Building a robust clinical diagnosis support system for childhood cancer using data mining methods

A Thesis Submitted for the Degree of
Doctor of Philosophy

By
Hamid Ghous

in

School of Software
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
JUNE 2016

CERTIFICATE

Date: **JUNE 2016**

Author: **Hamid Ghous**

Title: **Building a robust clinical diagnosis support system for childhood cancer using data mining methods**

Degree: **Ph.D.**

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.

Signature of Author

Acknowledgements

I am greatly indebted to my supervisor, Dr. Paul Kennedy for his continuous encouragement, advice, help and invaluable suggestions. He is such a nice, generous, helpful and kindhearted person. I feel really happy, comfortable and unconstrained with him during my PhD study. I owe my research achievements to his experienced supervision. Many thanks are also due to my co-supervisor, Dr. Daniel Catchpoole for his valued suggestions and constant support, and for the numerous conversations with him. I appreciate the travel support for attending the international conferences which I received from the School of Software and the Vice-Chancellor's Conference Fund. I wish to thank my fellow research students and the staff of the school for providing various assistance for the completion of this research work.

I would like to thank my family for their understanding and assistance. I also like to thank them for the freedom to study for the long time necessary to complete postgraduate studies. This thesis could not have been completed without the support and encouragement of my family and friends. I would also like to thank Aedan Roberts from The Children's Hospital at Westmead for his support.

To My Family and friends

Table of Contents

List of Tables

List of Figures

Abstract

Progress in understanding core pathways and processes of cancer requires thorough analysis of many coding and noncoding regions of the genome. Data mining and knowledge discovery have been applied to datasets across many industries, including bioinformatics. However, data mining faces a major challenge in its application to bioinformatics: the diversity and dimensionality of biomedical data. The term 'big data' was applied to the clinical domain by Yoo et al. (2014), specifically referring to single nucleotide polymorphism (SNP) and gene expression data. This research thesis focuses on three different types of data: gene-annotations, gene expression and single nucleotide polymorphisms.

Genetic association studies have led to the discovery of single genetic variants associated with common diseases. However, complex diseases are not caused by a single gene acting alone but are the result of complex linear and non-linear interactions among different types of microarray data. In this scenario, a single gene can have a small effect on disease but cannot be the major cause of the disease. For this reason there is a critical need to implement new approaches which take into account linear and non-linear gene-gene and patient-patient interactions that can eventually help in diagnosis and prognosis of complex diseases. Several computational methods have been developed to deal with gene annotations, gene expressions and SNP data of

complex diseases. However, analysis of every gene expression and SNP profile, and finding gene-to-gene relationships, is computationally infeasible because of the high-dimensionality of data. In addition, many computational methods have problems with scaling to large datasets, and with overfitting. Therefore, there is growing interest in applying data mining and machine learning approaches to understand different types of microarray data.

Cancer is the disease that kills the most children in Australia (Torre et al., 2015). Within this thesis, the focus is on childhood Acute Lymphoblastic Leukaemia. Acute Lymphoblastic Leukaemia is the most common childhood malignancy with 24% of all new cancers occurring in children within Australia (Coates et al., 2001). According to the American Cancer Society (2016), a total of 6,590 cases of ALL have been diagnosed across all age groups in USA and the expected deaths are 1,430 in 2016.

The project uses different data mining and visualisation methods applied on different types of biological data: gene annotations, gene expression and SNPs.

This thesis focuses on three main issues in genomic and transcriptomic data studies:

(i) Proposing, implementing and evaluating a novel framework to find functional relationships between genes from gene-annotation data.

(ii) Identifying an optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression.

(iii) Proposing, implementing and evaluating a novel feature selection approach to identify related metabolic pathways in ALL

This thesis proposes, implements and validates an efficient framework to find functional relationships between genes based on gene-annotation data. The framework is built on a binary matrix and a proximity matrix, where the binary matrix contains information related to genes and their functionality, while the proximity matrix shows similarity between different features.The framework retrieves gene functionality information from Gene Ontology (GO), a publicly available database, and visualises the functional related genes using singular value decomposition (SVD). From a simple list of gene-annotations, this thesis retrieves features (i.e Gene Ontology terms) related to each gene and calculates a similarity measure based on the distance between terms in the GO hierarchy. The distance measures are based on hierarchical structure of Gene Ontology and these distance measures are called similarity measures. In this framework, two different similarity measures are applied:

(i) A hop-based similarity measure where the distance is calculated based on the number of links between two terms.

(ii) An information-content similarity measure where the similarity between terms is based on the probability of GO terms in the gene dataset.

This framework also identifies which method performs better among these two similarity measures at identifying functional relationships between genes. Singular value decomposition method is used for visualisation, having the advantage that multiple types of relationships can be visualised simultaneously (gene-to-gene, term-to-term and gene-to-term)

In this thesis a novel framework is developed for visualizing patient-to-patient relationships using gene expression values. The framework builds on the random forest

feature selection method to filter gene expression values and then applies different linear and non-linear machine learning methods to them. The methods used in this framework are Principal Component Analysis (PCA), Kernel Principal Component Analysis (kPCA), Local Linear Embedding (LLE), Stochastic Neighbour Embedding (SNE) and Diffusion Maps. The framework compares these different machine learning methods by tuning different parameters to find the optimal method among them. Area under the curve (AUC) is used to rank the results and SVM is used to classify between relapsed and non-relapsed patients.

The final section of the thesis proposes, implements and validates a framework to find active metabolic pathways in ALL using single nucleotide polymorphism (SNP) profiles. The framework is based on the random forest feature selection method. A collected dataset of ALL patient and healthy controls is constructed and later random forest is applied using different parameters to find highly-ranked SNPs. The credibility of the model is assessed based on the error rate of the confusion matrix and kappa values. Selected high ranked SNPs are used to retrieve metabolic pathways related to ALL from the KEGG metabolic pathways database.

The methodologies and approaches presented in this thesis emphasise the critical role that different types of microarray data play in understanding complex diseases like ALL. The availability of flexible frameworks for the task of disease diagnosis and prognosis, as proposed in this thesis, will play an important role in understanding the genetic basis to common complex diseases.

This thesis contributes to knowledge in two ways:

(i) Providing novel data mining and visualisation frameworks to handle biological data.

(ii) Providing novel visualisations for microarray data to increase understanding of disease.

Author's Publications

(i) H. Ghous, P. J. Kennedy, N. Ho, and D. R. Catchpoole. Comparing functional visualisations of lists of genes using singular value decomposition. Journal of Research and Practice in IT, 2013. Accepted, 02/10/13.

(ii) H. Ghous, N. Ho, D. R. Catchpoole, and P. J. Kennedy. Functional visualisation of genes using singular value decomposition. In Y. Zhao, J. Li, P. J. Kennedy, and P. Christen, editors, Proceedings of the Tenth Australasian Data Mining Conference (AusDM 12), volume 134 of Conferences in Research and Practice in IT (CRPIT), pages 53-59. Australian Computer Society, 2012.

(iii) H. Ghous, N. Ho, D. R. Catchpoole, and P. J. Kennedy. Comparing functional visualizations of genes. In J. Maria Pena and F. Famili, editors, 5th Workshop on Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions, European Conferencce on Machine Learning and Principals and Practice of Knowledge Discovery in Databases 2011, pages 12-21. 5-9 September 2011.

(iv) H. Ghous, P. J. Kennedy, D. R. Catchpoole, and S. J. Simoff. Kernel-based visualisation of genes with the gene ontology. In J. F. Roddick, J. Li, P. Christen, and P. J. Kennedy, editors, Data Mining and Analytics 2008: proceedings of the

Seventh Australasian Data Mining Conference (AusDM08), volume 87 of Conferences in Research and Practice in IT (CRPIT), pages 133-140. Australian Computer Society, Sydney, 2008.

Abbreviations and Glossary

Term	**Abbreviation**	**Definition**
Acute Lymphoblastic Leukaemia	ALL	
Biased data		Data where cases of one class is very high compared to other class
Classification		A supervised learning process of assigning cases to target class
Data transformation		A process to transform or consolidate data for effective use
Diffusion Maps	DM	
Dimensionality reduction		Reducing the number of random variables in a dataset
Gene Ontology	GO	
High dimensionality		Data whose dimension is higher than dimensions considered in traditional multivariate analysis
Kernel Principal Component Analysis	kPCA	
Kyoto Encyclopaedia of Genes and Genomes	KEGG	
Local Linear Embedding	LLE	
Nonlinear Dimensionality Reduction	NLDR	
Prediction		The process of predicting the outcome of certain case
Principal Component Analysis	PCA	
Prognosis		Predicting likely outcome of patient
Random Forest	RF	
Singular Value Decomposition	SVD	
Single Nucleotide Polymorphism	SNPs	
Stochastic Neighbour Embedding	SNE	

Chapter 1

Introduction

The human genome is complex. Within this complexity lies information about the individual. Functionally, this complex genome underpins the biological mechanisms of a patient's disease. Hence, when building a framework for the personalised treatment of disease, the complexity of the genome must be captured in meaningful and actionable ways. Presently, however, the knowledge provided by the genome is inaccessible to a clinician making personalised patient management decisions, yet it is this information that allows us to identify patients as individuals.

Microarray and high-throughput technology has allowed for an individual patient's genome to be deciphered in rapid time. This microarray data can be very high-dimensional, biased, nonlinear and gathered in multiple forms such as gene-expression, gene annotations and Single Nucleotide Polymorphisms (SNPs). this thesis builds data analysis frameworks to handle complex genomic data of childhood cancer derived from an individual patient. The data collected from individual patients is high-dimensional data, biased, non-linear and contains missing values. This thesis proposes robust strategies to handle three different types of high-dimensional biological data: gene annotations, gene expression and SNPs, around principles in

data mining, visualisation, human interaction and interpretation for clinicians. The paradigm is the childhood cancer Acute Lymphoblastic Leukaemia (ALL), a type of leukaemia that is common in children.

The focus of this thesis is to visualise the functional relationships between genes using ALL gene annotation data, classify ALL patients using gene expression profiles and find metabolic pathways in ALL using SNPs profiles.

1.1 Research Objective

The primary objective of this thesis is to develop a robust data analysis framework to handle the complex genomic data of childhood cancer. This objective can be divided into three sub-objectives:

(i) Develop, implement and evaluate a novel framework to find functional relationships between genes from gene-annotation data.

(ii) Identifying an optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression.

(iii) Develop, implement and evaluate a novel feature selection approach to identify metabolic pathways in ALL using SNPs profiles.

1.2 Significance

Cancer is the disease that kills the most children in Australia (Torre et al., 2015). Within this thesis, the focus is on childhood Acute Lymphoblastic Leukaemia (ALL). According to the American Cancer Society (2016), a total of 6,590 cases of ALL have

been diagnosed across all age groups in USA and the expected deaths are 1,430 in 2016. Acute Lymphoblastic Leukaemia is the most common childhood malignancy in Australia accounting for 24% of all new cancers occurring in children (Coates et al., 2001). According to the National Cancer Institute (2013), ALL accounts for more than 5,000 new cases of leukaemia each year in the USA. Today, despite good treatment outcomes overall, the major obstacle to curing is the recurrence of evident disease in the patient or relapse (Henze et al., 1991). The likelihood of cure for ALL after relapse is poor. At present, treatment depends on diagnosis using a broad-brush allocation of patients to a few defined risk stratification groups, followed by the use of a complex combination of cytotoxic agents. Pollock et al. (2000) suggests that treatment should rather be determined by each patient‘s unique genetic makeup, leading to the implicit notion that patients with similar genotypes will have similar outcomes to similar treatments. this thesis takes such an approach and validates it in ALL cancer.

1.3 Problem statement

This thesis hypothesizes that integrating different types of biological data such gene annotations, gene expressions and genome-wide SNP profiles, will be effective for modeling, data mining and visualisation of gene-to-gene and patient-to-patient relationships. For modeling different aspects of any complex disease, there is a need to develop a framework that can use different types of data available to reliably identify patients at greater risk of catching a disease or not responding to current treatment. The ultimate goal of such work is to develop computational tools that will allow doctors and researchers to examine patients' genetic background to answer several

research questions. Questions such as finding functional relationships between genes, identifying those patients who have a high chance of relapse using established therapy, to learn why patients relapse, and to help choose treatment strategies which best suit each patient, more specifically, translating genomic knowledge into public health and medical care in the form of personalized medicine. Based on that, this research study endeavours to answer the following research questions:

(i) Can I develop, implement and evaluate computational framework based on data mining and visualisation methods to find relationships between ALL genes using gene annotation?

(ii) Can I find which dimensionality reduction method performs better for ALL gene expression high-dimensional data, to eventually help in finding patient-to-patient relationships?

(iii) Can SNPs profiles be used to find metabolic pathways related to ALL patients?

1.4 Contributions of the thesis

Among many challenges, there are two major challenges that biologists are currently facing: the variety of types of biological data available and the "curse of dimensionality", a term defined to handle data containing very high number of features Clarke et al. (2009). Biological data is challenging to analyse using traditional data mining approaches because it comes in a lot of different forms (e.g. expression and SNPs profiles) and it is limited in terms of datapoints (or observations), compared to other datatypes (Marx, 2013). This thesis will contribute to solving both challenges by using different types of biological data: gene annotations, gene expression and SNPs

profiles, and then proposing novel frameworks using data mining and visualisation methods to handle high-dimensional data.

In this work, data mining approaches have been chosen to deal with linear and non-linear interactions within the data. The proposed approaches can be used for resolving disease models including diagnosis and visualizing patient-to-patient relationships. The approaches developed in this thesis use data mining, dimensionality reduction and feature selection methods to find gene-to-gene and patient-to-patient relationships associated with ALL. Applying traditional dimensionality reduction methods does not guarantee best results as discussed in Section 3.2.2. There is a need to compare these methods and find the best method to handle the "curse of dimensionality" and nonlinearity of ALL patients' data.

The major contribution of this work is the application of data mining, machine learning and feature selection methods to ALL data. By analyzing the main challenges of the given domain, this thesis identifies three contributions to knowledge. These contributions are first listed and then explained below.

(i) Proposing, implementing and evaluating a novel framework to find functional relationships between genes from gene-annotation data.

(ii) Identifying the optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression.

(iii) Proposing, implementing and evaluating a novel feature selection approach to identify related metabolic pathways in ALL

Each of these contributions is validated by application to real-world problems of ALL studies.

1.4.1 Finding functional relationships between genes

The critical question that comes to mind in dealing with biological data is the type of data available to the biologist at the time of investigation; in this case, if gene annotations are available, then this thesis can help biologists find functional relationships between genes. Although there are many tools available to find functional relationship between genes, Huang et al. (2008) and recently Wagholikar et al. (2012) has put emphasis on new tools required in this field. The framework in Section 4.2 is based on a binary matrix showing relationships between genes and Gene Ontology terms (explained in detail in Section 2.2.3), and a proximity matrix that finds similarity between Gene Ontology terms. A visualisation method is used to explore genes and their features in details. This work directly contributes to helping biologists identify possible functional relationships between genes which will eventually help them in diagnosis and classification of disease.

Comparing different similarity measures

The diversity of microarray data leads to a question, Can we rely on one method for analysis and visualisation of complex data? To handle this problem, the experiment in Section 4.2.2 compares two datasets using two similarity measures. The proposed method will contribute directly in helping biologists to run dataset through different similarity measures to find the most likely functional relationship between genes. Comparing results of both datasets may give a better understanding of genes involved in disease and identify new relationships to explore.

1.4.2 Comparing different non-linear and linear dimensionality reduction methods

Gene expression data is high-dimensional microarray data affected by the "curse of high dimensionality" especially when patient numbers are small as in ALL. Lee and Verleysen (2007) have classified different available dimensionality reduction methods into hierarchy of linear and non-linear methods. This thesis investigates which method performs better on ALL gene expression data.

Gene expression data for a patient consists of hundreds of thousands of transcript abundance measures for a patient. Chapter 5 compares different available linear and nonlinear methods (describe in Section 5.2). The comparison is performed on ALL dataset derived from patients at the CHW. The framework proposed in Chapter 5 is based on a feature selection method and five different linear and nonlinear dimensionality reduction methods. Haghverdi et al. (2015) has also urged comparison between methods to find the best method. this thesis will contribute to knowledge in two ways, firstly by showing the comparison between different linear and nonlinear dimensionality reduction methods and determine the best method among them for ALL gene expression data.

Secondly, a novel framework is proposed for visualisation of gene expression data by applying a feature selection method first and then applying dimensionality reduction methods to fine tune the results. The importance of this work can be seen in the fact that Haghverdi et al. (2015) found that Diffusion Maps method performed best in their study while in this thesis Diffusion Maps performed the worst (discussed later in Chapter 5).

For the task of visualising patient-to-patient relationships, Chapter 5 employs

several data reduction methods for visualizing gene expression data. Information visualisation is considered as a direct way to help browse the datasets. The visualisation results are an important tool that can be used to assist clinicians and biomedical researchers in understanding the different structure of patients and to compare different clusters in the visualisation. The main challenge in visualizing gene expression datasets stems from the high dimensionality of the data. Different dimensionality reduction methods were applied to gene expression data of leukaemia patients to determine the best method for visualizing this type of data. Visualisation approaches were compared based on measures such as area under the curve. The resultant visualisations were more accurate and informative in discriminating the major characteristics of the dataset.

1.4.3 Finding metabolic pathways in ALL

The framework proposed in Chapter 6 applies the 'random forest' feature selection method on high dimensional ALL SNP data to find pathways related to ALL (as described in Section 6.3). The framework aims to find the high-ranked SNPs from the dataset and then retrieves pathways related to those SNPs using the KEGG database. In a later section, the proposed framework in chapter 4 (Section 4.2) is applied on genes related to these high-ranked SNPs for functional understanding. Chapter 6 contributes to knowledge by introducing a novel framework to extract metabolic pathways through SNP data using random forest method.

1.5 Thesis structure

This thesis has been divided into seven chapters:

Chapter 1 gives a brief introduction and background history of the work undertaken along with the key objectives and significance of this work.

Chapter 2 provides the background about ALL cancer, datasets, public databases and the methods used in this project.

Chapter 3 provides a detailed literature survey about the functional relationship between genes, linear and non-linear dimensionality reduction techniques and finding metabolic pathways using genome data. The literature survey identifies the gap in current research knowledge in terms of data analysis.

Chapter 4 describes a novel experimental design for finding gene-to-gene relationships in microarray data and compares different similarity measures with visualisation techniques.

Chapter 5 proposes a framework to handle the curse of dimensionality based on different dimensionality reduction techniques. A comparison is presented between these different dimensionality reduction techniques.

Chapter 6 proposes a novel framework to find metabolic pathways related to ALL using SNP data.

Chapter 7 concludes the thesis and summarizes the key findings of this work and discusses further courses of action in the light of results obtained.

Chapter 2

Background

It is becoming clear that progress towards new insights in cancer treatment requires a thorough analysis of many genes (Jones et al., 2008). The routine use of high–throughput technology in biomedical research has made large volumes of data available to biologists. However, the sheer scale of this data makes using it challenging. Also, as integration of multiple data types becomes commonplace, making sense of the data becomes even more difficult. Researchers are facing a significant challenge to process this biological data, and to retrieve information for diagnosis and prognosis of diseases. Many machine learning methods developed over past decades offer promising approaches to meeting this challenge, but systematic approaches to using these methods are required to handle this complex biological data. The focus of this chapter is to elaborate on cancer, types of biological data, Ontologies available to retrieve information related to data and methods used in this thesis.

2.1 Acute lymphoblastic leukaemia (ALL)

Leukaemia is a type of blood or bone marrow cancer characterized by an abnormal increase in immature white blood cells (blast cells) (Fauci, 2008). There are three

classes of blood cells: white blood cells, red blood cells and platelets. Normally white blood cells help the body fight against infection, red blood cells carry oxygen to tissues throughout the body and platelets help control bleeding. All these classes are differentiated from stem cells in bone marrow. These cells have a limited lifespan and are regenerated once they grow old, or get damaged and die. A leukaemia patient normally has over-production of white blood cells in an uncontrolled fashion which results in diffuse replacement of normal bone marrow with leukaemic cells. These leukaemic cells infiltrate organs such as the liver, spleen, lymph nodes, meninges and gonads, leading ultimately to bone marrow failure. Leukaemia affects people of all age groups: in 2013, approximately 48,610 children and adults in the United States developed some form of leukaemia and 23,720 died from it (National Cancer Institute, 2013).

2.1.1 Clinical and biological description of ALL

Leukaemia can be subdivided into acute and chronic forms (Fauci, 2008). In acute leukaemia the history is usually brief and life expectancy without treatment is short, and in chronic leukaemia the patient has been unwell for years, and survival is long (in years). There are different types of leukaemia but the focus of this thesis is on Acute Lymphocytic Leukaemia (ALL).

ALL affects lymphoid cells and is rapidly progressive. Lymphoid tissues are made up partially of lymphocyte cells, and these cells develop from lymphoblasts to become mature and infection-fighting cells. There are two main types of ALL based on lymphocytes cells; B-ALL and T-ALL (Fauci, 2008). B-ALL develops from B lymphocytes cells which are responsible for protecting the body from invading germs by

making antibodies. T-ALL develops from T lymphocytes cells which are capable of directly destroying germs or slowing the activity of the immune system. ALL caused from the early forms of these lymphocytes then spreads to different parts of the body through blood such as liver and lymph nodes (Pui and Evans, 1998).

According to ACS (2016), a total of 6,590 cases of ALL have been diagnosed across all age groups in USA and the expected deaths are 1,430 in 2016. Biologists can collect data from ALL patients in different forms such as gene annotations, gene exprssion values and SNPs. But this data is either very high-dimensional (based on very high number of attributes or contains alot of information about patient), biased (cases of one type of patients are higher than other types) or hard to understand. There is a need to analyse this data and provide visualisations to help biologists in better understanding of different types of data. This thesis focuses on finding functional relationship between gene-annotation, classifying ALL patients based on gene expression values and finding ALL metabolic pathways using SNPs profiles.

2.1.2 Genetic and molecular basis of ALL

The contribution of genetic factors to ALL has been thoroughly investigated by many biologists. These genetic factors can manifest as either over- or under- expression of genes, or hyper-activation of pathways. Vlierberghe and Ferrando (2012) have suggested that ALL patients share unique gene expression signatures and molecular alterations. These irregularities in gene expression profiling cause deregulation in cellular processes such as cell cycle signalling, cell growth and proliferation in ALL patients. Similarly, Forero et al. (2013) found that T-ALL is caused by deregulated expression of normal transcription factor proteins. They further discussed that the

rearranged alleles in TAL1-related genes transcribed in T-ALL cells can be a factor in T-ALL.

Similar to gene expression, pathway deregulation in cellular pathways has been associate with ALL (Collins-Underwood and Mullighan, 2010). Zhao (2010) suggested that the Phosphatidylinositol 3-kinase (PI3K)/Akt/mammalian target of rapamycin pathways plays an important role in cell proliferation and progress of cell cycle. Scientists have put emphasis on identifying and analysing these alterations in gene expression and pathways of ALL (Mullighan (2012),Collins-Underwood and Mullighan (2010) and Roberts and Mullighan (2015)).

Along with gene expression and pathways, genetic variations in the form of SNPs, discussed in Section 2.2.2, plays an important role in prognosis (Pui and Evans (2006) and Yang et al. (2009)) and molecular lesions (Mullighan et al., 2007) in ALL. This suggests that the genetic and molecular basis of ALL are highly dependent on gene-expression, pathways and SNPs of ALL patients.

2.2 Data relevant to studying ALL

A broad range of genomic and transcriptomic data types are available for this thesis to advance knowledge in diagnosis and prognosis of ALL. this thesis focuses on (i) gene expression data and (ii) genomic variation data in the form of patient-linked single nucleotide polymorphisms (SNPs).

2.2.1 Gene expression data

Proteins control cell function and genes encode proteins. A set of genes is encoded on a genome in the sequence of chemical building blocks called nucleotides. These nucleotides are arranged along a deoxyribonucleic acid (DNA) molecule and associated to ribonucleic acid (RNA). During the process of protein building, genetic information is transferred from DNA to messenger RNA (mRNA). The process of transcribing information from DNA to mRNA is called gene expression or transcription (O'Connor et al., 2010). These mRNA molecules are used as a template for the construction of new proteins, and this process is called translation.

Gene products are not needed all the time by organisms, nor are they needed in identical amounts, so every gene is not expressed all the time in a cell. Scientists have been able to find techniques to show which genes are "turned on" and which are "turned off" in a tissue at a given time. Microarray technologies have made it possible to study different diseases using gene expression data (Ziauddin and Sabatini, 2001). More specifically, these gene expression patterns can help scientists to explore the current status of disease, progress of disease, cellular response to stimuli and target drug identification.

cDNA microarray technologies measure the abundance of mRNA transcripts by binding them to small oligonucleotide sequences called "probes" (Fodor et al., 1993). These probes represent sections of sequences of known mRNA transcripts. Microarrays contain tens of thousands of probes printed onto a solid surface or onto polystyrene beads, with several copies of each probe located in different regions to minimise noise effects.

High throughput technologies like microarrays produce huge amounts of data

across a range of diseases, and to store this data scientists require databases from where they query, compare and retrieve information. Gene Expression Omnibus (GEO) is an example of a public repository designed to store high-throughput gene expression data (Edgar et al., 2002). GEO contains approximately over a billion individual gene expression profiles, derived from over 100 organisms (Barrett and Edgar, 2006). Thousands of gene expression values are gathered for a single patient which forms a huge dataset for any specific disease. In these datasets each gene expression value represents a dimension for a patient and since there are many tens of thousands of gene expression values, the resulting dataset is very high-dimensional. To find patterns within high-dimensional data users need data mining and visualisation methods to analyses this data (Clarke et al., 2009). The ALL dataset used in study of chapter 5 is high-dimensional gene expression dataset gathered from GEO. An example of gene expression data in GEO can be found in Fig. 2.1.

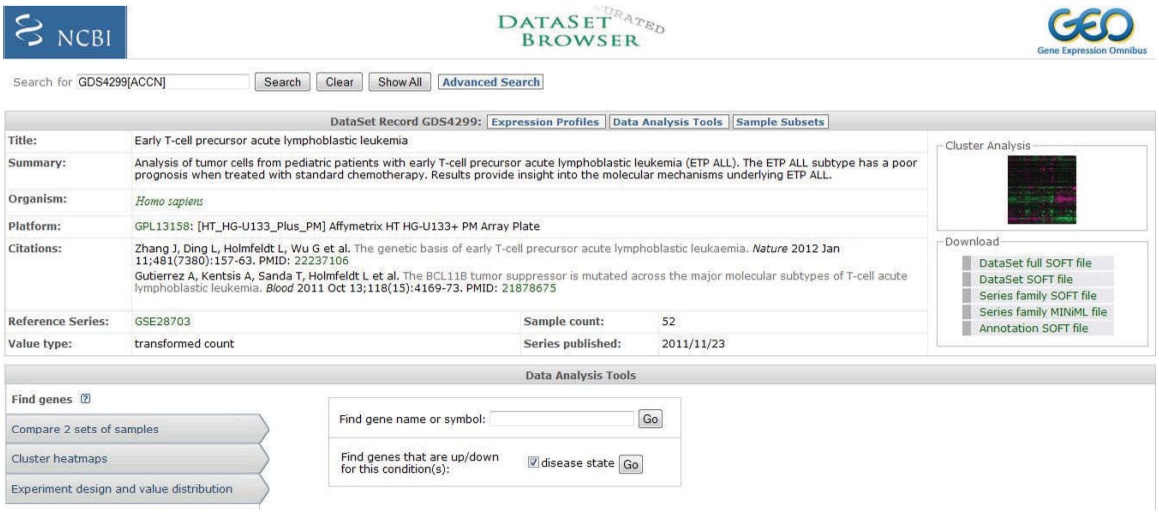


Figure 2.1: An example of a dataset record in GEO for the gene GDS4299.

2.2.2 Genomic variation data

The human genome contains 1 to 10 million genetic variations in the form of sequence polymorphisms: a polymorphism is a variation of a DNA sequence commonly defined as occurring in at least 1% of the population (Cavalli-Sforza and Bodmer, 1999). There are several types of polymorphism in the human genome, and one of the most commonly-studied are Single Nucleotide Polymorphisms (SNPs). SNPs are variations in single nucleotides between individuals or samples, and they are the simplest but most abundant type of genetic variation in humans. Genetic variations, especially

SNPs, are known to be a key feature in discovering disease-causing genes (Treviño et al., 2009).

As well as studying gene expression, microarrays can be used to study SNPs. For example, Illumina array-based technology uses oligonucleotide probes to measure more than 1 million SNPs on a single chip (Illumina, 2014), consisting of around 50-mer oligonucleotides per SNP. Because each genome contains 1-10 million SNPs, this data is also very high-dimensional. In Chapter 6, ALL SNP data is processed through feature selection methods to identify metabolic pathways associated with ALL.

2.2.3 Public Ontologies

High-throughput technologies are producing huge amounts of microarray data including gene expression and SNP data across a range of diseases. To record and store this data, scientists require databases from where they query, compare and retrieve information. Such databases include Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG).

Gene Ontology

The Gene Ontology (GO) provides a controlled vocabulary to describe genes and gene product attributes in many organisms (Ashburner et al., 2000). GO can provide ways to find features related to each genes which will be used in this thesis to find functional relationship between genes. GO is a collaborative effort beginning in 1998 and spans many organisms including, but not limited to, Drosophila, Saccharomyces, mouse and human. Gene Ontology can be broadly split into two parts:

(i) Ontology

(ii) Annotation

The Ontology part is divided into three disjoint trees or hierarchies: 'Cellular component', 'molecular function' and 'biological process'. Cellular component is associated with the physical structure and location of the genes or gene products. It describes both the external connection to the environment, and internal structure (Ashburner et al., 2000).

Molecular function describes the biochemical activity of gene products, but molecular function only defines what the gene product does without specifying location or context.

Biological process describes the biological objective to which the gene or gene product contributes. A biological process is accomplished via one or more ordered assemblies of molecular functions. This suggests that molecular functions are initiators of biological processes(Ashburner et al., 2000).

The second part is Annotation, the characterization of gene products using terms from the ontology. The building blocks of the Gene Ontology are the terms. Each entry in GO has (i) a unique alphanumerical identifier (GO:########); (ii) a term name, e.g. oxidoreductase activity; (iii) synonyms (if applicable); and (iv) a definition. Each term is also assigned to one of the three hierarchies, which are structured as directed acyclic graphs. Most terms have a textual definition, with references stating the source of the definition. If any clarification of the definition or remarks about term usage is required, these are held in a separate comments field. Each gene has one or more terms related to it and a term may have multiple parents on the tree. The terms provide us with a description of the functionality of a gene.

Table 2.1 shows three examples of genes with their related terms. Following each term name is the Gene Ontology accession number for the term. One of the challenges with using terms from the Gene Ontology is that each term may give different amounts of information. For example, the gene Aldh1a7 in Table 2.1 contains some very specific terms, such as Retinal metabolic process or Aldehyde dehydrogenase (NAD) activity which give specific and useful information along with other terms such as Cytoplasm or Metabolic process which are more general (high in the hierarchy) and shared by many other genes. These latter terms do not confer much useful information. Also, some genes have been investigated thoroughly and have many annotations (such as Aldh1a7) whilst others are not well annotated (such as Srpx2).

As illustrated in Fig. 2.2, GO terms are related in two main ways: "is–a" and "part–of". The "is–a" relationship is the main relationship seen in the Gene Ontology and represents a simple class–subclass relationship. For example, the figure shows that the term "extracellular space" is an "extracellular region part" and that an "extracellular region part" is a "cellular component". Cellular component is the root of the hierarchy. Less commonly seen is the "part–of" relationship which signals containment. If C is "part–of" D it means that whenever C is present, it is always a part of D, but that C does not always have to be present. For example, in the figure "extracellular region part" is part of "extracellular region".

The Gene Ontology database allows Structured Query Language (SQL) queries of the terms associated with genes, the relationships between terms (parent and child) as well as finding the distance between terms in number of hops. There are also many web-based tools available to query the databases. In this thesis, GO is used to find functional relationships between genes. SQL queries are used to retrieve features

from GO, and singular value decomposition (SVD) is then applied to find relationships between those genes.

Table 2.1: Example of three genes from the Gene Ontology.

Gene Name	**GO term name and accession**
Aldh1a7	cytoplasm (GO:0005737) oxidoreductase activity (GO:0016491) aldehyde dehydrogenase (NAD) activity (GO:0004029) metabolic process (GO:0008152) retinal metabolic process (GO:0042574)
Srpx2	electron transport (GO:0006118) extracellular region (GO:0005576)
Tspan7	biological process (GO:0008150) molecular function (GO:0003674) integral to membrane (GO:0016021) membrane attack complex (GO:0005579)

Kyoto Encyclopaedia of Genes and Genomes

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al., 2008) is a biological resource which aims to link genomes to the biological systems they govern. Like GO, KEGG also allows SQL queries to extract features from it.

In this thesis KEGG is involved in achieving two objectives. Objective 1, where this thesis identifies a gene-annotation dataset based on KEGG classes and finds functional relationship between those genes. For objective 3, the KEGG pathway database is used to retrieve pathways related to important SNPs found using the random forest method.

Kyoto Encyclopaedia of Genes and Genomes consists of a series of interconnected

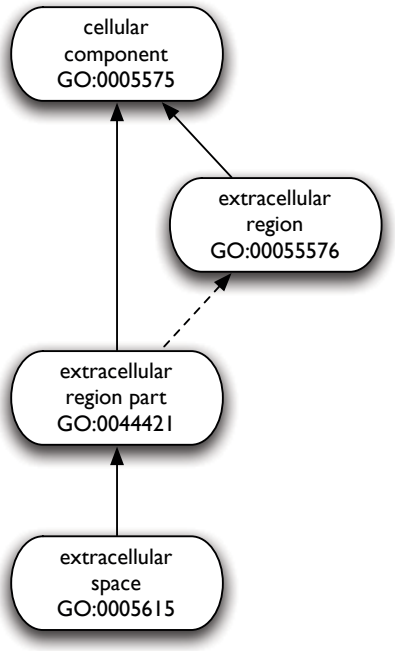


Figure 2.2: Example of a small part of the hierarchical structure of GO terms. Solid lines represent "is–a" relationships and dashed lines represent a "part–of" relationship.

databases of biological systems that interrelate (i) genes and proteins, (ii) chemical building blocks, (iii) molecular interaction pathway diagrams and (iv) hierarchies and relationships of biological objects. The last of these, KEGG BRITE, links genes into a functional hierarchy called the KEGG Orthology (KO). KEGG Orthology is a collection of gene catalogues for all complete genomes and some partial genomes generated from publicly available resources such as NCBI. KEGG contains information for over 826 organisms (Kanehisa et al., 2008). Each gene entry has unique indentifiers: "Entry", gene name, definition, orthology information, pathway and other related information (see in fig.2.3)

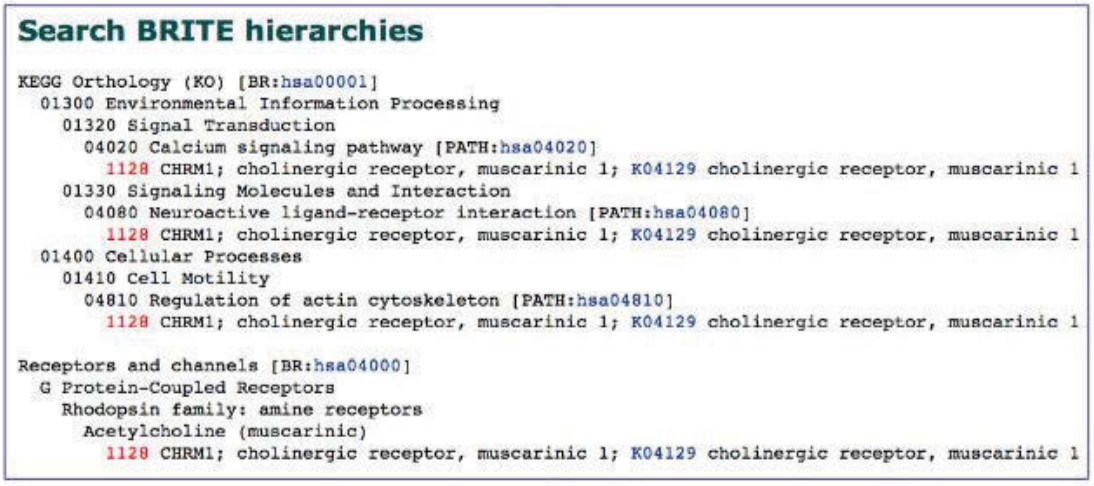


Figure 2.3: Example of the hierarchical structure of KEGG database.

KEGG metabolic pathway database

The KEGG pathway database consists of two sections: metabolic pathways and regulatory pathways. The first entries of metabolic pathway data were from the book "Metabolic Maps" (Nishizuka, 1980) and "Biochemical Pathways" (Gerhard, 1992). The information was then verified and updated with reference to Enzymes and Metabolic Pathways database (EMP) (Selkov et al., 1996) and the Enzyme Handbook volume 13 (Schomburg and Salzmann, 1991).

Kyoto Encyclopaedia of Genes and Genomes pathway consists of 10 categories which contain around 100 metabolic pathway diagrams. There are three main ways to retrieve metabolic diagrams: by using hierarchical text menu, by the hierarchically drawn graphics menu and by key word search using DBGET/LinkDB (Fujibuchi et al., 1998) which is a network based distributed database system which retrieves data from different molecular biology databases. An example of a metabolic pathway for cysteine metabolism is shown in Figure 2.4.

For computation, two main characteristics of the KEGG metabolic pathway diagrams are important. First, each pathway diagram has one reference diagram which is manually drawn and updated. All organism specific diagrams are computationally derived by matching genes in the gene catalogue and the enzyme objects between pathway diagram and its reference diagram. This process helps in immediately reconstructing organism specific pathways.

Second, each enzyme in a pathway is hyperlinked to the enzyme section of the LIGAND database (Suyama et al., 1993) which provides extra information such as the enzyme nomenclature, the reaction scheme, the chemical compounds involved, and additional links to molecular and biological information.

The pathway diagram drawn with KEGG metabolic pathways puts emphasis on two areas: the appearance of two consecutive enzymes in the pathway and the successive conversion of major compounds. This information can be encoded into binary form. In the past these binary form relations have been used in computing possible reaction paths (Goto et al., 1998)). Similarly Ogata et al. (1996) used these enzyme-enzyme binary form relations to find correlations between other types of molecular relations. In this thesis, the KEGG pathway database is used to retrieve pathways related to SNP variations relevant to ALL. Like GO, KEGG and KEGG pathways allow us to use SQL queries to retrieve information from their database.

To achieve Objective 1 of this thesis, the study in chapter 4 will functionally visualise genes using gene-annotation data. To achieve Objective 2 of this thesis, the study in chapter 5 will use ALL gene expression dataset for patient-patient analysis while for Objective 3 of this thesis, ALL SNPs data will be used to identify the active metabolic pathways involved in ALL

this thesis focuses on three types of data relevant to ALL: Gene expression data, SNP data, and gene annotation data. The information related to this data will be retrieved from two public ontologies, GO and KEGG. Once all the information is retrieved from GO and KEGG, data mining methods will be applied for analyses of these datasets.

2.3 Data mining

Data mining (Han and Kamber, 2006) involves finding useful and interesting patterns in large datasets. It has been a flourishing field in computing and has been widely applied to complex and data-intensive domains including insurance (Sumathi and

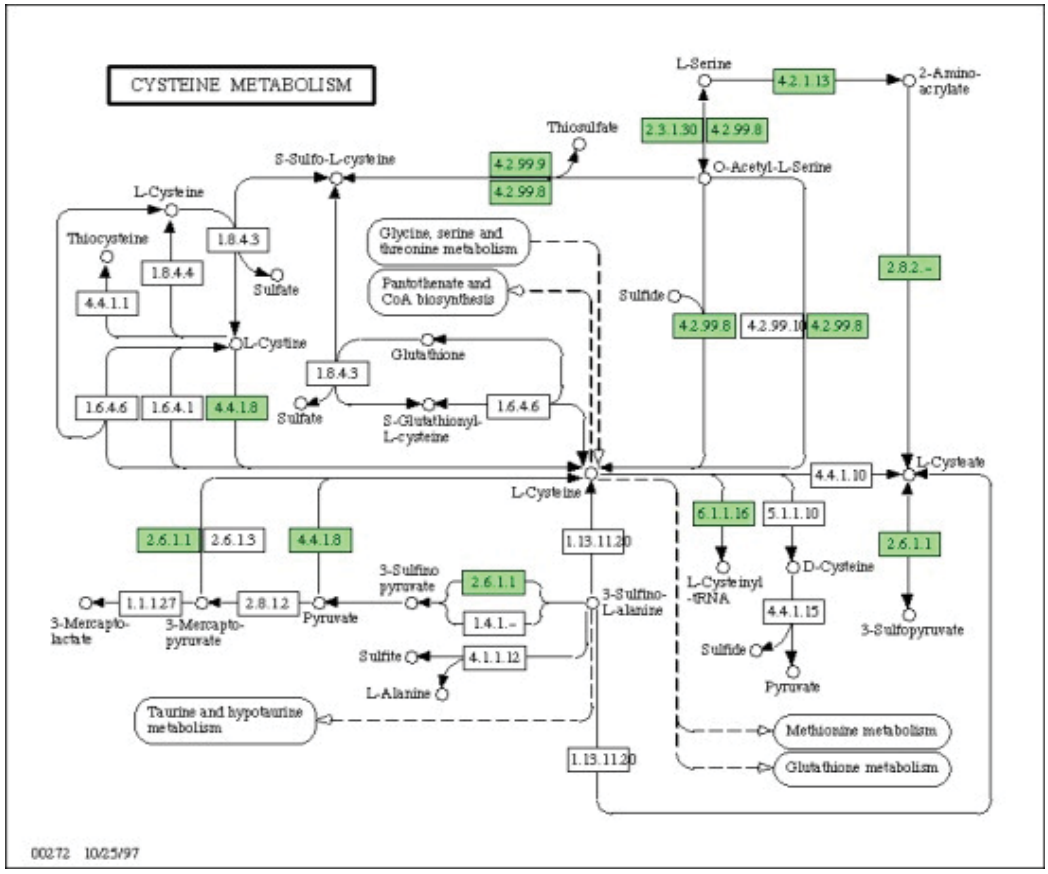


Figure 2.4: The pathway of cysteine metabolism in a KEGG pathway diagram showing the relationships of genes or gene products. A rectangle is a gene product (an enzyme). An enzyme is marked (shaded) when the corresponding gene is found in the genome, in this figure, *Escherichia coli*. (Ogata, 1998).

