IDENTIFICATION OF DISEASE RELATED SIGNIFICANT SNPs

by
CEYDA SOL

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
January 2010

IDENTIFICATION OF DISEASE RELATED SIGNIFICANT SNPs

APPROVED BY:

Assist. Prof. Dr. Nilay Noyan (Thesis Supervisor)	……
Assoc. Prof. Dr.Uğur Sezerman (Thesis Co-advisor)	……
Assist. Prof. Dr. Kemal Kılıç	……
Assoc. Prof. Dr. Ş. İlker Birbil	……
Assist. Prof. Dr. Yücel Saygın	……

DATE OF APPROVAL: ……20.01.2010……

To my family

Acknowledgments

I would like to thank all people who have helped and inspired me during my thesis study. I especially want to thank my advisors, Assist. Prof. Dr. Nilay Noyan and Assoc. Prof. Dr. Uğur Sezerman for their guidance and support from the initial level to the end. Assoc. Prof. Dr. Ş. İlker Birbil, Assist. Prof. Dr. Kemal Kılıç and Assist. Prof. Dr. Yücel Saygın deserve a special thanks as my thesis committee members. I am thankful to the genetic research specialist Deni Hogan for her consultancy and providing me a free access for SVS7 (SNP and Variation Suit) software. I would also like to thank Phil Sherrod to let me use the DTREG software freely during my research. My deepest gratitude goes to my family for their love and support throughout my life. I would also thank to my friends at my office, Belma Yelbay, Mahir Umman Yıldırım, Halil Şen, Cenk Cengiz, Tolga Dinçer and Ozan Erdem for their technical guidance and valuable friendship. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of my thesis.

IDENTIFICATION OF DISEASE RELATED SIGNIFICANT SNPs

Ceyda Sol

Industrial Engineering, Master of Science Thesis, 2010

Thesis Supervisor: Assist. Prof. Dr. Nilay Noyan,

Thesis Co-advisor: Assoc. Prof. Dr. Osman Uğur Sezerman

Keywords: genome wide association analysis, tag SNP selection, genetic algorithm, feature selection, association rule mining, SNP combination

Abstract

Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide in the genome sequence is altered. Since, variations in DNA sequence can have a major impact on complex human diseases such as obesity, epilepsy, type 2 diabetes, rheumatoid arthritis; SNPs have become increasingly significant in identification of such complex diseases. Recent biological studies point out that a single altered gene may have a small effect on a complex disease, whereas interactions between multiple genes may have a significant role. Therefore, identifying multiple genes associated with complex disorders is essential. In this spirit, combinations of multiple SNPs rather than individual SNPs should be analyzed. However, assessing a very large number of SNP combinations is computationally challenging and due to this challenge, in literature there exist a limited number of studies on extracting statistically significant SNP combinations. In this thesis work, we focus on this challenging problem and develop a five step "disease-associated multi-SNP combinations search procedure'' to identify statistically significant SNP combinations and the significant rules defining the associations between SNPs and a specified disease. The proposed five step multi-SNP combinations procedure is applied to the simulated rheumatoid arthritis data set provided by Genetic Analysis Workshop 15. In each step, statistically significant SNPs are extracted from the available set of SNPs that are not yet classified as significant or insignificant. In the first step, the genome wide association analysis (GWA) is performed on the original complete multi-family data set. Then, in the second step we use the tag SNP selection algorithm to find a smaller subset of informative SNP markers. In literature most tag SNP selection methods are based on the pair wise (two-markers) linkage disequilibrium (LD) measures. But in this thesis, both the pair wise and multiple marker LD measures have been incorporated to improve the genetic coverage. Up to the third step the procedure aims to identify individual significant SNPs. In the third step a genetic algorithm (GA)

based feature selection method is performed. It provides a significant combination of SNPs and the GA constructs this combination by maximizing the explanatory power of the selected SNPs while trying to decrease the number of selected SNPs dynamically. Since GA is a probabilistic search approach, at each execution it may provide different SNP combinations. We apply the GA several times to obtain multiple significant SNP combinations, and for each combination we calculate the associated pseudo r-square values and apply some statistical tests to check its significance. We also consider the union and intersection of the SNP combinations, identified by the GA, as potentially significant SNP combinations. After identifying multiple statistically significant SNP combinations, in the fourth and fifth steps we focus on extracting rules to explain the association between the SNPs and the disease. In the fourth step we apply a classification method, called Decision Tree Forest, to calculate the importance values of individual SNPs that belong to at least one of the SNP combinations found by the GA. Since each marker in a SNP combination is in bi-allelic form, genotypes of a SNP can affect the disease status. Different genotypes of SNPs are considered to define candidate rules. Then utilizing the calculated importance values and the occurrence percentage of the candidate rule in the data set, in the fifth step we perform our proposed rule extraction method to select the rules among the candidate ones. In literature there are many classification approaches such as the decision tree, decision forest and random forest. Each of these methods considers SNP interactions which are explanatory for a large subset of patients. However, in real life some SNP interactions that are observed only in a small subset of patients might cause the disease. The existing classification methods do not identify such interactions as significant. However, of the proposed five-step multi-SNP combinations procedure extracts these interactions as well as the others. This is a significant contribution to the research on identifying significant interactions that may cause a human to have the disease.

BİR HASTALIĞA İLİŞKİN ÖNEMLİ TEKLİ NÜKLEOTİD POLİMORFİZMLERİN BELİRLENMESİ

Ceyda Sol

Endüstri Mühendisliği, Fen Bilimleri Tezi, 2010

Tez Danışmanı: Yrd. Doç. Dr. Nilay Noyan

Yardımcı Tez Danışmanı: Doç. Dr. Osman Uğur Sezerman

Anahtar Kelimeler: genom ilişki analizi, genetik algoritma, tekli nükleotid polimorfizm (SNP), temsilci SNP seçimi, nitelik seçim metodu, kural madenciliği, SNP kombinasyonu

Özet

Genom dizilimindeki tek bir nükleotidin değişimi ile oluşan DNA dizilimindeki çeşitliliklere tekli nükleotid polimorfizm (SNP) denir. DNA dizilimdeki farklılıklar obezite, diyabet, romatoid artrit gibi kompleks hastalıkların oluşumunda önemli bir etkiye sahip olduğundan, SNP analizi kompleks hastalıkların tanımlanmasında giderek önem kazanmaktadır. Yakın zamandaki biyolojik çalışmalar, tek bir gendeki değişimin kompleks hastalıkların tanılanmasında zayıf olduğunu gösterirken, birden çok gen etkileşiminin önemli bir role sahip olduğunu işaret etmektedir. Bu nedenle, kompleks bir hastalığın teşhis edilmesinde hastalıkla ilişkili tek bir genden ziyade gen kombinasyonlarının incelenmesi gerekmektedir. Ancak insan genomunda çok fazla sayıda SNP bulunduğundan SNP kombinasyonlarının oluşturulması hesaplama açısından zor bir problemdir. Bu nedenle literatürde kompleks bir hastalıkla ilgili önemli SNP kombinasyonlarının çıkarılmasını ele alan çalışmaların sayısı oldukça sınırlıdır. Bu tez çalışmasının amacı bu zorlu problem üzerine yoğunlaşarak istatistiksel olarak önemli SNP kombinasyonlarını ve bu kombinasyonlardaki SNP'ler ile kompleks hastalık arasındaki ilişkiyi gösteren önemli ilişki kurallarının çıkarılmasıdır. Bu kapsamda beş aşamalı arama algoritması geliştirilmiş ve önerdiğimiz prosedür Genetic Analysis Workshop 15 tarafından sağlanan romatoid artrit SNP data setine uygulanmıştır. Prosedürün her bir aşamasında istatistiksel olarak önemli SNP'ler henüz önemli olup olmadığı belirlenmemiş mevcut SNP seti arasından seçilmektedir. Prosedürün ilk aşamasında orjinal SNP verisine genom ilişki analizi, ikinci aşamada ise daha küçük fakat daha bilgi verici SNP seti elde etmek için temsilci SNP seçim metodu uygulanmıştır. Literatürde birçok SNP seçim algoritması ikili bağlantı dengesizliği (pairwise linkage disequilibrium) ölçülerine dayalıdır. Bu tezde, en az sayıda SNP ile maksimum genetik bilgiye ulaşabilmek amacıyla hem ikili hem çoklu bağlantı

dengesizlik ölçü metotları kullanılmıştır. Üçüncü aşamaya kadar, önerdiğimiz prosedür SNP'lerin önemini bireysel olarak incelemektedir. Üçüncü aşamada ise genetik algoritmaya dayalı nitelik seçim metodu ile önemli SNP kombinasyonları elde edilmiştir. Genetik algoritma (GA), seçilen SNP sayısını dinamik olarak azaltmakta ve seçilen SNP'lerin açıklayıcı gücünü maksimize edecek şekilde SNP kombinasyonlarını oluşturmaktadır. GA olasılıklı arama yaklaşımı olduğu için algoritmanın her uygulanışında farklı SNP kombinasyonları elde edilebilir. Bu nedenle genetik algoritma birkaç kez uygulanmış ve birçok önemli SNP kombinasyonu elde edilmiştir. Daha sonra, her bir önemli SNP kombinasyonu için istatistik testleri ve ölçüm kriterleri (pseudo r^2) kullanılarak SNP kombinasyonlarının istatistiksel önemi kontrol edilmiştir. Ayrıca, belirlenmiş önemli SNP kombinasyonlarındaki ortak SNP'ler belirlenerek bu SNP'lerden yeni bir aday SNP kombinasyonu oluşturulmuştur. Dördüncü aşamada her bir kombinasyondaki en önemli 6 SNP'i belirlemek amacıyla karar ağacı ormanı sınıflandırma metodu uygulanmıştır. Kompleks bir hastalığın oluşumunda SNP genotiplerinin de önem taşıdığı düşünüldüğünden beşinci aşamada SNP'lerin farklı genotipleri aday kurallar olarak göz önüne alınmış ve önemli SNP kombinasyonlarındaki her bir SNP için aday SNP-genotip ilişki kuralları çıkarılmıştır. Beşinci aşamada aday ilişki kuralları arasından önemli kuralları seçmek için, hesaplanan önem değerlerinden ve aday kuralların görülme sıklığından yararlanılarak önerdiğimiz kural çıkarma metodu uygulanmıştır. Literatürde karar ağacı, karar ağacı ormanı, rassal orman gibi birçok sınıflandırma metodu kullanılmaktadır. Fakat bu metotların her birisi hasta insan populasyonunun çoğunluğunu açıklayan SNP etkileşimlerini dikkate almaktadır. Ancak gerçek hayatta bazı SNP etkileşimleri hasta insanların sadece çok küçük bir kısımda gözlemlenmektedir. Mevcut sınıflandırma metotları bu etkileşimleri tespit etmekte yetersiz kalmaktadır. Bizim önerdiğimiz beş aşamalı SNP kombinasyonu arama prosedürü ise hem bu ilişkileri hem de diğer sınıflandırma yöntemleri tarafından bulunan önemli ilişki kurallarını çıkarabilmektedir. Bu nedenle, önerdiğimiz beş aşamalı SNP kombinasyonu arama prosedürü ve ilişki kurallarının çıkarımı algoritması kompleks bir hastalığa neden olabilecek önemli SNP etkileşimlerinin incelenmesine ilişkin çalışmalara önemli bit katkı sağlamaktadır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Recently, SNP (single nucleotide polymorphisms) analyses have been receiving significant attention for developing new treatments against common complex diseases. A combination of genetic, environmental and even lifestyle factors may cause the complex disease. Thus, investigating the disease causing effects is not an easy task. Since complex diseases are not controlled by a single locus, analyzing SNP combinations would be more powerful to extract the susceptible gene or chromosomes related to the disease

In this study, we focus on the rheumatoid arthritis (RA) disease, which is a complex multi factorial disorder. It affects many joints and tissues and cause deformations of them. To determine possible genetic reasons of RA, we conducted a genome based analysis. Scientists have been investigating RA many years. According to these previous studies, we know some of the susceptible chromosomal regions which are associated with the disease. Although other chromosomes may affect the disease status, we just focus on chromosome 6 to test our results against the previous studies.

There is a wide literature on the SNP analysis for different objectives. For instance, the genome wide association or linkage based methods can be applied to determine the possible disease related SNPs from a SNP data (Freedman, 2004; Samani et al., 2007; Uh et al., 2007). In order to obtain a specified genetic coverage with the minimum number of SNPs a tag SNP selection method can be used (Gopalakrishnan, 2006; Sya et al., 2006; Hao, 2007; Wang et al., 2008). Data mining tools or classification methods can be performed to extract susceptible disease related genotypes (Murthy et al., 1995; Tong et al., 2003; Tong et al., 2004; Xie et al., 2005).

The aim of genome wide association (GWA) analysis is to determine disease susceptibility genes for complex disorders. By the help of this approach we can scan a large number of SNP markers in the human genome. The principle of GWA is based on

comparing allele, genotype or haplotype frequencies between patient and healthy people. In our study we scan 17821 SNP markers on chromosome 6 in human genome to detect RA disease related significant SNPs.

Tag SNP selection is an important method in designing case control association studies (Hao, 2007). Linkage disequilibrium measures which are based on pair wise correlation between SNPs are widely used for the purpose of designing association studies (Gupta, 2005). The goal is to minimize the number of markers selected for genotyping in a particular platform and therefore reduce the genotyping cost while simultaneously representing information provided by all other markers (Hao, 2007). Thus, the main advantage of tag SNP selection is obtaining a smaller set of SNPs, which includes most of the information in the original SNP set. In our study, we used Haploview-Tagger software for the tag SNP selection. The tag SNP selection algorithm of Tagger is based on both the pair-wise and multiple linkage disequilibrium.

Feature selection is a variable selection method which helps us to better understand the data and it is another powerful method to select a subset of disease relevant SNPs. This technique is also referred as the discriminative gene selection in the field of biology. Feature selection algorithms are used to determine influential genes related to the disease by removing most irrelevant and redundant SNPs from the data (Horne et al., 2004; Phuong et al., 2005; Saeys et al., 2007). In our study, our aim is to analyze disease susceptible SNP combinations not to analyze the effect of an individual SNP. Thus, we developed a feature selection method based on a genetic algorithm to determine disease related SNP combinations.

The machine learning techniques such as support vector machines, decision tree and decision forest are used to identify a set of disease causing SNPs. Machine learning is a scientific discipline that deals with the developing algorithms that let computers change behavior based on data. Among these techniques, decision tree and decision tree forest are widely used for the SNP classification, since they allow the use of both non-numerical and numerical values (Vlahou et al., 2003). Besides, the accuracy of decision forest and decision tree is higher than other methods (Murthy et al., 1995). Decision forest is a technique of combining the results of multiple classification models to produce a single prediction (Tong et al., 2003). Because most genetic data is noisy, a decision tree algorithm may not provide reasonable classification accuracy. However, when several decision trees are combined to produce a decision tree forest, classification accuracy considerably increases. Therefore, we preferred to use a decision

tree forest algorithm rather than a decision tree algorithm. We compute a significance value for each SNP of a SNP combination set by using the DTREG software. Consequently, we determine the most significant SNPs for each combination set.

In complex diseases, determining the most significant SNP combinations may not be adequate to explain the disease because different genotypes of a bi-allelic SNP may affect the disease status in a different way. While a homozygous genotype may be the reason of the disease, a heterozygote one may not. Thus, after determining significant SNP combinations, the genotype effect should be extracted. For this reason, we develop a decision rule procedure.

The reasons of having a complex disease have been studied for many years, but most of the studies focus on individual effects of SNPs. Since a complex disease is multi-factorial, a group of SNP effects should be investigated. Our genetic algorithm based feature selection method analyzes multiple SNPs simultaneously. Thus, our proposed approach is potentially more successful to explain the disease causing effects compared to individual SNP analysis methods. Besides, existing studies in general have computational difficulties to investigate more than two-SNP effects due to the memory and time limits. Fortunately, we are able to identify several-SNP effects in a reasonable time and without requiring too much memory. In addition, unlike the existing decision rule methods our method may detect rarely observed relations and so may provide a higher explanatory power. Moreover, there is no other study which combines all the bioinformatics approaches mentioned above; genome wide association analysis, optimal tag SNP selection, feature selection, decision tree forest and decision rule models. Thus, our study may be a useful guide for the complex disease analysis and contribute to literature and real-world practice.

CHAPTER 2

PREPROCESSING OF THE DATA: GENOME WIDE ASSOCIATION ANALYSIS AND RELATED WORK

The first step of our work is to apply genome wide association analysis (GWA) to determine disease susceptible SNPs and eliminate unrelated and redundant SNPs from the data. By applying GWA, we obtain a smaller set of potentially significant SNPs related to RA disease.

There are two different methods considering the whole genome to identify causative factors of a complex disease: genome wide linkage mapping and genome wide association analysis (GWA). Although genome wide linkage mapping is robust when two different alleles at a locus affect the disease susceptibility (allelic heterogeneity), it is not robust when two different alleles at different locus affect the disease susceptibility (locus heterogeneity). Linkage mapping is partially successful to determine the disease related genes or single nucleotide polymorphisms (SNPs) when heritability of a complex disease is low. Unlike the genome wide linkage mapping, the genome wide association analysis can be applied for both pedigree and case/control data sets. Risch et al. (1996) compare the two methods and mention that the genome wide association is a more powerful technique. Thus, we use the genome wide association method in our study instead of the linkage mapping. Before introducing GWA, a brief explanation of single nucleotide polymorphisms (SNPs) is given in below.

Single nucleotide polymorphism (SNP) is a variation in DNA sequence which occurs when a single nucleotide (A, T, C or G) in the genome differs between members of a species. For instance, two similar DNA sequences (AAGC**C**TA and AAGC**T**TA) are presented in Figure 2.1. The only difference in these sequences is the 5^{th} nucleotide (C and T). Each different sequence is called a SNP.

Study of SNPs is a key point in biomedical science to identify a function of a gene. In human genome, there are approximately 10 million SNPs some of which do not have

a significant role in developing the disease. Thus investigating whole SNPs allows us to identify associated SNPs with the risk of developing a disease.

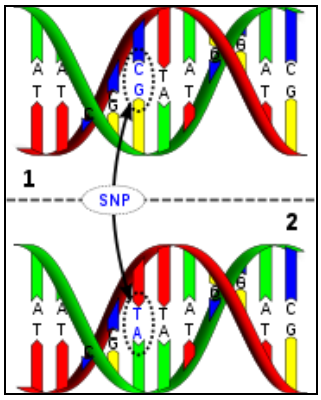

Figure 2.1. Two DNA molecules with a polymorphism

2.1. Genome Wide Association Analysis

GWA is a method to investigate millions of susceptible SNPs to associate them to a specific disease. GWA focuses on comparing the genetic variation between case (individuals having the specified disease) and control (individuals not having the disease) groups. It is based on the idea that if the genetic variation at a gene location is observed more frequently in case groups than in control groups, this variation is considered as strongly the reason of the disease. Currently, GWA has been applied for many complex diseases: obesity (Johansson et al., 2009), breast cancer (Zheng et al., 2009), type 2 diabetes (McCarthy et al., 2009), myocardial infarction (Kathiresan et al., 2009) and Alzheimer (Waring et al., 2008). Genome wide association analysis has six main steps:

- Collecting genomic data: selection of case and control groups
- DNA isolation, genotyping and quality control of SNPs
- Analysis of population stratification
- Statistical tests for SNP association
- Looking up potentially significant SNPs
- Replication of identified association in an independent population

In our study, GWA is applied by the help of the genome wide analysis module of SVS7 software (SNP and Variation Suit) which is developed by Golden Helix Team.

2.1.1. Collecting Genomic Data

The data in our hand is a simulated rheumatoid arthritis data which is provided by the Southwest Foundation for Biomedical Search Group (Genetic Analysis Workshop 15, 2006, GWA15). GWA team firstly generated a population including two million families each of which including 2 parents and 2 offspring with the RA status. 100 random samples, including 2000 controls (none of the individuals in the family has the disease status) and 1500 case families (including affecting sibling pair (ASP) and affected or unaffected parents), are created from the entire population. Each of 100 replicates (the random sample) includes all the individuals of 1500 case families and just one randomly selected individual of a control family.

In GWA analysis, the selection of case and control groups from the same population is a crucial issue. The previous related studies reveal that DR type at the HLA locus on chromosome 6 of human beings has strongly affected the RA status. Thus, we investigate a very dense map of 17820 SNPs on chromosome 6 rather than considering the whole chromosomes. We need to have three different data files, including phenotype, genotype and map information. Phenotype data consists of family id, individual id, father id, mother id, sex and rheumatoid arthritis affection status (2=affected, 1=unaffected). Individual IDs are unique integers within each replicate. All SNPs in the data are in diallelic form and are coded as 1 and 2. In the map data, chromosome number, marker name and physical location in base pairs are reported. There is no missing SNP information on all family members in the data.

Moreover, although the original data includes some genotyping errors, these errors are not modeled for 100 replicate samples. In addition, there is no false phenotype information. To upload our data to SVS, we first write a C++ code to convert the data to SVS7 input format.

2.1.2. Genotyping and Quality Control

2.1.2.1.1. Filtering Poor Quality SNPs

Before statistical testing, we filter poor quality SNPs from data according to some quality metrics: call rate, minor allele frequency and Hardy Weinberg Equilibrium (HWE).

Call rate: We drop SNPs that can not satisfy the specified call rate (0.90).
Minor allele frequency (MAF): MAF indicates the frequency of a less common allele of the SNP at a locus that is observed in a specific population. If we select SNPs having lower MAF values in the data, we need to select more tag-SNPs to capture the whole variation in the population. Since our aim is to find a minimum number of SNPs associated to the disease, we desire higher MAF values. Generally, the most appropriate MAF value is 0.01. Thus, we drop SNPs having a MAF value smaller than 0.01.

