Data Mining Methods for Single Nucleotide Polymorphisms Analysis in Computational Biology

LIU Yang

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Principal Supervisor: Prof. Michael K. Ng

Hong Kong Baptist University

August 2011
Single nucleotide polymorphism (SNP) is among the most common genetic variations since its widespread distributions. They are considered as abundant and invaluable markers in human genome, that is a potential powerful tool for both of genetic researches and applications in practice. A well-chosen set of SNPs can represent millions of common genetic variants throughout the genome and also give researchers a better understanding of disease association information. Hence, the real challenge in association studies lies in carefully selecting reliable marker alleles which are most likely responsible for disease and furthermore, well representing them in an intuitive manner.

This thesis addresses above problems by first presenting two data mining algorithms for the detection, selection, comparison and analysis of SNPs that associate with genetic disorders from genome-wide data. One of them is a clustering method, which employs a subspace categorical clustering algorithm to compute a weight for each SNP in the group of case samples and the group of control samples, and uses the weights to identify the subsets of relevant SNPs that categorize these two groups. The other one is a classification method called shrunken centroid method that can succinctly characterize each class (case and control) by shrinking each centroid with respect to the overall centroid by a certain threshold in a categorical manner, and detect association between a disease and multiple marker genotypes based on a set of relevant SNPs selected.

We also investigate the use of SNP networks for the interpretation of significant SNPs based on the detection of interacting and associated SNPs. In one aspect, SNP networks are constructed based on the selected ones obtained from our proposed shrunken centroid method. A statistical software PLINK is employed to compute the pair-wise SNP-SNP interactions, and pairs with smaller significant $P$-values than a defined threshold are chosen to identify an undirected and unweighted SNP network. Genes involved in this SNP network are further extracted. A gene-gene similarity value is computed using GOSemSim and gene pairs that have similarity values being greater than a threshold are selected to construct gene networks. Biological
relationships between these two forms of networks can be analyzed. In the other aspect, we present a novel method to mine, model and evaluate SNP sub-networks from a completed version of genome-wide network, which is constructed based on the SNP-SNP interactions from a logistic regression model implemented using PLINK. Then by using gene information, selected SNP seeds are employed to detect SNP sub-networks with a maximal modularity. Finally to identify functional role of each SNP sub-network, its gene association network is constructed and their functional similarity values are calculated to show the biological relevance. We also perform a classification analysis to demonstrate how the detected SNP sub-networks are used for disease association study.

Last, we incorporate DNA copy number variation (CNV) data derived from SNP arrays into a computational shrunken dissimilarity measure model and formalize the detection of copy number variations as a case-control classification problem. By shrinkage, the number of relevant CNVs to disease can be determined. The corresponding SNPs and genes will be investigated and SNP and gene networks can be constructed afterwards.
# Table of Contents

Declaration ........................................... i

Abstract .............................................. ii

Acknowledgements ........................................ iv

Table of Contents ........................................ v

List of Tables ........................................... viii

List of Figures .......................................... xvii

List of Abbreviations ..................................... xxiii

## Chapter 1 Introduction ................................ 1

1.1 SNP Overview ......................................... 1

1.2 Human Genome-wide Association Study .................. 3

1.3 Motivation ............................................ 5

1.4 Contribution ......................................... 6

1.5 Thesis Outline ........................................ 6

## Chapter 2 Literature Review ............................ 8

2.1 Computational Methods ................................ 8

2.1.1 Dimension Reduction ................................. 8

2.1.2 Logic Regression ..................................... 10

2.1.3 Support Vector Machine ............................... 12

2.1.4 Bayesian Statistical Methods ......................... 14
2.1.5 Random Forest .................................................. 16
2.2 Databases ......................................................... 17
2.3 Computational Tools ............................................. 19

Chapter 3 Data Mining Methodology for SNPs Analysis 21
3.1 Clustering: Subspace K-Mode (SKM-SNP) ......................... 21
  3.1.1 Introduction .................................................. 21
  3.1.2 Methods ..................................................... 23
  3.1.3 Experimental Results ....................................... 29
  3.1.4 Conclusions ................................................ 35
3.2 Classification: Shrunken Centroid Method ....................... 35
  3.2.1 Introduction ................................................ 35
  3.2.2 Methods ..................................................... 37
  3.2.3 Experimental Results ....................................... 40
  3.2.4 Conclusions ................................................ 48

Chapter 4 Construction and Analysis of SNP Networks 52
4.1 SNP Networks Constructed from Existing Results .................. 52
  4.1.1 Introduction ................................................ 52
  4.1.2 Methods ..................................................... 53
  4.1.3 Experimental Results ....................................... 55
  4.1.4 Conclusions ................................................ 59
4.2 Genome-wide SNP Networks ........................................ 60
  4.2.1 Introduction ................................................ 61
  4.2.2 Methods ..................................................... 64
  4.2.3 Experimental Results-Parkinson Disease (PD) Data .......... 67
  4.2.4 Experimental Results-WTCCC Coronary Artery Disease (CAD) Data ..................................................... 80
  4.2.5 Conclusions ................................................ 85

Chapter 5 Study of SNP and Gene Networks from Copy Number Variations 88
5.1 Introduction ....................................................... 89