Fiaschi, Linda (2011) Novel guidelines for the analysis of single nucleotide polymorphisms in disease association studies. PhD thesis, University of Nottingham.

Novel Guidelines for the Analysis of Single Nucleotide Polymorphisms in Disease Association Studies

By
Linda Fiaschi, BSc, MSc

Thesis submitted to The University of Nottingham
for the Degree of Doctor of Philosophy

School of Computer Science
The University of Nottingham
Nottingham, United Kingdom

February 2011

To my family and my Richard,

with love and pride.

Novel Guidelines for Single Nucleotide Polymorphisms Analysis in Disease Association Studies

Linda Fiaschi

Submitted for the degree of Doctor of Philosophy
February 2011

Abstract

How genetic mutations such as Single Nucleotide Polymorphisms (SNPs) affect the risk of contracting a specific disease is still an open question for numerous different medical conditions. Two problems related to SNPs analysis are (i) the selection of computational techniques to discover possible single and multiple SNP associations; and (ii) the size of the latest datasets, which may contain millions of SNPs.

In order to find associations between SNPs and diseases, two popular techniques are investigated and enhanced. Firstly, the 'Transmission Disequilibrium Test' for family-based analysis is considered. The fixed length of haplotypes provided by this approach represents a possible limit to the quality of the obtained results. For this reason, an adaptation is proposed to select the minimum number of SNPs that are responsible for disease predisposition. Secondly, decision tree algorithms for case-control analysis in situations of unrelated individuals are considered. The application of a single tool may lead to limited analysis of the genetic association to a specific condition. Thus, a novel consensus approach is proposed exploiting the strengths of three different algorithms, ADTree, C4.5 and Id3. Results obtained suggest the new approach achieves improved performance.

The recent explosive growth in size of current SNPs databases has highlighted limitations in current techniques. An example is 'Linkage Disequilibrium' which identifies redundancy in multiple SNPs. Despite the high accuracies obtained by this method, it exhibits poor scalability for large datasets, which severely impacts on its performance. Therefore, a new fast scalable tool based on 'Linkage Disequilibrium' is developed to reduce the size through the measurement and elimination of redundancy between SNPs included in the initial dataset. Experimental evidence validates the potentially improved performance of the new method.

Declaration

The work in this thesis is based on research carried out at the *Intelligent Modelling and Analysis* (and, originally, the *Automated Scheduling, Optimisation and Planning*) Research Group, the School of Computer Science, the University of Nottingham, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I would like to thank my supervisor Dr. Jonathan Garibaldi for giving me the opportunity to study for this PhD and for his valuable advice and guidance during my work.

I am grateful also to Dr. Linda Morgan and the Bioinformatics meetings group from Clinical Chemistry Division, Institute of Genetics at Queen's Medical Center, Nottingham, for their precious support and their meaningful discussions in the field of genetics.

My sincere gratitude goes also to all my friends and colleagues from Computer Science for their friendship and support during these years of study. A special thanks to Dr. Andrea Sackmann from Poznan University in Poland, for her cooperation and availability to discuss and share new ideas.

I would like to express my deep gratitude to my family in Italy for their continuous support, understanding and encouragement every time I needed during the development of this work. Finally a special thought goes to my little Richard who became the major inspiration of my journey.

This work was supported by Marie Curie Action MEST-FP6-2004-7597 under the Sixth Framework Program of the European Community.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

In April 2003, the Human Genome Project (HGP), a 13-year project coordinated by the U.S. Department of Energy and the National Institute of Health was completed.

During the early years of the HGP, the Wellcome Trust (U.K.) became a major partner together with additional contributions coming from Japan, France, Germany, China, among others. The main goals of this project were to identify all the genes in human DNA, determine the sequences of the chemical base pairs that constitute human DNA, store this information in databases and improve tools for data analysis. Following the scientific milestones of enormous proportions achieved by this world class research, many years will be spent in the analysis of these data, fostering the development of more competitive biotechnology systems and novel analysis tools for new medical applications.

Under this scenario, Bioinformatics, the application of information technology to the field of molecular biology, is established as one of the most challenging disciplines in medical studies. Since the late 1980s, Bioinformatics has focused on genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing. Nowadays its role concerns the creation and maintenance of databases, development of algorithms, computational and statistical techniques, and theories to address formal and practical problems arising from the management and analysis of biological data. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, genome-wide association studies and the modelling of evolution. The development and implementation of tools that enable efficient management and access to various types of information is therefore strictly related to this discipline. Along with these issues, a relevant point is the development of new algorithms and statistics. Such systems have the ability to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structures and functions, and cluster protein sequences into families of related sequences.

In general terms, the completion of human genome sequencing has opened up a long list of challenging problems to be resolved by the research community. Specifically, one challenge is how genetic mutations are responsible for a specific disease. The biological data must be combined to form a comprehensive picture of cellular activities in order to understand how a normal biological function can be altered in different disease states. Therefore, in this specific application, the most pressing task for Bioinformatics now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains and protein structures. The actual process of analysing and interpreting data is referred to as Computational Biology.