2.1.3. Detect Population Stratification

Since population stratification may cause false positive results in the analysis, assessing the impact of population stratification is a significant part of GWA analysis. Population stratification indicates the differences in allele frequencies between case and control groups resulted from different ancestries rather than the association between the diseases. Population stratification (population structure) is analyzed by comparing the observed association between SNPs and the disease with the expected association statistics under the null hypothesis of no association. The deviations from the null distribution are assessed by quantile-quantile plot (Q-Q plot). In y axes of Q-Q plots, the observed association statistics (chi-square statistic or $-\log_{10}p$) of each SNP are displayed in an increasing order. In x axes of Q-Q plots, expected association statistics under the null hypothesis (such as chi-square) are displayed. If there is a deviation from the identity line, either the assumed distribution is incorrect or the sample includes true associated SNPs.

In Figure 2.2.A the black line points out the expected chi-square statistics under the null hypothesis of no association. The dark blue line indicates the observed chi-square statistics including all SNPs and the light blue line shows the observed chi square statistics when the most strongly associated SNPs are excluded from the data. Figure 2.2.B refers to the observed and expected chi-square statistics of SNPs after the population stratification is adjusted. After adjustment, the observed chi-square statistics of SNPs converges to expected chi-square statistics which indicates the existence of population stratification in the data.

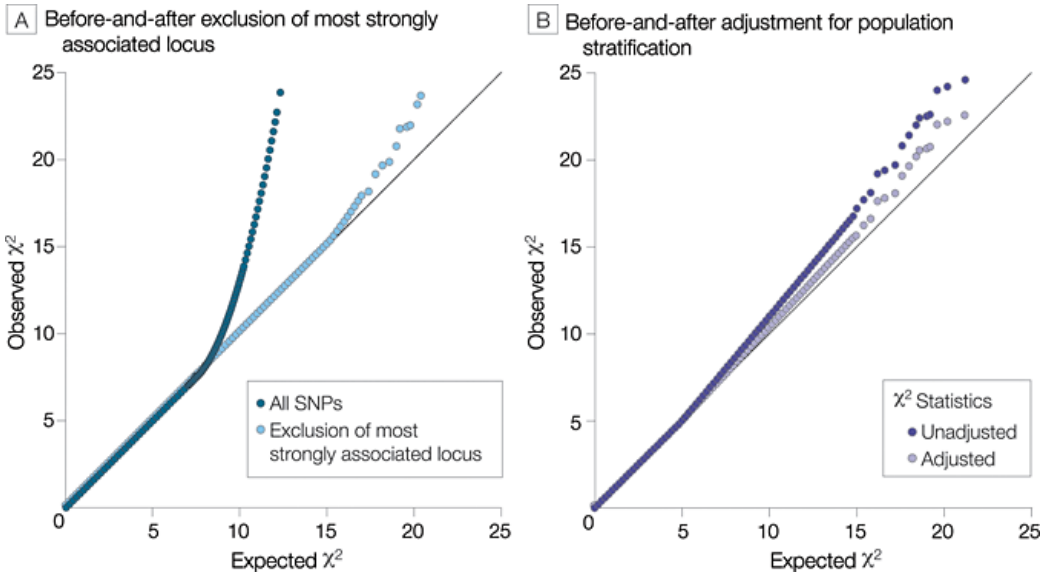

Figure 2.2. Quantile – Quantile Plots (A and B)

Another method to analyze the population stratification is the principal component analysis (PCA). Since we do not know the statistically significant SNPs in the beginning of the study, we applied the genotypic principal component analysis which uses the "EIGENSTART" PCA technique developed by Price (AL et al., 2006).

Firstly we compute the principal components by finding up to top 50 components. For further information about principal component formulas, you can read "SNP and Variation Suite (SVS)" Manual. We then plot eigenvalues of principal components to determine the number of principle components to be extracted from the data. The largest eigenvalues correspond to principal components. According to "EIGENSTRAT" PCA technique the first principal component or the first few principal components correspond directly to the stratification patterns. Therefore, after determining the top k (user defined value) principal components, these patterns should be removed from both the SNP data and dependent variable data by using vector-analysis related techniques. SVS automatically detects these patterns and removes them from the data and provides a corrected input data to the user. To be sure about removing the patterns, SVS also provides a PCA outlier removal option. To do this, we select the number of principal components involving in this process and standard deviation threshold to remove outliers. After correcting the population stratification, genotype association tests will be applied to the corrected data.

2.1.4. Genotype Association Testing

Although SVS provides many association tests, the only statistical test which is available for corrected data is the correlation trend test.

2.1.4.1. Correlation/Trend Test

Correlation/Trend test is used to test the significance of correlation between two numeric variables. Suppose that we have n pairs of observations, x_i for (i=1, 2…n) indicating the SNP value and y_i indicating the disease status. The correlation between x_i and y_i, denoted by R, is:

$$R = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} = \frac{\sum x_i y_i - \sum x_i \sum \frac{y_i}{n}}{\sqrt{\left(\sum x_i^2 - {(\sum x_i)^2}/{n}\right)\left(\sum y_i^2 - {(\sum y_i)^2}/{n}\right)}} \quad (2.1)$$

R^2 approximates a chi-square statistic with (n-1-k) degrees of freedom, where k is the number of principal components that is removed from the data. This chi-square statistics allow us to find a p value.

$$x^2 = (n - 1 - k)R^2 \quad (2.2)$$

2.1.4.2. Bonferroni Correction

Bonferroni correction is a method used for multiple dependent or independent hypothesis testing comparisons. According to Bonferroni rule, if we want the overall significance level of the whole set to be equal to α, each individual hypothesis must be tested at α/n significance level where n is the total number of hypothesis. By reducing the alpha value, we can avoid false positive results or in other words type 1 error. Type1 error is the rate of rejecting the null hypothesis when the null hypothesis is true. In our study, the null hypothesis refers to the case of not having the disease.

2.1.4.3. False Discovery Rate (FDR)

False discovery rate is the expectation of proportion of false positives to total positives in the data. FDR controls the type-1 errors in the analysis.

$$FDR = E\left[\frac{false\ positives}{false\ positives + true\ positives}\right] \quad (2.3)$$

2.1.5. Looking Up Potentially Significant SNPs

We firstly list the correlation/trend test p values in an increasing order and then select the SNPs having a p value smaller or equal to specified significance level (α/n) for further analysis. After determining the statistically significant SNPs, we isolate the non-significant ones from the data and construct a new subset of SNP data.

2.1.6. Replication of Identified Association in Independent Populations

The replication of genome wide association analysis in independent populations is significant to reduce the number of false-positive results. A false positive result refers to a SNP which is found to be related to the disease although it has no effect on developing the disease. To eliminate these results, we perform seven replication studies with different case and control populations. Each replication data includes the same SNP set (17820 SNPs on chromosome6).

CHAPTER 3

PREPROCESSING OF THE DATA: OPTIMAL TAG SNP SELECTION

The current genotyping technologies are not adequate to genotype all SNPs in all genes although the number of SNPs at a gene is finite (Nickerson et al., 2000). Thus, a set of informative SNPs should be chosen to use existing technology. Consequently, theoretical approaches have been developed for many years to choose a set of informative SNPs. Carlson et al. (2004) mention that investigating all SNPs is inefficient, because some of these SNPs are strongly correlated and they can provide the same information. The technique of selecting a set of minimum number of SNPs which provides maximum information about unselected SNPs in the data based on the correlation between SNPs is called the tag SNP selection procedure. There exist many publications about tag SNP selection based on linkage disequilibrium statistics (Gopalakrishnan et al., 2005; Syam et al., 2006; Hao K., 2007; Wang et al., 2008).

Pearson et al. (2008) state that SNPs which are located nearly each other are tend to be inherited together more often than expected by chance, and this nonrandom association is called the linkage disequilibrium. If a SNP has high linkage disequilibrium with another SNP, they are almost always inherited together. Thus, if we know the information that one of these SNPs is related to the disease, we can easily state that the other SNP may strongly be related to the disease as well. Linkage disequilibrium for a SNP pair is quantified by the help of a correlation measure. This correlation measure indicates the proportion of variation of one SNP explained by other SNP and it can only take the values between 0 and 1. If a SNP pair has a correlation value bigger than a pre-specified value (generally 0.8), those SNPs are supposed to be related to the disease. Linkage disequilibrium (D) and correlation (R^2) measures are calculated as in below.

$$D = p(AB) - p(A) * p(B$$

$$p(A) = probability\ of\ allele\ A\ at\ first\ SNP(marker)$$

$$p(a) = probability\ of\ allele\ a\ at\ first\ SNP(marker)$$

$$p(B) = probability\ of\ allele\ B\ at\ second\ SNP(marker)$$

$$p(b) = probability\ of\ allele\ b\ at\ second\ SNP(marker)$$

$$R^2 = {}^{D}\Big/\sqrt{p(A) * p(a) * p(B) * p(b)}$$

Most tag SNP selection studies are based on the pair-wise linkage disequilibrium. Shyam et al. (2006) study tag SNP selection based on the pair-wise linkage disequilibrium criteria to minimize the number of selected SNPs while obtaining maximum information provided by all SNPs. Although pair-wise linkage disequilibrium methods provide reasonable solutions, researchers have also focused on multiple linkage disequilibrium based tag SNP selection algorithms. Hao K. (2007) proposes a tag SNP selection method which is based on the multiple marker linkage disequilibrium. He develops Carlon's Greedy algorithm method (Carlson et al., 2003; Carlson et al., 2004). The proposed method by Hao includes both pair-wise and multiple SNP linkage disequilibrium of nearly located SNPs. Wang and Jiang (2008) propose a new greedy algorithm by considering the method of Hao. Their method is more efficient in terms of time and memory. While Hao's aim is to find a tag SNP set which can cover most of the data, Wang and Jiang can find a SNP set which covers all the SNP in the data with less time and memory usage. Barrett et al. (2005) also develop a tag SNP selection algorithm based on both the pair-wise and multiple correlations. This algorithm has been integrated in Haploview software which is developed by The Broad Institute of MIT and Harvard in 2004. We used Haploview-Tagger module to find optimal tag SNPs among the set of SNPs which are obtained at the end of genome wide association analysis.

3.1. Haploview Tagger Module

Haploview Tagger algorithm works in two steps. First, it selects tag SNPs based on the pair-wise linkage disequilibrium, which is similar to Carlson's Greedy approach. In the second step, it searches SNPs based on the multiple linkage disequilibrium (multi-marker haplotype) to improve tagging performance. Multi-marker correlation measures are calculated similar to the pair-wise correlation. The only difference is the multi-

marker approach uses haplotype instead of single SNPs. Thus, it calculates the correlation between haplotype blocks. A haplotype is a haploid genotype; it is a set of closely linked SNPs that are tend to be inherited together. Haploview Tagger has an option to force specific SNPs as tag SNPs not to exclude them from the further analysis. According to previous studies of RA disease and the results obtained for the GWA15 simulated data, SNP3437 is strongly related to the disease. Thus, in all tag SNP selection processes, we use this option not to exclude SNP3437 before implementing the genetic algorithm based feature selection process. Haploview Tagger algorithm needs haplotype blocks for multi-marker correlation calculations. Therefore, before running the Tagger algorithm, we form linkage disequilibrium blocks based on the Gabriel's algorithm (Gabriel et al., 2002). Then we determine the tag SNP selection criteria. We ignore pair-wise comparisons of SNPs which have a distance bigger than 300 kb apart. This avoids the selection of SNPs which are too far from each other. We also set correlation threshold as 0.8 and LOD (log of odd ratio) score as 3.0. LOD score is a statistical estimate of whether two loci are likely to lie near each other on a chromosome and are therefore likely to be inherited together as a package (Breiman, 1999). Finally, we set the minimum distance between tag SNPs as 0 bp and run the Tagger algorithm. The Haploview Tagger output provides us with the tag SNPs set, captured SNPs set and a coverage ratio. The captured SNPs are the SNPs which are not selected as the tag SNPs but can be explained by the tag SNP sets. The coverage is the percentage of alleles which are explained by the tag SNPs set. At this stage, we obtain the potentially informative disease related SNPs and the next step is to find the disease related SNP combinations. For this reason, we develop a genetic algorithm based feature selection method, which will be explained in detail in Chapter 6.

3.2. Detecting Colinearity between TAG SNPs

Since we select tag-SNPs according to linkage disequilibrium measures, it is most likely to include correlated SNPs in the constructed tag-SNP set because a tag SNP is highly correlated with its neighboring SNPs. In genetic algorithm based feature selection method, we use the method of logistic regression to construct SNP combinations. However, considering correlated SNPs as predictor variables in a regression analysis can lead misleading results. For example, some of the estimated coefficients in the regression equation can even have opposite signs. Thus, excluding correlated SNPs

from the further analysis is crucial to improve the statistical performance of a regression model. For this reason we calculated the pair-wise correlation of each SNP to extract the colinearity between SNPs and exclude SNPs which have a pair-wise correlation higher than 0.90. We also make a list of correlated SNPs to determine the excluded SNPs associated with each selected SNP.

CHAPTER 4

APPROACHES USED IN GENETIC ALGRORITM BASED FEATURE SELECTION METHOD

Before introducing our genetic algorithm based feature selection method, the utilized statistical techniques are briefly discussed in this section to provide better understanding of the proposed method.

4.1. Logistic Regression and Related Studies

In the field of bioinformatics, epidemiologic data sets include large number of genes (SNPs) and small number of data samples. This issue makes it difficult to classify and construct a model for the gene or SNP selection. However, logistic regression is an effective approach to analyze significant genes or SNPs in medical studies. For instance, Foraita et al. (2008) apply logistic regression for comparison of graphical chain models. After constructing several logistic models, another important issue is how to select one of the models. Therefore, two information criteria are proposed to select the best statistical model among a group of models: Akaike information criterion (AIC) and Bayesian information criterion (BIC). For instance, Stumpfl et al. (2005) apply AIC for statistical analysis of biological networks. Xiaobo et al. (2005) propose a logistic regression method based on AIC and BIC to identify important genes for the cancer classification. Li et al. (2001) apply a two stage variable selection method to the German asthma data set to find the variables that best explains the data set. In the following section, we present a brief explanation of logistic regression, the motivation of using logistic regression and the explanation of AIC and BIC criteria.

4.2. Introduction to Logistic Regression

Like many forms of regression analysis, logistic regression uses several predictor variables, but it specifically aims to estimate the probability of occurrence of an event. Our aim of using a logistic regression method is to construct a biologically reasonable model to explain the association between a dependent variable (the probability of having the disease) and many independent variables (a group of SNPs). In this section, we briefly introduce the univariate logistic regression method but in our study we apply the multiple logistic regression method and the presented techniques can be generalized for the multivariate case.

4.2.1. Logistic Regression Method

The mean value of the dependent variable given the independent variable is called the conditional mean and represented as "$E(Y/x)$". (x=independent variable, Y=dependent variable). In linear regression this conditional mean is explained by a linear equation:

$$E(Y/x) = \beta_0 + \beta_1 x. \tag{4.1}$$

where β_0 and β_1 indicate the model coefficient. For binary response variables, the conditional mean must be between 0 and 1. [$0 \leq E(Y/x) \leq 1$] like the cumulative distribution of a random variable. Thus, for the analysis of binary dependent variables, many distribution functions have been used. In our study, we used the logistic distribution. Let us denote the $E(Y/x)$ by $\pi(x)$. By using the logistic distribution; $\pi(x)$ is defined as;

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x.}}{1+e^{\beta_0+\beta_1 x.}} \tag{4.2}$$

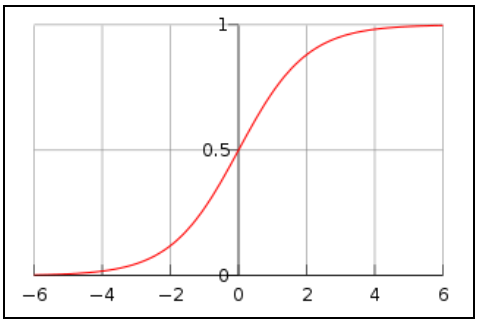

Figure 4.1. Logistic curve

As it can be seen from the figure of the logistic curve, input values for the logistic curve can take any value from -∞ to +∞. Since $\pi(x)$ can only take the values between 0 and 1, it must be converted to a real number in linear regression. This transformation is called "logit transformation". By transferring $\pi(x)$ to $g(x)$, we can obtain continuous values which can range from -∞ and +∞.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x. \tag{4.3}$$

The unknown model parameters (β_0, β_1) are estimated using the maximum likelihood estimation method. Thus, the maximum likelihood estimators, which maximize the likelihood function, are used to predict the probabilities of having the disease.

4.2.3. Testing the Significance of the Variables

The model parameters are estimated with and without the independent variables that are tested for the significance. These two sets of estimated parameters define two likelihood functions, which we refer to as LL_{fitted} and LL_{full}; LL_{fitted}: likelihood of the fitted model and LL_{full}: likelihood of the model including all parameters. The "likelihood ratio test" used the following statistic "D" to compare the difference between these two models:

$$D = -2\ln\left[\frac{LL_{fitted}}{LL_{full}}\right] \tag{4.4}$$

D is also called "deviance". Moreover, the distribution of D is known (approximately chi-square distribution) and therefore can be used for hypothesis testing.

4.3. Assessing the Fitness of the Model (Goodness of Fit Test)

By the goodness of fit test, we can test how effective a logistic model is. In our study, the statistical tests and pseudo r^2's are used for two purposes. The first purpose is to test the significance of a SNP-combination and the second purpose is to compare the significance of different SNP-combinations.

4.3.1. Classification Tables

A classification table displaying the results of correctly and misclassified instances is useful to understand how the model fits the data. We perform the following steps to find the classification error:

- Calculate the predicted response variables representing the probabilities of having a disease by applying the multiple logistic regression.
- Using the estimated function, calculate the predicted disease probability for each individual.
- Predict whether an individual has the disease or not based on the predicted probability. Set a cutoff value and if the predicted probability of an instance is bigger than that cutoff value, it is considered as case (has the disease) and takes the value of 1. If it is smaller than the cutoff value, it is considered as control (does not have the disease) and takes the value of 0.
- Compare actual disease status and predicted disease status and count the number of correctly classified instances.
- Divide the number of truly classified instances to the total number of instances to obtain the correct classification rate.

There are two measurements in a classification table: sensitivity and specificity. Let us denote the response variable as Y. Positive value of Y (Y=1) indicates cases and negative value of Y (Y=0) indicates controls.

$$Sensitivity = \frac{Correctly\ classified\ positives}{Total\ number\ of\ actual\ positives} \tag{4.5}$$

$$Specificity = \frac{Correctly\ classified\ negatives}{Total\ number\ of\ actual\ negatives} \tag{4.6}$$

In our study, our aim is to obtain the highest sensitivity with the constructed logistic model (SNP-combinations). We want to predict the disease status with minimum number of explanatory variables. However, just considering sensitivity can lead misleading results due to the fact that the constructed model (SNP-combinations) can also be explanatory for controls. Thus, we define a new measurement which we call "CAR (classification accuracy ratio)" to indicate the classification performance of the constructed model.

$$CAR = \frac{Correctly\ classified\ positive\ and\ negative\ instances}{Total\ number\ of\ instances} \tag{4.7}$$

4.3.2. Hosmer-Lemeshow Test

Hosmer and Lemeshow (2000) suggest dividing observations into groups according to their predicted probabilities to obtain a chi-square statistics. To use Hosmer-Lemeshow test we firstly list predicted probabilities in an ascending order. Then we divide these probabilities into 10 groups. The first group includes the observations which have the smallest predicted values and the last group includes the observations which have the highest predicted values. For each group, we compute a chi-square statistic by using the predicted and observed probabilities.

$$n_j = Number\ of\ observations\ in\ the\ j^{th}\ group$$

$$O_j = \sum_i y_{ij} = \ Observed\ number\ of\ cases\ in\ the\ j^{th}\ \text{group}$$

$$E_j = \sum_i \hat{p}_{ij} = Expected\ number\ of\ cases\ in\ the\ j^{th}\ group$$

$$G_{HL}^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j\left(1 - {}^{E_j}/_{n_j}\right)} \cong X_8^2 \qquad (4.8)$$

Then we construct a null hypothesis stating that there is no difference between the observed and predicted probabilities. If the p value of the statistic is smaller than 0.05, we reject the null hypothesis. Hence greater p value is desired not to reject the null hypothesis.

4.3.3. Likelihood Ratio Test (LRT)

LRT is another option to test the goodness of fit of the model obtained by the logistic regression. This test uses log likelihoods (LL) as a measurement. Since probability is smaller than 1, LL can take values between negative infinity and zero. Statistical packages like SPSS and STATA does not display LL. Since -2LL approximates a chi-square distribution, they provide -2*LL. We desire small values of -2LL for better prediction of response variable. Suppose a model h(x) with N predictors:

$$h(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x. + ...\beta_N x \qquad (4.9)$$

Then construct a null hypothesis (H_0) and compute the following measurements by using the equation 4.4.

Null Hypothesis: Ho: $\beta_0 = \beta_1 = ...\beta_N = 0$

$-2LL_{null}$ = model with only intercept

$-2LL_{model\ (N)}$ = model with intercept and N predictors

Model chi-square= $-2LL_{null}$-($-2LL_{model\ (N)}$) with N degrees of freedom

If the model p value is smaller than a pre-specified threshold value, we can reject the null hypothesis meaning that the model is statistically significant.

4.3.4. Scalar Measures of Fit: Pseudo R^2

Unlike linear regression, there is not only one coefficient of determination (R^2) defined for logistic regression. However, there are different pseudo R^2's which are constructed to measure the fitness of a logistic model. Although they are different, none of them is superior from each other. Besides, none of these pseudo R^2's represents the explained variance clearly. Hence they only provide partial information about the model.