Sivanandam, 2006), banking (Hormozi and Giles, 2004), astronomy (Borne, 2009) and telecommunications (Weiss, 2005). Data mining strategies are classified into two categories: (i) supervised and (ii) un-supervised methods.

(i) Supervised strategies

Supervised data mining builds models using a known target attribute. These models train on data that includes both the input and the desired output results. It requires careful partition of training, validation and test datasets, which, if not achieved, can result in over-fitting of the model, where the model achieves high accuracy on the training dataset but poor accuracy on other data.

Classification is a supervised technique of data mining, which is used to predict the class of a data instance on the basis of features. It has been used intensively in the past for disease classification, prediction and drug treatment using microarray data. An example of classification on a cancer dataset can be found in Ganesh Kumar et al. (2012) who applied fuzzy Genetic Swarm Algorithm to classify colon cancer, leukemia and lymphoma datasets. Specific descriptions of data mining methods and classification of cancer microarray data to help in diagnosis and prognosis are given later in this literature review. In this thesis, supervised methods are used for classification of ALL patients.

(ii) Unsupervised strategies

In unsupervised learning, there is no target variable and the aim is to identify the structure of the data or to visualise it. All the variables are treated in the same way, which allows data to cluster on the basis of their statistical properties only. Unsupervised learning models can be categorised into clustering, and

dimensionality reduction of data, where the number of attributes is reduced from many thousands in the case of genetic data to fewer while maintaining some aspects of the data such as the average distance between neighboring points. In this thesis, multiple unsupervised learning methods are used to reduce the dimensions of a high-dimensional ALL gene-expression dataset.

The supervised and unsupervised strategies related to this thesis are discussed in the next section.

2.3.1 Methods used in this thesis

Supervised and unsupervised data mining strategies selected for this thesis are the following: Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Kernel Principal Component Analysis (kPCA), Local Linear Embedding (LLE), Stochastic Neighbour Embedding (SNE), Diffusion Maps (DM) and Random Forest (RF).

SVD is used for gene-annotation analysis to find functional relationships between genes which will fulfill objective 1 of this work. PCA, kPCA, LLE, SNE and DM are used for classification of ALL patients. A comparison is made between these methods on ALL patient data. This will lead to objective 2 of this thesis while the random forest method is used for finding important SNPs from the ALL SNP dataset which eventually lead to finding pathways related to ALL. Objective 3 will be achieved using the random forest method.

2.3.2 Singular Value Decomposition

Singular value decomposition (Golub and Van Loan, 1996) is a method that transforms a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ into the orthogonal matrices $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{V} \in \mathbb{R}^{m \times r}$

and a diagonal matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$ where $r \leq m$ is the rank of $\mathbf{X}$.

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{2.3.1}$$

Row vectors of $\mathbf{U}$ relate to the original data points (rows of $\mathbf{X}$) and rows of $\mathbf{V}$ are associated with the data attributes (columns of $\mathbf{X}$). The columns of $\mathbf{U}$ are called the left singular vectors of $\mathbf{X}$ and columns of $\mathbf{V}$ are called the right singular vectors. The elements of $\mathbf{D}$ are termed the singular values of $\mathbf{X}$. Singular value decomposition has been used often in bioinformatics, for example, in visualisation of gene expression values (John Tomfohr and Kepler, 2005), but the novelty in this work is to augment lists of genes with knowledge from a domain ontology and to use the later principal components to extract better understanding.

Singular value decomposition is closely related to the eigendecomposition of the scaled correlation matrix $\mathbf{X}^T\mathbf{X}$ and, hence, is quite similar to the Principal Component Analysis (PCA) of $\mathbf{X}$. The eigenvectors of the scaled correlation matrix are the columns of $\mathbf{U}$ and the square roots of the eigenvalues are the elements of $\mathbf{D}$. However, using SVD allows visualisation of both data points (using rows of the $\mathbf{U}$ matrix) and attributes (using rows of $\mathbf{V}$) at the same time which makes it preferable to PCA.

Values of $\mathbf{D}_{ii}$ are associated with the amount of variance they explain in the data. Consequently, by ordering the columns of $\mathbf{U}$ and $\mathbf{V}$ according to decreasing values of $\mathbf{D}_{ii}$ and choosing the k highest ranked columns, it is possible to perform feature selection or visualisation of the data matrix $\mathbf{X}$. The matrix is

$$\mathbf{X}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T \tag{2.3.2}$$

where $\mathbf{U}_k$ and $\mathbf{V}_k$ are matrices composed of the k first columns of $\mathbf{U}$ and $\mathbf{V}$ respectively and $\mathbf{D}_k$ is the diagonal matrix with the k values of $\mathbf{D}_{ii}$ for $i \leq k$. The matrix

$\mathbf{X}_k$ is the best possible k rank approximation to $\mathbf{X}$ in the least squares sense (Jolliffe, 2004). Matrices $\mathbf{X}_2$ or $\mathbf{X}_3$ are commonly used to visualise the higher dimensional $\mathbf{X}$. Each column of $\mathbf{U}$ and $\mathbf{V}$ can be used to visualise an aspect of variance in the original data set $\mathbf{X}$. SVD will be used to achieve objective 1, to find functional relationships between genes in Chapter 4.

2.3.3 Principal Component Analysis (PCA)

Principal Component Analysis (Pearsons, 1901) is a well known data transformation method that rotates a dataset into a different orthogonal coordinate system such that the coordinates are ordered in decreasing order of the variance in the data. The coordinates of the transformed dataset (called principal components) are orthogonal linear combinations of the original coordinates. The principal components are ordered in descending order by the amount of variance they explain in the data.

Often, most of the variance in the dataset can be explained by many fewer coordinates than in the original dataset, with the last principal coordinates often associated with noise components of the original data. Consequently, PCA is often used for compression of data or feature selection. PCA allows visualisation of datasets by plotting the first two or three principal components of the data. However, due to the fact that the principal components are linear combinations of the original dataset, PCA has the limitation that it can model only linear relationships in the data.

When applying PCA the dataset can be viewed as a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ where n is the number of data items each containing d attributes and the m–dimensional row vector x_i represents data item i. The principal components of the dataset are the eigenvectors of the covariance or correlation matrix of $\mathbf{X}$ ordered by decreasing value

of the associated eigenvalue. The data is transformed into the principal component space by projecting each data item x_i along the principal components.

2.3.4 Kernel Principal Component Analysis

Several approaches have been devised to extend PCA to recognise nonlinear relationships among data attributes. One approach is kernel PCA (Haykin, 1999) which transforms the dataset $\mathbf{X}$ into a feature space using a (nonlinear) kernel function κ before the PCA is done. Kernel PCA returns the principal components of the data items in the feature space. The input to kPCA is a Gram kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ which is a representation of the original dataset transformed with the kernel function. Each element k_{ij} of the kernel matrix can be viewed as a kind of similarity between data items x_i and x_j and is defined as

$$k_{ij} = \kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{2.3.3}$$

where x_i and x_j are the data items, $\phi(x_i)$ is the transformation of x_i into the "feature" space and $\langle \cdot, \cdot \rangle$ is the dot product operator. Generally it is not necessary to compute $\phi(x_i)$ explicitly. Instead, $\mathbf{K}$ is computed directly from the dataset. This is called the "kernel trick" and it means that the feature space can be very large without making generation of $\mathbf{K}$ inefficient. It also means that non–vectorial data types can be handled using special kernels such as string kernels (e.g. Leslie et al. (2004)). In kPCA the principal components are the eigenvectors of the kernel matrix.

The projection of low-dimensional data representation $\mathbf{Y}$ can be defined as where k is the kernel function.

$$Y = \{ \sum_j \alpha_1 \kappa\big(x_j, x\big), \sum_j \alpha_2 \kappa\big(x_j, x\big)\big)....., \sum_j \alpha_d \kappa\big(x_j, x\big)\big)\} \tag{2.3.4}$$

$\alpha_1 ... \alpha_d$ represents the number of iterations for each k.

2.3.5 Local Linear Embedding (LLE)

Local Linear Embedding (Roweis and Saul, 2000) computes data to a low dimensional representation by preserving neighbourhood embedding of high input data. It learns the global structure of nonlinear input data by exploiting the local symmetries of linear reconstruction of data. The local properties of data are determined by placing the data points as a linear combination of their nearest neighbours.

$$\phi(Y) = \sum_i (y_i - \sum_{j=1}^{k} w_{ij} y_{ij})^2 \tag{2.3.5}$$

In high dimensional space LLE creates a linear hyperplane between point y_i and its neighbours and constructs a matrix W where W_i represents reconstruction weight for y_i and its neighbours. Similarly, in low-dimensional space Y, y_i is reconstructed from its neighbours.

2.3.6 Stochastic Neighbour Embedding (SNE)

Stochastic Neighbour Embedding (Hinton and Roweis, 2002) is an iterative nonlinear technique that attempts to retain pairwise distances in a low-dimensional representation of data. It preserves the mainly local properties of the manifold, because similarities of nearby points contribute more to the cost function.

$$\phi(Y) = \sum_{ij} P_{ij} \log \frac{P_{ij}}{q_{ij}} \tag{2.3.6}$$

In the first step, a matrix P is created with P_{ij} being the probability of data points x_i and x_j calculated by the Gaussian kernel for all combinations of data points. Another matrix Q is created with q_{ij} representing the probabilities for the low-dimensional space data points y_i and y_j generated with the same Gaussian kernel. The Kullback-Leibler divergence can be used to measure the difference between P and Q and SNE tries to minimize this difference. The minimisation of Y is performed using the gradient descent method but it can be performed in different ways (Hinton and Roweis, 2002).

2.3.7 Diffusion Maps (DM)

Diffusion Maps (Coifman and Lafon, 2006) is another nonlinear dimensionality reduction technique where the proximity between the data points is obtained on the basis of a Markov Random Walk algorithm where a number of steps are performed to obtain the diffusion distance between points on X, the graph of data . Diffusion maps calculates diffusion distance based on many paths through a graph G, representing the data. This makes diffusion maps more robust to noise.

In first the step, a graph G of the data is constructed (Van der Maaten, 2007) and a matrix W is computed with the weight of the edges in G using the Gaussian kernel function. In a second step a matrix P is generated as a normalization of matrix W. In the normalisation, rows sum to 1. This P matrix is considered as a 'Markov Matrix' that defines the forward transition probability matrix of a dynamical process. Matrix P represents the probability of a transition for point x_i from high-dimensional space to low-dimensional space in a single time step. So, the diffusion distance using the random walk forward probabilities $P(t)$ can be defined as

$$D^{(t)}(x_i, x_j) = \sum_k \frac{(P_{ik}^{(t)} - P_{jk}^{(t)})^2}{\psi(x_k)^{(0)}} \tag{2.3.7}$$

Where the forward probability matrices for t timesteps $p(t)$ is given by P_{ij}.

2.3.8 Random forest

Random forest (RF) (Breiman, 2001) is a classifier based on generating a large number of random trees, and this randomness can be bootstrapping or bagging, in which trees are trained on subsets sampled from the original dataset. Random forest also provides an option to decide how deep each of the resulting trees is allowed to go, and then takes information from all the combined trees that lead to improving the accuracy of distribution. The process of combining trees is done through voting; a given data point runs through each decision tree in the forest, and the number of votes decides the classification of that point. Random forest trains each tree independently and keeps records for each point classification after the voting process.

Classification algorithms that provide a metric for feature importance are of great interest for feature selection and feature prioritising. Random forest has been widely applied to microarray data and genetic variation data (Meng et al. (2009) and Bureau et al. (2005)). Random forest has also shown to be more robust in the presence of noise and missing data (Bureau et al. (2005) and Strobl et al. (2008)).

In this work, the random forest approach is mainly applied as a feature selection method. The importance measure generated by the random forest method was chosen for measuring the importance of each SNP or gene expression (weighting), and selection of an appropriate set of SNPs and gene expressions (feature selection).

2.4 Summary

Biological data is available in many kinds such as gene-annotations, gene expression profiles and single nucleotide polymorphisms. The focus of this chapter was to provide background of ALL cancer, the microarray data types, data visualisation, dimensionality reduction methods and feature selection method used in this thesis. The SVD explained in this chapter will be used in Chapter 4 to find functional relationships between genes using ALL gene-annotation data, while LLE, SNE, PCA, kPCA and DM will be used in Chapter 5 to classify the ALL patients using gene-expression data. The random forest method will be used in Chapter 6 to identify the pathways related to ALL. In the next chapter the literature will be reviewed related to methods and data types explained in this chapter.

Chapter 3

Literature review

The focus of previous the chapter was to define different biological data types and the methods used in this thesis. Researchers in the past have used these methods (described in Section 2.3.1) to handle large and complex biological data to retrieve information for diagnosis and prognosis of diseases. The focus of this chapter is to review the literature related to finding functional relationships between genes, microarray based classification of patients and microarray analysis of metabolic pathways.

3.1 Finding functional relationships between genes

Previous works in functional visualisation of genes are widespread but this thesis will specifically focus on functional visualisation of genes using GO. The literature in this area can be divided into two main areas: defining similarity measures using GO annotations, and applying unsupervised methods to visualise the functional relationships between genes.

3.1.1 Defining similarity measures using GO annotations

Similarity measures for genes listed in GO are often based on the degree of overlap of specific terms, or of the distance between those terms in the ontology. Sheehan et al. (2008) describe several approaches for similarity measures between GO annotations including those based on sets, vectors, graphs and terms, and propose an algorithm that finds specific common ancestors between terms over the hierarchical GO structure. Term overlap measures (Mistry and Pavlidis, 2008) are made up of a set of all the annotations related to a gene and all the parent terms, compared between genes. Similarity measures can also be computed across the three sub-ontologies by calculating similarity within a sub–ontology and then finding inter–gene relationships across the three sub–ontologies (Sanfilippo et al., 2007).

Graph-based methods use the distance between terms in the ontology as a partial measure of similarity, and find clusters of genes with similar terms or functionality using the hierarchical GO structure (Lee et al., 2004). Similarly Popescu et al. (2004) use GO terms to extract a functional summary of gene clusters. They identify the most frequent terms by applying fuzzy methods to clusters of genes and produce a hierarchical clustering of genes that results in clusters labelled with the "most representative term" of the contained genes.

Another common approach is to use information-theoretic measures to assess similarity between GO terms. Richards et al. (2010) assess functional coherence of a gene set using both a graph–based similarity measure and an information content similarity measure. Speer et al. (2005) and Fröhlich, Speer, Poustka, and Beissbarth (Fröhlich et al.) take a kernel-based approach and cluster genes with an information–theoretic

kernel function to calculate the similarity between genes over the GO. The motivation behind this approach, as opposed to a distance measure over the GO graph, is to handle the variable branching and density of GO more effectively. They derive gene clusters by applying a dual k–means clustering algorithm.

Recent discussion about the need for varied and flexible similarity measures in ontologies(Couto and Pinto, 2013) has argued that genes with similar structure do not always have similar functions, and there is a need to apply similarity measures to fill the gap between structure and function. The "GO Semantic Similarity Tool" (GOssTO)(Caniza et al., 2014) includes six similarity measures including both term-to-term and graph-based measures. The "Adaptable Gene Ontology semantic similarity-based Functional analysis" (A-DaGO-Fun)(Mazandu et al., 2016) allows users to use multiple term-to-term similarity measures to find the similarities in gene products of GO, based on the assumption that similarity between genes, as represented by these measures, plays a significant role in determining the functional relationship between genes.

The literature discussed above shows that there is a need to apply similarity measures to find functional relationships between genes. However, just retrieving information related to genes does not provide a meaningful application. In this thesis, gene-annotation ALL data is used and features of those genes are extracted from GO. Similarity and visualisation methods are then applied to find functional relationships between genes.

3.1.2 Applying unsupervised methods to visualise functional relationship between genes

Tools for functional analysis of large gene lists can be divided into three categories(Huang et al., 2008): singular enrichment analysis, gene set enrichment analysis and modular enrichment analysis.

Many unsupervised methods are made available through a web interface for researchers. 'DAVID'(Alvord et al., 2007) is a highly-cited method which finds functional relationships between a set of genes using a series of statistical methods such as heuristic fuzzy multiple–linkage partitioning. Similarly, GeneTrail (Backes et al., 2007) helps in finding functional enrichments in gene and protein data sets by using two statistical methods: over–representation methods and gene set enrichment analysis. FuncAssociate (Berriz et al., 2009) is also a web based tool that performs gene set enrichment analysis using the hierarchical structure of GO. Warde-Farley et al. (2010) is a web-based tool to predict gene functionality based on adaptive network weighting methods. The visualisation of their methods is shown in network form, while the focus of this thesis is to apply some visualisation methods like singular value decomposition (SVD) to visualise the functional relationship between genes and their features at the same time. Annotation, Visualisation and Impact Analysis web server (AVIA v.2.0) (Vuong et al., 2015) integrates with DAVID (Jiao et al., 2012) and provides a web-based API to access DAVID's tools for functional analysis and clustering of genes. However, this work does not allow users to visualise genes and their related terms on the same visualisation.

Some authors have used the GO hierarchical structure for clustering of GO terms. Lee et al. (2005) proposed a novel ontology based clustering algorithm CLUGO. They

considered hierarchical characteristics and the clustering of term distributions in GO by identifying the distribution of significant groups. Shi et al. (2014) has applied random forest with GOstat (Beissbarth and Speed, 2004) on leukemia and prostate cancer datasets to find functional relationships between genes. They have also used Gene Ontology to extract GO terms related to those genes, but they did not provide visualisations where users can visualise gene and terms together across higher dimensions. Tan et al. (2015) introduced a tool called Constellation Map to functionally visualise large sets of genes. Constellation Map generates a radial plot where each node of the plot represents a significantly enriched gene set. The angular distance between nodes represents the similarity between them and the distance between the node to origin suggest the association to the phenotype of interest.

Reviews of GO-term based visualisation tools include Schroeder et al. (2013) which reviewed different available tools for visualisation of multidimensional genomic data. They categorised tools into three categories: genomic coordinates, heatmaps and network. In their review, they have suggested that there is still a need for new visualisation tools which provide better visualisation, are easier to use and that do not require any computational expertise by users. They also suggested that there is a need for new genomic visualisation tools for cancer data to help clinicians in diagnosis and prognosis. Similarly, Supek and Škunca (2016) typically categorize them into five types: interactive GO browsers that consist of tools that are interactively browsing entire GO, network visualisation tools that display graph-like visualisation, GO visual overlays that can visualise interesting subsets of the GO, semantic similarity analysis tools that provide semantic similarity between terms and emerging methods that consist of methods which display trend underlying a group of GO terms. These

reviews frequently suggest that there is a need for visualisation tools that can handle redundant GO terms, handle more than a single list of terms, and extract patterns between those terms.

Although these tools have provided meaningful results, biological processes are complex and cannot be measured based on term-term relationships or solely in two dimensions. The visualisations described above focus mainly on the first two or three dimensions of data. The research gap can be divided into two major issues: First, GO based visualisation tools do not provide visualisation where a gene and its related terms can be visualised together. Second, none of the methods above provide visualisation of higher dimensions of the data past the first two or three. To gain more understanding from biological data, it is valuable to examine the relationship between genes and GO terms, and higher dimensions of the data (as described in Section 2.3.2), with different similarity measures. This literature review shows that there is a need for a framework that finds the similarity between terms based on different similarity measures and then analyzes the patterns in terms. These patterns then lead to identifying the functional relationship between genes, and this relationship can be visualised on different axes. In Objective 1 of this thesis, this approach is used to find functional relationships between genes in an ALL gene-annotations dataset.

3.2 Microarray based classification of patients

Data analysis and visualisation of ALL patients can be divided into two sections (i) feature selection (ii) classification of ALL patients using dimensionality reduction methods.

3.2.1 Feature selection

Feature selection is the process of decreasing the number of attributes in a dataset in a principled way to improve the visualisation of the data points. This section will focus on literature review of feature selection, dimensionality reduction methods used in the biological domain, and different dimensionality reduction methods used to date for classification of ALL patients.

Many feature selection techniques have been proposed (Guyon and Elisseeff, 2003), but all can be grouped into either feature ranking methods, subset selection methods or feature construction methods.

Feature ranking methods rank individual features using a metric such as the Bayesian Information Criterion (Friedman et al., 2001) and select the highest ranked. Generally these methods make the simplifying assumption that features are independent. Features can be ranked based on their correlation to an attribute of interest, such as the class of data object, or in an unsupervised way based on intrinsic properties of the attribute such as its variance or entropy.

Subset selection methods identify groups of attributes as a set rather than individual attributes. Due to the fact that there are many combinations of features that could be selected, and that this number increases exponentially with the number of attributes, subset selection techniques require a heuristic search through the feature space with a metric for the subset. These methods do not necessarily require the assumption of independence of features, but the search is more computationally demanding than methods which do. The choice of optimisation algorithm to drive the heuristic search is broad and can include greedy searches, as well as methods such as genetic algorithms or simulated annealing.

3.2.2 Dimensionality reduction methods with biological data

Dimensionality reduction approaches summarise a large number of data attributes into a smaller set with less redundancy or no redundancy at all. Lee and Verleysen (2007) categorise dimensionality reduction approaches as shown in Figure 3.1. First, all approaches are divided into linear and nonlinear methods. Linear approaches include matrix decomposition methods such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Independent Component Analysis (ICA) (Skillicorn, 2007). Linear methods are suited to data with many attributes and/or few data points because they are simple, fast, do not fall into local optima and involve no parameters. Conversely, nonlinear methods (Lee and Verleysen, 2007), should only be used when there is enough data to set algorithm parameters. Nonlinear methods can be classified based on the quantity preserved in the lower dimensional data projection: distances between data points, or topology of the dataset. Methods preserving distance between points can use both local and global information about data, whereas topology-preserving methods use local information (neighbourhood relationships). Distance-preserving methods can be subdivided based on the measure used: Euclidean, geodesic or other. Euclidean distance-preserving methods include multidimensional scaling (MDS); geodesic distance-preserving methods include Isomap; and "other" distance-preserving methods including kernel PCA (kPCA) and semidefinite embedding (SDE). Topology-preserving methods divide into those using a predefined lattice (the structure to which is mapped) such as Self-Organising Maps (SOM) and those deriving the lattice from the data, such as Local Linear Embedding (LLE). this thesis will compare LLE, DM, PCA, kPCA and SNE for classification of ALL patients.

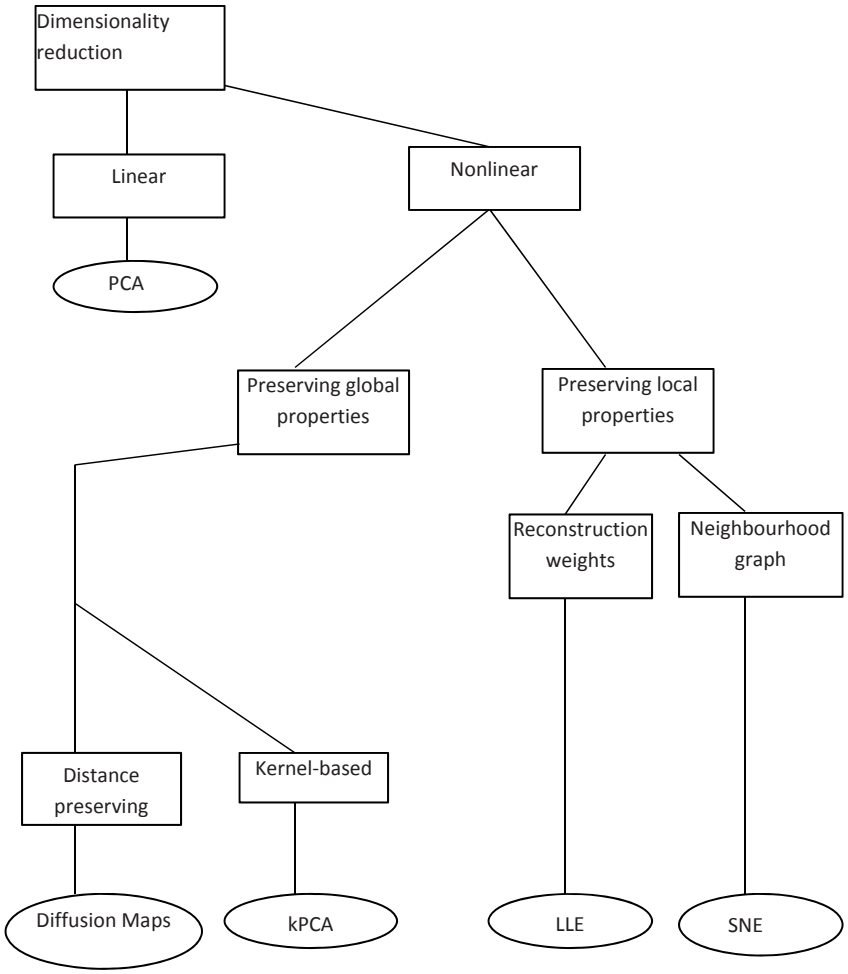


Figure 3.1: The figure shows the dimensionality reduction methods used in this thesis based on suggestions by Lee and Verleysen (2007). The rectangles represents categories and eclipses represents methods.

Nonlinear dimensionality reduction methods have been intensively used in the past; for example, DeMers and Cottrell (1993) introduced a novel network involving five layers; input, encoding layer, hidden layer, decoding layer and output layer. They successfully applied this method on different kinds of data such as: a closed 1-D manifold, time series data and 8-bit grey scale face images. Since then, the use of dimensionality reduction methods has increased in all kinds of data analysis such as human handwriting recognition (Tenenbaum et al., 2000), face recognition (Roweis and Saul, 2000), hyperspectral image processing (Kim and Finkel, 2003), motion data (Suo and Qian, 2010) and biological data for example Reutlinger and Schneider (2012) have recently used nonlinear dimensionality reduction methods such as Stochastic Neighbor Embedding, stochastic proximity embedding and self-organizing maps to map compound libraries for drug discovery.

In the past, linear and nonlinear dimensionality reduction methods have been applied to identification and classification of ALL, and microarray-based classifiers have been acknowledged as important for diagnosis of ALL (Bacher et al., 2010). A pioneering study for for leukaemia cancer diagnosis, prognosis or prediction based on microarray data (Golub et al., 1999) used a signal-to-noise statistic to select a small set of genes and developed a schema to identify and classify between ALL and AML. Following this work, many researchers have worked on ALL data for classification and diagnosis. Self-organizing maps (SOM) have been applied to underlie the pathogenesis of leukaemia (Ramaswamy et al., 2001), but the main focus of this work is clustering rather than classification.

Support vector machines have also been widely used for leukaemia classification. Yeoh et al. (2002) analyses the pattern of gene expression in leukaemia using PCA,

hierarchical clustering, discriminant analysis with variance (DAV), and self-organizing maps with classifiers like support vector machines (SVM). Similarly, Haferlach et al. (2005) used SVM and 10–fold cross validation to examine gene expression profiles in 937 bone marrow samples related to leukaemia subtypes. Given that their objective is classification, they argue that gene expression profiling can predict clinical sub-entities of leukaemia. Huang and Chang (2007) introduced a method called evolutionary support vector machines, which is a combination of automatic feature selection method, SVM classifier and k-fold cross validation methods. Li et al. (2009) also used Support Vector Machine–Recursive Feature Elimination (SVM-RFE) tool Bacher et al. (2010) worked on gene-expression profiling for diagnosis and sub-classification of Leukaemia. They used support vector machines to design a linear binary classifier with good accuracy. Chen and Huang (2010) applied multi-task support vector sample learning (MTSVSL) and Back Propagation Neural Network (BPNN) classifier techniques to classify leukaemia and prostate cancer. They performed k-fold cross validation and leave-one-out cross validation to validate their results, and found that for a leukaemia dataset the BPNN method had higher accuracy than MTSVSL method. However, this work only compared different types of SVM classifiers, and did not use dimensionality reduction methods on gene expression data, or different classifiers, which may have given them better results.

More recently Deeb et al. (2015) has applied SVM for classification of diffuse large B-cell lymphoma patients using protein expression profiles. However, the dataset only consists of 20 patients. While Zhao et al. (2016) also used SVM for classification of diffuse large B-cell lymphoma patients by eight gene expression profiles. They selected 8 genes from 414 patients treated with CHOP/R-CHOP chemotherapy and used

ROC for validation of results. However their work do not apply any dimensionality reduction methods.

Other than SVM, numerous researchers have worked on classification of leukemia cancer. Nilsson et al. (2004) used Isomap to identify the subtypes of lymphoma and lung cancer. Similarly, Zong et al. (2005) developed a tool to identify different types of white blood cell in a given blood sample. They used two approaches: a multidimensional space using artificial neural networks (ANN) and exploration of gene expression profiling of ALL to classify its six subtypes, ANN performed better on their ALL dataset. Tan et al. (2005) proposed a classifier called Total Principal Component Regression to classify human tumors such as ALL by extracting the latent variable structure and the errors in the microarray gene expression data. Wang and Gotoh (2009) built a classifier based on decision rules and rough sets to classify leukemia cancer. They argue that in order to build an accurate classifier for diagnosis of cancer, it is important to select important genes from a dataset. They build their classifier on the basis of single gene selection from a gene expression dataset. However, these classifiers are based on linear analysis. In this work feature selection is performed first and then non-linear dimensionality reduction methods are applied. This approach is used because the complexity of biological data means that non-linear methods uncover details better than linear methods in some cases.

Shyamala and Vijayakumar (2014) introduced a Modified Extreme learning Machine (MELM) for classification of gene expression data, performing testing on leukemia, lymphoma and small round blue cell tumors. This work claims to deal with local minima, improper learning rates and overfitting issues, but does not compare the published method with already existing methods.

Given the range of different methods, numerous comparison studies have been performed attempting to identify optimal methods for particular situations. Four examples are discussed below.

Shi and Luo (2010) compared three dimensionality reduction methods (PCA, LLE and Isomap) to cluster cancer tissue samples and found that LLE and Isomap perform better than PCA in classification and cluster validation. This work motivates the work done in this thesis to investigate the application of different linear and non-linear machine learning methods on ALL cancer data.

Orsenigo and Vercellis (2013) compared six dimensionality reduction methods: isometric feature mapping, locally linear embedding, Laplacian eigenmaps, Hessian eigenmaps, local tangent space alignment and maximum variance unfolding on different data, including leukaemia data. They found that LLE performed better than the other methods, while Isomap remained the second-best method on their datasets.

Musa (2014) assessed the performance of the principal component analysis, kernel principal component analysis and independent component analysis methods on Logistic regression (LR) and 1-regularized logistic regression. They compared the results over several datasets, including a leukaemia dataset. They found that PCA, kPCA and 1-regularized LR outperform ICA in some datasets, but kPCA was ineffective in some cases with larger datasets. The Musa (2014) study compared similar machine learning methods (based on component analysis), while the focus of this thesis will be to compare selective linear and nonlinear methods for classification of ALL.

Haghverdi et al. (2015) compared principal component analysis, t-distributed Stochastic Neighbour Embedding (t-SNE) and Diffusion Maps on qPCR data of mouse haematopoietic stem cells and RNA-Seq of human preimplantation embryo

datasets. They found that diffusion maps performed better than the other two methods, and suggested that there is a need to address this challenge of comparing different linear or nonlinear methods and determine the best method among them on specific dataset.

The literature above showed that researchers have been trying to find an optimal dimensionality reduction method for huge datasets and the issues raised by Wagholikar et al. (2012). Three specific deficiencies have been identified across the studies: with the tools for diagnosis of cancer.

1. Decision accuracy: There are very few applications that produce results with high accuracy, and so far there are not many applications that can match the diagnostic performance of human experts.

2. Usability: Many individuals using these tools are not expert computer users, and some of these tools are not user-friendly. If it is hard for clinicians to understand the interface, they may become frustrated and stop using the tool.

3. Explanation: Many tools don't provide enough supporting information about decisions. A tool will be more acceptable to clinicians if the tool provides some explanation of why it has suggested a particular diagnosis.

4. More recently, Nogueira and Brown (2016) have suggested that stability of feature selection methods play an important role in high-dimensional data.

The assembled literature shows that dimensionality reduction methods have not performed well on biological data due to its complexity. The amount of biological data is increasing on a daily basis, and biologists need tools which allow them to use

different data mining and visualisation methods so that they can compare results for improved diagnosis of cancer. This research gap leads to objective 2 of this thesis, which will be addressed in Chapter 5.

3.3 Microarray analysis of metabolic pathways

Schilling et al. (1999) has defined metabolic pathways as the set of chemical reactions occurring in a cell. The extensive map of pathways helps distribution and processing of metabolites. Metabolic pathways are involved in development of cells, cell division and growth. There are two types of metabolic processes:

Anabolic: Responsible for creating parts of cells through nutrition.

Catabolic: Responsible for obtaining energy by breaking down carbohydrates, lipids, proteins and nucleic acids.

Metabolic pathways play an important role in any organism. The data science literature for studying metabolic pathways using microarray data can be divided into two main areas: finding metabolic pathways associated to disease, and machine learning methods used for predicting metabolic pathways.

3.3.1 Finding metabolic pathways associated to disease

Microarray data is complex and requires preprocessing before any analysis. Many authors in this area have applied simple statistical methods such as z-score and p-value for pre-processing. The purpose of applying these methods is to find whether data is meaningful for pathway analysis. P-value based analysis tools include Association

LIst Go AnnoTatOR (Holmans et al., 2009) which defines significant SNPs by using a cut-off p-value and then counts significant genes in each pathway. Holmans et al. (2009) analyzed Wellcome Trust Case-Control Consortium Crohn's disease data and found pathways related to Crohn's disease. i-GSEA4GWAS (Zhang et al., 2010) performs SNP label permutation and assign SNPs to genes to calculate a modified Gene Set Enrichment Analysis (GSEA) enrichment score which leads to finding related pathways. i-GSEA4GWAS (Zhang et al., 2010) also uses a Manhattan plot of the gene set which helps users to compare association results of a given pathway. GSEA-SNP (Holden et al., 2008) calculates an enrichment score based on all SNPs in a given pathway without calculating gene-level test statistics. Instead they rank SNPs according to p-value and suggest that calculating p-values at a SNP level will show a higher degree of association to a binary phenotype than gene-level test statistics. They tested this algorithm on a cohort of patients with non-Hodgkin's lymphoma, finding biological relations to disease, but the study data contained only 1892 SNPs, which is a small sample size.

Gene Set-based Analysis of Polymorphisms (Medina et al., 2009) calculates an enrichment score by ranking the gene list and assigns the best SNP p-value to a gene. This analysis was applied to breast cancer data and found biological processes associated with risk of sporadic postmenopausal breast cancer. GSA-SNP (Nam et al., 2010) calculates the log of each p-value and then use the kth ($k = 1, 2, 3, 4$ or 5) best p-values in each gene to represent the gene which leads to symmetric distribution of the gene score. They compare the results with the best and second-best p-values to find the association of SNPs in a gene which helps to remove randomly-associated signals. One of the limitations with these methods is the arbitrary choice of cut-off

p-value. A standard cut-off value may not work with every dataset or type of disease. Another issue with these tools is that some authors select the minimum p-value in a gene as the representative p-value for a gene, which can eventually cause information loss. There is a need for a framework to perform SNP analysis without information loss. In this thesis, the feature selection method random forest is used to select the important SNPs and pathways related to them.

The literature other than p-value tests includes GenGen (Wang et al., 2007) assigns the best test statistic among SNPs in or near a gene to represent gene-level signals and then calculates a Kolmogorov-Smirnov-like enrichment score for a pathway analysis. They argue that pathway analysis can give more insights into diseases than individual gene analysis. The SNP ratio test (O'Dushlaine et al., 2009) calculates the number of significant SNPs in pathways using a p-value threshold and creates a small subset of the dataset. The authors applied this algorithm on a Parkinson's disease dataset and identified pathways, finding pathways related to the disease, but these small datasets can lead to loss of information and the relationship between different SNPs. GRASS (Chen et al., 2010) applied regularised regression to select representative 'eigen-SNPs' for each gene and then assess their joint association with disease risk. They tested their algorithm on colon cancer and identified the top enriched pathways.

3.3.2 Machine learning methods to predict metabolic pathways

Machine learning methods have been used intensively in past for predicting and finding pathways. The work related to predicting metabolic pathways includes Yamanishi

et al. (2005) applied individual kernels, or integrated kernels where the kernel matrix is a symmetric matrix where all the diagonal elements are 1. They compared kernel methods using area under curve (AUC) metrics and showed that kernel canonical correlation analysis (Akaho, 2001) performed better than other methods. Their comparison suggest that nonlinear kernel methods such as kPCA could be useful for metabolic pathway analysis. Dale et al. (2010) who used naive Bayes, decision trees and logistic regression, and compared the results with their previous work PathLogic (Paley and Karp, 2002) which finds pathways using enzyme information from databases. They constructed a dataset based on 5610 pathways and compared both algorithms. They found that machine learning-based methods performed slightly better than PathLogic, suggesting that applying machine learning-based methods on metabolic pathways can lead to better understanding of disease.

The literature finding pathways using microarray data includes Panteris et al. (2007) who used microarray data to find the behaviour of pathways. They selected a subset of genes of each pathway, and transformed them into pathway signatures specific to one disease or process. They used a dataset of *Escherichia coli* for this experiment and found active pathways related to it. Engreitz et al. (2010) used independent component analysis (ICA) to identify known pathways by analysing gene expression data in the preclinical anti-cancer drug parthenolide. Rapaport et al. (2007) integrate *a priori* knowledge of gene networks in the analysis of gene expression data. They used spectral decomposition of gene expression profiles with respect to eigenfunctions of the graph for classification. They performed experiments on a yeast dataset and investigated biological relevance through finding pathways. Curtis et al. (2005) reviewed the "Pathway Analysis" tool for microarray data analysis.

They have found that binomial distribution, z-score and gene set enrichment analysis performed well to identify mechanisms underlying diseases and adaptive physiological compensatory responses. Finding active pathways allows researchers to identify important pathways involved in a disease. However, there is no literature available related to feature selection of ALL SNPs using the random forest method and finding pathways related to ALL. One objective of this thesis is to find pathways involved in ALL using SNP dataset.

This literature shows that these tools simply display pathways that are associated to those genes and processes, while genetic variation data (SNPs) of a patients gathered in a a complex dataset with information related to a patient can be in thousands. These high numbers of SNPs for each patient makes the data hard to understand. There is a need to apply feature selection method to select the best SNPs and find pathways related to those SNPs. The focus of this thesis will be to find metabolic pathways that are associated to ALL using the random forest feature selection method. All the methods above do not address potential overfitting on large datasets (such as SNPs), and handling these large datasets requires the use of feature selection techniques.

A recent review by Jin et al. (2014) suggests the following challenges:

(i) A need for tools that can process gene expression and genetic variation data

(ii) A need of weighting scheme for SNP profiles which leads to finding highly informative genes and pathways.

(iii) Non-biased SNP selection, according to them, "permutation of SNPs, which is often used in P value-based approaches, can disrupt linkage disequilibrium

patterns between SNPs and may not generate the correct null distribution".

this thesis will address these challenges, and apply the random forest feature selection method to find high-weighted SNPs with independent and identical SNP sampling. this thesis uses kappa statistics for validation rather than P-value. Kappa statistics have already been used in biomedical studies (Sim and Wright, 2005).

Furthermore, Jones et al. (2008) suggest that despite the complex genomic landscape that is being revealed through cancer genome sequencing projects, the complexity can be summarised as causing defects in one or more of 12 well-defined signaling pathways. They performed genetic analysis of SNP probes related to 24 pancreatic cancers through next-generation sequencing-by-synthesis technologies. They suggested a thorough analysis is required to understand pathways related to any disease. this thesis will follow their lead to find metabolic pathways related to ALL.

Another key issue with metabolic pathway analysis is that many human genes are uncharacterized and not mapped in predicted pathway analysis, This is where our first objective (functional visualisation of genes) becomes useful to find functional relationships between genes and associate functionally related genes to find pathways. This forms Objective 3 of the thesis which will be addressed in Chapter 6.

3.4 Summary of research gap

This literature review has shown a significant research gap for functional visualisation of genes, visualisation of ALL patients and finding metabolic pathways. These gaps fall into three areas:

(i) Although there are many tools available related to functional visualisation of

genes (as described in Section 3.1.2), they are limited to visualising a few dimensions of data. In this thesis the focus will be to analyse higher dimensions of resultant eigenvectors of SVD. Most of the work in this area has also concentrated on one type of similarity measure. In this thesis, comparison is made between two similarity measures: 'hop-based' (Lord et al., 2003) and information content-based similarity measures (Fröhlich et al., 2006), compared with with Pearson correlation.

(ii) Researchers have recently compared different machine learning methods on microarray datasets (as described in Section 3.2.2), but gene expression data is high-dimensional data with many expression values for each patient and high variance between patients which makes classification difficult. The gene expression data for acute leukemia cancer is also high dimensional and exhibits high variance which makes it hard to classify patients between relapsed, patients who relapsed after treatment, and non-relapsed patients. There is a lack of work in comparing different visualisation methods. This thesis compares some linear and nonlinear machine learning methods based on area under the curve (AUC) metrics, and identifies the best methods for classification of ALL patients.

(iii) The literature on identifying pathways related to acute leukemia is very short. Recent work (described in Section 3.3) has shown that authors have applied machine learning methods to find metabolic pathways in different diseases. There is a need to apply feature selection method on SNP datasets to select important SNPs. There are no published studies using the random forest method on SNP data to identify pathways associated with ALL diagnosis and progression. The focus of this work is to apply random forest method on ALL SNPs dataset to

identify important SNPs and highlight pathways related to ALL.

