Modelling, Data Mining and Visualisation of Genetic Variation Data

A Thesis Submitted for the Degree of
Doctor of Philosophy

By

Ahmad A. Aloqaily

in

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY
UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA
2012

© Copyright by Ahmad A. Aloqaily, 2012

CERTIFICATE

Date: **2012**

Author: **Ahmad A. Aloqaily**

Title: **Modelling, Data Mining and Visualisation of Genetic Variation Data**

Degree: **Ph.D.**

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Author

Acknowledgements

I wish to express my gratitude to many people who have inspired me both personally and professionally. First and foremost, I would like to thank my supervisors, Dr. Paul Kennedy, Prof. Simeon Simoff and Dr. Daniel Catchpoole. They introduced me to the fields of bioinformatics. I never would have had the courage to go forward with the methodology and theoretical research without their tremendous confidence in me, much more than I had in myself.

They always encouraged and challenged me to think bigger and better. I will always be grateful for all the time and attention that they has invested in me. I can hardly think of anything that has been achieved without their help, as they offer me the freedom to explore new ideas independently.

I would also like to acknowledge the Faculty of Engineering and IT at UTS for offering me a good research environment. I am also grateful to my fellow students, for their encouragement and friendships. I would also like to thank Nicholas Ho, computational biologist at the Children's Cancer Research Unit at the Children's Hospital at Westmead, for assisting me with some of the work on biological insights.

I gratefully acknowledge the funding sources that made my PhD work possible. I was sponsored by the Faculty of IT, Hashemite University - Jordan. My work was also supported by the Australian Rotary Health Research Fund (ARHRF). In terms of external datasets that have been used in this research study, I acknowledge the use of genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. I also acknowledge the use of a genome-wide SNP dataset generated at the St Jude Children's Research Hospital, the dataset includes 242 patients with acute lymphoblastic leukaemia treated at St Jude Hospital, USA.

Lastly, my special thanks go to my parents, brother and sisters for their constant support and encouragement, and more importantly their faith in me over many years. My parents, I can never thank you enough for whatever you have done for me. Last, but not least, I wish to send personal thanks to my beloved wife Malak, who has been with me for nearly 5 years already and sacrificed a lot for helping me to pursue my academic pathway. Thank you for your patience and care, and for accompanying me to go through highs and lows accentuated by my research studies.

To My Family,

My Wife and Our Angel, Tala.

Table of Contents

Table of Contents **ix**

List of Tables **x**

List of Figures **xiii**

Abstract **xvi**

1 Introduction **1**

1.1 Genetic Variation Studies 5

1.2 Motivation and Challenges 6

1.3 Problem Statement 8

1.4 Scope and Contributions of the Thesis 9

1.5 Thesis Outline 17

2 Literature Review and Background **19**

2.1 Genetic Variations 24

2.1.1 Basic Concepts 24

Haplotype, Genotype and Phenotypes 24

Linkage Disequilibrium and Block Structure of Human Genome 26

2.1.2 Genome-wide Association Analysis 28

Linkage Analysis 29

Candidate-gene Studies 30
Genome-wide Association Studies 32
2.1.3 Genome-wide Association Study Approaches 33
Indirect Approaches 35
Gene-centric Approaches 36
2.1.4 Markers for Genome-wide Association Studies 38
2.2 Computational Analysis of GWA Studies 42
2.2.1 An Overview 42
2.2.2 Preliminary Analysis 45
2.2.3 Computational Haplotype Analyses 50
2.2.4 Tests for Association 52
Single-SNP Analysis 52
Multiple-SNPs Analysis 55
2.2.5 Evaluating the Statistical Significance of Putative Findings . . 58
2.3 Relevance to the Thesis 60
2.4 Data Mining and Machine Learning Methods 61
2.4.1 Current Directions in Genetic Variation Studies 62
2.4.2 Supervised-based Methods 66
Support Vector Machines 67
Random Forests 73
2.4.3 Unsupervised-based Methods 77
Principal Components Analysis 77
Multidimensional Scaling 78
Stochastic Neighbour Embedding 80
Curvilinear Component Analysis 83
Laplacian Eigenmap 86
Locally Linear Embedding 88
2.5 Case Study 90