Undoubtedly one of the most important threats to human health are genetic diseases. Such diseases are defined as a disorder caused by genetic factors, in particular abnormalities along the DNA chain. There are numerous different types of genetic disorders. Some of these changes in genome can cause specific advantages in certain environments (Darwinian Fitness) but many create abnormalities that result in destructive effects for a living being. In a single gene disorder the starting point is a mutation/change in one gene. Genes encode the proteins which are some of the most important functional components in living beings and which play a role in the structure of cells. The results of a mutation which occurs in the functional part of a gene that encodes a protein can create relevant problems. The protein is no longer functional and can result in severe consequences for the individual. Almost 6000 single gene disorders are known currently and it is estimated that 1 in 200 newborns face a genetic disorder related to a single gene. These disorders include sickle cell anaemia, cystic fibrosis, Aicardi Syndrome and Huntington's disease. Another class of human genetic diseases are caused by mutations in more than one gene. The appearance of a disease can also be shown by environmental factors which combine with mutations. It is clear that polygenic disorders are much more complicated than the single gene disorders. These abnormalities are also difficult to analyse as there are numerous factors that researchers should take into consideration in order to reach useful conclusions. Many well known chronic diseases are in fact multi-factorial genetic diseases including Alzheimer's disease, diabetes, obesity, arthritis and numerous types of cancer.

The issue of genetic disorders can be viewed in two different ways. In embryonic development and in infancy the alteration of the DNA chain can cause congenital malformations. Alternatively, different mutations can affect the individual's susceptibility to contracting a disease. Some information contained in the DNA chain may not result in an evident disease but just in the predisposition to be more likely than the rest of the population to develop a disorder, which may occur later in life.

The genetic alterations responsible for human disorders can be cataloged under different rules of classification. In this context the occurrence frequency is the key rule of partitioning. In particular, the specific genetic marker under analysis in this Thesis is the Single Nucleotide Polymorphism (SNP). This is a genetic change of a single nucleotide along the DNA chain that happens once every 100 to 300 bases. This type of mutation is already known to be responsible for an increased risk of contracting various diseases.

In this research, numerous considerations are illustrated which may be applied to a wide generic context. The new methodologies proposed are originally developed for and repeatedly tested on datasets related to the specific disease of Pre-eclampsia. This is a condition in pregnancy and in the post-partum period which affects both mothers and their babies. It manifests itself with different degrees of severity which can lead to different consequences for the mother and babies. While a percentage of cases can result in a stillbirth, one of the most common results is a pre-term delivery with underweight babies. Babies born in this manner are predisposed to experience various disorders later in life.

Currently there are numerous tools used to approach genetic data analysis. This work is focused on two of the main problems affecting the related research community in the effort to extract as much information as possible from human DNA.

The first problem is related to the presence of the so called 'linkage disequilibrium' between genetic markers. This hidden relationship is responsible for the co-inheritance of genetic markers and phenotypes in families over several generations. In simpler terms, this phenomenon explains (for instance) how the presence of a specific phenotype such as colour of the skin can be linked to other different phenotypes such as the shape of eyes, body size, or hair type and colour. Unfortunately, a proportion of the genetic information involved in this complex linkage is not clearly evident in discernible phenotypes. In some cases, a noticeable feature is linked to a hidden condition attributed to the linkage between their genetic markers. In other cases, the two linked markers do not reflect any evident feature but are still responsible for a condition or for a disease predisposition.

Researchers in the field of genetics are focused on resolving this issue through detecting groups of genetic markers which are linked together. Determining these would help to decrease the size of the datasets under analysis as lots of genetic information could be predicted by a single marker which would link them together. Under this scenario many solutions have been proposed in the literature, covering a large number of different hypotheses. One of the most common tools used for such purposes is the 'Linkage Disequilibrium' (LD) function, which has been implemented in different software languages. This function is based on the measure of the correlation coefficient between different pairs of SNPs. The typical genetic dataset is represented by a matrix composed by the attributes in columns and patients in rows. The attributes in turn consist of genetic information such as SNPs and other medical information such as clinical variables. The latest achievements in the realisation of SNP-chip technologies highlight the large amount of genetic markers that can now be analysed in a single dataset. This has reached a size of more than 1.2 million SNPs [5]. Considering a matrix composed of this many attributes, the application of the current LD function requires the measurement of the square correlation coefficient for every possible combination of two SNPs, taken from a set of 1.2 million. This results in a number of operations exceeding 10^{12}. Such computational complexity cannot be faced with most of the current machines used for these purposes. These limitations represent the starting point for the development of this work. A new solution is proposed to overcome this problem, based upon the current LD technique. An improved use of this tool is therefore presented and discussed. Several experiments have been carried out and discussed in this Thesis in order to motivate the final structure of the novel methodology proposed for redundancy reduction.

Conversely, the second important problem faced by data-miners in genetic analysis is the association between genetic markers and diseases. This time the link is not between DNA molecules but rather between genome mutations and diseases. Attention is focused on selecting a small number of SNPs from the initial dataset which can be proved to be responsible either for the presence of a disease or for an increased risk of disease. Also for these issues many solutions can be found in the literature, spanning the most common scenarios. In this research, two of these techniques are examined to identify possible improvements. These approaches are shown in two separate research studies applied to a relatively small sized dataset. Each of these studies is based upon a different method of analysis.