Chapter 4

Finding functional relationships between genes

4.1 Introduction

Advancements in DNA microarray technology have made progress towards new insights in cancer treatment, which require a methodical analysis of many genes. The routine use of microarray-based high-throughput technologies have already generated a massive amount of data about gene functionality, and gene expression patterns. High-throughput microarray technology is a mechanism to simultaneously measure the activity level of thousands of genes in a biological sample. Conducting microarray experiments for all patients in a cohort generates datasets of potentially hundreds of patients, and tens of thousands of genes, where each entry is a positive number indicating the activity of a particular gene in that sample. Researchers in the past have selected subsets of genes related to a targeted class, for instance, if they want to know whether the selected sample carries any disease or not by using methods such as random forest. The outcome may produce a number of lists of genes of interest, where each entry in that list is the name of the gene. However, it is quite difficult to

conclude the specific functionality of genes from these lists, as there are hundreds of such lists, and the fact that genes do not have a one-to-one mapping to phenotype i.e. genes highlighted by some experiments in one area of biology may have been discovered and annotated in another area. Consequently, the gene name on its own may not assist in understanding the function of the gene. For this reason, researchers have investigated ways of making sense of lists of genes by augmenting or enriching the data with functional information from databases such as the Gene Ontology (described in Section 2.2.3). This work directly contributes to help biologists identify possible functional relationships between genes which will eventually help them in diagnosis and classification of disease.

The focus of this chapter is to find functional visualisations of genes using singular value decomposition. The motivation for applying SVD compared to other dimensionality reduction methods such as Principal Component Analysis (PCA) is that genes and terms may be visualised on the same graph. This allows improved understanding of the biological function of genes. The approach is applied to two data sets: a data set used to validate the approach composed of genes selected from the KEGG database (described in Section 2.2.3) and a data set of genes highlighted from biological experiments in childhood cancer. This approach differs from the work described in Section 3.1.2 by recognising that functionality needs to be described over several 'axes'. Rather than looking at only two or three functional dimensions, this thesis find that it is valuable to also examine later dimensions that describe more subtle functional similarities between genes. Furthermore, this thesis compares a hop-based similarity measure of the set of approaches using shared common ancestors to measure similarity, to an information-theoretic similarity measure (Fröhlich et al., 2006)

using Pearson correlation.

4.2 Experimental Design

This section describes the datasets used in this thesis for the purpose of discovering functional information using SVD as shown in Figure 4.5.

4.2.1 Data Sets

Two datasets were interrogated in the present study: a validation set of genes selected from known classes and a data set of genes identified from an experiment in the cancer domain. In both the cases, the datasets consist of a list of gene names that are annotated with associated terms from the Gene Ontology.

KEGG data set

A set of genes has been selected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Section 2.2.3), which includes a functional classification of genes independent of GO. The rationale is to validate our approach with genes of known functional similarity. KEGG links genomes to their biological systems and is a series of interconnected databases that interrelate (i) genes and proteins, (ii) chemical building blocks, (iii) molecular interaction pathways and, (iv) hierarchies of biological objects. The last of these, KEGG BRITE, links genes into a functional hierarchy called the KEGG Orthology (KO). This hierarchy is different from the GO and has been constructed independently. The work in this chapter will validate the proposed approach by extracting genes from classes based on their KO terms and visualise them

using GO terms. The KEGG dataset (Section 4.1) contains genes (also in GO) from five KO classes: ribosome (ko03010, class 1), RNA polymerase (ko03020, class 2), transcription (ko01210, class 3), pentose phosphate pathway (ko00030, class 4) and pentose and glucoronate inter-conversions (ko00040, class 5). The expected result should see genes in classes 1, 2 and 3 would be identical, because all three classes are related to RNA activities, whereas, genes in classes 4 and 5 would be similar to one another, because they are both related to pentose, but different to the other classes.

The selected KEGG data set consists of a matrix of 67 rows, one for each gene, and 286 columns, one for each GO term. The set of 286 GO terms is the union of all GO terms directly associated with the genes. Each entry in the matrix is 1 if the gene is directly associated with the term, 0 otherwise. Figure 4.1 shows the number of terms associated with each gene in the KEGG dataset. Figure 4.2 shows the frequency of terms having direct association to various numbers of genes in the KEGG dataset. This shows that almost all terms are directly associated with very few genes and motivates our use of the proximity measures to relate terms using their relationships over the ontology.

Acute Lymphoblastic Leukaemia data set

Acute Lymphoblastic Leukaemia (ALL) is the most common childhood malignancy with around 250 children in Australia diagnosed annually. Microarray technology has been used extensively to identify the markers that are predictive of treatment outcome in ALL. The cancer dataset lists genes that are recognised as important in ALL.

The ALL dataset used in this thesis builds on previous work by Flotho et al. (2007)

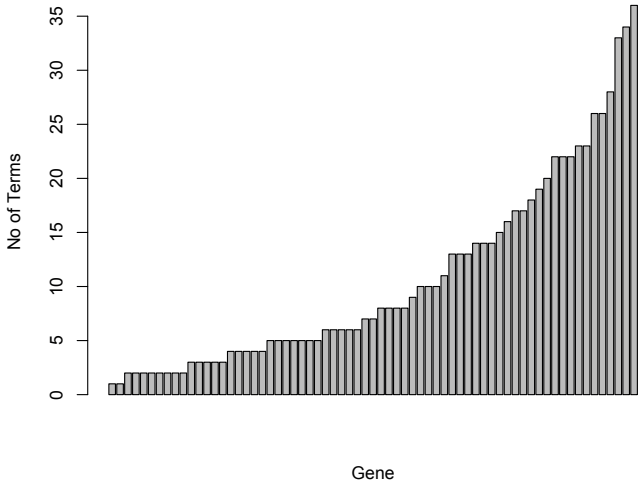


Figure 4.1: The number of terms associated with each gene in the KEGG dataset. Genes are ordered in increasing number of terms.

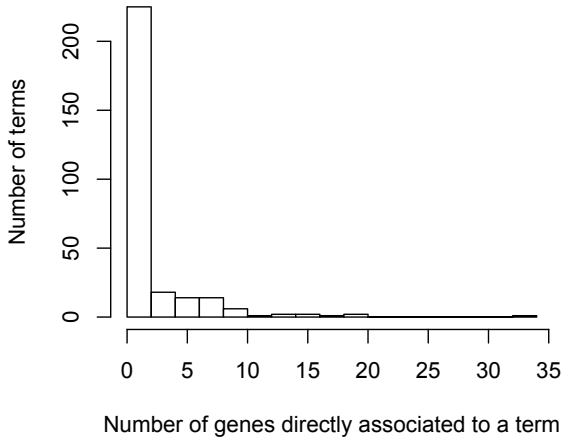


Figure 4.2: The frequency of terms having a direct association to various numbers of genes in the KEGG dataset.

Table 4.1: Genes in the KEGG dataset listed by class identifier. Column 1: class number Column 2: KO terms describing class and associated genes.

Class	**KO structure and list of genes used**
1	genetic information processing : translation : **ribosome** *rpsA*, *rpsB*, *rpsC*, *rpsD*, *rpsE*, *rpsF*, *rplB*, *rplC*, *rplD*, *rplE*, *rplF*, *RPS21*, *RPS23*, *RPS24*, *RPS25*, *rpmB*, *rpmC*, *rpmD*, *rpmE*, *rpmF*
2	genetic information processing : transcription : **RNA polymerase** *FLIA*, *RPOA*, *RPOB*, *RPOZ*, *RPOH*, *RPON*, *RPOD*, *RPB2*, *RPB1*, *RPB3*, *RPA49*, *RPA14*, *RPA34*, *RPA43*, *RPA12*, *RPC19*, *RPC25*, *RPB7*, *RPB4*
3	genetic information processing : **transcription** *GREA*, *GREB*, *NUSA*, *NUSB*, *NUSG*, *MBF1*, *Rcl1*, *RHO*, *ELP3*, *POLRMT*, *gtf2a2*
4	metabolism : carbohydrate metabolism : **pentose phosphate pathway** *pgl*, *zwf*, *edd*, *rpe*, *tktA*, *fbp*, *rpiA*, *gcd*, *rbsK*, *pgm*, *eda*
5	metabolism : carbohydrate metabolism : **pentose and glucoronate interconversions** *GUSB*, *galU*, *rpe*, *AKR1*, *mtlY*, *mtlD*, *clpX*

and Catchpoole et al. (2008). Flotho and colleagues (Flotho et al., 2007) identified a fourteen-gene signature with expression values able to separate a cohort of ALL patients into two groups that agreed with minimal residual disease (MRD) results. Minimal residual disease refers to small numbers of cancerous cells remaining after treatment (in the order of one cancerous cell in a million normal cells). It is used in oncology to know when a cancer has been eliminated and to compare therapies.

Catchpoole and coworkers examined these genes in a different cohort of ALL patients and also discovered a method for the separation of patients (Catchpoole et al., 2008). The data mining algorithm Random Forest (RF, described in Section 2.3.8), was used to identify other genes that supported the same separation of patients

as achieved by the Flothos gene signature. Patients were clustered using hierarchical clustering based on the gene signature from Flotho et al. (2007). The resulting two clusters formed the class labels for constructing our predictive model. A RF model of 50,000 trees was constructed on a gene expression dataset of 127 ALL patients and 22,280 probe sets which was generated using Affymetrix Human Genome microarrays on diagnostic bone marrow samples. Using this RF model, the 250 probe sets with the largest mean decrease in Gini index (effectively the 250 probe sets that contribute most to the RF model differentiating between the patients in the two clusters formed using Flothos signature) were selected to form the ALL gene list cancer dataset. Since some of these probe sets referred to the same gene, the dataset ended up with 195 unique genes. These 195 gene will be used as the ALL dataset for this thesis.

These 195 gene were then paired with their associated GO terms, giving a total cancer dataset of matrix of 195 rows, one for each gene, and 980 columns, one for each GO term. As aforementioned, the set of GO terms is the union of all of the terms directly associated with the genes. Entries in the matrix are 1 if the gene is directly associated with the term or 0 if not. Figure 4.3 shows the distribution of the number of terms of the genes in the cancer dataset. Figure 4.4(a) shows the frequency of terms having direct association with various numbers of genes in the cancer dataset. As mentioned earlier, almost all terms are directly associated with very few genes.

4.2.2 Incorporating functional information into the SVD

As shown in Figure 4.5, given a set of genes G, define T as the set of GO terms directly associated with any of the genes. From G we create a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ where n is the number of genes $|G|$ and m the number of GO terms $|T|$. Each element x_{ij} of $\mathbf{X}$

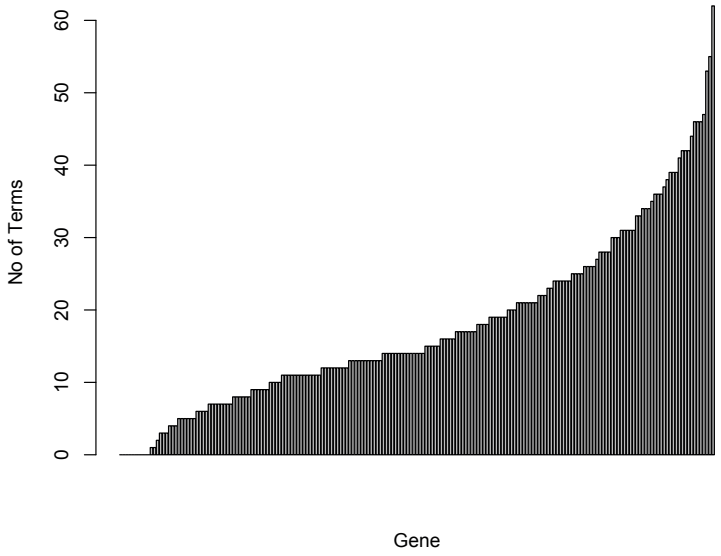
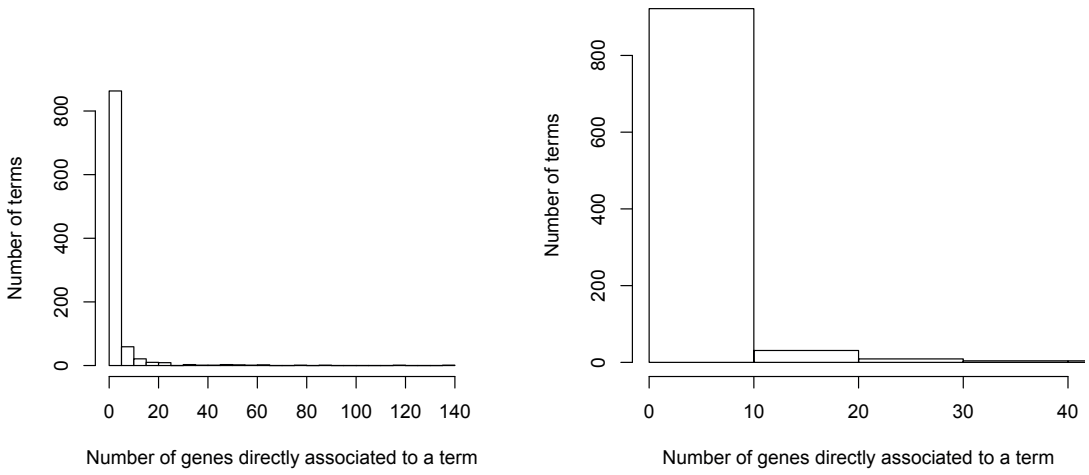


Figure 4.3: Distribution of the number of terms for genes in the cancer dataset. Genes are ordered by increasing number of terms.

has the value 1 if the gene i is directly associated with term j otherwise 0. This is similar to computational linguistics where "genes" are replaced by "documents".

This data matrix is augmented by information reflecting inter–term similarities. A symmetric proximity matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ is created with elements $0 \leq p_{ij} \leq 1$ representing the proximity (or similarity) between GO terms i and j. this thesis has explored two approaches for calculating values in the proximity matrix: a hop-based approach and an information-content based approach.



(a) (b)

Figure 4.4: (a) is showing frequency of terms having a direct association to various numbers of genes in the cancer dataset while figure (b) is representing top 40 genes with direct association to high frequency of terms in the cancer dataset.

Hop-based approach

The proximity between GO terms is based on the number of links (or distance) between them and is defined as $p_{ij} = (d_{ij} + 1)^{-1}$ where d_{ij} is the minimum distance between terms i and j over the hierarchy using "is-a" links which are more frequent than "part-of" relationships, extracted from GO using SQL. Elements of P are 0 P_{ij} 1. Terms i and j having a close relationship will have P_{ij} with a value near 1. The diagonal elements of P are P_{ii}= 1.

Information-content approach

The information content (IC) based proximity (Fröhlich et al., 2006) uses information content theory (Resnik, 1995) to calculate the semantic similarity between GO terms. It is based on the probability of GO terms in the gene dataset $\mathbf{X}$. The information content measure is defined as

$$IC(m) = -\log_2 P(m) \tag{4.2.1}$$

where $P(m)$ is the probability of term m in the data matrix and is calculated as $P(m) = \text{freq}(m)/N$ where N is the total number of GO terms in $\mathbf{X}$ and $\text{freq}(m)$ is the number of occurrences of m or any of its child terms. Similarity between terms i and j is defined as

$$p_{ij} = -\log_2 \min_{\hat{m} \in Q_a(i,j)} P(\hat{m}) = -\log_2 P_{ms}(i,j) \tag{4.2.2}$$

where $Q_a(i,j)$ is a function returning the set of common shared parent terms between terms i and j and P_{ms}, the probability of the minimum subsumer (Lord et al., 2003), is the minimum $P(\hat{m})$ if there is more than one parent. Values of p_{ij} using this scheme are not limited to the range $[0, 1]$.

The augmented data matrix is defined as $\mathbf{X}' = \mathbf{XP}$ with $\mathbf{P}$ being the term-to-term proximity matrix computed using either the hop-based or information-content based approach. SVD is applied to $\mathbf{X}'$ after centering and normalisation.

this thesis performed the visualisation of the genes (rows of $\mathbf{X}'$) by plotting rows of $\mathbf{U}$ using various columns. Similarly, the GO terms visualised (columns of $\mathbf{X}'$) on the same graph as the genes by plotting rows of $\mathbf{V}$ using the same columns as for $\mathbf{U}$. The first column is the projection of the data into the axis of most variation of the data. This is called the projection of the data into the first principal component

(PC1). The second column represents the projection into the axis related to the next largest amount of variation. This is the projection of the data into the second principal component (PC2). A similar scheme applies for the other columns of **U** and **V**.

In the final step, Pearson correlation was calculated between each column of **U** (the data projected to a principal component) and columns of **X** representing specific GO terms, as well as to a new column containing the number of GO terms associated with each gene (i.e. the sum of each row of **X**). Columns of **X** with high absolute correlation to a particular column of **U** are GO terms that explain the meaning of

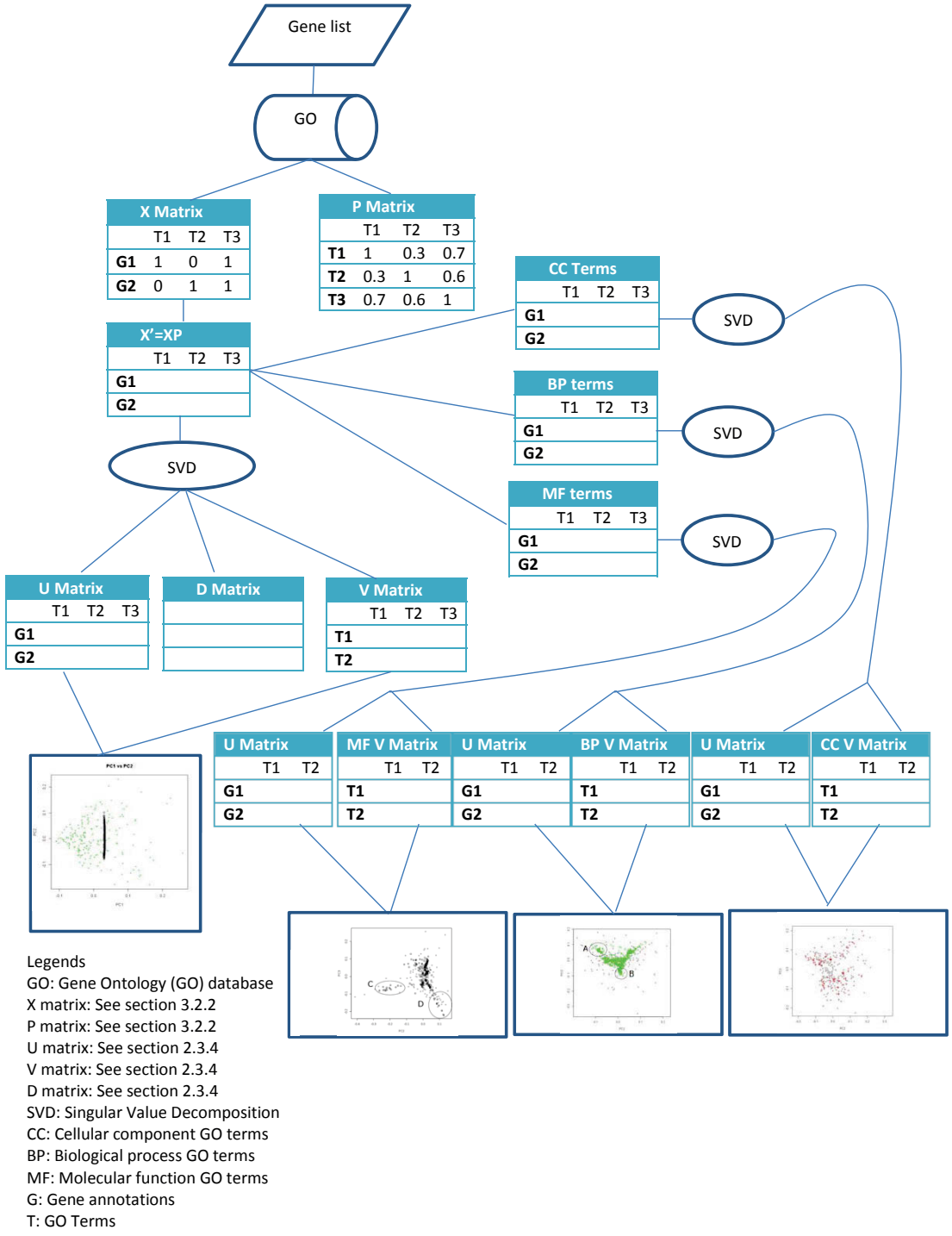


Figure 4.5: Experimental design for finding functional relationship between genes.

the respective principal component. The framework of this experiment is presented in Figure 4.5.

4.3 Results and Discussion

Firstly, the distribution pattern of the proximity matrices are presented. Figure 4.6 shows box plots for the values in the upper triangular section of the hop-based and IC proximity matrices for the datasets.

Next, the study explored the KEGG dataset with and without using a proximity matrix, to show that the proximity matrix is required for visualisation. Figure 4.7(a) shows genes projected into PC1 and PC2 without using a proximity matrix. That is, simply applying SVD to the $\mathbf{X}$ matrix. The plot revealed that because similar terms cannot be correctly determined, then, apart from a few genes, all genes grouped together. This is in contrast with the clear spread of genes when using proximity matrices as shown in Figures 4.7(b) and 4.8(a).

After transformation of the KEGG dataset with SVD, the Pearson correlation was calculated between the data projected to principal components and to the association of GO terms to genes (i.e., $\mathbf{X}$) with both similarity measures, the total number of GO terms for each gene and to the gene class. The strongest relationship expected in the KEGG data set is that comparing genetic information processing genes (those in classes 1, 2 and 3) with carbohydrate metabolism genes (those in classes 4 and 5).

There was a very strong correlation of 0.995 between the data projected into principal component 1 (denoted as PC1 in this chapter) and the number of terms associated with each gene suggesting that this principal component is a "size" component (Jolliffe, 2004). It seems reasonable that the most variation in the dataset

is based on the number of terms for genes. When plotted, as in Figure 4.8(a), PC1 shows variation in the genes, but not much in the terms. This is due to the relative amounts of variation in genes compared to terms. When the terms are plotted on their own as in Figure 4.9 the relationships are clearer. Figure 4.9(a) shows that PC1 does not separate based on the sub-ontologies for the hop-based measure (although PC2 does to a small extent), but there is a separation for the information–content measure (Fig. 4.9(b)).

Principal component 2, associated with the next largest variance, generally contrasts the genetic information processing genes with the carbohydrate metabolism genes as can be seen in Figure 4.10(a) with the hop–based similarity measure, where PC2 denotes the axis for principal component 2. However, it is not a completely clear division, and there is some overlap. The outlier (circled) with high PC2 and PC3 values is the gene *RHO* which is associated with the largest number of terms in the data. Table 4.2 shows that the highest correlation to PC2 is with the class label, which represents the expected differences between genes in this validation dataset, followed by strong positive correlations to GO terms describing carbohydrate metabolism and negative correlations to terms associated with ribosomes. The IC method separates genetic information processing and carbohydrate metabolism genes only at PC5 (Figure 4.10(b)) demonstrating that both methods eventually find the expected functional relationship between genes, but that they are clearer with the hop–based approach.

Apart from the outlier *RHO* in the top right hand corner, Figure 4.11 shows that PCs 2 and 4 separate the different kinds of genetic information processing genes as expected because there are more of these than the carbohydrate processing genes. Again, the separation involves some overlap between the classes. this thesis does

not present PCs 5 and greater for the hop-based approach because the expected structure has been explained using PCs 2–4. Correlations for PCs 2 through 4 are given in Table 4.3 showing a mixture of some expected (e.g. for PC3) and unexpected GO terms (e.g. those for PC2 and 4).

Table 4.2: GO term name and accession for terms with Pearson correlation > 0.5 to PC2 values for KEGG data with hop–based similarity measure. "Class" refers to the class identifier for the gene.

Term name and accession	**Correlation**
Class	0.55
Carbon utilization by utilization of organic compounds (GO:0015978)	0.54
Cellular catabolic process (GO:0044248)	0.54
Ribosome (GO:0005840)	-0.63
Ribonucleoprotein complex (GO:0030529)	-0.63
Intracellular (GO:0005622)	-0.60
Structural constituent of ribosome (GO:0003735)	-0.60
Translation (GO:0006412)	-0.60
Cytosolic small ribosomal subunit sensu Eukaryota (GO:0005843)	-0.58

4.3.1 Visualising cancer data set

As with the KEGG dataset, the distribution of the proximity matrices for the cancer dataset was examined first. Figure 4.6 shows box plots for the values in the upper triangular section of each proximity matrix. Next, similarly to the KEGG data, the cancer dataset was explored with and without using a proximity matrix, to show that the proximity matrix is required for visualisation. Figure 4.12 shows genes projected into PC1 and PC2 without using a proximity matrix (top) and with using the IC similarity measure (bottom). That is, simply applying SVD to the $\mathbf{X}$ matrix. The figure shows that, because similar terms cannot be ascertained correctly, genes do not cluster meaningfully.

Table 4.3: GO term name and accession number for those terms with absolute value of Pearson correlation > 0.25 for PC2–PC4 values for the KEGG data set with information content similarity measure.

Term name and accession	**Correlation**
PC2	
GO:0007165 (signal transduction)	0.45
GO:0006955 (immune response)	0.45
GO:0005102 (receptor binding)	0.45
GO:0005164 (tumor necrosis factor receptor binding)	0.45
GO:0008675 (2-dehydro-3-deoxy-phosphogluconate aldolase activity)	0.45
GO:0005576 (extracellular region)	0.45
PC3	
GO:0005886 (plasma membrane)	0.47
GO:0005624 (membrane fraction)	0.47
GO:0005622 (intracellular)	0.46
GO:0030529 (ribonucleoprotein complex)	0.43
GO:0005840 (ribosome)	0.43
GO:0006414 (translational elongation)	0.42
PC4	
GO:0016021 (integral to membrane)	-0.68
GO:0007165 (signal transduction)	-0.62
GO:0006955 (immune response)	-0.62
GO:0005102 (receptor binding)	-0.62

As with the KEGG dataset, the hop-based and IC approaches both show a strong correlation (0.997 and 0.988 respectively) between the number of GO terms and projected points to PC1. Overall, the distributions of GO terms for PC2 and PC3 (or PC1) make three clusters, associated with each GO sub-ontology (see Figure 4.13) as evidenced by the distribution of intra- and inter-cluster distances shown in Figure 4.14. To untangle the relationships across the sub-ontologies, SVD was applied to the terms from each sub-ontology separately. The found clusters were examined through correlation and by listing the terms in each cluster.

For the Cellular Component GO terms, the hop-based approach revealed a separation between cytoplasmic structure terms and DNA replication terms along the PC3 axis as shown in Figure 4.15(a). It also highlighted a cluster of terms associated with the membrane on the negative end of PC2 and a cluster of tubulin and kinesin GO terms towards the positive end of PC3 (see Figure 4.15(a) clusters A and B respectively). PC2 in the IC approach reveals four small distinct clusters: a cluster of membrane and extracellular-matrix-related terms, a cluster of terms associated to organelles, protein-complex-related terms and a cluster of cell-division-apparatus-related terms as shown in Figure 4.15(b) as clusters A, B, C and D respectively. Some of the terms in these clusters for the IC measure are listed in Table 4.4.

For the Biological Process GO terms, PC3 in the hop-based approach reveals a cluster of terms associated with development (e.g. embryonic development, notochord development, forebrain development, embryonic axis specification) as shown in Figure 4.16(a). PC2 in the IC approach identifies five tight clusters (shown in Figure 4.16(b)): cluster A relates to morphogenesis and early development, cluster B to homeostasis and response to stimulus (within which is a subgroup related to molecular transport in the cell), cluster C relates to gene expression regulation and metabolism, cluster D to differentiation and cluster E to DNA metabolism and function along with a number of small subgroups e.g vesicle transport. As before, some of the terms found in these clusters (IC measure) are listed in Table 4.5.

For the Molecular Function terms, PC2 in the hop-based approach identifies a cluster of terms associated with DNA helicase activity. In close proximity to this cluster is a loosely packed cluster of six genes that code for mini chromosome maintenance proteins (MCM2, MCM3, MCM4, MCM5, MCM6 and MCM7) as shown

in Figure 4.17(a). Both MCM proteins and replicative helicase play integral roles in eukaryotic DNA replication. The IC approach also identified this loose cluster of MCM genes and the GO term for DNA helicase activity. Across PC2, the rest of the clusters relate to enzyme activity No. 1, enzyme activity No. 2 and non-enzymatic molecular interactions as shown in Figure 4.17(b) as A, B and C respectively. Some terms from these clusters (IC measure) are in Table 4.6.

A summary of the GO term clusters shown in Figures 4.15(b), 4.16(b) and 4.17(b) is presented in Table 4.7 using IC method for Cellular Components (CC), Biological Process (BP) and Molecular Function (MF) of GO based on correlation results with their biological interpretation.

SVD visualisation of the cancer data results in a meaningful functional visualisation of the genes, particularly when limited to terms in sub–ontologies. Clusters of terms highlight functional groupings of genes, and the genes themselves cluster "behind" the terms that describe them. Correlations describe the PC axes. Each PC describes a different functional aspect of the gene set.

4.4 Summary

In this chapter, singular value decomposition has been applied to the lists of genes augmented with GO terms and inter–term similarities. Two datasets were visualised: validation data from KEGG, and a set of genes identified experimentally in Section 4.2.1. Results showed that principal component 1 measured the number of terms associated with genes. Later principal components(PCs) allowed visualisation of genes according to their functional information, but the meaning of PCs varied depending on the underlying genes. For the KEGG data, PCs described gene functionality.

Table 4.4: GO terms from the cellular component sub-ontology with absolute value of Pearson correlation > 0.35 for PC1–4 values from the cancer data set for the IC measure.

PC	**GO term name and accession**	**Correlation**
1	Number of terms	0.86
2	GO:0000777 (condensed chromosome kinetochore)	0.54
	GO:0000775 (chromosome, centromeric region)	0.50
	GO:0000776 (kinetochore)	0.46
	GO:0000778 (condensed nuclear chromosome kinetochore)	0.42
3	GO:0005856 (cytoskeleton)	0.46
	GO:0005874 (microtubule)	0.41
	GO:0005819 (spindle)	0.39
	GO:0031298 (replication fork protection complex)	-0.38
	GO:0042555 (MCM complex)	-0.36
4	GO:0005737 (cytoplasm)	0.38
	GO:0005730 (nucleolus)	0.37
	GO:0005634 (nucleus)	0.36

For the larger cancer dataset, the early PCs simply identified known hierarchies. Separate visualisation using terms from the individual sub-ontologies was more informative. The correlation between GO terms and PCs improved understanding of the functional meaning of the PCs. These results show that our approach can bring meaningful biological interpretation to gene lists. Users should not expect that the meaning of the PCs should generalise from one gene list to another, apart from gross patterns such as the sub-ontologies. This is because different sets of GO terms will be associated with lists of genes and SVD will focus on those that explain the most variance. In practice, our approach should be applied to specific gene lists of interest to explore only the functional characteristics of those genes.

The results in this chapter of finding functional relationships between genes using different similarity measures and visualisation address Objective 1 in Section 1.1 and

Table 4.5: GO terms from the biological process sub-ontology with absolute value of Pearson correlation > 0.35 for PC1–4 values for the cancer data set for the IC measure.

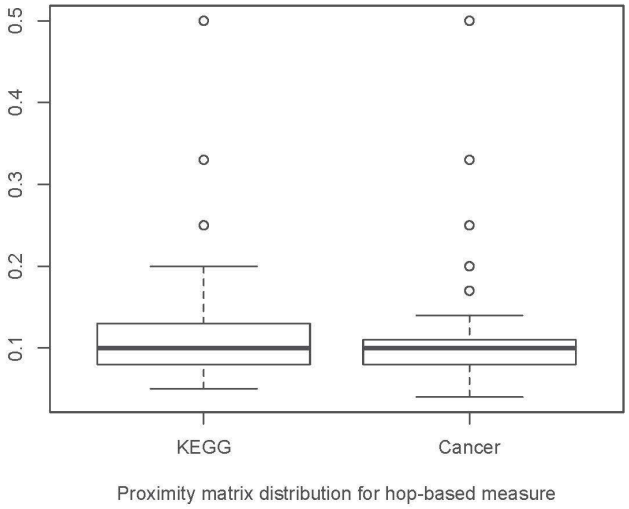
PC	**GO term name and accession**	**Correlation**
1	Number of terms	0.95
2	GO:0007067 (mitosis)	0.67
	GO:0051301 (cell division)	0.66
	GO:0007049 (cell cycle)	0.44
	GO:0006260 (DNA replication)	-0.45
3	GO:0009790 (embryonic development)	-0.35
4	GO:0006281 (DNA repair)	0.59
	GO:0006974 (response to DNA damage stimulus)	0.44
	GO:0000724 (double-strand break repair)	0.39
	GO:0006350 (transcription)	-0.49
	GO:0045449 (regulation of transcription)	-0.49

Thesis Contribution 1 in Section 1.4. In this chapter, the focus was on list of gene annotations. In the next chapter, the focus will be on how to handle ALL gene expression data and in Chapter 6 a case study will be conducted on ALL SNP data.

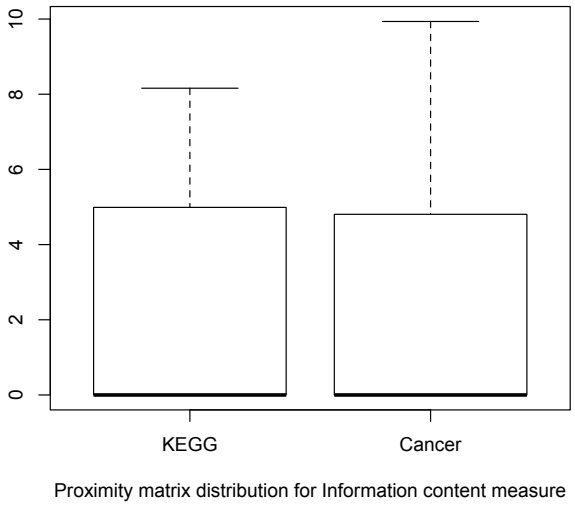
Note: Work in this chapter has been published as: H. Ghous, N. Ho, D. R. Catchpoole, and P. J. Kennedy. Comparing functional visualisations of genes. In J. Maria Pena and F. Famili, editors, 5th Workshop on Data Mining in Functional Genomics and Proteomics: Current Trends and Future Directions, European Conferencce on Machine Learning and Principals and Practice of Knowledge Discovery in Databases 2011, pages 12-21. 5-9 September 2011.

Table 4.6: GO terms from the molecular function sub-ontology with absolute value of Pearson correlation > 0.35 for PC1–4 values for the cancer data set for the IC measure.

PC	**GO term name and accession**	**Correlation**
1	Number of terms	-0.87
2	GO:0043140 (ATP-dependent 3'-5' DNA helicase activity)	-0.60
	GO:0003678 (DNA helicase activity)	-0.58
	GO:0004003 (ATP-dependent DNA helicase activity)	-0.57
	GO:0009378 (four-way junction helicase activity)	-0.53
	GO:0003697 (single-stranded DNA binding)	-0.56
3	GO:0016301 (kinase activity)	-0.56
	GO:0004672 (protein kinase activity)	-0.53
	GO:0004674 (threonine kinase activity)	-0.57
4	GO:0004518 (nuclease activity)	0.67
	GO:0004527 (exonuclease activity)	0.65
	GO:0004523 (ribonuclease H activity)	0.59
	GO:0008409 (5'-3' exonuclease activity)	0.56

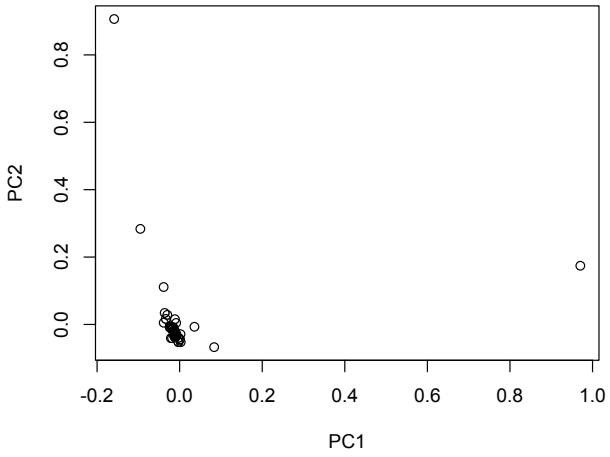


(a)

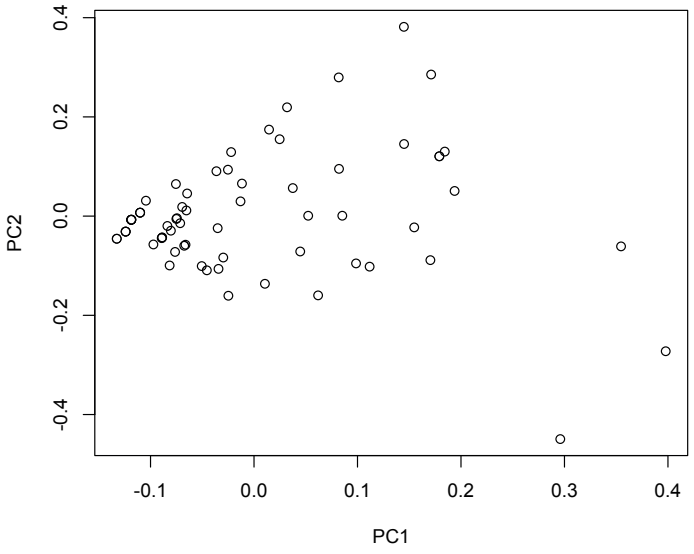


(b)

Figure 4.6: Distribution of values in the proximity matrices. (a) hop–based proximity matrix (b) information-content proximity matrix.

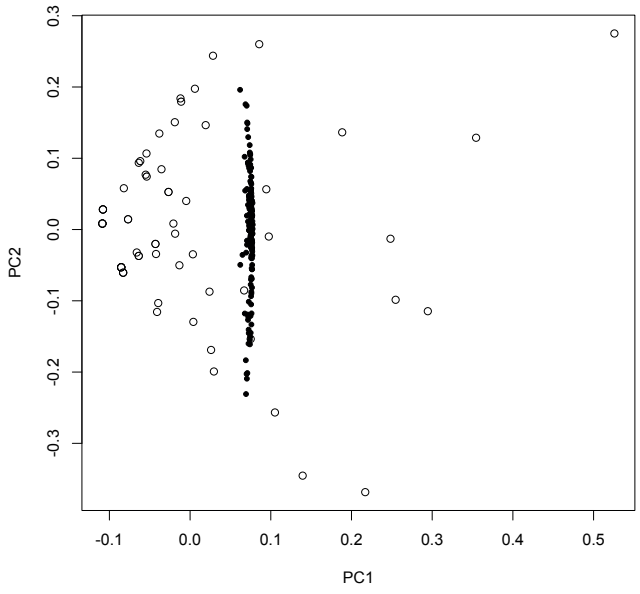


(a)

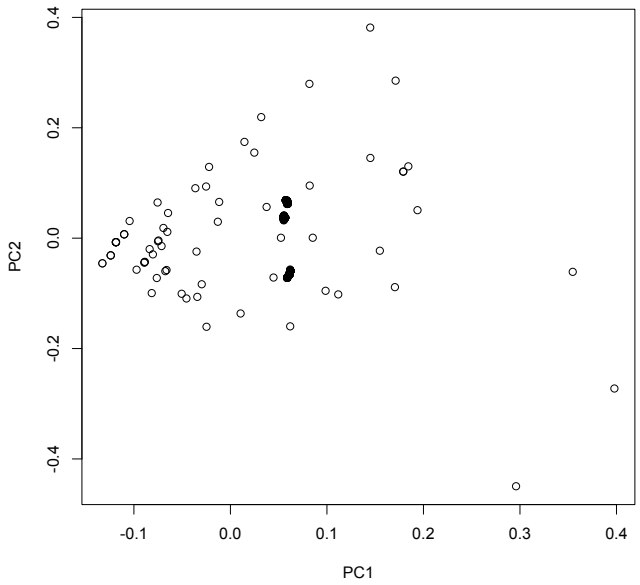


(b)

Figure 4.7: Plot of genes from KEGG dataset projected into PC1 and PC2 (a) without using a proximity matrix, (b) using the information-content proximity matrix. The comparison shows the importance of proximity matrix for visualisation. Legend: ∘ is gene.

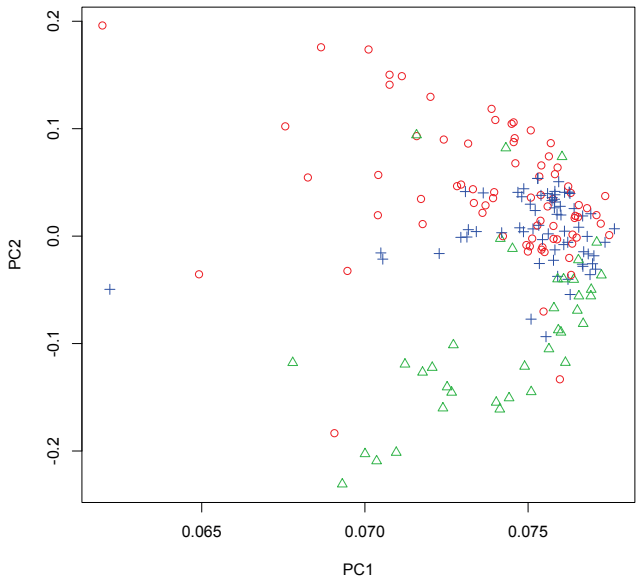


(a)

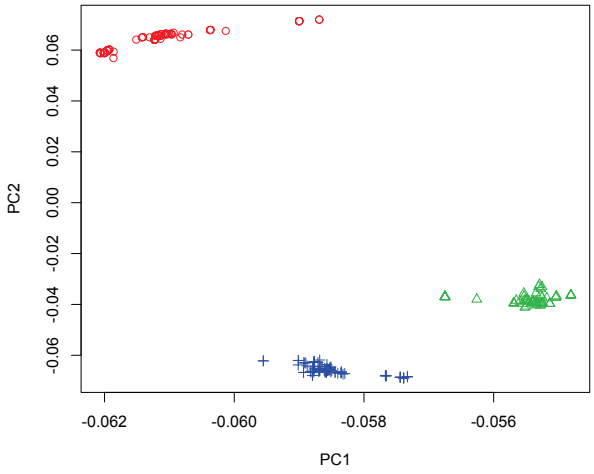


(b)

Figure 4.8: Plot for PC1 and PC2 for both methods using KEGG dataset. (a) Hop-based method where GO terms are making arc shape with no clear separation. (b) IC similarity measure where GO terms are clustered based on sub-ontologies. Legend: ∘ is gene, • is term.