4.3.4.1. Efron's Pseudo R^2

Efron (1978) suggested a pseudo-R^2 for binary response variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{\theta}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{\bar{y}})^2}$$

$\hat{\theta}_i$ = *model predicted probabilities*

$\bar{\bar{y}}$ = *average probability of all instances*

y_i = *actual probability of an individual* (y_i is a binary variable)

4.3.4.2. McFadden's Pseudo R^2

McFadden (1973) proposed a pseudo R^2 for models whose parameters are estimated by a maximum likelihood method. This pseudo R^2 also called "likelihood ratio index".

- Calculate the log likelihood LL_{full} of the model with all parameters in the regression model.
- Calculate the log likelihood LL_{null} of the model with only the intercept.

$$The\log likelihood\ ratio\ = 1 - \left({LL_{full}}/{LL_{null}} \right)$$

To avoid overfitting, McFadden's pseudo R^2 is adjusted by including a penalty parameter (K) which indicates the number of predictors in the model.

$$Adjusted\ pseudo\ R^2 = 1 - \left({(LL_{full} - K)}/{LL_{null}} \right)$$

4.3.4.3. Cox and Snell Pseudo R^2

Most statistical packages like SPSS provide Cox and Snell pseudo R^2 in logistic regression outputs. We also compute this measure. Let N be the total number of observations in the data set, then Cox and Snell pseudo R^2 is given by;

$$R^2 = 1 - \left\{ {LL_{null}}/{LL_{full}} \right\}^{2/N}$$

4.3.4.4. Nagelkerke Pseudo R^2

Since Cox and Snell pseudo R^2 can never take the value of 1, Nagelkerke modified it and suggested the following pseudo-R^2 by dividing the Cox and Snell pseudo R^2 by its maximum possible value.

$$R^2 = \frac{1 - \left\{ {LL_{null}}/{LL_{full}} \right\}^{2/N}}{1 - LL_{null}^{\ 2/N}}$$

4.3.4. Information Measures

To compare and select logistic models including different number of parameters, information measures like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) have been recently used in literature. Model selection criteria of AIC/BIC recently applied to epidemiology (Li et al., 2001); microarray data analysis (Nyholt et al., 2001) and DNA sequence analysis. The advantage of using such information measures is that we can use them for both nested and nonnested regression models. A nested regression model refers to two regression models which are identical except one variable. Nonnested models define any regression models that include more

than one different variable with the other model. Although the aim of AIC and BIC is the same (finding a good model), they differ in a theoretical sense. This difference can lead the selection of a different model among the same model set by each criterion. Despite their difference, there is not a clear explanation that one criterion is superior to the other. Selection of a good logistic model depends on the data set on hand. For different data sets, sometimes one criterion may find a better model than the other. Hence we consider both criteria.

4.3.5.1. Akaike Information Criterion (AIC)

The objective of AIC model selection is to find a model that best explains the data with the least number of independent variables. AIC is just a model selection tool rather than a hypothesis testing. Adding variables can fit the data perfectly and increases the likelihood but it can cause over fitting. To avoid this problem, AIC includes a penalty parameter which is an increasing function of the number of parameters in the model. Among a several competing models which are obtained from the same data set, the one with the lowest AIC value is the best. AIC is based on the theory of information gain "Kullback-Leibler information". Information gain is a measure of the difference between two probability distributions. More detailed information about the mathematical derivation of AIC and Kullback-Leibler information are given in (Burnham and Anderson, 2002). AIC is calculated by the following formula (Akaike, 1987).

$$AIC = 2k - 2\ln(L) \tag{4.10}$$

$$k = number\ of\ parameters\ in\ the\ model$$

$$\ln(\mathrm{L}) = \text{maximum log}\,likelihood\ function$$

4.3.5.2. Bayesian Information Criterion (BIC)

Schwarz (1978) proposes Bayesian information criterion for model selection. BIC is based on Bayes Rule and it is an approximation of the Bayes Factor. Similar to AIC, BIC includes a stronger penalty term to deal with the over fitting problems. Since the penalty term of BIC is stronger, it generally selects less complex models than AIC. Besides, BIC also includes sample size in the penalty term. BIC is computed by the following formula:

$$BIC = -2\ln(L) + K\log(n) \tag{4.11}$$

$$n = sample\ size$$

$$\ln(L) = maximum\ \log likeli\ hood$$

$$K = number\ of\ parameters\ in\ the\ model$$

The first term in the model indicates the deviance that measures the difference between the log-likelihood of the best fitting model and the log-likelihood of the model under consideration. As more parameters are added to the model, this term gets larger. The second term represents the penalty. For the models with too many parameters, the penalty term increases. For the models with too few parameters, the deviation increases. By combining these two terms, we balance over fitting and under fitting problems.

4.3.5.3. Comparison of AIC and BIC

The comparison of AIC versus BIC is very difficult since they are based on different theory. BIC assumes that the true generation model is in the set of candidate models and it assumes that there was a true model which is independent of the sample size in the model set, thus BIC tries to select this true model as the sample size goes to infinity with probability one. Unlike BIC, AIC does not assume that the true model is in the candidate models. It just selects the best model among a group of models. Most simulations that show BIC to perform better than AIC assume that the true model is in the candidate set and that it is relatively low dimensional. In contrast, most simulations that favor AIC over BIC assume that the true model is infinitely dimensional, and hence it isn't in the candidate set. Wagenmakers et al. (2004) state that AIC selects a specific model for the sample size at hand, but BIC does not.

CHAPTER 5

LITERATURE REVIEW: FEATURE SELECTION ALGORITMS

In bioinformatics field, the data often consists of large number of features and comparably very few number of samples. In such cases, the method of feature selection is very useful to improve the classification accuracy. The aim of feature selection is to select the most informative feature subset from the original data by providing reasonable prediction accuracy (Koller and Sahami, 1996). The main advantage of a feature selection method is reducing the problem dimension by not deteriorating the prediction performance. Silverman (1986) determines the required sample size for problems having different dimensions. As it is shown in Table 5.1, even for small dimensionality, the required number of sample is very huge. Thus, the search space of feature selection is very high and the problem is NP-hard. Moreover, collecting the genetic data requires high technology and budget, due to this problem achieving the required sample size is generally impossible. To deal with this problem, reducing the feature dimension is crucial to decrease the required amount of time and memory by the learning algorithms (Steinbach et al., 2006).

Table 5.1. Required Sample Size for Given Number of Dimensions

Dimensionality	Required Sample Size
1	4
2	19
5	786
7	10,700
10	842,000

Dash and Liu (1997) propose that in a typical feature selection method, there are four basic steps: a generation procedure, an evaluation function, a stopping criterion, a validation procedure.

- generation procedure is used for producing candidate subsets iteratively;
- an evaluation function investigates the feature subset under examination;
- a stopping criterion is used to decide when to stop; and
- a validation procedure is needed to test the validity of the feature subset.

The initial step of a feature selection algorithm, called generation procedure, is searching for a feature subset (Siedlecki et al., 1988; Langley, 1994). *The generation process can start with no feature, with all features or a random subset of features. In the first two cases, features are iteratively added or removed, whereas in the last case, features are either iteratively added or removed or produced randomly thereafter* (Langley, 1994; Dash and Liu, 1997).

The second step is measuring the goodness of a generated subset and comparing it with the goodness of the previous best subset by using the evaluation function. If the current subset is better, then it is replaced with the previous best subset.

To execute the feature selection algorithm in a reasonable time, there is a need for stopping criterion. Stopping criterion can be based either on the generation procedure or the evaluation function. If the selected feature number or the iteration number reaches to a predefined value or if deleting or adding features does not provide a better subset or the optimal subset is obtained, the algorithm stops.

The validation step is not part of a feature selection process but it is strongly recommended to be applied to test the prediction power of the selected subset using independent populations. Figure 5.1 represents the feature selection process with validation (Langley, 1994; Dash and Liu, 1997).

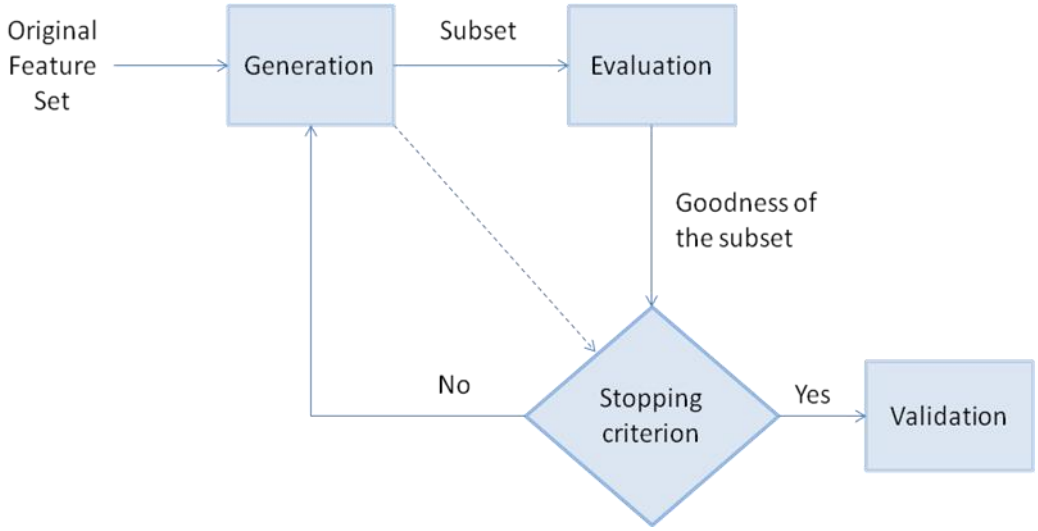

Figure 5.1. General Feature Selection Process with Validation

Feature selection methods can be applied to supervised (classification) or unsupervised (clustering) learning. For unsupervised learning, the feature selection method is applied to group features to find a good feature subset that provides a high cluster quality. In unsupervised learning, the feature selection aims to find a feature subset that provides higher classification accuracy (Kim and et al., 2003). Feature selection techniques are categorized into three groups (filter, wrapper and embedded) based on the integration of feature selection search to the classification model (Saeys et al., 2007).

5.1. Feature Selection Methods

5.1.1. Filter method

In the filter method, each feature is ranked according to some univariate metric. Features which have the highest rank are used for further analysis and the others are eliminated from the data (Ahmad et al., 2008). Filter approach considers all features and put them in a filter to output a subset of good features. Then this feature subset is used as an input for the classification algorithm. This method searches the feature subset independent of the classifier. Since feature selection is independent of the classification algorithm, the subset selection is performed only once and various classifiers are obtained (Saeys et al., 2007). Thus, it is faster than wrapper and embedded methods (Guyan and Elisseeff, 2003). Most filter approaches use univariate filter metrics like chi-square (Forman, 2003),

Euclidean distance and information gain (Ben-Bassat, 1982). These metrics investigate the power of each feature individually by ignoring the feature dependencies. Thus, filter methods cannot detect the features which are not individually informative but can be informative when it is combined with other features. In order to tackle this problem, multivariate search methods are developed: Markov blanket filter (Koller and Sahami, 1996), correlation based feature selection (Hall, 1999), Pearson correlation coefficient (Cho and Won, 2003) and fast correlation based feature selection (Yu and Liu, 2004).

5.1.2. Wrapper method

Wrapper method considers all features and generates some subsets of candidate features and passes them to the predictor. The predictor makes training and computes the prediction power of the feature subset. A new feature subset is generated until the optimum or near-optimal feature subset is obtained. There are two wrapper search methods; deterministic and randomized. Sequential forward selection (Kitler, 1978), sequential backward elimination (Kittler, 1978) and beam search (Siedelecky and Sklansky, 1988) are some examples of the deterministic search methods. Simulated annealing, genetic algorithm (Holland, 1975) and randomized hill climbing (Skalak, 1994) are randomized search techniques. In wrapper techniques, feature subset search is integrated with the classifier, in other words it considers feature dependencies. The main disadvantages of a wrapper approach are its risk of overfitting and intensive computational time (Saeys et al., 2007).

5.1.3. Embedded method

Embedded methods perform variable selection in the process of training and are usually specific to given learning machines (Elisseef and Guyon, 2003). Like wrapper techniques, embedded approaches are specific to a given learning algorithm. Decision trees, weighted naive bayes (Duda et al., 2001) and random forest (Guyon et al., 2002; Weston et al., 2003) are some examples of embedded feature selection techniques. Embedded methods are much faster than wrapper methods (Saeys et al., 2007).

5.2. Available Feature Selection Algorithms

Feature selection algorithms may be based on the statistical pattern recognition (SPR) classification techniques (supervised and unsupervised) or they can use artificial neural networks (ANN). An artificial neural network (ANN) is a nonlinear statistical data modeling tool for simulating biological neural networks. An artificial neural network consists of interconnected group of neurons. SPR techniques are categorized into two groups based on the optimality of the solutions. It can provide either optimal or suboptimal feature sets. Suboptimal solutions can be divided into two categories based on the number of feature subsets on a given solution. A suboptimal solution has either single solution obtained at the beginning of the algorithm and improves this solution iteratively or a population of different feature subsets each time the selection is applied. To generate a feature subset, deterministic or randomized feature selection techniques can be used. Deterministic models are the algorithms that give the same feature subset each time the feature selection is performed. Stochastic models are the ones that provide different feature subsets for each application of the algorithm.

Deterministic single solution methods firstly construct one feature subset and add or remove features iteratively until a stopping condition is satisfied. Deterministic single solution algorithms do not guarantee optimal solutions due to the fact that they do not search for all possible subsets. Beam and best first search are examples of multiple solution deterministic feature selection models.

The most widely used stochastic multiple solution feature selection method is genetic algorithm which is introduced by Siedlecki and Sklansky in 1989. Branch and bound method is an optimal solution search method that is proposed by Narenda and Fukunaga (1977). Optimal search algorithms are impractical for even small sample problems because the complexity of such algorithms is exponential in the worst case scenarios.

Feature selection methodology can also be based on node pruning. A node pruning algorithm firstly trains the data, removes least prominent node and iterates this procedure until reaching the specified node size or classification accuracy. Figure 5.2 displays the categorized feature selection methods (Jain and Zongker, 1997).

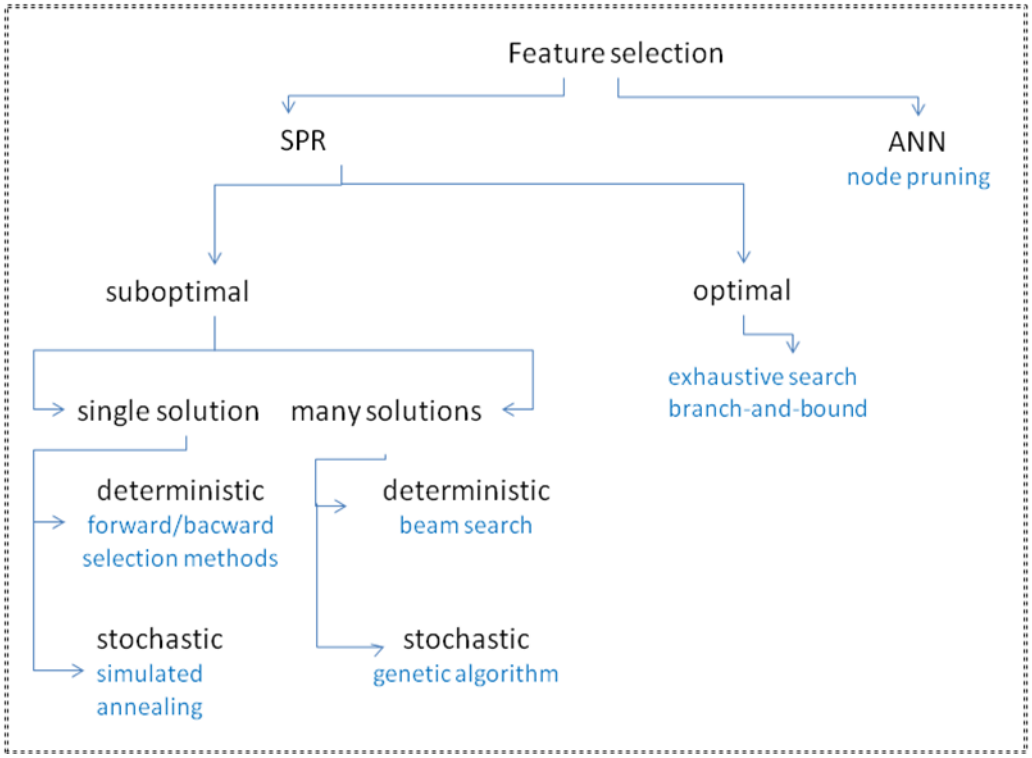

Figure 5.2. Taxonomy of feature selection algorithms

In bioinformatics, feature selection algorithms are applied for sequence, microarray, mass spectra and single nucleotide polymorphism analysis (Daly et al., 2001; Gabriel et al., 2002; Carlson et al., 2004). Since exhaustive search techniques are not practical to implement, researchers often prefer evolutionary algorithms (simulated annealing, genetic algorithm) to solve optimization and machine learning problems (Segal and Zhang, 2006).

Shah and Kusiak (2004) develop a genetic algorithm based feature selection method to identify gene/SNP patterns. They use a global search mechanism, weighted decision tree, decision-tree based wrapper, a correlation-based heuristic to select the most significant genes. Wu and et al. (2008) propose a heuristic based on genetic algorithm to assemble single individual SNP haplotypes. Chang et al. (2008) develop an odd ratio based genetic algorithm procedure to produce SNP barcodes of genotypes to measure the disease risk among many SNP combinations. Nakamichi et al. (2004) propose a combination of logistic regression and genetic algorithm model to investigate the association between a combination of SNPs and a disease. Gong et al. (2005) develop a data reduction technique based on a genetic algorithm and support vector machines to identify the key SNP features. Ooi and Tan (2003) apply a genetic

algorithm based gene selection method for a multi-class prediction problem. Ooi and Tan (2003) also propose that genetic algorithm based techniques may be powerful tools to analyze complex multi-class gene expression data. Liu et al. (2005) combine genetic algorithm and support vector machine methods for multi-class cancer classification. Handels et al. (1999) apply various feature selection algorithms to optimize skin tumor recognition by using greedy algorithms. According to Handels et al. (1999) among all available feature selection techniques, genetic algorithm gives the best results in terms of the classification rate. In this study, we also focus on genetic algorithm based feature selection methods.

CHAPTER 6

PROPOSED METHOD: GENETIC ALGORITHM BASED FEATURE SELECTION METHOD

The goal of this research is developing a feature (SNP) selection method to identify significant SNP combinations related to a complex disease. In literature different feature selection approaches exist such as the principal component analysis, genetic algorithm (GA) and decision tree. Among these techniques, the GA is an efficient and effective method to analyze millions of SNPs. In this study, we develop a genetic algorithm based feature selection method to maximize the explanatory power of the selected SNPs while trying to decrease the number of SNPs dynamically. The proposed GA is applied to the simulated rheumatoid arthritis data set provided by Genetic Analysis Workshop 15 and at each execution the algorithm constructs a set (combination) of SNPs with the minimum cardinality and the highest explanatory power in a logistic regression model. Before introducing the proposed GA, we provide a short summary of the genetic algorithm approach.

6.1. Steps of the Genetic Algorithm

- Generate an initial population (by random selection of individual solutions).
- Calculate the fitness function of each solution in the population.
- Apply reproduction, crossover and mutation operators.
- Determine the fitness value (score) of each newly generated solution.
- Remove solutions, which have unsatisfactory fitness value from the population.
- Repeat this process until a termination condition has been satisfied.

In the first step of the algorithm, an initial solution set, which is called population, is constructed (generation). A population includes valid alternative candidate solutions,

which are called individuals. The initial population is used to produce a new generation, which is called offspring. In the second step, fitness function values are calculated to determine the quality of the solutions in the population. The new population is expected to be better than the old one in terms of fitness values. The third step (reproduction) includes two main processes, which are crossover and mutation. In the crossover process, two chromosomes are paired and two new chromosomes (solutions) are obtained. After generating new solutions, mutations of chromosomes occur according to the mutation probability (rate). If the mutation rate is met for a chromosome, the associated solution is obtained by changing one/more genes in that chromosome. In the fourth step, the fitness value of each newly generated solution in the population is calculated. In the fifth step the unsatisfactory solutions are removed from the population to make the population better in terms of the fitness values. As the search iterates through multiple generations, fitter solutions increase in the population, and less fit solutions decrease in the population. As a result the final population would be the best of all populations considered through the algorithm.

6.2. Introduction to Proposed Genetic Algorithm based Feature Selection Method

We determine the population size by the formula that is proposed by Küçükural (2009). Let X denote the population size (the number of parents in one generation), Y denote the size of the tag-SNPs set, P denote the desired number of occurrence of a SNP in one population (or called feature coverage) and K denote the number of SNPs used to represent a parent. Let W denote the number of individuals who have the worst fitness scores in the population. X and W are calculated by the following formula.

$$X = (Y * P)/K \tag{6.1}$$

$$W = 0.20 * X \tag{6.2}$$

In the proposed algorithm, three different methods to generate populations are described in detail. The *first method* is used to generate the initial population and the *second method* is implemented M1 times to generate ``better'' populations (in terms of significance of SNPs) iteratively. At each implementation of the second method, the current population is used and improved to generate the next population. For example, the initial population is used to generate the second one, and the M^{th} population is used to generate the $(M1+1)^{th}$ population. Then the $(M1+2)^{th}$ population is obtained by

applying the *third proposed generation* method. Finally, additional populations are generated by reapplying the *second method M2 times to further improve the generated populations*. After a total of M1+M2+2 generation of populations, a final ``population'' is obtained and the rest of study focuses only this population.

The first population generation method constructs each individual by *random selection of SNPs.* The second population generation method constitutes each individual by using the survival probabilities of *individuals* in the previous population. The third population generation method uses survival probabilities of *SNPs* in the previous population to form a parent.

