water partition coefficient, logK\textsubscript{OW}) from the structural description of the compound without actually carrying out the experiment is of great significance. The accuracy of such prediction depends on the degree of sophistication of the structural descriptors used and the quality of limited experimental data that is available to build the model. In the example of predicting octanol-water partition coefficient (logK\textsubscript{OW}), a model that is built merely on the counts of hydrophilic and hydrophobic features is less sophisticated than a model that assigns a weight to each structural feature. Further sophistication in the model would account for not only the weighted presence or absence of structural features but also their orientation in 3-dimensional space, the spatial distance separating any two features and so on. The measurement of these properties are sometimes difficult and may require time consuming experiments, however, they can be predicted based on the structural information of the
compound. Thus, given informative structural descriptors, a compound can be characterized for its physical, biological and chemical properties.

The information content conveyed by a molecular descriptor depends on the kind of representation used. There are simple representations that count atom types or structural fragments in a molecule. Other descriptors like 2D-descriptors are derived from molecular graphs representing topological features. There are other kinds of descriptors called geometrical or 3D-descriptors derived from a geometrical representation. The simplest forms of molecular descriptors are linear notations that convert chemical structure information to a string or a sequence of letters, using a set of predefined rules. One of the most popular linear descriptors is Simplified Molecular Input Line Entry System (SMILES). SMILES system is based on principles of molecular graph theory, where elements are represented by standard atom symbols. It allows structure specification by a very small set of rules that encode for type of atoms (aliphatic, aromatic etc.), chemical bonds between atoms (single bond, double bond, triple bond etc.) and branching. The simplicity of notation makes SMILES representation well suited for high-speed machine processing tasks including structure storage, structure searching and retrieval. 2D-descriptors like molecular graphs represent additional information in the form of topological indices than one-dimensional linear strings. The topological description of a molecule includes information on atom-atom connectivity, the size, shape, branching, heteroatoms and presence of multiple bonds in a molecule. Topological indices are the most widely used descriptors and have found applications in almost all of the stages of chemoinformatics where structural information is processed. These include quantitative structure activity relationship (QSAR) studies, quantitative structure property
relationship (QSPR) studies and similarity/diversity screening studies of virtual combinatorial libraries. More recently, the focus has shifted to geometrical representation of structures, which involves the knowledge of relative positions of atoms in 3D-space. 66

3D-descriptors provide additional discriminatory information than topological 2D-descriptors for compounds with similar structures and molecular orientation. Although, geometrical descriptors provide additional information, it comes at an increased computational overhead to the programs that deal with 3D-descriptors. Therefore, linear and 2D-descriptors are used for preliminary screening of large databases of molecules and establishing relationships between structures and physical, biological and chemical properties with reasonable accuracy, while the more complex 3D-descriptors are being preferred to establish precise relationships between molecular structures and more sensitive properties.

2.3.1 Quantitative Structure Activity Relationships

Quantitative structure-activity relationships (QSAR) broadly represent the methods to correlate structural or physico-chemical property descriptors of compounds with chemical/biological activities. The physico-chemical properties such as measures of solubility, electronic properties, and steric effects can themselves be predicted from structural features using empirical or computational methods. These properties can subsequently be used to develop models that predict activities. Activities used in QSAR include chemical measurements and biological assays. QSAR methods currently are being applied in many disciplines, with the vast number of applications lying in the areas
of drug design and environmental risk assessment. Examples include the prediction of lipophilicity parameters like octanol-water partition coefficients, electronic and steric constants, polar interaction parameters like molar refractivity indices, environmental toxicity and mutagenic potencies of certain chemical compounds.

One of the first relationships developed for correlating structure with activity is attributed to Hammett. In his work, Hammett proposed a reaction constant to describe the reactivity of aromatic systems. The reactivity was expressed in the form of rate constants and a substituent parameter that describes the influence of aromatic substituents. As the substituents changed within the same parental system, so did the influence on reactivity and yet the effect of substituents was more or less conserved across different systems. This observation formed the basis of modern QSAR – that substituents can be assigned a share of the activity exhibited by the entire molecule and then the activity of an entirely different molecule be predicted by the contribution of the substituent activities in an additive linear form or non-linear form whichever is appropriate. The next major contribution in QSAR came from Hansch and Fujita in which they proposed an extra-thermodynamic model to correlate biological activities with physico-chemical properties. Later, Free and Wilson formulated an additive model based on group contribution approach, where each substituent contributes its own share of activity to the parent structure. Many QSAR models have been developed ever since using a variety of approaches including conventional techniques like linear regression, partial least squares and numerous machine-learning techniques like neural networks, genetic algorithms, fuzzy set and logic theory. In such models, the response is usually a continuous variable while the predictor variables are discontinuous structural descriptors.
At times, it is useful to group the continuous response variable into discrete categories (e.g. toxicity measure of a compound, a continuous property can be transformed into a discontinuous variable with non-toxic and toxic categories segregated by some cut-off value). The advantage of being able to predict activities and physico-chemical properties of compounds from structural descriptors without actually carrying out experiments comes at the risk of producing chance correlations because of the large number of predictor variables involved in the analyses. The number of structural descriptors required to represent all the compounds in a chemical library of a decent size can get astronomically high. Often, the number of such structural descriptors is in several thousands and not all of them are useful in the analysis because some are redundant, some are correlated with each other and too many descriptors increase computational costs, make it difficult to identify discriminating features and obscure the model from interpretation. Most of the recently developed QSAR methods address this issue by reducing the dimension of descriptor variables by the use of techniques like multidimensional scaling, self-organizing maps, principal component analysis, factor analysis and structure visualization from graphed data points.¹

Multidimensional scaling (MDS) is a way to rearrange objects from a given dimensional space to a lower dimensional space such that the distances between the objects are reproduced approximately in both the spaces. MDS is a minimization algorithm that evaluates different configurations with the goal of minimizing the lack of fit between the distances of objects in the two spaces. Self-organizing maps are one of the machine-learning techniques where the new reduced dimensional space is built such that approximate input patterns are reproducible.¹⁷,¹⁸ Principal component analysis and factor
analysis are techniques that transform original variables into new independent variables based on variance-covariance structure.\textsuperscript{61,62} By visualizing structures as two-dimensional graphed data points, it is possible to identify and screen out the redundant structural descriptors.\textsuperscript{1} After appropriate descriptor selection and development of QSAR model, the properties or activities being predicted by the model have to be validated. Confidence in prediction and the model can be built-up by evaluating the performance of the model on test sets of compounds that were not included in building the model. It is useful to apply cross-validation techniques such as the ‘leave-one-out’ cross-validation where each compound is left out in model building and subsequently the same model is used to predict the activity of the left-out compound.

2.3.2 Chemical Compound Classification

One reason why compounds in a chemical library need to be classified into different categories is to identify the profiles unique to a class of compounds so that when a new compound is to be designated a class label, it is grouped with the appropriate class. It is of great significance to identify the class of compounds exhibiting toxicity from those that do not and to accurately label a new compound as either toxic or non-toxic without experimentation that is destructive to the experimental specimens. Identifying unique structural signatures not only help categorize a new compound as either toxic or non-toxic but also help in the rational design of drugs by providing insights into the relationships between certain structural features and their effect on toxicity. In addition to toxicity, structural profiling helps optimize several other key properties of drugs such as
absorption, distribution, metabolism and elimination. The ability to predict measures of above-mentioned properties from structural profiles avoids expensive experimentation and helps identify, at an early stage, those compounds that would fail as potential drugs for not having one of the key physico-chemical or biological properties in acceptable range. Identifying such candidates early would also allow optimization of these properties by structural alterations.

Many pharmaceutical organizations pursuing discovery of new drugs have adopted high-throughput technologies at primarily two levels – screening targets for a given disease and screening chemical compounds for any given target. Although, high-throughput screening (HTS) allows for searching of vast chemical space for ‘hits’ on a given target, it would not be feasible to test each compound in the combinatorial library in the pursuit of leads (promising drug candidates) that exhibit potency or activity against a given biological target. Therefore, classification of chemical compounds is a necessary step to select representative yet diverse compounds spanning the entire chemical space of the library. Expensive and time-consuming experimentation prohibits the evaluation of every compound, rather only a small representative set of compounds are selected for evaluation. Moreover, most pharmaceutical organizations would like to acquire compounds from external sources that offer complimentary diversity to their existing collection of libraries. For this purpose, the compounds need to be classified into various categories based on similarities and differences between existing compounds and to be newly acquired compounds. Diversity of chemical libraries increases the chance of identifying all possible diverse leads in the screening process against a given biological target. Pursuing structurally diverse leads increases the success rate of converting leads
into final drugs with optimized pharmacogenomic properties as certain leads cannot be optimized for certain properties although they are potent and active against a given target. Another reason for classifying compounds into different categories is that QSAR models developed for correlating activity and structural descriptors can be validated using similar compounds in a class. For example, after dividing a library into separate classes, each class can be broken down into training and test sets. Training sets are used for model development while the model makes prediction on the compounds in test set, thereby allowing model validation.

Accurate classification of compounds depends on the quality of molecular descriptors used to represent the structural information, the ability of the QSAR method to summarize the information contained in large number of descriptors and to predict the class membership of new compounds based on profiling compounds in the learning set. A number of techniques have been developed to achieve classification of compounds based on structural descriptors and some of the existing techniques are presented in the next section.

2.3.3 Current Methodologies

Existing techniques for analysis of large chemical data sets have evolved from simple linear regression models to highly sophisticated artificial-intelligence based models. The data sets involved in the drug discovery process are not only large as in the case of gene expression studies but also sparse and involve discontinuous type of variables. These data sets contain two types of data – the predictor variables and the
response variables. The predictor variables often encountered are of two types – (i) counts of descriptors and (ii) continuous values representing physico-chemical properties. Examples of such structural descriptor counts include the number of specific benzene groups, amino groups, carbonyl groups etc., the number of hydrogen bond acceptors, hydrogen bond donors, the number of rotatable bonds, scores for Lipinski’s rule of five and so on. Examples of continuous properties include molecular weight, molar refractivity index, a measure of solubility (e.g., logK_{OW}), polar surface area and so on. The response variable can represent class labels identifying the class membership of the compound (e.g., toxic vs. non-toxic, different levels of activity against a biological target – low, medium and high) or continuous properties like measures of chemical toxicity and biological assay end-points (e.g., lethal dosage levels expressed as LD_{50} values or infective dosage levels expressed as ID_{50} values). The key objectives of the analyses are to classify compounds with respect to the response variable of interest and to identify features in the predictive set of descriptors that are most responsible for observed classification.

Some of the existing techniques for class prediction and feature identification include multiple linear regression, hierarchical cluster analysis, support vector machines, recursive partitioning, neural networks and partial least squares. Most of these have been already mentioned in the “existing techniques” section on analysis of gene expression data sets. Recursive partitioning (RP) is one of the non-linear approaches to partition the data. RP method works by first splitting the entire data set into two groups – those compounds that are represented by a molecular descriptor and those that do not. The average responses of the two groups are calculated and this process is repeated for
each molecular descriptor in the predictive set. The descriptor that has the highest average difference between both the groups represented and not represented by the descriptor is identified as the first node on which the data has to be partitioned. Each subset is further partitioned based on the rest of the descriptors recursively.

2.4 Identification of Active Chemical Features

Classification of compounds based on a specific response (e.g., mutagenic potency) can be attributed to a unique combination of structural features whose presence or absence makes the compound classify into one of the several categories (e.g., mutagenic or non-mutagenic) of that response variable. Identifying this unique combination of features establishes a reference pattern, which can be used to classify new compounds into their appropriate classes for a specific response. Relating structural features with certain properties like biological or chemical activity paves the way for rational design of drugs. Instead of searching large number of compounds for a given pattern, compounds can be built-up to have these patterns or compounds having part of the pattern can be modified to obtain specific combinations of features that yield required properties of interest.

The task of identifying feature combinations responsible for activity is very similar to the task of identifying gene patterns that characterize a biological condition. Some of the same techniques like principal component analysis, partial least square regression, recursive partitioning are used to filter out features that have the most influence in class prediction in the context of the response variable of interest. In case of
principal component analysis and partial least square analysis, the components that captured highest amount of variability in the data set are used to identify differentiating chemical features. The features that have a high degree of correlation (Pearson correlation) or high amount of loading are identified as most strongly related to the component and therefore for the observed classification. In case of recursive partitioning, the output of the analysis is a dendrogram (tree) in which predictors are used to progressively split the data set into smaller and more homogeneous subsets. If a node in the dendrogram contains mainly active compounds, then the features that split the tree in that path correspond to the features that are associated with activity. The path to a node whose compounds are predominantly inactive correspond to the feature that does not have a strong effect on activity.

2.5 Conclusions

Informatics methods for characterization of biological samples and chemical compounds alike have gained great importance in the modern drug discovery paradigm. The large amounts of data generated either from high-throughput experimentation or obtained historically have valuable information hidden in them, which needs to be extracted and used rationally for modern drug discovery. Microarray technologies have had a deep impact on characterizing biological states at the genetic level. These technologies have made it possible to study thousands of genes simultaneously for changes in expression levels. Diagnosis of several biological conditions including different types of tumor is possible by identifying groups of differentially expressed
genes and assigning unique molecular signatures. This technology has not only augmented conventional diagnosis, but also has provided insights into the genetic causes of disease states. Genes identified with differential expression patterns are explored for potential drug targets and groups of genes having correlated expression patterns are studied for understanding gene networks.

In high-throughput screening (HTS) experiments, a large number of chemical compounds are screened against disease targets in order to identify biologically active compounds. The data generated from these experiments serves as the starting point in the long process of drug discovery. The compounds used in HTS experiments are analyzed to identify the differential structural features that distinguish a compound for its observed activity from others. Identification of differential features aids in the rational design of drugs and in optimization of several key properties of the potential drug compounds. There are some similarities between the gene expression studies and HTS experiments for screening chemical compounds. Both the experiments generate large amounts of data with the main goal of discovering knowledge from raw data i.e., to classify samples into different categories and identify the underlying features most responsible for the observed classification. On the other hand, the difference between the two lies in the type of data generated – in case of gene expression studies, the predictor variables are continuous type whereas in case of HTS studies, the predictor variables are mostly discontinuous type. Nevertheless, the methods used for the analyses of the two types of data sets can be mutually adapted to fit the needs of the analysis. From the perspective of drug discovery, the same informatics tools developed at target identification stage are also useful at lead identification and optimization stages.
CHAPTER 3

SYSTEMS AND METHODS

3.1 Introduction

The art of classifying distinct objects based on their properties is employed in virtually all kinds of applications. An insurance company may classify customers into different groups of risk based on the customer’s income, age, past credit history and family size. The admissions office at a university may classify an applicant as ‘likely to graduate successfully’ or ‘unlikely to graduate’ based on the applicant’s test scores and past academic record. A semi-conductor manufacturer may classify an electronic component as ‘acceptable’ or ‘unacceptable’ based on whether the component’s material strength, electrical conductance and physical dimensions fall within specifications. Although the actual determination of whether the insurance company’s customer was a good risk or whether the university’s admitted applicant was successful at the end of the program or whether the semiconductor plant’s electronic component really performed up to the expectation can only be made after a long period of time. However, the ability to predict beforehand is essential to make these critical decisions. Similarly, most medical problems can only be identified in a conclusive manner by way of expensive operations.
However, it is more practical to diagnose a medical condition from measuring other indirect but easily observed properties. Likewise, the toxicity of a potential drug can be established by thorough testing on animal models followed by clinical and pre-clinical trials in different populations. However, the ability to predict beforehand the ballpark level of toxicity of a new drug is precious.

Appropriate classification of objects into various groups is a twofold process – separating distinct objects and allocating new objects to previously defined groups. In the first step, when a population is being separated into distinct groups for the first time, the emphasis is on obtaining as accurate a classification as possible and avoiding costly misclassifications. All potential variables that are believed to influence the class membership of objects are considered and carefully analyzed. In the second step, when a new object is to be allocated to one of the previously defined groups, the same variables that were used in classification are measured for the new object and their profile compared to the class profiles in order to identify the best possible match. The profile of variables that characterize a class of objects is identified and used in the allocation of new objects to distinct classes.

In certain classification tasks, there is factual response data available based on past historical observations. A list of university students who were admitted and successfully completed the degree program and a list of those who were admitted but did not successfully complete the degree. When this type of actual response data is available, the first part of the classification problem does not apply; rather these classifications are true and can be used for learning and extracting characteristic profiles of variables that can distinguish future students. The historic data is treated as a training set and predictive
models are built to make predictions on new objects. In such a case, where prior factual knowledge of class distinctions is available, the task of classifying a new object is often referred to as supervised classification. In the absence of any prior knowledge of class distinctions, the task of classifying a new object is referred to as unsupervised classification.

Several supervised techniques have been demonstrated to work remarkably well for binary classification and a few of these techniques are reasonably good at supervised multi-class predictions. However, techniques for unsupervised binary and multi-class classification have not been fully developed. In this work, we present a technique that uses principal component analysis along with $k$-means clustering on differentially weighted principal components to address unsupervised classification for both binary and multi-class problems. This unsupervised methodology is extended to cases where prior knowledge of class distinctions are available i.e., to solve supervised classification problems. We demonstrate our methods using two different data sets: (a) gene expression data sets to classify cancers (b) chemical data sets to classify toxicity/activity of compounds. Methods to identify the differentiating profiles that are most responsible for observed classifications are also discussed.

3.2 Unsupervised Classification Methodology

Absence of any prior knowledge of class distinctions makes unsupervised classification a harder problem to solve than supervised classification. In data sets such as gene expression studies and chemical data sets, the number of variables that are believed
to influence class membership is usually large. As already mentioned earlier, not all of the variables are useful for predicting class membership because some are redundant, some are correlated with each other and some may be irrelevant adding nothing but noise to the data. Therefore, techniques that can reduce high dimensionality of the variables by weeding out the unnecessary ones and selecting the differentiating variables are used as the first step. We chose to use principal component analysis for this purpose.

3.2.1 Principal Component Analysis and Selection

In many experimental conditions including gene expression studies and high-throughput screening experiments, it often happens that more variables than required are included in the study. This situation arises when there are no clear rules on pre-selecting variables to be included in the experiment and it is relatively simple to study a large number of variables in one experimental run. In addition, by considering as many variables as possible, the chance of leaving any key variables out of the study is minimized. Although a large number of variables can be measured effortlessly in such high-throughput experiments, it comes at the expense of dealing with the large amount of resulting data. Principal component analysis (PCA) is a dimension reduction technique applied to data sets containing large number (on the order of $\sim 10^3$ or more) of inter-dependent variables. PCA aims at explaining the variance-covariance structure of this large number of variables using a smaller set (on the order of $\sim 10^1$) of new independent variables known as principal components (PCs), which are linear combinations of the original variables. The smaller set of PCs that replace the larger set of original variables tends to retain as much information as possible about individual differences in objects or
experimental units. PCA usually serves as an intermediate data reduction step in much larger investigations. The derived principal components are used as inputs to further analyses such as regression, discriminant analysis or cluster analysis. PCA often reveals relationships between original variables that would have been otherwise difficult to uncover.

