

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

© Copyright by Hamid Ghous, 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 Catchpole 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

Table of Contents	viii
List of Tables	ix
List of Figures	xii
Abstract	1
Author's Publications	6
Abbreviations and Glossary	9
1 Introduction	10
1.1 Research Objective	11
1.2 Significance	11
1.3 Problem statement	12
1.4 Contributions of the thesis	13
1.4.1 Finding functional relationships between genes	15
1.4.2 Comparing different non-linear and linear dimensionality reduction methods	16
1.4.3 Finding metabolic pathways in ALL	17
1.5 Thesis structure	18
2 Background	19
2.1 Acute lymphoblastic leukaemia (ALL)	19
2.1.1 Clinical and biological description of ALL	20
2.1.2 Genetic and molecular basis of ALL	21
2.2 Data relevant to studying ALL	22
2.2.1 Gene expression data	23

2.2.2	Genomic variation data	25
2.2.3	Public Ontologies	26
2.3	Data mining	33
2.3.1	Methods used in this thesis	36
2.3.2	Singular Value Decomposition	36
2.3.3	Principal Component Analysis (PCA)	38
2.3.4	Kernel Principal Component Analysis	39
2.3.5	Local Linear Embedding (LLE)	40
2.3.6	Stochastic Neighbour Embedding (SNE)	40
2.3.7	Diffusion Maps (DM)	41
2.3.8	Random forest	42
2.4	Summary	43
3	Literature review	44
3.1	Finding functional relationships between genes	44
3.1.1	Defining similarity measures using GO annotations	45
3.1.2	Applying unsupervised methods to visualise functional relationship between genes	47
3.2	Microarray based classification of patients	49
3.2.1	Feature selection	50
3.2.2	Dimensionality reduction methods with biological data	51
3.3	Microarray analysis of metabolic pathways	58
3.3.1	Finding metabolic pathways associated to disease	58
3.3.2	Machine learning methods to predict metabolic pathways	60
3.4	Summary of research gap	63
4	Finding functional relationships between genes	66
4.1	Introduction	66
4.2	Experimental Design	68
4.2.1	Data Sets	68
4.2.2	Incorporating functional information into the SVD	72
4.3	Results and Discussion	78
4.3.1	Visualising cancer data set	80
4.4	Summary	83
5	Visualising Leukaemia Cancer Dataset using NLDR	100
5.1	Introduction	100
5.2	Experimental Design	101
5.2.1	Dataset	102

5.2.2	Attribute selection	105
5.3	Results and Discussion	106
5.4	Identification of the most suitable method on the basis of AUC results	106
5.4.1	Biological interpretation of results	122
5.4.2	Impact of PCA application prior to nonlinear dimensionality reduction on results	127
5.5	Summary	128
6	Case Study: Finding pathways related to ALL from SNPs using random forest	130
6.1	Introduction	130
6.2	Dataset	131
6.3	Experimental Design	131
6.4	Results	136
6.4.1	Finding high-ranked SNPs using random forest	136
6.4.2	Finding pathways related to ALL	144
6.4.3	Functional visualisation of genes	151
6.5	Summary	159
7	Conclusion	162
7.1	Functional relationship between genes	163
7.2	Visualising leukaemia cancer dataset using dimensionality reduction methods	165
7.3	Case Study: Finding pathways related to ALL using random forest .	168
7.4	Limitations and Future suggestions	169
7.4.1	Limitations	170
7.4.2	Future directions	171
	Appendix	174
	Bibliography	190

List of Tables

2.1	Example of three genes from the Gene Ontology.	29
4.1	Genes in the KEGG dataset listed by class identifier. Column 1: class number Column 2: KO terms describing class and associated genes. .	71
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.	80
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.	81
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.	84
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.	85
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.	86
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.	99

5.1	Summary of gene expression dataset for ALL patients used in this chapter.	104
5.2	AUC values calculated with different percentage of training, validation and test data based on mean	109
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	120
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	121
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	121
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	121
5.7	Biological data for interesting patients	123
5.8	Regression results for patient pairs of GSM180185, GSM180188 and GSM180178 and GSM180179	127
6.1	kappa values for different training and test set	139
6.2	Confusion matrix for 30% test set for healthy control and ALL patients respectively, predicted by the random forest method.	139
6.3	Confusion matrix for 20% test set for healthy control and ALL patients respectively, predicted by the random forest method.	140
6.4	Kappa values for different training and test set	140
6.5	Confusion matrix for test data with 200 trees for healthy control and ALL patients respectively predicted by the random forest method. . .	141
6.6	Confusion matrix for test data with 500 trees for healthy control and ALL patients respectively, predicted by the random forest method. . .	141