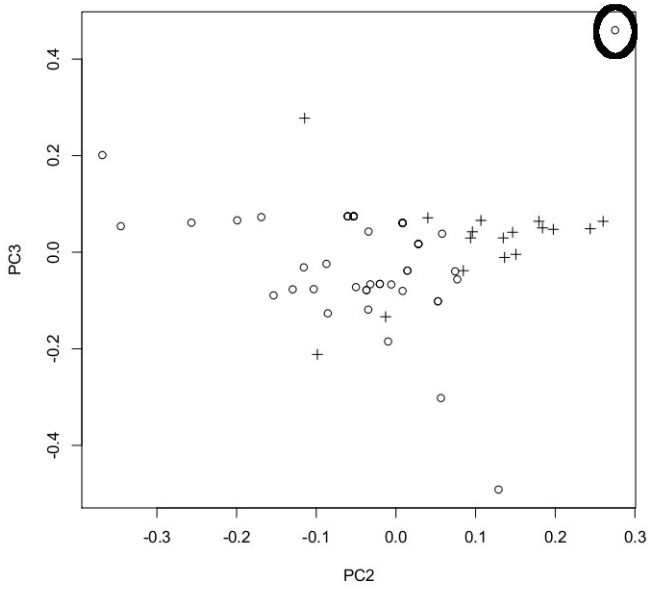


(a)

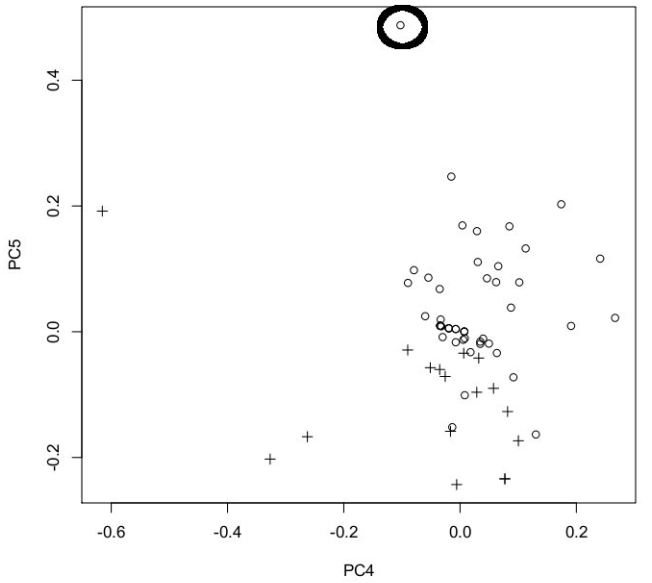


(b)

Figure 4.9: Plot of terms from the KEGG dataset projected into PC1 and PC2 (a) for the hop–based proximity matrix, (b) using the information-content proximity matrix. Legend: + is molecular function GO term, ∘ is biological process GO term and △ is cellular component term.



(a)



(b)

Figure 4.10: (a) Principal components (PC) 2 and 3 for hop-based method and (b) PC4 and 5 for IC method. Legend: (∘) is genetic information processing genes and (+) represent carbohydrate metabolism genes.

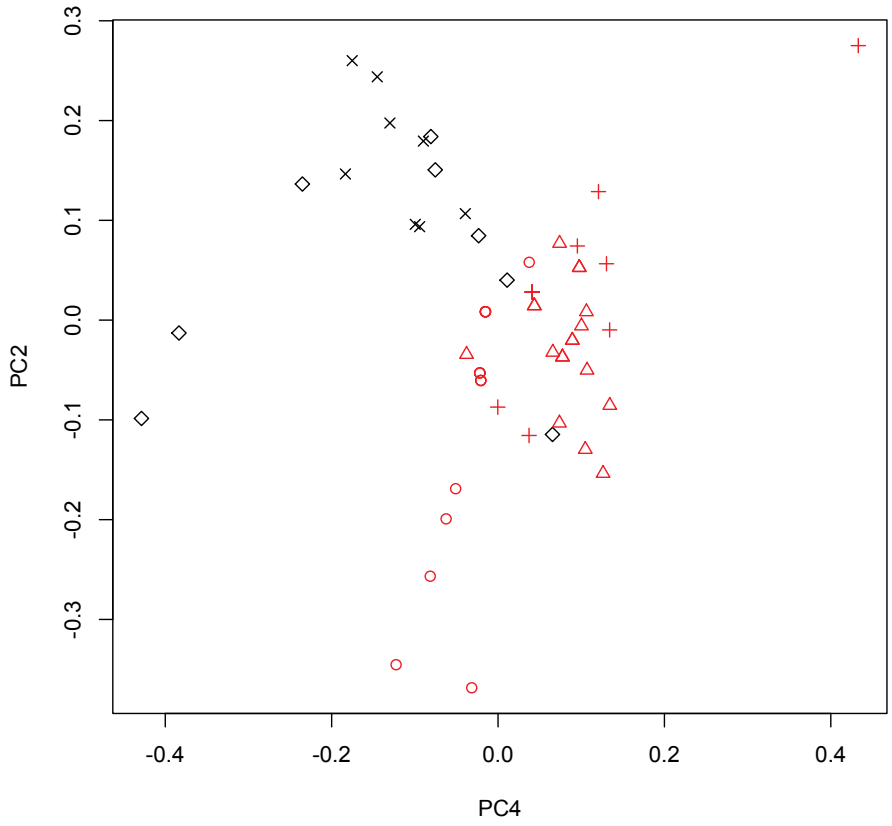
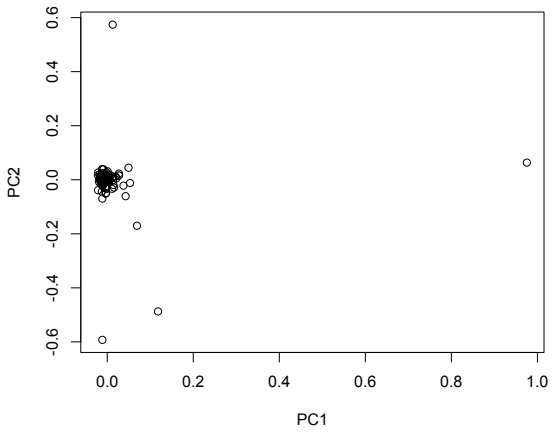
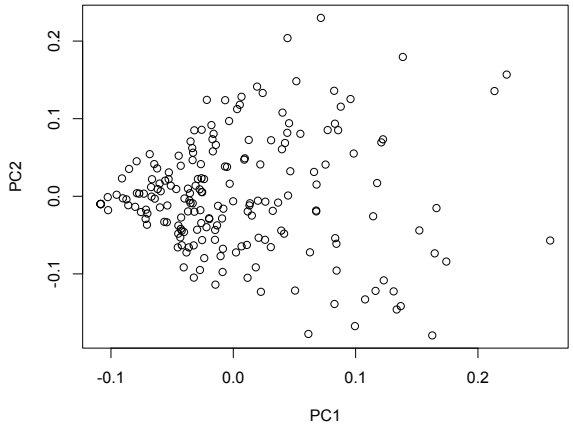


Figure 4.11: Plot of principal components 4 and 2 for **U** matrix (genes) for the KEGG dataset using the hop–based approach. Genes are related to KEGG categories for ribosomes (∘), RNA polymerase (△), transcription (+), pentose phosphate pathway (×) and pentose and glucoronate interconversions (⋄).

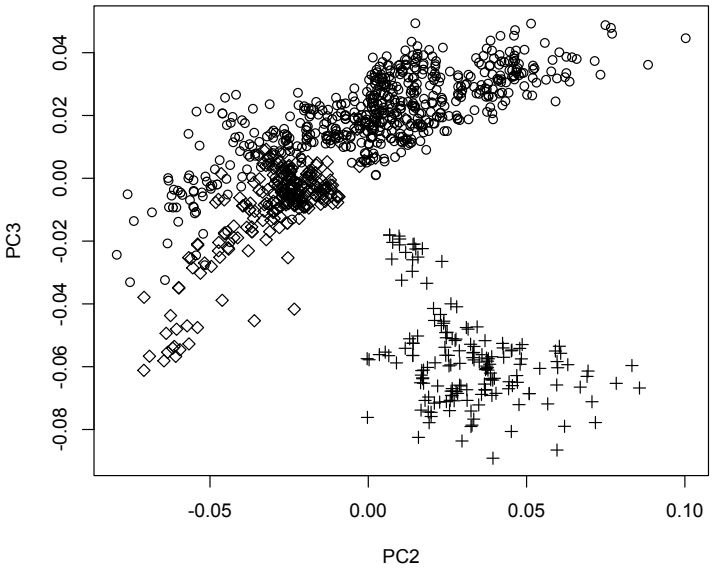


(a)

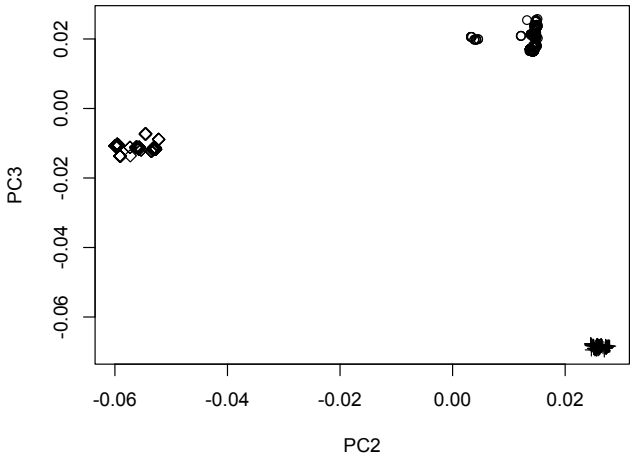


(b)

Figure 4.12: Plot of genes from the cancer dataset projected into PC1 and PC2 (a) without using a proximity matrix, (b) using the information-content proximity matrix. Legend: ∘ is gene.

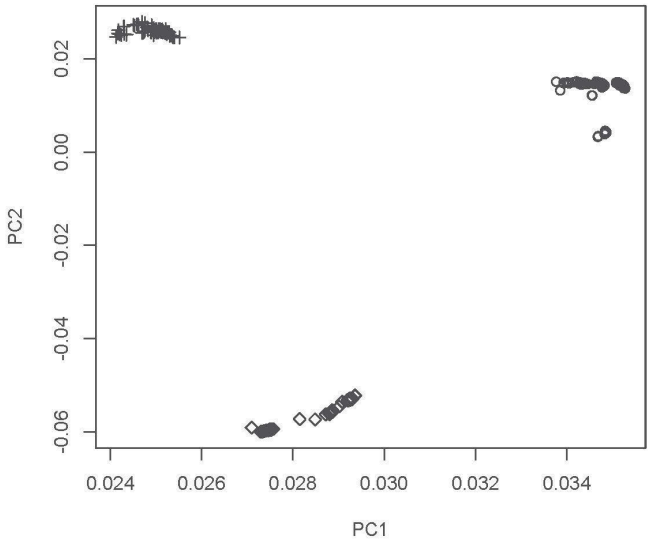


(a)

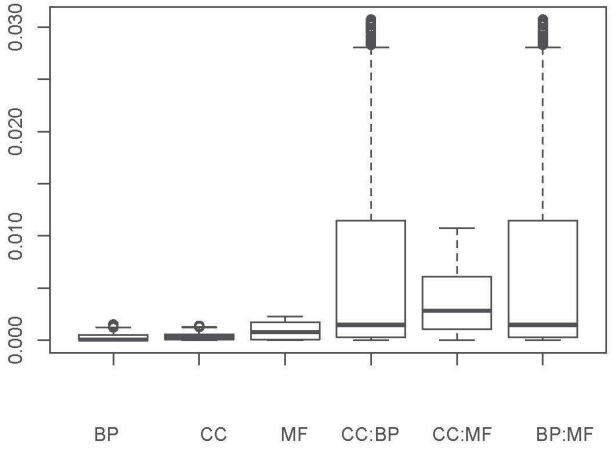


(b)

Figure 4.13: Plot of terms from the cancer dataset projected into PC2 and PC3 (a) using hop–based measure, (b) using the information-content based measure. Legend: $\diamond$ is molecular function GO term, $\circ$ is biological process GO term and $+$ is cellular component term.

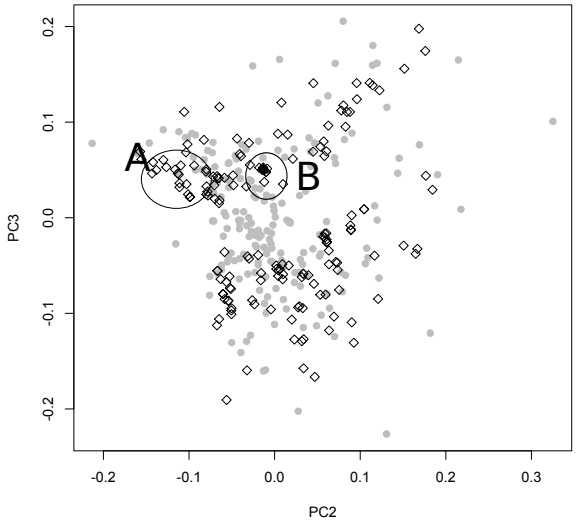


(a)

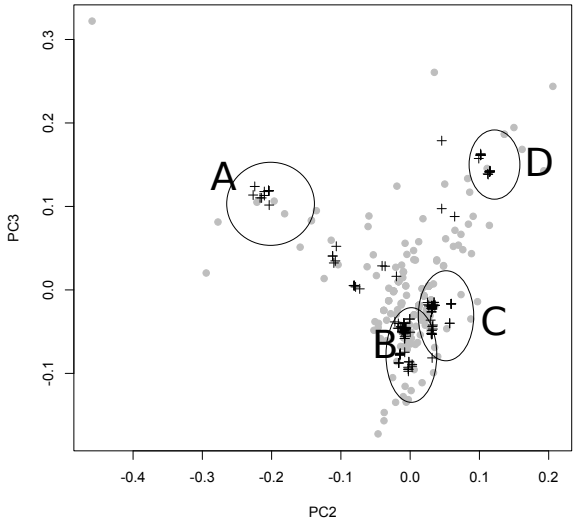


(b)

Figure 4.14: Terms from the cancer dataset projected into PC1 and PC2 using the IC similarity measure form clusters associated with the sub-ontology. (a) Plot of terms projected to PC1 and PC2, Legend: $\diamond$ is molecular function GO term, $\circ$ is biological process GO term and $+$ is cellular component term (b) Distributions of distances inside and between clusters over PC1 and PC2.

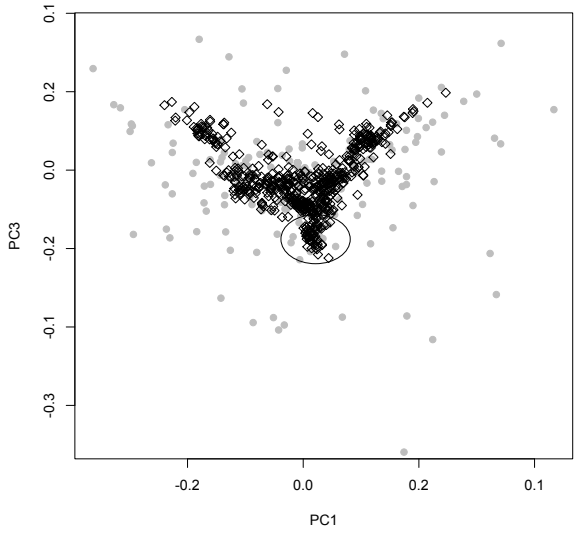


(a)

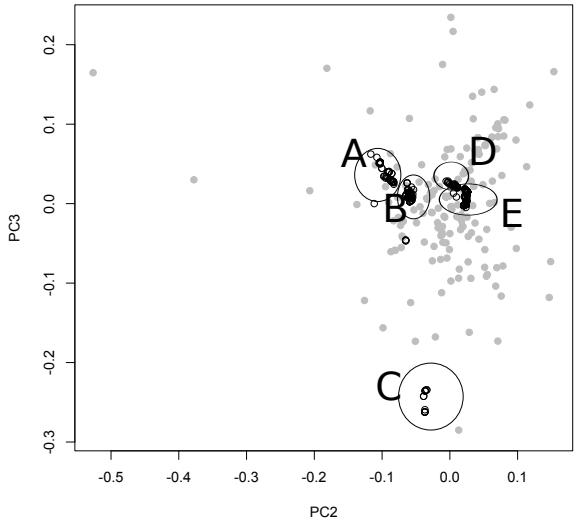


(b)

Figure 4.15: Plot of principal components 2 and 3 of cancer dataset with cellular component (CC) terms. (a) Hop-based similarity measure. Legend: (•) is genes and (⋄) is CC terms.(b) IC similarity measure. Legend: (•) is genes and (+) is CC terms.

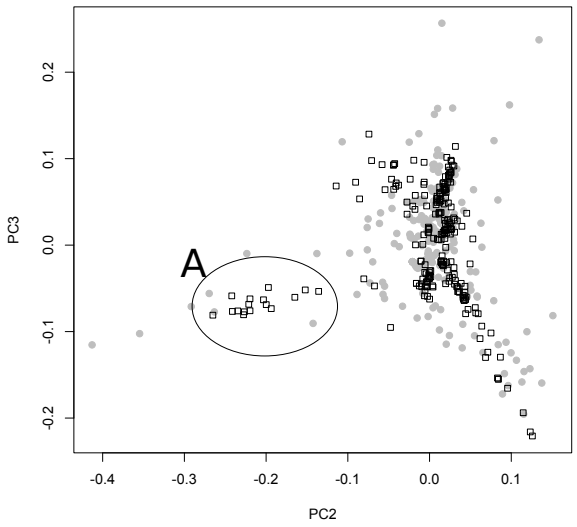


(a)

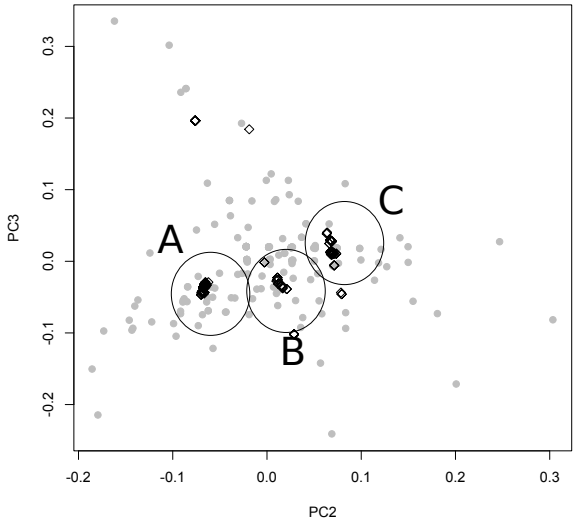


(b)

Figure 4.16: Plot of principal components 2 and 3 of cancer dataset with biological process(BP) terms. (a) Hop based similarity measure. Legend: (•) is genes and (◇) is BP terms (b) IC similarity measure. Legend: (•) is genes and (○) is BP terms.



(a)



(b)

Figure 4.17: Plot of principal components 2 and 3 of cancer dataset with molecular function terms. Legend: (•) is genes and (⋄) is molecular function terms. (a) Hop based similarity measure (b) IC similarity measure.

Table 4.7: GO term clusters using IC method for Cellular Components(CC), Biological Process(BP) and Molecular Function(MF) of GO based on correlation results.

Clusters	Example Terms	Description
CC Terms		
A	GO:0042175/nuclear envelope-endoplasmic reticulum network, GO:0005887/integral to plasma membrane	Cluster of membrane and extracellular matrix
B	GO:0005635/nuclear envelope, GO:0030117/membrane coat and GO:0000324/fungal-type vacuole	Organelles
C	GO:0042719/mitochondrial inter-membrane space protein transporter complex, GO:0005760/gamma DNA polymerase complex and GO:0031588/AMP-activated protein kinase complex	Protein complexes
D	GO:0044430/cytoskeletal part, GO:0031616/spindle pole centrosome and GO:0000922/spindle pole	Cell division apparatus
BP Terms		
A	GO:0001658/branching involved in ureteric bud morphogenesis, GO:0048754/branching morphogenesis of a tube and GO:0001947 heart looping	Morphogenesis and Early Development (Stem Cells)
B	GO:0006974/response to DNA damage stimulus, GO:0007548/sex differentiation and GO:0007276/gamete generation	Response to Stimulus Transport or Homeostasis
C	GO:0010468/regulation of gene expression GO:0010628/positive regulation of gene expression and GO:0005975/carbohydrate metabolic process	Gene expression regulation and metabolism
D	GO:0048676/axon extension involved in development, GO:0045467/R7 cell development and GO:0007409/axonogenesis	Differentiation
E	GO:0000718/nucleotide-excision repair, DNA damage removal, GO:0000720/pyrimidine dimer repair by nucleotide-excision repair, GO:0000724/double strand break repair via homologous recombination	DNA metabolism and function with a number of small subgroups e.g. vesicle transport
MF Terms		
A	GO:0003678/DNA helicase activity, GO:0004003/ATP-dependent DNA helicase activity and GO:0008026/ATP-dependent helicase activity	Enzyme activity No.1
B	GO:0003777/microtubule motor activity, GO:0003774/motor activity and GO:0003924/GTPase activity	Enzyme activity No.2
C	GO:0016853/isomerase activity, GO:0003689/DNA clamp loader activity and GO:0003916/DNA topoisomerase activity	Molecular interactions non-enzymatic

Chapter 5

Visualising Leukaemia Cancer Dataset using NLDR

5.1 Introduction

As described in Chapters 1 and 2, human genomic and transcriptomic data relevant to cancer is high-dimensional. Meanwhile, visualisation of high-dimensional data has become an important part of life sciences research, and now affects the design of treatment methodologies. Many visualisation techniques exist to reduce the dimensionality in a certain set of data. These techniques classify the data, correlate the data and visualise sub-classes within a class. Generally, dimensionality reduction methods (DRM) transform a high-dimensional dataset into a lower number of dimensions, for example two-or three-dimensional data, which could be displayed in a scatter plot. The key aim of dimensionality reduction approaches is to summarize a large number of data attributes into a smaller set with no redundancy, or less redundancy.

The focus of this chapter is to apply the dimensionality reduction and machine learning methods described in Section 2.3.1 to the ALL gene expression dataset described in Section 5.2.1, and to compare these linear and nonlinear feature selection

methods based on parameter choices in terms of their effect on classification accuracy between two patient groups (relapsing and non-relapsing). Later in this chapter, PCA will be used as pre-processing (as suggested by Lee and Verleysen (2007)) to these linear and non-linear methods and the results will be compared.

5.2 Experimental Design

this thesis has proposed a data analysis framework to analyse, visualise and classify gene expression data of childhood cancer, acute lymphoblastic leukaemia (ALL). In this chapter, linear and nonlinear dimensionality reduction methods and parameter choices have been applied to classify between relapsed and non-relapsed patients. The framework consists of three steps.

In the first step of this thesis, the random forest method was performed on a raw dataset of 99 patients and 55000 gene expression profiles to find important variables (in this case gene expressions). In the second step, data was centered and scaled, which was suggested by Park et al. (2003) before applying Principal Component Analysis (PCA), kPCA, Local Linear Embedding (LLE), Stochastic Neighbour Embedding (SNE) and Diffusion Maps (see details in Section 2.3.1). Each of these methods finds different structures in high-dimensional data. For instance, LLE preserves local properties of data while PCA, kPCA, DM and SNE preserves global properties of data. Principal Component Analysis finds linear transformation based on variance while kPCA finds the non-linear mapping based on kernel function. Diffusion Maps calculates distance based on a Markov random walk on the graph of the data to certain number of timestamps while SNE is an iterative technique which is a global method which preserves local properties because of its cost function. In the third

and final step, classification was performed using Support Vector Machines (SVM), a hyperplane-based classifier which chooses the maximal margin hyperplane in feature space to minimize the risk of overfitting (Cortes and Vapnik, 1995) and classification accuracy was assessed using an area under curve (AUC) approach, to determine which methods perform best on classifying patients as relapsing or non-relapsing on the basis of the gene expression data. The proposed framework is presented in Figure 5.1.

5.2.1 Dataset

The dataset used in this chapter is a gene expression dataset retrieved from the NCBI Gene Expression Omnibus (accession number GSE7440). It consists of 99 patients treated under the COG1961 protocol and whose diagnostic bone marrow samples were hybridised to Affymetrix U133 Plus 2.0 microarrays (Bhojwani et al., 2008). This dataset has 54,675 probes per patient. The information available for these patients are: sex, age, white blood cell count, translocation and description. The description consists of 4 values: Rapid early responder (RER), slow early responder (SER), complete continuous remission and relapsed. There were some patients who were RER or SER but relapsed. Relapsed are the patients who went through the treatment but still relapsed later. The summary of the dataset is shown in Table 5.1 and a density plot of the data is shown in Figure 5.2. A complete list can be found in Appendix Table 1). In this thesis, the patients are classified based on relapsed and non-relapsed.

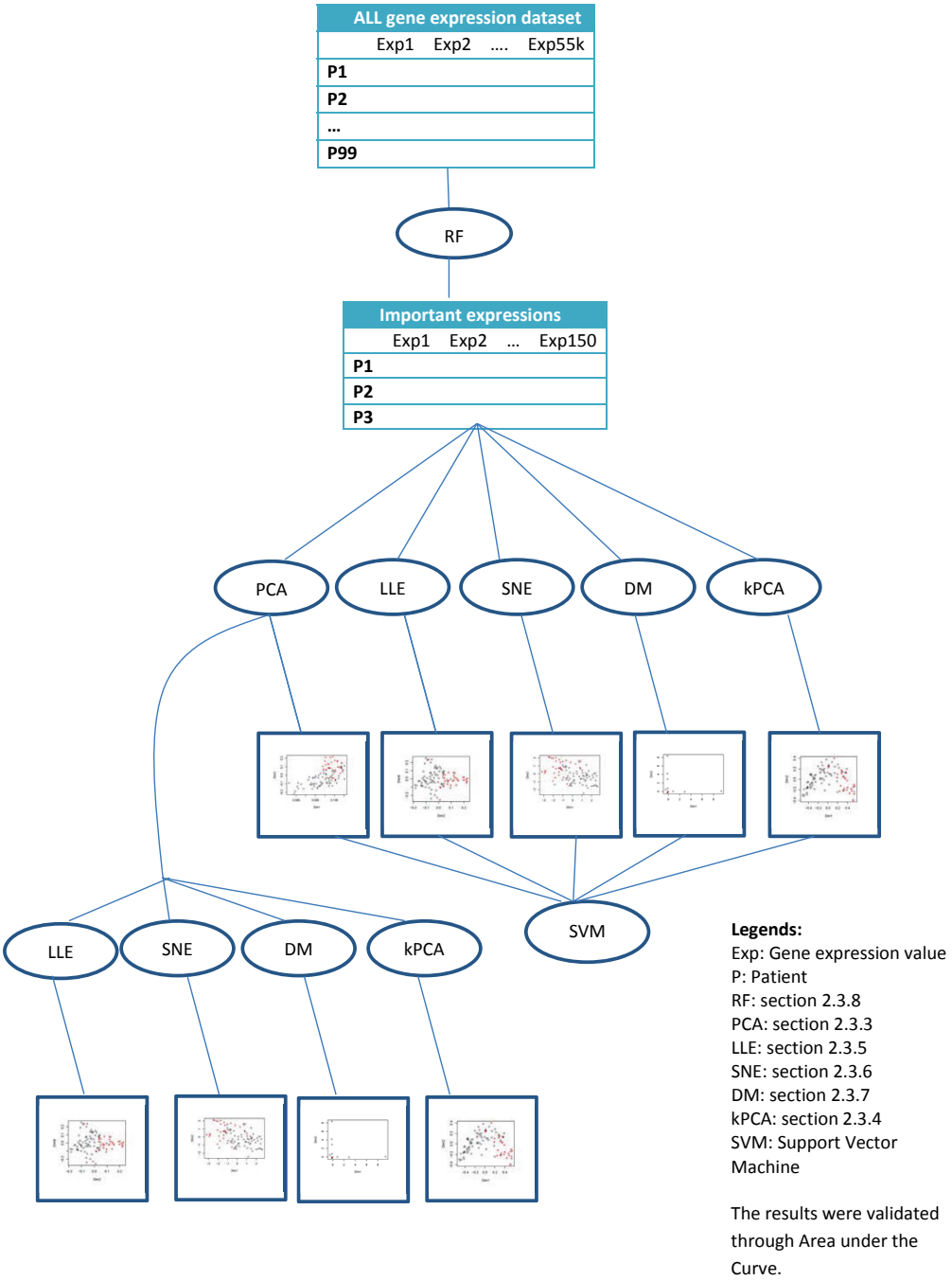


Figure 5.1: Experimental design for classification of ALL patients based on relapse status.

Table 5.1: Summary of gene expression dataset for ALL patients used in this chapter.

Male Female	61 38
Age (months)	Minimum =12 Maximum = 225 Average = 122.42 Median = 134 Standard deviation = 61.55
White blood cells(109/L)	Minimum =1800 Maximum =732000 Average =101601.4 Median = 65800 Standard deviation = 120627.4
Translocation information	Yes = 27 No= 72
Relapsed or non-relapsed count	Relapsed= 31 Non-relapsed= 68

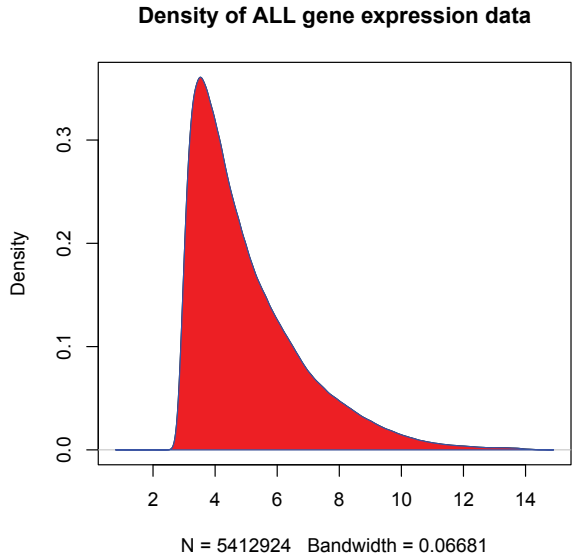


Figure 5.2: Density plot of ALL gene expression data where N is total number of probesets and bandwidth is based on minimum and maximum gene expression value.

5.2.2 Attribute selection

A Random Forest (Breiman, 2001) model of 50,000 trees was constructed with the dataset using the randomForest package (Liaw and Wiener, 2002) on R (R Core Team, 2015). The treatment outcome of the patients (i.e. whether the patient experienced a relapse or not) was used as the class labels. The 150 probes with the largest mean decrease in Gini index, an impurity measurement that indicates the probes' importance to the classification problem, were selected for further analysis.

5.3 Results and Discussion

The experimental results obtained in this thesis can be discussed in three parts as given below:

(i) Identification of the most suitable method on the basis of AUC results

(ii) Biological interpretation of the results

(iii) Using PCA as preprocessing to nonlinear dimensionality reduction

5.4 Identification of the most suitable method on the basis of AUC results

As shown in Figure 5.1, this analysis compared a range of linear and nonlinear dimensionality reduction methods applied to a reduced dataset of 150 probes. After dimensionality reduction, SVM using a Radial Basis Function (rbfdot) kernel was used for classification of patients, and performance of the different methods was assessed by AUC. The partition of the dataset was 70% training, 15% validation and 15% test dataset.

For diffusion maps (DM), this thesis tuned parameters based on different numbers of dimensions with different number of timestamps, an obtained measure for the proximity of the data points. It was found that DM did not perform well overall on the dataset, as the AUC on test dataset is 0.5 with minimum value 0.3889 (98 dimensions over 30 timestamps), representing performance equivalent to random selection. The maximum value 0.625 was achieved when dimensions were reduced to 50 dimensions

and used 30 timestamps. There was no clear relationship between the number of dimensions and the timestamps as shown in Table 5.2.

For kPCA using the linear kernel, better performance was observed with lower numbers of dimensions. When reduced to 50 dimensions, it shows AUC value of 0.7500 on test data while the best results 0.9167 for kPCA were achieved when dimensions were reduced to 10. While the the AUC value on the test dataset was 0.6667 as shown in Table 5.2.

LLE depends on the number of neighbors per hyperplane. In this thesis different numbers of dimensions were selected with different numbers of neighbors per hyperplane. It was found that average AUC results on test data were around 0.78 with minimum value achieved with 97 dimensions and 30 neighbors per hyperplane. The results show that working with lower number of neighbors per hyperplane performs better than higher numbers of neighbors per hyperplane. The best result (0.9028) on test data was achieved when dimensions were reduced to 10 with 30 neighbors per hyperplane.

SNE has shown the best results across all the methods compared in this thesis. Because SNE does not require specific parameter settings, evaluation was done based on numbers of dimensions. The AUC achieved through SNE was 0.9861 when the dimension were reduced to 10. A similar AUC value was also achieved with 2 dimensions as shown in Table 5.2.

In the next step, optimal method settings were used with varying percentages of training,validation and test data as shown in Figure 5.2. Diffusion Maps (reduced to 50 dimensions) performed best with 25% test dataset (0.6316) while SNE (reduced to

10 dimensions) performed best with 15% test dataset (0.9861 respectively). Subsequently PCA, kPCA (reduced to 10 dimension) and LLE (reduced to 10 dimensions) achieved best results with 20%test dataset 0.7265, 0.9829 and 0.9658 respectively. So this thesis suggests using 65% training, 15% validation and 20% test dataset.

For further validation, an average AUC was collected for 30 randomly-selected samples of different training and test datasets. The reason for selecting 30 was j to choose a representative number for this experiment. Support vector machine and AUC were performed on these sample datasets for 15% and 20% validation and test datasets respectively. The box-plot results have been presented in Figures 5.3 and 5.4. These box-plots show that kPCA, LLE and SNE have much higher values than Diffusion Maps and PCA.

The dimensionality reduction methods showed varying levels of success in visualising the genetic distinction between relapsed and non-relapsed patients. The patients on the first dimensions for each nonlinear method, and the degree of separation observed, was similar to the differences between methods found on basis of AUC values.The distinction between relapsed and non-relapsed patients can be seen at the first dimension in PCA, SNE, kPCA and LLE as shown in Figures 5.5, 5.6, 5.7 and 5.8 respectively.

However, the distinctions lacked clear boundaries, and the cluster of two patients are visible in kPCA, SNE, LLE and PCA as shown in Figures 5.9, 5.10, 5.11, 5.12. Diffusion Maps were not able to show this distinction at all. The patients in the PCA, SNE, kPCA and LLE plots are more spread than DM (See Figures 5.5, 5.6, 5.7 and 5.8 respectively). This spread of patients in the space may reflect the overall genetic diversity of the cohort while in kPCA as shown in Figure 5.7 the separation

between patients could be seen making an arc. The second dimension in PCA and LLE provided the best separation between patients relapsed (+) and non-relapsed (o) while the same type of separation can been seen in dimension one for SNE and kPCA.

Table 5.2: AUC values calculated with different percentage of training, validation and test data based on mean

	60/15/25	50/15/35	65/15/20	70/15/15
DM	0.63	0.32	0.53	0.62
PCA	0.73	0.7	0.72	0.67
kPCA	0.90	0.91	0.98	0.91
LLE	0.92	0.91	0.96	0.90
SNE	0.92	0.92	0.98	0.98

In these plots, it can be seen that some relapsed patients (red, +) are overlapping with non-relapsed patients (black, o). Because of the spread between patients, the Euclidean distance between the patients was computed to see how far apart they are in lower-dimensional space. It was found that the distance between patients is not very high in PCA as compared to SNE, LLE and kPCA suggesting that patients are very close to each other in PCA. A sample of the distance between patients can be seen in Tables 5.3, 5.4, 5.5 and 5.6.

Next, patients from differing classes who were close together in the lower-dimensional space, and patients from the same class who were far apart in the lower-dimensional space, were investigated on the basis of their gene expression. It was found that a two-patient cluster, GSM180178 (non-relapsed) and GSM180179 (relapsed), were

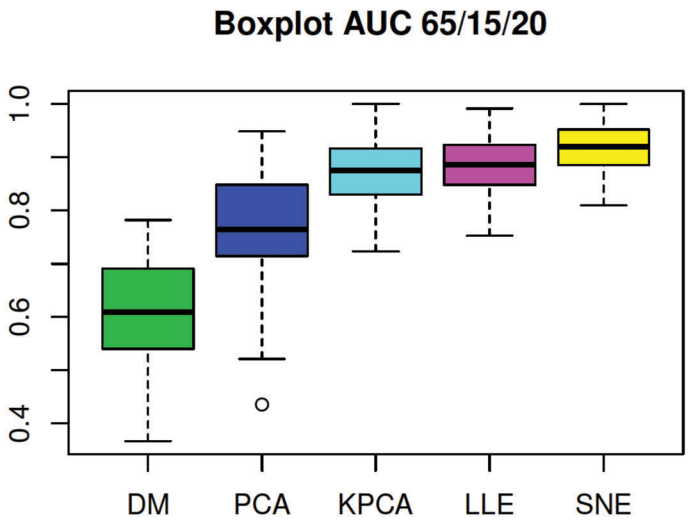


Figure 5.3: Box-plot comparison of results with data distribution as 65% training, 15% validation and 20% test data for all the methods.

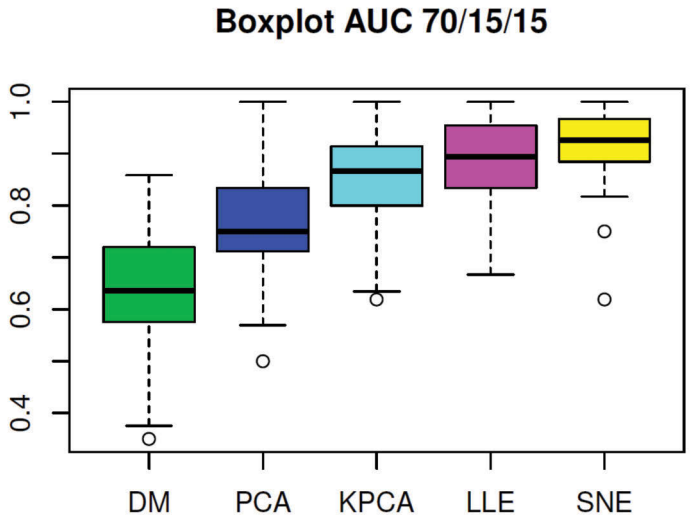


Figure 5.4: Box-plot comparison of results with data distribution as 70% training, 15% validation and 15% test data over 30 random values between 1 to 10000.