Before describing the theory of PCA, it would be worthwhile to consider how some supervised binary classification techniques rely on an approach that reduces the dimensionality in the data set by pre-selecting variables based on a two-sample $t$-test statistic. In most supervised approaches, a subset of the samples in the data set is assigned as training set. The value of $t$-statistic is computed as:

$$t = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{s_1^2}{c_1} + \frac{s_2^2}{c_2}}}$$  \hspace{1cm} \text{[eqn. 3.1]}

where,

$\bar{y}_i$ is the average response belonging to $i^{th}$ class in a training set; ($i=1,2$ for binary)

$s_i^2$ is the sample variance of response values belonging to $i^{th}$ class in training set;

$c_i$ is the number of samples belonging to $i^{th}$ class in training set;

For each predictor variable, a $t$-statistic is computed based on its ability to differentiate the binary samples in training data set. Predictor variables are sorted by their
$t$-statistic values and a small subset of $p$ predictors is selected by taking the $p/2$ predictors with the largest positive $t$-values and $p/2$ predictors with the smallest negative $t$-values.\textsuperscript{54}

In gene-expression studies, this technique can be misleading because it completely ignores the interaction of genes and selects them based on an individual basis. It is unlikely that isolated genes selected in such a manner without any inter-dependence capture all the information for class distinction. Moreover, the subset of genes selected in such a manner may contain highly correlated genes carry redundant information. Besides, using a $t$-statistic to pre-select predictors inherently assumes normal distribution, which may not be true on all occasions. Assuming normal distribution when it is not applicable leads to unreliable $t$-test results. Some supervised multi-class classification techniques also rely on a similar approach based on ordering and pre-selecting predictors. In this case, the predictor selection is based on the number of times a pair-wise difference for various pairs of categories in training set exceeds a critical score.\textsuperscript{53}

In the approach presented in our work, principal components are used to reduce the dimensionality of predictor variables. PCA is a multivariate approach, not merely multi-variable in nature, in which each of the new variables (PCs) is obtained by taking all of the original variables into account simultaneously. The fact that each of the PC is obtained from linear combinations of all the original variables, accounts for the multivariate dependence within the original variables that is sometimes lost if they are treated as multiple but independent variables. In the computation of PCs, the original variables (e.g., expression levels of $n$ genes) can be thought of as $n$ random variables $X_1, X_2, \ldots, X_n$ measured for each of the $m$ samples. Thus, we have a data matrix $\tilde{X}_{m,n}$ with $m$ rows and $n$ columns. If we let the random matrix, $\tilde{X}_{m,n} = [X_1, X_2,$
... have the covariance matrix \( \Sigma \) with eigenvalues \( \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0 \), then the corresponding eigenvalue-eigenvector pairs will be \( (\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_n, e_n) \) and the principal components are given by:

\[
P_C_i = e_i X_{m,n} = e_{i,1} X_1 + e_{i,2} X_2 + ... + e_{i,n} X_n
\]

\[i = 1, 2, ..., n\]

[eqn. 3.2]

such that,

\[
\text{Var}(P_C_i) = e_i \Sigma e_i = \lambda_i
\]

\[i = 1, 2, ..., n\]

[eqn. 3.3]

\[
\text{Cov}(P_C_i, P_C_k) = e_i \Sigma e_k = 0
\]

\[i \neq k\]

[eqn. 3.4]

The maximum possible number of principal components equals the rank \( r \) of the data matrix, and therefore \( r \leq \min (m, n) \). Generally a large proportion of the overall variability is accounted for by a small number \( p \) of the principal components, where \( p << r \). Thus the original data set of \( m \) samples and \( n \) variables could now be replaced by a new data set of \( m \) samples and \( p \) variables without losing much information, where \( p << n \). It should be noted that these \( p \) PCs are orthogonal to each other and hence are uncorrelated and do not carry any redundant information. The most commonly used method for PCA is singular value decomposition, which is implemented in many mathematical software packages including MATLAB®.72 Since the PCs are linear combinations of all of the
original variables (e.g., gene expressions), each of the original variables contributes to a PC in the presence of other variables making this a truly multivariate approach. Thus, by using PCs as a small number of new informative variables, it is possible to account for the interaction of predictors in each PC and by selecting different PCs that are independent of each other, to overcome the problem of selecting redundant information at the expense of independent information. It is also reasonable to hypothesize that the class distinctions correspond to changes in a group of predictors (e.g., genes exhibiting group behavior) and not due to changes in a single isolated predictor. Therefore, techniques that consider the multivariate nature of variables should do at least as much and in most cases better than the techniques that ignore this possibility.

An important topic within the application of PCA that often goes without careful consideration is the choice of using covariance matrix or correlation matrix for the computation of PCs. In most of the applications including bioinformatics and chemoinformatics, it has become a standard practice to use correlation matrix in the computation of principal components. Although the use of correlation matrix is necessary in certain cases, it is equally arbitrary in others. The decision warrants careful consideration depending on the nature of the variables involved in the experiment. The PCs obtained from covariance matrix are entirely different from those obtained via correlation matrix and there is no straightforward relationship between them.\(^6\) If the predictor variables involved in the analysis consist of measurements on the same scale then it makes sense to derive principal components based on eigenvectors of the covariance matrix. However, if the scales of measurement are arbitrary, then the principal components should be derived based on eigenvectors of the correlation matrix. In this
case, the variables have widely differing variances and need to be standardized to have unit variances to carry out meaningful analyses. The use of correlation matrix is equivalent to standardizing the original variables. Nevertheless, it should be pointed out that when applicable, it is advantageous to use covariance matrix for the derivation of PCs.\textsuperscript{73} In the special case when all variables are measured in the same units (e.g., gene expression measurements), it is advantageous to use the covariance matrix in computation of PCs. In this case, if the correlation matrix is preferred over covariance matrix, it is equivalent to standardizing the original variables. The transformation to standardize variables without necessity is an arbitrary choice and unjustified without doubt. It is well known that unnecessary transformations of variables lead to the loss of accurate dependence structure in variables and makes the interpretation difficult and unreliable.

The main objective of PCA is to reduce the dimensionality and capture as much variation in the data set as possible. Therefore, using PCA it is possible to select a small number $p$ of PCs ($p \ll n$) that explain most of the variation in the data set. However, it may not be that all the $p$ PCs capture information about the individual samples in the data set. Sometimes, a PC might merely be a weighted average of original variables, indicating that it does not capture any contrasts. Therefore, if the goal is to classify distinct objects, it is necessary to identify those PCs that could be used as predictors in a clustering algorithm and not those that merely capture a high percentage of variation in the data set.

For this purpose, we analyze the eigenvectors of the $p$ PCs. The distribution of the eigenvector coefficients is informative for this analysis (Figure 3.1). Those eigenvectors
that have a bell shaped distribution centered on zero have approximately the same number of positive and negative coefficients. Such eigenvectors are of interest because they lead to PCs that are linear combinations of original variables capturing contrasts in the samples. In Figure 3.1(a), the distribution of eigenvector coefficients is skewed to the positive side, indicating that most of the coefficients are positive. The PC corresponding to this positively skewed example eigenvector would not be a good predictor for classification purposes. In Figure 3.1(b), the distribution of eigenvector coefficients is centered on zero and there are approximately the same number of positive and negative coefficients (with proportion close to 0.5). The PC corresponding to this example eigenvector would be a good predictor for classification purposes. These informative eigenvectors (and hence the corresponding PCs) can be identified quantitatively by examining the proportion of eigenvector coefficients that are negative (or alternatively, positive) in an eigenvector. Those eigenvectors that have a proportion close to 0.5 correspond to PCs that capture the contrasts in the data set and should be good predictors of class distinction. Those that have a proportion close to 0 or 1 (farthest from 0.5) are not as contrasting and would not be good predictors. It is especially important to exclude these PCs as predictors if they account for a high percent of overall variability in the data set because they are capable in that case of masking the class distinction. Thus, from amongst the $p$ PCs that explain most of the variation in the data set, those $q$ PCs ($q \leq p$) are selected that have their proportions close to 0.5 as explained above. Although the primary purpose of PCA is data reduction, careful selection of influential principal
Figure 3.1: Schematic of eigenvector distributions: (a) poor predictor for classification purpose. (b) good predictor for classification purpose.
components can be made so that they capture not only the variation in the data set but also differentiate the samples in the data set. Influential principal components thus selected can be used as predictive variables in an unsupervised grouping method such as \( k \)-means clustering. Thus, the \( q \) PCs act as predictors in a clustering algorithm, such as \( k \)-means clustering.\(^ {74}\)

3.2.2 Hierarchical \( K \)-means Clustering using Influential Principal Components

\( K \)-means is a clustering method that relies on some sort of distance metric to measure similarity between two objects. The distance is calculated in the dimensional space of selected principal components. The value of \( k \), the expected number of clusters, need to be supplied as an input to the clustering algorithm. \( K \)-means algorithm treats the samples as data points in the principal component space and works by initially partitioning these data points into \( k \) clusters at random. For the first iteration of the partition, centroids for each of the \( k \) clusters are calculated. The sum of distances of all data points in a cluster from the cluster centroid is calculated. Now the data points are adjusted amongst the \( k \) partitions, by re-assigning the points in a cluster to other clusters, so that the overall sum of point-to-cluster centroid distances is minimized in all the clusters. In each of the subsequent iterations, each of the data points is re-assigned to the cluster whose centroid it is nearest to, and after the assignment of samples, the cluster centroids are re-calculated. This process is iteratively repeated until the sum of point-to-cluster centroid distances is minimized in all the clusters.\(^ {58,74}\) Figure 3.2 illustrates \( k \)-means clustering process with the help of a cartoon example for classifying 15 data points into two clusters (\( k = 2 \)) in two-dimensional space.
Figure 3.2: Cartoon illustration of $k$-means clustering process. (a) 15 data points in 2-dimensional space; (b) selection of random seeds; (c) assignment of data points to one of the two clusters; (d) re-assignment of data points.
In Figure 3.2(a), the 15 data points to be partitioned in 2-dimensional space are shown. The objective is to partition these data points into two-clusters \(k=2\). The two cluster separation process starts by choosing two random seeds that act as initial cluster centroids as shown in Figure 3.2(b). Each of the data points is assigned to one of the two clusters based on which centroid the data point is nearest to. This produces the initial set of clusters as shown in Figure 3.2(c). Now, the initial random seeds are dropped as the cluster centroids and based on the data points belonging to two clusters in Figure 3.2(c), two new cluster centroids are obtained. Distance of each data point is calculated with respect to the new cluster centroids and the data points are re-assigned to one of the new centroids the data point is nearest to. This re-assignment produces new clusters as shown in Figure 3.2(d), but this time the sum of distances of data points to their own cluster centroids is less than the corresponding sum in the first iteration with random cluster centroids. The centroids shown in Figure 3.2(d) are again dropped and new centroids calculated based on the assignment of data points to the two clusters in the Figure 3.2(d). This process is iteratively repeated until the sum of the distances is minimized optimally.

The above example illustrates the \(k\)-means clustering process when the objects have to be partitioned into two clusters. \(K\)-means clustering requires the pre-specification of the expected number of clusters, \(k\). However, in many practical applications of the clustering process, more than two clusters are possible and the number of clusters is unknown ahead of time. Therefore, we apply \(k\)-means by specifying the parameter, \(k = 2\). This produces two clusters in the first run and then within each of the clusters, \(k\)-means clustering process is applied. This produces two more sub-clusters in a binary tree-like
structure until the resulting clusters fail to separate clearly. The end of classification is characterized when the clustering produces unstable clusters or clusters with one group having extremely small number of samples in it. Unstable clusters are obtained when samples are not assignable consistently to the same clusters if the process is repeated several thousand times. Clusters with one group having extremely small number of samples are obtained when the differences in the samples are not significant and the small number of samples are also not consistently produced if the process is repeated several thousand times. Thus, \( k \)-means is applied recursively to achieve binary classification at each stage, ultimately leading to solve an unsupervised multi-class classification problem as shown in Figure 3.3. After the samples are classified into separate groups, it is possible to work backwards by relating the influential PCs responsible for classification of objects to the features that have a high correlation or loadings with the PCs to obtain a potentially non-redundant pool of differentiating features.
Figure 3.3: Schematic of hierarchical $K$-means clustering.
3.2.3 Identification of key features

Even though principal component analysis (PCA) is primarily a dimension reduction technique, it reveals relationships between the variables involved in the analysis, which, would not have been otherwise discovered. It would be a formidable task to pick out differentiating features from among thousands of variables. We hypothesized that if the discovered classes at the end of hierarchical $k$-means clustering represent true structure of differences, then the principal components that produced these classes capture key information required to distinguish the classes. We propose two methods to identify the differentiating features.

One way to identify the differentiating features is to measure correlation of original features with the influential PCs. Original features that have a high correlation with the scores of influential principal components are expected to contain the information related to class distinctions within samples. These correlations are measured for all of the original features for each of the influential PCs. Therefore, it is possible to end up in a situation where the original features identified as differentiating ones by comparing the correlations of one PC (say, PC$_1$) are entirely different from the original features identified via the correlations of another PC (say, PC$_2$). Even though it appears paradoxical to come up with different lists of differentiating features, this is in fact the very solution to the problem of being trapped in the horde of redundant features at the expense of identifying all the possible non-redundant differentiating ones. In most cluster analyses, one of the objectives is to identify differentiating features with the hope of gaining an insight into all the possible mechanisms of interplay that differentiate one class of objects from another. If a group of features connected in a network end up to be
strongly differentiating, then the likelihood of identifying features belonging to the same network as strongly differentiating is maximized. In this case, features carrying redundant information from the same network are identified as significant features at the expense of features that belong to different networks that differentiate the objects too but not as strongly. Failure to identify features from all the differentiating networks is sub-optimal in gaining complete insight of all possible underlying mechanisms in class distinction. It is very much the case with gene expression studies, where networks of genes are involved in differentiating the biological samples. If the analyses only identify genes from one or two most strongly differentiating networks, then valuable information on other differentiating gene networks is lost. The genes identified from same network usually have redundant information in them. Sometimes, these genes do not turn out to be practical targets for therapy and the medical condition remains unresolved, despite the existence of other potential targets that were left undiscovered. Therefore, the method of identifying distinct lists of differentiating features based on each influential PC enables complete exploration of feature space in the search for differentiating features. In order to obtain a consolidated list of important features from all of these individual lists, we propose to calculate an overall rank for each feature. A feature is sorted (ranked) based on its magnitude of correlation in each list associated with a PC. The overall rank of a feature (R) is obtained through the summation of all the individual ranks weighted by proportions of variability explained by the corresponding PCs:

\[ R_j = r_{j,1} V_1 + r_{j,2} V_2 + \ldots + r_{j,n} V_n \]  
[eqn. 3.5]
Where,

\[ R_j \] is the overall rank of \( j^{\text{th}} \) feature;

\[ r_{j,i} \] is the rank of \( j^{\text{th}} \) feature in \( i^{\text{th}} \) PC list;

\[ V_i \] is the proportion of variability accounted by \( i^{\text{th}} \) PC; \((i = 1 \text{ to } n)\).

The second way to identify the differentiating features would be by examining the feature loadings in each PC. It is worthwhile to review the equation (eqn. 3.2) showing PCs as linear combination of original features:

\[
PC_i = e_{i,1}X_1 + e_{i,2}X_2 + \ldots + e_{i,n}X_n
\]  

[eqn. 3.2]

In the equation above, a PC is constructed by taking the linear combinations of original features and the relative contribution of each feature is given by the eigenvector coefficients of that PC, which load the original features. Thus, if a PC is influential in differentiating the classes, then the original features that make the most contribution to that PC would also be influential in differentiating the classes. The feature loadings are obtained using multivariate PCA and therefore represent the contribution of a feature to the PC in presence of other features. Therefore, features identified from this manner have the advantage of being extracted from a multivariate approach in which the interaction of features is accounted for. Like in the correlation method described above, features can be sorted in a list for each influential PC. Overall ranks of features can be obtained in a
similar manner to equation 3.5 as described above. We consider both the methods to be useful and complementary in exploring the features space to identify differentiating features. The correlation method treats the features as univariate in calculating the correlation coefficients whereas the feature-loading method treats the features as multivariate, thus complementing each other.

3.3 Supervised Classification Methodology

When prior knowledge of class distinctions is available, the classification is usually referred to as supervised classification. In a supervised method, the samples to be categorized are divided into two groups – a training data set and a test data set. Each sample in the training set is pre-assigned to a class label based on prior knowledge of class distinction. This assignment of the classes to different samples in a supervised method builds critical knowledge into the system to predict class labels for each of the samples in the test set and thus supervises the analysis. Such supervised methods extract patterns from the training data and build a model to make predictions on new samples. The availability of reliable knowledge concerning class distinctions builds confidence and increases accuracy of the predictive model. On the other hand, one needs to be cautious in choosing the samples that constitute the training set because the rules to differentiate future samples are derived from the sample differences in the training data. Therefore, it is essential to ensure the quality and reliability of samples that are assigned to the training set. Presence of outliers and unreliable samples can adversely affect the prediction accuracy. If the training data is unreliable, the prediction accuracy of an unsupervised method might surpass that of a supervised one. Nonetheless, if prior
knowledge of class distinctions is accessible, supervised methods should be considered first.

The methods proposed by us for unsupervised classification and feature identification could be adapted to fit the needs of solving a supervised classification problem. In the supervised scenario, we propose to utilize partial least squares (PLS) components as a replacement for principal components in the unsupervised methodology. PLS is very close to principal component regression in which the focus is on explaining as much predictor variation as possible. PLS on the other hand places more emphasis on response variables because in addition to the predictor variables, data is also available on the response variables.

3.3.1 PLS Component Analysis and Selection

PLS is a quantitative approach for modeling complicated relationships between predictors (X) and the responses (Y). Traditionally, the modeling of Y by means of X is done using multiple linear regression (MLR). However, MLR does not work well if X-variables (predictors) are numerous and highly correlated. PLS, on the other hand can handle data with strongly collinear, noisy and numerous X-variables. PLS was developed by Herman Wold in 1960s as an econometric technique, but later became one of the standard methods for Chemometric analysis. 22,75,76

The basic idea of PLS is that for all the predictors in the model, there exist only a few latent variables that account for most of the variation in the response. PLS aims at extracting these latent factors that account for as much of predictor (X-variables)
variation as possible while placing the primary emphasis on modeling the responses. The linear PLS model works by finding the estimates of the latent variables. These estimates are often referred as the X-scores and are represented in the matrix notation using the symbol, \( T \). The X-scores also act as predictors of \( Y \). Therefore, both \( Y \) and \( X \) are modeled by X-scores \((T)\) and thus the same underlying latent variables using:

\[
X = TP' + E \quad \text{[eqn. 3.6]}
\]

\[
Y = TC' + F \quad \text{[eqn. 3.7]}
\]

Where,

\( P \) denotes the loading of \( X \) so that the X-residuals, \( E \) are minimized;

\( C \) denotes the loading of \( Y \) so that the Y-residuals, \( F \) are minimized.