The first one is a family-based approach, called the ‘Transmission Disequilibrium Test’, which requires the provision of an accordingly structured database. One of the possible solutions, proposed in the literature for the implementation of this technique, is software called TRANSMIT. The input data consists of a population of families, containing the offspring together with the parents. In this work, experimental research is carried out upon the application of this tool to a relatively small dataset in order to highlight the pitfalls of the method. The string of SNPs analysed by this method is composed of a fixed number genetic markers. However the actual number of SNPs directly involved in an increase of disease risk may result in a subset of this haplotype. This represents a limitation for the current solution as irrelevant SNPs could be brought forward for further analysis, affecting the performance and the quality of results. This pitfall has provided motivation for the subsequent development of a new approach for the TRANSMIT software. This solution is discussed and the improvements in its performance validated through experimental results.

The second method of analysis is the case-control study and is applied to a population of independent individuals. The Decision Tree technique is a method used commonly by the genetic community and therefore is investigated in this Thesis. In particular, attention is focused on three different algorithms commonly used for analysis of this nature. As per the previous technique, the advantages and disadvantages of these solutions are highlighted and discussed in detail in an attempt to provide researchers with a novel improved proposal for future association studies. This is realised through a combined analysis of the three algorithms under scrutiny to overcome the limitations that each single technique may yield. Three different steps form the structure of the proposed approach: the pre-processing of the data, the statistical assessment of the results and the results analysis. The pre-processing stage covers major issues such as missing values, balancing of the dataset and choice of the predictable variable. The proper analysis is carried out by a comparison of three different solutions in the decision tree range, ADTree, ID3 and C4.5. A statistical analysis on the significance of the results obtained from these three different options is the relevant tool for making the final decision. In particular, the experiments performed in this context are based on a continuous variable as the outcome choice. This requires a further detailed analysis of different options for converting the outcome variable to a Boolean variable by setting different thresholds.

1.2 Aims of this Thesis

Different methods are available for the analysis of SNPs data associated with the probability to contract a disease. Each one is based on different strategies, algorithms or data structures. Nevertheless, there are still numerous limitations and pitfalls that every technique presents in a specific angle of the problem.

The first key aim of this work is the effective disease association analysis performed with the Transmission Disequilibrium Test(TDT) for family based studies and the Decision Tree approach for case-control analysis. Different experiments are performed with these two techniques to highlight their drawbacks, together with an analysis of their current state of art. The new idea which inspired the realisation of the second component of this research is based on the exploitation of the positive aspects of two different techniques which are usually employed for this type of analysis. These approaches are based on different statistical methods and databases features. The TRANSMIT software, employed for a TDT analysis can only work with a fixed length of haplotype, affecting the quality of the final results. Alternatively, the single application of a decision tree algorithm provides a limited analysis of a more complex and articulated problem. These hypotheses have arisen the motivation for realising possible solutions able to overcome the specified limitations, leading to the realisation for both methods of an alternative way of use, finalised to optimise their performance. Both the solutions proposed for optimisation of the current techniques are based in the assumption that duplicate have been SNPs removed beforehand.

Current techniques to detect and remove redundancies, which achieve a good accuracy of results, are both too slow and have an high computational complexity. Thus these techniques are not widely usable or affordable. Hence, a problem exists when different methods are combined, affecting the performance of the analysis due to the large size of the available data. Thus, there is a need to reduce the size of databases, which is continuously increasing as a result of improved technology for genetic applications. In this manner, it is possible to contain the computational complexity of the analysis required for current state of the art research. Reduction of the size of a large dataset can be achieved by the elimination of attributes of the initial dataset, representing the columns in an input matrix. In this specific study, these attributes consist of SNPs data. Irrelevant SNPs need to be selected and deleted from the dataset as this can impede the final results. Under this hypothesis, the second aim of this work is to improve the currently used techniques for detecting redundant SNPs (LD function) in order to select a small number of SNPs that can represent the rest of the original set, with a high accuracy and low computational complexity.

In conclusion the following objectives of the thesis were identified:

(1) Improving the existing software for disease association study (for family and population based datasets) in order to rapidly identify the relevant SNPs, after the elimination of redundant SNPs.

(2) Developing a new fast scalable tool able to approximate the LD function outcome.

(3) Combining the two previous points, realisation of novel guidelines for genetic data analysis, able to fulfil the current demanding requirements of increasing database sizes.

1.3 Organisation of the Thesis

Following this Introduction, Chapter 2 covers a wide range of the problems faced in this area of research. In order to give a basic support to the readers for a better understanding of the subjects presented, a number of definitions and notions are reported from the literature in the medical, genetic and computer science fields. Furthermore, the different methods of analysis commonly used by the research community are also analysed and discussed in this Chapter. This is presented to build the foundations for a clearer discussion of the novel methodologies that are proposed. This includes up-to-date tools, strategies and software which have been recently published in the analysis of SNPs. Together with the scientific aspects of the problem, a part of this Chapter addresses the state of the art in knowledge regarding Pre-eclampsia which is the specific disease under examination in this thesis. This is performed in order to give an overview of the actual achievements relevant in the field.

Within disease association studies, the first method investigated is the family-base analysis Transmission Disequilibrium Test (TDT), illustrated in Chapter 3. A general overview of the method is followed by a description of the algorithm used for its implementation, namely the TRANSMIT software. The original version of this tool is employed in a set of experiments that highlight the pitfalls. An advanced version of this technique is proposed and discussed through further experiments showing an improvement in the results achieved. Whereas the original software analyses a fixed length of haplotype, the new approach performs a separate test for different sequences of SNPs taken from the original set for different sizes. Due to scalability constraints, this method is supposed to be applied following an elimination of the SNPs not relevant for the analysis.