The first method is applied at the beginning of the algorithm and all tag SNPs have the same importance score to be involved in a SNP combination (individual). Thus, in order to generate an individual, we apply random selection of tag SNPs. In the second method, we obtain next generation by applying crossover between two individuals obtained from the previous population. Since the individuals who have a higher fitness score can generate a better individual, we use survival probabilities of individuals in the previous population. The individuals having a higher survival probability refer to the individuals which have a higher chance to be a patient. Applying this procedure to several times, we can acquire information about significance of SNPs by counting the frequency of each SNP in the population. The more frequently observed SNPs have a higher survival probability meaning that they may be the disease causing SNPs. If a SNP has a higher survival probability, it has a higher chance to be transferred to the next generation. Therefore, we firstly apply second method M1 times to obtain highly observed SNPs. Then we apply the third method which uses SNP survival probabilities for once to generate an individual by not applying crossover. Thus, in that way we integrate this information about SNPs to the algorithm. Then we continue to apply second method to reach the best population.

6.3. Outline of the Proposed Algorithm

As it is mentioned before to compare SNP combination models we use AIC and BIC measures. Since these two measures do not give the same solution, we coded developed algorithm two times. The first one only uses AIC measure as a fitness score and second one uses only BIC. In the next section, the outline of our feature selection method will be given in terms of AIC measure. The genetic algorithm based feature

selection method which uses BIC measure is also the same. Thus we represent fitness score as AIC/BIC to mention the usage of two measures. The flowchart of the proposed algorithm is given in Figure 6.1.

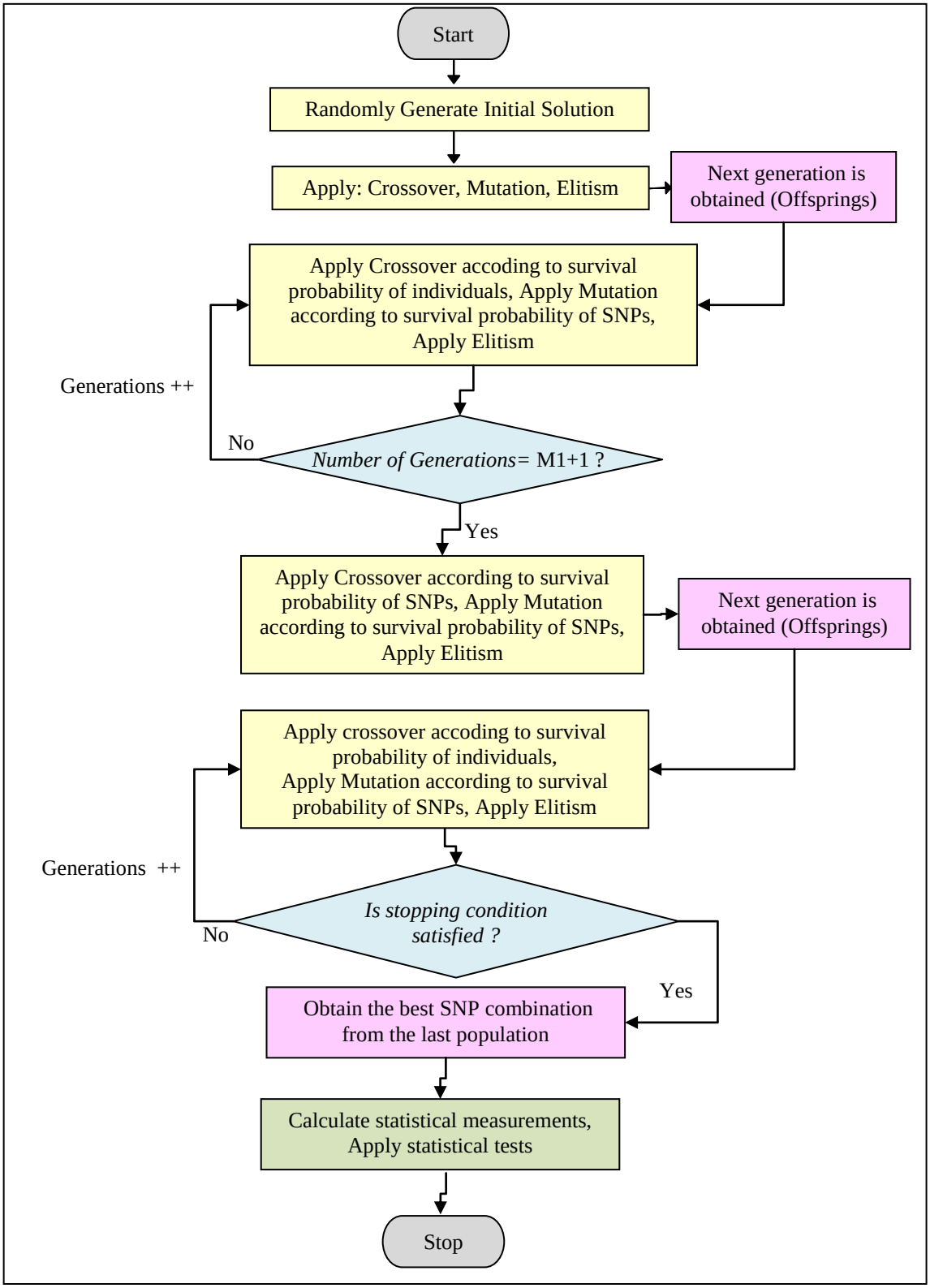

Figure 6.1. Flow chart of the genetic algorithm based feature selection method

I. CONSTRUCT THE FINAL PROPOSED POPULATION

A. GENERATE INITIAL (PARENTS') POPULATION (POPULATION):

first generation method

1. Generate the initial population: Each population includes X individuals which are called parents. Repeat the procedure of creating a parent, which is explained below, X times to generate the initial population.
    a. Select a SNP randomly from the tag-SNPs set. (Generate a random number between 1 and Y.)
    b. Select (K/2-1) SNPs from the left side and 15 SNPs from the right side of the SNP selected in Step (a) to obtain a K-feature-sized-parent. If there are not enough SNPs in one of the sides, select all SNPs in that side and select remaining ones from the other side. Thus, (K-1) SNPs are selected from the neighborhood of the chosen tag-SNP and a parent with K features is obtained. K should be an even number.
    c. After generating a parent, calculate its fitness score. To do it, apply the method of multiple logistic regression to the original multiple family data. In this application of the multiple logistic regression, the K features associated to that parent are considered as explanatory variables and the response variable takes value one if the case has the disease and zero otherwise. Then, calculate fitness scores: the values of the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)
    d. Execute Steps (a)-(b) and (c) X times to obtain the initial population.
2. Calculate the estimated survival probability of each parent in the initial population:
    a. Calculate the total fitness score of the initial population. (Sum the fitness value of each parent).
    b. Divide the fitness value of each parent by the population fitness value.
    c. Calculate the survival probability of each parent by subtracting the fitness score from 1. The higher the survival probability, denoted by

θ_i for the i[th] parent (i=1,…,X), the higher the significance of a parent (the combination of SNPs), since a low AIC/BIC value (fitness score) indicates a high significance in the logistic regression.

B. OBTAIN THE NEXT (OFFSPRINGS') POPULATION: ***second generation method***

1. Apply Crossover:
    a. Select two parents from the initial population based on the associated survival probabilities. We perform a random selection as described below:

    Construct intervals corresponding to each parent in a way that the length of the interval for the i[th] parent is equal to the survival probability θ_i: $[\{(0,\theta_1),(\theta_1,\theta_1+\theta_2),...,(\sum_{i=1}^{X-1}\theta_i,\sum_{i=1}^{X}\theta_i)\}$. Then generate a random number from the interval $(0,\sum_{i=1}^{X}\theta_i)$ and if the value belongs to the interval $(\sum_{i=1}^{i-1}\theta_i,\sum_{i=1}^{i}\theta_i)$, we select the i[th] parent.

    b. Generate two random numbers between 1 and Y to determine the starting and ending points of the crossover region and apply the two-point crossover to construct two new children. For each child, if there exist multiple copies of a SNP, just keep one and delete the other copies. Thus, the generated children do not have repetitive SNPs.
2. Apply the multiple logistic regression for each new child and calculate their AIC and BIC values.
3. Execute Steps 1-3 X/2 times iteratively to obtain the next population with X parents.
4. Apply mutation operator for each parent in the population that we are currently generating.
    a. Assign a mutation rate by generating a random number between 0 and 1. If the mutation rate is smaller than 0.05, then the mutation is applied. Otherwise mutation is not applied.
    b. If the mutation is decided to be applied, select a SNP to be mutated by generating a random number between 1 and K. Thus, at most one mutation is allowed to occur for each parent.

c. For each SNP in the tag-SNP set (with Y SNPs), calculate the associated survival probability (not for parents, for individual SNPs).
d. Use the method described in Step B.1 for selecting a SNP to be used as the new value of the mutated SNP. Here different than Step B.1 we consider the survival probability of each SNP in the tag-SNP set.
e. After the mutation, make sure that the new value, the SNP selected in Step 4.e, has at most one copy in the parent.

5. Apply multiple logistic regression to each parent in the current population and calculate their fitness scores.
6. Elitism of Best Childs
   a. Find the W worst parents (children of the previous population) in the current population according to their fitness scores.
   b. Find the W best parents in the previous population according to their fitness values.
   c. Replace the worst parents in the current population with the best parents in the previous population.
   d. Calculate fitness score of the new population obtained after the replacement.
   e. Calculate the survival probability of each parent in the population.

C. REPEAT STEPS B.1-B.6 M1-1 TIMES TO CONSTRUCT THE $(M1+1)^{th}$ POPULATION (use the second generation method)

At this stage of the algorithm, the $(M1+1)^{th}$ population is the best one. However, to further improve it, we employ the third generation method.

D. CREATE A NEW PARENT POPULATION BY USING SNP SURVIVAL PROBABILITIES: *third generation method*

1. Calculate the survival probabilities of individual SNPs based on the number of occurrence in the $(M+1)^{th}$ population.
2. Use the method described in Step B.1 for selecting K SNPs to construct a potentially significant combination of SNPs (a parent). Here different than Step B.1 we consider the survival probability of each SNP based on the number of occurrences.
3. Repeat Step D.2 X times to form a population including X parents.

4. Apply the multiple logistic regression to the generated population and calculate the fitness score of each parent.
5. Calculate the survival probability of each parent in the population.
6. Implement Steps B.1-B.6 to obtain the $(M1+2)^{th}$ generation.

E. GENERATE ADDITIONAL M2 POPULATIONS iteratively applying the second generation method M2 times. (UB on M2=500)

We set an upper bound on the value of M2. If the algorithm can keep improving the available populations, this upper bound is attained. Otherwise, the algorithm stops if there is no improvement in the best fitness score of the populations generated in consecutive 50 iterations.

II. CALCULATING SOME STATISTICAL MEASURES FOR THE FINAL POPULATION

We apply hypothesis testing and calculate pseudo r-squares to determine the statistical significance of a combination.

A. Classification Table (Prediction Performance)

As mentioned before, the final population includes the best SNP combinations in terms of goodness of fitness. We picked the SNP combination in the last population which best fits the data and prepare a classification table for that SNP combination. We tried different cutoff values (from 0.3 to 0.7) to predict whether an individual has the disease or not based on predicted probabilities. According to numerical results, the value 0.5 performs well in terms of correct prediction percentage. Therefore, if it is not stated otherwise we take the cutoff value as 0.5 in our computational study.

B. Apply McFadden's Pseudo R-Square

C. Apply Adjusted McFadden's Pseudo R-Square

D. Apply Cox and Snell Pseudo R-Square

E. Apply Nagelkerke Pseudo R-Square

F. Apply Efron's Pseudo R-Square

G. Apply Likelihood Ratio Test

H. Apply Hosmer and Lemeshow Chi-Square Test (Goodness of Fit Test)

III. REITERATE ALL STEPS FROM I TO III

To test the validity of the proposed approach, we run the whole algorithm 5 times.

IV. FIND OCCURANCE NUMBER OF EACH DISTINCTIVE SNP by CONSIDERING ALL RUNS

V. EXTRACT SNPs WHOSE OCCURANCE RATIO IS BIGGER THAN T (0.85)

Find "intersection SNP set".

At the end of the whole algorithm, we obtain potentially significant SNP combinations which are supposed to lead to the disease.

After obtaining five significant SNP combinations and one intersection SNP set from seven runs of the algorithm, we construct a new SNP combination from the intersection SNP set. Therefore, we have seven alternative SNP combinations for a population.

Since we calculate the statistical measures (pseudo-r^2s) for the seven potentially significant SNP combinations, we then use these measures in order to decide whether the SNP combination is really significant or not. If a SNP combination satisfies the threshold values of all pseudo-r^2 measures, then this SNP combination is considered as significant and otherwise it is eliminated from the study. Satisfying all pseudo-r^2 thresholds allow us to choose a SNP combination independent of the measure. However, a decision maker can specify a set of r^2s that have to be satisfied in his study to determine the significant SNP combinations.

In the next step, since a SNP combination includes more than six SNPs, we apply DTREG decision tree forest algorithm for each SNP combination to reduce the number of SNPs into six in a SNP combination because, our rule extraction method considers at most six SNPs.

DTREG decision tree forest algorithm assigns an importance value (over 100) to each SNP in the SNP combination set. Then we select the top six SNPs according to their importance value. We then apply decision rule extraction method for each possible SNP combination to determine the classification performance of each SNP combination.

Next, we pick the SNP combination which has the highest classification performance for this population and denote it as the “best SNP combination of the population”. (In this procedure we apply decision rule extraction method for six times for both AIC based solutions and BIC based solutions, in total 12 times).

We apply the same steps (preprocessing and genetic algorithm based feature selection method) for seven different populations and obtain the best SNP combination for each population. Thus, we have seven best SNP combinations in total. These seven SNP combinations provide very similar classification accuracy and each of them is significant.

Although the populations are different, the best SNP combinations consist of some common SNPs. In order to find the mostly observed SNPs in best SNP combinations, we determine the occurrence of each SNP in seven alternative SNP combinations. We then select the top six SNPs which are mostly observed and construct a new SNP combination from these SNPs. We then compared the classification accuracy of that SNP combination with the other seven best SNP combinations. According to experimental results, there is not a considerable difference in terms of classification accuracy of the newly generated SNP combination and the SNP combinations obtained from genetic algorithm. Thus we will apply decision rule extraction method to any of that SNP combination to find the related SNP-genotype relations. We selected the newly generated SNP combination to extract genotype-SNP relations.

CHAPTER 7

APPLICATION OF DECISION TREE FOREST ALGORITHM TO OBTAIN THE BEST SET OF SIGNIFICANT SNP COMBINATIONS

Decision tree learning is a widely used decision support tool in data mining field. It uses a decision tree as a predictive model to classify an instance. A decision tree method produces IF-THEN expressions to classify an instance. These If-THEN structures are very helpful to get intuitive interpretation of biological questions. Although decision tree is an effective tool for rule extraction, it cannot provide reasonable classification accuracy for a noisy data like the one in our study. However, for many years techniques to combine the results of multiple classification models have been investigated to make a single prediction from many decision trees which are called decision tree forest (Tong et al., 2003). Decision tree forest is a technique that combines similar single decision trees to provide higher classification accuracy compared to the single decision tree models. Tong et al. (2003) suggest a decision forest algorithm to classify 232 chemicals into two categories (estrogen and non-estrogen receptor-binding). They compared the model performance between a decision tree and a decision tree forest. They conclude that decision forest provides a higher classification accuracy for both testing and validation samples. According to Tong et al. (2004) combining several identical decision tree models produces no gain thus provides a more accurate prediction ratio. Xie at al. (2005) examine the association between esophageal cancer risk and 61 SNPs in a case/control study by developing a decision tree forest method. Like decision trees, decision tree forest algorithms also uses IF-THEN rules for classification of observations but unfortunately, it does not list the produced rules as an output. However, it assigns an importance value for each variable in the data set so that we can be aware of the most significant variables in the data set. Since our aim is to reduce the size of a SNP combination before applying our rule extraction method, we use a

decision tree forest algorithm to select the most significant six SNPs from a SNP combination. We use DTREG decision tree forest algorithm which is developed by Phil Sherrod who integrates the random forest algorithm of Breiman (1999) into DTREG. Decision tree forest models are so far among the most accurate models invented (Sherrod P., 2009). One advantage of decision tree forest is that without using a separate data set validation can be done by using out of bag data rows. However, the main disadvantage of a decision tree forest is that the model is too complex and it includes many decision trees. Thus, the decision tree cannot be visualized. The outline of the algorithm to construct a decision tree forest is given in below:

- Assume that the data set includes N observations and m variables.
- The first step is the selection of N observations from the data set with replacement (bagging). Approximately 2/3 of the rows are selected as a test sample. The remaining 1/3 of the rows are called "out of bag" rows and these rows are used as a validation sample. For each time that a new tree is created, this random selection is repeated.
- The second step is constructing a decision tree by the use of selected rows in step1. To split a node in a tree, only a group of variables (k) is chosen randomly among m variables ($k < m$). For each time that a node is splitted, we randomly select a new variable set from m variables.
- By repeating steps 1 and 2, we obtain a large decision tree forest.

After constructing the decision forest, we run the rows through each tree in the forest and record the predicted value. Then we use the predicted categories for each tree as votes and assign the category with the most votes as the predicted category for the row (Sherrod P., 2009). DTREG assigns each variable (SNP) an importance ratio by applying the following steps: For each tree in the forest, DTREG puts down the out of bag observations and counts the number of truly classified instances. Then it randomly permutes the values of variable m in the out of bag observations and runs them through the decision tree. After that it counts the number of truly classified instances for the permuted data. Next it subtracts the vote of out of bag data from the vote of the permuted one. It iterates this procedure for each tree sums the differences and then takes the average of differences. This number is the raw importance score for variable m. After calculating the importance ratio for each variable, we then select the top six variables to reduce the size of the SNP combination.

CHAPTER 8

PROPOSED DECISION RULE EXTRACTION METHOD

After obtaining significant SNP combinations from the genetic algorithm, we developed a rule extraction method to analyze significant genotypes related to the disease. A rule refers to a SNP combination with genotype information. Although determining significant SNP combinations is a very significant issue, it is not enough to understand the structure of a complex disease. Different genotypes of a SNP may result the disease status in a different way. Thus, the affect of genotypes should also be assessed. Different softwares are publicly available to extract decision rules according to genotype information of SNPs like DTREG, Weka and RapidMiner. These tools use decision tree algorithms or special rule mining methods. Since most genetic data are very dense, decision tree algorithms do not provide an adequate classification ratio for such data. They extract the rules which are mostly observed in the population. However, in real life, some patients may have a common genotype related to the disease but the ratio of these people in the population may be very rare. Most of the existing softwares do not detect such relations. Our developed rule extraction method can detect both rarely and mostly observed relations in the population. Hence, it provides higher classification accuracy than other well known methods.

8.1. Outline of the Proposed Decision Rule Extraction Method

General outline of our developed method is given in Figure 8.1. We take the significant SNP combinations identified using the GA as an input and provide significant SNP relations (rules) which are associated with the disease, as outputs. As it can be understood from the figure, our rule extraction method has three main stages:

- Association rule mining,

- Selection of significant rules,
- Determination of minimum number of significant rules.

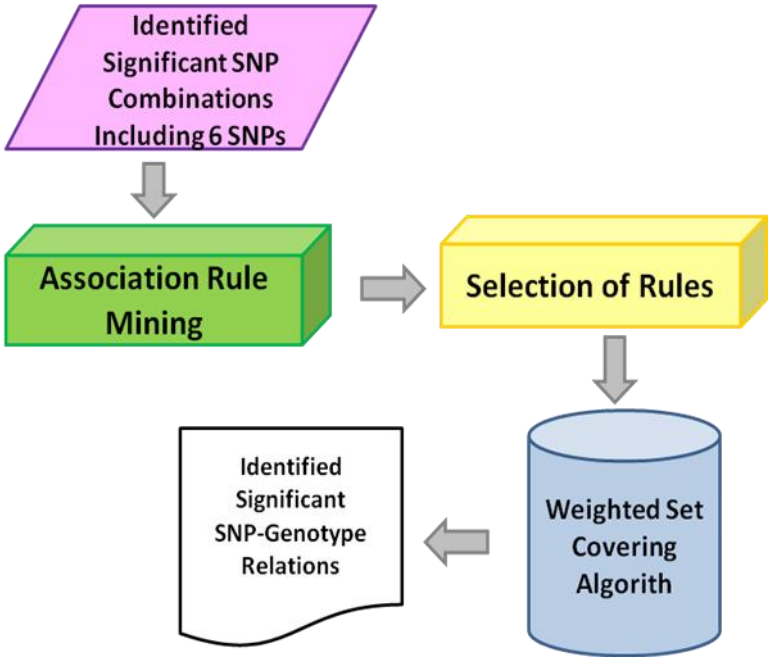

Figure 8.1. Representation of the Proposed Decision Rule Extraction

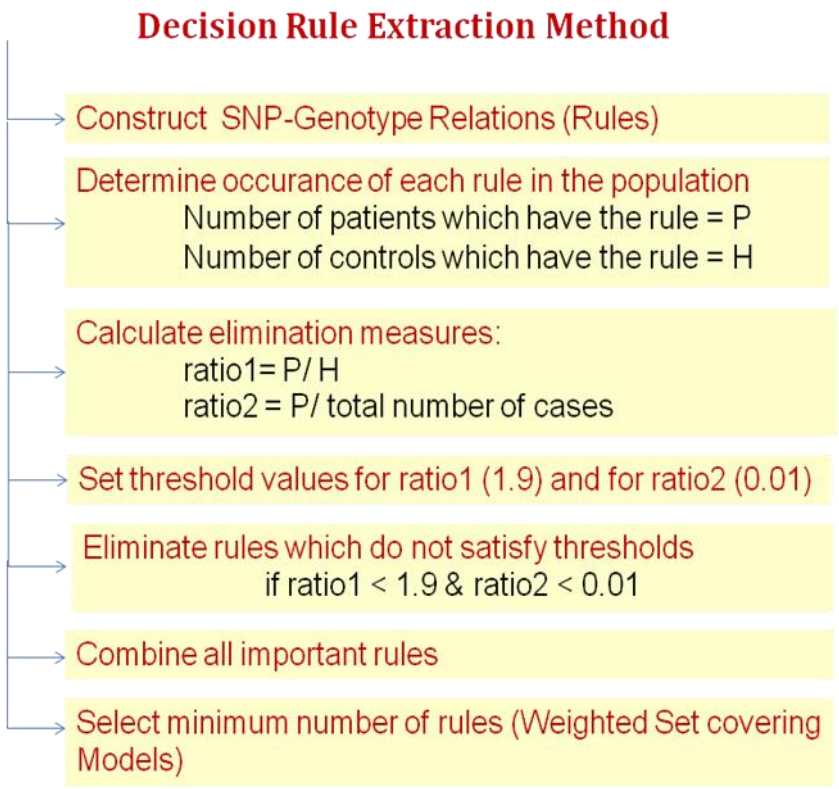

Figure 8.2. Detailed Outline of the Proposed Rule Extraction Method

By using DTREG decision tree forest algorithm, we obtain the most significant six SNPs for each alternative SNP combinations. However, all the patients may not have all these six SNPs. Thus, in the rule extraction stage we construct new SNP combinations with the additional genotype information by using these six SNPs. We allow a combination include at least one and at most six SNPs.