The critical step of PLS is the estimation of X-scores, which are estimated as linear combinations of the original variables with linear coefficients (weights) referred to as \( W \) in matrix notation. If we let the predictor matrix, \( X = \tilde{X}_{m,n} = [X_1, X_2, \ldots, X_n] \) with \( m \) rows corresponding to \( m \) samples and \( n \) columns corresponding to \( n \) predictor variables, then:

\[
T = XW \quad \text{[eqn. 3.8]}
\]

Such that the X-score vector \((t)\) and X-weight vector \((w)\) are related as:

\[
t_i = w_i X_{m,n} = w_{i,1} X_1 + w_{i,2} X_2 + \ldots + w_{i,n} X_n \quad i = 1, 2, \ldots, p
\]

\[\text{[eqn. 3.9]}\]
Recalling, that equation 3.2 is:

\[ \text{PC}_i = e_i X_{m,n} = e_{i,1} X_1 + e_{i,2} X_2 + \ldots + e_{i,n} X_n \quad i = 1, 2, \ldots, n \]

[eqn. 3.2]

Equation 3.2 for the derivation of principal components is presented here to compare the similarity with equation 3.9 for the derivation of PLS X-scores. The X-score vector, \( t_i \) corresponds to the principal component vector \( \text{PC}_i \) and the weight vector \( w_i \) corresponds to the eigen vector \( e_i \). Moreover, like the principal components, the X-score vectors are also orthogonal. The working of PLS is based on an algorithm known as NIPALS (Non-linear Iterative Partial Least Squares). It derives its name because the critical part of the estimation is based on a simple bivariate regression i.e., least squares regression between two variable vectors (one Y-vector and another X-vector). The PLS X-weights (\( w \)) are estimated from Y-scores (\( u \)) as:

\[ w = X'u / (u'u) \quad [\text{eqn. 3.10}] \]

In the above equation 3.10, \( u \) takes the place of the independent variable while \( X \) takes the place of the dependent variable and \( w \) takes the place of regression coefficients in a typical bivariate regression framework.\(^{77}\) In order to start the algorithm, it is assumed that \( u = y \) (where, \( y \) is one of the \( Y \) columns and is treated as the initial guess of the Y-score vector for the corresponding \( Y \) column), and the first X-weights vector, \( w \) is calculated using equation 3.10. Now that the X-weights are available, the X-scores vector, \( t \) can be estimated using:

\[ t = Xw \quad [\text{eqn. 3.11}] \]
Since the same X-scores are also used to predict Y using equation 3.7, the corresponding vector of Y-weights can be estimates as:

\[ c = \frac{Y't}{(t't)} \]  
[eqn. 3.12]

Now that the Y-weights are available, the updated Y-scores for the initially chosen Y column can be estimated using:

\[ u = \frac{Yc}{(c'c)} \]  
[eqn. 3.13]

This new estimate of Y-scores is substituted back in equation 3.10 to reiterate the algorithm. The process is terminated by verifying some sort of convergence criteria based on the relative change in the X-score vector, t:

\[ \left| \frac{(t_{\text{old}} - t_{\text{new}})}{t_{\text{new}}} \right| < \varepsilon; \text{ where } \varepsilon \text{ is small, e.g., } 10^{-6} \text{ or } 10^{-8}. \]  
[eqn. 3.14]

At the end of the above-described process, one component of the X-score matrix i.e., one X-score vector, say \( t_1 \) as described by equation 3.9 is obtained. This X-score vector \( t_1 \) corresponds to the selected Y column as the initial guess for \( u \). The X-score vector \( t_1 \) obtained via equation 3.9 in PLS analysis is equivalent to the first principal component \( \text{PC}_1 \) obtained via equation 3.2 in PCA and the first X-weight vector \( w_1 \) in PLS analysis corresponds to the first eigenvector \( e_1 \) in the principal component analysis. After the first PLS component i.e., the first X-score vector (\( t_1 \)) is obtained, the X and Y matrices are deflated by subtracting the corresponding contribution of the first component as follows:

\[ p = \frac{X't_1}{(t_1't_1)} \]  
[eqn. 3.15]

\[ X_{\text{deflated}} = X - t_1p' \]  
[eqn. 3.16]

\[ Y_{\text{deflated}} = X - t_1c' \]  
[eqn. 3.17]
From equation 3.10, it can be seen that the first weight vector \( w_1 \) is actually the first eigenvector of the combined variance-covariance matrix \( (X'Y Y'X) \) and the following weight vectors are the eigenvectors of the deflated versions of the combined variance-covariance matrix \( (X_{deflated}' Y_{deflated} Y_{deflated}' X_{deflated}) \). The entire algorithm using equations 3.10 to equation 3.17 is executed repeatedly to extract one PLS component after another until no more significant information is contained in \( X \) about \( Y \). The maximum number of PLS components \( (p) \) extracted is usually determined with the help of a cut-off on the percent residual variation left over in \( X \) and \( Y \). Most standard statistical packages including SAS® and MATLAB® implement PLS algorithm. There is much similitude between PLS and PCA and it is easy to see that a logical extension of PCA in the presence of response data is PLS.

The PLS components are obtained by maximizing the variance-covariance structure between predictor and response variables. In our proposed supervised classification method here, we extract PLS components and evaluate each PLS component for its predictive ability to distinguish various classes in a manner similar to that detailed in section 3.2.1 in the unsupervised classification method. To identify the influential PLS components that capture the information for differentiating the samples, we examine the distribution and the proportion of negative (or alternatively positive) coefficients of the \( X \)-weight vector corresponding to the PLS component of interest. Those weight vectors that have a proportion close to 0.5 correspond to PLS components that capture the contrasts in the response data and should be good predictors of class distinction. Those that have a proportion close to 0 or 1 (farthest from 0.5) are not as
contrasting and would not be good predictors. It is especially important to exclude these PLS components as predictors if they account for a high percent of overall variability in the data set because they are capable in that case of masking the class distinction. Thus, from amongst the $p$ PLS components that explain most of the variation in the data set, those $q$ PCs ($q \leq p$) are selected that have their proportions close to 0.5 as explained above.

### 3.3.2 Hierarchical $K$-means Clustering using Influential PLS Components

We propose to use the $K$-means clustering method in a manner similar to that of the method described in section 3.2.2 with the only exception being that selected influential PLS components are used instead of the influential principal components. We expect that for the same data set, improved results would be obtained with hierarchical $k$-means clustering using PLS components than with those obtained using corresponding principal components. At the same time, we also expect that number of PLS components turning out to be not influential will be much less than the corresponding principal components for the same data set under investigation. We expect as above because PLS components are obtained by maximizing the variance-covariance structure between predictor and response variables while the principal components are obtained from maximizing the predictor variance alone. Therefore, PLS components would incorporate more information relevant to class differences in response data than the corresponding principal components.
3.3.3 **Identification of key features**

PLS is essentially a prediction algorithm that focuses on predicting responses and not necessarily on understanding the relative roles of predictor variables in predicting the response. However, we believe that the similarity of PLS structure to the PCA can be exploited to identify the key features that are responsible for differentiating the classes. Correlation coefficient method and the factor loading method described in section 3.2.3 for identification of differentiating features in PCA will be employed to identify key features in PLS analysis. Correlation coefficients for features would be calculated with respect to influential PLS components and X-loadings of influential PLS components examined in order to identify the key features in a manner similar to that of unsupervised analysis.

3.4 **Conclusions**

In this work, we present a technique that uses principal component analysis along with hierarchical $k$-means clustering on differentially weighted principal components to address unsupervised classification for both binary and multi-class problems. This unsupervised methodology is extended to cases where prior knowledge of class distinctions are available i.e., to solve supervised classification problems. In the supervised scenario, PCA is replaced by PLS. We demonstrate our methods using two different data sets: (a) gene expression data sets to classify cancers (b) chemical data sets to classify toxicity/activity of compounds. Two methods to identify the differentiating features most responsible for the observed classifications are also discussed.
4.1 Introduction

In this section, we apply the proposed classification methodology to characterize the cancerous state of biological samples and demonstrate the performance of our methods with the help of two case studies. More specifically, in the first case study, we apply our methods to classify different types of leukemia based on gene expression studies carried out using microarray experiments. In the second case study, we demonstrate our methods on a gene expression data set for classifying the NCI 60 cell lines. We also identify the differentiating genes most responsible for observed classification using our methods and compare them with other techniques described in the literature.

4.2 Unsupervised Analysis: Leukemia data set

In the first case study, we consider the acute leukemia data set published by Golub et al.15 As mentioned earlier, leukemia is a type of cancer that affects the white
blood cells with an unchecked growth or proliferation of white blood cells in the bone marrow. The highly proliferating white blood cells remain immature and as a result lose their key function of immune response. White blood cells are of different types – myeloid type, lymphoid type, and lymphoid cells in turn are further subdivided into B-cell lymphocytes and T-cell lymphocytes. Each of these sub categories has a distinct role in mediating immune response and the cancer condition depends on the types of white blood cells that are involved. Although, all of the above diseased conditions are referred to as leukemia in general, different forms of leukemia require different approaches for optimal therapeutic results. Thus, one of the key challenges of cancer treatment is its accurate diagnosis, which facilitates targeting of specific therapies to distinct tumor types, thereby maximizing the efficacy of the treatment and minimizing the unwanted side effects. Conventionally, the diagnosis of cancer into its subtype is primarily based on the morphological appearance of the sample tumor tissue extracted from a patient. Sometimes, the morphological appearance of a tumor is atypical and the sample cannot be assigned to a particular subtype unambiguously or at other times, two samples present similar morphological appearance despite their distinct biological origin. In such cases, the overall gene expression profile of the tissue sample is studied using microarray experiments to aid the diagnosis.

Microarray experiments permit the measurement of thousands of genes simultaneously that are suspected to be involved in the medical condition under investigation. The acute leukemia data set in this case study contains gene expression measurements on tissue samples taken from 72 leukemia patients diagnosed with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). For each of the
tumor samples, 7129 measurements were made including measurements that correspond to 6817 genes and certain control sequences. It is reasonable to expect that although expression levels are measured for thousands of genes, only a few sets of genes typically carry relevant information to differentiate the tumor samples. The rest of the genes do not contribute to the variation across samples and if used in the analysis directly they could potentially degrade the performance of the classification strategy by masking the contribution of the few important genes. Therefore, the task of classifying tumor samples and identifying differentiating gene expression patterns amidst thousands of genes fits the framework of the methodology developed by us.

4.2.1 Principal Component Analysis and Selection

To demonstrate the unsupervised analysis, we shall not use the response data i.e., whether a particular tissue sample came from ALL or AML patient is unknown. This leaves us with the task of classifying tissue samples into their appropriate tumor type in the absence of any prior knowledge of class distinctions. Therefore, we shall apply the unsupervised classification methodology using principal component analysis and hierarchical \( k \)-means clustering as described in section 3.2. The analysis begins by performing PCA on the gene expression data matrix with \( m = 72 \) (tissue samples) and \( n = 7129 \) (gene expression measurements) dimensions. In general, PCs could be computed based on either the covariance matrix or the correlation matrix of the \( n \) variables. We believe that it is more appropriate to use covariance matrix in this case because all the elements of the gene expression data matrix are measured in the same units. Using a
correlation matrix instead, by standardizing the data is equivalent to making an arbitrary choice of measurement units. Transforming original variables could sometimes lead to loss of information, for example, it could scale down the very differences across samples that are being detected. Therefore, such transformations including standardization of covariance matrix should be used only when there is a compelling evidence to believe that the original variables are measured on scales with widely differing ranges or if the units are not commensurate due to different types of variables being involved in the analysis. Moreover, the gene expression data obtained from the raw scanned images are already normalized inorder to remove any systematic variations in the measuring system and therefore there is no pressing need to transform the original variables. Thus, it is preferable to work with the original variables and the covariance matrix for computing PCs in the case study presented here.

The maximum possible number of principal components equals the rank \((r)\) of the data matrix, and hence \(r \leq \min (m, n)\). Therefore, the rank of leukemia data set has to be less than or equal to 72. Although, as many as \(n = 7129\) PCs can be extracted, it is only meaningful to focus on the first 72. Thus, PCA is applied to obtain a set of 72 eigenvector-eigenvalue pairs for the given leukemia data such that \(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{72} \geq 0.\) PCA is implemented using the MATLAB® software (sample MATLAB programs are made available in the appendix A) to obtain \(\text{PC}_1\) corresponding to \((\lambda_1, e_1)\) that captures the most variability in the data set followed by \(\text{PC}_2\), followed by \(\text{PC}_3\) and so on. A plot depicting the cumulative percent variability explained by the PCs is shown in Figure 4.1.
Figure 4.1: A plot of cumulative percent variability accounted vs. the number of principal components in the leukemia data set for ALL/AML distinction.
The plot in Figure 4.1 shows that the first 60 PCs combined together explain about 98% of the variability in the leukemia data set; the first 20 PCs account for nearly 80% of the variability leaving out a residual variability of only about 20%. The eigenvectors corresponding to the first 60 PCs were investigated to determine those that capture contrasts in the samples and as a result could be used as predictors to classify the tumor samples. Table 4.1 and Figure 4.2 present the analysis of the first 10 eigenvectors corresponding to the 10 largest eigenvalues. Eigenvectors that have a proportion close to 0.5 correspond to PCs that capture the contrasts in the data set and should be good predictors of class distinction. Those PCs that have a proportion farthest from 0.5 are not as contrasting and therefore were not used as predictors.

<table>
<thead>
<tr>
<th>Eigenvector</th>
<th>Percent variability captured by the corresponding PC</th>
<th>Proportion of Negative coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.43</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>10.34</td>
<td>0.53</td>
</tr>
<tr>
<td>3</td>
<td>7.69</td>
<td>0.73</td>
</tr>
<tr>
<td>4</td>
<td>5.85</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>5.78</td>
<td>0.65</td>
</tr>
<tr>
<td>6</td>
<td>5.02</td>
<td>0.68</td>
</tr>
<tr>
<td>7</td>
<td>4.56</td>
<td>0.23</td>
</tr>
<tr>
<td>8</td>
<td>3.24</td>
<td>0.65</td>
</tr>
<tr>
<td>9</td>
<td>2.85</td>
<td>0.47</td>
</tr>
<tr>
<td>10</td>
<td>2.41</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 4.1: Percent variability accounted by the corresponding PC and the proportion of negative coefficients for the first 10 eigenvectors in the case of ALL/AML distinction.
Figure 4.2 shows the proportion of negative coefficients for the first 10 eigenvectors plotted against the percent variability accounted by their corresponding principal components. From Figure 4.2, it can be seen that the principal component corresponding to the first eigenvector (PC₁) would not be a good predictor of tumor classes. This could be identified by examining the proportion of negative coefficients in the eigenvector and the distribution of the eigenvector coefficients in Figure 4.3. From Figure 4.3, it can be seen that the eigenvector 2 has a bell shaped distribution centered on zero indicating that it has about the same number of positive and negative coefficients, while eigenvector 1 has most of its significant coefficients on one side of the scale (greater than zero). To select the predictive and influential PCs, we use a conservative cut-off criterion of [0.4, 0.6] for the proportion of coefficients with one sign or the other. The proportion of negative (or alternatively positive) coefficients can be seen as a measure of the predictive capability of a PC. In Figure 4.2, eigenvectors whose proportions of negative coefficients fall outside the interval [0.4, 0.6] are regarded as poor predictors of class distinction and were excluded as predictors. Eigenvectors #1, 7, 3, 6, 8 and 5 violate the cut-off criteria and hence their corresponding PCs were excluded as predictors. In the set of eigenvalue-eigenvector pairs (λᵢ, eᵢ), the most important eigenvectors correspond to the largest eigenvalues with the magnitude of an eigenvalue representing the amount of variability accounted by the respective principal component in the PC-dimensional space. In this case, the most important eigenvector is the eigenvector corresponding to PC₁ having the largest eigenvalue and therefore the largest share of variability (17.43%) in the data set as shown in Table 4.1.
Figure 4.2: A plot showing proportion of negative coefficients for the first 10 eigenvectors vs. the percent variability accounted by their corresponding principal components.
Figure 4.3: A plot showing the distribution of eigenvector coefficients for eigenvectors 1 and 2 in the ALL/AML case study.
However, $\mathbf{PC}_1$ has the proportion of negative coefficients equal to 0.11 indicating that nearly 90% of its eigenvector coefficients are all positive. Thus, it is necessary to exclude this PC ($\mathbf{PC}_1$) from the set of predictors for class distinction because this PC does not contribute significantly to the contrasts in the data set, although it carries the most percent variability in the data set. In fact, if included $\mathbf{PC}_1$ may have the undesired effect of masking the true classification. On the other hand, $\mathbf{PC}_2$ corresponding to second eigenvector captures about 10.34% of the total variability and has a proportion of negative coefficients equal to 0.53, which suggests that it should be retained in the set of predictive PCs. In fact, $\mathbf{PC}_2$ will be the most important PC for classification purposes because it is the PC with next largest eigenvalue, after excluding $\mathbf{PC}_1$. Although the importance of such non-differentiating principal components is mentioned in theory\textsuperscript{61,62}, we have observed that this is generally ignored in many bioinformatics applications of PCA.

### 4.2.2 Classification Results and Discussion

From the first 60 PCs that explain about 98% of the variability in the leukemia data set, a predictive subset of PCs was selected as outlined above, which were used as predictors in a $k$-means clustering algorithm to partition the samples into two ($k = 2$) groups. The two groups were compared to the known classification of ALL and AML classes and it was found that the 72 samples separate with a reasonably high accuracy of 93% (with 5 misclassifications) even in the mere two dimensional space formed by $\mathbf{PC}_2$ and $\mathbf{PC}_4$. As seen in the Figure 4.4, the five misclassified samples are shown in
diamonds. The clustering results were found to remain stable with the variation in the number of PCs included from the predictive subset. The tumor samples were grouped in the same way no matter the first 2, 10 or 40 PCs from the predictive subset was used. In fact, the accuracy improved to 97% (2 misclassifications) when 30 or more predictive PCs were used. K-means clustering was repeated several thousand times with different random seeds in order to avoid being trapped in a local optimum and to ensure reproducibility. The algorithm converged to a clustering result that was obtained with the highest frequency in these several thousand repeats and was accepted as the nearest optimal solution.

![Plot showing prediction results for unsupervised classification of ALL/AML tumor samples in the dimensional space of PC2 and PC4. The five misclassified samples are shown in diamonds.](image)

Figure 4.4: A plot showing prediction results for unsupervised classification of ALL/AML tumor samples in the dimensional space of $\text{PC}_2$ and $\text{PC}_4$. The five misclassified samples are shown in diamonds.
A summary of the ALL/AML class prediction results obtained through our approach is presented in Tables 4.2 and 4.3. Table 4.2 shows that all the ALL samples were predicted accurately while five AML samples (samples #28, 29, 35, 38 and 66) were misclassified and grouped with ALL instead. This shows that our unsupervised classification method predicts with 93% accuracy. Next, we carried out validation of the above results through leave-one out cross validation method. Here, we left out a tissue sample out of our data set (hence, \( m = 71 \)) and carried out the principal component analysis as before. We computed the principal components, examined the eigenvector coefficients and selected predictive principal components. The left out sample was transformed from the gene-space to the PC-space using equation 3.2. The predictive principal components were used as predictors to separate the 71 samples, excluding the sample that was left out, into two clusters. Now, the distance of the left-out sample in the predictive PC-space is computed from both the cluster centroids and the sample is grouped along with the cluster to which it is nearest to. The left out sample was evaluated to see if it classified with its appropriate class. This process was repeated 72 times leaving out one tissue sample each time and predicting its class membership. Table 4.3 presents the results of leave-one-out cross validation prediction. It was found that a 90% leave-one-out cross validation accuracy is obtained (The 7 misclassified AML samples: #28, 29, 35, 38, 62, 64 and 66 were grouped with ALL). Given the fact that our approach did not use any prior knowledge of class distinction, the results obtained are highly encouraging. The high prediction accuracy of this unsupervised technique can be attributed to the inherent multivariate nature of PCA and the critical step of proper selection of predictive subset of PCs that capture the class distinction in the data, not
merely the overall variability. The kappa measures of agreement were 0.94 and 0.88 for regular and cross-validation prediction contingency tables respectively. The gamma measures of agreement were 1 for both regular and cross-validation prediction contingency tables.

<table>
<thead>
<tr>
<th>PREDICTED CLASSIFICATION</th>
<th>TRUE CLASSIFICATION</th>
<th>ALL</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>47</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td>0</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Prediction results for ALL/AML classification using unsupervised method.