6.7	Confusion matrix for test data with 1000 trees for healthy control and ALL patients respectively, predicted by the random forest method. . .	142
6.8	Count of ‘0’, ‘1’, ‘2’ and ‘-1’ for top 14 SNPs in ALL dataset and healthy control dataset.	144
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.	145
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.	146
6.11	Top 14 SNPs found through the random forest method with gene annotation, gene description and diseases associated to them.	148
6.12	Pathways found related to leukaemia cancer from top 80 genes.	150
6.13	GO Terms found in clusters A and B during analysis of PC3 and PC4 as shown in Figure 6.7.	158
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.	159

List of Figures

2.1	An example of a dataset record in GEO for the gene GDS4299. . . .	25
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.	30
2.3	Example of the hierarchical structure of KEGG database.	31
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, <i>Escherichia coli</i> . (Ogata, 1998).	34
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.	52
4.1	The number of terms associated with each gene in the KEGG dataset. Genes are ordered in increasing number of terms.	70
4.2	The frequency of terms having a direct association to various numbers of genes in the KEGG dataset.	70
4.3	Distribution of the number of terms for genes in the cancer dataset. Genes are ordered by increasing number of terms.	73

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.	74
4.5	Experimental design for finding functional relationship between genes.	77
4.6	Distribution of values in the proximity matrices. (a) hop-based proximity matrix (b) information-content proximity matrix.	87
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: \circ is gene.	88
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: \circ is gene, \bullet is term.	89
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, \circ is biological process GO term and Δ is cellular component term.	90
4.10	(a) Principal components (PC) 2 and 3 for hop-based method and (b) PC4 and 5 for IC method. Legend: (\circ) is genetic information processing genes and $(+)$ represent carbohydrate metabolism genes.	91
4.11	Plot of principal components 4 and 2 for \mathbf{U} matrix (genes) for the KEGG dataset using the hop-based approach. Genes are related to KEGG categories for ribosomes (\circ), RNA polymerase (Δ), transcription ($+$), pentose phosphate pathway (\times) and pentose and glucuronate interconversions (\diamond).	92

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: \circ is gene.	93
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.	94
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.	95
4.15	Plot of principal components 2 and 3 of cancer dataset with cellular component (CC) terms. (a) Hop-based similarity measure. Legend: (\bullet) is genes and (\diamond) is CC terms.(b) IC similarity measure. Legend: (\bullet) is genes and ($+$) is CC terms.	96
4.16	Plot of principal components 2 and 3 of cancer dataset with biological process(BP) terms. (a) Hop based similarity measure. Legend: (\bullet) is genes and (\diamond) is BP terms (b) IC similarity measure. Legend: (\bullet) is genes and (\circ) is BP terms.	97
4.17	Plot of principal components 2 and 3 of cancer dataset with molecular function terms. Legend: (\bullet) is genes and (\diamond) is molecular function terms. (a) Hop based similarity measure (b) IC similarity measure.	98
5.1	Experimental design for classification of ALL patients based on relapse status.	103
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.	105

5.3	Box-plot comparison of results with data distribution as 65% training, 15% validation and 20% test data for all the methods.	110
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.	110
5.5	The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for PCA. Legend: + is relapsed patients, o is non-relapsed patients	111
5.6	The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for SNE. Legend: + is relapsed patients, o is non-relapsed patients	112
5.7	The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for kPCA. Legend: + is relapsed patients, o is non-relapsed patients	113
5.8	The resulting plot of patients into dimension1 (Dim1) and dimension 9 (Dim9) for LLE. Legend: + is relapsed patients, o is non-relapsed patients	114
5.9	Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for kPCA. Legend: + is relapsed patients, o is non-relapsed patients	116
5.10	Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for SNE. Legend: + is relapsed patients, o is non-relapsed patients	117
5.11	Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 9 (Dim9) for LLE. Legend: + is relapsed patients, o is non-relapsed patients	118
5.12	Highlighted interesting patient pairs in resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for PCA. Legend: + is relapsed patients, o is non-relapsed patients	119

5.13	The resulting plot of patients into dimension1 (Dim1) and dimension 2 (Dim2) for DM. Legend: + is relapsed patients, o is non-relapsed patients	120
5.14	The regression plot of gene expression values for patients GSM180178 and GSM180179. The o (black) are gene expression values for each probe	124
5.15	The regression plot of gene expression values of patients GSM180185 and GSM180188. The o (black) are gene expression values for each probe	125
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	126
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.	133
6.3	Experimental design for finding pathways related to ALL.	135
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	138
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.	143
6.6	GO terms projected into PC2 and PC3 form clusters associated with the sub-ontology. Legend: \diamond (black) is the molecular function GO term, o (green) is the biological process GO term, and + (red) is the cellular component term	153

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: \diamond (black) is the molecular function GO term, \circ (green) is the biological process GO term and + (red) is the cellular component term	154
6.8	Two outliers ‘TGF β 1’ and ‘NOS1’ found through gene-set enrichment analysis. Legend: \circ is gene.	155
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: \circ is gene.	156
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: \circ is gene.	157

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 Conference on Machine Learning and Principles 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	