8.2. Steps of the Proposed Decision Rule Extraction Method

8.2.1. Association Rule Mining

In our study, all SNPs are in bi-allelic form and bi-allelic SNPs have three possible genotypes (AA, Aa, aa). We firstly construct all possible SNP-genotype combinations including at most six SNPs. Thus, we allow a combination with at most six significant SNPs and at least one significant SNP. Combinations of larger number of significant SNPs can also be considered. But in our experiments we observed that considering at most six SNPs is good enough in terms of the classification accuracy. Then for each constructed rule, we firstly identify the patients, who have these rules and compute the number of such patients in each group (case and control). Totally a large number of different rules can be obtained. For example, in our study we obtain 4094 rules in total with at least one and at most six SNP combinations. But some of these rules may not be observed or may be rarely observed in the population. Thus, a rule selection procedure is required to extract significant rules among the rule set. For this reason we developed a selection procedure.

8.2.2. Selection of Significant Decision Rules

To select significant rules among the whole rule set, we determine two selection criteria. The first criterion is called “Ratio1” and the second is called “Ratio2”. How to calculate these measures is given in 8.1 and 8.2.

P = the number of cases which have the rule

H= the number of controls which have the rule

R= the number of cases which have the rule

T = total number of cases

$$\text{Ratio1} = P / H \quad (8.1)$$

$$\text{Ratio2} = R / T \quad (8.2)$$

We set thresholds values for Ratio1 and Ratio2 which are 1.9 and 0.03 respectively to eliminate non-significant rules. If a rule satisfies both of the criteria threshold values, this rule is selected as a significant rule. Because our aim is to extract the rules which are mostly observed in patients, we desire a P value which is 1.9 times bigger than H. Besides, a rule can be observed in cases more than controls but it may have a very low frequency. Hence, we define the second criterion (Ratio2) to select a rule which is observed at least 3 percentages of the patients. We especially set a lower threshold value for Ratio2 not to exclude the rarely observed significant rules. By eliminating rules which do not satisfy these thresholds, we obtain a set of significant rules.

8.2.3. Determining Minimum Number of Significant Rules

Among significant rule set, some rules may explain the same patients. Our objective is to extract the smallest number of rules to explain the status of all the patients. This rule selection problem can be modeled as a general weighted set covering problem to find an optimal set of rules.

8.2.3.1. General Weighted Set Covering Model

For each rule, we define a set (rule set) including the patients which have that rule. Then considering these rule sets, we want to determine the minimum number of rule sets for which the union covers (contains) all the patients. We consider alternative methods considering different criteria to select the rules. These criteria are incorporated into the models by defining weights associated with the rule sets. Then we solve the corresponding weighted set covering problem, which can be formulated as a binary linear programming problem. The general weighted set covering model is given in below:

Min objective $\min z = \sum_{i=1}^{K} (1/W_i) R_i$ (8.3)

Subject to

$$\sum_{i=1}^{K} P_j * R_i \geq 1 \; for \; j = 1, 2, \ldots J$$

R_i= 0 or 1

Parameters:

$U = set\ of\ all\ different\ patients\ covered\ by\ the\ set\ of\ all\ rules\ (K)$

$J = |U|\ , number\ of\ patients\ covered\ by\ U\ set$

$K = total\ number\ of\ rules$

$P_j = [P_{1j}\quad P_{2j}\ \dots\ P_{Kj}]\ is\ a\ 1xK\ vector$

$where\ P_{ji} = 1\ if\ rule\ i\ covers\ patient\ j\ and\ zero\ otherwise\ 0$

Decision variables:

$$R_i = \begin{Bmatrix} 1\ , & if\ rule\ i\ is\ chosen \\ 0\ , & otherwise \end{Bmatrix}\ for\ i = 1,2\ \dots K$$

We solve the general weighted set covering model for three different types of weights. The models according to these three different criteria are explained in below.

First criterion: Giving equal importance to each rule

The first objective is selecting the optimal number of rules by giving each rule the same importance value. To do this, we assign the value of 1 as the weight of a rule and so the objective function coefficients in 8.3 are

$$W_i = 1\ for\ every\ i = 1, 2, \dots . K \tag{8.4}$$

Second criterion: Maximum cardinality

We define the weight of each significant rule based on the cardinality, i.e., the number of patients covered by the rule set. A weight of a rule is calculated as:

$$C_i = cardinality\ of\ rule\ i\ (number\ of\ patients\ covered) \tag{8.5.}$$

$$W_i = C_i$$

Selecting rules based on maximum cardinality can allow a researcher to analyze the rules which are observed in the majority of patients. This information can be useful for developing cures and drugs that most of the patients can respond.

Third criterion: Maximum ratio1

We also define the weight of a rule based on maximum ratio1. Ratio1 is the proportion of the rules which are observed in cases divided by the rules which are observed in controls. A weight of a rule based on maximum ratio1 criterion is calculated as:

$$T_i = ratio1\ of\ rule\ i\ \left(ratio1 = \frac{cases\ having\ the\ rule}{controls\ having\ the\ rule}\right) \qquad (8.6)$$
$$W_i = T_i$$

Giving priority for the rules having the maximum ratio1 value in rule selection can allow a researcher to determine the most different SNP-genotype combinations between cases and controls. This information can be useful for the diagnosis of the disease.

The solution of the weighted set covering problem based on different criteria might provide different sets of rules due to the different objective functions with the general set covering algorithm.

8.3. Extracting Significant Genotype of Each Significant SNP in the Significant SNP Combination

After determining optimal significant rule sets based on different criteria, the next step is to extract the significant genotypes of each SNP in the significant SNP combination. Since not all genotypes of a SNP can lead to the disease status, we should analyze the genotypes that are observed in optimal rule sets. Thus, we investigate the genotype of each SNP in each rule in the optimal rule sets so we can determine the relationship between the disease and the genotype of a SNP (being homozygote or heterozygote).

CHAPTER 9

EXPERIMENTAL RESULTS

We repeated our study seven times by using different population samples to test the reliability of the proposed methods. As it is mentioned before, our analysis was based on the simulated rheumatoid arthritis data provided by Genetic Analysis Workshop 15. Since we know the most likely disease causing genome interval (chromosome 6) from the previous studies (Uh et al., 2007; Zhang et al., 2007), we apply our methods to 17820 SNPs on chromosome 6. Each replication data set consists of 8000 individuals, but they include different number of cases and controls. For each replication, first we apply preprocessing steps (genome wide association analysis and tag SNP selection) to 17820 SNPs. The number of SNPs remained after the end of preprocessing steps is listed in below for each replication.

Table 9.1. Number of potentially significant SNPs remained after preprocessing

	Data Set Name	Number of Significant SNPs determined by GWA	Number of Selected Tag-SNPs determined by Haploview
Replicate 1	rep0001	165	148
Replicate 2	rep0002	171	154
Replicate 3	rep0004	160	142
Replicate 4	rep0005	189	111
Replicate 5	rep00052	145	89
Replicate 6	rep00053	188	130
Replicate 7	rep00054	148	108

Although the initial data sets include large number of SNPs (17820 SNPs), the number of potentially significant SNPs after preprocessing is not very huge. Thus, eliminating insignificant SNPs from the analysis is a very crucial step to save time. After applying genome wide association tests, we extract the potentially significant SNPs whose p value are smaller than Bonferroni adjusted significance level (0.05/number of SNPs

included in the association analysis of GWA after removing the poor quality SNPs).Then we apply a tag SNP selection algorithm to choose optimal SNP set.

We use tag SNPs as an input for genetic algorithm based feature selection algorithm code. For each replication, we run MATLAB code five times and obtain five significant SNP combinations. Then we find the intersection of SNPs that are observed in these five significant SNP combinations and construct a new SNP combination by using these SNPs. Thus, we have six alternative significant SNP combinations for each replicate. We have 84 (42 of them are obtained from BIC based GA, the others are obtained from AIC based GA) alternative significant SNP combinations in total.

To test whether a SNP combination is reasonable in terms of goodness of fitness, we calculate statistical measures which are mentioned in Chapter 4. The results of statistical measures are given in the Appendix A. All alternative significant SNP combinations have very similar statistical measurement values and all pseudo-r^2 values are bigger than 0.3 indicating the goodness of fitness of the alternative models (SNP combinations). The size of each significant SNP combination obtained from GA and average SNP size are given in Table 9.2.

Table 9.2. Size of each SNP significant SNP combination obtained from GA

SIZE	AIC	BIC
Population1	21	18
Population2	21	19
Population3	23	21
Population4	22	19
Population5	22	22
Population6	20	20
Population7	20	20
Average	21	20

Because alternative SNP combinations include more than six SNPs, we apply a decision tree forest algorithm to find the most significant six SNPs. (We construct 200 decision trees for each decision tree forest). Then we apply our decision rule extraction algorithm to that six-SNP combination. Each alternative solution in a replicate approximately provide the same sensitivity value which is always approximately 0.85. The intersection SNP set always provides a better sensitivity value (approximately 0.90-0.94) for each replicate. As an example the sensitivity values of alternative combinations for replicate 5 are given in Table 9.3. Each solution refers to a significant

SNP combination-including six SNPs-obtained from genetic algorithm based feature selection method.

Table 9.3. Sensitivity value of each solution of a genetic algorithm based feature selection method

Replicate_0005	Sensitivity – AIC based GA (%)	Sensitivity – BIC based GA (%)
SOLUTION1	84.78	87.51
SOLUTION2	84.38	84.58
SOLUTION3	84.72	84.50
SOLUTION4	85.38	84.78
SOLUTION5	84.78	89.73
INTERSECTION	91.2	91.2

In the second and third column of Table 9.3 the sensitivity values of solutions obtained from genetic algorithm based feature selection method which considers AIC and BIC as a rule selection criterion are listed respectively. Since intersection SNP combinations provide higher sensitivity values, we consider SNP intersection sets for each replicate for further analysis. Hence, we have seven alternative solutions having a sensitivity value more than 0.90. If a replication includes more than six SNPs in the intersection SNP set, we apply DTREG to reduce the SNP number into six. We next execute our decision rule code for each intersection SNP set. The sensitivity value of each intersection SNP set is given in Table 9.4.

Table 9.4. Sensitivity value of each solution of a genetic algorithm based feature selection method for seven replications

	Sensitivity (%) – AIC based GA	Sensitivity (%) – BIC based GA
Replicate 1	90.51	90.0
Replicate 2	92.54	93.17
Replicate 3	93.82	93.60
Replicate 4	91.55	93.34
Replicate 5	91.59	93.46
Replicate 6	90.70	90.84
Replicate 7	90.23	90.34
Average Sensitivity	91.56	92.11

These seven alternative SNP combinations include common SNPs providing a very close sensitivity values (min: 90.23 - max: 93.82). Moreover, using Akaike or Bayesian

information criterion as a fitness score does not affect the results considerably. Both methods give very similar results including at most two different SNPs.

Although the data sets are different for each replicate, the SNP combinations consist of some common SNPs. To determine the mostly observed SNPs, we also counted the occurrence of each SNP placed in seven alternative combination sets (intersection SNP set). The mostly observed SNP is denseSNP6_3437 which is mentioned by Uh et al. (2007) as the most significant SNP leading to the disease. Moreover, according to the data answers provided us by the Genetic Analysis Workshop, DR type at the HLA locus on chromosome 6 is the trait locus and includes denseSNP6_3734. In the following table, we listed mostly observed SNPs in descending order. The second column indicates the names of the important SNPs; the third column indicates the occurrence of each SNP in seven alternative SNP combinations; the fourth column displays SNPs that are correlated with the SNP in the second column. PF denotes the previously determined significant SNPs.

Table 9.5. The most significant SNPs obtained from seven replications

			CORELATED SNPs
PF	denseSNP6_3437	16	denseSNP6_3413 (2), denseSNP6_3419 (1) denseSNP6_3416 (3), denseSNP6_3818 (2)
PF	denseSNP6_3430	8	denseSNP6_3414 (2),denseSNP6_3427 (3)
	denseSNP6_3446	7	
	denseSNP6_3429	7	denseSNP6_3415 (1)
	denseSNP6_3434	7	
PF	denseSNP6_3440	6	
	denseSNP6_3438	5	
PF	denseSNP6_3439	5	denseSNP6_3437, denseSNP6_3430
PF	denseSNP6_3426	4	
PF	denseSNP6_3442	4	
	denseSNP6_3947	4	
	denseSNP6_3870	4	
	denseSNP6_3443	3	
PF	denseSNP6_3436	2	denseSNP6_3429
	denseSNP6_3435	2	

In the following table, all red shaded SNPs (denseSNP6_3437, denseSN6_3430, denseSNP6_3440, denseSNP6_3438, denseSNP6_3439) are highly correlated to each other (Correlation > 0.85). The black shaded SNPs do not have a high correlation with any other SNP. DenseSNP6_3429 are highly correlated with denseSNP6_3436.

DenseSNP6_3437 and denseSNP_3430, denseSNP_3446, denseSNP6_3429 and denseSNP6_3434 are observed in all solutions. While counting the occurrence of SNPs, we consider the correlated SNPs of each SNP because we deleted the correlated SNPs from the data before applying the genetic algorithm based feature selection method. Thus, some of the SNPs cannot be observed in some populations, but instead of that SNP, the correlated SNP can be chosen as a significant SNP. This is the reason why some SNPs have an occurrence number bigger than seven.

Zhang et al. (2007) proposes a nonparametric association analysis and combines family and case control genotype data. The test that they propose performs better than traditional case-control chi-square test and transmission disequilibrium test in terms of type 1 error rate. They apply their method to the same GWA 15 simulated data set considering just chromosome 6. According to their results, the most likely interval for a major gene is between 49.4262 cm and 49.5184 cm on chromosome 6. They found denseSNP6_3439, denseSNP6_3442, denseSNP6_3437, denseSNP6_3436, denseSNP6_3440, denseSNP6_3430 and denseSNP6_3426 as the most significant SNPs. As it can be seen from the table above, we can detect the previously determined significant SNPs as well as the new significant SNPs.

In literature, there is only one study which constructs SNP combinations. Uh et al. (2007) develop a Bayesian variable-selection logistic regression model to find the disease causing SNPs combinations. They apply their method to the SNPs on chromosome 6 of GWA 15 simulated data. They find just one significant SNP combination including denseSNP6_3437 and denseSNP6_3439. They also investigated the average prediction error of that SNP combination and find the best prediction performance as 86.94 %. Since we can classify patients more accurately by adding additional SNPs to the SNP combinations, our method performs better in terms of classification accuracy (>90 %). Moreover, Zhang et al. (2007) investigate SNPs individually and can find just six important SNPs which are listed in below, but we investigate SNP combinations and thus can extract more significant SNPs (15 SNPs).

To compare our results with the previous works, we apply our decision rule algorithm to the independent populations by selecting just the previously determined SNPs and newly found SNPs. For this reason we pick the mostly observed six SNPs from our analysis. These SNPs are: denseSNP6_3429, denseSNP6_3430, denseSNP6_3434, denseSNP6_3437, denseSNP6_3440, denseSP6_3446. The selected SNPs and the related sensitivity ratio are given for each population in Table 9.6.

According to Table 9.6, newly determined SNPs; denseSNP6_3429, denseSNP6_3434 and denseSNP6_3446 provide a higher prediction ratio. This indicates that the SNP combinations including denseSNP6_3429, denseSNP6_3434 and denseSNP6_3446 are more powerful than the combinations having denseSNP6_3439, denseSNP6_3426 and denseSNP6_3442. Moreover, while Zhang et al. (2007) find just six significant SNPs by investigating SNPs individually, we find 15 significant SNPs by constructing SNP combinations. This reveals the fact that investigating SNPs individually can lead some SNPs to be disregarded. Our method, thus, can find more powerful SNP combinations than the previously mentioned SNP combination (just including two SNPs) and individually significant SNPs (six SNPs).

Table 9.6. Comparison of newly and previously detected SNPs

Previously detected SNPs	
3437, 3430, 3440, 3439, 3426, 3442, 3436	Sensitivity Ratio (%)
Population1	91,23
Population2	92,12
Population3	91,91
Population4	89,08
Population5	89,74
Population6	91,23
Average Sensitivity Ratio	90.88
Newly detected SNPs	
3429, 3430, 3434, 3437, 3440, 3446	Sensitivity Ratio (%)
Population1	89,5
Population2	92,77
Population3	92,51
Population4	92,97
Population5	92,55
Population6	91,92
Average Sensitivity Ratio	92,03

In the literature, in order to construct decision rules, decision tree algorithms are mostly applied. However, most genetic data are noisy and decision tree algorithms are inefficient to classify a case/control data. For this reason, scientists have been developing a decision tree forest algorithm for biological data recently (Tong et al., 2004). However, while decision tree algorithms apply for the decision rules to be an

output, decision tree forest algorithms do not. Thus, we develop our own decision rule extraction method. To test the performance of our decision rule extraction method, first we apply DTREG decision tree forest and single decision tree algorithms to the tag SNPs set (the SNP set used in genetic algorithm based feature selection method as an input) for the same six populations.

The reason of comparing our results with DTREG is that it is the only software that includes decision tree forest module. Moreover, DTREG single decision tree algorithm provides a better classification performance than other single decision tree tools like Weka and RapidMiner. The sensitivity results of each population are displayed in Table 9.7.

Table 9.7. Sensitivity value of solutions obtained from DTREG

SNPs found by DTREG - Decision Tree Forest	
changes in each repetition	Sensitivity Ratio (%)
Population1	78,6
Population2	79,41
Population3	78,52
Population4	79,46
Population5	78,94
Population6	79,54
Average Sensitivity Ratio	79,1
SNPs found by DTREG - Single Decision Tree	
changes in each repetition	Sensitivity Ratio (%)
Population1	72,66
Population2	71,21
Population3	75,39
Population4	76,36
Population5	74,71
Population6	79,11
Average Sensitivity Ratio	74,9

As it can be seen from Table 9.7, although DTREG decision tree forest algorithm provides a higher sensitivity value than single decision tree algorithm, it is still smaller than the sensitivity value obtained from our decision rule extraction method. While DTREG can find an average sensitivity value approximately as 0.80, our decision rule extraction method can find average sensitivity as 0.92. Besides, while denseSNP6_3437 are observed in our all alternative SNP combinations, DTREG decision tree and decision tree forest algorithms do not detect it as a significant SNP for each population.

Next, to determine the genotype-SNP effect, we investigated the output of our decision rule extraction method to a random population by selecting mostly observed six-SNPs (denseSNP6_3429, denseSNP6_3430, denseSNP6_3434, denseSNP6_3437, denseSNP6_3440, denseSNP6_3446). Although set covering algorithm gives the best genotype-SNP rule set, a researcher want to learn the SNP combinations which are mostly observed in cases. Thus he/she can consider ratio1 as a rule selection criterion in his analysis. Another scientist may want to learn the rules which can explain most of the patients and may select maximum cardinality as a rule selection criterion. Therefore we modify the weights in the objective function of the weighted set covering model for three different aims. But all of the rule sets based on different criterion give the same classification accuracy because they consider the same patient set covered by all rules. For each criterion the number of selected rules is listed with respect to the number of SNPs in a rule in Table 9.8.

Since weighted set covering algorithm is an optimal search method, it selects the minimum number of rules. Rules including six-SNPs are rarely selected. Thus, investigating SNP combinations including more than six SNPs may be unnecessary.

Table 9.8. Number of selected rules according to each criterion

	General Set Covering Alg.	Set Covering Alg. Based on Max Ratio1	Set Covering Alg. Based on Max. Cardinality	Total Number of Rules
Population1	4	4	4	12
Population2	5	5	6	16
Population3	8	8	8	24
Population4	6	6	8	20
Population5	8	7	9	24
Population6	5	5	6	16
min;max	4;8	4;8	4;8	112

The rules selected by weighted set covering algorithm is given in Table 9.9, Table 9.10 and Table 9.11. According to the optimal rule set obtained from weighted set covering model when all the weights are equal to one, all SNPs are in homozygote form. However, considering other rule sets which are based on maximum cardinality and maximum ratio1 criteria some of the SNPs can be in heterozygote form. Thus, we considered all rules in all rule sets and extracted the genotypes of each SNP.