The second method examined for association studies is based on the case-control analysis as presented and discussed in Chapter 4. A framework methodology for SNP data mining is illustrated and is centred around decision tree algorithms. A complete description is given of the main steps that the research process is expected to follow within the exploration of a genetic dataset. Several experiments are also performed in order to highlight strengths and limitations of this proposed technique.

The possibility to access a large variety of data is a vital component for any research actualisation. As this is a main issue relevant to the scope of this Thesis, Chapter 5 is dedicated to the creation of an application for generating artificial SNPs datasets for decision tree analysis. This generic medical dataset is examined from a number of perspectives and attention is specifically addressed to the genetic information of SNPs attributes. Different parameters can be set for this application such as the number of patients in the database, probability of occurrence of a certain allele, or a couple of alleles, and the amount of SNPs under analysis. Additionally, the probability of contracting a disease for each genetic marker present in the dataset or disease model can also be set in different ways. Several experiments are performed in which the three decision tree algorithms studied in this work (namely ADTree, ID3 and C4.5), are applied to identify their strengths and limitations in this context.

Another major issue that is examined in this work is the detection of linkage disequilibrium (LD) between SNPs. A detailed overview of the problem is given in Chapter 6. The new idea described for improving the performance of the original LD function is focused on the reduction of the time and computational complexity that this kind of task usually requires. Several experiments with real and artificial datasets have been carried out with the aim of showing how different settings of the involved parameters provide different qualities of results. In particular, experimental evidence demonstrates how improved results can be obtained by setting the parameters to their best value and through choosing the most functional techniques for clustering purposes.

The new methodology proposed in this Thesis is described and discussed in Chapter 7. This is a technique built for providing researchers with a new tool helpful in the difficult task of large SNPs databases analysis. As the source is genetic information, the problem is decomposed into two different issues, namely the linkage disequilibrium detection between markers and the disease risk association with SNPs. The first aspect, expounded as redundancy elimination is resolved with the proposal of the novel technique, called RD-snp, for redundancy detection in SNPs datasets. The second problem is, in turn, divided in two different contexts, dependent on the dataset format: the family based analysis and the case control study. For both of these options a novel solution is proposed. This analytical process is applied to the specific medical application of risk disease association with a explicit reference to Pre-eclampsia disease.

The conclusive Chapter 8 completes the overview of this research. Discussion on the effectiveness and any possible improvement that this work can bring to this field of study is presented. Further directions are also included which can be considered for the stimulation of future research. Subsequent resulting aspects are highlighted which may be interesting to pursue using extensive and additional analysis. These directions are presented and discussed.

Chapter 2

Literature Review

2.1 Introduction

In this introductory chapter an overview of several concepts from different backgrounds is given as, according to the scope of the project, this thesis is the result of multidisciplinary work.

This Chapter is divided in four main sections. In the first section an overview of the genetic and medical background is given in order to introduce the state of the art on the disease Pre-eclampsia and highlight the reasons why further research is needed in this field. The second section shows the two main methods of analysis for disease association study, case-control and family based analysis. The current tools available in the literature, which are used for these purposes are discussed, together with a critical assessment of them, in order to highlight their weaknesses and hence to provide the motivations for the improvements proposed in this work. The third section provides an overview of the pressing problem of SNPs dataset size reduction. The various feature selection solutions that can be found in the literature for the specific field of SNPs reduction are shown and their pitfalls are highlighted. The motivation for the new proposed technique is therefore provided. The final section gives an overview of the software tools used in this work.

2.2 Glossary

In this section the relevant technical terms are listed with their respective meanings. All uses of these words in the rest of Thesis are thus referred to these definitions [6].

- *DNA*: double-stranded macromolecule consisting of two chains running in opposite directions and called deoxyribonucleic acid.
- *Gene*: small units of the DNA chain.
- *Chromosome*: genes units composed by pairs set of nucleotides.
- *Nucleotide*: the basic building block of nucleic acids, such as DNA and RNA.
- *Somatic cells*: all body cells of an organism, apart from the sperm and egg cells.
- *Germ cell*: the reproductive cells (sperm and egg cells) in multicellular organisms.
- *Allele*: one member of a pair of genes occupying a specific spot on a chromosome that controls the same trait.
- *Locus*: the location of a gene (or of a significant sequence) on a chromosome, as in genetic locus.
- *Phase*: the information that is needed to determine the two haplotypes that underlie a multi-locus genotype within a chromosomal segment.
- *Trait*: an attribute of phenotype.
- *Phenotype*: a physical appearance or biochemical characteristic of an organism as a result of the interaction of its genotype and the environment.
- *Haplotype*: the set of alleles on one chromosome.
- *Homozygous*: having two identical alleles that code for the same trait.
- *Heterozygous*: having dissimilar alleles that code for the same trait.
- *Diploid*: a cell or an organism consisting of two sets of chromosomes.

- *Haploid*: a cell or an organism having half of the number of chromosomes in somatic cells.

- *Gamete*: a reproductive cell or sex cell that contains the haploid set of chromosomes.

- *Autosome*: any chromosome not considered as a sex chromosome, or not involved in sex determination.

- *Single Nucleotide Polymorphism (SNP)*: a mutation of a single base of DNA.