<table>
<thead>
<tr>
<th>PREDICTED CLASSIFICATION</th>
<th>TRUE CLASSIFICATION</th>
<th>ALL</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>47</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td>0</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Leave-one-out cross validation prediction results for ALL/AML classification using unsupervised method.
In the Table 4.3, the subset of 54 samples predicted as ALL was now treated in the same manner as the original data set of 72 samples to see if the 54 samples can be separated. Repeating the process of PCA, followed by principal component selection and \( k \)-means clustering led to the separation of 54 samples into two subsets – one consisting of 47 samples that truly belong to ALL and the other consisting of 7 AML samples that were misclassified as ALL samples in the first separation step. Now, the subset of 47 samples was further considered for analysis. The 47 tumor samples that correspond to ALL patients were now analyzed to see if any further sub-classification is possible. PCA was done on the original data matrix of 47 samples and 7129 gene expression measurements. The first 10 PCs that captured 70% of the variation in the data set were further screened to select a subset of predictive PCs. The plot in Figure 4.5 suggests that \( \text{PC}_1, \text{PC}_3, \text{PC}_4 \) and \( \text{PC}_6 \) should be removed from the predictive subset for the purpose of clustering the samples as they have the proportion of negative coefficients farthest from 0.5 and outside the cut-off criteria of \([0.4,0.6]\) used in our approach. \( K \)-means clustering was performed with \( k = 2 \) groups to partition the 47 samples into two clusters. The classification summary is shown in Table 4.4. Once again a high prediction accuracy of 91% was found for this subset of data (Kappa = .70 and Gamma = .97). The two groups partitioned here closely matched with the samples that differed in their lineage. One group had a B-cell lineage and the other a T-cell lineage. Lymphocytes are divided into B-cell and T-cell origin and although both play a role in the immune response of the body, there are marked differences in the way they mediate the response. Thus, it was possible to identify biological differences through genetic studies, which could be explored to target specific therapies to distinct tumor types in order to improve efficacy and minimize toxicity.
Figure 4.5: A plot showing proportion of negative coefficients for the first 10 eigenvector vs. the percent variability accounted by their corresponding principal components for the subset of 47 ALL samples.
In Figure 4.5, it should be noted that $\text{PC}_1$ and $\text{PC}_3$ capture about 21% and 7% of the total variability respectively and therefore are capable of masking the clustering results if not excluded as predictors in $k$-means clustering process. Despite the fact that $\text{PC}_1$ accounts for nearly 21% of the total variability in the data set, 70% of the eigenvector coefficients for this PC are negative. This is yet another example when a PC does not capture contrasts in the data despite being able to account for a large percent variability in the data.

<table>
<thead>
<tr>
<th>PREDICTED CLASSIFICATION</th>
<th>TRUE CLASSIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B-CELL ALL</td>
</tr>
<tr>
<td>B-CELL ALL</td>
<td>37</td>
</tr>
<tr>
<td>T-CELL ALL</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.4: Prediction results for B-cell ALL/T-cell ALL class distinction within the 47 ALL samples using unsupervised method.
As expected, the prediction accuracy falls steeply when we include influential yet poor predictors in the predictive set of PCs for the purpose of predicting class membership. As shown in Table 4.5, in the case of ALL/AML distinction, the prediction accuracy falls from 93% to 60% (29 misclassifications) with the mere inclusion of $\text{PC}_1$ in the predictive set. While, in the case of B-cell/T-cell ALL distinction, the prediction accuracy falls from 91% to 66% (16 misclassifications) as shown in Table 4.6 with the mere inclusion of $\text{PC}_1$ in the predictive set. In both the cases, $\text{PC}_1$ is predominantly responsible for masking the true classifications as it accounts for nearly 17% of the variability in ALL/AML data set while $\text{PC}_1$ accounts for 21% variability in the B-cell/T-cell ALL data set respectively.

Prediction results presented in Tables 4.5 and 4.6 underscore the importance of evaluating the predictive ability of each principal component that is used as an input to the $k$-means classification step.

The prediction accuracies of this unsupervised method are comparable with most of the supervised classification schemes presented in prior literature for this same data set.\textsuperscript{15,53,54} Thus, we demonstrate the effective use of PCA combined with recursive $k$-means partitioning to address the unsupervised multi-class classification problem. It should be re-iterated that one of the key steps in our approach is the evaluation of principal components in order to identify good predictors of class distinction i.e., those principal components that not only account for a high percent variability in the data set but also contain information to differentiate samples based on their underlying class distinctions.
**Table 4.5:** Prediction results for ALL/AML classification after the inclusion of PC\(_1\) in the predictive subset of PCs.

<table>
<thead>
<tr>
<th>TRUE CLASSIFICATION</th>
<th>ALL</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>33</td>
<td>15</td>
</tr>
<tr>
<td>AML</td>
<td>14</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table 4.6:** Prediction results for B-cell ALL/T-cell ALL classification after the inclusion of PC\(_1\) in the predictive subset of PCs.

<table>
<thead>
<tr>
<th>TRUE CLASSIFICATION</th>
<th>B-CELL ALL</th>
<th>T-CELL ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-CELL ALL</td>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>T-CELL ALL</td>
<td>13</td>
<td>6</td>
</tr>
</tbody>
</table>
4.2.3 Identification of Differentiating Genes

Having achieved success with the prediction of the leukemia sub-classes, the attention was turned towards identifying important genes that are responsible for differentiating the tumor subtypes. It was hypothesized, that if the discovered classes reflect true underlying biological differences, then, the principal component(s) that classified them should capture the information required for class distinction. Therefore, genes that have a high correlation coefficient or PC-loadings between expression level and principal component scores are expected to be either under- or over-expressed in a particular tumor type and vice-versa, thereby carrying in them important information that contrasts the tumor samples in the dataset.

First, we applied the correlation method to identify the differentiating features. For each of the PCs in the predictive subset, a list of genes having high linear correlations was identified. A list of such genes was compiled to produce a list of differentiating genes using equation 3.5 as described in the chapter, “Systems and Methods”. Genes are sorted by weighting the absolute correlation between a gene and a particular PC in the predictive subset with the corresponding percent variability accounted by that PC. Since this list is compiled from the genes identified via each of the independent predictive principal components, the corresponding genes should capture a wide variety of cancer targets for therapeutic intervention. Table 4.7 shows a partial list of most differentiating genes identified and sorted with an overall rank #1 to 20 by the correlation method along with their corresponding overall ranks obtained by the feature loading method.
<table>
<thead>
<tr>
<th>Gene Accession Number</th>
<th>Gene Description</th>
<th>Rank (Correlation Method)</th>
<th>Rank (Loading Method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X95735_at</td>
<td>Zyxin</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>X14008_rn</td>
<td>Lysozyme gene (EC 3.2.1.17)</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Lysozyme</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M27891_at</td>
<td>CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>M19045_f_at</td>
<td>LYZ Lysozyme</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>L09209_s_at</td>
<td>APLP2 Amyloid beta (A4) precursor-like protein 2</td>
<td>5</td>
<td>63</td>
</tr>
<tr>
<td>J03801_f_at</td>
<td>LYZ Lysozyme</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>M16038_at</td>
<td>LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog</td>
<td>7</td>
<td>138</td>
</tr>
<tr>
<td>X52056_at</td>
<td>SPI1 Spleen focus forming virus (SFFV) proviral integration oncogene spi1</td>
<td>8</td>
<td>98</td>
</tr>
<tr>
<td>X62320_at</td>
<td>GRN Granulin</td>
<td>9</td>
<td>29</td>
</tr>
</tbody>
</table>

Continued

Table 4.7: A partial list of 20 most differentiating genes identified through overall correlation rank method for ALL/AML classification.
Table 4.7 continued

<table>
<thead>
<tr>
<th>Probe ID</th>
<th>Gene Description</th>
<th>Value_1</th>
<th>Value_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>X64072_s_</td>
<td>SELL Leukocyte adhesion protein beta subunit</td>
<td>10</td>
<td>46</td>
</tr>
<tr>
<td>M11147_at</td>
<td>FTL Ferritin, light polypeptide</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>X61587_at</td>
<td>ARHG Ras homolog gene family, member G (rho G)</td>
<td>12</td>
<td>108</td>
</tr>
<tr>
<td>M63138_at</td>
<td>CTSD Cathepsin D (lysosomal aspartyl protease)</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>M81695_s_</td>
<td>ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)</td>
<td>14</td>
<td>195</td>
</tr>
<tr>
<td>X16546_at</td>
<td>RNS2 Ribonuclease 2 (eosinophil-derived neurotoxin; EDN)</td>
<td>15</td>
<td>217</td>
</tr>
<tr>
<td>M62762_at</td>
<td>ATP6C Vacuolar H+ ATPase proton channel subunit</td>
<td>16</td>
<td>105</td>
</tr>
<tr>
<td>M32304_s_</td>
<td>TIMP2 Tissue inhibitor of metalloproteinase 2</td>
<td>17</td>
<td>200</td>
</tr>
</tbody>
</table>

Continued
In the second method using feature loadings, we identified a list of genes that were heavily loaded for each of the PCs in the predictive subset. Like in the correlation method described above, genes were sorted for each influential PC and overall ranks of genes were computed in a manner similar to that of correlation method using equation 3.5 as described earlier. Table 4.8 shows a partial list of most differentiating genes identified and sorted with an overall rank #1 to 20 by the feature loading method along with their corresponding overall ranks obtained by the correlation method. Comparing Tables 4.7 and 4.8, one can see that there is some overlap between the ranks of genes but for the most part, different genes were identified using both the methods. As explained earlier in the “Systems and Methods” section, the two methods complement each other because correlation method relies on the univariate treatment of genes while the feature loading method works by the multivariate treatment of genes. Thus, significant diversity in the biomarker space can be achieved, thereby avoiding the problem of identifying highly correlated genes that contain redundant information.
<table>
<thead>
<tr>
<th>Gene Accession Number</th>
<th>Gene Description</th>
<th>Rank (Loading Method)</th>
<th>Rank (Correlation Method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M27891_at</td>
<td>CST3 Cystatin C</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(amyloid angiopathy and cerebral hemorrhage)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M11147_at</td>
<td>FTL Ferritin, light polypeptide</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>M26311_s_at</td>
<td>GB DEF = Cystic fibrosis antigen mRNA</td>
<td>3</td>
<td>105</td>
</tr>
<tr>
<td>Y00787_s_at</td>
<td>INTERLEUKIN-8 PRECURSOR</td>
<td>4</td>
<td>69</td>
</tr>
<tr>
<td>Z19554_s_at</td>
<td>VIM Vimentin</td>
<td>5</td>
<td>234</td>
</tr>
<tr>
<td>Y00433_at</td>
<td>GPX1 Glutathione peroxidase 1</td>
<td>6</td>
<td>96</td>
</tr>
<tr>
<td>X14008_rna1_f_at</td>
<td>Lysozyme gene (EC 3.2.1.17)</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>V00594_s_at</td>
<td>Metallothionein isoform 2</td>
<td>8</td>
<td>239</td>
</tr>
<tr>
<td>M19045_f_at</td>
<td>LYZ Lysozyme</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>M63438_s_at</td>
<td>GLUL Glutamate-ammonia ligase (glutamine synthase)</td>
<td>10</td>
<td>301</td>
</tr>
</tbody>
</table>

Continued

Table 4.8: A partial list of 20 most differentiating genes identified through feature loading method for ALL/AML classification.
Table 4.8 continued

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Gene</th>
<th>Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>M96326_rna1_at</td>
<td>Azurocidin gene</td>
<td>11</td>
<td>280</td>
</tr>
<tr>
<td>J03077_s_at</td>
<td>PSAP Sulfated glycoprotein 1</td>
<td>12</td>
<td>53</td>
</tr>
<tr>
<td>J03801_f_at</td>
<td>LYZ Lysozyme</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>S71043_rna1_s_at</td>
<td>Ig alpha 2=immunoglobulin A heavy chain allotype 2 {constant region, germ line}</td>
<td>14</td>
<td>157</td>
</tr>
<tr>
<td>M69043_at</td>
<td>MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3</td>
<td>15</td>
<td>129</td>
</tr>
<tr>
<td>M19507_at</td>
<td>MPO Myeloperoxidase</td>
<td>16</td>
<td>408</td>
</tr>
<tr>
<td>M21119_s_at</td>
<td>LYZ Lysozyme</td>
<td>17</td>
<td>35</td>
</tr>
<tr>
<td>M84526_at</td>
<td>DF D component of complement (adipsin)</td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td>J04456_at</td>
<td>LGALS1 Ubiquinol-cytochrome c reductase core protein II</td>
<td>19</td>
<td>211</td>
</tr>
<tr>
<td>M87789_s_at</td>
<td>(hybridoma H210) anti-hepatitis A IgG variable region, constant region, compleme</td>
<td>20</td>
<td>379</td>
</tr>
</tbody>
</table>
One of the most important genes for AML/ALL class distinction, according to the correlation method was found to be to the focal adhesion protein, zyxin (Gene Accession Number: X95735). From previous literature, this gene was identified as not only an informative gene\textsuperscript{78,79} for AML/ALL distinction, but also the most important gene in some studies aimed at finding the optimal and best prediction models for AML/ALL distinction\textsuperscript{80,81}. Another gene identified at the top of the list through both correlation and feature loading methods, cystatin C, a proteolytic inhibitor is known to be a differential gene in large granular lymphocyte (LGL) leukemia\textsuperscript{82} and was featured as one of the three genes that are good targets for investigating basic biology of ALL/AML\textsuperscript{83}. Granulin, also at the top of our list, was identified as a gene significantly under expressed in a study of translocation-associated fusion genes in AML\textsuperscript{84}. Recent studies have found the structural basis for the co-activation of protein kinase-B involving pleckstrin homologous domains that bind to T-cell Leukemia-1 (TCL1) family proto-oncogenes\textsuperscript{85}. Our model results support these experimental findings, suggesting that pleckstrin homologous domain could be an attractive drug target against leukemia caused by TCL1 proteins.

Other genes identified also include the primary leukemia target, topoisomerase II beta and a host of other clinically used marker genes for leukemia diagnosis. These include CD 19, a marker gene for B-cell ALL; CD7, a highly sensitive gene for T-cell ALL; CD9 antigen which, is expressed during early stages of B-cell differentiation or activation; interleukin 1 beta involved in B-cell maturation and proliferation\textsuperscript{86}; HOXB (HOXB3 and HOXB2) genes which are known to be expressed in AML and not expressed in chronic myeloid leukemia (CML)\textsuperscript{87}; CD33 and CD14 which are expressed in AML\textsuperscript{88}; CD63 antigen which is associated with early stages of tumor progression and
onco-genes like c-MYB and E2A. For comparison, the list of genes identified here using correlations with $\text{PC}_2$ alone contained 19 out of the top 50 genes (38%) predicted by Golub’s supervised method as differentiating the AML/ALL cancer types.

Additionally, adenosine deaminase (ADA) which has been proposed as a novel marker for chronic lymphocytic leukemia is also identified. Other genes identified by this approach that have been reported to overexpress in leukemia from previous literature include HMG1 and chemokine receptor-4. Also, genes that are generally over expressed in cancerous tissues (but not specific to any cancer subtype) in comparison to normal tissues were also identified. These include Prothymosin alpha (PTMA) which is associated with cell proliferation, Amyloid beta (A4) precursor-like protein 2 (APLP2). Glycerol 3-phosphate dehydrogenase 1 (GDP1) was also identified which is found to be generally down regulated in the tumor tissues.

### 4.2.4 Unsupervised Analysis: NCI 60 Data Set

After successful demonstration of our methods using the acute leukemia data set, we focused on validating our methods by applying them to a different data set. In the second case study, we selected the data set published by Ross et al at the National Cancer Institute. We used this gene expression data set to illustrate that the 60 cell lines could be separated with reasonable accuracy by applying the recursive k-means clustering with a selected subset of PCs in a similar manner as outlined for the acute leukemia data set. The NCI 60 data set consists of 60 cancer cell lines from the following subgroups: breast (BRE), central nervous system (CNS), colon (COL), leukemia (LEU), lymph node.
system (LNS), melanoma (MEL), ovarian (OVA), prostate (PRO), and renal (REN). For each of the cell lines, measurements were made for over 7000 genes but here we focus on a subset of 3747 gene expressions. For the unsupervised analysis, we shall pretend that the class labels for the samples are unknown and try to predict their class membership. Here, \( m = 60 \) and \( n = 3747 \) data matrix was the starting point and we computed the 60 eigenvectors and the corresponding PCs using the covariance matrix of the data set. The \( p = 10 \) eigenvectors captured about 46% of the variation in the data set and these were subsequently analyzed for their predictive capability by examining their proportion of negative eigenvector coefficients in a similar manner described in the acute leukemia data set. Figure 4.6 illustrates the analysis of the first 10 eigenvectors corresponding to the first 10 PCs. None of the eigenvectors, except \( \text{PC}_5 \), violates the cutoff criteria of \([0.4, 0.6]\) for the proportion of eigenvector coefficients with one sign or the other. Hence, all of the top 10 PCs, except \( \text{PC}_5 \), were included in the predictive subset of NCI 60 cell lines. The rest of the nine PCs out of the first ten PCs contribute about 43% of the variability in this data set and all of these appear in the predictive subset. Even the \( \text{PC}_5 \), that is excluded from the predictive set of PCs falls outside the cutoff criteria of \([0.4, 0.6]\) for the proportion of eigenvector coefficients only by a small margin. In addition, \( \text{PC}_5 \), accounts for only 3.5% variability in the data set suggesting that it might not be very influential even if included in the predictive sub set of PCs for the purpose of classification. Therefore, we expect that the classification of samples would remain relatively unchanged with either the inclusion of exclusion of \( \text{PC}_5 \) from the predictive sub set. However, it is better to exclude it from the predictive sub set as it violates the cutoff criteria.
Figure 4.6: A plot showing proportion of negative coefficients for the first 10 eigenvector vs. the percent variability accounted by their corresponding principal components for the NCI 60 data set.
Next, the $k$-means clustering algorithm was applied using the predictive set of PCs and divided the 60 cell lines into two ($k = 2$) groups. Each of the subgroups was then further divided into two groups using $k$-means clustering. This process is recursively applied in a binary fashion, where the samples are grouped into one of the two groups at each stage until there are no more stable clusters possible. The end of the recursive process was signaled when the two subgroups were obtained with one group having extremely small number (in this case, one of the subgroups had only one observation) of observations in it. At the end of the recursive $k$-means clustering, the 60 cell lines were partitioned into meaningful groups (BRE, CNS, COL, LEU, LNS, MEL, OVA, PRO, and REN) with few misclassifications. A summary of the recursive $k$-means partitioning is shown in Figure 4.7 below. At each stage of sub-division, PCA was performed to obtain the new PCs corresponding to the subgroup and an analysis of eigenvectors was carried out to determine the predictive subset of PCs. $K$-means was applied at each stage using the corresponding predictive subset.