Table 9.9. Selected rules according to general set covering algorithm

	RULE SETS ACCORDING TO GENERAL SET COVERING ALGORITHM								
	SNP	**G**	**SNP**	**G**	**SNP**	**G**	**SNP**	G	G: GENOTYPE
RULE	3437	dd							POPULATINON1
RULE	3440	dd	3446	dd					
RULE	3430	dd	3446	dd					
RULE	3430	dd	3440	dd					
RULE	3437	dd							POPULATINON2
RULE	3434	dd	3446	dd					
RULE	3429	dd	3446	dd					
RULE	3434	dd	3434	dd					
RULE	3446	dd	3440	dd					
RULE	3437	dd							POPULATION3
RULE	3434	dd	3446	dd					
RULE	3446	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3429	dd	3440	dd					
RULE	3437	dd							POPULATION4
RULE	3434	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3434	dd					
RULE	3430	dd	3440	dd	3446	dd			
RULE	3437	dd							POPULATION5
RULE	3434	dd	3446	dd					
RULE	3440	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3434	dd					
RULE	3430	dd	3440	dd					
RULE	3429	dd	3430	dd					
RULE	3437	dd							POPULATION6
RULE	3434	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3429	dd	3430	dd	3440	dd			
	D=MINOR ALLELE, d=MAJOR ALLELE								

Table 9.10. Selected rules based on maximum ratio1 criterion

	RULE SETS ACCORDING TO SET COVERING ALGORITHM BASED ON MAXIMIM RATIO CRITERION								
	SNP	**G**	**SNP**	**G**	**SNP**	**G**	**SNP**	G	G: GENOTYPE
RULE	3437	dd							POPULATINON1
RULE	3440	dd	3446	dd					
RULE	3430	dd	3446	dd					
RULE	3430	dd	3440	dd					
RULE	3437	dd							POPULATINON2
RULE	3440	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3429	Dd	3434	dd	3446	dd			
RULE	3437	dd							POPULATION3
RULE	3434	dd	3446	dd					
RULE	3440	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3446	dd					
RULE	3430	dd	3434	dd					
RULE	3430	dd	3440	dd					
RULE	3437	dd							POPULATION4
RULE	3434	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3434	dd					
RULE	3430	dd	3440	dd					
RULE	3437	dd							POPULATION5
RULE	3440	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3440	dd					
RULE	3429	dd	3430	dd					
RULE	3434	dd	3440	Dd	3446	dd			
RULE	3437	dd							POPULATION6
RULE	3434	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3430	dd	3440	dd					
	D=MINOR ALLELE, d=MAJOR ALLELE								

Table 9.11.Selected rules according to set covering algorithm based on max. cardinality

	RULE SETS BASED ON MAXIMIM CARDINALITY								
	SNP	**G**	**SNP**	**G**	**SNP**	**G**	**SNP**	G	GENOTYPE
RULE	3437	dd							POPULATINON1
RULE	3440	dd	3446	dd					
RULE	3430	dd	3440	dd					
RULE	3429	dd	3430	dd	3440	Dd	3446	dd	
RULE	3437	dd	3440	Dd					POPULATINON2
RULE	3440	Dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3449	dd					
RULE	3437	dd	3440	dd					
RULE	3429	dd	3434	dd	3446	dd			
RULE	3437	dd							POPULATION3
RULE	3434	dd	3446	dd					
RULE	3440	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3429	Dd	3430	dd	3434	dd			
RULE	3429	dd	3430	dd	3440	dd			
RULE	3429	dd	3430	dd	3440	dd	3446	dd	
RULE	3434	dd	3446	dd					POPULATION4
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3437	dd	3440	dd					
RULE	3437	dd	3440	Dd	3446	Dd			
RULE	3429	Dd	3430	dd	3434	dd			
RULE	3437	dd	3440	Dd	3446	dd			
RULE	3430	dd	3440	dd	3446	dd			
RULE	3437	dd	3440	Dd					POPULATION5
RULE	3440	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3437	dd	3440	dd					
RULE	3430	dd	3440	dd					
RULE	3429	dd	3430	dd					
RULE	3429	Dd	3430	dd	3434	dd			
RULE	3434	dd	3440	Dd	3446	dd			
RULE	3437	dd	3440	Dd					POPULATION6
RULE	3434	dd	3446	dd					
RULE	3429	dd	3434	dd					
RULE	3434	dd	3440	dd					
RULE	3437	dd	3440	dd					
RULE	3437	dd	3430	dd	3440	dd			

Table 9.12. Significant genotype of each significant SNP

SNP name	Genotypes	Occurrence	Percentage of Occurrence
denseSNP6_3434	dd	53	47,3 %
denseSNP6_3440	Dd, dd	52	46,4 %
denseSNP6_3446	Dd, dd	39	34,8 %
denseSNP6_3430	dd	30	26,7 %
denseSNP6_3429	Dd, dd	29	25,8 %
denseSNP6_3437	dd	24	21,4 %

To test the performance of our decision rule extraction method, we also apply DTREG-single decision tree algorithm to the six-SNP combination determined by our feature selection method. While DTREG single decision tree algorithm can classify instances with an average 76.36 % prediction accuracy, our decision rule extraction method can provide higher prediction accuracy (90-92%). In table 9.13 the sensitivity values of each population that is calculated according to DTREG single decision tree algorithm are listed.

Table 9.13. Sensitivity values calculated by DTREG-single decision tree

DTREG - Single Decision Tree	
3429, 3430, 3434, 3437, 3440, 3446	Sensitivity Ratio (%)
Population1	80.74
Population2	81.14
Population3	80.74
Population4	72.20
Population5	70.73
Population6	72.66
Average Specificity Ratio	76.36

CHAPTER 10

CONCLUSION AND FUTURE RESEARCH

In this thesis, we propose a genetic algorithm based feature selection method and decision rule extraction method in order to determine the significant SNP combinations and significant SNP-genotype relations (rules). Our experimental results show that the proposed algorithm provides better classification accuracy than previous works. Moreover, the significant SNP combinations determined by us explain more patients than other method which considers the SNP combinations.

In conclusion, our genetic algorithm based feature selection method can construct equally significant SNP combinations which provide better classification accuracy than decision tree forest and single decision tree algorithms. Moreover, since we consider SNP combinations, we can detect the power of SNP groups to explain the disease. While investigating SNPs individually can only find six important SNPs, our feature selection method can detect fifteen significant SNPs. Any six-SNP combinations by using fifteen important SNPs in Table 8.4 can lead to similar classification accuracy because there is a little difference with the SNP combinations. While the previous work (Uh. Et al., 2007) can detect only one SNP combination including two SNPs with a lower prediction performance (at most 86.94 %); our genetic algorithm based feature selection method can detect more powerful SNP combinations. Besides, our decision rule extraction method also performs better than current decision tree and decision forest algorithms of DTREG. While DTREG single decision tree algorithm can detect rules with average 76.36 % classification accuracy including six significant SNPs which are determined by us, we can provide 92.03% classification accuracy with the same SNP combination.

Since the genetic factors are not the only reason of a complex disease, further research may focus on constructing SNP combinations by not only considering genetic factors but also including environmental factors to the model to better explain the disease.

Bibliography

[1] Ahmad, F. K., N. M. Norwawi, S. Deris, N. H. Othman. 2008. A review of feature selection techniques via gene expression profiles. International Symposium on Information Technology. **2** 1-7.

[2] Akaike, H. 1987. Factor Analysis and AIC. Psychometrika. **52** 317-332.

[3] AL, P., N. J. Patterson, R. M. Plenge, W. E. Weinblatt, N. A. Shadick, D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. **38**(8) 904-909.

[4] Aldrich, H. J., F. D. Nelson. 1995. Linear probability, logit, and probit models, Newbury Park, Sage.

[5] Silverman, B. W. 1986. Density estimation for statistics and data analysis. 1st edition, Chapman and Hall, CRC Press.

[6] Barrett, J. C., B. Fry, J. Maller, M. J. Daly. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. **21** 263-265.

[7] Bekkerman, R., R. El-Yaniv, N. Tishby. 2003. Winter. Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research, (JMLR). **3** 1183-1208.

[8] Ben-Bassat, M. 1982. Pattern recognition and reduction of dimensionality. In: Handbook of Statistics, North Holland, 773–791.

[9] Bozdoğan, H. 1987. Model selection and unified Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika, **52** 345-370.

[10] Breiman, L. 1999. Random forests, random features. Technical Report 567, Department of Statistics, University of California, Berkeley.

[11] Burnham, K. P., D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd edition, Springer-Verlag.

[12] Burnham, K. P., D. R. Anderson. 2004. Multimodel inference understanding AIC and BIC in model selection, Thousand Oaks: Sage.

[13] Carlson, C. S., A. Michael, J. R. Mark, J. D. Smith, L. Kruglyak, D. A. Nickerson. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nature Genetics. **33** 518–521.

[14] Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, D.A. Nickerson. 2003. Selecting a maximally informative set of Single-Nucleotide Polymorphisms for association analyses using linkage disequilibrium. The American Journal of Human Genetics. **74**(1) 106-120.

[15] Caruana, R., V. De Sa. 2003. Benefitting from the variables that variable selection discards. of Machine Learning Research (JMLR). **3** 1245-1264.

[16] Chang, H. W., L. Y. Chuang, C. H. Ho, P. L. Chang, C. H. Yang. 2008. Odds ratio-based genetic algorithms for generating SNP barcodes of genotypes to predict disease susceptibility. OMICS. **12**(1) 71-81.

[17] Chatterjee, S., A. Hadi, B. Price. 1999. Regression analysis by example. New York, Wiley.

[18] Cho, S. B., H. H. Won. 2003. Machine learning in DNA microarray analysis for cancer classification. In Proceedings of the First Asia-Pacific Bioinformatics Conference.

[19] Daly, J. M., J. D. Rioux, S. F. Schaffner, T. J. Hudson, E S. Lander. 2001. High-resolution haplotype structure in the human genome. Nature Genetics. **29** 229-232.

[20] Schaid, D. J., A. J. Batzler, G. D. Jenkins, M. A. Hildebrandt. 2006. Exact tests of Hardy Weinberg equilibrium and homogeneity of disequilibrium across Strata. The American Journal of Human Genetics. **79**(6) 1071-1080.

[21] Doak, J. 1992. An evaluation of feature selection methods and their application to computer security. Technical report, Davis, CA: University of California, Department of Computer Science.

[22] Duda, R. O., P. E. Hart, D. G. Stork. 2001. Pattern classification. Wiley Interscience, New York.

[23] Foraita, R., K. Bammann, I. Pigeot. 2008. Modeling gene-gene interactions using graphical chain models. Human Heredity. **65** 47-56.

[24] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classifiation. Journal of Machine Learning Research (JMLR). **3** 1289-1306.

[25] Freedman, M. 2004. Assessing the impact of population stratification on genetic association studies. Nature Genetics. **36** 4.

[26] Gabriel, S. B., S. F. Schaffner, H. Nyguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. A. Lochner, M. Faggart, S. N. Liu-Cordero, C. R. A.

Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, D. Altshuler. 2003. The structure of haplotype blocks in the human genome. Science Express. **296**(5576) 2225-2229.

[27] Globerson, A., N. Tishby. 2003. Sufficient dimensionality reduction. Journal of Machine Learning Research (JMLR). **3** 1307-1331.

[28] Gong, B., Z. Guo, J. Liu, G. Zhu, S. Lv, S. Rao, X. Li. 2005. Application of a genetic algorithm-suport vector machine hybrid for prediction of clinical phenotypes based on genome wide snp profiles of sib pairs. Fuzzy Systems and Knowledge Discovery. **3614** 830-835.

[29] Gopalakrishnan, S., Z. Qin. 2006. Tag SNP Selection based on pairwise LD criteria and power analysis in association studies. Pacific Symposium on Biocomput. 511-22.

[30] Guyon, I., A. Elisseeff. 2003. An introduction to variable and feature selection, Journal of Machine Learning Research (JMLR). **3** 1157-1182.

[31] Hao, K. 2007. Genome-wide selection of tag SNPs using multiple-marker correlation. Bioinformatics. **23** 3178-3184.

[32] Handels, H., J. Kreusch, S. J. Pöppl. 1999. Feature selection for optimized skin tumor recognition using genetic algorithms. Artif Intelligence in Medicine. **16**(3) 283-97.

[33] Hosmer, D. W., S. Lemeshow. 2000. Applied logistic regression (2nd Edition). New York, Wiley.

[34] Horne, B. D., N. J. Camp. 2004. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. Genetic Epidemiology. **26**(1) 11-21.

[35] Horng, J. T., K. C. Hu, L. C. Wu, H. D. Huango, H. C. Lai, T. Y. Chu. 2004. Identifying the combination of genetic factors that determine susceptibility to cervical cancer. Bioinformatics and Bioengineering. BIBE 2004. Proceedings, Fourth IEEE Symposium on 19-21 May. 73 – 78.

[36] Johansson, A., F. Marroni, C. Hayward, C. S. Franklin, A. V. Kririchenko, I. Jonasson, A. A. Hicks, S. H. Wild, I. V. Zorloltseva, J. F. Wilson, I. Rudan, H. Campbell, C. Pattaro, P. Pramstaller, B. A. Oostra, A. F. Wright, C. M. Duijn, Y. S. Auichenko, U. Gyllensten. 2009. Linkage and genome wide association analysis of obesity-related phenotypes: Association of weight with the MGAT1 Gene. Obesity (2009). doi:10.1038/oby.2009.359

[37] Kathiresan, S., B. f. Voight, S. Purcell, K. Musunuru, D. Ardissino, P. M. Mannucci , S. Anand, J. C. Engert, N. J. Samani, H. Schunkert, J. Erdmann, M. P. Reilly, D. J. Rader, T. Morgan, J. A. Spertus, M. Stoll, D. Girelli, P. P. McKeown, C. C. Patterson, D. S. Siscovick, C. J. O'Donnell, R. Elosua, L. Peltonen, V. Salomaa, S. M. Schwartz, O. Melander, D. Altshuler, D. Ardissino,

P. A. Merlini, C. Berzuini, L. Bernardinelli, F. Peyvandi, M. Tubaro, P. Celli, M. Ferrario, R. Fetiveau, N. Marziliano, G. Casari, M. Galli, F. Ribichini, M. Rossi, F. Bernardi, P. Zonzin, A. Piazza, P. M. Mannucci, S. M. Schwartz, D. S. Siscovick, J. Yee, Y. Friedlander, R. Elosua, J. Marrugat, G. Lucas, I. Subirana, J. Sala, R. Ramos, S. Kathiresan, J. B. Meigs, G. Williams, D. M. Nathan, C. A. MacRae, C. J. O'Donnell, V. Salomaa, A. S. Havulinna, L. Peltonen, O. Melander, G. Berglund, B. F. Voight, J. N. Hirschhorn, R. Asselta, S. Duga, M. Spreafico, K. Musunuru, M. J. Daly, S. Purcell, B. F. Voight, S. Purcell, J. Nemesh, J. M. Korn, S. A. McCarroll, S. M. Schwartz, J. Yee, S. Kathiresan, G. Lucas, I. Subirana, A. Surti, C. Guiducci, L. Gianniny, D. Mirel, M. Parkin, N. Burtt, S. B. Gabriel, J. R. Thompson, P. S. Braund, B. J. Wright, A. J. Balmforth, S. G. Ball, A. S. Hall, H. Schunkert, J. Erdmann, P. Linsel-Nitschke, W. Lieb, A. Ziegler, I. König, C. Hengstenberg, M. Fischer, K. Stark, A. Grosshennig, M. Preuss, H. E. Wichmann, S. Schreiber, H. Schunkert, N. J. Samani, J. Erdmann, W. Ouwehand, C. Hengstenberg, P. Deloukas, M. Scholz, F. Cambien, M. P. Reilly, M. Li, Z. Chen, R. Wilensky, W. Matthai, A. Qasim, H. H. Hakonarson, J. Devaney, M. S. Burnett, A. D. Pichard, K. M. Kent, L. Satler, J. M. Lindsay, R. Waksman, S. E. Epstein, D. J. Rader,, T. Scheffold, K. Berger, M. Stoll, A. Huge, D. Girelli, N. Martinelli, O. Olivieri, R. Corrocher, T. Morgan, J. A. Spertus, P. McKeown, C. C. Patterson, H. Schunkert , E. Erdmann, P. Linsel-Nitschke, W. Lieb, A. Ziegler, IR. König, C. Hengstenberg, M. Fischer, K. Stark, A. Grosshennig, M. Preuss, H. E. Wichmann, S. Schreiber, H. Hólm, G. Thorleifsson, U. Thorsteinsdottir, K. Stefansson, J. C. Engert, R. Do, C. Xie, S. Anand, S. Kathiresan, D. Ardissino, P. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphism and copy number variants. Nature Genetics. **41**(3) 334-41.

[38] Kim, YS., W. N. Street, F. Menczer. 2003. Data mining: opportunities and challenges. 80–105.

[39] Koller, D., M. Sahami. 1996. Toward optimal feature selection, in Proceedings of International Conference on Machine Learning.

[40] Küçükural, A., C. Meydan, D. Yörükoğlu, U. Sezerman. 2009. Biomarker detection for Hexachlorobenzene toxicity using genetic algorithms on gene expression data. (submitted).

[41] Langley, P. 1994. Selection of relevant features in machine learning. In Proceedings of the AAAI Fall Symposium on Relevance. 1–5.

[42] Li, W., D. R. Nyholt. 2001. Marker selection by Akaike information criterion and Bayesian information criterion. Genetic Epidemiology. **21**(1) 272-277.

[43] Liao, F.T. 1994. Interpreting probability models, logit, probit, and other generalized linear models. Thousand Oaks. CA. Sage.

[44] Liu, J. J., G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, X. B. Ling. 2005. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics **21**(11) 2691-2697.

[45] Long, J. S. 1997. Regression models for categorical and limited dependent variables. Thousand Oaks. CA. Sage.

[46] Long, J. S., F. Jeremy. 2006. Regression models for categorical dependent variables using Stata. College Station. Stata Press.

[47] McCarthy, M. I., E. Zeggini. 2009. Genome-wide association studies in type 2 diabetes. Current Diabetes Reports. **9**(2) 164-171.

[48] Menard, S. W. 2002. Applied logistic regression analysis. Thousand Oaks. Sage.

[49] Murthy, S. K. 1995. On growing better decision trees from data. PhD thesis. Johns Hopkins University, Baltimore, Maryland.

[50] Mao, J., K. Mohiuddin, A. K. Jain. 1994. Parsimonious network design and feature selection through node pruning. In Proceedings of 12th ICPR, Jerusalem, 662-624.

[51] Nakamichi, R., S. Imoto, S. Miyano. 2004. Logistic regression model of binary disease trait for case/control study considering interactions between SNPs and environments. Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04). 73.

[52] Narendra, P. M., K. Fukunaga. 1977. A branch and bound algorithm for feature subset selection. IEEE Transactions and Computers. **26**(9) 917 922.

[53] Nickerson, D. A., S. L. Taylor, S. F. Fullerton, K. M. Weiss, A. G. Clark, J. H. Stengard, V. Salomaa, E. Boerwinkle, C. F. Sing. 2000. Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Research. **10** 532–1545.

[54] Nyholt, R.D., W. Li. 2001. Marker selection by AIC and BIC. Genetic Epidemiology. **21**(l1) 8272-8277.

[55] O'Connel, A. A. 2006. Logistic regression models for ordinal response variables. Thousand Oaks. Sage.

[56] Ooi, C. H., P. Tan. 2003. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics. **19(**1) 37-44.

[57] Pearson, T. A., T. A. Manolio. 2008. How to interpret a genome-wide association study. The Journal of American Medical Association (Jama). **299**(11) 1335-1344.

[58] Phuong, T. M., Z. L. Russ, B. Altman. 2005. Choosing SNPs using feature selection. IEEE Computational Systems Bioinformatics Conference (CSB'05). 301-309.

[59] Risch, N., K. Merikangas. 1996. The future of genetic studies of complex Human Diseases. Sciences. **273**(5281) 1516-7.

[60] Saeys, Y., I. Inza, L. Pedro. 2007. A review of future selection techniques in bioinformatics. Bioinformatics. **23**(19) 2507-2517.

[61] Samani, N.J., N. J. Samani, F. M. Sci, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B.Mayer, R. J. Dixon, T. Meitinger, P. Braund, H. Wichmann, J. H. Barrett, I. R. König, S. E. Stevens, S. Szymczak, D. A. Tregouet, M. M. Iles, F. Pahlke, H. Pollard, W. Lieb, F. Cambien, M. Fischer, W. Ouwehand, S. Blankenberg, A. J. Balmforth, A. Baessler, S. G. Ball, T. M. Strom, I. Brænne, C. Gieger, P. Deloukas, M. D. Tobin, A. Ziegler, J. R. Thompson, H. Schunkert. 2007. Genome wide association analysis of coronary artery disease. New England Journal of Medicine. **357** 443-453.

[62] Segall, R.S., Q. Zhang. 2006. Applications of neural network and genetic algorithm data mining techniques in bioinformatics knowledge discovery - a preliminary study. Proceedings of Southwest Decision Sciences Institute. March 1-4, Oklahoma City.

[63] Schwarz, G.E. 1978. Estimating the dimension of a model. Annals of Statistics. **6**(2) 461–464.

[64] Siedlecki, W., J. Sklansky. 1989. A note on genetic algrorithms for large scale features selection. Pattern recognition letters. **10** 335 347.

[65] Siedlecki, W., J. Sklansky. 1988. An automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence. **2** 197–220.

[65] Shah, S. C., A. A. Kusiak. 2004. A data mining and genetic algorithm based gene/SNP selection. Artificial Intelligence in Medicine. **31** 183-196.

[67] Stumpf, M. P. H., P.J. Ingram, I. Nouvel, C. Wiuf. 2005. Statistical model selection methods applied to biological networks. Lecture Notes in Computer Science. Springer Berlin/Heidelberg.