- *Admixture*: when two or more subpopulations inbreed, so that two randomly chosen individuals in the population might have different degrees of genetic heritage from the original subpopulations.

- *Population Stratification*: the presence in a population of distinct strata or groups that show limited inbreeding.

- *Linkage Disequilibrium*: the occurrence of some genes together, more often than would be expected by chance.

- *Association*: the occurrence together of two or more phenotypic characteristics more often than would be expected by chance.

- *Odds Ratio*: measure of effect size, describing the strength of association or non-independence between two binary data values.

- *Type I error*: when a null hypothesis is incorrectly rejected when it is in fact true (also known as a false positive).

- *Type II error*: when a null hypothesis is not rejected despite being false (also known as a false negative).

2.3 Genetic and Medical Background

Diseases and their possible association with the human genome are becoming one of the most attractive topics for Medicine and Bioinformatics. Lots of research findings have highlighted the link between these two aspects but still lot of work needs to be done for many of the both common and rare diseases.

2.3.1 Diseases and Genetics

The principles of inheritance, described by Mendel's work in 1900 have been subsequently analyzed by genetic research groups which confirmed their universal significance among plants and animals. Because of the discovery that genetic mechanisms were the same in most organisms, at the beginning geneticists were particularly focused on little animals such as mice and flies, due to their short life cycles, huge numbers of offspring, easy genetic analysis and for being easily growing species. Their mutations were carefully analysed, characterised and mapped, making these specimens become the so-called 'model organisms' as used for studies of basic biological processes. Gradually scientists added other species to their collection of organisms, such as viruses and microorganisms, then plants and more complex animals. The subsequent development of DNA technology and genome sequencing results confirmed that all life has common origin. In other terms, genes from different organisms but with similar functions are very similar in structure and DNA sequences. Thus, studying the genetic information of simple species brings a good understanding of the human genome and a possible solution for developing new drugs for treating human diseases.

The excitement and explosion of information generated by the genetic discoveries from the beginning of the twentieth century up to the present has no competitive comparison with any other scientific discipline, as the long list of Nobel Prizes confirms. In the 'Age of Genetics' all aspect of modern life are affected by the continuous discovery from human DNA studies in medicine, agriculture, biotechnology, law, pharmaceutical industry etc. Diagnosis and prediction of the course of a disease, detection of genetic defects 'in utero', disease resistant plants, more productive animals, paternity testing and murder investigations are a few examples of applications.

Human genetic disorders are commonly associated with mutations that alter the codification of a gene, the number of chromosomes or the protein structures which affect the delicate balance of gene expression. However, gene expression is a more complex process which involves a number of sequential steps. Mutations, which may affect all the steps,

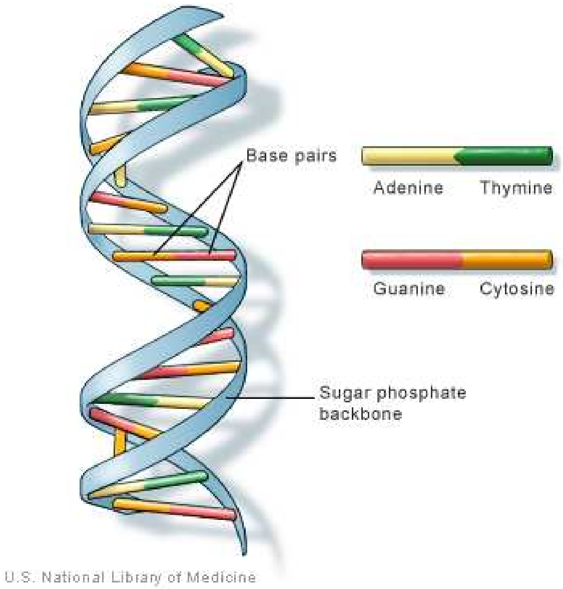


Figure 2.1: Internal structure of the DNA chain. Credit: U.S. National Library of Medicine

can result in genetic disorders. The work with model organisms thus has proven very useful in revealing how genetic factor contribute to the phenotypes of complex diseases [7].

2.3.2 The Human Genome

All the biological information needed to build and maintain a living example of an organism is enclosed in a double-stranded macromolecule consisting of two chains running in opposite directions, called deoxyribonucleic acid (DNA) [8]. DNA essentially encodes a sequence of four types of nucleotides or nitrogenous bases, abbreviated as A (Adenine), G (Guanine), T (Thymine) and C (Cytosine) [9]. These four bases in their different combinations specify most of the amino acid sequences of proteins.

The long chain of DNA is divided into smaller units better known as ‘genes’. According to the official Guidelines for Human Gene Nomenclature, a gene is defined as ‘a DNA segment that contributes to phenotype/function: in the absence of demonstrated function a gene may be characterized by sequence, transcription or homology’. In simpler terms, these are the basic biological units of heredity, a segment of nucleotides needed to contribute to a function (Figure 2.2).