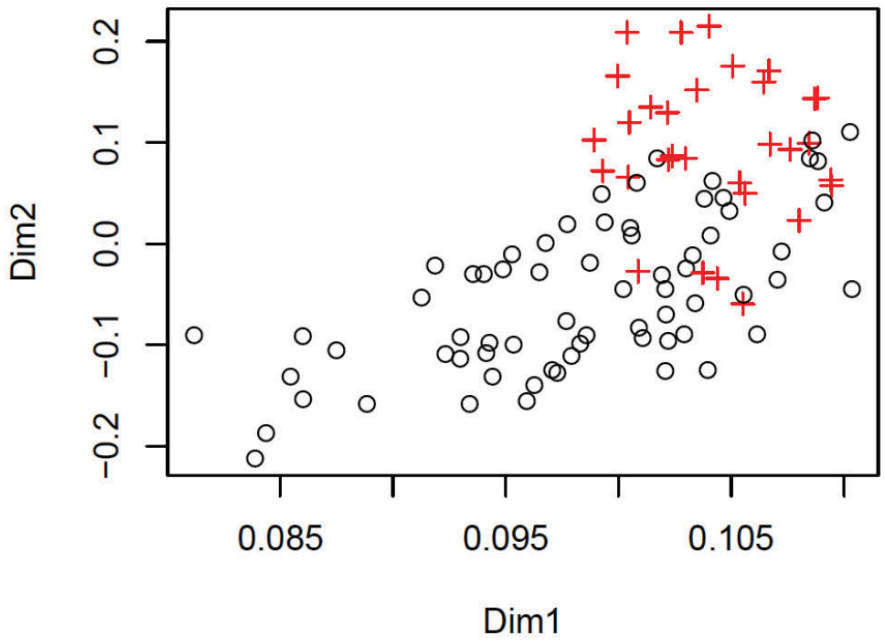


Figure 5.5: The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for PCA. Legend: + is relapsed patients, ∘ is non-relapsed patients

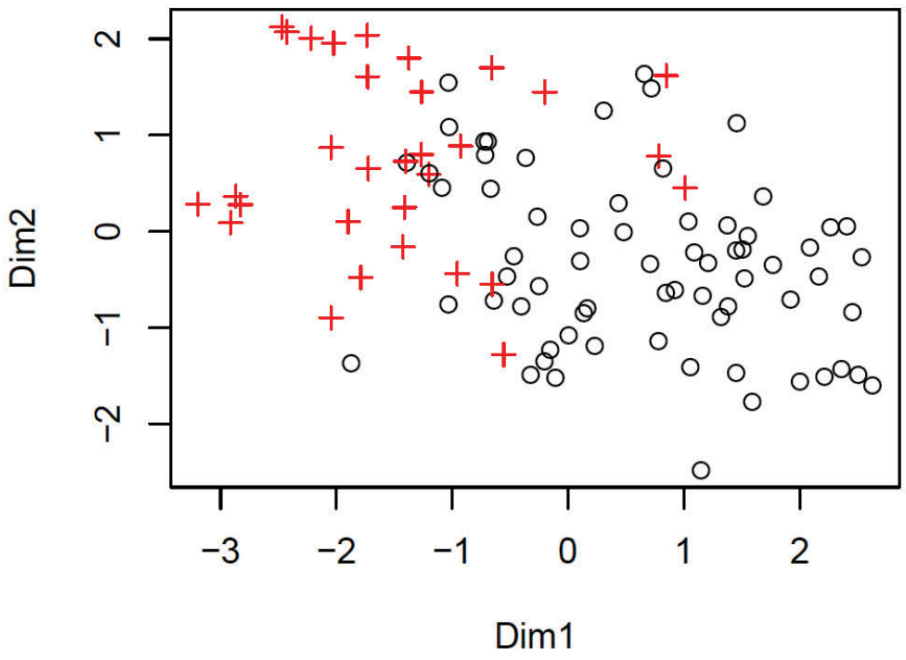


Figure 5.6: The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for SNE. Legend: + is relapsed patients, ∘ is non-relapsed patients

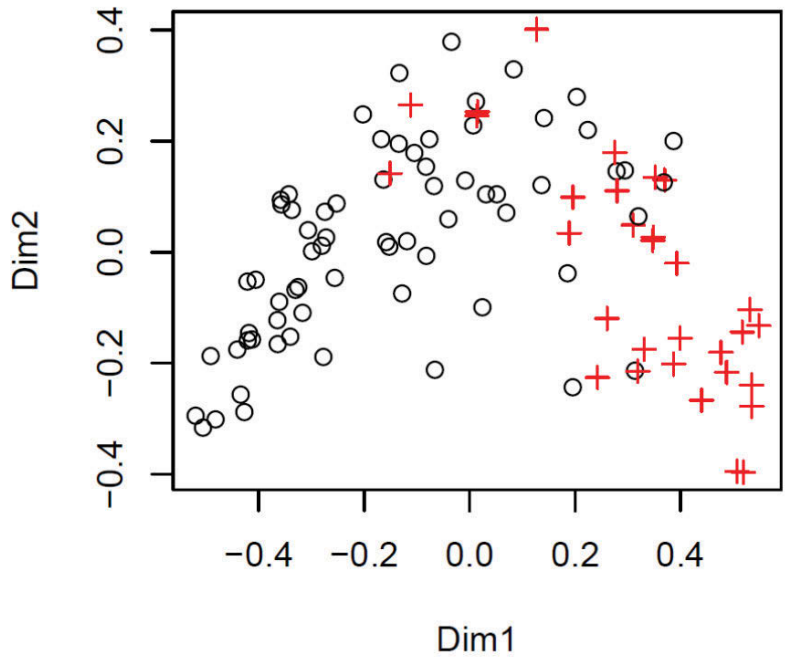


Figure 5.7: The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for kPCA. Legend: + is relapsed patients, ∘ is non-relapsed patients

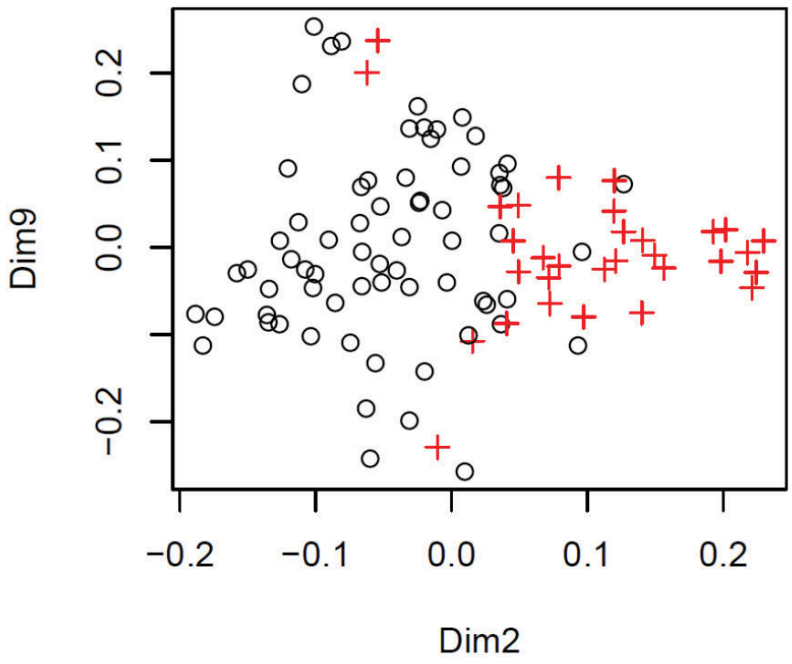


Figure 5.8: The resulting plot of patients into dimension1 (Dim1) and dimension 9 (Dim9) for LLE. Legend: + is relapsed patients, ∘ is non-relapsed patients

seen in visualisations of PCA, SNE, kPCA and LLE as shown in Figures 5.9, 5.10, 5.11, 5.10 and 5.12.

Similarly, patients GSM180185 (relapsed) and GSM180188 (non-relapsed) were also found together in all visualisations except DM. PCA highlighted GSM180186 (replased) close to GSM180205 (non-relapsed) which cannot be seen together in other visualisations, suggesting that their relationship is not based on variance of data (shown in Figure 5.12). Similarly, kPCA identified GSM180155 (non-relapsed) and GSM180180 (relapsed) close to each other (shown in Figure 5.9) while LLE identified GSM180202 (replased) and GSM180198 (non-relapsed) together (shown in Figure 5.11). The classification of relapse and non-relapsed patients is more clear in SNE than in other visualisations, only one pair of opposite class (GSM180184, relapsed) and GSM180139, non-relapsed) is clustered, other than the two common pairs described above (shown in Figure 5.10). These results show the diversity of dimensionality reduction methods used in this chapter. The biological finding of why these patients have been clustered together will be presented in next section.

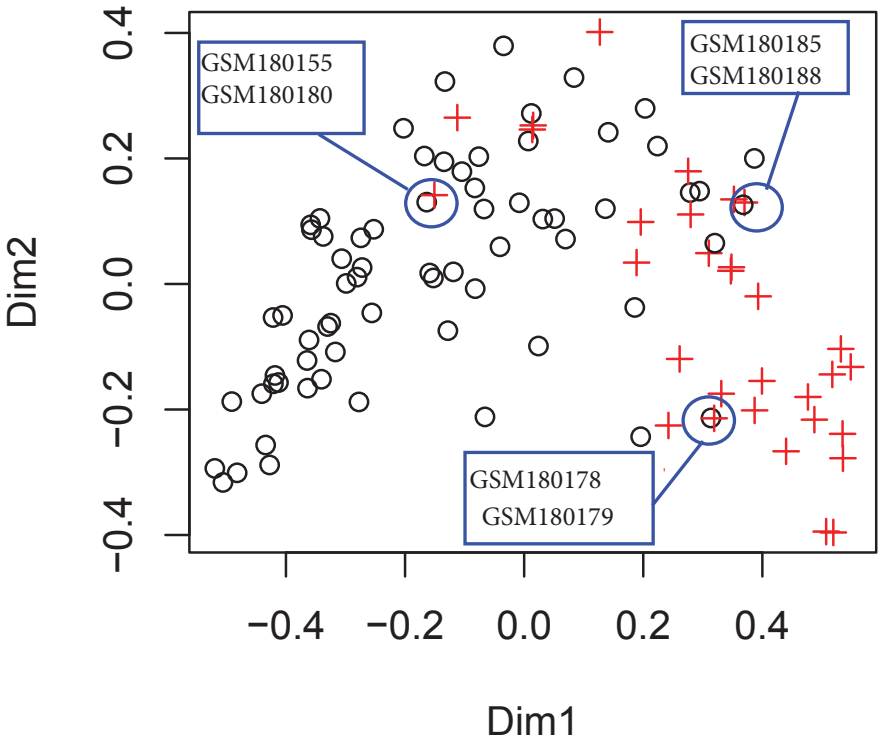


Figure 5.9: Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for kPCA. Legend: + is relapsed patients, ∘ is non-relapsed patients

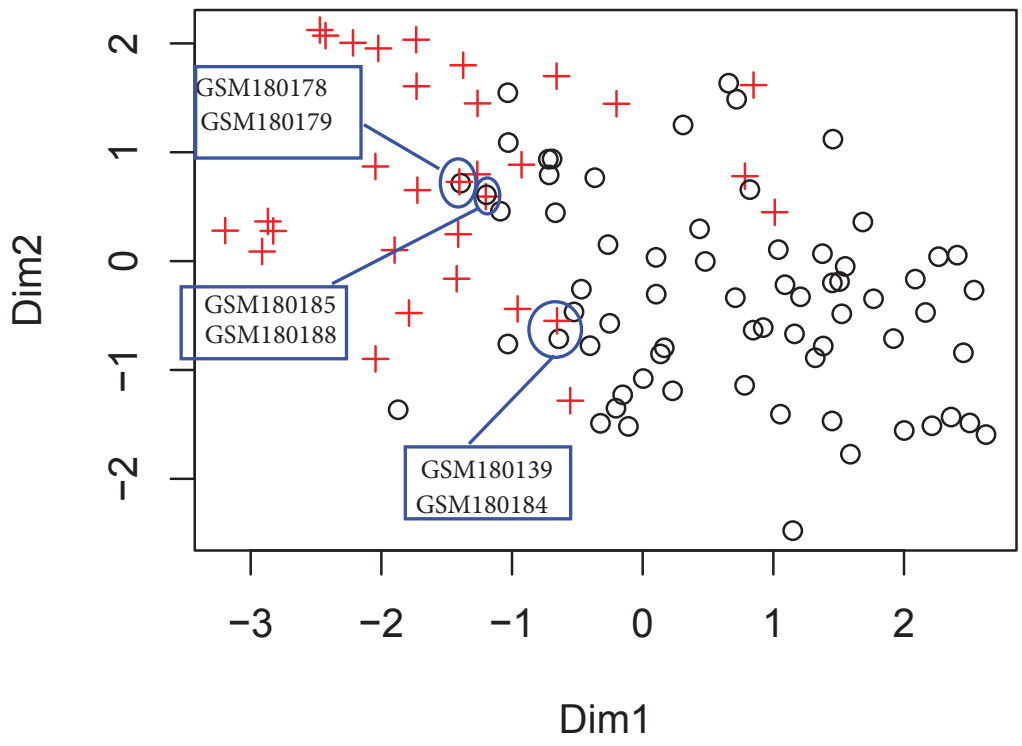


Figure 5.10: Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for SNE. Legend: + is relapsed patients, ∘ is non-relapsed patients

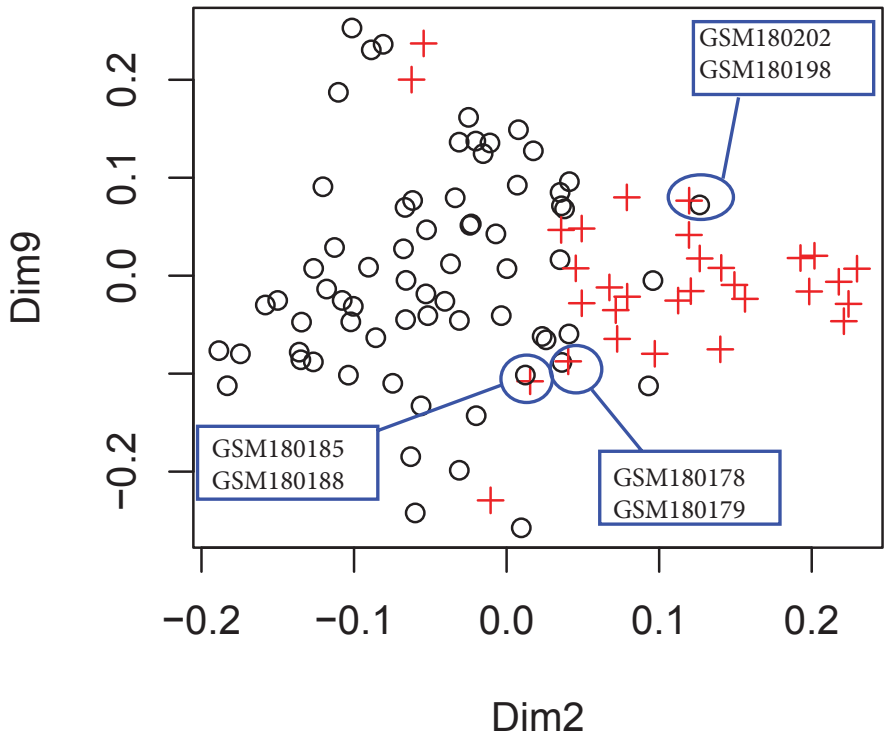


Figure 5.11: Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 9 (Dim9) for LLE. Legend: + is relapsed patients, ∘ is non-relapsed patients

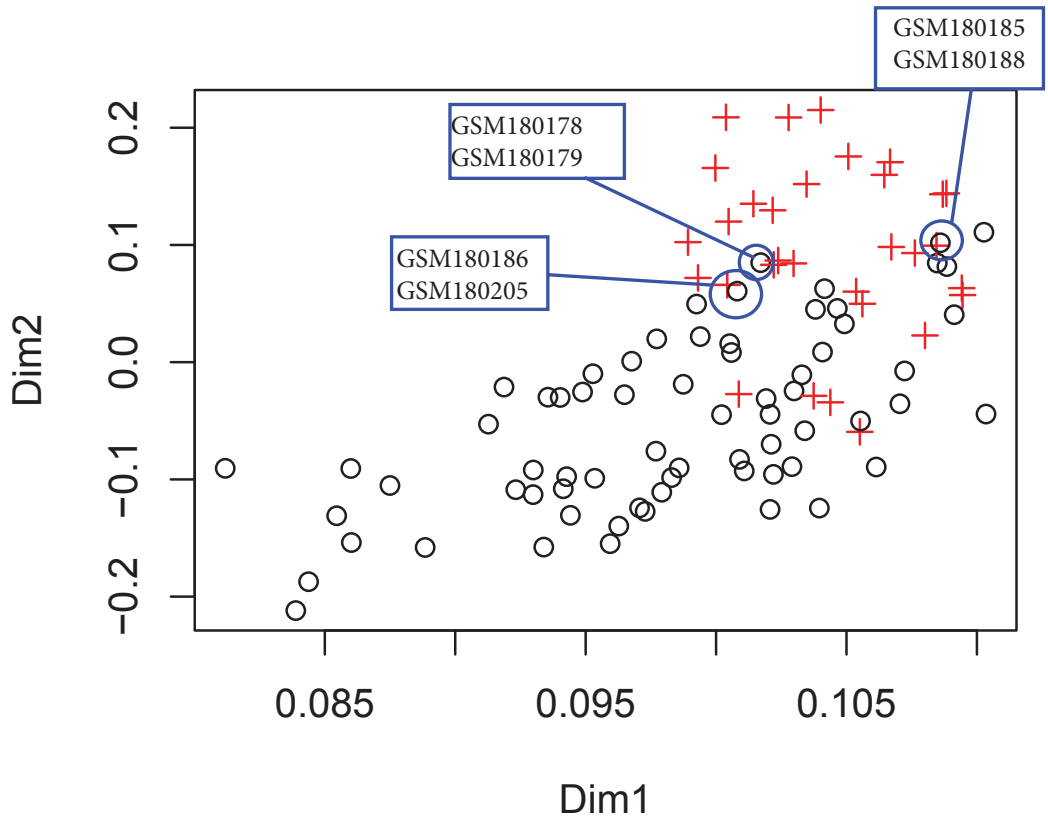


Figure 5.12: Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for PCA. Legend: + is relapsed patients, ∘ is non-relapsed patients

Diffusion Maps (DM) found a tight cluster of all patients, and two potential outliers orthogonal to each other as shown in Figure 5.13). In the lower dimensions, the two outliers drifted closer towards the main cluster of patients which itself has spread into indistinct clusters. DM did not show convincing separation compared to other visualisation methods.

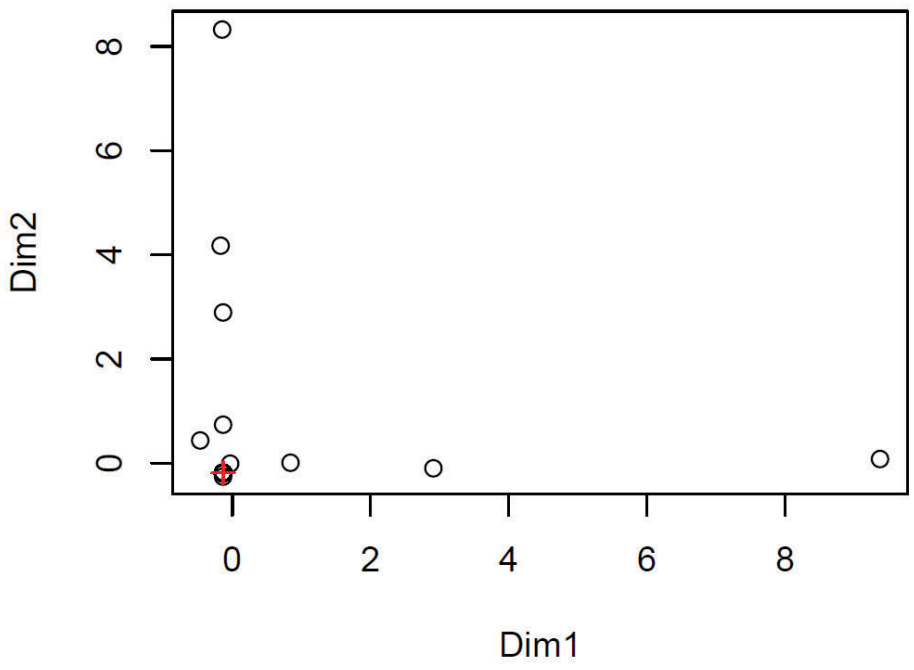


Figure 5.13: The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for DM. Legend: + is relapsed patients, ∘ is non-relapsed patients

Table 5.3: Distance calculated between points in PCA feature space where CP1 is class for patient 1 and CP2 is class for patient 2. P1 is point 1 and P2 is point 2

Patient 1	Patient 2	CP1	CP2	Point 1	Point 2	Distance P1 and P2
GSM180188	GSM180185	Non-relapsed	Relapsed	0.09	0.10	1.41
GSM180178	GSM180179	Non-relapsed	Relapsed	0.10	0.102	1.41
GSM180136	GSM180141	Non-relapsed	Non-relapsed	0.09	0.086	1.41

Table 5.4: Distance calculated between points in SNE feature space where CP1 is class for patient 1 and CP2 is class for patient 2. P1 is point 1 and P2 is point 2

Patient 1	Patient 2	CP1	CP2	Point 1	Point 2	Distance P1 and P2
GSM180178	GSM180179	Non-relapsed	Relapsed	-1.39	-1.40	0.014
GSM180185	GSM180188	Relapsed	Non-relapsed	-1.20	-1.19	0.018
GSM180145	GSM180178	Non-relapsed	Non-relapsed	2.62	-1.39	4.63

Table 5.5: Distance calculated between points in kPCA feature space where CP1 is class for patient 1 and CP2 is class for patient 2. P1 is point 1 and P2 is point 2

Patient 1	Patient 2	CP1	CP2	Point 1	Point 2	Distance P1 and P2
GSM180178	GSM180179	Non-relapsed	Relapsed	0.31	0.31	0.009
GSM180185	GSM180188	Relapsed	Non-relapsed	0.37	0.36	0.022
GSM180183	GSM180216	Relapsed	Relapsed	0.48	0.01	1.133

Table 5.6: Distance calculated between points in LLE feature space where CP1 is class for patient 1 and CP2 is class for patient 2. P1 is point 1 and P2 is point 2

Patient 1	Patient 2	CP1	CP2	Point 1	Point 2	Distance P1 and P2
GSM180178	GSM180179	Non-relapsed	Relapsed	-0.43	-0.43	0.01
GSM180185	GSM180188	Relapsed	Non-relapsed	-0.01	-0.007	0.03
GSM180178	GSM180188	Non-relapsed	Non-relapsed	-0.43	-0.007	0.87
GSM180179	GSM180185	Relapsed	Relapsed	-0.43	-0.01	0.85

5.4.1 Biological interpretation of results

In this section, the similarities in gene expression profiles between patients of differing classes were investigated. Further investigation was done on two pairs of patients (patients GSM180178 (non-relapsed), patient GSM180179 (relapsed) and GSM180185 (relapsed), GSM180188 (non-relapsed)). Biological attributes of these patients are presented in Table 5.7.

Patient GSM180185 differs from patient GSM180188 based on sex, age and white blood cell count. To further test these results, regression analysis was performed between these two patients and found that the correlation between these patients is 0.9981079, which shows that these patients are very similar to each other, while covariance is 3.1701. The R-squared value is 0.995 while the adjusted R-squared value is the same (0.995). These results consolidate our findings that these patients are very similar to each other, but the sex of patient may have played role in GSM180185 to relapse as shown in Table 5.8 and Figure 5.15.

Similarly, patients GSM180178 and patient GSM180179 are very close in age and both are male, but the difference in white blood cells is very significant. Nguyen et al. (2002) have suggested that white blood cells play a significant role in prognosis of leukemia. So white blood cells may have caused the patient GSM180179 to relapse. Regression analysis validated these results by achieving the correlation value 0.9981079 and R-squared value is 0.9962 as shown in Table 5.8 and Figure 5.14. Figure 5.14 shows that the values of both patients are very close to each other which supports that they are very similar patients but white blood cells may have caused GSM180179 to relapse.

Table 5.7: Biological data for interesting patients

Sample	Sex	Age(months)	WBC(109/L)	Description
GSM180178	Male	167	4400	SER;CCR
GSM180179	Male	176	93500	RER;relapse
GSM180185	Male	158	164000	SER;relapse
GSM180188	Female	41	64500	RER;CCR

Hence, it was found that patients closer together have almost the same gene expression profile as shown in Table 5.7 and high R-squared value as shown in Table 5.8.

In further analysis, these patients were compared with other patients through Pearson correlation. A comparison was performed between Euclidean distance calculated between each pair of patients in the previous section, and Pearson correlation calculated between all patients as shown in Figure 5.16. The comparison was performed on all methods used in this chapter. SNE achieved the best results using 2 dimensions and 10 dimensions. The highlighted points represent comparisons between patients with similar expression profiles but different outcomes (red) described above, or patients with dissimilar expression profiles, but similar outcomes (blue).

The comparison in Figure 5.16 also shows that SNE 10 (with 10 dimensions), SNE 2 (with 2 dimensions) and kPCA found patients with similar expression profiles at the top left corner while PCA found the same patients on the far left hand side. The comparison shows that the patients are more spread through distance in PCA and LLE than kPCA and SNE. The correlation in kPCA, SNE (10) and SNE (2) is more linear than in LLE and PCA. Comparison between PCA and LLE shows that PCA

is more spread than LLE but the distance between points is minimal while the points are more spread in LLE.

These graphs support our finding that SNE has performed the best among all methods, with kPCA marginally behind, while LLE and PCA found the same patients close to each other but overall they are more spread out than SNE and kPCA. PCA is a method that spreads the data as described in Section 2.3.1. The DM did not show any kind of spread and no evidence of patients from different classes being together which also support the findings that DM did not perform well on this dataset.

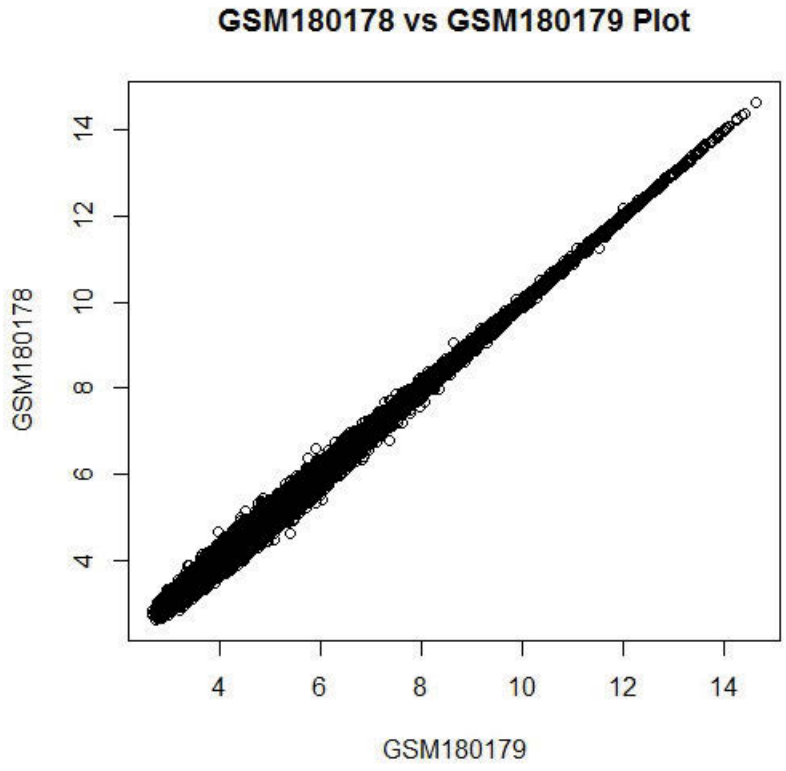


Figure 5.14: The regression plot of gene expression values for patients GSM180178 and GSM180179. The ∘ (black) are gene expression values for each probe

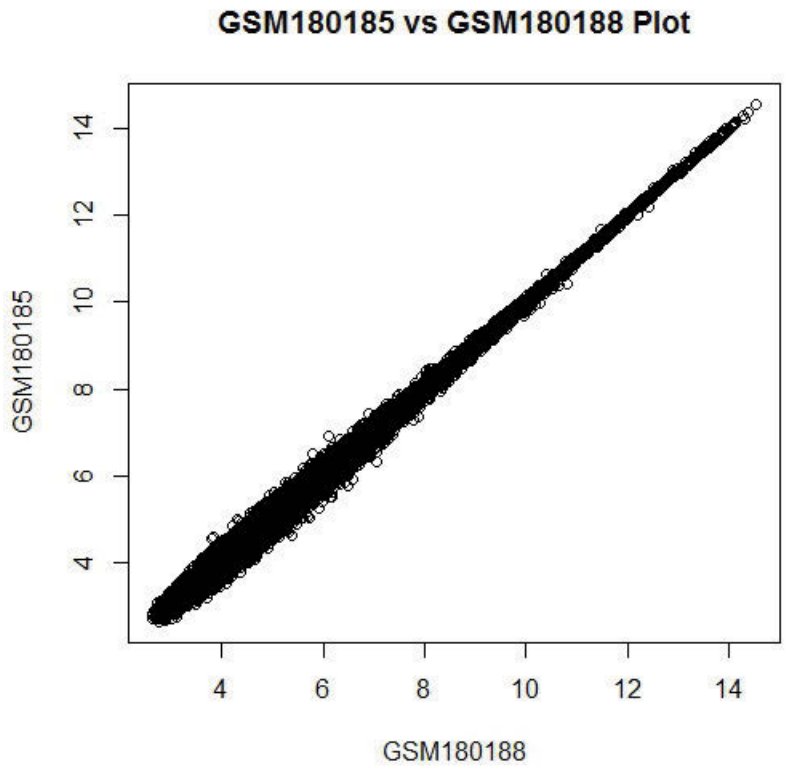


Figure 5.15: The regression plot of gene expression values of patients GSM180185 and GSM180188. The ∘ (black) are gene expression values for each probe

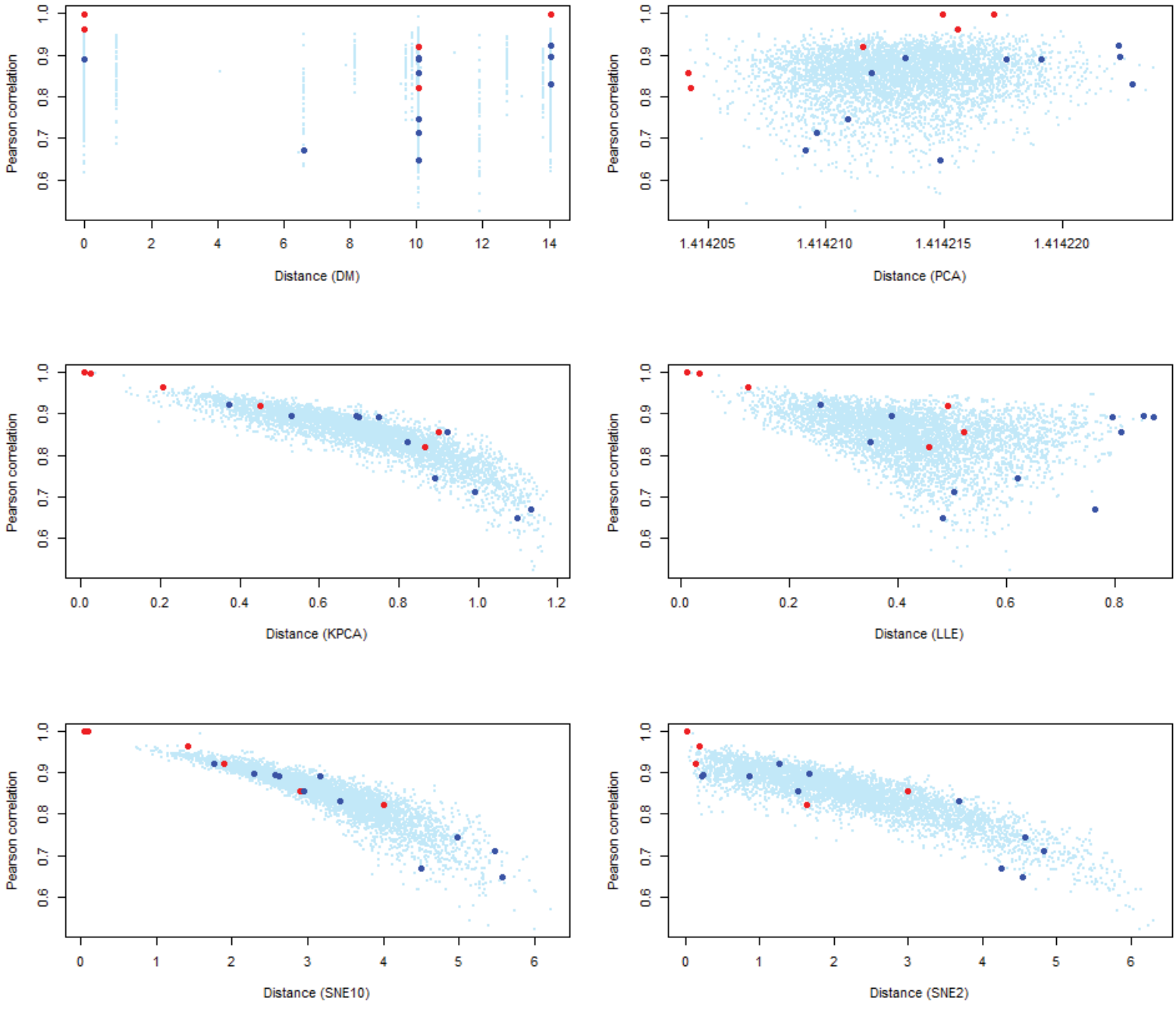


Figure 5.16: Pearson correlation for the 150 gene values for each pairwise comparison with the distance measure for each pair. Patients close together but different outcomes are in red, while patients far apart with same outcome are represented in blue

Table 5.8: Regression results for patient pairs of GSM180185, GSM180188 and GSM180178 and GSM180179

Patients	Min	Median	Max	R-squared	Adjusted R-squared
GSM180185, GSM180188	-0.76034	0.00197	0.82201	0.995	0.995
GSM180178, GSM180179	-0.77955	0.00050	0.70165	0.9962	0.9962

5.4.2 Impact of PCA application prior to nonlinear dimensionality reduction on results

Lee and Verleysen (2007) has suggested that using PCA as preprocessing before dimensionality reduction methods may improve the functionality of methods while dealing with high-dimensional datasets. In the last part of this experiment, PCA was used as a preprocessing technique before applying DM, SNE, LLE and kPCA. Similar to the previous experiment in Section 5.2, the parameters of each method were tuned to find the optimal method.

Using DM, the results were similar to those obtained without PCA (as discussed in Section 5.4), with no real clusters highlighted between patients. For DM, the best results were achieved with 50 dimensions and 30 timestamps, producing an AUC of 0.625, while kPCA achieved an AUC of 0.6944 when reduced to 2 dimensions, which is lower than achieved through kPCA in the previous experiment Section 5.4. Local Linear Embedding did not perform well in this scenario, with many data points that were lost for all dimensions, while SNE performed similar to kPCA. SNE achieved the best AUC of 0.6944 when dimensions were reduced to 2, which is lower than the value achieved in previous Section 5.4. The visualisation of these methods did not show

any clear classification of relapsed and non-relapsed patients. These results suggest that at least for the experiment in this chapter, applying PCA before other machine learning methods does not improve performance on an ALL gene expression dataset.

5.5 Summary

The purpose of this work was to compare various nonlinear dimensionality reduction techniques on an ALL gene expression dataset to classify between relapsed and non-relapsed patients. The methods used in this chapter consisted of Principal Component Analysis (PCA), kPCA, Local Linear Embedding (LLE), Stochastic Neighbour Embedding (SNE) and Diffusion Maps. Each of these methods finds different structures in high-dimensional data. For instance, LLE preserves local properties of data, while PCA, kPCA, DM and SNE preserves global properties of data. Principal Component Analysis finds a linear transformation based on variance, while kPCA finds a non-linear mapping based on kernel function. Diffusion Maps calculates distance based on Markov random walk on the graph of the data to a certain number of timestamps, while SNE is an iterative technique which is a global method, but it does preserve local properties because of its cost function. The parameters in these methods were tuned to find the optimal method to classify relapsed and non-relapsed ALL patients. The classification was performed using SVM while the classification accuracy was assessed using area under the curve (AUC).

The results showed that SNE produced the highest AUC value 0.9306 among all the methods, while LLE and kPCA were closest to SNE. kPCA achieved the best AUC when dimensions were reduced to 10 with 0.916, while LLE produced AUC 0.902 when dimensions were reduced to 10 with 30 neighbors per hyperplane. PCA

and DM did not achieve high accuracy as compared to other methods with AUC values 0.666 and 0.5 respectively.

The visualisations of SNE, kPCA, LLE and PCA showed varying levels of success for genetic distinction between relapsed and non-relapsed patients. The separation of patients from both classes (class represents relapse or non-relapse) were apparent, but some patients were overlapping. Interestingly, two pairs of patients, GSM180185, GSM180188, and GSM180178 and GSM180179, were highlighted in SNE, kPCA, LLE and PCA. Both pairs made separate clusters in these visualisations. The gene expression profiles for these patients were investigated further. Patients GSM180185 and GSM180188 have very similar gene expression profiles, although GSM180188 mainly differ from GSM180185 based on age, sex and white blood cells. Similarly, the comparison between GSM180178 and GSM180179 shows similar gene expression values except the white blood cells count, in which GSM180179 has higher number of white blood cells which may have caused the patient to relapse.

The results in this chapter have demonstrated that nonlinear dimensionality reduction methods achieve higher classification accuracy for relapse in the dataset than linear methods, and hence would result in improved classification of ALL classes. This analysis addresses Objective 2, identifying an optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression, as described in Section 1.1 and the Thesis Contribution 2 in Section 1.4. The next chapter will focus on data mining analysis of SNP (genomic) data in ALL.

Chapter 6

Case Study: Finding pathways related to ALL from SNPs using random forest

6.1 Introduction

Metabolic pathways play an important role in the human body, and microarray data can be used to find pathways related to disease (as described in Section 3.3). The SNP dataset is a high-dimensional nonlinear dataset which leads to problems like missing values, unbalanced classes and feature selection as described in Section 6.2. The focus of this chapter is to use feature selection methods such as random forest on an ALL SNP dataset gathered at The Children's Hospital at Westmead, and a healthy control dataset from Wellcome Trust UK, to find metabolic pathways related to ALL. Random forest is a method that can help in finding the top-ranked variables in high-dimensional data as described in Section 2.3.8. The relevant SNPs are found through the random forest method, then further processed to find the genes related to them, and eventually to identify pathways related to ALL. Genes related to high-ranked SNPs are then further evaluated to find a functional relationship between

them using the framework described in Section 4.2

6.2 Dataset

The ALL SNPs dataset contains a large number of SNP values. In this work, the SNP dataset consists of 4 values, '0', '1', '2' and '-1'. Values '0' through '2' represents the number of minor alleles present in a specific SNP for a patient, while the value '-1' means missing values. Summary of the dataset is presented in Figure 6.1.

The dataset used for this experiment contains SNP data generated from samples taken from ALL patients at The Children's Hospital at Westmead, and healthy controls from Wellcome Trust UK. The dataset of ALL patients contains 139 patients and 13980 SNPs while healthy control data consists of 476 samples with values for the same SNPs. The collated dataset is 615 rows (with the number of healthy controls and ALL patients) by 13980 columns (the number of SNPs).

6.3 Experimental Design

The framework for analysing the SNP data consists of five steps as shown in Figure 6.3. In the first step, a dataset of SNPs from ALL patients and healthy controls was constructed.

In the second step, a feature selection method, random forest (described in Section 2.3.8) was applied to find the highest ranked SNPs. The random forest method ranks the contribution of variables to a classification problem, based on mean decrease of the Gini index (described in Section 2.3.8). The accuracy of the results achieved from random forest was based on the error rate of the confusion matrix and kappa statistics

values.

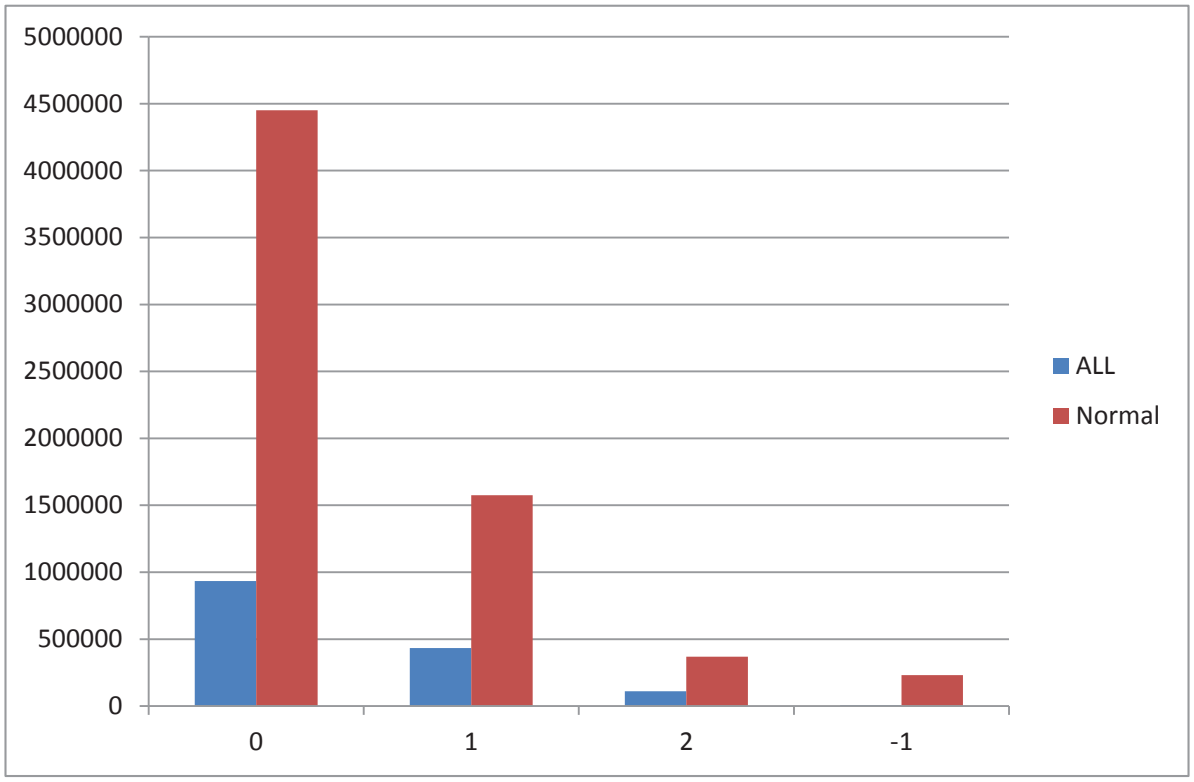


Figure 6.1: Distribution of data values 0, 1, 2 and -1, i.e minor alleles present in ALL and healthy control SNP datasets where blue bars represents ALL data and red bars are healthy control data. The y-axis represents the total number of values while the x-axis represents the number of minor alleles. The relative plot of both datasets can be seen in Figure 6.2.

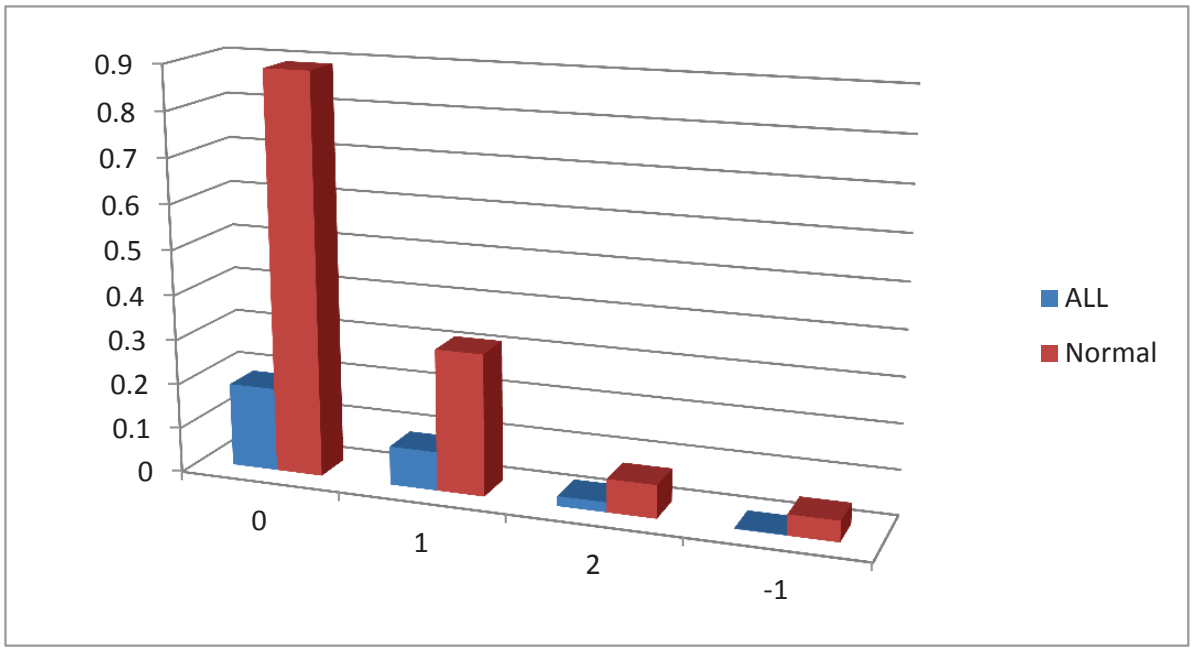


Figure 6.2: The relative frequency plot of both datasets ALL and healthy controls. The total number of minor alleles 0, 1, 2 and -1 is normalised between 0 to 1 to provide the distribution comparison between two dataset. The y-axis represents normalised value from 0 to 1 while the x-axis represents data values 0, 1, 2 and -1.