Figure 4.7 illustrates the results from the unsupervised classification of the 60 cell lines into various groups. At the first step, COL and LEU (Group_1) separate from the rest of the cell lines (Group_2). At this point, BRE-MCF7, BRE-T-47D and LNS-NCI-H522 are also grouped along with COL and LEU. At the next sub division however, LEU (Group_1_2) separate out clearly from the rest (Group_1_1). Further subdividing Group_1_1, we found that COL (Group_1_1_2) separate out clearly from the rest (Group_1_1_1). If we try to further subdivide the COL cell lines, they do not separate out well. One of the resulting subgroups has just one observation (COL_HT29), indicating that the possible end of recursive clustering for this node.
Figure 4.7: Partitioning of the NCI 60 cell lines using recursive $k$-means clustering process.
The rest of the binary tree was completed in a similar manner and the separation of the 60 cell lines was observed with few misclassifications. Thus, we have demonstrated the applicability of the proposed unsupervised classification technique to the NCI 60 cancer cell lines data set. Genes that are most responsible for differentiating the classes are identified in a similar manner to that of the leukemia data set using the two methods – correlation method and the feature loading method. The flowchart in Figure 4.8 briefly summarizes the key steps of our unsupervised methodology. The unsupervised methodology begins by performing principal component analysis on the original predictor variables. A small number ($p$) of principal components are selected such that they account for a reasonable amount of variation in the original predictor variables. These $p$ principal components would then be scrutinized for their predictive ability by examining the proportion of eigenvector coefficients to identify a sub set of $q$ principal components that contain the information to distinguish the samples. The $q$ principal components that are good predictors act as inputs to the $k$-means clustering process, which is performed recursively. The supervised methodology as described in the “Systems and Methods” chapter is quite similar to the unsupervised methodology except for the first step where PLS is used in place of PCA to benefit from the availability of response data. The supervised methodology is demonstrated using the acute leukemia data set in the next section of this chapter.
Step 1: Perform PCA on raw data

Step 2: Select $p$ PCs that account for a reasonable amount of variation in the dataset.

Step 3: Screen those PCs out, which have a proportion of negative (positive) coefficients far from 0.5 using a cutoff criteria. As a result, obtain the predictive subset of $q$ PCs.

Step 4: Use the predictive subset of $q$ PCs to group the samples via $k$-means clustering with $k=2$ groups. Repeat $k$-means several thousand times with random seeds to ensure cluster stability. As a result, either stable sub-clusters are obtained or no more clustering is indicated if sub-clusters are unstable for this particular node.

Step 5: Within each of the sub-clusters identified in Step 4, repeat steps 1-4 until no more stable clusters are obtained.

Step 6: To identify the genes responsible for an obtained classification, select the genes that have a high correlation or loadings with the principal component scores at that stage.

Figure 4.8: Flowchart depicting our unsupervised methodology of recursive $k$-means clustering using predictive subset of PCs.


4.4 Supervised Analysis: Leukemia data set

The leukemia data set that was used to demonstrate our unsupervised methodology is reconsidered here in the context of supervised analysis. This demonstration of the supervised methodology would be presented in a concise manner compared to the unsupervised methodology presented in earlier section because the outline of both the methods is similar with the exception of the first step where principal component analysis is replaced with PLS analysis. This being the only difference in the supervised and unsupervised methodologies, it makes the transition between the two methods straightforward. However, it is also the key step in the supervised scenario as it deals with the multivariate analysis of relating the predictor variables to the response variables. Partial least squares (PLS) aims at extracting latent factors that account for as much of predictor (X-variables) variation as possible while placing the primary emphasis on modeling the responses. As described in the section 3.3.1, the X-score vector, \( t_i \) corresponds to the principal component vector \( PC_i \) and the weight vector \( w_i \) corresponds to the eigen vector \( e_i \). Moreover, like the principal components, the X-score vectors are also orthogonal.

We implemented PLS algorithm using JMP® software available from the SAS Institute Inc. First, we extracted PLS components from the expression patterns of all the 72 samples. These PLS components would then serve as predictors to classify any new samples. Table 4.9 presents a summary of the extracted PLS components. It can be seen that the first 10 PLS components account for nearly 43% variability in \( X \) and 100% variability in \( Y \). In fact, there is no more significant information contained in \( X \) about \( Y \) after the extraction of the first four PLS components.
<table>
<thead>
<tr>
<th>PLS Component Number</th>
<th>Percent Variability Accounted in X</th>
<th>Cumulative Percent Variability Accounted in X</th>
<th>Percent Variability Accounted in Y</th>
<th>Cumulative Percent Variability Accounted in Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.625</td>
<td>9.625</td>
<td>64.23</td>
<td>64.23</td>
</tr>
<tr>
<td>2</td>
<td>9.528</td>
<td>19.15</td>
<td>19.54</td>
<td>83.77</td>
</tr>
<tr>
<td>3</td>
<td>8.585</td>
<td>27.74</td>
<td>6.662</td>
<td>90.43</td>
</tr>
<tr>
<td>4</td>
<td>2.52</td>
<td>30.26</td>
<td>7.344</td>
<td>97.78</td>
</tr>
<tr>
<td>5</td>
<td>3.135</td>
<td>33.39</td>
<td>1.389</td>
<td>99.17</td>
</tr>
<tr>
<td>6</td>
<td>2.159</td>
<td>35.55</td>
<td>0.663</td>
<td>99.83</td>
</tr>
<tr>
<td>7</td>
<td>2.14</td>
<td>37.69</td>
<td>0.116</td>
<td>99.94</td>
</tr>
<tr>
<td>8</td>
<td>2.079</td>
<td>39.77</td>
<td>0.041</td>
<td>99.99</td>
</tr>
<tr>
<td>9</td>
<td>1.998</td>
<td>41.77</td>
<td>0.01</td>
<td>100</td>
</tr>
<tr>
<td>10</td>
<td>1.562</td>
<td>43.33</td>
<td>0.003</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.9: A list of first 10 PLS components and their corresponding percent variation accounted in predictor (X) and response (Y) variables for the acute leukemia data set.
Nevertheless, we will examine all the 10 PLS components for their predictive ability by examining the proportion of weight vector coefficients for each PLS component. Like in the unsupervised scenario, we shall use a selection criteria of $[0.4, 0.6]$ for the proportion. Any PLS component whose proportion of negative (or alternatively positive) weight vector coefficients fall outside the cutoff criteria of $[0.4, 0.6]$ would not be selected into the predictive subset of PLS components. The effect of including such PLS components into the predictive subset and excluding them from the predictive subset will be investigated by comparing the predictive accuracy of the two cases. Figure 4.9 shows a plot of the proportion of negative coefficients for the first 10 weight vectors corresponding to the PLS components. It can be seen that PLS components #3 and #4 violate the cutoff criteria. Consequently, we expect that these two components would not be good predictors of class distinction. From Table 4.9, it should also be noted that these two components put together account for 14% of the variability in response data as compared to 84% variability accounted by the first two PLS components (#1 and #2) together. Therefore, we expect that the components #3 and #4 would be relatively less influential when compared to components #1 and #2. The rest of the six PLS components (#5 to #10) fall well within the cutoff criteria and therefore should be good predictors of class distinction. However, all of these six components put together account for mere 2% variability in the response data suggesting that they would not be highly influential on the class predictions.
Figure 4.9: A plot showing proportion of negative coefficients for the first 10 weight vectors corresponding to the PLS components in the supervised analysis of leukemia data.
First, we performed the $k$-means clustering using all the first 10 PLS components (including components #3 and #4) as the predictors. $K$-means clustering is performed with $k = 2$ groups as usual with several thousand random seeds to identify stable clusters and ensure reproducibility. The prediction results as shown in Table 4.10 indicate that the samples are predicted with 100% accuracy even with 10,000 randomization runs. Not even a single sample is misclassified. Now, we perform the $k$-means clustering using only eight out of the first ten PLS components (excluding components #3 and #4) as the predictors. The process is repeated in exact same manner as above i.e., $k$-means clustering is performed with $k = 2$ groups with ten thousand random seeds to identify stable clusters and ensure reproducibility. The prediction results as shown in Table 4.10 indicate that there are no misclassifications even in this case and all the 72 samples are predicted with 100% accuracy. Although, PLS components #3 and #4 were included, they did not affect the classification accuracy of our method and it is reasonable to expect so because these two components are not as influential as the first two components (#1 and #2) due to the relative amounts of variability they account for in the response data. Moreover, the prediction accuracy remained the same (100%) even with the inclusion or exclusion of the components #5 to #10. This result is also not surprising because these six components account for only a minor percent variability (2%) and are not influential when compared to the first two PLS components. In fact, the first two components are the only two influential PLS components that are required to make good predictions on any new samples in future. The rest of the components do not carry a significant amount of information to distinguish the tumor samples.
Table 4.10: Prediction results for ALL/AML classification using all the first 10 PLS components as predictors in the supervised method.

Table 4.11: Prediction results for ALL/AML classification with PLS components #3 and #4 excluded from the first 10 PLS components as predictors in the supervised method.
Now, we turned our attention to validate the prediction accuracy of our supervised method when new samples are presented for classification. As for the most supervised techniques, the leukemia data was split up into two sets – a training set and a test set. For the purpose of comparison, we partitioned the data set in a similar fashion as mentioned in Golub et al with 38 tumor samples allocated to the training set and 34 samples to the test set. The PLS components were extracted using the training set data and predictions were made on the test set. Table 4.12 presents a summary of the extracted PLS components. It can be seen that the first nine PLS components account for nearly 49% variability in $X$ (predictor variables) and 100% variability in $Y$ (response variable). In fact, there is no more significant information contained in $X$ about $Y$ after the extraction of first four PLS components. Nevertheless, we will examine all the 9 PLS components for their predictive ability by examining the proportion of weight vector coefficients. Like in the unsupervised scenario, we shall use a selection criteria of [0.4, 0.6] for the proportion. Figure 4.10 shows a plot of the proportion of negative coefficients for the first 9 weight vectors corresponding to the PLS components. It can be seen that PLS components #3 and #4 violate the cutoff criteria. Consequently, we expect that these two components would not be good predictors of class distinction. From Table 4.12, it should also be noted that these two components put together account for 8% of the variability in response data as compared to 91% variability accounted by the first two PLS components (#1 and #2) together. Therefore, we also expect that the components #3 and #4 would be relatively less influential when compared to components #1 and #2. The rest of the five PLS components (#5 to #9) fall well within the cutoff criteria and therefore should be good predictors of class distinction. However, all of these five components put together
<table>
<thead>
<tr>
<th>PLS Component Number</th>
<th>Percent Variability Accounted in X</th>
<th>Cumulative Percent Variability Accounted in X</th>
<th>Percent Variability Accounted in Y</th>
<th>Cumulative Percent Variability Accounted in Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10.71</td>
<td>10.71</td>
<td>74.25</td>
<td>74.25</td>
</tr>
<tr>
<td>2</td>
<td>9.439</td>
<td>20.15</td>
<td>16.36</td>
<td>90.61</td>
</tr>
<tr>
<td>3</td>
<td>8.514</td>
<td>28.67</td>
<td>5.68</td>
<td>96.29</td>
</tr>
<tr>
<td>4</td>
<td>7.082</td>
<td>35.75</td>
<td>2.905</td>
<td>99.2</td>
</tr>
<tr>
<td>5</td>
<td>2.172</td>
<td>37.92</td>
<td>0.708</td>
<td>99.91</td>
</tr>
<tr>
<td>6</td>
<td>2.81</td>
<td>40.73</td>
<td>0.073</td>
<td>99.98</td>
</tr>
<tr>
<td>7</td>
<td>2.946</td>
<td>43.68</td>
<td>0.015</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>2.947</td>
<td>46.62</td>
<td>0.004</td>
<td>100</td>
</tr>
<tr>
<td>9</td>
<td>2.6</td>
<td>49.22</td>
<td>6e-4</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4.12: A list of 9 PLS components and their corresponding percent variation accounted in predictor (X) and response (Y) variables for the 38 samples in the training set of acute leukemia data.
Figure 4.10: A plot showing proportion of negative coefficients for the first 9 weight vectors corresponding to the PLS components of the 38 samples in the training set of acute leukemia data.
account for mere 1% variability in the response data, suggesting that they would not be highly influential on the class predictions. This is the exact same trend as was observed in the entire leukemia data set consisting of all the 72 samples. From this observation, we can infer that the training set is a good representation of the original data set and therefore expect that the PLS components extracted from the training set would perform in a similar manner as to the components extracted from the entire leukemia data set. The prediction results as shown in Table 4.13 indicate that there is one misclassification in this case. The sole misclassification was for sample #66, which was truly an AML sample but was grouped with other ALL samples in the prediction. Inclusion or exclusion of PLS components #3 and #4 did not affect the classification efficiency of our method as expected, because these two components are not as influential as the first two components (#1 and #2) due to the relative amounts of variability they account for in the response data. Moreover, the prediction accuracy remained the same (97%) even with the inclusion or exclusion of the components #5 to #9. This result is not surprising because these six components account for only a minor percent variability (1%) and are not influential when compared to the first two PLS components. Now, we repeated the above process by making 20 random selections of the training and test sets with the number of training samples fixed at 38 and test samples fixed at 34 respectively. The average prediction efficiency for these 20 randomizations is shown in Table 4.14, where we obtain a mean prediction efficiency of 96% (with 1.4 average misclassifications out of 34). In all the 20 randomization runs, it was the AML samples that misclassified as ALL.
<table>
<thead>
<tr>
<th>TRUE CLASSIFICATION</th>
<th>ALL</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>AML</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

Table 4.13: Supervised prediction results of ALL/AML classification for a test set of 34 samples using PLS components derived from a training set of 38 samples.

<table>
<thead>
<tr>
<th>TRUE CLASSIFICATION</th>
<th>ALL</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>20</td>
<td>1.4</td>
</tr>
<tr>
<td>AML</td>
<td>0</td>
<td>12.6</td>
</tr>
</tbody>
</table>

Table 4.14: Supervised prediction results of ALL/AML classification for 20 random runs on a test set of 34 samples using PLS components derived from a training set of 38 samples.
4.5 Conclusions

Classification of cancer based on gene expression studies is demonstrated using the unsupervised and supervised methods developed by us. Both the methods are analogous to each other except for the first step of reducing the dimensionality of data. Unsupervised methodology uses PCA to extract few principal components that aim at capturing the variance-covariance structure of all the original predictor data. On the other hand, the supervised method uses PLS to extract few components that capture the variance-covariance structure of original predictor and response data together. We analyzed two data sets – acute leukemia data set and NCI 60 data set. In the demonstration of the unsupervised method, the leukemia data is used as the first case study and the method is illustrated in detail including the extraction of principal components, selection of predictive components, prediction of class membership using recursive \( k \)-means clustering and finally the identification of genes most responsible for the observed classification. In the second case study, NCI 60 data set is used to validate the unsupervised method by predicting the class membership of cancer samples. Leukemia data set is once again considered for the demonstration of supervised method. Both unsupervised and supervised methods produced results with good prediction accuracies when compared with other existing methods in the literature. We were able to identify most of the genes linked to leukemia diagnosis that were already known in literature and also identified some genes presented in Table 4.8 that are less well known in the literature. Such genes identified via the multivariate approach may hold the keys to the complete exploration of the diverse biomarker space for treating any given medical condition.
CHAPTER 5

CLASSIFICATION OF CHEMICAL COMPOUNDS

5.1 Introduction

In high-throughput screening (HTS) experiments, a large number of chemical compounds are screened against disease targets in order to identify biologically active compounds. The data generated from these experiments serves as the starting point in the lengthy drug discovery process. The compounds used in HTS experiments are analyzed to identify structural features that distinguish a compound for its observed activity from others. Identification of differential features facilitates the rational design of drugs and in optimization of several key properties of the potential drug compounds. There are some similarities between the gene expression studies and HTS experiments for screening chemical compounds. Both the experiments generate large amounts of data with the main goal of discovering knowledge from raw data i.e., to classify samples into different categories and identify the underlying features most responsible for the observed classification. On the other hand, the difference between the two lies in the type of data generated – in case of gene expression studies, the predictor variables are continuous type whereas in case of HTS studies, the predictor variables are mostly discontinuous type. In
a DNA microarray data set, the continuous gene expression values are encountered, whereas, in a HTS data set, the discrete counts of structural descriptors are encountered. Examples of such structural descriptor counts include the number of specific benzene groups, amino groups, carbonyl groups etc., the number of hydrogen bond acceptors, hydrogen bond donors, the number of rotatable bonds, scores for Lipinski’s rule of five and so on. In some studies involving HTS studies, certain continuous variables are also included such as molecular weight, molar refractivity index, a measure of solubility (e.g., logK<sub>ow</sub>), polar surface area and so on. However, most of these continuous properties can be predicted from the discrete structural descriptors with reasonable accuracy. If continuous variables were involved as predictor variables then the methodological treatment for the classification of compounds into various classes of activity (or potency) would be exactly similar to that of the methodology presented in gene expression analysis. If however, the predictor variables are discrete, then there are certain differences in the methodology for classification. Nevertheless, the methods used for the analyses of the two types of data sets can be mutually adapted to fit the needs of the analysis. From the perspective of drug discovery, the same informatics tools developed at target identification stage are also useful at lead identification and optimization stages.

In this chapter, we shall consider four data sets with the goal of classifying compounds into their appropriate levels of activity. In the first case study, we shall demonstrate our methodology on the mutagenicity data set collected and cleaned by Feng et al<sup>69</sup>. The first case study shall be presented in a detailed manner to demonstrate our unsupervised and supervised methodologies while the other three case studies which
involve the use of proprietary data, shall be presented in a concise form manner with only the presentation of results.

5.2 Unsupervised Analysis: Mutagenicity Data Set

The mutagenicity data set was collected by Feng et al from public sources like EPA and NIH and originally had 2018 compounds but the authors reduced the number of compounds to 1863 because the software used in their study could not represent some of the compounds as a collection of structural descriptors. We shall also consider the same number of compounds (i.e., 1863) to demonstrate our methodology for the purpose of comparison. However, we shall consider two types of descriptors – (i) FRAG descriptors used by Feng et al and (ii) Leadscope® descriptors from Leadscope Inc. The response data in this data set is a binary response (mutagen or non-mutagen) based on the Ames test for mutagenicity where four strains of bacteria are tested with or without metabolic activation and if any of the eight tests is positive, the compound is considered positive for the test.

First, we shall illustrate our methodology using the FRAG descriptors, which are based on the kinds of fragments and the number of each kind present in the molecule. Some of these descriptor fragments reflect the physiochemical properties like logP and aromatic index. There were 197 FRAG descriptors, of which, 187 were structural descriptors and the rest were continuous variable type property descriptors. We considered the two sets of predictor variables separately with the greater focus on structural descriptors because the continuous property variables could themselves be
predicted from structural fragments. In the unsupervised methodology developed here, the first step was to collapse the dimensionality of the FRAG descriptors using PCA. The computation of PCs here is different from that described in the gene expression data sets. The computation of variance-covariance matrix is the key difference between the two data sets, since the variables involved here are not continuous. Linear correlation coefficients (e.g., Pearson correlation coefficients) cannot be defined for discrete variables like counts of descriptors, but a measure of correlation can be obtained indirectly from Tanimoto similarity measures of the two variables. The Tanimoto similarity measure between two compounds A and B is defined as:

\[
T\text{animoto similarity, } t_{A,B} = \frac{S_A}{(S_A + S_B - S_C)} \quad \text{[eqn. 5.1]}
\]

Where,

\[
S_A = \text{sum of counts of all descriptors present in a compound A};
\]

\[
S_B = \text{sum of counts of all descriptors present in a compound A};
\]

\[
S_C = \text{sum of counts of descriptors present both in compounds A and B};
\]

A sample MATLAB® program to compute Tanimoto similarity coefficient is provided in Appendix B. Thus, principal components can be extracted using the correlation matrix built from Tanimoto similarity measures. Table 5.1 shows a summary of the amount of variability accounted by the number of principal components. It can be seen that the first few PCs could not account for large percentage variability in the data set. The first 10 PCs account for only 17% variability in the data set and first 100 PCs are
required to account for 86% variability in the data. This can be explained by the diversity built into the descriptors, which in turn is attributed to the ability of the software used to obtain such diverse FRAG descriptors. In fact, it is a desirable attribute of any software to extract as diverse structural descriptors as possible. Figure 5.1 plots the proportion of negative eigenvector coefficients for the first 10 PCs in this analysis.