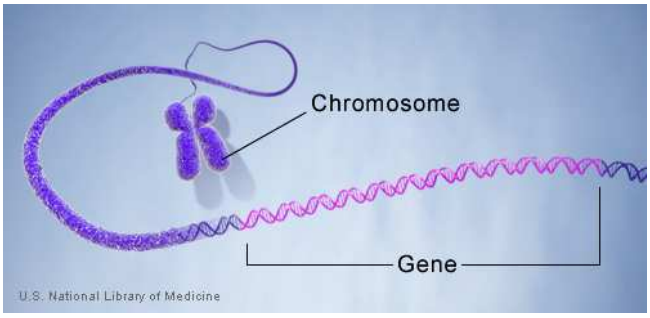


Figure 2.2: Gene definition. Credit: U.S. National Library of Medicine

In April 2003, the sequencing of the human genome was finished but the exact number of genes encoded by the genome is still unknown. Different estimations have been available from different institutions, research groups and studies, at different points in time. They range from the first one amounting to 1000,000 revealed in October 1996 [10] down to the most recent limited to 20,000-25,000 genes estimated by The International Human Genome Sequencing Consortium, led in the United States by the National Human Genome Research Institute (NHGRI) and the Department of Energy (DOE) in October 2004 [11]. It will still take some time before the real gene count will be discovered. All the uncertainty and variety about the predicted number is due to different computational methods and gene-finding programs. Some of them count genes from detection of their beginning and ending, whereas other distinguish them by sequence comparison between new and known segments. If the former tends to overestimate the gene number by counting also segments that look like genes, the latter tends to underestimate the number as the count is limited to a comparison of only known sequences. Beside this, small genes are difficult to estimate, there are genes that code for several proteins, some only for RNA, some overlap and so on. Before the real and final answer, intensive and extensive laboratory work will have to be carried out by the scientific community [12, 13].

Genes are arranged in precise units, each one composed by pairs set of nucleotides: the chromosomes, see Figure 2.3. One chromosome in each pair comes from the mother and the other from the father. The chromosomes in any particular pair look like each other, except in the male gender. There is one pair of chromosomes, which indeed codes for the sex of the individual. This pair has two X chromosomes in females whereas one X and one Y chromosome in males.

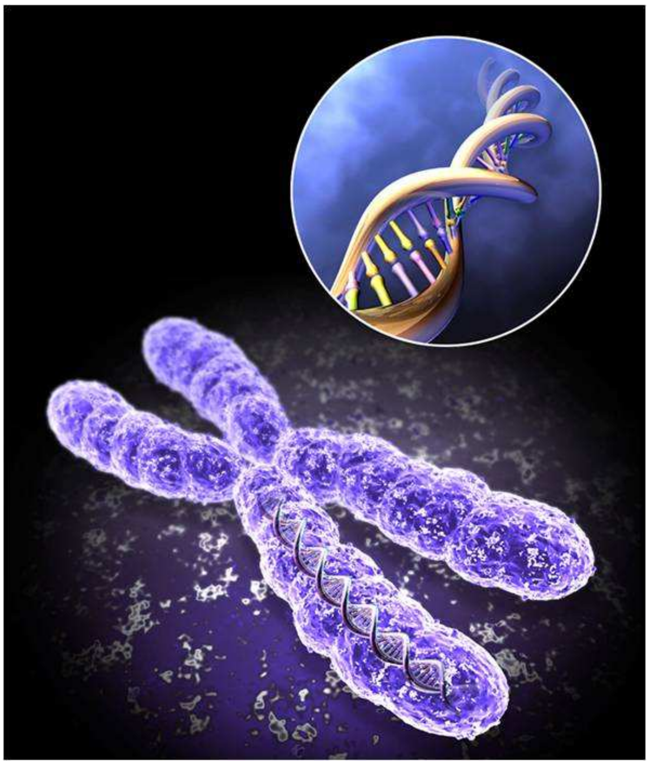


Figure 2.3: Chromosomes. Credit: Jeff Johnson, Biological and medical visuals.

The whole human genome is divided in a certain number of genes which are grouped together in 23 different pairs of chromosomes and the position of the gene within the chromosome is called the locus (Figure 2.4). One member of a pair of genes occupying a specific spot on a chromosome that controls the same trait is called an allele. Even if there are more than two type of alleles in a population, in an individual only two allele can be present and they have the same probability to be inherited (except for special conditions), see Figure 2.5.

Most of the cells have a set of two homologous chromosomes and for this they are called diploid. Only the sexual cells, gametes, have only a single set of chromosomes and are called haploid. Each trait is affected by the two allele inherited one from each parent and that exhibit a feature dominant, co-dominant or recessive. If both allele are the same, the gene is called homozygous, whereas if they are different the gene is said to be heterozygous. Sometimes an allele overcomes the effect of others in affecting the traits, sometimes it depends on the gene (in homozygous or heterozygous) and other time traits are combinations of alleles from different genes.