For validation leave-one-out cross validation is used. Arlot and Celisse (2010) has compared different cross validation methods and suggested that the cross validation method should be chosen based on data complexity and the model. The reasons to select leave-one-out cross validation method is because of bias, as in this experiment

the number of healthy control cases are 476 while ALL patients are 139 as shown in the figure 6.1, the high-dimensionality of the dataset, in this experiment total number of SNPs are 13980, and the complexity of random forest method required to process this data (as described in Section 2.3.8). The random forest method contains a 'number of trees' parameter for each forest, and a varying number of trees were selected to tune the random forest model. The best random forest model is selected based on a high kappa statistics value among all computed models. Kappa statistics are used to measure agreement between two classes on nominal scales (Cohen, 1960). Fleiss et al. (2013) has suggested kappa over 0.75 as excellent, 0.40 to 0.75 as good and below 0.40 as poor. The kappa value is measured between 0 to 1 where 1 is the highest value that can be achieved and 0 is the minimum value.

In the third step, top ranked SNPs were selected based on mean decrease of Gini index found by random forest, and genes were retrieved from the KEGG database (Section 2.2.3). The SNPs are located in a coding region of the genome for a particular gene.

In the fourth step, the metabolic pathways linked to those genes were retrieved from KEGG the metabolic pathways database (described in Section 2.2.3).

In the final step, the genes related to high-ranked SNPs found in the third step were evaluated through the framework described in Section 4.2, for finding functional relationships between them.

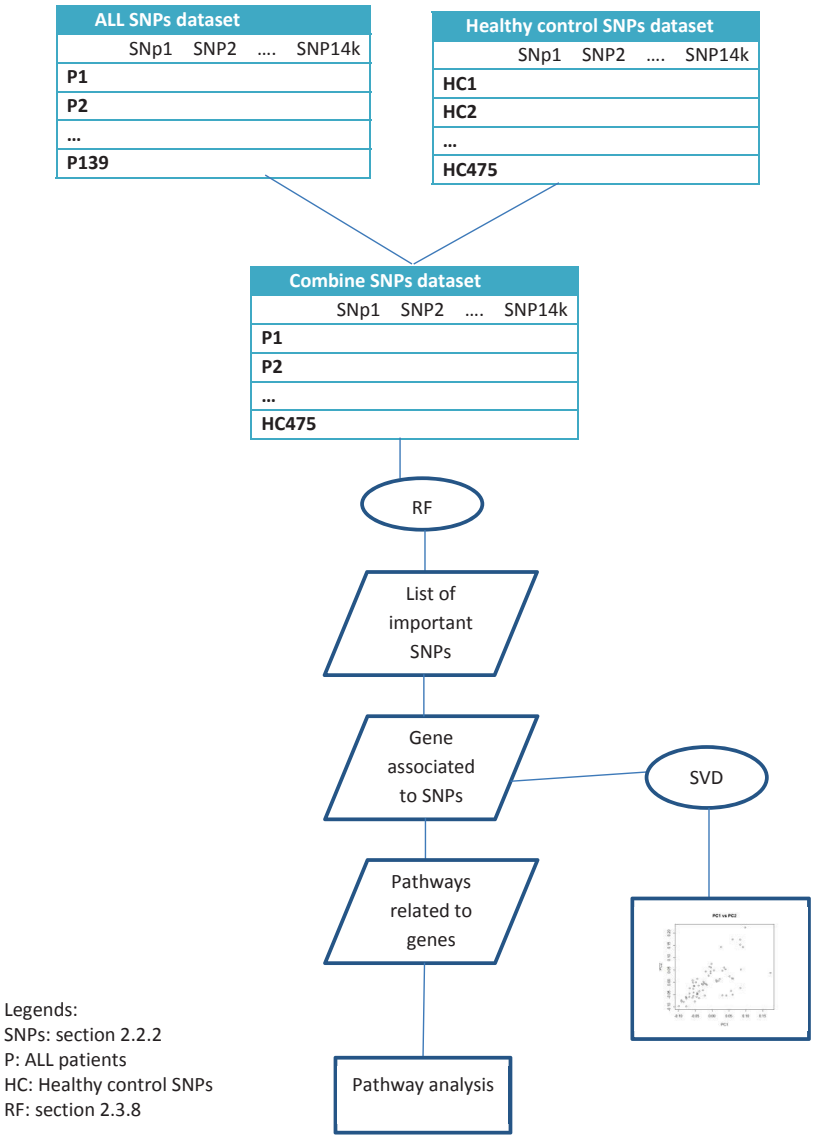


Figure 6.3: Experimental design for finding pathways related to ALL.

6.4 Results

The results section is divided into two sections:

(i) Finding the high-ranked SNPs by tuning the parameters of the random forest method

(ii) Finding pathways related to ALL

6.4.1 Finding high-ranked SNPs using random forest

The focus of finding high-ranked SNPs using random forest is further divided into three steps:

(i) Solving the missing values problem

(ii) Identifying the best percentage of training data and test dataset for this experiment,

(iii) Find the optimal number of trees in a forest for the random forest method

Missing values problem

The handling of missing values from dataset can be divided in two steps.

As shown in Figure 6.3, in the first step, the distribution of values of datasets of ALL patients and healthy controls were plotted to determine the trend of data. Figure 6.1 shows distribution of '0', '1', '2' and -1 in dataset. Random forest was applied on the dataset by using the default number of trees (100) and the partition of data was set to the default value at 70% training set and 30% test set. The kappa accuracy achieved in results was 1.00. Figure 6.4 represents the important variables

of the dataset. It was found that the top variables have 476 missing values in healthy control dataset. These results indicate that random forest was selecting the most important variables based on missing values.

In the last step, SNPs containing either all or a large number of missing values were removed from both healthy control and ALL dataset and missing values -1 were replaced with NA. To mitigate the effect of missing values, the random forest method was applied with default number of trees (100) after removing SNPs with all missing values for normal or ALL patients. The kappa statistics gain was still 1 and the random forest method was still selecting the top-ranked SNPs based on missing values. During several experiments, SNPs with missing values 300, 250, 150 and 50 were removed from dataset but the result was still the same (kappa value 1 and the top-ranked SNPs were influenced by missing values). When SNPs containing missing values 15 or more were removed, the kappa value gained was 0.6985 which suggests that missing values influence the results with 15 or more number of NA values. In the next step, SNPs with 15 or more missing values were removed from the dataset. The combined dataset was reduced to 139 ALL patients and 476 healthy control with 11469 SNPs (reduced from 13980). This dataset is then used for tuning the parameters of random forest to achieve the best model for this dataset.

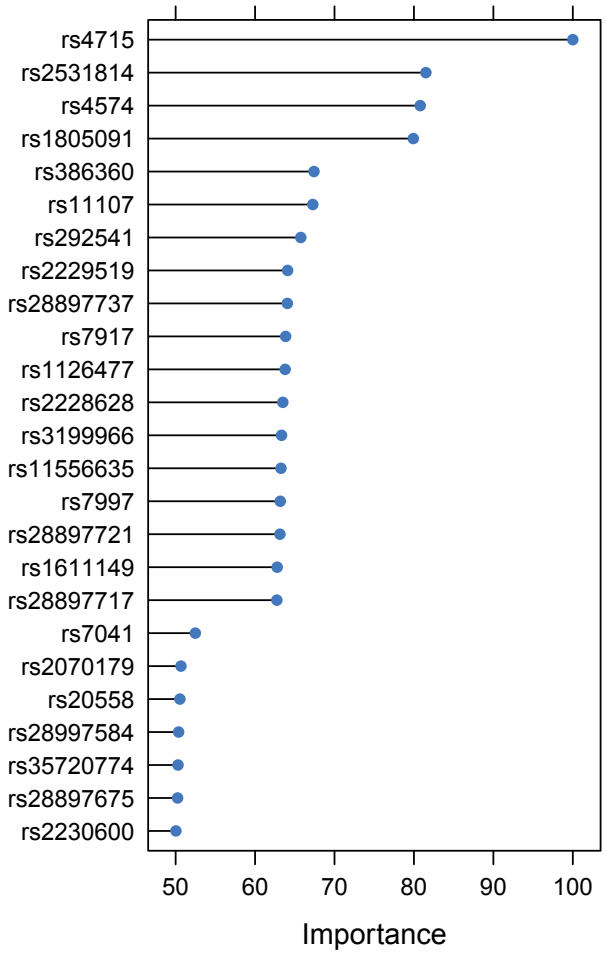


Figure 6.4: Top 20 important SNPs based on mean decreases Gini-index extracted when random forest was ran on the combined dataset of ALL patients and healthy control SNPs. dataset

Training dataset tuning

Selecting the right amount of training and test data is necessary for classification methods, and distribution of data depends on the complexity of model and dataset

(Bickel et al., 2007). The dataset was run through default parameters of the random forest with 100 trees on 70% training set and 30% test set. The results show kappa value at 0.6450 on the training set and 0.6985 on the test set with confusion matrix correctly predicting 26 cases of ALL patients out of 43 patients. A confusion matrix of the results can be seen in Table 6.2.

In next step, the percentage of training set was changed to 80% and test set to 20%. The result shows that kappa value on the training set is 0.6483 while on test set 0.6575, which is lower than 70% training set. The confusion matrix for test set showed that 15 cases were correctly predicted out of 23 ALL patients as shown in Table 6.2. These results suggest that although there is not much difference in performance on the training set, a partition of 30% test set has performed better than a partition of 20% test set as shown in Table 6.4. So this thesis will focus on 70% training set and 30%test set.

Table 6.1: kappa values for different training and test set

	Training	Test
70% training and 30% test	0.6450	**0.6985**
80% training and 20% test	0.6483	0.6575

Table 6.2: Confusion matrix for 30% test set for healthy control and ALL patients respectively, predicted by the random forest method.

	Healthy control	ALL
Healthy control	141	14
ALL	3	26

Table 6.3: Confusion matrix for 20% test set for healthy control and ALL patients respectively, predicted by the random forest method.

	Healthy controll	ALL
Healthy control	95	10
ALL	2	15

Table 6.4: Kappa values for different training and test set

	Training	Test
70% training and 30% test	0.6450	**0.6985**
80% training and 20% test	0.6483	0.6575

Number of trees

Huang and Boutros (2016) has suggested that the parameters in random forest method should be tuned to achieve optimal results. One of the essential parameters in the random forest model is the number of trees selected for each forest (Breiman, 2001). In this section, the focus of the study is to find the best 'number of trees' for the dataset in this thesis. In the first instance, the experiment was run through 70% training and 30% test set with the default number of trees (100). The results achieved 0.6450 kappa value on training set while 0.6985 on the test set. The model correctly classified 26 ALL patients out of 43 total. Oshiro et al. (2012) have found that increasing the number of trees can increase the accuracy of the model.

In the next step, the number of trees was increased to 200, 500 and 1000 with leaving the rest of the parameters the same. The kappa value achieved on training

set were 0.6893, 0.6773 and 0.6726 respectively. The kappa accuracy on the test was 0.5578, 0.9372 and 0.692 respectively.

The confusion matrices of all three experiments show that with 200 trees, only 23 out of 43 ALL patients were correctly classified, while for 500 trees the model has correctly classified 39 out of 43 ALL patients as shown in Tables 6.5 and 6.6. For 1000 trees the correctly classified patients were also 23 ALL patients (as shown in table 6.7) which suggest that increasing the 'number of trees' in RF does not always achieve good results, as suggested by Oshiro et al. (2012). Based on these results, 500 trees were used for future analysis.

Table 6.5: Confusion matrix for test data with 200 trees for healthy control and ALL patients respectively predicted by the random forest method.

	Healthy control	ALL
Healthy control	135	22
ALL	4	23

Table 6.6: Confusion matrix for test data with 500 trees for healthy control and ALL patients respectively, predicted by the random forest method.

	Healthy control	ALL
Normal	141	3
ALL	1	39

Table 6.7: Confusion matrix for test data with 1000 trees for healthy control and ALL patients respectively, predicted by the random forest method.

	Healthy control	ALL
Healthy control	145	14
ALL	2	23

Summary of results

After finding the correct data partition and number of trees, results suggest that the random forest method can achieve best results when data is partitioned as 70% training and 30% test with 500 trees on the combined dataset of ALL and healthy control SNPs .

In the next step this experiment extracted the top SNPs based on Gini-index. The importance plot of mean decrease in Gini-index is shown in Figure 6.5. Figure 6.5 shows there are two particularly high important SNPs, rs11147977 and rs299284, and then the importance decreases gradually. this thesis selected the top 14 SNPs for further biological investigation. The cut-off for mean decrease in Gini index was set to 10, considering after that the difference between mean decrease Gini-index is very small between SNPs. A list of the top 14 SNPs can be found in Table 2.

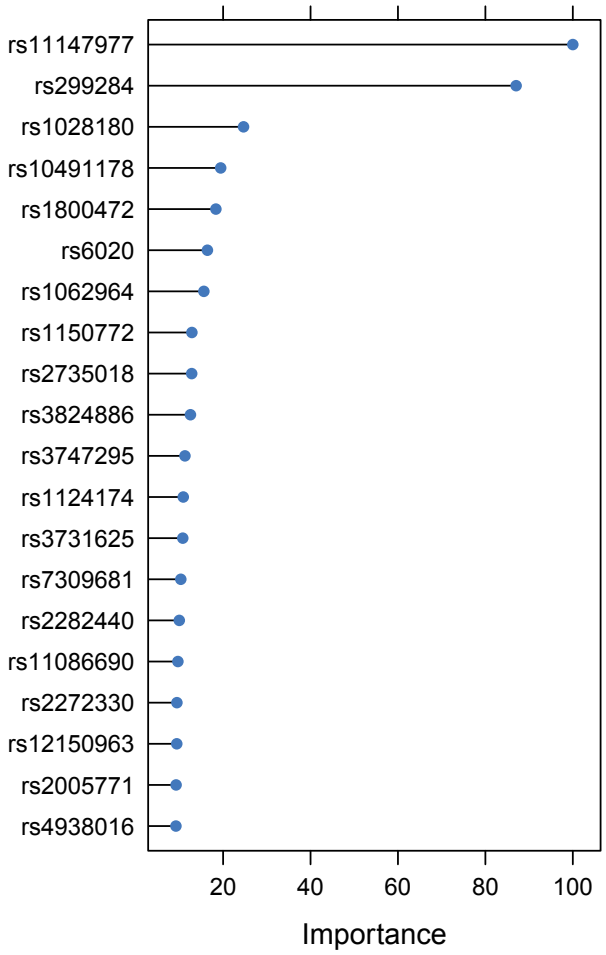


Figure 6.5: Top 20 SNPs selected by random forest from dataset constructed without 15 or more missing values using 500 trees based on mean decrease in Gini-index. Where rs11147977 has achieved a mean decrease in Gini-index of 100, rs4938016 has achieved 9.209. The mean decrease in Gini-index for top 14 SNPs are shown in Table 2.

6.4.2 Finding pathways related to ALL

Further analysis of important SNPs selected by random forest in the ALL dataset show high influence of homozygous SNPs (i.e the minor allele 0 or 2) as shown in Table 6.9. Values '0' through '2' represent the number of minor alleles present in a specific SNP for a patient (as described in Section 6.2) which suggest they are always homozygous for the minor allele. Assié et al. (2008) and Orloff et al. (2012) have described that homozygosity plays an important role in cancer predisposition, which supports the findings of this thesis. The first two most important SNPs, rs11147977 and rs299284, do not contain any '1' in ALL dataset while in the healthy control SNP dataset rs11147977 contains 105 '1's and no '2' values. Similarly rs299284 contains 93 '1's and only 4 '2s as shown in Table 6.10.

The total count of '2' in ALL dataset is high (39 and 37) in rs11147977 and rs299284 compared to the rest of the SNPs as shown in Table 6.9.The detail count of minor alleles for the top 14 SNPs can be found in Tables 6.8. These findings suggest that in childhood ALL, a homozygous genotype for these SNPs are predictive that a patient may get ALL, especially if it is the minor allele 2. Similarly, rs10491178, rs1800472, rs1062964 are also homozygous as they do not contain any '1' values.

Table 6.8: Count of '0', '1', '2' and '-1' for top 14 SNPs in ALL dataset and healthy control dataset.

	0	1	2	-1
ALL	1580	187	174	5
Healthy control	5886	749	11	18

Table 6.9: The count of minor allele in ALL dataset for top 14 SNPs shows that rs11147977 and rs299284 have a higher frequency of '2' minor allele than do the rest of the SNPs.

SNPs	2 minor allele	1 minor allele	0 minor allele
rs11147977	39	0	97
rs299284	37	0	101
rs1028180	10	25	104
rs10491178	19	0	119
rs1800472	6	0	133
rs6020	9	20	110
rs1062964	11	0	128
rs1150772	4	27	108
rs2735018	4	15	120
rs3824886	9	21	109
rs3747295	11	12	116
rs1124174	4	29	106
rs3731625	4	25	110
rs7309681	7	13	119

Table 6.10: The count of minor allele in healthy control dataset for the top 14 SNPs showing that homozygous SNPs have a high frequency of 1 minor allele and low number of 2 minor allele.

SNPs	2 minor allele	1 minor allele	0 minor allele
rs11147977	0	105	369
rs299284	4	93	379
rs1028180	0	7	469
rs10491178	2	46	428
rs1800472	0	16	460
rs6020	0	3	473
rs1062964	1	53	419
rs1150772	1	115	352
rs2735018	0	93	379
rs3824886	0	44	432
rs3747295	0	14	462
rs1124174	0	93	383
rs3731625	2	67	406
rs7309681	1	0	476

Further analysis of the SNP 'rs11147977' revealed its association with the gene SKA3, spindle and kinetochore associated complex subunit 3. SKA3 is a component of the spindle and kinetochore associated complex, which plays an important role in cell division and chromosome segregation (Schmidt et al., 2012). Chromosome segregation, a process in which two sister chromatids segregate from each other during

meiosis, is involved in leukemia (Bogdanov, 2009). SKA3 has also been involved in diseases like breast cancer (Jiao et al., 2013) and prostate cancer (Lee et al., 2015).

The SNP 'rs299284' is related to gene HMMR, hyaluronan-mediated motility receptor. HMMR is involved in cell motility, the ability of cells to move, which helps in muscle contraction and wound healing (Mitchison and Cramer, 1996). HMMR is involved in breast cancer (Akenteva et al., 2015) and leukemia (Hatfield et al., 2014). A summary of the 14 selected SNPs and associated genes can be found in Table 6.11 and more details are shown in Appendix Table 3.

Table 6.11: Top 14 SNPs found through the random forest method with gene annotation, gene description and diseases associated to them.

SNPs	Genes	Description	Disease
rs11147977	SKA3	Spindle and kinetochore-associated protein 3	Breast cancer (Jiao et al., 2013) and prostate cancer (Lee et al., 2015)
rs299284	HMMR	Hyaluronan mediated motility receptor	Breast cancer (Akenteva et al., 2015) and leukemia cancer (Hatfield et al., 2014)
rs1028180	BLZF1	Golgin-45 and basic leucine zipper nuclear factor 1	Leukemia (Duprez et al., 1997)
rs10491178	ABCA10	ABC transporter A family member 10	
rs1800472	TGFβ 1	Transforming growth factor beta-1	Camurati-Engelmann (Janssens et al., 2000)
rs6020	F5	Coagulation factor V	
rs1062964	TACSTD2	Tumor-associated calcium signal transducer 2	Head and neck cancer (Nakanishi et al., 2014)
rs1150772	RPP21	Protein subunit of nuclear ribonuclease P	
rs2735018	HLA-G	HLA class I histocompatibility antigen, alpha chain G	
rs3824886	HTATIP2	Oxidoreductase HTATIP2	lung carcinoma (Shtivelman, 1997)
rs3747295	NHS	Nance-Horan syndrome protein	Dental anomalies, and mental retardation (Burdon et al., 2003)
rs1124174	INSIG1	Insulin-induced gene 1 protein	
rs3731625	ITSN2	Intersectin-2	General cancer cells (Rush et al., 2005)
rs7309681	MAPKAPK5	MAP kinase-activated protein kinase 5	Cancer genomes (Greenman et al., 2007)

Pathways related to ALL

Jones et al. (2008) has argued that despite the individual variability in the cancer genome, the changes summarised represent 12 major pathways as being involved in 23 pancreatic tumours. They analysed 80 genes to extract those pathways. this thesis chose top 80 SNPs found through random forest and extract genes related to them. The gene-annotations and description of these genes can be found in Table 4 in the Appendix.

For further analysis metabolic pathways related to these 80 genes were extracted from KEGG metabolic database and Ingenuity pathway analysis (QIAGEN Redwood City).

The pathways directly or indirectly associated to leukemia that were found in this analysis are: chronic myeloid leukemia signaling (Ramsay et al., 2013), Rho signaling (Ellenbroek and Collard, 2007), FAK signaling (Siesser and Hanks, 2006), systemic lupus erythematosus signaling (Flores-Borja et al., 2007), Cdc42 signaling (Su et al., 2005), Ephrin B signaling (Noren and Pasquale, 2007), MAPK Signaling (Wu et al., 2011), p38 MAPK signaling (Yu et al., 2004), HMGB1 signaling (Jia et al., 2014), hepatic fibrosis / hepatic stellate cell (Vick et al., 2009), IL-12 signaling (Oppmann et al., 2000) and production in macrop, TNF signaling pathway (Aggarwal, 2003), TGF-β signaling (Kitisin et al., 2007), protein kinase A signaling (Brazil and Hemmings, 2001), Wnt-catenin signaling (Garcia-Manero et al., 2009), Aryl hydrocarbon receptor signaling (Hayashibara et al., 2003), role of osteoblasts (Taichman, 2005), osteoclasts and cho (Yokota et al., 2010), and T helper cell differentiation pathways (Maillard et al., 2005). A list of these pathways can be found in Table 6.12.

Table 6.12: Pathways found related to leukaemia cancer from top 80 genes.

Pathways related to leukaemia cancer
Chronic myeloid leukemia signaling (Ramsay et al., 2013)
Rho signaling (Ellenbroek and Collard, 2007)
FAK signaling (Siesser and Hanks, 2006)
Systemic lupus erythematosus signaling (Flores-Borja et al., 2007)
Cdc42 signaling (Su et al., 2005)
Ephrin B signaling (Noren and Pasquale, 2007)
MAPK Signaling (Wu et al., 2011)
p38 MAPK signaling (Yu et al., 2004)
HMGB1 signaling (Jia et al., 2014)
Hepatic fibrosis / hepatic stellate cell (Vick et al., 2009)
IL-12 signaling and production in macrop (Oppmann et al., 2000)
TNF signaling pathway (Aggarwal, 2003)
TGF-β signaling (Kitisin et al., 2007)
Protein kinase A signaling (Brazil and Hemmings, 2001)
Wnt -catenin signaling (Garcia-Manero et al., 2009)
Aryl hydrocarbon receptor signaling (Hayashibara et al., 2003)
Role of osteoblasts (Taichman, 2005)
Osteoclasts and cho (Yokota et al., 2010)
T helper cell differentiation pathways (Maillard et al., 2005)

6.4.3 Functional visualisation of genes

For further analysis, gene-set enrichment analysis (as described in Chapter 4) using singular value decomposition(SVD) was performed on the top 80 genes found in the last section. The similarity measure is hop-based (described in Section 4.2.2). SVD allows analysis of genes and terms separately.

When only terms were plotted, the second principal component((PC2) shows separation between three subontologies of GO terms as expected: cellular component, biological process and molecular function, as shown in Figure 6.6. The third principal component(PC3) for terms analysis highlighted two spread molecular function term clusters; cluster A and B as shown in Figure 6.7. Cluster A consists of dehydrogenase activity terms and nicotinamide adenine dinucleotide phosphate activity terms, while cluster B contains transferase activity terms; glucosyltransferase activity, galactosyltransferase activity, fucosyltransferase activity and phenanthrol glycosyltransferase activity. The dehydrogenase activity plays an important role in thyroid tumors Mirebeau-Prunier et al. (2013) and breast cancer (Savolainen-Peltonen et al., 2014) while transferase activity has been involved in hepatitis C (Everhart and Wright, 2013). The full list of terms found in these clusters can been seen in Table 6.13.

Gene based analysis shows a very strong Pearson correlation of 0.999 between the data projected into principal component 1 (PC1) and the number of terms associated with each gene, suggesting that this principal component is a "size" component (Jolliffe, 2004). PC1 also highlighted two outliers, NOS1 and TGFβ 1, as shown in Figure 6.8. These two genes associated with many more GO terms (123 and 234 respectively) than the rest of the genes, where the average number of terms associated to genes is 21.82. This suggests that these genes are very common in many diseases, for example

ALL (Kitisin et al., 2007). After ignoring these two outliers PC3 highlighted clusters A,B,C, D and E as shown in Figure 6.10.

Cluster A consists of six genes: OR3A4, C13orf3, C10orf53, MGC20470, IL17R and C10orf113. The functionality of these genes are unknown to GO, which leads to no terms related to these genes.

Cluster B consists of the genes TACSTD2, tumor associated calcium signal transducer 2 and ATXN7, ataxin 7 genes. These genes are associated with head and neck cancer (Nakanishi et al. (2014) and Mirghani et al. (2014)).

Cluster C contains the genes ABBC5, ATP-binding cassette subfamily CFTR (cystic fibrosis transmembrane conductance regulator) member5 and SDC3, syndecan3. Gene ABBC5 has been involved in treatment of ALL (Krajinovic et al., 2015) while SDC3 have been found in myeloma leukemia (Fadnes et al., 2012).

Cluster D consists of the genes FGD6, FYVE, RhoGEF And PH domain containing 6 and MYOM1, Myomesin 1. These genes have been related to breast cancer (Fan et al. (2014) and Lee et al. (2013)).

Cluster E is more spread than the other clusters consisting of three genes. WASF3, a member of the Wiskott-Aldrich family of proteins, CDH26, Cadherin 26, and ANKRD6, Ankyrin Repeat Domain 6. CDH26 has been involved in many cancers such as leukemia (Croce and Calin, 2012) and breast cancer (Haverty et al., 2008) while WASF3 and ANKRD6 are associated to breast cancer (Cvetković et al. (2013) and Sato et al. (2013)). The summary of these clusters is presented in Table 6.14.

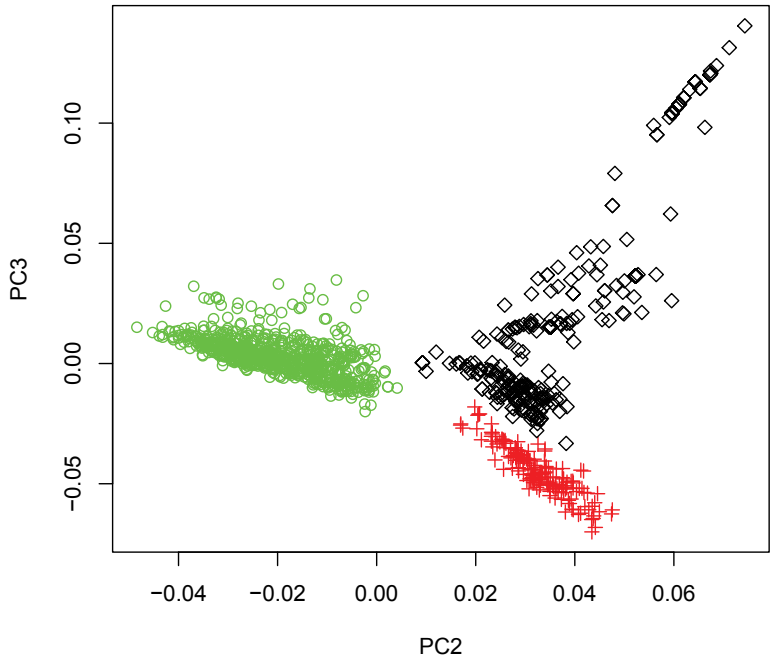


Figure 6.6: GO terms projected into PC2 and PC3 form clusters associated with the sub-ontology. Legend: ◇ (black) is the molecular function GO term, ○ (green) is the biological process GO term, and + (red) is the cellular component term

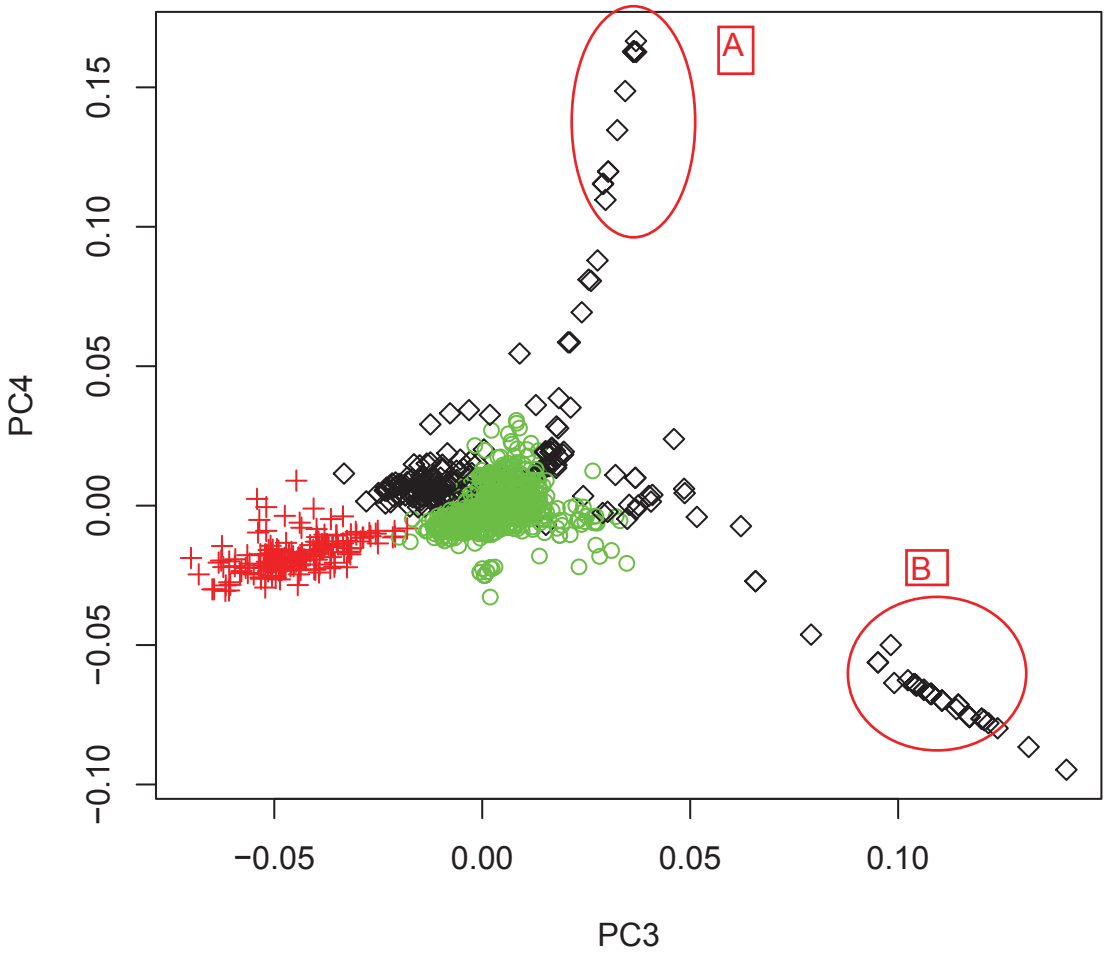


Figure 6.7: Two 'arms' of the distribution of molecular function terms found during analysis of terms. The details of these clusters is shown in Table 6.13. Legend: ⋄ (black) is the molecular function GO term, ∘ (green) is the biological process GO term and + (red) is the cellular component term

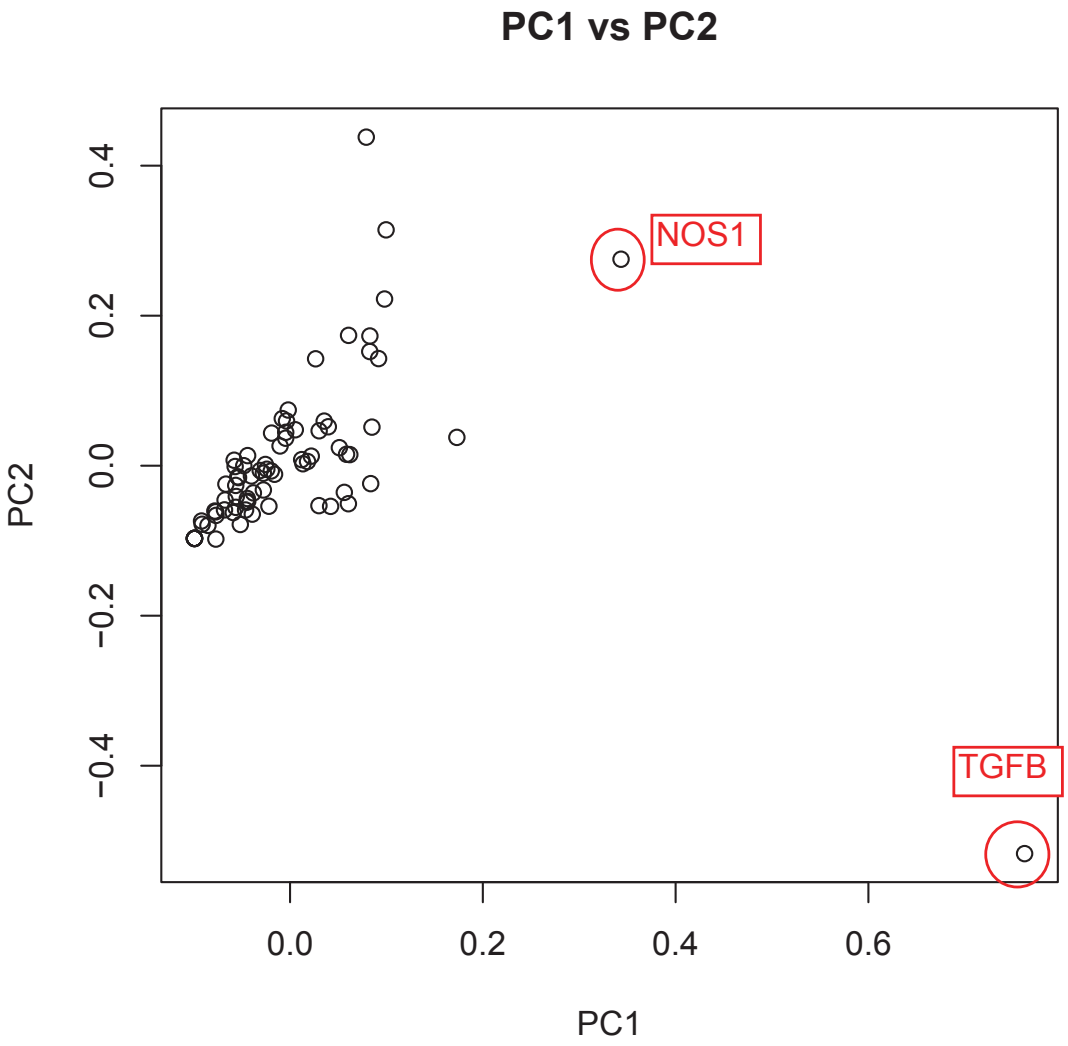


Figure 6.8: Two outliers 'TGFβ 1' and 'NOS1' found through gene-set enrichment analysis. Legend: ∘ is gene.

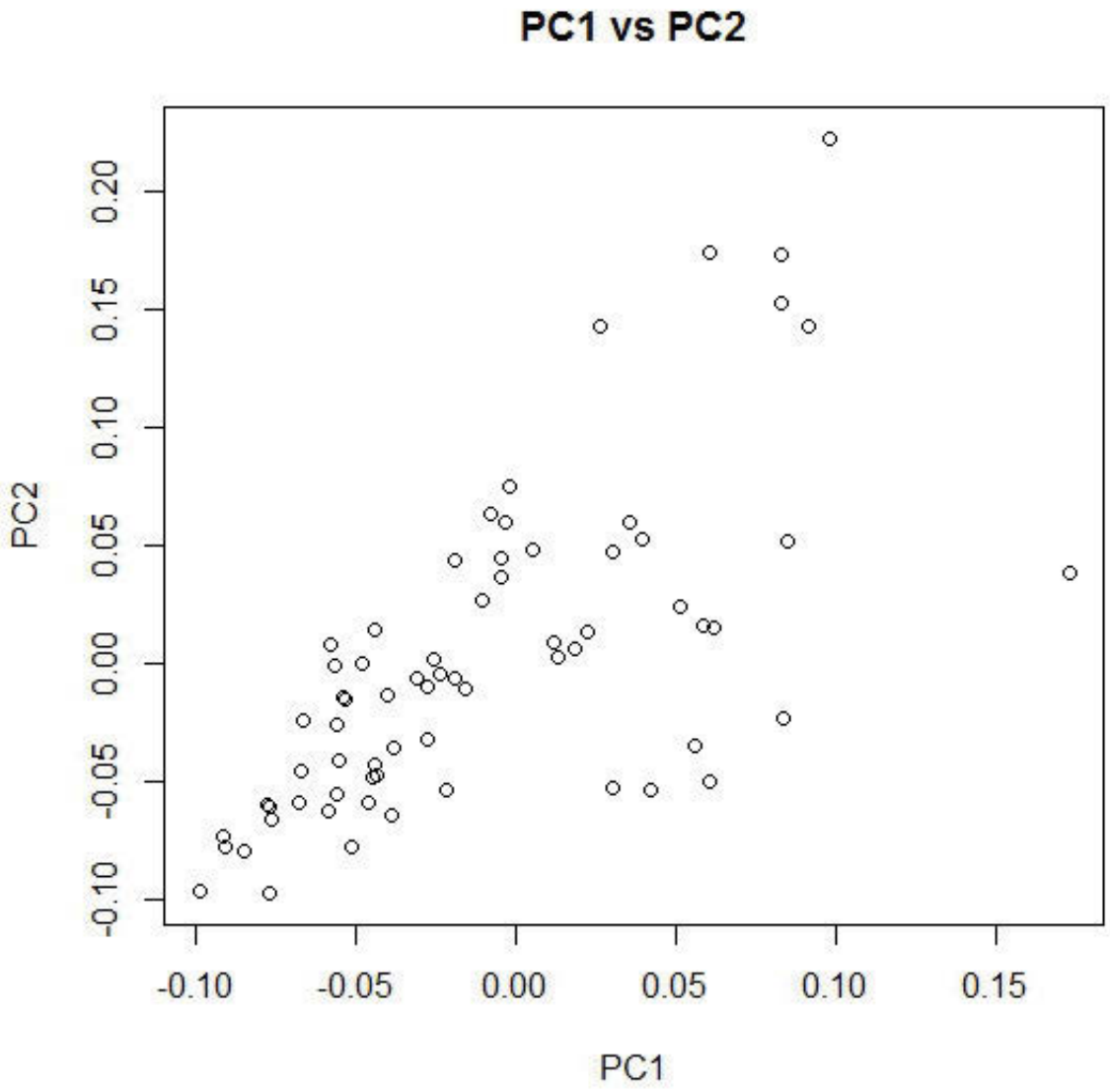


Figure 6.9: Plot of PC1 vs PC2 where PC1 shows high correlation to number of terms associated to genes. Genes on the bottom left are associated with the lowest number of terms and genes on the far right are associated with the highest number of terms. Legend: ∘ is gene.

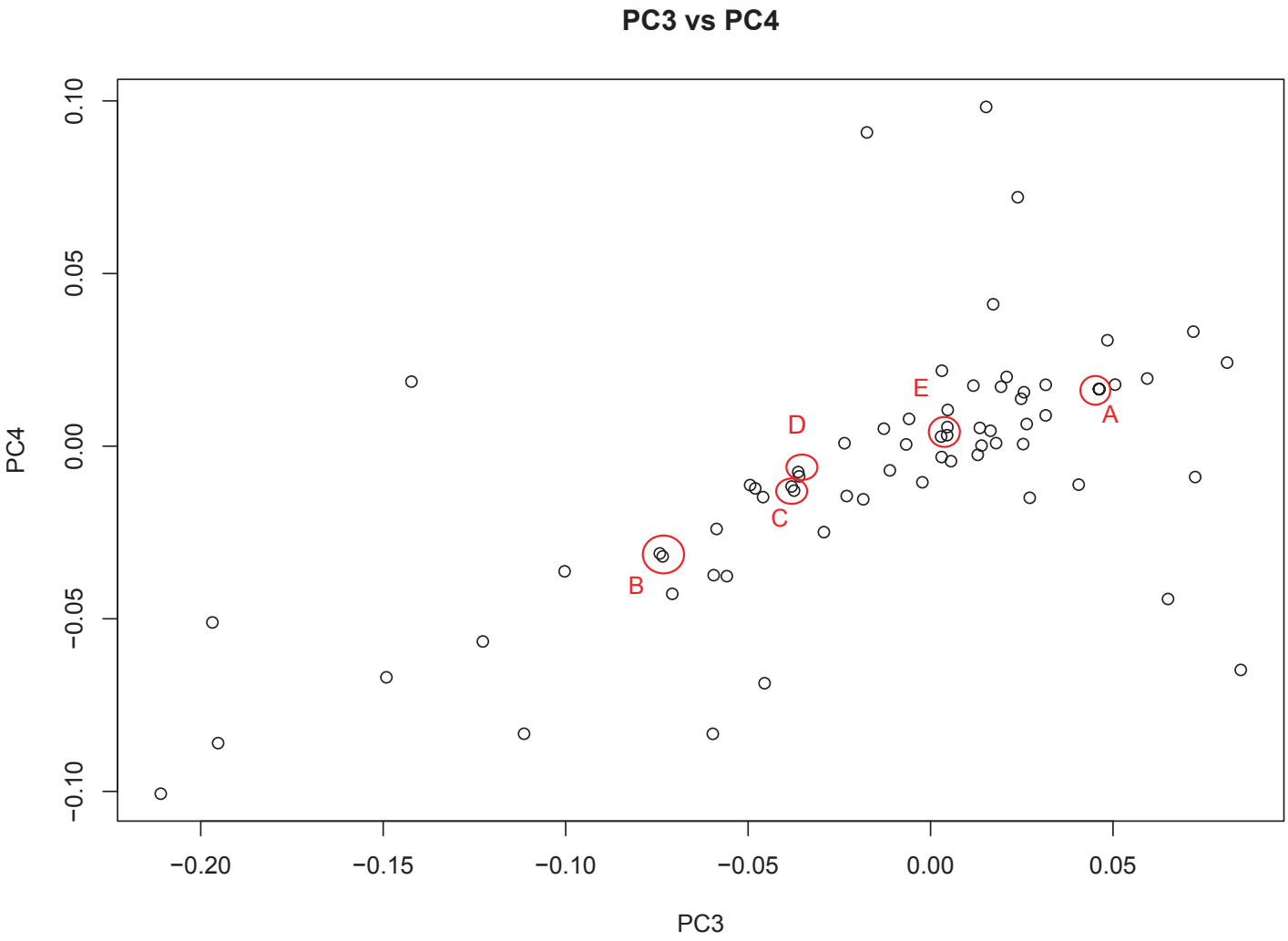


Figure 6.10: Clusters A, B, C, D and E were found during analysis of PC3 and PC4. The detail of these clusters is presented in Table 6.14. Legend: ○ is gene.