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>Percent Variability Accounted</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC-1</td>
<td>2.9796</td>
</tr>
<tr>
<td>PC-2</td>
<td>2.0387</td>
</tr>
<tr>
<td>PC-3</td>
<td>1.7622</td>
</tr>
<tr>
<td>PC-4</td>
<td>1.6722</td>
</tr>
<tr>
<td>PC-5</td>
<td>1.6537</td>
</tr>
<tr>
<td>PC-6</td>
<td>1.5435</td>
</tr>
<tr>
<td>PC-7</td>
<td>1.5326</td>
</tr>
<tr>
<td>PC-8</td>
<td>1.373</td>
</tr>
<tr>
<td>PC-9</td>
<td>1.3313</td>
</tr>
<tr>
<td>PC-10</td>
<td>1.3006</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Principal Components</th>
<th>Cumulative Percent Variability Accounted</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC-1 to PC-10</td>
<td>17.19</td>
</tr>
<tr>
<td>PC-2 to PC-20</td>
<td>28.82</td>
</tr>
<tr>
<td>PC-3 to PC-30</td>
<td>38.56</td>
</tr>
<tr>
<td>PC-4 to PC-40</td>
<td>47.16</td>
</tr>
<tr>
<td>PC-5 to PC-50</td>
<td>55.37</td>
</tr>
<tr>
<td>PC-6 to PC-100</td>
<td>85.97</td>
</tr>
<tr>
<td>PC-7 to PC-150</td>
<td>98.32</td>
</tr>
</tbody>
</table>

(b)

Table 5.1: (a) Percent variability accounted by the first 10 PCs; (b) Cumulative percent variability accounted by first 150 PCs.
Figure 5.1: A plot showing proportion of negative coefficients for the first 10 eigenvector vs. the percent variability accounted by their corresponding principal components for the mutagenicity data set using FRAG descriptors.
From Figure 5.1, it can be seen that principal components #1 and #2 will be poor predictors of class membership. $\textbf{PC}_1$ has all of its corresponding eigenvector coefficients as positive, while $\textbf{PC}_2$ has nearly 70% of its corresponding eigenvector coefficients as negative. $K$-means clustering was used with $k = 2$ groups as usual to separate the compounds into two classes using the predictive set of principal components. Prediction results are shown in Table 5.2, first with the inclusion of $\textbf{PC}_1$ and $\textbf{PC}_2$ in the predictive set of principal components.

<table>
<thead>
<tr>
<th>PREDICTED CLASSIFICATION</th>
<th>TRUE CLASSIFICATION</th>
<th>Mutagen</th>
<th>Non-Mutagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutagen</td>
<td>683</td>
<td>625</td>
<td></td>
</tr>
<tr>
<td>Non-Mutagen</td>
<td>215</td>
<td>340</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Unsupervised prediction results for mutagenicity prediction using FRAG descriptors with the inclusion of $\textbf{PC}_1$ and $\textbf{PC}_2$ in the predictive subset of PCs.
From Table 5.2, it can be seen that a prediction accuracy of 54.91% is obtained using the predictive set of principal components PC₁ and PC₂ are included as predictors. When we excluded PC₁ from the predictive subset of PCs, the prediction accuracy rose to 56.4% and with the additional exclusion of PC₂, the prediction accuracy rose to 57.31%. Unlike the case of gene expression analysis, exclusion of PC₁ and PC₂ did not have a dramatic effect on the prediction accuracy here. The marginal increase in prediction accuracy can be explained by the fact that PC₁ and PC₂ together account for mere 5% variability in the data set, which makes them relatively less influential. Nevertheless, this case stands as testimony to the finding that not all principal components contribute toward accurate prediction of class membership, whether it be classification of cancers of classification of chemical compounds. Although the increase in prediction accuracy is only marginal (2.3% increase), it underscores the importance of selecting predictive principal components that will be used as inputs to any clustering algorithm.

Next, we consider the Leadscope® descriptors to represent the chemical compounds in the mutagenicity data set and repeat the unsupervised analysis. A total of 3404 structural descriptors were obtained from Leadscope® software, which were collapsed dimensionally using PCA. We measured the similarity between compounds using Tanimoto coefficients and extracted principal components from the correlation matrix of Tanimoto similarity coefficients. The principal components were analyzed based on the proportion of eigenvector coefficients to identify predictive principal components. Figure 5.2 shows a plot of the proportion of negative eigenvector coefficients for the first 25 PCs in this analysis.
Figure 5.2: A plot showing proportion of negative coefficients for the first 25 eigenvector vs. the percent variability accounted by their corresponding principal components for the mutagenicity data set using Leadscope® descriptors.
From Figure 5.2 it can be seen that six out of the first 25 principal components (#1, 2, 3, 4, 11 and 14) will be poor predictors of class membership. We carried out $k$-means clustering with and without these six principal components. When we included these six principal components in the predictive subset of PCs, the prediction accuracy was 51.58% and when we excluded these six components, the prediction accuracy rose to 52.76%. Once again, the prediction accuracy increased only by a marginal amount. This can be explained by the fact that the six principal components put together account for less than 3% variability in the data set. This offered more evidence to the hypothesis that not all principal components can be used as predictors for classifying objects and the choice of principal components can affect the prediction accuracy. In the next section, we present the supervised methodology to classify compounds in the mutagenicity data set.

5.3 Supervised Analysis: Mutagenicity Data Set

Supervised methodology is demonstrated using both the FRAG descriptors and Leadscope® descriptors. We begin with the analysis of FRAG descriptors to obtain the PLS components. The PLS components were analyzed for their predictive ability by examining the proportion of negative weight vector coefficients. Figure 5.3 plots the proportion of negative coefficients for the first 10 weight vector vs. the percent variability accounted by their corresponding PLS components. It can be seen that all the top 10 PLS components except component #4 would be good predictors of class membership.
Figure 5.3: A plot showing proportion of negative coefficients for the first 10 weight vector vs. the percent variability accounted by their corresponding PLS components for the mutagenicity data set using FRAG descriptors.
When the PLS component #4 is included in the predictive subset of components and used as one of the inputs to the \(k\)-means clustering process, we obtained a prediction accuracy of 76.11%. When the PLS component #4 was excluded in the predictive subset of components and the rest of the components were used as inputs to the \(k\)-means clustering process, we obtained a prediction accuracy of 81.05%. The increase in prediction accuracy is only marginal (5%) because the fourth PLS component accounted for approximately 4% variability in the response data. If we excluded PLS component #1 from the predictive subset, the prediction accuracy fell steeply to 51.64%.

Next, we used the Leadscope® software to obtain Leadscope® structural descriptors to represent the chemical compounds in the mutagenicity data set. We used several different combinations of settings to generate the descriptors. One combination specified a minimum of 10 compounds per descriptor scaffold, minimum of 10 atoms per scaffold and a maximum of 999 rotatable bonds per scaffold. This combination generated 61 descriptors, which were large fragments common to most compounds in the data set. When these 61 descriptors were used in PLS analysis, we found that these descriptors are not able to capture the variation in response well. It is reasonable to expect that if the descriptors are large fragments, then they might not be discriminatory in nature and merely represent the more common inactive fragment across chemical compounds. Therefore, we explored several combinations of settings that generate a large number of descriptors. Generating a large number of descriptors would make it possible to identify sufficiently smaller yet discriminating structural features in the compounds. One such combination specified a minimum of one compound per descriptor scaffold, minimum of three atoms per scaffold and a maximum of 999 rotatable bonds per scaffold. This
combination generated 1191 descriptors to represent all the compounds in the data set. These descriptors were then transformed into reduced dimensional space of PLS components. We extracted 15 PLS components that accounted for nearly 74% variability in the response data. The PLS components were analyzed for their predictive ability by examining the proportion of negative weight vector coefficients. Figure 5.4 plots the proportion of negative coefficients for the first 15 weight vector vs. the percent variability accounted by their corresponding PLS components. It can be seen that all the 15 PLS components would be good predictors of class membership. We performed $k$-means clustering as usual with $k = 2$ groups using all the 15 PLS components. The prediction results shown in Table 5.3 indicate that a prediction accuracy of 91.57% is obtained.

<table>
<thead>
<tr>
<th>PREDICTED CLASSIFICATION</th>
<th>TRUE CLASSIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mutagen</td>
</tr>
<tr>
<td>Mutagen</td>
<td>824</td>
</tr>
<tr>
<td>Non-Mutagen</td>
<td>74</td>
</tr>
</tbody>
</table>

Table 5.3: Supervised prediction results for mutagenicity prediction using Leadscope® descriptors.
Figure 5.4: A plot showing proportion of negative coefficients for the first 15 weight vectors vs. the percent variability accounted by their corresponding PLS components for the mutagenicity data set using Leadscope® descriptors.
5.4 Conclusions

In this chapter, we applied our unsupervised and supervised methodologies to the task of classifying chemical compounds based on structural descriptors. We observed that selection of principal components or alternatively PLS components is critical to achieve good prediction accuracy when classifying compounds. We also observed that inclusion or exclusion of certain principal components does not affect the prediction accuracy dramatically. The reason for this observation is that these principal components do not account for a high percentage of variability in the data unlike in the case of gene expression analysis. These observations can be attributed to the nature of structural descriptors, which tend to be diverse and uncorrelated when compared to gene expression data that is highly correlated.

We have validated our methodology on three additional proprietary data sets and found similar observations with all of them. In one data set, local lymph node analysis was performed on 225 compounds, which tested as either potent or non-potent. Unsupervised analysis achieved a prediction accuracy of 55.11% while supervised analysis achieved a prediction accuracy of 81.64%. In the second data set, 353 compounds were classified for their toxicity level with a binary response. Unsupervised analysis achieved a prediction accuracy of 50.14% while supervised analysis achieved a classification efficiency of 81.8%. In the last data set, 297 compounds were categorized based on a response that indicates whether or not a compound caused a chromosomal aberration in the study. The unsupervised analysis predicted with an accuracy of 67.68% while the supervised method predicted with an accuracy of 93.14%. The task of
identifying structural descriptors that differentiate the compounds most strongly for their activity is similar to that of identifying differential genes. We used two methods – correlation method and the feature loading method to identify the most differentiating structural descriptors in a manner similar to the identification of differentiating genes.
CHAPTER 6

PREDICTION OF OCTANOL-WATER PARTITION COEFFICIENT

6.1 Introduction

When a solute is introduced in a liquid mixture containing two immiscible solvents, the solute distributes itself between the two phases. If sufficient time is allowed to attain equilibrium, the ratio of concentrations (or mole fractions) of this solute between the two phases assumes a certain value, which was found to be relatively insensitive to variations in temperature and concentration\(^95\). This ratio is called “Partition Coefficient” of the solute for a given liquid mixture. It is often reported as the logarithm of the ratio (logP or logK\(_{ow}\)).

Octanol-water partition coefficient is of special interest, as it is believed to be an important property for understanding the partitioning of small molecules (e.g., drugs, toxins) into cell membranes\(^67\). Although methods have been developed to experimentally measure logP and this data exists for more than 18000 organic chemicals\(^96\), there is experimental difficulty in the measurements at low concentrations of the hydrophobic solutes. Moreover, there is uncertainty involved in experimental measurements at low concentrations of the solute. Thus, there are many chemicals whose logP has not been
experimentally measured. Therefore, methods for calculating or predicting logP without performing experiments are of interest. There are several methods already in existence to predict logP values for most compounds. These methods can be broadly grouped into the following two categories: (i) Fragment or Group Contribution methods (ii) Whole molecule approaches. A brief discussion of these two methods follows in the next section.

6.2 Existing Methods

We shall present the fragment or group contribution methods first followed by the whole molecule approaches. In a fragment or group contribution approach, the logP of a compound is viewed as a collection of the logP values of the constituent fragments of the compound. Fujita et al. observed that the difference between logP of benzene (R-H) and simply substituted benzenes (R-X) was a constant quantity $\pi_X$ characteristic of the substituent X:

$$\log P(R-X) = \log P(R-H) + \pi_X \quad \text{[eqn. 6.1]}$$

and

$$\log P(Y-R-X) = \log P(H-R-H) + \pi_X + \pi_Y \quad \text{[eqn. 6.2]}$$

A list of $\pi$ values has been tabulated for a variety of substituents from experimental data. This was one of the earliest methods developed to predict logP and
hence is rather crude and limited in its applicability. The main limitation of this method is that the π value of a substituent is not universally correct irrespective of the compound in which the substituent is present. The π value of the same substituent is likely to be influenced by the specific environment in which the substituent is present. A modification of the above simplistic approach was proposed by Hansch and Leo to account for the deviations in fragmental contribution \( (f_n) \) by adding the correction factors \( (F_m) \). A number of effects including electronic delocalization, steric, ortho-meta-para and intramolecular effects were accounted in the correction factors. The modified equation is written as:

\[
\log P = \sum (a_n f_n) + \sum (b_m F_m) \quad \text{[eqn. 6.3]}
\]

Where, \( a \) is the number of occurrences of fragment of type \( n \) and \( b \) is the number of occurrences of correction factor \( F \) of type \( m \). This algorithm is called CLogP (available in computerized form) and is very complicated to implement manually. Hansch and Leo listed 200 fragments and 25 correction factors in 1979 and the list is updated as new fragment constants and correction factors are determined.\(^9\) A similar approach, proposed by Rekker \textit{et al.}, exists with 136 fragments and 10 structural correction factors. It is available in a computerized version called PrologP. It is not updated on a regular basis.\(^9\)

The SciFinder® search of Chemical Abstracts Services® database implements a more recent algorithm known as ACD LogP® (Advanced Chemistry Development LogP). It calculates a logP value for the neutral form of the molecule (with an error
margin of ±0.3 logP units). The algorithm is similar to the ones described above but uses a database of experimentally measured logP values to extract the contribution of pre-defined functional groups along with appropriate correction factors. The original correlations were based on a database of 3600 structures with 532 group contributions, 21 carbon atom type contributions and 2206 intramolecular correction factors. The functional groups in ACD® database differ from each other in their chemical structure, attachment to the hydrocarbon skeleton, cyclization and aromaticity. Carbon atom-type contributions differ from each other in their state of hybridization, number of attached hydrogens, branching, cyclization and aromaticity. Intramolecular correction factors account for the type of interacting groups, the length and type of the fragmental system in between the interacting groups (ACD®/LogP Technical Information). The database has been improved to include 18,400 structures and 1200 functional groups in the recent release. If a molecule contains groups or intramolecular interactions not included in the database, then a number of secondary algorithms estimate them. Although this algorithm is not described in detail in scientific literature, it is updated as new information becomes available. However, ACD® program does not take into account the specific features of different geometric isomers, stereoisomers, conformers, isotopes and structures with non-covalent bonds. At present, none of the fragment based methods address the issue of tautomeric forms of a given compound.

In the second method involving whole molecule approaches, the basis is that when a molecule enters a solvent, it enters it as a whole entity, not as a series of fragments and therefore, logP values based on whole molecule approaches should be more reliable and accurate than those derived from fragment based ones. These
approaches are more recent compared to the fragment based approaches and the most popular among these is the Linear Solvation Energy Relationships (LSER) method. In this approach, logP is estimated from the sum of three types of solute-solvent interactions – (i) repulsive (cavity formation), (ii) non-specific (dipolarity/polarizability) and (iii) specific attractive (hydrogen bonding). The repulsive contribution is determined from the molar volume of the solute, and the other two contributions from what are known as solvatochromic parameters that are determined from the ultraviolet spectrum of the $\pi \rightarrow \pi^*$ electronic transition. Alternatively, these solvatochromic parameters could be replaced with other descriptors obtained from quantum mechanical calculations.

We used a fragment-based approach called UNIFAC, which in itself is built upon an activity coefficient model known as UNIQUAC (UNIversal QUAsi-Chemical) model developed by Abrams and Prausnitz. By predicting the activity coefficient of a compound in octanol and water mixture, the partition coefficient can be estimated. We present the derivation and use of this non-ideal solution activity coefficient approach to predict logP in the following section.

6.3 Non-ideal Solution Activity Coefficient Approach

Let a solute $i$ be introduced in a binary liquid solution with two immiscible phases ($\alpha$ and $\beta$) at temperature $T$ and pressure $P$. The solute distributes itself in two phases until the driving force (chemical potential, $\mu$) for that solute in two immiscible phases becomes equal, attaining chemical equilibrium. Therefore at chemical equilibrium:
\[ \mu_i^\alpha = \mu_i^\beta \quad \text{[eqn. 6.4]} \]

Equivalently, this can be expressed in terms of the auxiliary function, fugacity:

\[ f_i^\alpha = f_i^\beta \quad \text{[eqn. 6.5]} \]

Dividing both sides by the standard state fugacity of pure \( i \) at system temperature \( (f_i^o) \), we have:

\[ \frac{f_i^\alpha}{f_i^o} = \frac{f_i^\beta}{f_i^o} \quad \text{[eqn. 6.6]} \]

This ratio of \( (f/f_o) \) is called the activity \( (a_i) \), which gives an indication of how “active” a substance is relative to its standard state.

\[ a_i^\alpha = a_i^\beta \quad \text{[eqn. 6.7]} \]

By definition, activity coefficient of a species \( i \), \( \gamma_i \) is the ratio of the activity of \( i \) to a measure of concentration of \( i \) (say \( x_i \))

\[ \gamma_i = a_i / x_i \quad \text{[eqn. 6.8]} \]

i.e.,

\[ \gamma_i = a_i / x_i \quad \text{[eqn. 6.8]} \]

Where, \( x_i \) is the mole fraction of i.

Substituting equation 6.8 in equation6.7 , we get:

\[ \gamma_i x_i^\alpha = \gamma_i x_i^\beta \quad \text{[eqn. 6.9]} \]
Or, \[ x_i^\alpha / x_i^\beta = \gamma_i^\beta / \gamma_i^\alpha \]  

[eqn. 6.10]

The ratio above, which is the mole fraction of solute \( i \) in \( \alpha \)-phase to \( \beta \)-phase at equilibrium is called Partition Coefficient (\( K_{\alpha\beta} \)).

i.e., \[ K_{\alpha\beta,i} = x_i^\alpha / x_i^\beta = \gamma_i^\beta / \gamma_i^\alpha \]  

[eqn. 6.11]

Therefore, the partition of any compound (solute \( i \)) can be calculated in principle if we know the corresponding activity coefficients in both the phases. For octanol-water mixture, the partition coefficients (\( K_{ow} \)) of any solute is,

\[ K_{ow,i} = x_i^o / x_i^w = \gamma_i^w / \gamma_i^o \]  

[eqn. 6.12]

Activity coefficient (\( \gamma_i \)) of species \( i \) can be obtained using the following notation:

\[ U = \text{Internal energy} \]

\[ S = \text{Entropy} \]

\[ V = \text{Volume} \]

\[ N_i = \text{Number of moles of } i \]

\[ G = \text{Gibbs free energy} \]
Fundamental equation in terms of $U$ is given by:

$$dU = TdS - PdV + \sum_i (\mu_i dN_i) \quad [\text{eqn. 6.13}]$$

Integrating equation 6.13, from a state of zero mass ($U = S = V = N_i = 0$) to a state of finite mass ($U, S, V, N_i$) at constant temperature ($T$), pressure ($P$) and composition, we get:

$$U = TS - PV + \sum_i (\mu_i N_i) \quad [\text{eqn. 6.14}]$$

Differentiation of the above equation to obtain a general equation for $dU$, we get:

$$dU = TdS + SdT - PdV - VdP + \sum_i (\mu_i dN_i) + \sum_i (N_id\mu_i) \quad [\text{eqn. 6.15}]$$

Comparing equation 6.13 and equation 6.15, we obtain the famous, Gibbs-Duhem equation:

$$SdT - VdP + \sum_i (N_id\mu_i) = 0 \quad [\text{eqn. 6.16}]$$

Now, for a pure species $i$,

$$SdT - VdP + N_id\mu_i = 0 \quad [\text{eqn. 6.17}]$$

Or alternatively,

$$d\mu_i = (-S/N_i) dT + (V/N_i) dP \quad [\text{eqn. 6.18}]$$

From equation 6.18, it can be seen that

$$(d\mu_i/dP)_T = V/N_i = v_i \text{ (molar volume of } i) \quad [\text{eqn. 6.19}]$$
For ideal gases,

\[ V_i = \frac{(RT)}{P} \quad \text{[eqn. 6.20]} \]

Substituting equation 6.20 in equation 6.19 and integrating at constant temperature and from a standard state (\( ^o \)),

\[ \mu_i - \mu_i^o = RT \log\left(\frac{P}{P^o}\right) \quad \text{[eqn. 6.21]} \]

Generalizing the above,

\[ \mu_i - \mu_i^o = RT \log\left(\frac{f_i}{f_i^o}\right) \quad \text{[eqn. 6.22]} \]

Where \( f_i \) is the fugacity of species \( i \).