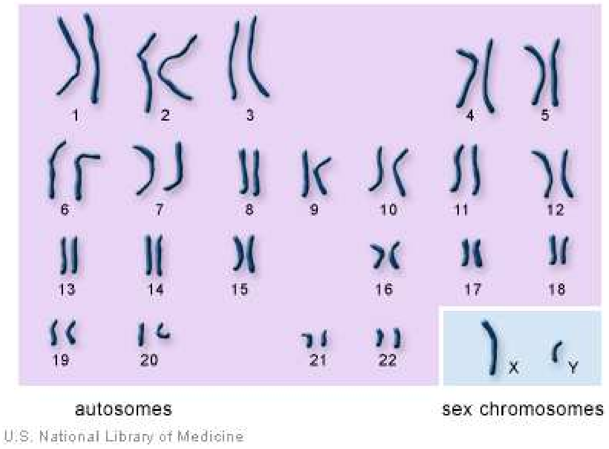


Figure 2.4: Human Chromosomes. Credit: U.S. National Library of Medicine

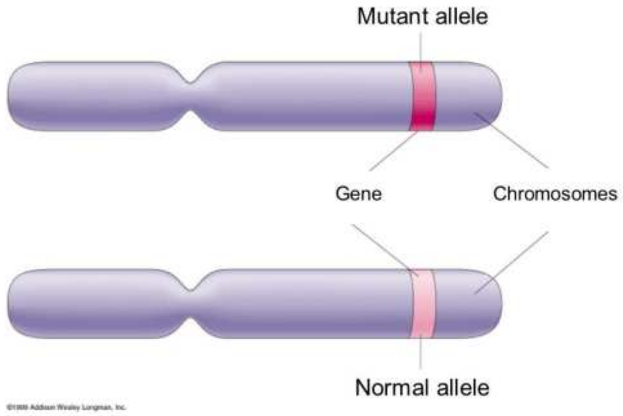


Figure 2.5: Allele definition. Credit: 1999 Addison Wesley Longman, Inc.

Human DNA is estimated to comprise around 3 billion base-pairs, of which around 99.9% are the same — there is only a small percentage that makes the difference between individuals [7, 14].

2.3.3 Genetic Mutations

In 2009, Klug et al. provided an extensive overview of genetic concepts from which the relevant topics related to this work can be summarised in this section, [7]. The essence of genetic function is the storage, replication and transmission of the DNA macromolecules. However, the capacity of DNA to make mistakes is an equally important factor. Changes in DNA are responsible for different phenotypes, adaptation and environmental diversity

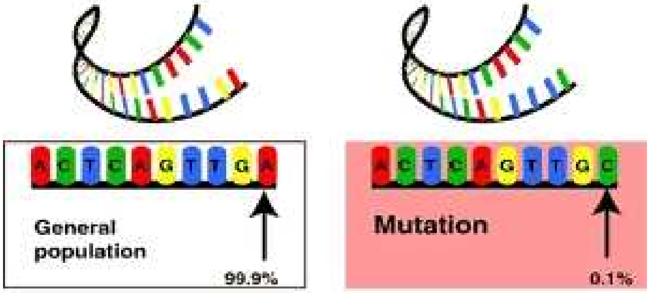


Figure 2.6: Example of mutation. Credit: [1]

and, most of all, evolution. On the other hand, they are also a cause of cell death, genetic diseases and cancer. Mutations also act as identifying ‘markers’ for genes, making feasible their tracing from parents to offspring and providing therefore the basis for genetic analysis. There can be mutations in large regions of chromosomes, called chromosomal mutations, which are different from the gene mutations which occur in the base-pair sequences of DNA within single genes.

A gene mutation is defined as an alteration of DNA sequence which consists of any base-pair change in any part of the DNA molecule. This includes single-base-pair substitutions, deletion or insertion of one or more base pairs up to major alteration in the chromosome structure. An example is shown in Figure 2.6. Mutations can occur within or without regions of gene that code for proteins. They may or may not affect phenotypes, depending on where they occur and to which degree. If they occur in somatic cells, they may lead to cellular dysfunctions or tumours but they are not transmitted through generations. Whereas if they occur in germ cells, they are heritable, causing genetic diversity and evolution.

Different classifications exist based on different effects of mutations. For instance they can be spontaneous or induced. In the former, no cause is known about their presence and therefore they are assumed accidental. In the latter, instead they are affected by extraneous factors, natural or artificial, such as radiation or chemical agents. The spontaneous mutations are low in muber, and vary between organisms and genes. A lively debate is still intriguing researchers on the possibility of adaptive mutations. The capability for an organism to induce a set of mutations as result of environmental pressure is an open and controversial question.

An alternative classification is based on the location of the mutation. The ones occurring in the gametes are called germ-line mutations and are transmitted to the offspring. Whereas those located in the other cells are called somatic mutations are not transmitted to future generations. Then there are Autosomal and X-linked mutations which occur respectively in autosomes and X-chromosome. Depending on the type of location, the mutation can bring different phenotypic defects with different degree. There are loss-of-function mutations which reduce or eliminate the function of a gene product. On the contrary, the gain-of-function mutations result in a gene product with new or enhanced function. The neutral mutations don't affect gene products because they occur in the part of the genome that do not contain genes. The visible ones are recognized by different phenotype and some other examples are the nutritional, behavioural, regulatory, conditional and lethal ones.

Within the several types of classifications, finally there is one based on type of molecular change. If one base pair changes in the DNA molecule, this is called point mutation or base substitution and includes replacement, insertion or deletion of one or more nucleotides. Within the single base change, the Single Nucleotide Polymorphisms, better known as SNPs are one of the most commonly studied kind of mutation by geneticists [7].

2.3.4 Single Nucleotide Polymorphisms (SNPs)

A point mutation or SNP involves the substitution of a base during the replication process. As the enzyme DNA polymerase cuts down one side of a DNA molecule, forming base pairs to build a new complementary strand, it occasionally adds the wrong base. However, DNA polymerase makes very few errors and it corrects most of these quickly. There is a further check, performed by other enzymes, in order to make sure that the new nucleotides are actually complementary to the template strand. Any misfits are then detected and replaced with the proper base. This impressive procedure guarantees a replication of the DNA with less than one mistake per billion nucleotides.