Table 6.13: GO Terms found in clusters A and B during analysis of PC3 and PC4 as shown in Figure 6.7.

Clusters	GO Terms	Disease
A	GO:0016229:Steroid dehydrogenase activity, GO:0051990:(R)-2-hydroxyglutarate dehydrogenase activity, GO:0004303: Estradiol 17-beta-dehydrogenase activity, GO:0035410: Dihydrotestosterone 17-beta-dehydrogenase activity, GO:0003857: 3-hydroxyacyl-CoA dehydrogenase activity, GO:0033764/Steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor, GO:0000252: C-3 sterol dehydrogenase (C-4 sterol decarboxylase) activity	Thyroid tumors (Mirebeau-Prunier et al., 2013) Breast cancer (Savolainen-Peltonen et al., 2014)
B	GO:0046527: Glucosyltransferase activity GO:0000030: Mannosyltransferase activity GO:0008378: Galactosyltransferase activity GO:0008417: Fucosyltransferase activity GO:0019112: Phenanthrol glycosyltransferase activity GO:0008375: Acetylglucosaminyltransferase activity GO:0004583: Dolichyl-phosphate-glucose-glycolipid alpha-glucosyltransferase activity GO:0003980: UDP-glucose:glycoprotein glucosyltransferase activity GO:0018715: 9-phenanthrol UDP-glucuronosyltransferase activity GO:0001962: alpha-1,3-galactosyltransferase activity	Hepatitis C (Everhart and Wright, 2013)

Table 6.14: Gene clusters found through gene-set enrichment analysis using the top 80 genes found through the random forest method as shown in Figure 6.10.

Clusters	Genes	Disease
A	OR3A4, C13orf3, C10orf53, MGC20470, IL17R and C10orf113	Functionality unknown to GO
B	TACSTD2 and ATXN7	Head and neck cancer
C	ABBC5 and SDC3	Myeloma leukemia cancer
D	FGD6 and MYOM1	Breast cancer
E	WASF3, CDH26 and ANKRD6	Leukemia and breast cancer

6.5 Summary

In this chapter, the random forest feature selection method was applied to a high-dimensional SNP dataset to find metabolic pathways related to ALL. The aim of this thesis was to show that the random forest method can be used for selection of influential SNPs from SNP datasets. The dataset used in this chapter was constructed from two sources, an ALL SNP dataset from The Children's Hospital at Westmead and a healthy control SNP dataset from the Wellcome Trust UK. The combined dataset was then run through the random forest method using different parameters. The genes and pathways related to the most influential SNPs selected by the random forest method were then further investigated. The pathways retrieved from important SNPs were then also compared with pathways defined by Jones et al. (2008) as the

major pathways in 23 pancreatic cancers.

The results show that the random forest method selected two SNPs, rs11147977 and rs299284 as the most important SNPs compared to the rest of the dataset as shown in Figure 6.5. Further investigation of these SNPs indicated that they are totally homozygous SNPs. Assié et al. (2008) and Orloff et al. (2012) have described that homozygosity plays important role in cancer predisposition, which supports the findings of this thesis. It was found that 'rs11147977' is related to SKA3 , a component of the spindle and kinetochore associated complex which plays an important role in cell division and chromosome segregation, is involved in leukemia (Bogdanov, 2009).

The SNP 'rs299284' is related to gene HMMR, hyaluronan-mediated motility receptor. HMMR is involved in cell motility and is is also involved in leukemia (Hatfield et al., 2014). This chapter then selected the top 14 SNPs by mean decrease in Gini index for further biological investigation. The cut-off for mean decrease in Gini-index was set to 10, because after that the difference between mean decrease in Gini-index is very small between SNPs. A list of top 14 SNPs can be found in Table 2. The genes were retrieved for the top 14 SNPs and it was found that some of these genes were related to leukemia, breast, prostate and other cancers as shown in table 6.11.

Jones et al. (2008) have highlighted 12 pathways based on 80 genes that play an important role in 23 pancreatic tumours. In the next step of this thesis, genes were retrieved related to the top 80 SNPs identified by the random forest method and pathways were extracted from KEGG and Ingenuity Pathways Analysis databases related to those 80 genes. The list of pathways shown in Table 6.12 include chronic myeloid leukemia signaling, Rho signaling, FAK signaling, systemic lupus erythematosus signaling, Cdc42 signaling, Ephrin B signaling, MAPK Signaling, p38 MAPK signaling,

HMGB1 signaling, hepatic fibrosis / hepatic stellate cell, IL-12 signaling and production in macrop, TNF signaling pathway, TGF-β signaling, protein kinase a signaling, Wnt -catenin signaling, aryl hydrocarbon receptor signaling, role of osteoblasts, osteoclasts and CHO and T helper cell differentiation pathways, which are related to ALL.

For further investigation, functional visualisation of the top 80 genes was performed using SVD as described in Chapter 3. The results identified two outliers, TGFβ 1 and NOS1, which are related to high numbers of GO terms (234 and 123), compared to rest of the genes where the average number of terms associated to genes is 21.82. The later principal components identified 5 small clusters of genes (A, B, C, D and E) as shown in Figure 6.10. Cluster A consists of genes whose functionality is unknown to GO, cluster B consists of genes related to head and neck cancer, cluster C of genes associated with myeloma leukemia, cluster D consists of genes related to breast cancer and cluster E genes are associated to leukemia and breast cancer. Further details of these clusters can be found in Table 6.14.

This chapter achieves Objective 3: develop, implement and evaluate a novel feature selection approach to identify related metabolic pathways in ALL, as described in Section 1.1 and Thesis Contribution 3 in Section 1.4. In the next chapter the thesis will be concluded.

Chapter 7

Conclusion

The objective of this thesis is to use data mining and machine learning approaches to propose, develop, implement and evaluate methods and techniques for the modelling of different aspects of a complex disease (ALL) based on genome-wide SNP data, gene expression data and gene annotation data.

The work presented in this thesis focuses on three main aspects of investigating the difficulties in biological data studies.

(i) A novel framework proposed, implemented and evaluated to find functional relationships between genes from gene-annotation data.

(ii) Identified an optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression.

(iii) A novel feature selection approach proposed, implemented and evaluated to identify related metabolic pathways in ALL

All these tasks can help in disease diagnosis and prognosis. The novelty of the proposed approaches in this thesis are based on the data mining and visualisation methods including, feature selection and machine learning techniques.

A primary goal of data mining and visualisation studies of biomolecular data is to identify different aspects of biological function that can contribute to diagnosis and prognosis of complex diseases such as ALL. The focus of Chapter 4 was to implement and evaluate a framework to retrieve features from Gene Ontology (GO) and apply singular value decomposition to find functional relationships between genes. Using this approach, clusters of genes and their annotations (from GO) can be displayed in the same visualisation.

Similar to gene annotation data, gene expression data from ALL patients can be used to model disease diagnosis and prognosis as described in Section 2.2.1. In Chapter 5, ALL patients were classified between relapsed and non-relapsed through dimensionality reduction methods using gene expression dataset. The purpose of the study is to provide a visualisation of patients based on their relapse status in a low-dimensional space. There are many linear, and non-linear methods available (Lee and Verleysen, 2007). The focus of Chapter 5 was to find an optimal method from selected linear and non-linear dimensionality reduction methods to classify ALL patients based on an ALL gene expression dataset.

To find subsets of features which have predictive power to find pathways related to ALL in SNP data, a feature selection method is used, and those selected SNPs are then used to find pathways related to ALL. The following sections are a summary of previous chapters regarding the contributions of the thesis.

7.1 Functional relationship between genes

The critical question that comes to mind in dealing with biological data is the type of data available to the biologist at the time of investigation: if gene annotations are

available then this thesis can help them find functional relationships between genes. This experiment answered the Research Question 1, develop, implement and evaluate a novel computational framework based on data mining and visualisation methods to find relationships between ALL genes using gene annotation as described in Section 1.3. Although there are many tools available to find functional relationships between genes, Huang et al. (2008) and recently Wagholikar et al. (2012) have put emphasis on new tools required in this field. The framework proposed in Chapter 4 is based on a binary matrix, as discussed in Section 4.2, that shows a relationship between genes and Gene Ontology terms, and a proximity matrix that finds similarity between Gene Ontology terms. A visualisation method (singular value decomposition) is used to visualise genes and their features in details. This work will help biologists in finding functional relationships between genes which eventually lead them to improve diagnosis and prognosis of diseases.

The framework in Chapter 4 (discussed in Section 4.2) also compared different datasets using different similarity measures. The proposed method will contribute to knowledge by helping biologist to run datasets through different similarity measures to find functional relationships between genes which eventually allows an investigator to check which similarity measure performs better for his dataset.

In Chapter 4, singular value decomposition was applied to gene-annotation data augmented with GO terms and inter–term similarities. Two datasets were visualised: validation data from KEGG and a set of genes related to ALL as described in Section 4.2.1. Results showed that Principal Component(PC) 1 measured the number of terms associated with genes as shown in Section 4.3. Later components allowed visualisation of genes according to their functional information, but the meaning of

PCs varied depending on the underlying genes. For the KEGG data, PCs described gene functionality as expected by separating gene classes based on their functionality as shown in see Section 4.3.

For the larger cancer dataset, the early PCs identified known GO sub-ontologies: molecular function, biological process and cellular component, as discussed in Section 4.3.1. Visualisation of terms from the individual sub-ontologies found informative clusters, as shown in Section 4.3.1. The correlation between GO terms and PCs improved understanding of the functional meaning of the PCs. These results show that approach described in Chapter 4 can bring meaningful biological interpretation to gene lists. Users should not expect that the meaning of the PCs should generalise from one gene list to another, apart from gross patterns such as the sub-ontologies. This is because different sets of GO terms will be associated with different lists of genes and SVD will focus on those that explain the most variance. In practice, this approach should be applied to specific gene lists of interest to explore the functional characteristics of those specific genes.

The results in this chapter of finding functional relationships between genes using different similarity measures and visualisation address Objective 1 in Section 1.1 and Thesis Contribution 1 in section 1.4.

7.2 Visualising leukaemia cancer dataset using dimensionality reduction methods

Another type of data is gene expression data. With gene expression data comes the curse of high dimensionality (Clarke et al., 2009). A patient can have hundreds of

thousands of gene expression values. Lee and Verleysen (2007) have classified different available dimensionality reduction methods into linear and non-linear methods, and this thesis explores the question of which method performs better for classification of ALL patients using gene expression data. The experiment in Chapter 5 answered Research Question 2: Is there a way to find which dimensionality reduction method performs better for ALL gene expression high-dimensional data, to eventually help in finding patient-to-patient relationships?, as described in Section 1.3.

Information visualisation is regarded as a direct way to help browse datasets (Chen, 2013). Results from information visualisation can be used to assist clinicians and biomedical researchers in understanding the different structure of patients and to compare different clusters in a visualisation. The main challenge in visualising gene expression datasets stems from the high dimensionality of the data. To deal with large amounts of gene expression data, different dimensionality reduction methods were applied to gene expression data of ALL patients to determine the best method for visualizing this type of data. Visualisation approaches were compared based on area under the curve.

In Chapter 5, the best dimensionality reduction method is found between Principal Component Analysis (PCA), kPCA, Local Linear Embedding (LLE), Stochastic Neighbour Embedding (SNE) and Diffusion Maps. Each of these methods finds different structures in high-dimensional data as discussed in Section 2.3.1. The parameters in these methods were tuned to find the optimal method to classify relapsed and non-relapsed ALL patients. The classification was performed using SVM while the classification accuracy was assessed using area under the curve (AUC) as discussed in Section 5.4. The comparison is performed on a clinical ALL dataset. Haghverdi

et al. (2015) has also urged comparison between dimensionality reduction methods to find optimal techniques.

this thesis contributed to knowledge in two ways: First, by showing the comparison between different linear and nonlinear dimensionality reduction methods and determining the optimal method among them for ALL gene expression data.

The visualisations of SNE, kPCA, LLE and PCA showed varying levels of success for genetic distinction between relapsed and non-relapsed patients as shown in section 5.4. The separation of patients from both classes (class represents relapse or non-relapse) were apparent, but some patients were overlapping. Interestingly, two pairs of patients, GSM180185, GSM180188, and GSM180178 and GSM180179, were highlighted in SNE, kPCA, LLE and PCA. Both pairs clustered separately in these visualisations. The gene expression profiles for these patients were investigated further. Patients GSM180185 and GSM180188 have very similar gene expression profiles, except GSM180188 mainly differs from GSM180185 based on age, sex and white blood cells as discussed in Section 5.4.2. Similarly, the comparison between GSM180178 and GSM180179 shows similar gene expression values except the white blood cell count, in which GSM180179 has a higher number of white blood cells which may have caused the patient to relapse.

The results in Chapter 5 have demonstrated that SNE, kPCA, LLE have achieved higher classification accuracy than DM for classification of relapsed patients in the ALL dataset, and hence would result in improved classification of ALL classes. This analysis addresses Objective 2, identifying an optimal dimensionality reduction method to classify between relapsed and non-relapsed ALL patients using gene expression, as described in Section 1.1 and Thesis Contribution 2 in Section 1.4.

7.3 Case Study: Finding pathways related to ALL using random forest

Finding pathways using SNPs data has been an active area of study in the past (discussed in Section 3.3). In a recent review, Jin et al. (2014) suggested that there is a need of weighting scheme for SNP profiles which leads to finding informative genes and pathways. following challenges in pathways analysis.

In Chapter 6, a robust framework was implemented on an ALL SNP dataset using the feature selection method ' Random Forest (RF)' to find important SNPs based on mean decrease in Gini-index. The aim of this thesis was to show that the random forest method can be used for selection of influential SNPs from SNP datasets. The dataset used in Chapter 6 was constructed from two sources: an ALL SNP dataset from The Children's Hospital at Westmead, and a healthy control SNP dataset from the Wellcome Trust UK.

The results showed that the high importance SNPs based on mean decrease in Gini-index were related to the homozygous genotype as described in Section 6.4.2. A list of the top 14 SNPs is presented in Table 2. Associated genes were retrieved for the top 14 SNPs, and it was found that some of these genes were related to leukaemia, breast, prostate and other cancers as shown in Table 6.11.

Furthermore, a case study is performed based on the hypothesis suggested by Jones et al. (2008), that 12 major pathways are involved in 23 pancreatic cancers. They used 80 genes for their study, so following their lead the top 80 genes were selected for pathway analysis as discussed in Section 6.4.2. The retrieved pathways identified pathways related to ALL as shown in Table 6.12.

For further investigation, functional visualisation of the top 80 genes was performed using SVD as described in Chapter 4. The results identified two outliers related to high numbers of GO terms compared to the rest of the genes as shown in Section 6.4.3 . The later principal components identified 5 small clusters of genes (A, B, C, D and E) as shown in Figure 6.10. Cluster A consists of genes whose functionality is unknown to GO, cluster B consists of genes related to head and neck cancer, cluster C genes associated with myeloma leukaemia, cluster D consists of genes related to breast cancer and cluster E genes are associated to leukaemia and breast cancer. Further details of these clusters can be found in Table 6.14.

This case study contributed to knowledge by introducing a novel framework based on random forest to rank SNPs based on importance of mean decrease in Gini-index, and extract metabolic pathways using KEGG database. Chapter 6 achieves Objective 3: develop, implement and evaluate a novel feature selection approach to identify related metabolic pathways in ALL as described in Section 1.1 and Thesis Contribution 3 in Section 1.4. .

7.4 Limitations and Future suggestions

This thesis provides novel frameworks to visualise gene annotations, gene expression data and SNP data. However there are three limitations with these techniques that should be addressed in future research.

7.4.1 Limitations

(i) Scaling for increased data: Biological data is increasing in size and scale on a daily basis, as is the information in public ontologies (Gene Ontology and KEGG). The frameworks proposed in Chapter 4 and Chapter 6 rely heavily on information from these databases. With the increase in information, it will be interesting to see how these frameworks cope at runtime in future with huge lists of genes (for example 10K genes annotations). One way to resolve this limitation is construct a large gene-annotation dataset as a validation dataset, related to different diseases and test these frameworks to find whether genes are separated based on their association with diseases. With time, that dataset will be tested again and again.

(ii) Although the framework in chapter 6 handles missing values using the mean value of the attributes i.e SNPs, there is still room for improvement in finding a better way to handle missing values, especially when the dataset is based on minor-allele frequency where values are 0, 1 and 2. One way to resolve this limitation may be to use most frequent value to handle missing values in categorical datasets.

(iii) Public ontologies are still lacking information related to SNPs and genes. Many genes do not have pathways associated with them. This limitation is likely to change over time, as the study in Chapter 6 could provide more comprehensive results if there were pathways available related to every genes. There is a possibility that there may be an ontology which can provide information for missing pathways from the KEGG pathway ontology. One way to resolve this

limitation is to create a centralised ontology search engine. Such a search engine would allow users to retrieve information from all available ontologies through sql queries. The pathways related to a gene will be retrieved from multiple ontologies and duplicate results would be removed before presentation to users.

7.4.2 Future directions

The work provided in this thesis answers a specific set of research gaps in the area of data mining in bioinformatics. However, in future, this work can be developed in the following ways.

(i) One of the major tasks in the future is to apply the described frameworks in Chapter 4, Chapter 5 and Chapter 6 on datasets of other diseases. The focus of this thesis was on ALL, but biological data is available for other diseases such as lung cancer, breast cancer and many more, which can be used with the frameworks described in chapters 4, 5 and 6. This should preferably be done using a dataset of gene-annotation, gene expression and SNPs gathered from the same patients, so that the framework implemented in this thesis can be tested on the same patients' dataset.

(ii) Developing a data mining and visualisation tool where the proposed, implemented and evaluated methods in chapters 4, 5 and 6 will be available to biologists for real-time use, allowing researchers to analyse gene-annotation, gene-expression and SNP data related to any disease. The tool will provide biological data analysis using many dimensionality reduction, feature selection, validation and visualisation techniques.

The tool can be a web-based tool but for big data analysis it will be challenging, a desktop based application would provide a better solution for this task. Lee and Verleysen (2007) categorised many dimensionality reduction methods; such a tool could provide all of them and also classification methods like support vector machines and decision trees. Similarly, visualisation functionality can be provided by providing different types of plots, scatter, histogram and box plots. This would be a comprehensive tool that will provide all types of functionality related to any kind of data analysis.

(iii) Supek and Škunca (2016) compared gene-set enrichment analysis methods on a standard list of GO terms. The framework implemented and evaluated in Chapter 4 will be tested on their list of terms, to compare the framework with already existing tools on a standard dataset, in order to enhance the reliability of the framework. There is a need for a standard dataset and standard validation method that can be used by researchers to compare their model with other available tools.

(iv) Also in the future, the validation of experiments in Chapter 4, Chapter 5 and Chapter 6 will be applied through cross-validation and 100xV-fold cross validation (Jayawardana et al., 2015).

(v) In Chapters 6, Pathway information was retrieved from KEGG database. In the future, other databases such as BIOCRATA will be used to find the pathways related to ALL.

Finally, microarray and high throughput technology has allowed for an individual patient's genome to be sequenced rapidly. This data can be very high-dimensional,

biased, nonlinear and gathered in multiple forms such as gene-expression, gene annotations and Single Nucleotide Polymorphisms (SNPs). this thesis proposed, implemented and validated data analysis frameworks to handle complex genomic data associated with childhood cancer derived from an individual patient. The data collected from individual patients is high-dimensional data, biased, non-linear and contains missing values. This thesis implemented and validated robust strategies to handle three different types of high-dimensional biological data: gene annotations, gene expression and SNPs ALL datasets, around principles in data mining, visualisation. This work provided novel strategies to visualise the functional relationship between genes using ALL gene annotation data, classify ALL patients using gene expression profiles and find active metabolic pathways in ALL using SNPs profiles that can help biologists in diagnosis and prognosis of ALL.

Appendix

Table 1: The description of patient datasets described in Section 5.2.1 where RER is 'rapid early responder', CCR is 'complete continuous remission' and SER is 'slow early responder'. Class is assigned based on relapsed and non-relapsed patients. WBC represents 'white blood cell' count.

Sample	**Sex**	**Age(months)**	**WBC(109/L)**	**Translocation**	**Description**	**Class**
GSM180132	Male	46	59300	t(1;19)	RER	1
GSM180133	Female	134	8700	t(4;11)	RER;CCR	1
GSM180134	Male	25	70000	t(12;21)	SER	1
GSM180135	Male	170	732000	t(12;21)	RER	1
GSM180136	Female	208	314600	t(9;22)	SER	1
GSM180137	Female	161	2100	t(1;19)	RER	1
GSM180138	Female	132	99500	t(1;19)	SER	1
GSM180139	Male	118	66700	t(1;19)	RER	1
GSM180140	Female	19	71500	t(1;19)	SER	1
GSM180141	Female	27	65200	t(1;19)	SER	1
GSM180142	Female	198	44580	t(9;22)	RER	1
GSM180143	Female	16	161700	t(9;22)	SER	1
GSM180144	Male	61	65800	t(12;21)	RER	1
GSM180145	Male	187	2950	t(12;21)	SER	1
GSM180146	Male	202	68000	t(12;21)	RER	1
GSM180147	Male	80	144000	t(12;21)	RER	1
GSM180148	Female	18	64100	t(9;22)	SER	1
GSM180149	Female	31	92800	t(4;11)	RER	1
GSM180150	Female	171	9800	t(1;19)	RER	1
GSM180151	Male	177	61800	t(12;21)	SER;relapse	2
GSM180152	Female	129	2500	t(12;21)	RER	1
GSM180153	Female	209	4000	t(9;22)	RER	1
GSM180154	Male	133	8000	t(4;11)	RER	1
GSM180155	Female	134	30700	t(9;22)	RER	1
GSM180156	Male	144	15600	t(1;19)	SER	1
GSM180157	Female	191	10500	t(1;19)	RER	1
GSM180158	Male	34	84700	t(4;11)	RER	1
GSM180159	Female	39	97000		SER	1
GSM180160	Female	16	191000		RER	1
GSM180162	Male	158	50000		SER	1
GSM180163	Female	117	300000		SER	1
GSM180164	Female	14	279000		RER	1
GSM180165	Male	213	68300		SER	1
GSM180166	Female	39	64900		SER	1
GSM180167	Male	66	88100		SER	1
GSM180168	Female	150	34400		RER	1
GSM180169	Male	154	54000		SER;relapse	2
GSM180170	Female	50	76400		RER;CCR	1

Continued ..

Sample	Sex	Age(months)	WBC(109/L)	Translocation	Description	Class
GSM180171	Female	136	50300		SER;CCR	1
GSM180172	Female	125	6900		RER;CCR	1
GSM180173	Male	126	10700		SER;relapse	2
GSM180174	Female	129	87600		SER;relapse	2
GSM180176	Male	177	45900		SER;CCR	1
GSM180177	Female	41	90800		SER;CCR	1
GSM180178	Male	167	4400		SER;CCR	1
GSM180179	Male	176	93500		RER;relapse	2
GSM180180	Male	52	165000		SER;relapse	2
GSM180181	Male	70	121500		SER;relapse	2
GSM180182	Male	39	86700		RER	1
GSM180183	Male	109	253100		RER;relapse	2
GSM180184	Male	128	68100		RER;relapse	2
GSM180185	Male	158	164000		SER;relapse	2
GSM180186	Male	193	28000		RER;relapse	2
GSM180187	Female	185	1800		RER;relapse	2
GSM180188	Female	41	64500		RER;CCR	1
GSM180189	Male	83	65000		SER;CCR	1
GSM180190	Female	29	178000		SER;CCR	1
GSM180191	Male	139	9440		RER;CCR	1
GSM180192	Male	101	58700		RER;relapse	2
GSM180193	Male	40	106000		SER;CCR	1
GSM180194	Male	225	262800		SER;relapse	2
GSM180195	Female	125	12600		RER;CCR	1
GSM180196	Male	12	189300		SER;relapse	2
GSM180197	Male	135	71670		SER;relapse	2
GSM180198	Female	109	672000		SER;CCR	1
GSM180199	Male	189	82400		RER;relapse	2
GSM180200	Male	199	6000		RER;CCR	1
GSM180201	Female	116	158000		SER;relapse	2
GSM180202	Male	191	91800		RER;relapse	2
GSM180203	Male	183	36000		RER;CCR	1
GSM180204	Male	188	303900		SER;relapse	2
GSM180205	Female	126	165900		RER	1
GSM180206	Male	169	4600		RER;CCR	1
GSM180207	Male	106	271700		SER;relapse	2
GSM180208	Male	138	9900		RER	1
GSM180209	Male	80	55000		SER	1
GSM180210	Male	214	17900		RER	1

Continued ...

Sample	**Sex**	**Age(months)**	**WBC(109/L)**	**Translocation**	**Description**	**Class**
GSM180211	Female	158	6700		SER	1
GSM180212	Male	153	62800		SER;relapse	2
GSM180213	Male	75	51100		SER;CCR	1
GSM180214	Male	191	31100		RER;CCR	1
GSM180215	Female	19	138550		CCR	1
GSM180216	Male	179	4250		Relapse	2
GSM180217	Male	79	325900		Relapse	2
GSM180218	Male	27	209000		Relapse	2
GSM180219	Male	36	113000		CCR	1
GSM180220	Female	146	315200		Relapse	2
GSM180221	Male	141	19900		CCR	1
GSM180222	Male	146	158000		Relapse	2
GSM180223	Female	148	28400		CCR	1
GSM180224	Male	173	44400		CCR	1
GSM180225	Male	213	98600		Relapse	2
GSM180226	Male	138	1800		CCR	1
GSM180227	Male	125	12900		RER;CCR	1
GSM180228	Male	126	2200		CCR	1
GSM180229	Male	208	108000		Relapse	2
GSM180230	Male	152	170400		Relapse	2
GSM180231	Female	189	60200		CCR	1
GSM180232	Male	178	260500		Relapse	2

Table 2: Top 14 important SNPs selected by random forest though mean decrease in Gini-index

SNPs	Mean decrease Gini-index
rs11147977	100
rs299284	87.03649
rs1028180	24.6727
rs10491178	19.44876
rs1800472	18.34746
rs6020	16.41
rs1062964	15.59911
rs1150772	12.85912
rs2735018	12.81387
rs3824886	12.52663
rs3747295	11.27293
rs1124174	10.88839
rs3731625	10.76915
rs7309681	10.32153

Table 3: The gene annotations, gene description and pathways related to the top 14 important SNPs found through the random forest method in Section 6.4.2

SNPs	**Genes**	**Gene description**	**KEGG pathways**	**Ingenuity pathway analysis**
rs11147977	C13orf3	This gene encodes a component of the spindle and kinetochore-associated protein complex that regulates microtubule attachment to the kinetochores during mitosis.		
rs299284	HMMR	The protein encoded by this gene is involved in cell motility. It is expressed in breast tissue and together with other proteins	Glioma Invasiveness Signaling, FAK Signaling	
rs1028180	BLZF1	basic leucine zipper nuclear factor 1		
rs10491178	ABCA10	The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters	ABC transporters	
rs6020	F5	This gene encodes an essential cofactor of the blood coagulation cascade.	Coagulation System, Intrinsic Prothrombin Activation Pathway, Extrinsic Prothrombin Activation Pathway	
rs1062964	TACSTD2	This intronless gene encodes a carcinoma-associated antigen. This antigen is a cell surface receptor that transduces calcium signals.		
rs1150772	RPP21	RPP21 is a protein subunit of nuclear ribonuclease P, which processes the 5-prime leader sequence of precursor tRNAs	RNA transport	

Continued ...

SNPs	**Genes**	**Gene description**	**KEGG pathways**	**Ingenuity pathway analysis**
rs1800472	TGFβ 1	This gene encodes a member of the transforming growth factor beta (TGFβ) family of cytokines, which are multifunctional peptides that regulate proliferation,	Role of Cytokines in Mediating Communica Hepatic Fibrosis / Hepatic Stellate Cell IL-12 Signaling and Production in Macrop Molecular Mechanisms of Cancer Altered T Cell and B Cell Signaling in R Renal Cell Carcinoma Signaling Cardiac Hypertrophy Signaling Tight Junction Signaling HMGB1 Signaling Factors Promoting Cardiogenesis in Verte RAR Activation, TGF-β Signaling , p38 MAPK Signaling, Protein Kinase A Signaling, Pancreatic Adenocarcinoma Signaling, Wnt catenin Signaling, Mitotic Roles of Polo-Like Kinase, Role of Pattern Recognition Receptors in, Hepatic Cholestasis, Cyclins and Cell Cycle Regulation, Regulation of the Epithelial-Mesenchymal, Chronic Myeloid Leukemia Signaling, Aryl Hydrocarbon Receptor Signaling, Human Embryonic Stem Cell Pluripotency, Cell Cycle: G1/S Checkpoint Regulation, Role of Macrophages, Fibroblasts and End, Role of Osteoblasts, Osteoclasts and Cho, Colorectal Cancer Metastasis Signaling, PPAR/RXR Activation, Role of NFAT in Cardiac Hypertrophy, Inhibition of Angiogenesis by TSP1, Antiproliferative Role of TOB in T Cell , Adipogenesis pathway, Regulation of IL-2 Expression in Activat, Atherosclerosis Signaling, T Helper Cell Differentiation, Germ Cell-Sertoli Cell Junction Signalin, Glucocorticoid Receptor Signaling	
rs3824886	HTATIP2	HIV-1 Tat interactive protein 2, 30kDa		

Continued ...

SNPs	**Genes**	**Gene description**	**KEGG pathways**	**Ingenuity pathway analysis**
rs3747295	NHS	This gene encodes a protein containing four conserved nuclear localization signals. The encoded protein functions in eye, tooth, craniofacial and brain development, and it can regulate actin remodeling and cell morphology. Mutations in this gene have been shown to cause Nance-Horan syndrome, and also X-linked cataract-40		
rs2735018	HLA-G	Allograft rejection, HLA-G belongs to the HLA class I heavy chain paralogues. This class I molecule is a heterodimer consisting of a heavy chain and a light chain (beta-2 microglobulin)	Cell adhesion molecules (CAMs), Viral carcinogenesis, Epstein-Barr virus infection, Natural killer cell mediated, Phagosome cytotoxicity, Autoimmune thyroid disease, Herpes simplex infection, Type I diabetes mellitus, Endocytosis, Viral myocarditis, Antigen processing and presentation, HTLV-I infection, Graft-versus-host disease	Systemic Lupus Erythematosus Signaling, Antigen Presentation Pathway, Type I Diabetes Mellitus Signaling, Cdc42 Signaling, Neuroprotective Role of THOP1 in Alzheim, Crosstalk between Dendritic Cells and Na, Autoimmune Thyroid Disease Signaling, Allograft Rejection Signaling, OX40 Signaling Pathway, Communication between Innate and Adaptiv, Graft-versus-Host Disease Signaling,
rs1124174	INSIG1	Oxysterols regulate cholesterol homeostasis through the liver X receptor (LXR)- and sterol regulatory element-binding protein (SREBP)-mediated signaling pathways.	Unfolded protein response	

Continued ...

SNPs	**Genes**	**Gene description**	**KEGG pathways**	**Ingenuity pathway analysis**
rs3731625	ITSN2	This gene encodes a cytoplasmic protein which contains SH3 domains. This protein is a member of a family of proteins involved in clathrin-mediated endocytosis. Intersectin 2 is thought to regulate the formation of clathrin-coated vesicles and also may function in the induction of T cell antigen receptor (TCR) endocytosis.	Ephrin B Signaling	
rs7309681	MAPKAPK5	The protein encoded by this gene is a tumor suppressor and member of the serine/threonine kinase family. In response to cellular stress and proinflammatory cytokines, this kinase is activated through its phosphorylation by MAP kinases including MAPK1/ERK, MAPK14/p38-alpha, and MAPK11/p38-beta.	Cell adhesion molecules (CAMs)	ERK/MAPK Signaling, p38 MAPK Signaling

Table 4: The gene annotations and gene description of the top 80 genes as described in Section 6.4.2

Genes	**Gene description**
C13orf3	This gene encodes a component of the spindle and kinetochore-associated protein complex that regulates microtubule attachment to the kinetochores during mitosis.
HMMR	The protein encoded by this gene is involved in cell motility. It is expressed in breast tissue and together with other proteins
BLZF1	basic leucine zipper nuclear factor 1
ABCA10	The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters
TGFβ1	This gene encodes a member of the transforming growth factor β (TGFβ) family of cytokines, which are multifunctional peptides that regulate proliferation,
F5	This gene encodes an essential cofactor of the blood coagulation cascade.
TACSTD2	This intronless gene encodes a carcinoma-associated antigen. This antigen is a cell surface receptor that transduces calcium signals.
RPP21	RPP21 is a protein subunit of nuclear ribonuclease P, which processes the 5-prime leader sequence of precursor tRNAs
HLA-G	Allograft rejection, HLA-G belongs to the HLA class I heavy chain paralogues. This class I molecule is a heterodimer consisting of a heavy chain and a light chain (beta-2 microglobulin)
HTATIP2	HIV-1 Tat interactive protein 2, 30kDa
NHS	This gene encodes a protein containing four conserved nuclear localization signals. The encoded protein functions in eye, tooth, craniofacial and brain development.
INSIG1	Oxysterols regulate cholesterol homeostasis through the liver X receptor (LXR)- and sterol regulatory element-binding protein (SREBP)-mediated signaling pathways.
ITSN2	This gene encodes a cytoplasmic protein which contains SH3 domains.
MAPKAPK5	The protein encoded by this gene is a tumor suppressor and member of the serine/threonine kinase family.
SDC3	The protein encoded by this gene belongs to the syndecan proteoglycan family. It may play a role in the organization of cell shape by affecting the actin cytoskeleton.
CDH26	Cadherins are a family of adhesion molecules that mediate Ca2+-dependent cell-cell adhesion in all solid tissues and modulate a wide variety of processes, including cell polarization and migration.
MGAM	This gene encodes maltase-glucoamylase, which is a brush border membrane enzyme that plays a role in the final steps of digestion of starch.
TSPAN18	tetraspanin 18
KIAA1377	centrosomal protein 126kDa
ART5	he protein encoded by this gene belongs to the ARG-specific ADP-ribosyltransferase family.
MS4A6A	This gene encodes a member of the membrane-spanning 4A gene family. Members of this nascent protein family are characterized by common structural features and similar intron/exon splice boundaries and display unique expression patterns among hematopoietic cells and nonlymphoid tissues.
Continued ...	

Genes	**Gene description**
IL17R	interleukin 17 receptor C
FBN3	This gene encodes a protein that belongs to the fibrillin gene family. Fibrillins are extracellular matrix molecules that assemble into microfibrils in many connective tissues.
PASK	This gene encodes a member of the serine/threonine kinase family that contains two PAS domains.
ANKK1	The protein encoded by this gene belongs to the Ser/Thr protein kinase family, and protein kinase superfamily involved in signal transduction pathways
SLC31A1	The protein encoded by this gene is a high-affinity copper transporter found in the cell membrane.
BZRAP1	benzodiazepine receptor (peripheral) associated protein 1
WASF3	This gene encodes a member of the Wiskott-Aldrich syndrome protein family. The gene product is a protein that forms a multiprotein complex that links receptor kinases and actin.
ZAN	This gene encodes a protein that functions in the species specificity of sperm adhesion to the egg zona pellucida.
PAPLN	papilin, proteoglycan-like sulfated glycoprotein
PMFBP1	polyamine modulated factor 1 binding protein 1
ANKRD6	ankyrin repeat domain 6
ANK1	Ankyrins plays key roles in activities such as cell motility, activation, proliferation, contact and the maintenance of specialized membrane domains
OR52H1	Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell.
FUT2	The protein encoded by this gene is a Golgi stack membrane protein that is involved in the creation of a precursor of the H antigen, which is required for the final step in the soluble A and B antigen synthesis pathway.
VPS52	This gene encodes a protein that is similar to the yeast suppressor of actin mutations 2 gene.
MYST4	The protein encoded by this gene is a histone acetyltransferase and component of the MOZ/MORF protein complex.
OR3A4	Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell.
TRIM24	The protein encoded by this gene mediates transcriptional control by interaction with the activation function 2 (AF2) region of several nuclear receptors, including the estrogen, retinoic acid, and vitamin D3 receptors.
PASK	his gene encodes a member of the serine/threonine kinase family that contains two PAS domains. Expression of this gene is regulated by glucose, and the encoded protein plays a role in the regulation of insulin gene expression.
DNAH9	This gene encodes the heavy chain subunit of axonemal dynein, a large multi-subunit molecular motor. Axonemal dynein attaches to microtubules and hydrolyzes ATP to mediate the movement of cilia and flagella.
Continued ...	

Genes	**Gene description**
SYNE1	This gene encodes a spectrin repeat containing protein expressed in skeletal and smooth muscle, and peripheral blood lymphocytes, that localizes to the nuclear membrane.
ATXN7	The autosomal dominant cerebellar ataxias (ADCA) are a heterogeneous group of neurodegenerative disorders characterized by progressive degeneration of the cerebellum, brain stem and spinal cord.
TLR6	The protein encoded by this gene is a member of the Toll-like receptor (TLR) family which plays a fundamental role in pathogen recognition and activation of innate immunity.
NOS1	The protein encoded by this gene belongs to the family of nitric oxide synthases, which synthesize nitric oxide from L-arginine.
COX4I2	Cytochrome c oxidase (COX), the terminal enzyme of the mitochondrial respiratory chain, catalyzes the electron transfer from reduced cytochrome c to oxygen.
MDC1	The protein encoded by this gene contains an N-terminal forkhead domain, two BRCA1 C-terminal (BRCT) motifs and a central domain with 13 repetitions of an approximately 41-amino acid sequence.
TNNI3K	This gene encodes a protein that belongs to the MAP kinase kinase kinase (MAPKKK) family of protein kinases.
IMPACT	impact RWD domain protein, impact, RWD domain protein
ABCC5	The protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes.
ACP5	This gene encodes an iron containing glycoprotein which catalyzes the conversion of orthophosphoric monoester to alcohol and orthophosphate.
LCE4A	late cornified envelope 4A
FGD6	FYVE, RhoGEF and PH domain containing 6
RPL35	Ribosomes, the organelles that catalyze protein synthesis, consist of a small 40S subunit and a large 60S subunit. Together these subunits are composed of 4 RNA species and approximately 80 structurally distinct proteins.
TRIM15	The protein encoded by this gene is a member of the tripartite motif (TRIM) family. The TRIM motif includes three zinc-binding domains, a RING, a B-box type 1 and a B-box type 2, and a coiled-coil region.
MICB	This gene encodes a heavily glycosylated protein which is a ligand for the NKG2D type II receptor. Binding of the ligand activates the cytolytic response of natural killer (NK) cells, CD8 alphabeta T cells, and gammadelta T cells which express the receptor.
C10orf113	chromosome 10 open reading frame 113
MGC32020	
RAI14	retinoic acid induced 14
IFNGR1	This gene (IFNGR1) encodes the ligand-binding chain (alpha) of the gamma interferon receptor. Human interferon-gamma receptor is a heterodimer of IFNGR1 and IFNGR2.
MYOM1	The giant protein titin, together with its associated proteins, interconnects the major structure of sarcomeres, the M bands and Z discs.
RNASE3	The protein encoded by this gene belongs to the pancreatic ribonuclease family, a subset of the ribonuclease A superfamily.
MAP1D	methionyl aminopeptidase type 1D (mitochondrial)
Continued ...	

Genes	**Gene description**
HSD17B8	In mice, the Ke6 protein is a 17-beta-hydroxysteroid dehydrogenase that can regulate the concentration of biologically active estrogens and androgens.
TEGT	transmembrane BAX inhibitor motif containing 6
PBOV1	This intronless gene encodes a protein of unknown function. Its expression is up-regulated in some types of cancer, including prostate, breast, and bladder cancer.
MAP2K3	The protein encoded by this gene is a dual specificity protein kinase that belongs to the MAP kinase kinase family. This kinase is activated by mitogenic and environmental stress, and participates in the MAP kinase-mediated signaling cascade.
SHD	Src homology 2 domain containing transforming protein D
OFCC1	orofacial cleft 1 candidate 1
OR13F1	Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell.
TLR1	The protein encoded by this gene is a member of the Toll-like receptor (TLR) family which plays a fundamental role in pathogen recognition and activation of innate immunity.
C10orf53	chromosome 10 open reading frame 53
PIGB	his gene encodes a transmembrane protein that is located in the endoplasmic reticulum and is involved in GPI-anchor biosynthesis.
SPINK5	This gene encodes a multidomain serine protease inhibitor that contains 15 potential inhibitory domains. The inhibitor may play a role in skin and hair morphogenesis and anti-inflammatory and/or antimicrobial protection of mucous epithelia.
POU5F1	This gene encodes a transcription factor containing a POU homeodomain that plays a key role in embryonic development and stem cell pluripotency. Aberrant expression of this gene in adult tissues is associated with tumorigenesis.
MGC20470	
MRGPRX4	MAS-related GPR, member X4
DFNB31	This gene is thought to function in the organization and stabilization of sterocilia elongation and actin cystoskeletal assembly, based on studies of the related mouse gene.
KIAA1618	This gene encodes a protein containing a C3HC4-type RING finger domain, which is a specialized type of Zn-finger that binds two atoms of zinc and is thought to be involved in mediating protein-protein interactions.
PDCD4	This gene is a tumor suppressor and encodes a protein that binds to the eukaryotic translation initiation factor 4A1 and inhibits its function by preventing RNA binding.