3 Feature Selection, Weighting, Prioritizing and Distance Metric Measure for SNP Data 91

3.1 Dealing with Genetic Variation Data: The Proposed Approaches . . . 94

3.2 SNP Selection and Weighting Based on Random Forests 95

3.2.1 A New Proposed Approach Based on Random Forests 97

3.3 Prioritizing SNPs for Evaluating Interaction Effects 101

3.3.1 A New Measure for Prioritizing SNPs 102

3.3.2 Selecting Markers Involved in Gene-Gene Interactions 107

3.4 Feature Construction Using Feature Induction 110

3.5 Distance Measure Calculation 113

3.5.1 A Random Forest Based Kernel Function 117

3.5.2 Minor-allele Frequency and Entropy Based Kernel Functions . 118

3.6 Discussion 119

4 Disease Classification Models for Patient Diagnosis and Prognosis Based on Genome-wide SNP Profiles 121

4.1 Problem Formulation and Description 125

4.2 Methods and Approaches 127

4.2.1 Disease Classification: a New Computational Framework . . . 128

4.3 Experimental Design 138

4.3.1 Genome-wide SNP Data 138

4.3.2 Acute Lymphoblastic Leukaemia Datasets 140

4.3.3 Data Preprocessing 143

4.3.4 Experimental Procedures 144

Application to the Westmead Dataset 145

Application to the St Jude Dataset 147

4.4 Experimental Results 148

4.4.1 Genetic Variation Profile as Diagnostic Tool 148

Feature selection results 149
Classification results and comparisons 152
4.4.2 Genetic Variation Profile as Prognostic Tool 155
Feature selection results 156
Classification results 158
Comparing the classification performance 160
4.5 Biological Insights 165
4.6 Discussion and Conclusion 179

5 Visualizing Genome-wide SNP Profiles 183
5.1 Datasets and Data Preprocessing 185
5.2 The SNP Visualization: Problem and Approaches 187
5.2.1 Comparing Visualisations 189
5.3 Experimental Procedures 193
5.3.1 Experimental Settings 195
5.4 Results 196
5.4.1 Unsupervised Visualization of the Westmead dataset 196
Trustworthiness and Continuity 196
Sensitivity of NeRV and LocalMDS Methods 198
Distance Measures 200
Quality of Visualizations 200
5.4.2 Supervised Visualization of the Case-control Dataset 204
5.5 Summary and Discussion 209

6 Conclusion 213
6.1 Summary of Contributions 215
6.2 Limitations and Future Directions 220

Bibliography 245

List of Tables

2.1 Types of SNPs and their properties (from Tabor et al. (2002)). 40

2.2 Typology of SNPs and their occurrence (from Risch (2000)). 42

4.1 Summary of used ALL datasets 140

4.2 Subtype distribution of the St Jude ALL cases. 143

4.3 The excluded SNPs for both the Westmead and the St Jude datasets. 144

4.4 Feature construction and selection for each way of interactions using the highest 1000 IE SNPs. 151

4.5 Feature selected for marginal effects and each level of interaction. . . 152

4.6 Statistical comparison of performance measures for the Westmead dataset including selected SNPs based on marginal effects and combined marginal and interaction sets. Values are mean (Standard Deviation) of each estimated performance measure over 10 runs 154

4.7 Statistical comparison of performance measures for the Westmead dataset including selected SNPs based combined marginal and interaction sets, and the same features without constructing interaction SNPs. Values are mean (standard deviation) of each estimated performance measure 155

4.8 Summary of selected SNPs for each of ALL subgroup applied to the St Jude dataset 158

4.9 Prediction performance of SVM models for ALL subgroups based on three feature selection methods, namely, the proposed RF-RFE, SVM-RFE and VarSelRF methods. Using 10-fold cross validation, the average accuracy based on training and test data is reported. Values are mean (standard deviation) of the reported 10-fold accuracies 163

4.10 The GO-BP analysis of genes associated with reported SNPs. Biological functions were based on edited terms from the GO-BP database . 167

4.11 The KEGG analysis of genes associated with reported SNPs. Biological functions were based on edited terms from the KEGG database . . . 171

4.12 The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, T-ALL, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them. 172

4.13 The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, TEL-AML, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them. 174

4.14 The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, Hyperdiploid >50C, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them. 175

4.15 The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, E2A-PBX1, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them. 176

4.16 The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, MLL, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them. . . 177

4.17 The functional annotation analyses based on DAVID database of genes associated with reported SNPs for ALL subtype, BCR-ABL, including SNP name, chromosome location, gene symbol and gene description. The cells with "-" are for those SNPs without genes attached to them. 178

List of Figures

2.1 Haplotypes and Genotypes 26

2.2 General Framework of GWA studies 44

2.3 The optimal separating hyperplane of a SVM model in a linearly separable case (in the case of 2 dimension feature space). The optimal separating hyperplane is the solid line. Support vectors are the data points that lie on hyperplanes (the dashed lines) with maximal distance to the optimal separating hyperplane. 70