[68] Syam, G., S. Q. Zhaohui. 2006. Tag SNP selection based on pairwise LD criteria and power analysis in association studies. Pacific Symposium on Biocomputing. **11** 511-522.

[69] Tan, P. N., M. Steinbach, V. Kumar. 2006. Introduction to data mining. Addison-Wesley.

[70] Tong, W., H. Hong, H. Fang, Q. Xie, R. Perkins. 2003. Decision Forest: Combining the predictions of multiple independent decision tree models. Journal of Chemical Information and Modeling Science. **43** 525-531.

[71] Tong, W., Q. Xie, H. Hong, H. Fang, L. Shi, R. Perkins, E. F. Petricoin. 2004. Using decision forest to classify Prostate cancer samples on the basis of seldi-tof ms data: Assessing chance correlation and prediction confidence. Environmental Health Perspectives. **112** 16.

[72] Uh, H. W., B. J. A. Mertens, H. Wijk, H. Putter, H. Houwelingen, J. Houwing-Duistermaat. 2007. Model selection based on logistic regression in a highly correlated candidate gene region. BMC Proceedings. I(Suppl I):SII4.

[73] Wagenmakers, E. J., S. Farrell. 2004. AIC model selection using Akaike weights. Psychonomic Bulletin and Review. **11**(1) 192-196.

[74] Wang, W. B., T. Jiang. 2008. A new model of multi-marker correlation for genome-wide tag SNP selection. Genome Inform. **21** 27-41.

[75] Waring, S. C., R. N. Rosenberg. 2008. Genome-wide association studies in Alzheimer disease. Archives of Neurology. **65**(3) 329-34.

[76] Wu, J., J. Wang, J. Chen. 2008. A genetic algorithm for single individual SNP haplotype assembly. Proceedings of the 2008 the 9th International Conference for Young Computer Scientists. 1012-1017.

[77] Weston, J., A. Elisseeff, B. Schölkopf. 2003. Use of zero norm with linear models and kernel methods. Journal of Machine Learning Research (JMLR). **3** 1439-1461.

[78] Xie, Q. et al., L. D. Ratnasinghe, H. Hong, R. Perkins, Z. Z. Tang, N. Hu, P. R. Taylor, W. Tong. 2005. Decision forest analysis of 61 Single Nucleotide Polymorphism in a case-control study of Esophageal cancer; a novel method. BMC Bioinformatics. **6**(2), S4.

[79] Zhang, J., X. Zhu, R. S. Cooper. 2007. An integrated genome-wide association analysis on rheumatoid arthritis data. BMC Bioinformatics Proceedings. 1 S35.

[80] Zheng, W., J. Long, Y. T. Gao, C. Li, Y. Zheng, Y. B. Xiang, W. Wen, S. Levy, S. L. Deming, J. L. Haines, K. Gu, A. M. Fair, Q. Cai, W. Lu, X. O. Shu. 2009. Genome-wide association study identified a new breast cancer susceptibility locus at 6q25.1. Nature Genetics. **41**(3) 324-328.

[81] Zhou, X., X. Wang, E. Dougherty. 2005. Gene selection using logistic regression based on AIC, BIC and MDL criteria. New Mathematics and Natural Computation. **1**(1) 129-145.

[82] http://www.ats.ucla.edu/stat/mult_pkg/fag/general/Psuedo_RSquareds.htm

[83] http://www.biostat.wisc.edu/~cook/642.tex/notes0412.pdf

[84] http://en.wikipedia.org/wiki/

[85] http://www.everythingbio.com/glos/definition.php?word=haplotype

[86] http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm#classification

[87] http://www.medterms.com/script/main/art.asp?articlekey=13053

[88] http://people.exeter.ac.uk/SEGLea/multvar2/disclogi.html

[89] http://www.orbel.be/workshops/dmor05/DMOR05Bontempi.pdf

[89] http://scholar.lib.vt.edu./theses/available/etd032799154323/unrestricted/chptr4.pdf

Appendix A

Results of the statistical measurements of significant SNP combinations

Table A.1. Statistical results of solutions obtained from population1 (replicate1)

	METHOD _ AIC			
REPLICATE_001	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,7815	0,308240342	0,304404173	0,34416313
SOLUTION2	0,77825	0,307522221	0,303503378	0,343518266
SOLUTION3	0,782875	0,309411939	0,30557577	0,345213851
SOLUTION4	0,7805	0,308479623	0,305008803	0,344377861
SOLUTION5	0,782125	0,308870511	0,305034342	0,344728493
REPLICATE_001	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,34416313	0,368544182	0	0,022825919
SOLUTION2	0,343518266	0,367699563	0	0,111738662
SOLUTION3	0,345213851	0,370214334	0	0,082744536
SOLUTION4	0,344377861	0,368517489	0	0,082827236
SOLUTION5	0,344728493	0,369457419	0	0,107442131
	METHOD _ BIC			
REPLICATE_001	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,7815	0,308240342	0,304404173	0,34416313
SOLUTION2	0,77825	0,307522221	0,303503378	0,343518266
SOLUTION3	0,782875	0,309411939	0,30557577	0,345213851
SOLUTION4	0,7805	0,308479623	0,305008803	0,344377861
SOLUTION5	0,782125	0,308870511	0,305034342	0,344728493
REPLICATE_001	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,34416313	0,368544182	0	0,022825919
SOLUTION2	0,343518266	0,367699563	0	0,111738662
SOLUTION3	0,345213851	0,370214334	0	0,082744536
SOLUTION4	0,344377861	0,368517489	0	0,082827236
SOLUTION5	0,344728493	0,369457419	0	0,107442131

Table A.2. Statistical results of solutions obtained from population2 (replicate2)

	METHOD _ AIC			
REPLICATE_002	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,788125	0,322484239	0,317194937	0,35726727
SOLUTION2	0,786375	0,320137711	0,315942747	0,355196678
SOLUTION3	0,78675	0,32210978	0,317732426	0,356937291
SOLUTION4	0,7875	0,320556472	0,315996729	0,355566685
SOLUTION5	0,78775	0,322236603	0,317859249	0,357049068
REPLICATE_002	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,35726727	0,387113717	0	0,111948468
SOLUTION2	0,355196678	0,38439133	0	0,051983622
SOLUTION3	0,356937291	0,385649198	0	0,022790705
SOLUTION4	0,355566685	0,385172403	0	0,081847452
SOLUTION5	0,357049068	0,386745295	0	0,091650651
	METHOD _ BIC			
REPLICATE_002	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,786625	0,320010117	0,316544712	0,355083898
SOLUTION2	0,785875	0,320210006	0,316015042	0,355260571
SOLUTION3	0,786	0,318547384	0,315081979	0,353789573
SOLUTION4	0,783875	0,317000778	0,312441034	0,352418206
SOLUTION5	0,786	0,319255618	0,315790213	0,35441659
REPLICATE_002	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,355083898	0,384023873	0	0,004032994
SOLUTION2	0,355260571	0,384206913	0	0,061208955
SOLUTION3	0,353789573	0,382268502	0	0,110441881
SOLUTION4	0,352418206	0,380622698	0	0,003093263
SOLUTION5	0,35441659	0,383164292	0	0,081373446

Table A.3. Statistical results of solutions obtained from population3 (replicate3)

	METHOD _ AIC			
REPLICATE_003	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,78325	0,30204463	0,298580639	0,339117229
SOLUTION2	0,784125	0,302425027	0,298596405	0,339461868
SOLUTION3	0,783625	0,301228178	0,297217241	0,338376917
SOLUTION4	0,783375	0,301834894	0,298006271	0,338927131
SOLUTION5	0,782875	0,30106378	0,296870526	0,338227749
REPLICATE_003	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,339117229	0,365382135	0	0,016311758
SOLUTION2	0,339461868	0,365604338	0	0,041311121
SOLUTION3	0,338376917	0,364451353	0	0,009502415
SOLUTION4	0,338927131	0,365130023	0	0,023577975
SOLUTION5	0,338227749	0,363792078	0	0,000991894
	METHOD _ BIC			
REPLICATE_003	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,7825	0,301135357	0,297853681	0,338292699
SOLUTION2	0,783625	0,30188284	0,298418848	0,338970592
SOLUTION3	0,78475	0,30141786	0,298136184	0,338548984
SOLUTION4	0,781	0,297861334	0,29330345	0,335315282
SOLUTION5	0,782375	0,300943505	0,296020991	0,338118597
REPLICATE_003	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,338292699	0,363912998	0	0,014165057
SOLUTION2	0,338970592	0,364526448	0	0,003017833
SOLUTION3	0,338548984	0,364580586	0	0,004809922
SOLUTION4	0,335315282	0,360313102	0	0,014270119
SOLUTION5	0,338118597	0,362910213	0	0,036030485

Table A.4. Statistical results of solutions obtained from population4 (replicate4)

	METHOD _ AIC			
REPLICATE_004	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,785	0,315050523	0,311222755	0,35086178
SOLUTION2	0,785	0,316002671	0,312721727	0,351708953
SOLUTION3	0,784	0,31582147	0,312175977	0,351547815
SOLUTION4	0,786	0,3167447	0,312916932	0,352368406
SOLUTION5	0,786125	0,316231661	0,312221619	0,351912532
REPLICATE_004	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,35086178	0,378355803	0	0,026951945
SOLUTION2	0,351708953	0,380238015	0	0,037179525
SOLUTION3	0,351547815	0,37999165	0	0,092557399
SOLUTION4	0,352368406	0,381489537	0	0,087296621
SOLUTION5	0,351912532	0,380776852	0	0,01028793
	METHOD _ BIC			
REPLICATE_004	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,78225	0,30984439	0,305105249	0,346210017
SOLUTION2	0,78425	0,314533384	0,31125244	0,350401193
SOLUTION3	0,784875	0,314843925	0,311562982	0,350677815
SOLUTION4	0,78125	0,313482388	0,310201444	0,349464121
SOLUTION5	0,783375	0,314001	0,310537782	0,349926687
REPLICATE_004	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,346210017	0,373247359	0	0,099197342
SOLUTION2	0,350401193	0,378684798	0	0,093137769
SOLUTION3	0,350677815	0,379384471	0	0,0482096
SOLUTION4	0,349464121	0,376767159	0	0,089466495
SOLUTION5	0,349926687	0,377727112	0	0,075221967

Table A.5. Statistical results of solutions obtained from population5 (replicate5)

	METHOD _ AIC			
REPLICATE_005	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,787875	0,318154678	0,313416375	0,353669533
SOLUTION2	0,7865	0,31773173	0,313722397	0,353294423
SOLUTION3	0,787625	0,317672578	0,31329876	0,353241945
SOLUTION4	0,787375	0,316673136	0,312299318	0,352354609
SOLUTION5	0,78875	0,317352284	0,313342951	0,352957711
REPLICATE_005	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,353669533	0,381735973	0	0,020186465
SOLUTION2	0,353294423	0,380701349	0	0,031627066
SOLUTION3	0,353241945	0,380931079	0	0,044766034
SOLUTION4	0,352354609	0,37993917	0	0,00694565
SOLUTION5	0,352957711	0,380756262	0	0,010218295
	METHOD _ BIC			
REPLICATE_005	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,787375	0,315709466	0,311882376	0,351497881
SOLUTION2	0,788875	0,314939583	0,31165922	0,35081262
SOLUTION3	0,787	0,312618838	0,309338474	0,348742576
SOLUTION4	0,786625	0,315136821	0,311127488	0,350988247
SOLUTION5	0,788375	0,316574626	0,312565292	0,352267083
REPLICATE_005	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,351497881	0,379536849	0	0,069769564
SOLUTION2	0,35081262	0,378601899	0	0,000382614
SOLUTION3	0,348742576	0,375513272	0	0,003861529
SOLUTION4	0,350988247	0,378519802	0	0,005419507
SOLUTION5	0,352267083	0,379680076	0	0,006262061

Table A.6. Statistical results of solutions obtained from population6 (replicate6)

	METHOD _ AIC			
REPLICATE_006	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,786125	0,325520079	0,321327762	0,360116532
SOLUTION2	0,78975	0,3267724	0,322215534	0,361214671
SOLUTION3	0,787125	0,324711139	0,321247921	0,359406183
SOLUTION4	0,785375	0,32499768	0,320805363	0,359657892
SOLUTION5	0,78725	0,324219284	0,319115594	0,358973888
REPLICATE_006	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,360116532	0,386159917	0	0,002332275
SOLUTION2	0,361214671	0,387021524	0	0,000699506
SOLUTION3	0,359406183	0,38523544	0	0,018978741
SOLUTION4	0,359657892	0,385495438	0	0,000870237
SOLUTION5	0,358973888	0,384734352	0	0,001100371
	METHOD _ BIC			
REPLICATE_006	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,787125	0,324246026	0,321329632	0,358997399
SOLUTION2	0,78525	0,324121318	0,319929001	0,35888775
SOLUTION3	0,787125	0,324448583	0,32080309	0,359175457
SOLUTION4	0,788125	0,324501075	0,320673307	0,359221592
SOLUTION5	0,78825	0,324876629	0,321231136	0,359551568
REPLICATE_006	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,358997399	0,385030617	0	0,004475674
SOLUTION2	0,35888775	0,384536195	0	0,006077299
SOLUTION3	0,359175457	0,384903453	0	0,002091754
SOLUTION4	0,359221592	0,385166	0	0,004994042
SOLUTION5	0,359551568	0,385441476	0	0,002198742

Table A.7. Statistical results of solutions obtained from population7 (replicate7)

	METHOD _ AIC			
REPLICATE_007	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,79	0,324352704	0,319973122	0,358766226
SOLUTION2	0,7875	0,32382673	0,319264666	0,358303998
SOLUTION3	0,7865	0,323169062	0,318971962	0,357725569
SOLUTION4	0,7885	0,323149194	0,318587129	0,357708087
SOLUTION5	0,789125	0,324044482	0,3196649	0,3584954
REPLICATE_007	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,358766226	0,387149983	0	0,03666968
SOLUTION2	0,358303998	0,386362818	0	0,058766173
SOLUTION3	0,357725569	0,385850109	0	0,082979186
SOLUTION4	0,357708087	0,385753048	0	0,104263826
SOLUTION5	0,3584954	0,386870941	0	0,076075663
	METHOD _ BIC			
REPLICATE_007	**CAR**	**McFaddens_R^2**	**McFaddens_R^2**	**RoxSnell_R^2**
SOLUTION1	0,78775	0,322589214	0,318574597	0,357215151
SOLUTION2	0,787	0,321951574	0,31884937	0,356653393
SOLUTION3	0,787125	0,321208021	0,316645956	0,355997705
SOLUTION4	0,788375	0,323389956	0,319922787	0,357919907
SOLUTION5	0,789125	0,322439695	0,318790044	0,35708347
REPLICATE_007	**Nagelkerke_R^2**	**Efron's_R2**	**LRT p value**	**Pro_Hosmer Test**
SOLUTION1	0,357215151	0,385322992	0	0,029711858
SOLUTION2	0,356653393	0,384461221	0	0,089505084
SOLUTION3	0,355997705	0,384245915	0	0,09758638
SOLUTION4	0,357919907	0,386200689	0	0,074513843
SOLUTION5	0,35708347	0,385419802	0	0,07351742

Appendix B

Detailed results of tag-SNPs selection

Table B.1. Tag SNPs of each population (replication – rep)

TAG SNPs						
REP1	REP2	REP3	REP4	REP5	REP6	REP7
SNP6_3281	SNP6_3437	SNP6_3437	SNP6_3437	SNP6_3020	SNP6_3437	SNP6_3437
SNP6_3427	SNP6_3353	SNP6_3434	SNP6_3776	SNP6_3765	SNP6_3026	SNP6_3428
SNP6_3418	SNP6_3773	SNP6_3407	SNP6_3028	SNP6_3049	SNP6_3428	SNP6_3031
SNP6_3781	SNP6_3406	SNP6_3262	SNP6_909	SNP6_3428	SNP6_3017	SNP6_2862
SNP6_3449	SNP6_3407	SNP6_3304	SNP6_3049	SNP6_3007	SNP6_2871	SNP6_3767
SNP6_3483	SNP6_3424	SNP6_3421	SNP6_3428	SNP6_2871	SNP6_3453	SNP6_3308
SNP6_3198	SNP6_3429	SNP6_3416	SNP6_3484	SNP6_2875	SNP6_3777	SNP6_3239
SNP6_3650	SNP6_3425	SNP6_2862	SNP6_2793	SNP6_3118	SNP6_2870	SNP6_2870
SNP6_3406	SNP6_3291	SNP6_3773	SNP6_3687	SNP6_3407	SNP6_3239	SNP6_3084
SNP6_3081	SNP6_3478	SNP6_3460	SNP6_2863	SNP6_3353	SNP6_3006	SNP6_3420
SNP6_3318	SNP6_3416	SNP6_3580	SNP6_3691	SNP6_3580	SNP6_3330	SNP6_3454
SNP6_3759	SNP6_3430	SNP6_3662	SNP6_3084	SNP6_3083	SNP6_3221	SNP6_3434
SNP6_2850	SNP6_3443	SNP6_3353	SNP6_3017	SNP6_3437	SNP6_3423	SNP6_3406
SNP6_2873	SNP6_2947	SNP6_3031	SNP6_3119	SNP6_3454	SNP6_3165	SNP6_3580
SNP6_3197	SNP6_3772	SNP6_3654	SNP6_3306	SNP6_3017	SNP6_3434	SNP6_2874
SNP6_3479	SNP6_3467	SNP6_3285	SNP6_3434	SNP6_3423	SNP6_3580	SNP6_3083
SNP6_3454	SNP6_3580	SNP6_3366	SNP6_3197	SNP6_3434	SNP6_3477	SNP6_3479
SNP6_2862	SNP6_3191	SNP6_3454	SNP6_3421	SNP6_2874	SNP6_3525	SNP6_3763
SNP6_3407	SNP6_2781	SNP6_2870	SNP6_3455	SNP6_3479	SNP6_3407	SNP6_2723
SNP6_3463	SNP6_3763	SNP6_2863	SNP6_3081	SNP6_3416	SNP6_3083	SNP6_3293
SNP6_3656	SNP6_2863	SNP6_3572	SNP6_3407	SNP6_3325	SNP6_3196	SNP6_3533
SNP6_3576	SNP6_3466	SNP6_3286	SNP6_3580	SNP6_3417	SNP6_3534	SNP6_3359
SNP6_3325	SNP6_3154	SNP6_3307	SNP6_2909	SNP6_3086	SNP6_3211	SNP6_3416
SNP6_3384	SNP6_3359	SNP6_3055	SNP6_3781	SNP6_3466	SNP6_3935	SNP6_3443
SNP6_3307	SNP6_3021	SNP6_3534	SNP6_3065	SNP6_3292	SNP6_2723	SNP6_3535
SNP6_3375	SNP6_3014	SNP6_3430	SNP6_3423	SNP6_3378	SNP6_3918	SNP6_3426
SNP6_3777	SNP6_3321	SNP6_3308	SNP6_2912	SNP6_3430	SNP6_3430	SNP6_3463
SNP6_3309	SNP6_2714	SNP6_3261	SNP6_3462	SNP6_3413	SNP6_3759	SNP6_3493
SNP6_2707	SNP6_3428	SNP6_3338	SNP6_3331	SNP6_3443	SNP6_2549	SNP6_147
SNP6_3436	SNP6_2721	SNP6_3293	SNP6_3430	SNP6_3763	SNP6_2874	SNP6_3309
SNP6_3662	SNP6_2705	SNP6_3870	SNP6_3739	SNP6_3544	SNP6_3375	SNP6_3497
SNP6_159	SNP6_3271	SNP6_2984	SNP6_3072	SNP6_3366	SNP6_3456	SNP6_3252
SNP6_3437	SNP6_3567	SNP6_2723	SNP6_3460	SNP6_3584	SNP6_3440	SNP6_3413
SNP6_2848	SNP6_3286	SNP6_3432	SNP6_3375	SNP6_3429	SNP6_2724	SNP6_3203
SNP6_3272	SNP6_2723	SNP6_3543	SNP6_3572	SNP6_3533	SNP6_3463	SNP6_3494
SNP6_3433	SNP6_3221	SNP6_3309	SNP6_3579	SNP6_3465	SNP6_3546	SNP6_3086
SNP6_3221	SNP6_3573	SNP6_2947	SNP6_3761	SNP6_3426	SNP6_3439	SNP6_3378