Table 5: Pathways found related to the top 80 genes from KEGG and Ingenuity Pathways Analysis as described in Section 6.4.2

KEGG pathways	**Jones et al. (2008) Pathways**	**Ingenuity pathway analysis**
ABC transporters	Apoptosis	Glioma Invasiveness Signaling, FAK Signaling
RNA transport	DNA damage control	Role of Cytokines in Mediating Communication
Cell adhesion molecules (CAMs), Viral carcinogenesis, Epstein-Barr virus infection, Natural killer cell mediated, Phagosome cytotoxicity, Autoimmune thyroid disease, Herpes simplex infection, Type I diabetes mellitus, Endocytosis, Viral myocarditis, Antigen processing and presentation, HTLV-I infection, Graft-versus-host disease	Regulation of G1/S phase transition	Coagulation System, Intrinsic Prothrombin Activation Pathway, Extrinsic Prothrombin Activation Pathway
Cell adhesion molecules (CAMs), Natural killer cell mediated cytotoxicity	Hedgehog signaling	Systemic Lupus Erythematosus Signaling, Antigen Presentation Pathway, Type I Diabetes Mellitus Signaling, Cdc42 Signaling, Neuroprotective Role of THOP1 in Alzheim, Crosstalk between Dendritic Cells and Na, Autoimmune Thyroid Disease Signaling, Allograft Rejection Signaling, OX40 Signaling Pathway, Communication between Innate and Adaptiv, Graft-versus-Host Disease Signaling,
Starch and sucrose metabolism, Galactose metabolism, Metabolic pathways, Carbohydrate digestion and absorption	Homophilic cell adhesion	Unfolded protein response
Mineral absorption	Integrin signaling	Ephrin B Signaling
Adherens junction, Fc gamma R-mediated phagocytosis, Choline metabolism in cancer	c-Jun N-terminal kinase signaling	ERK/MAPK Signaling, p38 MAPK Signaling
Olfactory transduction	KRAS signaling	PCP pathway, Granulocyte Adhesion and Diapedesis
Glycosphingolipid biosynthesis - lacto and neolacto series, Metabolic pathways, Glycosphingolipid biosynthesis - globo series	Regulation of invasion	
Continued ...		

KEGG pathways	**Jones et al. (2008) Pathways**	**Ingenuity pathway analysis**
Non-alcoholic fatty liver disease (NAFLD),Huntington's disease, Huntington's disease, Metabolic pathways, Parkinson's disease, Oxidative phosphorylation, Cardiac muscle contraction,Alzheimer's disease	Small GTPasedependent signaling (other than KRAS)	Glycogen Degradation III
Osteoclast differentiation, Rheumatoid arthritis, Lysosome	TGF-β signaling	Signaling by Rho Family GTPases
Ribosome	Wnt/Notch signaling	Sperm Motility
Toxoplasmosis		Rac Signaling
Huntington's disease, Phagosome, Tuberculosis, Toll-like receptor signaling pathway, Chagas disease (American trypanosomiasis), Signaling pathways regulating pluripotency of stem cells		Type I Diabetes Mellitus Signaling, Production of Nitric Oxide and Reactive , HMGB1 Signaling, Hepatic Fibrosis / Hepatic Stellate Cell , IL-12 Signaling and Production in Macrop , iNOS Signaling, T Helper Cell DifferentiationRole of JAK1, JAK2 and TYK2 in InterferoColorectal Cancer Metastasis SignalingInterferon Signaling
Ribosome biogenesis in eukaryotes, Proteoglycans in cancer		Crosstalk between Dendritic Cells and Na
Metabolic pathways, Steroid hormone biosynthesis		
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis		
Continued ...		

KEGG pathways	**Jones et al. (2008) Pathways**	**Ingenuity pathway analysis**
Rap1 signaling pathway, Toxoplasmosis, MAPK signaling pathway, Epstein-Barr virus infection, Fc epsilon RI signaling pathway, Toll-like receptor signaling pathway, TNF signaling pathway, Influenza A, Inflammatory mediator regulation of TRP channels, Amyotrophic lateral sclerosis (ALS), GnRH signaling pathway		Hepatic Fibrosis / Hepatic Stellate Cell IL-12 Signaling and Production in Macrop Molecular Mechanisms of Cancer Altered T Cell and B Cell Signaling in R Renal Cell Carcinoma Signaling Cardiac Hypertrophy Signaling Tight Junction Signaling HMGB1 Signaling Factors Promoting Cardiogenesis in Verte RAR Activation, TGF-β Signaling , p38 MAPK Signaling, Protein Kinase A Signaling, Pancreatic Adenocarcinoma Signaling, Wnt/-catenin Signaling, Mitotic Roles of Polo-Like Kinase, Role of Pattern Recognition Receptors in, Hepatic Cholestasis, Cyclins and Cell Cycle Regulation, Regulation of the Epithelial-Mesenchymal, Chronic Myeloid Leukemia Signaling, Aryl Hydrocarbon Receptor Signaling, Human Embryonic Stem Cell Pluripotency, Cell Cycle: G1/S Checkpoint Regulation, Role of Macrophages, Fibroblasts and End, Role of Osteoblasts, Osteoclasts and Cho, Colorectal Cancer Metastasis Signaling, PPAR/RXR Activation, Role of NFAT in Cardiac Hypertrophy, Inhibition of Angiogenesis by TSP1, Antiproliferative Role of TOB in T Cell , Adipogenesis pathway, Regulation of IL-2 Expression in Activat, Atherosclerosis Signaling, T Helper Cell Differentiation, Germ Cell-Sertoli Cell Junction Signalin, Glucocorticoid Receptor Signaling

Bibliography

Aggarwal, B. B. (2003). Signalling pathways of the TNF superfamily: a double-edged sword. *Nature Reviews Immunology 3*(9), 745–756.

Akenteva, N., S. Shushanov, and A. Kotelnikov (2015). Effects of RHAMM/HMMR-selective peptides on survival of breast cancer cells. *Bulletin of experimental biology and medicine 159*(5), 658–661.

Alvord, G., J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology 8*(9), R183.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet 25*(1), 25–9.

Assié, G., T. LaFramboise, P. Platzer, and C. Eng (2008). Frequency of germline genomic homozygosity associated with cancer cases. *Jama 299*(12), 1437–1445.

Bacher, U., A. Kohlmann, and T. Haferlach (2010). Gene expression profiling for diagnosis and therapy in acute leukaemia and other haematologic malignancies. *Cancer treatment reviews 36*(8), 637–646.

Backes, C., A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Müller, E. Meese, and H.-P. Lenhof (2007). GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Research 35*(suppl 2), W186–W192.

Barrett, T. and R. Edgar (2006). Gene Expression Omnibus: Microarray data storage, submission, retrieval, and analysis. *Methods in enzymology 411*, 352–369.

Beissbarth, R. and T. Speed (2004). GOstat: finding statistically over expressed Gene Ontologies within groups of genes. *Bioinformatics 20*(9), 1464–1465.

Berriz, G., J. E. Beaver, C. Cenik, M. Tasan, and F. P. Roth (2009). Next generation software for functional trend analysis. *Bioinformatics 25*(22), 3043.

Bhojwani, D., H. Kang, R. Menezes, W. Yang, H. Sather, N. Moskowitz, D. Min, J. Potter, R. Harvey, S. Hunger, N. Seibel, E. A. Raetz, R. Pieters, M. A. Horstmann, M. V. Relling, M. L. den Boer, C. L. Willman, and W. L. Carroll (2008). Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: A Children's Oncology Group Study. *Journal of Clinical Oncology 26*(27), 4376–4384.

Bickel, S., M. Brückner, and T. Scheffer (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine learning*, pp. 81–88. ACM.

Bogdanov, K. (2009). Significance of mitotic spindle checkpoint genes in leukemia. *Cell and Tissue Biology 3*(6), 503–510.

Borne, K. (2009). Scientific data mining in astronomy. *arXiv preprint arXiv:0911.0505*.

Brazil, D. P. and B. A. Hemmings (2001). Ten years of protein kinase B signalling: a hard Akt to follow. *Trends in biochemical sciences 26*(11), 657–664.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Burdon, K. P., J. D. McKay, M. M. Sale, I. M. Russell-Eggitt, D. A. Mackey, M. G. Wirth, J. E. Elder, A. Nicoll, M. P. Clarke, and L. M. FitzGerald (2003). Mutations in a novel gene, NHS, cause the pleiotropic effects of nance-horan syndrome, including severe congenital cataract, dental anomalies, and mental retardation. *The American Journal of Human Genetics 73*(5), 1120–1130.

Bureau, A., J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic epidemiology 28*(2), 171–182.

Caniza, H., A. E. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, M. Mesiti, G. Valentini, and A. Paccanaro (2014). GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. *Bioinformatics 30*(15), 2235–2236.

Catchpoole, D., D. Guo, H. Jiang, and C. Biesheuvel (2008). Predicting outcome in childhood acute lymphoblastic leukemia using gene expression profiling: Prognostication or protocol selection? *Blood 111*(4), 2486–2487.

Cavalli-Sforza, L. L. and W. F. Bodmer (1999). *The genetics of human populations*. Courier Dover Publications.

Chen, A. H. and Z.-W. Huang (2010). A new multi-task learning technique to predict classification of leukemia and prostate cancer. In *Medical Biometrics*, pp. 11–20. Springer.

Chen, C. (2013). *Information visualisation and virtual environments*. Springer Science & Business Media.

Chen, L. S., C. M. Hutter, J. D. Potter, Y. Liu, R. L. Prentice, U. Peters, and L. Hsu (2010). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *The American Journal of Human Genetics 86*(6), 860–871.

Clarke, B., E. Fokoue, and H. H. Zhang (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media.

Coates, M. et al. (2001). *NSW Cancer Registry*. Cancer Research and Registers Division, NSW Cancer Council.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement 20*(1), 37–46.

Coifman, R. and S. Lafon (2006). Diffusion maps. *Applied and Computational Harmonic Analysis 21*(1), 5–30.

Collins-Underwood, J. and C. Mullighan (2010). Genomic profiling of high-risk acute lymphoblastic leukemia. *Leukemia 24*(10), 1676–1685.

Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine learning 20*(3), 273–297.

Couto, F. M. and H. S. Pinto (2013). The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of bioinformatics and computational biology 11*(05), 1371001.

Croce, C. M. and G. A. Calin (2012, June 28). MicroRNA signatures associated with human Chronic Lymphocytic Leukemia (CLL) and uses Thereof. US Patent App. 13/536,837.

Curtis, R. K., M. Orešič, and A. Vidal-Puig (2005). Pathways to the analysis of microarray data. *TRENDS in Biotechnology 23*(8), 429–435.

Cvetković, D., A. V. Babwah, and M. Bhattacharya (2013). Kisspeptin/KISS1R system in breast cancer. *J Cancer 4*(8), 653.

Dale, J. M., L. Popescu, and P. D. Karp (2010). Machine learning methods for metabolic pathway prediction. *BMC Bioinformatics 11*(1), 15.

Deeb, S. J., S. Tyanova, M. Hummel, M. Schmidt-Supprian, J. Cox, and M. Mann (2015). Machine learning-based classification of diffuse large b-cell lymphoma patients by their protein expression profiles. *Molecular & Cellular Proteomics 14*(11), 2947–2960.

DeMers, D. and G. Cottrell (1993). Non-linear dimensionality reduction. *Advances in neural information processing systems*, 580–580.

Duprez, E., J. Tong, J. Derre, S. Chen, R. Berger, Z. Chen, and M. Lanotte (1997). JEM-1, a novel gene encoding a leucine-zipper nuclear factor upregulated during retinoid-induced maturation of NB4 promyelocytic leukaemia. *Oncogene 14*(13), 1563–1570.

Edgar, R., M. Domrachev, and A. E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research 30*(1), 207–210.

Ellenbroek, S. I. and J. G. Collard (2007). Rho GTPases: functions and association with cancer. *Clinical & Experimental Metastasis 24*(8), 657–672.

Engreitz, J. M., B. J. Daigle Jr, J. J. Marshall, and R. B. Altman (2010). Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of biomedical informatics 43*(6), 932–944.

Everhart, J. E. and E. C. Wright (2013). Association of γ-glutamyl transferase (GGT) activity with treatment and clinical outcomes in chronic hepatitis C (HCV). *Hepatology 57*(5), 1725–1733.

Fadnes, B., A. Husebekk, G. Svineng, Ø. Rekdal, M. Yanagishita, S. O. Kolset, and L. Uhlin-Hansen (2012). The proteoglycan repertoire of lymphoid cells. *Glycoconjugate journal 29*(7), 513–523.

Fan, P., H. E. Cunliffe, O. L. Griffith, F. A. Agboke, P. Ramos, J. W. Gray, and V. C. Jordan (2014). Identification of gene regulation patterns underlying both oestrogen- and tamoxifen-stimulated cell growth through global gene expression profiling in breast cancer cells. *European Journal of Cancer 50*(16), 2877–2886.

Fauci, A. S. (2008). *Harrison's principles of internal medicine*, Volume 2. McGraw-Hill Medical New York.

Fleiss, J. L., B. Levin, and M. C. Paik (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Flores-Borja, F., P. S. Kabouridis, E. C. Jury, D. A. Isenberg, and R. A. Mageed (2007). Altered lipid raft–associated proximal signaling and translocation of CD45 tyrosine phosphatase in B lymphocytes from patients with systemic lupus erythematosus. *Arthritis & Rheumatism 56*(1), 291–302.

Flotho, C., E. Coustan-Smith, D. Pei, C. Cheng, G. Song, C.-H. Pui, J. R. Downing, and D. Campana (2007). A set of genes that regulate cell proliferation predicts treatment outcome in childhood acute lymphoblastic leukemia. *Blood 110*(4), 1271.

Fodor, S., R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams (1993). Multiplexed biochemical assays with biological chips. *Nature 364*, 555–556.

Forero, R. M., M. Hernández, and J. M. Hernández-Rivas (2013). Genetics of acute lymphoblastic leukemia. *Edited by Margarita Guenova and Gueorgui Balatzenko*, 1.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer Series in Statistics.

Fröhlich, H., N. Speer, A. Poustka, and T. Beissbarth.

Fröhlich, H., N. Speer, C. Spieth, and A. Zell (2006). Kernel Based Functional Gene Grouping. *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, 3580–3585.

Fujibuchi, W., S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and M. Kanehisa (1998). DBGET/LinkDB: an integrated database retrieval system. In *Pac. Symp. Biocomput*, Volume 98, pp. 683–694.

Ganesh Kumar, P., T. Aruldoss Albert Victoire, P. Renukadevi, and D. Devaraj (2012). Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. *Expert Systems with Applications 39*(2), 1811–1821.

Garcia-Manero, G., H. Yang, S.-Q. Kuang, S. O'Brien, D. Thomas, and H. Kantarjian (2009). Epigenetics of acute lymphocytic leukemia. In *Seminars in hematology*, Volume 46, pp. 24–32. Elsevier.

Gerhard, M. (1992). *Biological Pathways*. Boehringer Mannheim.

Golub, G. and C. Van Loan (1996). *Matrix computations*. Johns Hopkins University Press.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*(5439), 531–537.

Goto, S., T. Nishioka, and M. Kanehisa (1998). Ligand: chemical database for enzyme reactions. *Bioinformatics 14*(7), 591–599.

Greenman, C., P. Stephens, R. Smith, G. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, and M. Stratton (2007). Patterns of somatic mutation in human cancer genomes. *Nature 446*(7132), 153–158.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research 3*, 1157–1182.

Haferlach, T., A. Kohlmann, S. Schnittger, M. Dugas, W. Hiddemann, W. Kern, and C. Schoch (2005). Global approach to the diagnosis of leukemia using gene expression profiling. *Blood 106*(4), 1189–1198.

Haghverdi, L., F. Buettner, and F. J. Theis (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics 31*(18), 2989–2998.

Han, J. and M. Kamber (2006). Data mining: Concepts and techniques.

Hatfield, K. J., H. Reikvam, and Ø. Bruserud (2014). Identification of a subset of patients with acute myeloid leukemia characterized by long-term in vitro proliferation and altered cell cycle regulation of the leukemic cells. *Expert opinion on therapeutic targets 18*(11), 1237–1251.

Haverty, P. M., J. Fridlyand, L. Li, G. Getz, R. Beroukhim, S. Lohr, T. D. Wu, G. Cavet, Z. Zhang, and J. Chant (2008). High-resolution genomic and expression analyses of copy number alterations in breast tumors. *Genes, Chromosomes and Cancer 47*(6), 530–542.

Hayashibara, T., Y. Yamada, N. Mori, H. Harasawa, K. Sugahara, T. Miyanishi, S. Kamihira, and M. Tomonaga (2003). Possible involvement of Aryl hydrocarbon Receptor (Ahr) in adult T-cell leukemia (ATL) leukemogenesis: constitutive activation of Ahr in ATL. *Biochemical and biophysical research communications 300*(1), 128–134.

Haykin, S. (1999). *Neural networks: a comprehensive foundation* (2nd ed.). Prentice–Hall.

Henze, G., R. Fengler, R. Hartmann, B. Kornhuber, G. Janka-Schaub, D. Niethammer, and H. Riehm (1991). Six-year experience with a comprehensive approach to the treatment of recurrent childhood acute lymphoblastic leukemia (ALL-REZ BFM 85). a relapse study of the BFM group. *Blood 78*(5), 1166–1172.

Hinton, G. and S. Roweis (2002). Stochastic neighbor embedding. *Advances in neural information processing systems 15*, 833–840.

Holden, M., S. Deng, L. Wojnowski, and B. Kulle (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics 24*(23), 2784–2785.

Holmans, P., E. K. Green, J. S. Pahwa, M. A. Ferreira, S. M. Purcell, P. Sklar, M. J. Owen, M. C. O'Donovan, and N. Craddock (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *The American Journal of Human Genetics 85*(1), 13–24.

Hormozi, A. M. and S. Giles (2004). Data mining: a competitive weapon for banking and retail industries. *Information systems management 21*(2), 62–71.

Huang, B. F. and P. C. Boutros (2016). The parameter sensitivity of random forests. *BMC bioinformatics 17*(1), 331.

Huang, D., B. T. Sherman, and R. A. Lempicki (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research 37*(1), 1–13.

Huang, H.-L. and F.-L. Chang (2007). ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. *Biosystems 90*(2), 516–528.

Illumina (2014). Illumina. http://www.illumina.com/. Viewed at: 2014-06-05.

Janssens, K., R. Gershoni-Baruch, N. Guañabens, N. Migone, S. Ralston, M. Bonduelle, W. Lissens, L. Van Maldergem, F. Vanhoenacker, L. Verbruggen, et al. (2000). Mutations in the gene encoding the latency-associated peptide of tgf-β1 cause camurati-engelmann disease. *Nature genetics 26*(3).

Jayawardana, K., S.-J. Schramm, L. Haydu, J. F. Thompson, R. A. Scolyer, G. J. Mann, S. Müller, and J. Y. H. Yang (2015). Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mrna, microrna, and protein information. *International Journal of Cancer 136*(4), 863–874.

Jia, L., A. Clear, F.-T. Liu, J. Matthews, N. Uddin, A. McCarthy, E. Hoxha, C. Durance, S. Iqbal, and J. G. Gribben (2014). Extracellular HMGB1 promotes differentiation of nurse-like cells in chronic lymphocytic leukemia. *Blood 123*(11), 1709–1719.

Jiao, X., S. D. Hooper, T. Djureinovic, C. Larsson, F. Wärnberg, C. Tellgren-Roth, J. Botling, and T. Sjöblom (2013). Gene rearrangements in hormone receptor negative breast cancers revealed by mate pair sequencing. *BMC genomics 14*(1), 1.

Jiao, X., B. T. Sherman, D. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki (2012). David-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics 28*(13), 1805–1806.

Jin, L., X.-Y. Zuo, W.-Y. Su, X.-L. Zhao, M.-Q. Yuan, L.-Z. Han, X. Zhao, Y.-D. Chen, and S.-Q. Rao (2014). Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & Bioinformatics 12*(5), 210–220.

John Tomfohr, J. L. and T. B. Kepler (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics 6*(1), 225–234.

Jolliffe, I. T. (2004). *Principal Component Analysis* (Second ed.). Springer Series in Statistics. New York: Springer.

Jones, S., X. Zhang, D. W. Parsons, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S.-M. Hong, B. Fu, M.-T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, and K. W. Kinzler (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science 321*(5897), 1801–1806.

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research 36*, 480–484.

Kim, D. and L. Finkel (2003). Hyperspectral image processing using locally linear embedding. In *Neural Engineering, 2003. First International IEEE EMBS Conference Proceedings*, pp. 316–319. IEEE.

Kitisin, K., T. Saha, T. Blake, N. Golestaneh, M. Deng, C. Kim, Y. Tang, K. Shetty, B. Mishra, and L. Mishra (2007). TGF-β signaling in development. *Sci Stke 399*.

Krajinovic, M., J. Elbared, S. Drouin, L. Bertout, A. Rezgui, M. Ansari, M. Raboisson, S. Lipshultz, L. Silverman, S. Sallan, D. S. Neuberg, J. L. Kutok, C. Laverdiere, D. Sinnett, and G. Andelfinger (2015). Polymorphisms of ABCC5 and NOS3 genes influence doxorubicin cardiotoxicity in survivors of childhood acute lymphoblastic leukemia. *The Pharmacogenomics Journal*.

Lee, C.-H., W.-H. Kuo, C.-C. Lin, Y.-J. Oyang, H.-C. Huang, and H.-F. Juan (2013). MicroRNA-regulated protein-protein interaction networks and their functions in breast cancer. *International Journal Of Molecular Sciences 14*(6), 11560–11606.

Lee, I.-Y., J.-M. Ho, and M.-S. Chen (2005). CLUGO: a clustering algorithm for automated functional annotations based on Gene Ontology. In *Fifth Data Mining IEEE International Conference on*, pp. 4–pp. IEEE.

Lee, J. and M. Verleysen (2007). *Nonlinear dimensionality reduction*. Springer Verlag.

Lee, M., K. A. Williams, Y. Hu, J. Andreas, S. J. Patel, S. Zhang, and N. P. Crawford (2015). GNL3 and SKA3 are novel prostate cancer metastasis susceptibility genes. *Clinical & experimental metastasis 32*(8), 769–782.

Lee, S. G., J. U. Hur, and Y. S. Kim (2004). A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics 20*(3), 381–388.

Leslie, C., R. Kuang, and E. Eskin (2004). Inexact matching string kernels for protein classification. In B. Schölkopf et al. (Eds.), *Kernel methods in computational biology*, pp. 95–112. MIT Press.

Li, Z., W. Zhang, M. Wu, S. Zhu, C. Gao, L. Sun, R. Zhang, N. Qiao, H. Xue, Y. Hu, S. Bao, H. Zheng, and J. D. Han (2009). Gene expression–based classification and regulatory networks of pediatric acute lymphoblastic leukemia. *Blood 114*(20), 4486–4493.

Liaw, A. and M. Wiener (2002). Classification and Regression by randomForest. *R news 2*(3), 18–22.

Lord, P. W., R. D. Stevens, A. Brass, and C. A. Goble (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics 19*(10), 1275–1283.

Maillard, I., T. Fang, and W. S. Pear (2005). Regulation of lymphoid development, differentiation, and function by the Notch pathway. *Annu. Rev. Immunol. 23*, 945–974.

Marx, V. (2013). Biology: The big challenges of big data. *Nature 498*(7453), 255–260.

Mazandu, G. K., E. R. Chimusa, M. Mbiyavanga, and N. J. Mulder (2016). A-DaGO-Fun: an adaptable Gene Ontology semantic similarity-based functional analysis tool. *Bioinformatics*, 477–479.

Medina, I., D. Montaner, N. Bonifaci, M. A. Pujana, J. Carbonell, J. Tarraga, F. Al-Shahrour, and J. Dopazo (2009). Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic acids research 37*(suppl 2), W340–W344.

Meng, Y. A., Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta (2009). Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics 10*(1), 1.

Mirebeau-Prunier, D., S. Le Pennec, C. Jacques, J.-F. Fontaine, N. Gueguen, N. Boutet-Bouzamondo, A. Donnart, Y. Malthièry, and F. Savagner (2013). Estrogen-related receptor alpha modulates lactate dehydrogenase activity in thyroid tumors. *PloS one 8*(3), e58683.

Mirghani, H., N. Ugolin, C. Ory, M. Lefèvre, S. Baulande, P. Hofman, J. L. St Guily, S. Chevillard, and R. Lacave (2014). A predictive transcriptomic signature of oropharyngeal cancer according to HPV16 status exclusively. *Oral Oncology 50*(11), 1025–1034.

Mistry, M. and P. Pavlidis (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics 9*(1), 327.

Mitchison, T. and L. Cramer (1996). Actin-based cell motility and cell locomotion. *Cell 84*(3), 371–379.

Mullighan, C. G. (2012). The molecular genetic makeup of acute lymphoblastic leukemia. *ASH Education Program Book 2012*(1), 389–396.

Mullighan, C. G., S. Goorha, I. Radtke, C. B. Miller, E. Coustan-Smith, J. D. Dalton, K. Girtman, S. Mathew, J. Ma, S. B. Pounds, X. Su, C. H. Pui, M. V. Relling, W. E. Evans, S. A. Shurtleff, and J. R. Downing (2007). Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature 446*(7137), 758–764.

Musa, A. B. (2014). A comparison of 1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression. *International Journal of Machine Learning and Cybernetics 5*(6), 861–873.

Nakanishi, H., C. Taccioli, J. Palatini, C. Fernandez-Cymering, R. Cui, T. Kim, S. Volinia, and C. M. Croce (2014). Loss of miR-125b-1 contributes to head and neck cancer development by dysregulating TACSTD2 and MAPK pathway. *Oncogene 33*(6), 702–712.

Nam, D., J. Kim, S.-Y. Kim, and S. Kim (2010). GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic acids research*, gkq428.

National Cancer Institute (2013). What you need to know about leukemia. http://www.cancer.gov/. Viewed at: 2015-12-01.

Nguyen, S., T. Leblanc, P. Fenaux, F. Witz, D. Blaise, A. Pigneux, X. Thomas, F. Rigal-Huguet, B. Lioure, and F. D. R. J. C. S. L. G. H. J. L. S. G. D. H. Auvrignon, Anne (2002). A white blood cell index as the main prognostic factor in t (8; 21) acute myeloid leukemia (AML): a survey of 161 cases from the French AML intergroup. *Blood 99*(10), 3517–3523.

Nilsson, J., T. Fioretos, M. Höglund, and M. Fontes (2004). Approximate geodesic distances reveal biologically relevant structures in microarray data. *Bioinformatics 20*(6), 874–880.

Nishizuka, T. (1980). Metabolic maps. *Biochemical Society of Japan*.

Nogueira, S. and G. Brown (2016). Measuring the stability of feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 442–457. Springer.

Noren, N. K. and E. B. Pasquale (2007). Paradoxes of the EphB4 receptor in cancer. *Cancer research 67*(9), 3994–3997.

O'Connor, C. M., J. U. Adams, and J. Fairman (2010). Essentials of cell biology. *NPG Education, Cambridge*.

O'Dushlaine, C., E. Kenny, E. A. Heron, R. Segurado, M. Gill, D. W. Morris, and A. Corvin (2009). The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics 25*(20), 2762–2763.

Ogata, H., W. Fujibuchi, H. Bono, S. Goto, and M. Kanehisa (1996). Analysis of binary relations and hierarchies of enzymes in the metabolic pathways. *Genome Informatics 7*, 128–136.

Oppmann, B., R. Lesley, B. Blom, J. C. Timans, Y. Xu, B. Hunte, F. Vega, N. Yu, J. Wang, K. Singh, F. Zonin, E. Vaisberg, T. Churakova, M. Liu, D. Gorman, J. Wagner, S. Zurawski, Y. Liu, J. S. Abrams, K. W. Moore, D. Rennick, R. de Waal-Malefyt, C. Hannum, J. F. Bazan, and R. A. Kastelein (2000). Novel p19 protein engages IL-12p40 to form a cytokine, IL-23, with biological activities similar as well as distinct from IL-12. *Immunity 13*(5), 715–725.

Orloff, M. S., L. Zhang, G. Bebek, and C. Eng (2012). Integrative genomic analysis reveals extended germline homozygosity with lung cancer risk in the PLCO cohort. *PloS one 7*(2).

Orsenigo, C. and C. Vercellis (2013). A comparative study of nonlinear manifold learning methods for cancer microarray data classification. *Expert Systems with Applications 40*(6), 2189–2197.

Oshiro, T. M., P. S. Perez, and J. A. Baranauskas (2012). How many trees in a random forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 154–168. Springer.

Paley, S. M. and P. D. Karp (2002). Evaluation of computational metabolic-pathway predictions for helicobacter pylori. *Bioinformatics 18*(5), 715–724.

Panteris, E., S. Swift, A. Payne, and X. Liu (2007). Mining pathway signatures from microarray data and relevant biological knowledge. *Journal of Biomedical Informatics 40*(6), 698–706.

Park, T., S.-G. Yi, S.-H. Kang, S. Lee, Y.-S. Lee, and R. Simon (2003). Evaluation of normalization methods for microarray data. *BMC Bioinformatics 4*(1), 1.

Pearsons, K. (1901). On lines and planes of closest fit to systems of point in space. *Philosophical Magazine 2*, 559–572.

Pollock, B. H., M. R. DeBaun, B. M. Camitta, J. J. Shuster, Y. Ravindranath, D. J. Pullen, V. J. Land, D. H. Mahoney, S. J. Lauer, and S. B. Murphy (2000). Racial differences in the survival of childhood b-precursor acute lymphoblastic leukemia: a Pediatric Oncology Group study. *Journal of Clinical Oncology 18*(4), 813–813.

Popescu, M., J. Keller, J. Mitchell, and J. Bezdek (2004). Functional summarization of gene product clusters using Gene Ontology similarity measures. In *Proceedings of*

IEEE Intelligent Sensors, Sensor Networks and Information Processing Conference, pp. 553–558. IEEE.

Pui, C.-H. and W. E. Evans (1998). Acute lymphoblastic leukemia. *New England Journal of Medicine 339*(9), 605–615.

Pui, C.-H. and W. E. Evans (2006). Treatment of acute lymphoblastic leukemia. *New England Journal of Medicine 354*(2), 166–178.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences 98*(26), 15149–15154.

Ramsay, A. G., R. Evans, S. Kiaii, L. Svensson, N. Hogg, and J. G. Gribben (2013). Chronic lymphocytic leukemia cells induce defective LFA-1–directed T-cell motility by altering Rho GTPase signaling that is reversible with lenalidomide. *Blood 121*(14), 2704–2714.

Rapaport, F., A. Zinovyev, M. Dutreix, E. Barillot, and J.-P. Vert (2007). Classification of microarray data using gene networks. *BMC Bioinformatics 8*(1), 35.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Int. Joint Conference on Artificial Intelligence*, pp. 448–453.

Reutlinger, M. and G. Schneider (2012). Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *Journal of Molecular Graphics and Modelling 34*, 108–117.

Richards, A., B. Muller, M. Shotwell, A. Cowart, B. Rohrer, and X. Lu (2010). Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph. *Bioinformatics 26*(12), i79.

Roberts, K. G. and C. G. Mullighan (2015). Genomics in acute lymphoblastic leukaemia: insights and treatment implications. *Nature Reviews Clinical Oncology 12*(6), 344–357.

Roweis, S. T. and L. K. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*(5500), 2323–2326.

Rush, J., A. Moritz, K. A. Lee, A. Guo, V. L. Goss, E. J. Spek, H. Zhang, X.-M. Zha, R. D. Polakiewicz, and M. J. Comb (2005). Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nature biotechnology 23*(1), 94–101.

Sanfilippo, A., C. Posse, B. Gopalan, R. Riensche, N. Beagley, and B. Baddeley (2007). Combining Hierarchical and Associative Gene Ontology Relations With Textual Evidence in Estimating Gene and Gene Product Similarity. *IEEE Transactions on Nanobioscience 6*(1), 51–59.

Sato, T., T. H. Tran, A. R. Peck, C. Liu, A. Ertel, J. Lin, L. M. Neilson, and H. Rui (2013). Global profiling of prolactin-modulated transcripts in breast cancer in vivo. *Molecular cancer 12*(1), 1.

Savolainen-Peltonen, H., V. Vihma, M. Leidenius, F. Wang, U. Turpeinen, E. Hämäläinen, M. J. Tikkanen, and T. S. Mikkola (2014). Breast adipose tissue estrogen metabolism in postmenopausal women with or without breast cancer. *The Journal of Clinical Endocrinology & Metabolism 99*(12), E2661–E2667.

Schilling, C. H., S. Schuster, B. O. Palsson, and R. Heinrich (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnology progress 15*(3), 296–303.

Schmidt, J. C., H. Arthanari, A. Boeszoermenyi, N. M. Dashkevich, E. M. Wilson-Kubalek, N. Monnier, M. Markus, M. Oberer, R. A. Milligan, M. Bathe, G. Wagner, E. L. Grishchuk, and I. M. Cheeseman (2012). The kinetochore-bound SKA1 complex tracks depolymerizing microtubules and binds to curved protofilaments. *Developmental cell 23*(5), 968–980.

Schomburg, D. and M. Salzmann (1991). *Enzyme handbook*. Springer.

Schroeder, M. P., A. Gonzalez-Perez, and N. Lopez-Bigas (2013). Visualizing multidimensional cancer genomics data. *Genome Med 5*(9), 10–1186.

Selkov, E., S. Basmanova, T. Gaasterland, I. Goryanin, Y. Gretchkin, N. Maltsev, V. Nenashev, R. Overbeek, E. Panyushkina, L. Pronevitch, E. Selkov, and I. Yunus (1996). The Metabolic Pathway Collection From Emp: The Enzymes and Metabolic Pathways Database. *Nucleic acids research 24*(1), 26–28.

Sheehan, Q. B., G. A., D. B, and S. (2008). A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics 9*(1), 468.

Shi, J. and Z. Luo (2010). Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Computers in biology and medicine 40*(8), 723–732.

Shi, T. W., K. Moorthy, M. S. Mohamad, S. Deris, S. Omatu, and M. Yoshioka (2014). Random forest and gene ontology for functional analysis of microarray data. In *Computational Intelligence and Applications (IWCIA), 2014 IEEE 7th International Workshop on*, pp. 29–34. IEEE.

Shtivelman, E. (1997). A link between metastasis and resistance to apoptosis of variant small cell lung carcinoma. *Oncogene 14*(18), 2167–2178.

Shyamala, N. and K. Vijayakumar (2014). Microarray Gene Expression Cancer Diagnosis Using Modified Extreme Learning Machine Classification. *Artificial Intelligent Systems and Machine Learning 6*(8), 293–296.

Siesser, P. M. and S. K. Hanks (2006). The signaling and biological implications of FAK overexpression in cancer. *Clinical Cancer Research 12*(11), 3233–3237.

Sim, J. and C. C. Wright (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy 85*(3), 257–268.

Skillicorn, D. (2007). *Understanding complex datasets: data mining with matrix decompositions*. CRC press.

Speer, N., H. Frohlich, C. Spieth, and A. Zell (2005). Functional grouping of genes using spectral clustering and gene ontology. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 298–303.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9*(1).

Su, J.-L., M.-T. Lin, C.-C. Hong, C.-C. Chang, S.-G. Shiah, C.-W. Wu, S.-T. Chen, Y.-P. Chau, and M.-L. Kuo (2005). Resveratrol induces Fasl-related apoptosis through Cdc42 activation of ASK1/JNK-dependent signaling pathway in human leukemia HL-60 cells. *Carcinogenesis 26*(1), 1–10.

Sumathi, S. and S. Sivanandam (2006). Data mining for insurance. *Introduction to Data Mining and its Applications*, 473–498.

Suo, N. and X. Qian (2010). Motion retrieval using isomap. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, pp. 1202–1204. IEEE.

Supek, F. and N. Škunca (2016). Visualizing Gene Ontology annotations. *arXiv preprint arXiv:1602.07103*.

Suyama, M., A. Ogiwara, T. Nishioka, and J. Oda (1993). Searching for amino acid sequence motifs among enzymes: the enzyme–reaction database. *Computer applications in the biosciences: CABIOS 9*(1), 9–15.

Taichman, R. S. (2005). Blood and bone: two tissues whose fates are intertwined to create the hematopoietic stem-cell niche. *Blood 105*(7), 2631–2639.

Tan, Y., L. Shi, W. Tong, and C. Wang (2005). Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. *Nucleic Acids Research 33*(1), 56–65.

Tan, Y., F. Wu, P. Tamayo, W. N. Haining, and J. P. Mesirov (2015). Constellation Map: Downstream visualization and interpretation of gene set enrichment results. *F1000Research 4*.

Tenenbaum, J. B., V. De Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*(5500), 2319–2323.

Torre, L. A., F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians 65*(2), 87–108.

Treviño, L. R., W. Yang, D. French, S. P. Hunger, W. L. Carroll, M. Devidas, C. Willman, G. Neale, J. Downing, S. C. Raimondi, C.-H. Pui, W. E. Evans, and M. V. Relling (2009). Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nature genetics 41*(9), 1001–1005.

Van der Maaten, L. (2007). An introduction to dimensionality reduction using MATLAB. *Report 1201*, 07–07.

Vick, B., A. Weber, T. Urbanik, T. Maass, A. Teufel, P. H. Krammer, J. T. Opferman, M. Schuchmann, P. R. Galle, and H. Schulze-Bergkamen (2009). Knockout of myeloid cell leukemia-1 induces liver damage and increases apoptosis susceptibility of murine hepatocytes. *Hepatology 49*(2), 627–636.

Vlierberghe, P. V. and A. Ferrando (2012, 10). The molecular basis of t cell acute lymphoblastic leukemia. *The Journal of Clinical Investigation 122*(10), 3398–3406.

Vuong, H., A. Che, S. Ravichandran, B. T. Luke, J. R. Collins, and U. S. Mudunuri (2015). AVIA v2. 0: annotation, visualization and impact analysis of genomic variants and genes. *Bioinformatics*, 2748–2750.

Wagholikar, K. B., V. Sundararajan, and A. W. Deshpande (2012). Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of medical systems 36*(5), 3029–3049.

Wang, K., M. Li, and M. Bucan (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics 81*(6), 1278–1283.

Wang, X. and O. Gotoh (2009). Cancer classification using single genes. *Genome Informatics 23*(1), 176–88.

Warde-Farley, D., S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris (2010). The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research 38*(suppl 2), W214–W220.

Weiss, G. M. (2005). Data mining in telecommunications. In *Data Mining and Knowledge Discovery Handbook*, pp. 1189–1201. Springer.

Wu, J.-C., C.-S. Lai, V. Badmaev, K. Nagabhushanam, C.-T. Ho, and M.-H. Pan (2011). Tetrahydrocurcumin, a major metabolite of curcumin, induced autophagic cell death through coordinative modulation of PI3K/Akt-mTOR and MAPK signaling pathways in human leukemia HL-60 cells. *Molecular nutrition & food research 55*(11), 1646–1654.

Yamanishi, Y., J.-P. Vert, and M. Kanehisa (2005). Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics 21*(suppl 1), i468–i477.

Yang, J. J., C. Cheng, W. Yang, D. Pei, X. Cao, Y. Fan, S. B. Pounds, G. Neale, L. R. Trevino, D. French, D. Campana, J. R. Downing, W. E. Evans, C. H. Pui, M. Devidas, W. P. Bowman, B. M. Camitta, C. L. Willman, S. M. Davies, M. J. Borowitz, W. L. Carroll, S. P. Hunger, and M. V. Relling (2009). Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *Jama 301*(4), 393–403.

Yeoh, E.-J., M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell 1*(2), 133–143.

Yokota, A., S. Kimura, R. Tanaka, M. Takeuchi, H. Yao, K. Sakai, R. Nagao, J. Kuroda, Y. Kamitsuji, E. Kawata, E. Ashihara, and T. Maekawa (2010). Osteoclasts are involved in the maintenance of dormant leukemic cells. *Leukemia research 34*(6), 793–799.

Yoo, C., L. Ramirez, and J. Liuzzi (2014). Big data analysis using modern statistical and machine learning methods in medicine. *International neurourology journal 18*(2), 50–57.

Yu, C., M. Rahmani, P. Dent, and S. Grant (2004). The hierarchical relationship between MAPK signaling and ROS generation in human leukemia cells undergoing apoptosis in response to the proteasome inhibitor bortezomib. *Experimental cell research 295*(2), 555–566.

Zhang, K., S. Cui, S. Chang, L. Zhang, and J. Wang (2010). i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic acids research 38*(suppl 2), W90–W95.

Zhao, S., X. Dong, W. Shen, Z. Ye, and R. Xiang (2016). Machine learning-based classification of diffuse large b-cell lymphoma patients by eight gene expression profiles. *Cancer medicine*.

Zhao, W. (2010). Targeted therapy in T-cell malignancies: dysregulation of the cellular signaling pathways. *Leukemia 24*(1), 13–21.

Ziauddin, J. and D. M. Sabatini (2001). Microarrays of cells expressing defined cdnas. *Nature 411*(6833), 107–110.

Zong, N., M. Adjouadi, and M. Ayala (2005). Artificial neural networks approaches for multidimensional classification of Acute Lymphoblastic Leukemia gene expression data. *WSEAS Transactions on Information Science and Applications 2*(8), 1071–1078.