Equation 6.22 is valid for both ideal and non-ideal systems (solid, liquid or gas, pure or mixed). Now, if we consider Gibbs free energy, which is defined as

\[ G = U - TS + PV \quad \text{[eqn. 6.23]} \]

Differentiating the equation 6.23,

\[ dG = dU - TdS - SdT + PdV + VdP \quad \text{[eqn. 6.24]} \]

Substituting equation 6.15 in equation 6.24, we get:

\[ dG = -SdT + VdP + \Sigma \mu_i dN_i \quad \text{[eqn. 6.25]} \]

From this, we see that:

\[ g_i = \left(\frac{dG}{dN_i}\right)_{T,P,N_{j\neq i}} = \mu_i \quad \text{[eqn. 6.26]} \]
Also, excess functions are defined as thermodynamic properties of solutions that are in excess of those of an ideal solution at the same conditions of temperature, pressure and composition. Therefore, excess Gibbs free energy is defined as:

\[ G^E = G_{\text{actual solution at } T, P \text{ and } x} - G_{\text{ideal solution at } T, P \text{ and } x} \quad \text{[eqn. 6.27]} \]

In terms of partial molar properties,

\[ g_i^E = g_i(\text{real}) - g_i(\text{ideal}) \quad \text{[eqn. 6.28]} \]

Combining equation 6.22 and equation 6.28, we get:

\[ g_i^E = g_i(\text{real}) - g_i(\text{ideal}) = RT \log(f_i(\text{real}) / f_i(\text{ideal})) \quad \text{[eqn. 6.29]} \]

Also, for an ideal solution, the fugacity of every component is proportional to some measure of its concentration (usually the mole fraction):

i.e., \[ f_i(\text{ideal}) = R_i x_i \quad \text{[eqn. 6.30]} \]

Where \( R_i \) is a proportionality constant.

If we set \( R_i = f_i^o \), the standard state fugacity, then

\[ f_i(\text{ideal}) = f_i^o x_i \quad \text{[eqn. 6.31]} \]

But, by definition of activity (equations 6.6 & 6.7),

\[ a_i(\text{ideal}) = f_i(\text{ideal}) / f_i^o = x_i \quad \text{[eqn. 6.32]} \]

Note that by comparing equation 6.32 and equation 6.8 the activity coefficient of \( i \) in ideal solution is:
Using equation 6.30 in equation 6.29, we get:

\[ g_i^E = RT \log(f_{i(\text{real})}/R_i x_i) \]  
[eqn. 6.34]

Moreover, setting \( R_i = f_i^0 \), we obtain:

\[ a_{i(\text{real})} = \gamma_i x_i = f_{i(\text{real})}/R_i \]  
[eqn. 6.35]

Or alternatively,

\[ \gamma_i = f_{i(\text{real})}/R_i x_i \]  
[eqn. 6.36]

Substituting equation 6.36 in equation 6.34, we get

\[ g_i^E = RT \log\gamma_i \]  
[eqn. 6.37]

Thus, activity coefficient of solute \( i (\gamma_i) \) in a liquid binary mixture can be obtained from excess Gibbs energy function. There are several expressions that are proposed for \( g^E \) such that it satisfies the constraints

\[ g^E \to 0 \text{ when } x_i \to 0 \text{ or } x_i \to 1 (\text{pure species i.e. ideal solution}) \]

One such model is the UNIQUAC model given by the addition of combinatorial and residual contributions:

\[ \log\gamma_i = \log\gamma_i (\text{combinatorial}) + \log\gamma_i (\text{residual}) \]  
[eqn. 6.38]

Where,
\[ \log \gamma_i \text{ (combinatorial)} = \log(\frac{\phi_i}{x_i}) + \frac{Z}{2}q_i \log(\frac{\theta_i}{\phi_i}) + \frac{1}{\gamma_i} \sum_j(x_j l_j) \]

\[ \log \gamma_i \text{ (residual)} = -q_i \left[ 1 - \log \left( \sum_j \theta_{ji} \tau_{ji} \right) - \left( \sum_j \frac{\theta_{ji} \tau_{ji}}{\theta_{ki} \tau_{ki}} \right) \right] \]

\[ l_i = \frac{(r_i - q_i)}{Z^2} - (r_i - 1) ; Z=10 \]

\( \gamma_i \) and \( q_i \) evaluated from molecular structure

\( \tau \)'s are interaction parameters.

\[ \theta_i = \frac{(q_i x_i)}{\sum_j (q_j x_j)} \quad \phi_i = \frac{(\gamma_i x_i)}{\sum_j (\gamma_j x_j)} \quad \tau_{ji} = \exp \left( \frac{(u_{ji} - u_{ii})}{RT} \right) \]

\( \gamma_i \) and \( q_i \) are measures of molecular vanderwaals volumes and molecular surface areas; \( \tau \)'s are evaluated from experimental phase equilibrium data.

Another model, which is built on the UNIQUAC model, is the UNIFAC model. It is similar to UNIQUAC as described in system of equations in 6.38, with the exception that the residual component is calculated as follows:

\[ \log \gamma_i \text{ (residual)} = \sum_{k, \text{all groups}} v_k \left( \log \Gamma_k - \log \Gamma_{k(i)} \right) \quad [\text{eqn. 6.39}] \]

\[ \log \Gamma_k = Q_k \left[ 1 - \log(\sum_m \theta_{mk} \phi_{mk}) - \sum_m \frac{\theta_{mk} \phi_{km}}{\sum_n \theta_{nm} \phi_{nm}} \right] \quad [\text{eqn. 6.40}] \]

Where, \( \log \Gamma_{k(i)} \) is the residual activity coefficient of group \( k \) in a reference solution containing only molecules of type \( i \).

As per the definition of the partition coefficient \( (k_{ow,i}) \) of a solute \( i \) in an octanol-water mixture, we can write:

\[ K_{ow,i} = \frac{C_i^{OR}}{C_i^{WR}} \quad [\text{eqn. 6.41}] \]
Where,

\[ C_i^{\text{OR}} = \text{concentration of } i \text{ in octanol rich phase} \]

\[ C_i^{\text{WR}} = \text{concentration of } i \text{ in water rich phase} \]

Typically, the amount of solute added to octanol-water mixture is on the order of \( \sim 10^{-3}\text{M} \), so that \( K_{\text{OW}} \) can be reduced in terms of infinite dilution activity coefficients. The water rich phase (WR) is essentially water, but the octanol rich phase (OR) is a mixture of 1-octanol and water (79.3 mol\% 1-octanol). Therefore,

\[ K_{\text{OW},i} = \frac{(C_{\text{tot}}^{\text{OR}} \cdot X_i^{\text{OR}})}{(C_{\text{tot}}^{\text{WR}} \cdot X_i^{\text{WR}})} = \frac{(C_{\text{tot}}^{\text{OR}} \cdot \gamma_i^{\text{WR},\infty})}{(C_{\text{tot}}^{\text{WR}} \cdot \gamma_i^{\text{OR},\infty})} \]  
[eqn. 6.40]

Using the liquid molar volumes at 25°C, the total molar concentration of octanol rich phase, \( C_{\text{tot}}^{\text{OR}} = 8.378 \text{ mol/L} \) and \( C_{\text{tot}}^{\text{WR}} = 55.56 \text{ mol/L} \). Therefore equation 6.40 reduces to:

\[ K_{\text{OW},i} = 0.1508 \frac{\gamma_i^{\text{WR},\infty}}{\gamma_i^{\text{OR},\infty}} \]  
[eqn. 6.41]

Where,

\[ \gamma_i^{\text{WR},\infty} = \text{Infinite dilution activity coefficient of } i \text{ in pure water} \]

\[ \gamma_i^{\text{OR},\infty} = \text{Infinite dilution activity coefficient of } i \text{ in ‘OR’ phase.} \]

We developed a MATLAB® program that calculates these \( \gamma \) values using UNIFAC group contribution approach and thus enables the prediction of \( \log K_{\text{OW}} \). The program and selected results are made available in the Appendix C.
CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

In the current study, most of our research effort was focused on developing informatics methodology that is applicable at two critical stages in drug discovery process – (i) target identification and (ii) lead compound identification and optimization. At the target identification stage, a methodology was developed to classify biological samples based on genetic fingerprints and to identify those biological targets that are most responsible for the observed classes. More specifically, we focused on developing a methodology to analyze microarray data for classification of cancers into different subtypes when no prior knowledge of differences in classification is available i.e., unsupervised classification. Our approach was extended to the supervised scenario, where prior knowledge of differences could be used to build the predictive system. We demonstrated our methodology with improved and comparable accuracies to other existing techniques for both unsupervised and supervised classifications. The genes identified via our approach as potential targets were validated with supporting studies already published in literature. Finally, the methodology was adapted and extended to the
chemoinformatics framework at the lead identification and optimization stage of drug discovery, where the objectives were to identify and classify chemical compounds into varying levels of activity against a given biological target and to identify structural features most responsible for the observed classification of compounds.

In this work, we proposed to apply the multivariate framework of principal component analysis in the case of unsupervised scenario and partial least squares in the case of supervised scenario. PCA and PLS were used to reduce the dimensionality of large number of variables involved in the analysis and to effectively summarize the inter-relationships between these predictor variables that are essential to differentiate the samples. The importance of selecting principal components and PLS components to identify influential as well as predictive components is described. Results comparing the prediction accuracies of including the influential yet non-predictive principal components and excluding them from the predictive subset were presented. The components selected in this manner were used as inputs to the $k$-means clustering algorithm to separate the clusters into two groups at each iteration. $K$-means was used recursively to solve the multi-class classification problem. After the successful demonstration of our methodology to predict class memberships, we also illustrated two methods – correlation method and feature loadings method to identify the predictors that most strongly differentiate between the samples. We validated the genes identified via these methods in the case of identifying biomarkers that differentiate leukemia sub-types. The methodology is similar and easily adapted in the case of identifying structural descriptors that most strongly differentiate between the classes of compounds based on activity.
7.2 Impact of Current Research

The central focus of this work has been on enhancing drug discovery by overcoming two major bottlenecks in the process – (i) target identification and (ii) lead optimization. It is estimated that the current drug therapy is based on less than 500 molecular targets while the potential drug targets may lie between 5000 and 10,000. Therefore, there are at least 10 times as many molecular targets to be exploited for future therapies than are being used as of today. Unlike most of the prior methods that only focus on identifying highly differential genes irrespective of their diversity, the methodology developed here aims at identifying not only the highly differential genes but also strives to ensure that the genes are non-redundant in their information content. Identifying all the diverse genes responsible for characterizing a medical condition provides alternate routes for therapeutic intervention and allows for a thorough exploration of target space for a given disease.

The cost of discovering a new drug has sharply risen to $802 million in the year 2002 with a timeline of nearly 13.6 years. Despite the large investments in time and money, drugs that were successfully introduced in the market had to be withdrawn later due to efficacy (38%) and safety (20%) reasons. Increasing the clinical success rate to 1 in 3 (from today’s 1 in 10) would lower the costs by $217 million. This underscores the importance of predicting potential failures at an early stage prior to the initiation of expensive clinical trials. Moreover, cutting drug development time in half would lower the costs from $802 million to $568 million, which can be realized by improving the current understanding of structure activity relationships. In this work, we develop a novel informatics methodology that addresses the issue of predicting drug safety from structural
features at an early stage and also improves the understanding of structural effects on efficacy. The current work attempts to reduce the time and cost associated with drug discovery.

In the present work, we have developed a novel unsupervised method that incorporates the multivariate treatment of variables using PCA and cleverly uses \( K \)-means clustering in a hierarchical manner to solve the unsupervised multi-class classification problem. Unsupervised techniques for classification are required not only when the response data is unavailable, but also in cases where the response data is available. Unsupervised methods are employed to select a diverse yet representative set of candidates from a vast candidate space for future experimentation when it is not feasible (due to time and expense involved) to test each and every candidate in the candidate space. Moreover, unsupervised grouping of candidates into several categories facilitates model validation using candidates that were not used to build the model. In addition, this work makes a unique contribution towards the fundamental understanding of principal component analysis when the extracted principal components are intended to be used as predictors of class membership. Present work demonstrates that not all principal components could be used as predictors for classification purposes and outlines a methodology to select not only influential but also good predictors of class distinction at the same time.
7.3 Future Work and Recommendations

The genes identified via the correlation rank method matched well with the results already published in literature. However, the loading rank method identified several genes that were not described in the leukemia related literature. This could be because of the multivariate treatment of genes in the loading rank method as opposed to the univariate treatment of genes in correlation rank method. Establishing the biological relevance of such less popular genes identified via the loading method provide significant insights into the understanding of new molecular targets for the disease. The Gene Ontology (GO)\textsuperscript{106} consortium is developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions. A gene product has one or more molecular functions and is used in one or more biological processes and it might be associated with one or more cellular components. Molecular function describes activities, such as catalytic or binding activities, at the molecular level. A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. A cellular component is a component of a cell but with the condition that it is part of some larger object, which may be an anatomical structure. The task of evaluating the identified genes in the light of GO information is a logical next step towards the extension of this work.

The methodology developed in this work for the unsupervised scenario attains high prediction accuracies in special cases where the influential principal components are included in the predictive subset. In cases where, the influential principal components that account for a large percentage of variability in the data set turn out to be poor
predictors of class distinction, the prediction accuracies are not very high. Therefore, we clearly can identify that principal component analysis being a linear approach is not able to capture the interactions of the original variables in a manner that differentiates the samples with high prediction accuracy. In such a scenario, one obvious solution is to use partial least squares when the response data is available. If such data is not available and one is forced to apply unsupervised analysis, then applications of non-linear principal components would be a promising area of future research to explore. Before exploring nonlinear PCA, it would be interesting to evaluate the extent to which linear PCA fails to account for the nonlinearities in the data. This could be verified by reconstructing the original data using the small set of linear PCs and comparing the reconstructed data with the true data. In the ALL/AML tumor classification results presented earlier, it is likely that although the PCs capture reasonable amount of variability in the dataset, they fail to account for the nonlinearities in the observed data and hence do not turn out to be good predictors. A large deviation in the original data and the reconstructed data from linear principal components will suggest that the linear PCA is inadequate. We carried out some preliminary research to evaluate the possibility of significant lack of fit in the case of gene expression data, but faced the problem of dealing with large sums of squares, which could be either due to the large number of variables involved or due to the existence of true lack of fit. Studies to identify each case more accurately are needed in future.

One possible extension to the current work is in the context of determining the cut-off criteria for the selection of predictive principal components in unsupervised analysis and predictive PLS components in supervised analysis. In the present work, we used a conservative cut-off criterion of [0.4, 0.6] to screen the good predictors of class
distinction from the poor ones. However, this criterion is rather subjective and there is a need to systematically define the selection criteria. The cross-validation approach can be utilized to determine the selection criteria in a more objective manner. Leaving out components and determining the influence of the left-out component on prediction accuracy can provide insights into selecting the interval for cut-off criteria. Moreover, possible modification in the component selection metric (the proportion of eigenvector coefficients having negative or positive sign) is possible to make the selection more robust. The current method involves the examination of the proportion of negative (or alternatively positive) eigenvector coefficients to determine the good predictors from poor ones. Incorporating additional information on the magnitude of the eigenvector coefficient along with its sign in making the selection of predictive components is a possible extension to the current component selection step. A method that examines the sign of the coefficient based on the weighted magnitude of the eigenvector coefficient could be explored for this purpose.