2.4 A SVM model with an optimal separating hyperplane in a linearly non-separable case. 71

2.5 The concept of a SVM mapping procedure, which maps training data non-linearly into a higher dimensional feature space (a case of mapping 2-D input space to 3-D feature space). 73

3.1 The allelic distributions of two SNPs. Both of these SNPs are reported to have no association with phenotype 103

3.2 The allelic distributions of two SNPs that have no association to a disease but with different allelic distributions. 106

3.3 The allelic distributions of two SNPs that have associations to a disease but with different allelic distributions. 106

3.4 Summary of steps involved in constructing a new multi-locus attributes using the MDR method: each multi-factor cell in n-dimensional space is labelled as either “high risk” or “low risk” based on the case to control ratios. For each multi-factor combination, distributions of cases (left bars in boxes) and of controls (right bars) are shown. 112

4.1 Disease diagnosis and prognosis classification Framework 130

4.2 The OOB error rates of the RF-RFE procedure applied to the Westmead dataset, as a function of number of SNPs maintained at each iteration of built models 150

4.3 ROC curves of the Westmead dataset based on the feature selection procedures (marginal, common and combined feature sets) 156

4.4 Box plots of AUC results of 10 runs based on marginal, common and combined feature sets 157

4.5 Box plots of the accuracies of the SVM classifiers based on 10 runs. The results are based on SVM-RFE and RF-RFE feature selection methods for T-ALL, TEL-AML and Hyper>50 subtypes 164

4.6 Box plots of the accuracies of the SVM classifiers based on 10 runs. The results are based on SVM-RFE and RF-RFE feature selection methods for E2A-PBX1, MLL and BCR-ABL subtypes 165

5.1 Trustworthiness of the mapping as a function of k that applied to the Westmead dataset, where k is the size of neighbourhood. Small neighbourhood sizes are the most important ones. PCA: Principal Component Analysis, LLE: Locally Linear Embedding, NeRV: Neighbour Retrieval Visualizer, LocalMD: Local Multidimensional scaling, LE: Laplacian Eigenmap and Rand_map: Random mapping. 198

5.2 Continuity of the mapping as a function of k applied to the Westmead dataset, where k is the size of neighbourhood. Small neighbourhood sizes are the most important ones. PCA: Principal Component Analysis, LLE: Locally Linear Embedding, NeRV: neighbour Retrieval Visualizer, LocalMDS: Local Multidimensional scaling, LE: Laplacian Eigenmap and Rand_map: Random mapping. 199

5.3 Trustworthiness of NeRV mapping as a function of k applied to the Westmead dataset, where k, the neighbourhood's size, is set by trustworthiness. The neighbourhood size, N, used by NeRV is ranging from 5 to 30. 201

5.4 Trustworthiness of LocalMDS mapping as a function of k applied to the Westmead dataset, where k, the neighbourhood's size, set by trustworthiness. The neighbourhood size, N, used by LocalMDS is ranging from 5 to 30. 202

5.5 Trustworthiness of NeRV mapping as a function of k applied to the Westmead dataset, where k is the size of neighbourhood. The lambda used by NeRV is ranging from 0 to 1. 203

5.6 Visualization of the Westmead data using NeRV method with $N = 30$ and $\lambda = 0.3$. 204

5.7 Visualization of the Westmead data using LocalMDS method with $N = 15$ and $\lambda = 0.2$ 205

5.8 The visualization of the case-control dataset based on the whole feature set 207

5.9 The visualization of the case-control dataset based on the marginally selected feature set 208

5.10 The visualization of the case-control dataset based on both selected marginal and interaction effect SNPs 209

Abstract

Data mining and knowledge discovery have been applied to datasets in various industries including biomedical informatics. The major challenges in data mining in the area stem from the fact that biomedical data comes in many forms with a highly dimensional nature. This research thesis focuses on one specific biomedical dataset, termed as genetic variation data in the form of genome-wide single nucleotide polymorphisms (SNPs) datasets.

Advances in single nucleotide polymorphism genotyping technologies have revolutionised our ability to explore the genetic architecture and models underlying complex diseases by conducting studies based on the whole genome. These studies are called genome-wide association studies. The basic strategy used in these studies is to examine the relationship between the disease of interest and genetic markers across the whole genome.

Many association studies have led to the discovery of single genetic variants associated with common diseases. However, complex diseases are not caused by single genes acting alone but are the result of complex non-linear interactions among genetic factors, with each gene having a small effect on disease risk. For this reason there is a critical need to implement new approaches that can take into account non-linear gene-gene interactions in searching for markers that jointly cause complex diseases.