SNP6_3396	SNP6_3440	SNP6_3359	SNP6_3422	SNP6_3460	SNP6_3396	SNP6_3379
SNP6_3541	SNP6_3071	SNP6_3191	SNP6_3467	SNP6_3478	SNP6_3311	SNP6_3072
SNP6_3058	SNP6_3099	SNP6_3651	SNP6_3103	SNP6_2937	SNP6_3325	SNP6_3429
SNP6_3557	SNP6_2866	SNP6_3283	SNP6_3478	SNP6_3572	SNP6_3425	SNP6_3446
SNP6_3455	SNP6_3479	SNP6_3347	SNP6_3812	SNP6_3057	SNP6_147	SNP6_3396
SNP6_3038	SNP6_3771	SNP6_2705	SNP6_2772	SNP6_3421	SNP6_3415	SNP6_3164
SNP6_3580	SNP6_3444	SNP6_3494	SNP6_3797	SNP6_3467	SNP6_3086	SNP6_3425
SNP6_2706	SNP6_3454	SNP6_3287	SNP6_3818	SNP6_3444	SNP6_3289	SNP6_3439
SNP6_2795	SNP6_3331	SNP6_3761	SNP6_3446	SNP6_3236	SNP6_3191	SNP6_2713
SNP6_3037	SNP6_2800	SNP6_2724	SNP6_3337	SNP6_3797	SNP6_3459	SNP6_3560
SNP6_3025	SNP6_3025	SNP6_3655	SNP6_3396	SNP6_3546	SNP6_3467	SNP6_3347
SNP6_2724	SNP6_3775	SNP6_3038	SNP6_3173	SNP6_3037	SNP6_2713	SNP6_3460
SNP6_3462	SNP6_3294	SNP6_3264	SNP6_3413	SNP6_3462	SNP6_3449	SNP6_3279
SNP6_3572	SNP6_3471	SNP6_3584	SNP6_3763	SNP6_3195	SNP6_3331	SNP6_3594
SNP6_3366	SNP6_2707	SNP6_3014	SNP6_3424	SNP6_3440	SNP6_3203	SNP6_3447
SNP6_3425	SNP6_3026	SNP6_3424	SNP6_3449	SNP6_3321	SNP6_3402	SNP6_3188
SNP6_3467	SNP6_3584	SNP6_3778	SNP6_3465	SNP6_3446	SNP6_3567	SNP6_3338
SNP6_3417	SNP6_2722	SNP6_3763	SNP6_3272	SNP6_3447	SNP6_3236	SNP6_3058
SNP6_3083	SNP6_3413	SNP6_3533	SNP6_3443	SNP6_3579	SNP6_3286	SNP6_3325
SNP6_3402	SNP6_3055	SNP6_3337	SNP6_2724	SNP6_3567	SNP6_3272	SNP6_2724
SNP6_3654	SNP6_3426	SNP6_3378	SNP6_3438	SNP6_3759	SNP6_3544	SNP6_3384
SNP6_3556	SNP6_2914	SNP6_2875	SNP6_3385	SNP6_3099	SNP6_2859	SNP6_3236
SNP6_2844	SNP6_3056	SNP6_3772	SNP6_2848	SNP6_3442	SNP6_3770	SNP6_2894
SNP6_2717	SNP6_2870	SNP6_3384	SNP6_3184	SNP6_2848	SNP6_3447	SNP6_3344
SNP6_2723	SNP6_3382	SNP6_3265	SNP6_3236	SNP6_3396	SNP6_3359	SNP6_3366
SNP6_3533	SNP6_3475	SNP6_3770	SNP6_3540	SNP6_3103	SNP6_3533	SNP6_3478
SNP6_3440	SNP6_3066	SNP6_3656	SNP6_3466	SNP6_3415	SNP6_3543	SNP6_3417
SNP6_3772	SNP6_2849	SNP6_3422	SNP6_3338	SNP6_3449	SNP6_3200	SNP6_2910
SNP6_3426	SNP6_3533	SNP6_3546	SNP6_3293	SNP6_2724	SNP6_3460	SNP6_3259
SNP6_3379	SNP6_2913	SNP6_3415	SNP6_3471	SNP6_3197	SNP6_3443	SNP6_2947
SNP6_3330	SNP6_3385	SNP6_3428	SNP6_3533	SNP6_3947	SNP6_2984	SNP6_3465
SNP6_3017	SNP6_3276	SNP6_3429	SNP6_3584	SNP6_3425	SNP6_3465	SNP6_3572
SNP6_3338	SNP6_3378	SNP6_3466	SNP6_3325	SNP6_3304	SNP6_2848	SNP6_3770
SNP6_3154	SNP6_3427	SNP6_3083	SNP6_3543	SNP6_3535	SNP6_2910	SNP6_2714
SNP6_3413	SNP6_3197	SNP6_3197	SNP6_3285	SNP6_2549	SNP6_3292	SNP6_3912
SNP6_3442	SNP6_2937	SNP6_3417	SNP6_3327	SNP6_3293	SNP6_3204	SNP6_3584
SNP6_3432	SNP6_3765	SNP6_3442	SNP6_3353	SNP6_3384	SNP6_2705	SNP6_2705
SNP6_3579	SNP6_3418	SNP6_2706	SNP6_3463	SNP6_3494	SNP6_3309	SNP6_3353
SNP6_2863	SNP6_3777	SNP6_3777	SNP6_3056	SNP6_3359	SNP6_3366	SNP6_3037
SNP6_3534	SNP6_2802	SNP6_2848	SNP6_3440	SNP6_3812	SNP6_3379	SNP6_2937
SNP6_2871	SNP6_3571	SNP6_3306	SNP6_3494	SNP6_3200	SNP6_3912	SNP6_3190
SNP6_3195	SNP6_3058	SNP6_3195	SNP6_2714	SNP6_3338	SNP6_3327	SNP6_3292
SNP6_3424	SNP6_3483	SNP6_2861	SNP6_2937	SNP6_3422	SNP6_3416	SNP6_3422
SNP6_3765	SNP6_3228	SNP6_3756	SNP6_2705	SNP6_3327	SNP6_3572	SNP6_3430
SNP6_3382	SNP6_3031	SNP6_3330	SNP6_2910	SNP6_3193	SNP6_3429	SNP6_3546
SNP6_3189	SNP6_3049	SNP6_3657	SNP6_3347	SNP6_3309	SNP6_3812	SNP6_3223

SNP6_3581	SNP6_3525	SNP6_3912	SNP6_3416	SNP6_3379	SNP6_3353	SNP6_3532
SNP6_2705	SNP6_2900	SNP6_3193	SNP6_2713	SNP6_2723	SNP6_3455	SNP6_3761
SNP6_3415	SNP6_3275	SNP6_3771	SNP6_3444	SNP6_2859	SNP6_3037	SNP6_3191
SNP6_3771	SNP6_3366	SNP6_3462	SNP6_3378	SNP6_3543	SNP6_3038	SNP6_3197
SNP6_3767	SNP6_3325	SNP6_3054	SNP6_3544	SNP6_3524	SNP6_3304	SNP6_3304
SNP6_3678	SNP6_3327	SNP6_3435	SNP6_3359	SNP6_3770	SNP6_3797	SNP6_3462
SNP6_3021	SNP6_3330	SNP6_3535	SNP6_3304		SNP6_3426	SNP6_3534
SNP6_3049	SNP6_3017	SNP6_2707	SNP6_3759		SNP6_3509	SNP6_3331
SNP6_3465	SNP6_3581	SNP6_3440	SNP6_2865		SNP6_2900	SNP6_3038
SNP6_3438	SNP6_3449	SNP6_3544	SNP6_2717		SNP6_3338	SNP6_3440
SNP6_3584	SNP6_3338	SNP6_3025	SNP6_2743		SNP6_3756	SNP6_3449
SNP6_3775	SNP6_2894	SNP6_3331	SNP6_3429		SNP6_3251	SNP6_3285
SNP6_3652	SNP6_3433	SNP6_3678	SNP6_3088		SNP6_3482	SNP6_3266
SNP6_3327	SNP6_3038	SNP6_3455	SNP6_3546		SNP6_3413	SNP6_3483
SNP6_3086	SNP6_3535	SNP6_3017	SNP6_3417		SNP6_3197	SNP6_3524
SNP6_3214	SNP6_3272	SNP6_3767	SNP6_3309		SNP6_3057	SNP6_3402
SNP6_3055	SNP6_147	SNP6_2982	SNP6_3442		SNP6_3077	SNP6_3870
SNP6_3060	SNP6_3322	SNP6_3436	SNP6_2706		SNP6_3306	SNP6_3444
SNP6_3494	SNP6_3214	SNP6_3200	SNP6_3200		SNP6_2995	SNP6_3573
SNP6_3439	SNP6_2797	SNP6_3759	SNP6_3191		SNP6_3384	SNP6_3543
SNP6_3773	SNP6_3060	SNP6_3433	SNP6_3479		SNP6_3237	SNP6_3375
SNP6_3014	SNP6_2910	SNP6_3325	SNP6_3447		SNP6_3422	SNP6_3158
SNP6_3353	SNP6_3447	SNP6_3037	SNP6_3402		SNP6_3207	SNP6_2859
SNP6_3676	SNP6_2862	SNP6_3081	SNP6_3545		SNP6_3080	SNP6_3759
SNP6_3378	SNP6_3375	SNP6_3664	SNP6_3058		SNP6_3321	SNP6_3471
SNP6_3761	SNP6_3063	SNP6_3426	SNP6_2723		SNP6_3444	
SNP6_3482	SNP6_3041	SNP6_3322	SNP6_3770		SNP6_3462	
SNP6_3031	SNP6_3200	SNP6_3282	SNP6_3321		SNP6_3763	
SNP6_3657	SNP6_3767	SNP6_3418			SNP6_3584	
SNP6_3054	SNP6_3778	SNP6_3413			SNP6_3510	
SNP6_3655	SNP6_3434	SNP6_3058			SNP6_2894	
SNP6_2947	SNP6_3509	SNP6_3745			SNP6_3446	
SNP6_2937	SNP6_3546	SNP6_3414			SNP6_2714	
SNP6_3322	SNP6_3438	SNP6_3266			SNP6_3213	
SNP6_3423	SNP6_2875	SNP6_2713			SNP6_3478	
SNP6_3184	SNP6_3103	SNP6_3775			SNP6_3055	
SNP6_3444	SNP6_3306	SNP6_3467			SNP6_2937	
SNP6_3664	SNP6_3292	SNP6_3385			SNP6_3579	
SNP6_3012	SNP6_3384	SNP6_3402			SNP6_3184	
SNP6_3193	SNP6_3308	SNP6_3644			SNP6_3227	
SNP6_3200	SNP6_3051	SNP6_3652			SNP6_3531	
SNP6_3478	SNP6_3465	SNP6_3375			SNP6_3378	
SNP6_3447	SNP6_2724	SNP6_2860			SNP6_3344	
SNP6_3236	SNP6_3476	SNP6_3327			SNP6_3399	
SNP6_3422	SNP6_3293	SNP6_3439			SNP6_3060	
SNP6_3414	SNP6_3439	SNP6_3184			SNP6_3479	

SNP6_3535	SNP6_3000	SNP6_3947	SNP6_3870
SNP6_3385	SNP6_3524	SNP6_3396	
SNP6_3649	SNP6_3307	SNP6_3236	
SNP6_3774	SNP6_3759	SNP6_3446	
SNP6_3331	SNP6_3463	SNP6_3479	
SNP6_3778	SNP6_3442	SNP6_3227	
SNP6_3430	SNP6_2873	SNP6_2937	
SNP6_3466	SNP6_3534	SNP6_3650	
SNP6_3460	SNP6_3309	SNP6_3483	
SNP6_3443	SNP6_2850	SNP6_3406	
SNP6_3429	SNP6_3761	SNP6_3465	
SNP6_3228	SNP6_3396	SNP6_2866	
SNP6_3416	SNP6_3493	SNP6_3478	
SNP6_2910	SNP6_3204		
SNP6_3546	SNP6_3572		
SNP6_3446	SNP6_3012		
SNP6_3763	SNP6_2901		
SNP6_3191	SNP6_2795		
SNP6_3359	SNP6_3037		
	SNP6_3432		
	SNP6_3494		
	SNP6_3446		
	SNP6_2871		
	SNP6_3532		
	SNP6_14754		

Appendix C

Detailed results of DTREG

Table C.1. Important SNPs when the full tag-SNPs set is given to DTREG-Single Decision Tree as an input for replication1

REPLICATION 1	
======= Overall Importance of Variables =======	
Variable	Importance
denseSNP6_3439	100.000
denseSNP6_3437	8.894
denseSNP6_3327	0.438
denseSNP6_3576	0.437
denseSNP6_3759	0.398
-------- Training Data --------	
Sensitivity = 72.66%	
Specificity = 83.34%	
Geometric mean of sensitivity and specificity = 77.82%	
Positive Predictive Value (PPV) = 76.95%	
Negative Predictive Value (NPV) = 79.94%	
Geometric mean of PPV and NPV = 78.43%	
-------- Validation Data --------	
Sensitivity = 71.80%	
Specificity = 83.38%	
Geometric mean of sensitivity and specificity = 77.38%	
Positive Predictive Value (PPV) = 76.78%	
Negative Predictive Value (NPV) = 79.44%	
Geometric mean of PPV and NPV = 78.10%	

Table C.2. Important SNPs when the full tag-SNPs set are given to DTREG-Single Decision Tree as an input for replication2

REPLICATION 2
======= Overall Importance of Variables =======
Variable Importance denseSNP6_3439 100.000 denseSNP6_3437 8.558
-------- Training Data --------
Sensitivity = 71.21% Specificity = 84.37% Geometric mean of sensitivity and specificity = 77.51% Positive Predictive Value (PPV) = 78.00% Negative Predictive Value (NPV) = 79.02% Geometric mean of PPV and NPV = 78.51%
-------- Validation Data --------
Sensitivity = 71.21% Specificity = 84.37% Geometric mean of sensitivity and specificity = 77.51% Positive Predictive Value (PPV) = 78.00% Negative Predictive Value (NPV) = 79.02% Geometric mean of PPV and NPV = 78.51%

Table C.3. Important SNPs when the full tag-SNPs set are given to DTREG-Single Decision Tree as an input for replication3

REPLICATION 3
======== Overall Importance of Variables ========
Variable Importance denseSNP6_3439 100.000 denseSNP6_3437 11.093
-------- Training Data --------
Sensitivity = 71.94% Specificity = 83.01% Geometric mean of sensitivity and specificity = 77.27% Positive Predictive Value (PPV) = 76.79% Negative Predictive Value (NPV) = 79.10% Geometric mean of PPV and NPV = 77.94%
-------- Validation Data --------
Sensitivity = 75.39% Specificity = 83.01% Geometric mean of sensitivity and specificity = 77.27% Positive Predictive Value (PPV) = 76.79% Negative Predictive Value (NPV) = 79.10% Geometric mean of PPV and NPV = 77.94%

Table C.4. Important SNPs when the full tag-SNPs set are given to DTREG-Single Decision Tree as an input for replication4

REPLICATION 4	
======= Overall Importance of Variables =========	
Variable	Importance
denseSNP6_3442	100.000
denseSNP6_3440	19.202
denseSNP6_3430	3.553
denseSNP6_3449	2.786
denseSNP6_3437	1.452
-------- Training Data --------	
Sensitivity = 76.36%	
Specificity = 81.76%	
Geometric mean of sensitivity and specificity = 79.01%	
Positive Predictive Value (PPV) = 76.64%	
Negative Predictive Value (NPV) = 81.53%	
Geometric mean of PPV and NPV = 79.05%	
-------- Validation Data --------	
Sensitivity = 75.39%	
Specificity = 80.20%	
Geometric mean of sensitivity and specificity = 77.76%	
Positive Predictive Value (PPV) = 74.90%	
Negative Predictive Value (NPV) = 80.61%	
Geometric mean of PPV and NPV = 77.71%	

Table C.5. Important SNPs when the full tag-SNPs set are given to DTREG-Single Decision Tree as an input for replication5

REPLICATION 5	
======== Overall Importance of Variables =========	
Variable	Importance
denseSNP6_3437	100.000
denseSNP6_3434	17.278
denseSNP6_3430	3.995
denseSNP6_3947	2.712
denseSNP6_3467	0.654
-------- Training Data --------	
Sensitivity = 74.71%	
Specificity = 82.62%	
Geometric mean of sensitivity and specificity = 78.56%	
Positive Predictive Value (PPV) = 77.14%	
Negative Predictive Value (NPV) = 80.62%	
Geometric mean of PPV and NPV = 78.86%	
-------- Validation Data --------	
Sensitivity = 74.11%	
Specificity = 82.77%	
Geometric mean of sensitivity and specificity = 78.32%	
Positive Predictive Value (PPV) = 77.16%	
Negative Predictive Value (NPV) = 80.28%	
Geometric mean of PPV and NPV = 78.71%	

Table C.6. Important SNPs when the full tag-SNPs set are given to DTREG-Single Decision Tree as an input for replication6

REPLICATION 6	
====== Overall Importance of Variables ==========	
Variable	Importance
denseSNP6_3439	100.000
denseSNP6_3437	19.179
denseSNP6_3479	0.880
denseSNP6_3912	0.707
denseSNP6_3304	0.598
-------- Training Data --------	
Sensitivity = 82.98%	
Specificity = 79.71%	
Geometric mean of sensitivity and specificity = 81.33%	
Positive Predictive Value (PPV) = 75.99%	
Negative Predictive Value (NPV) = 85.82%	
Geometric mean of PPV and NPV = 80.76%	
-------- Validation Data --------	
Sensitivity = 79.11%	
Specificity = 78.47%	
Geometric mean of sensitivity and specificity = 78.79%	
Positive Predictive Value (PPV) = 73.98%	
Negative Predictive Value (NPV) = 82.92%	
Geometric mean of PPV and NPV = 78.32%	

Appendix D

Detailed results of DTREG-Single Decision Tree

Table D.1. Sensitivity values when only the significant SNP combination is given to DTREG-Single Decision Tree as an input for replication1

REPLICATION 1
====== Overall Importance of Variables =========
Variable Importance denseSNP6_3437 100.000 denseSNP6_3430 3.719 denseSNP6_3429 0.604 denseSNP6_3446 0.601 denseSNP6_3440 0.308
-------- Training Data --------
Sensitivity = 81.20% Specificity = 71.85% Geometric mean of sensitivity and specificity = 76.38% Positive Predictive Value (PPV) = 69.28% Negative Predictive Value (NPV) = 83.02% Geometric mean of PPV and NPV = 75.84%
-------- Validation Data --------
Sensitivity = 80.74% Specificity = 71.87% Geometric mean of sensitivity and specificity = 76.18% Positive Predictive Value (PPV) = 69.17% Negative Predictive Value (NPV) = 82.68% Geometric mean of PPV and NPV = 75.63%

Table D.2. Sensitivity values when only the significant SNP combination is given to DTREG-Single Decision Tree as an input for replication2

REPLICATION 2
===== Overall Importance of Variables ==========
Variable Importance denseSNP6_3437 100.000 denseSNP6_3430 4.732
-------- Training Data --------
Sensitivity = 81.14% Specificity = 72.97% Geometric mean of sensitivity and specificity = 76.95% Positive Predictive Value (PPV) = 69.67% Negative Predictive Value (NPV) = 83.49% Geometric mean of PPV and NPV = 76.27%
-------- Validation Data --------
Sensitivity = 81.14% Specificity = 72.97% Geometric mean of sensitivity and specificity = 76.95% Positive Predictive Value (PPV) = 69.67% Negative Predictive Value (NPV) = 83.49% Geometric mean of PPV and NPV = 76.27%

Table D.3. Sensitivity values when only the significant SNP combination is given to DTREG-Single Decision Tree as an input for replication3

REPLICATION 3	
====== Overall Importance of Variables ==========	
Variable	Importance
denseSNP6_3437	100.000
denseSNP6_3430	3.719
denseSNP6_3429	0.604
denseSNP6_3446	0.601
denseSNP6_3440	0.308
-------- Training Data --------	
Sensitivity = 81.20%	
Specificity = 71.85%	
Geometric mean of sensitivity and specificity = 76.38%	
Positive Predictive Value (PPV) = 69.28%	
Negative Predictive Value (NPV) = 83.02%	
Geometric mean of PPV and NPV = 75.84%	
-------- Validation Data --------	
Sensitivity = 80.74%	
Specificity = 71.87%	
Geometric mean of sensitivity and specificity = 76.18%	
Positive Predictive Value (PPV) = 69.17%	
Negative Predictive Value (NPV) = 82.68%	
Geometric mean of PPV and NPV = 75.63%	

Table D.4. Sensitivity values when only the significant SNP combination is given to DTREG-Single Decision Tree as an input for replication4

REPLICATION 4	
====== Overall Importance of Variables ========	
Variable	Importance
denseSNP6_3437	100.000
denseSNP6_3434	18.147
denseSNP6_3430	4.705
denseSNP6_3446	0.987
-------- Training Data --------	
Sensitivity = 72.20%	
Specificity = 83.39%	
Geometric mean of sensitivity and specificity = 77.60%	
Positive Predictive Value (PPV) = 77.31%	
Negative Predictive Value (NPV) = 79.29%	
Geometric mean of PPV and NPV = 78.29%	
-------- Validation Data --------	
Sensitivity = 72.20%	
Specificity = 83.39%	
Geometric mean of sensitivity and specificity = 77.60%	
Positive Predictive Value (PPV) = 77.31%	
Negative Predictive Value (NPV) = 79.29%	
Geometric mean of PPV and NPV = 78.29%	

Table D.5. Sensitivity values when only the significant SNP combination is given to DTREG-Single Decision Tree as an input for replication5

REPLICATION 5
======= Overall Importance of Variables ========
Variable Importance denseSNP6_3437 100.000 denseSNP6_3434 17.278 denseSNP6_3430 3.995
-------- Training Data --------
Sensitivity = 70.73% Specificity = 83.87% Geometric mean of sensitivity and specificity = 77.02% Positive Predictive Value (PPV) = 77.49% Negative Predictive Value (NPV) = 78.49% Geometric mean of PPV and NPV = 77.99%
-------- Validation Data --------
Sensitivity = 70.73% Specificity = 83.87% Geometric mean of sensitivity and specificity = 77.02% Positive Predictive Value (PPV) = 77.49% Negative Predictive Value (NPV) = 78.49% Geometric mean of PPV and NPV = 77.99%

Table D.6. Sensitivity values when only the significant SNP combination is given to DTREG-Single Decision Tree as an input for replication6

REPLICATION 6	
==== Overall Importance of Variables ==========	
Variable	Importance
denseSNP6_3437	100.000
denseSNP6_3434	17.143
denseSNP6_3430	5.001
denseSNP6_3446	0.315
-------- Training Data --------	
Sensitivity = 72.75%	
Specificity = 83.37%	
Geometric mean of sensitivity and specificity = 77.88%	
Positive Predictive Value (PPV) = 77.20%	
Negative Predictive Value (NPV) = 79.81%	
Geometric mean of PPV and NPV = 78.49%	
-------- Validation Data --------	
Sensitivity = 72.66%	
Specificity = 83.35%	
Geometric mean of sensitivity and specificity = 77.82%	
Positive Predictive Value (PPV) = 77.15%	
Negative Predictive Value (NPV) = 79.76%	
Geometric mean of PPV and NPV = 78.44%	