Another area where future research is promising is in the generation of structural descriptors for representing chemical compounds. 3D-descriptors provide additional discriminatory information than topological 2D-descriptors for compounds with similar structures and molecular orientation. Although, geometrical or 3D-descriptors provide additional information, it comes at an increased computational overhead to the programs that deal with 3D-descriptors. Therefore, methods to efficiently represent a compound in 3D-descriptors are needed which reduce the computational time and cost in future analysis. Group contribution models (e.g., modified UNIFAC) as in the case of predicting octanol-water partition coefficients can be evaluated to generate descriptors that contain
more information than the present descriptors. Descriptors that are based on not only the presence or absence of a structural feature but also take into account the local environment around a feature will be more sophisticated in terms of information they carry. Such sophisticated chemical information about the presence of a structural feature and its interaction with the local environment would help characterize the chemical compounds more precisely and facilitate the optimization of physico-chemical properties to an improved level. The whole molecule approach to predict logP presents similar ideas and hence could be leveraged in the context of enhancing the information content in descriptors that account for the influence of the specific environment in which the descriptor is present. Also, the fact that structural features that describe chemical compounds have much less inter-dependence than genes that describe the biological samples need to be kept in mind when defining the cut-off criteria for selecting predictive components.
APPENDIX A

MATLAB PROGRAM FOR GENE EXPRESSION ANALYSIS
/* BEGIN: Sample program to obtain PCs */
tic;
clear;
pack;
cle;

xtest = load('xtest.txt');
xtrain = load('xtrain.txt');
x = [xtrain;xtest];
z = zscore(x);

% PCA using corr matrix of x or cov of z.
cov_z = cov(z);
[V D] = eig(cov_z);
pc = z*V;
eig_values = sum(D);

fprintf('Contribution of first 10 PCs : %f
', sum(eig_values(1:10))*100/sum(eig_values));
fprintf('Contribution of first 20 PCs : %f
', sum(eig_values(1:20))*100/sum(eig_values));
fprintf('Contribution of first 30 PCs : %f
', sum(eig_values(1:30))*100/sum(eig_values));
fprintf('Contribution of first 40 PCs : %f
', sum(eig_values(1:40))*100/sum(eig_values));
fprintf('Contribution of first 50 PCs : %f
', sum(eig_values(1:50))*100/sum(eig_values));
fprintf('Contribution of first 60 PCs : %f
', sum(eig_values(1:60))*100/sum(eig_values));
toc;
t=toc

/* END: Sample program to obtain PCs */

/* BEGIN: Sample program to perform K-means clustering*/
%K-means clustering using first 50 PCs:
for i=1:1000
    [idx,C,sumd,DD] = kmeans(pc(:,138:186),2, 'replicates', 1);
    ctoc(i) = sum((C(1,:)-C(2,:)).^2);
    sumd(i) = sum(sumd);
end
/* END: Sample program to perform K-means clustering*/

/* BEGIN: Sample program to perform cross validation */
% Performing Cross_Validation leaving out Sample #2.
tic;
clear;
pack;
clc;

xtest = load('xtest.txt');
xtrain = load('xtrain.txt');
x = [xtrain;xtest];
i = 2; %Leaving out sample #2
x(i,:) = [];

157
% PCA using cov matrix of x.
cov_x = cov(x);
[V D] = eig(cov_x);
pc = x*V;
eig_values = sum(D);

% fprintf('Contribution of last 10 PCs : %f\n',
sum(eig_values(7119:7129))*100/sum(eig_values));

[idx,C,sumd,D] = kmeans(pc(:,7069:7128),2, 'replicates', 1); %Using 60 top PCs.
savefile = 'output_XVal_2_1.mat';
save(savefile, 'pc','V');

savefile = 'output_XVal_2_2.mat';
save(savefile, 'C', 'sumd', 'D');

toc;
t=toc
/* END: Sample program to perform cross validation */
APPENDIX B

MATLAB PROGRAM FOR CLASSIFYING CHEMICAL COMPOUNDS
/ BEGIN: Sample program to perform PCA and K-means clustering for classification of chemical compounds*/

load('x.txt');

% If a particular column has all entries as 0s, then that column(feature)
% is not present in any of the compounds(rows). These columns are not of
% interest and are removed prior to calculation of tanimoto correlation.

a = isinf(1./sum(x));
b = isinf(1./a);
x=x(:,b);
tan_corr = tanimoto(x);
[V D] = eig(tan_corr);
pc = x*V;
eig_values = sum(D);

fprintf('Contribution of last 10 PCs : %f
', sum(eig_values(178:187))*100/sum(eig_values));

fprintf('Contribution of last 20 PCs : %f
', sum(eig_values(168:187))*100/sum(eig_values));

fprintf('Contribution of last 30 PCs : %f
', sum(eig_values(158:187))*100/sum(eig_values));

fprintf('Contribution of last 40 PCs : %f
', sum(eig_values(148:187))*100/sum(eig_values));

fprintf('Contribution of last 50 PCs : %f
', sum(eig_values(138:187))*100/sum(eig_values));

fprintf('Contribution of last 100 PCs : %f
', sum(eig_values(88:187))*100/sum(eig_values));

fprintf('Contribution of last 150 PCs : %f
', sum(eig_values(38:187))*100/sum(eig_values));
% Proportion of eigenvector coefficients that are negative for first 50 EVs:
% prop=sum(V(:,187:-1:138)<0)/187;

%K-means clustering using first 50 PCs:
for i=1:1000
    [idx,C,sumd,DD] = kmeans(pc(:,138:186),2, 'replicates', 1);
    ctoc(i) = sum((C(1,:)-C(2,:)).^2);
    sumd(i) = sum(sumd);
end
/* END: Sample program to perform PCA and K-means clustering for classification of
chemical compounds*/

/*BEGIN: Program to compute Tanimoto similarity coefficient */
function y = tanimoto(x)
    [r,c] = size(x);
    y = zeros(c,c);
    sumx = sum(x);
    for i = 1:c
        A = sumx(i);
        for j = i:c
            B = sumx(j);
            C = sum(x(:,i) & x(:,j));
            y(i,j) = C/(A + B - C);
        end
    end
    y = triu(y) + tril(y+y',-1); %create symmetric correlation matrix
/*END: Program to compute Tanimoto similarity coefficient */
APPENDIX C

MATLAB PROGRAM FOR PREDICTING LOGK_{OW}
% This program estimates the LogP value for a drug compound in a octanol-water mixture.

% ----------------- Variable Declarations -----------------
% drug_name;
% num_of_comps_in_soln;
% comp_names_in_soln = cell(1, num_of_comps_in_soln);
% x_soln = cell(1, num_of_comps_in_soln);
% x_drug;
% T_kelvin;
% master_compound='master_compound_list.txt';
% mc_htable = java.util.Properties;
% z = 10;
% master_fp;
% drug_fp_index ;
% fp_drug;
% soln_fp_index(1,num_of_comps_in_soln);
% fp_soln(i,num_of_comps_in_soln);
% drug_activity_coeff;
%----------------- Variable Declarations -----------------

% Reading Drug info, Temp and number of compounds in the solution
clc;
clear all;

prompts_1 = {'Name of the compound (Drug) whose LogP value is to be estimated',...
'Concentration of the drug compound',...
'Temperature of the system in kelvin';

title_1 = 'Inputs for Activity Coefficient Calculator';
line_numbers_1 = [1 1 1];
line_numbers_1 = line_numbers_1';
default_inputs_1 = {'MIANSERIN-', '0', '298'};
add_opts_1.Resize='on';
add_opts_1.WindowStyle='normal';
add_opts_1.Interpreter='tex';

user_inputs_1 =
    inputdlg(prompts_1,title_1,line_numbers_1,default_inputs_1,add_opts_1);

drug_name = char(upper(user_inputs_1(1,1)));
x_drug = str2num(user_inputs_1{2,1});
T_kelvin = str2num(user_inputs_1{3,1});

% Reading input for compounds other than drug

% prompts_2 = {'Name of the compound in solution other than drug',...
% 'Concentration of the compound'};
% %
% % for i = 1 : num_of_comps_in_soln
% %    user_inputs_2 = inputdlg(prompts_2,title_1);
% %    comp_names_in_soln{1,i} = char(upper(user_inputs_2(1,1)));
% %    x_soln(1,i) = str2num(user_inputs_2{2,1});
% % end;
% coordination number assignment

z = 10;

% hash table to load fp values by reading in the names of the compounds

master_compound='master_compound_list.txt';
mc_htable = java.util.Properties;

try
    FIS = java.io.FileInputStream(master_compound);
catch
    error(sprintf('Failed to open %s for reading.', master_compound));
end;

mc_htable.load(FIS);
FIS.close;

load('master_fp.txt');

drug_fp_index = str2num(mc_htable.get(char(drug_name)));
fp_drug = master_fp(drug_fp_index, 2:end);

% for i = 1 : num_of_comps_in_soln;
%    soln_fp_index(1,i) = str2num(mc_htable.get(char(comp_names_in_soln(1,i))));
%    fp_soln(i,:) = master_fp(soln_fp_index(1,i), 2:end);
% end;

% obtaining fp vector for WATER
x_soln(1,1) = 1;
soxn_fp_index(1,1) = str2num(mchhtable.get(char('WATER')));
fp_soln(1,:) = master_fp(soxn_fp_index(1,1), 2: end);

% Invoking get_unifac_gamma.m for aqueous phase
disp(' AQUEOUS PHASE');
[drug_activity_coeff total_a missing_a] = get_unifac_gamma(T_kelvin, x_drug, x_soln, fp_drug, fp_soln, z);
gamma_aqueous_phase = drug_activity_coeff(1);
fprintf('gamma_aqueous_phase Value : %f\n', gamma_aqueous_phase);

% obtaining fp vector for WATER(.26) and OCTANOL(.74) mixture
x_soln(1,1) = .207;
x_soln(1,2) = .793;
soxn_fp_index(1,1) = str2num(mchhtable.get(char('WATER')));
fp_soln(1,:) = master_fp(soxn_fp_index(1,1), 2: end);
soxn_fp_index(1,2) = str2num(mchhtable.get(char('OCTANOL')));
fp_soln(2,:) = master_fp(soxn_fp_index(1,2), 2: end);

% Invoking get_unifac_gamma.m for tissue phase
disp('******************************************************');
disp(' TISSUE PHASE');

166
[drug_activity_coeff total_a missing_a] = get_unifac_gamma(T_kelvin, x_drug, x_soln, fp_drug, fp_soln, z);

gamma_tissue_phase = drug_activity_coeff(1);
fprintf('gamma_tissue_phase Value : %f\n', gamma_tissue_phase);
logP = log10(0.1508*gamma_aqueous_phase/gamma_tissue_phase);

% Displaying output

disp('******************************************************
');
disp(' ');
fprintf('Drug Name: %s\n', drug_name);
fprintf('LogP Value : %f\n', logP);
disp(' ');
disp('******************************************************
');

/*END: Program to estimate octanol-water partition coefficient*/
<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Activity Coefficient of Aqueous Phase</th>
<th>Activity Coefficient of Tissue Phase</th>
<th>LogK_{OW}</th>
<th>ACD logP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Octane</td>
<td>1.10E+05</td>
<td>3.6664</td>
<td>3.6562</td>
<td>5.006</td>
</tr>
<tr>
<td>Alprenolol</td>
<td>2.03E+05</td>
<td>0.036132</td>
<td>5.9274</td>
<td>2.884</td>
</tr>
<tr>
<td>Nonane</td>
<td>3.47E+05</td>
<td>4.2095</td>
<td>4.0945</td>
<td>5.538</td>
</tr>
<tr>
<td>Naproxen-2</td>
<td>51633</td>
<td>1.1626</td>
<td>3.8259</td>
<td>2.99</td>
</tr>
<tr>
<td>Naproxen-1</td>
<td>46775</td>
<td>1.1805</td>
<td>3.7763</td>
<td>2.99</td>
</tr>
<tr>
<td>Corticosterone</td>
<td>4.48E+05</td>
<td>0.44177</td>
<td>5.1843</td>
<td>1.758</td>
</tr>
<tr>
<td>Pindolol</td>
<td>14305</td>
<td>0.29904</td>
<td>3.8582</td>
<td>1.97</td>
</tr>
<tr>
<td>Ethane</td>
<td>70.351</td>
<td>0.99882</td>
<td>1.0262</td>
<td>1.818</td>
</tr>
<tr>
<td>Cimetidine</td>
<td>2920.9</td>
<td>0.01895</td>
<td>4.3663</td>
<td>0.203</td>
</tr>
<tr>
<td>Metformin</td>
<td>4.10E-03</td>
<td>0.00058473</td>
<td>0.0239</td>
<td>-2.313</td>
</tr>
<tr>
<td>Progesterone</td>
<td>1.72E+08</td>
<td>4.0424</td>
<td>6.8079</td>
<td>4.037</td>
</tr>
<tr>
<td>Buspirone</td>
<td>1.92E+01</td>
<td>2.08E-06</td>
<td>6.1422</td>
<td>3.434</td>
</tr>
<tr>
<td>Hydrocortisone</td>
<td>39306</td>
<td>0.41005</td>
<td>4.16</td>
<td>1.426</td>
</tr>
<tr>
<td>Promazine</td>
<td>5.57E+01</td>
<td>4.95E-05</td>
<td>5.2295</td>
<td>4.633</td>
</tr>
<tr>
<td>Diclofenac</td>
<td>1.64E+06</td>
<td>3.9778</td>
<td>4.7945</td>
<td>3.284</td>
</tr>
<tr>
<td>Aminoantipyrine</td>
<td>1061.7</td>
<td>0.095848</td>
<td>3.2228</td>
<td>-0.4</td>
</tr>
</tbody>
</table>

Table C1: A partial list of octanol-water partition coefficients predicted by our program along with the ACD® logP values from SciFinder® database.
<table>
<thead>
<tr>
<th>Substance</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testosterone</td>
<td>6.21E+06</td>
<td>1.3711</td>
<td>5.8342</td>
<td>3.475</td>
</tr>
<tr>
<td>Beta-estradiol</td>
<td>3.16E+05</td>
<td>0.1376</td>
<td>5.5389</td>
<td>4.131</td>
</tr>
<tr>
<td>Sotalol</td>
<td>12804</td>
<td>0.47912</td>
<td>3.6053</td>
<td>0.317</td>
</tr>
<tr>
<td>Octanol</td>
<td>3171</td>
<td>1.0394</td>
<td>2.6628</td>
<td>3.001</td>
</tr>
<tr>
<td>Pregnanolone</td>
<td>1.82E+09</td>
<td>1.2668</td>
<td>8.3354</td>
<td>4.891</td>
</tr>
<tr>
<td>Quinine</td>
<td>3.59E+07</td>
<td>0.98963</td>
<td>6.7381</td>
<td>3.44</td>
</tr>
<tr>
<td>Prednisolone</td>
<td>51194</td>
<td>0.3304</td>
<td>4.3686</td>
<td>1.69</td>
</tr>
<tr>
<td>Metoprolol-8</td>
<td>12862</td>
<td>0.096848</td>
<td>4.3016</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-7</td>
<td>6891.1</td>
<td>0.076448</td>
<td>4.1333</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-6</td>
<td>16993</td>
<td>0.10163</td>
<td>4.4017</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-5</td>
<td>9091.2</td>
<td>0.080103</td>
<td>4.2334</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-4</td>
<td>3125.2</td>
<td>0.053758</td>
<td>3.9428</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-3</td>
<td>1693.1</td>
<td>0.042909</td>
<td>3.7745</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-2</td>
<td>4169</td>
<td>0.056957</td>
<td>4.0429</td>
<td>1.789</td>
</tr>
<tr>
<td>Metoprolol-1</td>
<td>2254.8</td>
<td>0.045384</td>
<td>3.8746</td>
<td>1.789</td>
</tr>
<tr>
<td>Mianserin-21</td>
<td>37194</td>
<td>0.03732</td>
<td>5.1769</td>
<td>3.433</td>
</tr>
<tr>
<td>Mianserin-20</td>
<td>3.25E+05</td>
<td>0.12452</td>
<td>5.595</td>
<td>3.433</td>
</tr>
<tr>
<td>Chlordiazepoxide</td>
<td>788.71</td>
<td>0.043219</td>
<td>3.4396</td>
<td>2.49</td>
</tr>
<tr>
<td>Ioxamic</td>
<td>11347</td>
<td>0.095407</td>
<td>4.2537</td>
<td>1.38</td>
</tr>
</tbody>
</table>

Continued
Table C1 continued

<table>
<thead>
<tr>
<th>Compound</th>
<th>Value</th>
<th>Exponent</th>
<th>Exponent 1</th>
<th>Exponent 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ibuprofen</td>
<td>1.71E+05</td>
<td>1.2302</td>
<td>4.3213</td>
<td>3.722</td>
</tr>
<tr>
<td>Aniline</td>
<td>43.513</td>
<td>0.6057</td>
<td>1.0348</td>
<td>0.936</td>
</tr>
<tr>
<td>Propranolol</td>
<td>22890</td>
<td>0.19225</td>
<td>4.2542</td>
<td>3.097</td>
</tr>
<tr>
<td>Aldosterone</td>
<td>86699</td>
<td>0.44665</td>
<td>4.4664</td>
<td>0.458</td>
</tr>
<tr>
<td>Thioridazine</td>
<td>7.66E+05</td>
<td>0.034611</td>
<td>6.5233</td>
<td>6.129</td>
</tr>
<tr>
<td>Alfaxalone</td>
<td>8.04E+06</td>
<td>1.4434</td>
<td>5.9244</td>
<td>2.566</td>
</tr>
<tr>
<td>Mianserin-15</td>
<td>39195</td>
<td>0.035085</td>
<td>5.2265</td>
<td>3.433</td>
</tr>
<tr>
<td>Mianserin-14</td>
<td>3.47E+05</td>
<td>0.11866</td>
<td>5.6445</td>
<td>3.433</td>
</tr>
<tr>
<td>Metoprolol</td>
<td>4169</td>
<td>0.056957</td>
<td>4.0429</td>
<td>1.789</td>
</tr>
<tr>
<td>Mianserin-9</td>
<td>42941</td>
<td>0.03422</td>
<td>5.277</td>
<td>3.433</td>
</tr>
<tr>
<td>Mianserin-8</td>
<td>3.83E+05</td>
<td>0.11658</td>
<td>5.695</td>
<td>3.433</td>
</tr>
<tr>
<td>Valproic_acid</td>
<td>210.87</td>
<td>0.82202</td>
<td>1.5875</td>
<td>2.719</td>
</tr>
<tr>
<td>Mianserin-3</td>
<td>45117</td>
<td>0.032076</td>
<td>5.3266</td>
<td>3.433</td>
</tr>
<tr>
<td>Promethazine</td>
<td>15929</td>
<td>0.027721</td>
<td>4.9378</td>
<td>4.686</td>
</tr>
<tr>
<td>Mianserin-2</td>
<td>4.07E+05</td>
<td>0.11063</td>
<td>5.7446</td>
<td>3.433</td>
</tr>
<tr>
<td>Perphenazine</td>
<td>66103</td>
<td>0.0067242</td>
<td>6.171</td>
<td>4.494</td>
</tr>
<tr>
<td>Ethane</td>
<td>70.351</td>
<td>0.99882</td>
<td>1.0262</td>
<td>1.818</td>
</tr>
<tr>
<td>Propane</td>
<td>266.57</td>
<td>1.3794</td>
<td>1.4645</td>
<td>2.35</td>
</tr>
<tr>
<td>Butane</td>
<td>945.89</td>
<td>1.784</td>
<td>1.9028</td>
<td>2.881</td>
</tr>
<tr>
<td>Pentane</td>
<td>3220.4</td>
<td>2.2138</td>
<td>2.3412</td>
<td>3.412</td>
</tr>
<tr>
<td>Hexane</td>
<td>10656</td>
<td>2.6699</td>
<td>2.7795</td>
<td>3.944</td>
</tr>
</tbody>
</table>

Continued
<table>
<thead>
<tr>
<th>Substance</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heptane</td>
<td>34534</td>
<td>3.1536</td>
<td>3.2178</td>
<td>4.475</td>
</tr>
<tr>
<td>Octane</td>
<td>1.10E+05</td>
<td>3.6664</td>
<td>3.6562</td>
<td>5.006</td>
</tr>
<tr>
<td>Decane</td>
<td>1.08E+06</td>
<td>4.7844</td>
<td>4.5328</td>
<td>6.069</td>
</tr>
<tr>
<td>Ethanol</td>
<td>7.6261</td>
<td>1.0663</td>
<td>0.0328</td>
<td>-0.187</td>
</tr>
<tr>
<td>Propanol</td>
<td>20.077</td>
<td>1.0232</td>
<td>0.4712</td>
<td>0.344</td>
</tr>
<tr>
<td>Butanol</td>
<td>54.116</td>
<td>1.0052</td>
<td>0.9095</td>
<td>0.875</td>
</tr>
<tr>
<td>Pentanol</td>
<td>147.98</td>
<td>1.0018</td>
<td>1.3478</td>
<td>1.407</td>
</tr>
<tr>
<td>Hexanol</td>
<td>408.51</td>
<td>1.008</td>
<td>1.7861</td>
<td>1.938</td>
</tr>
<tr>
<td>Heptanol</td>
<td>1135.4</td>
<td>1.0211</td>
<td>2.2245</td>
<td>2.469</td>
</tr>
<tr>
<td>Octanol</td>
<td>3171</td>
<td>1.0394</td>
<td>2.6628</td>
<td>3.001</td>
</tr>
<tr>
<td>Nonanol</td>
<td>8889.6</td>
<td>1.0621</td>
<td>3.1011</td>
<td>3.532</td>
</tr>
<tr>
<td>Decanol</td>
<td>24993</td>
<td>1.0883</td>
<td>3.5395</td>
<td>4.063</td>
</tr>
</tbody>
</table>
LIST OF REFERENCES


MATLAB Software. The Math Works Inc., 24 Prime Park Way, Natick, MA 01760-1500, USA.


ACD®/LogP Technical Information.  