Several computational methods have been developed to deal with the genetic complexity of complex diseases. However, testing each SNP for main effects and different orders of gene-gene interaction is computationally infeasible for such high-dimensional data. Also, these methods do not scale well. Therefore, there is growing interest in applying non-parametric predictive models including data mining and machine learning approaches to understand genetic variation data.

This thesis constructs models which incorporate genetic variation data in a manner that will alleviate the error induced by the high dimensionality of such data. Data mining approaches, specifically non-parametric ones, are developed for the modelling, exploration and visualization of patient-to-patient relationships based on genome-wide SNP data. This thesis focuses on three main issues in genetic variation studies: (1) feature selection and distance calculations, (2) framework for the task of disease diagnosis and prognosis, and (3) models for the comparison and visualisation of patient-to-patient relationships based on genome-wide SNP profiles.

This thesis proposes efficient feature selection approaches to find an optimal subset of markers with the highest predictive power for the disease of interest, while managing the large search space required. The proposed approaches select genetic markers for marginal effects as well as gene-gene interaction effects. Markers with marginal effects are selected with an iterative random forest (RF) based procedure, called RF-RFE. The importance measure generated by random forest was chosen for estimating the importance of each SNP (weighting) and facilitates the selection of an appropriate set of SNPs. To deal with the large search space involved in detecting gene-gene interactions, putative markers are prioritized in the search using a new measure, called Interaction Effect (IE), that quantifies the potential for a SNP to be involved in gene-gene interaction. This measure can also be used as a splitting criterion in random forest construction to define a cut-off value of a ranked list of SNPs. The prioritized SNP set is used to construct new combined features, which carry the information to account for gene-gene interactions.

This thesis proposes three new methods for calculating distances between genotype profiles based on a kernel-based weighting function including: RFK, using the RF variable importance measure; MAFK, based on the minor allele frequency measure and EK, using the entropy measure. The distances can be subsequently incorporated for the purpose of disease classifications, cluster analyses and visualizations.

The feasibility of using genetic variation data for disease diagnosis and prognosis is explored with a new computational framework. The framework demonstrates the use of different phases of data processing and modelling to build reliable disease diagnostic and prognostic models using genetic variation data. The proposed feature selection approaches are incorporated in the framework to select an optimal subset of SNPs with the highest predictive power.

The proposed framework is empirically evaluated using two case studies of acute lymphoblastic leukaemia. The results demonstrate that the framework can produce highly accurate diagnosis and prognosis models. This thesis shows that a significant improvement of models' performance requires including interaction markers. The results are consistent with known biology while the accuracy of the produced models is also high.

Finally, several data reduction methods are used to visualize genetic variation data. For unsupervised-based visualization, they are compared based on the trustworthiness metric. For the supervised-based visualization, the performance is compared based on class discrimination. This thesis finds that the Neighbour Retrieval Visualizer method shows the best results for unsupervised-based visualization. Furthermore, in the supervised-based approach, the results highlight the importance of using feature selection to remove insignificant features. The visualization has the potential to assist clinicians and biomedical researchers in understanding relationships between patients and has the potential to lead to delivery of advanced personalized medicine.

The methodologies and approaches presented in this thesis emphasise the critical role that genetic variation data plays in understanding complex disease. The availability of a flexible framework 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. A comprehensive validation of the methods and approaches embedded in the framework is a matter of applying this framework to other complex diseases.

Publications

Al-Oqaily, A., Tafavogh, S., Catchpoole, D., and Kennedy, P. 'A new computational framework for the task of disease diagnosis and prognosis', In preparation for submission to BMC Bioinformatics journal.

Al-Oqaily, A., Kennedy, P., Catchpoole, D., and S. Simoff, S., (2008). 'Comparison of visualization methods of genome-wide SNP profiles in childhood acute lymphoblastic leukemia', *Data Mining and Analytics 2008: proceedings of the Seventh Australasian Data Mining Conference (AusDM'08)*, **87**, pp. 111-121, Australian Computer Society.

Kennedy, P., Simoff, S., Catchpoole, D., Skillicorn, D., Ubaudi, F. and **Al-Oqaily, A.** (2008). 'Integrative visual data mining of biomedical data: Investigating cases in chronic fatigue syndrome and acute lymphoblastic leukaemia'. In S. J. Simoff, M. H. Boehlen, and A. Mazeika, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Springer-Verlag New York Inc, pp. 367-388.

The work in chapters 3 and 4 was turned into a project plan for a successful grant application: 'Implementing personalized medicine using global genomic similarity', Cancer Institute NSW Research Innovation Grant 2011, 10/RFG/2-23, Daniel Catchpoole, Paul Kennedy, $50,000.