Nonetheless, this type of mutation does occur and it is responsible for the many subtle and not so subtle variations found within and among species. In terms of number, while at most positions in human DNA the same base is found, approximately once every 100 to 300 bases a wrong base may be found, constituting a SNP (Figure 2.7).

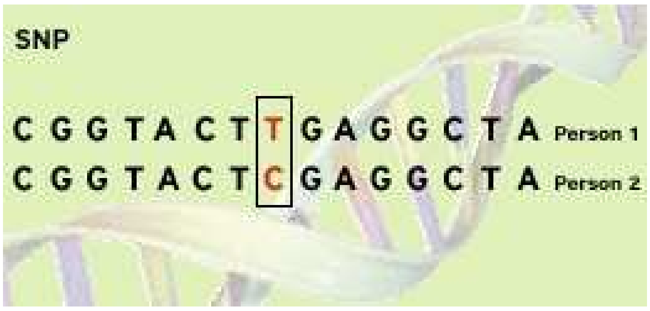

SNP
C G G T A C T T G A G G C T A Person 1
C G G T A C T C G A G G C T A Person 2

Figure 2.7: SNP Definition. Credit: [2]

Figure 2.8: Pre-eclampsia - evident effects of inhibited growth in a baby. Credit: [1]

The majority of these changes have no effect or at least effects which are not yet known, but others can cause subtle differences in physical or psychological characteristics. Some of them may actually affect a person's response to drug therapy and even confer a personal susceptibility or resistance to a certain disease, determining then the severity or progression of it. For this reason, analysis of SNPs has become the subject of extensive research [15–17].

2.3.5 Pre-eclampsia

Within the diseases considered to be related to genetic causes, there is one called pre-eclampsia (PE) which is currently under genetic analysis for any heritable association [18–23]. PE is a progressive disorder which occurs during pregnancy and in the period soon after the birth and it affects both the mother and the baby. The major symptoms are high blood pressure, swelling, proteins in the urine and problems with vision. It is one of the leading cause of death and disability in mothers and babies. The most evident effect for babies is a shorter gestational age and lower birth weight, Figure 2.8.

Occurring in around 5-8% of all pregnancies, pre-eclampsia affects four million women worldwide each year. Moreover, together with other disorders of high blood pressure during pregnancy, it is responsible globally for an estimated 76,000 maternal and 500,000 infant deaths each year [24, 25]. In particular, PE complicates 2-3% of women at first pregnancy and 5-7% of women who have never given birth before. Extensive vascular alterations which take place in the spiral arteries suppling maternal blood to the placenta is a process that usually takes place during pregnancy. In the presence of pre-eclampsia, this important physiological change is substantially restricted. Extensive clinical studies have led to the conclusion that pregnancy-induced hypertension is a complex process, usually commencing in the early stages and depending on different physiological maternal responses. This multi-factorial disease presents a syndrome of symptoms and signs, with haematological and biochemical abnormalities. Hypertension and proteinuria are considered the hallmarks, but the clinical manifestation of this disease are heterogeneous. Some women experience severe symptoms requiring intensive care whereas other remain asymptomatic. Usually, one in six babies is very pre-term, whereas two-thirds are born after the 37th week and are normally grown [25].

Pre-eclampsia: State Of The Art

The heritable aspects of pre-eclampsia are complex and difficult to detect. One of the patterns to be identified is the tendency for the risk of pre-eclampsia to be passed from mother to daughter, but recent studies have shown that an increased risk of pre-eclampsia could also be transmitted through the father [23].

In order to assess the pre-eclampsia risk associated with genes transmitted from the parents to the offspring, a cohort study has been performed on a Norwegian population, composed of 438,597 mother-offspring units and 286,945 father-offspring units. The interesting results show that the daughters of women who had pre-eclampsia during pregnancy had more than twice the risk of pre-eclampsia themselves (odds ratio 2.2, 95% confidence interval 2.0 to 2.4) compared with other women. Men born after a pregnancy complicated by pre-eclampsia had a moderately increased risk of fathering a pre-eclamptic pregnancy (1.5, CI = 1.3 to 1.7). Sisters of affected men or women, who were themselves born after pregnancies not complicated by pre-eclampsia, also had an increased risk (2.0, CI = 1.7 to 2.3). Women and men born after pre-eclamptic pregnancies were more likely to trigger severe pre-eclampsia in their own (or their partner's) pregnancy (3.0, CI = 2.4 to 3.7, for mothers and 1.9, CI = 1.4 to 2.5, for fathers). In conclusion, fetal genes from both mother or father can trigger PE even if maternal association is stronger than fetal association [23].

Different experiments have provided evidence that more than one SNP is associated with an increased risk of pre-eclampsia. A case control study over 72 cases and 70 controls, within women resident in Hungary, shows that "HSPA1B (1267)GG and HSPA1L (2437)CC genotypes are more frequent among preeclamptic than control patients, suggesting that these genotypes may play a role in the susceptibility for preeclampsia" [19]. "PE is associated with IL-10-(1082) polymorphism" was found in a study over 189 cases and 151 controls [18]. "The higher frequency of IL-10 -1082 G allele in preeclamptic patients compared to controls may be considered as a genetic susceptibility factor for the development of PE" has been shown in a study of 134 preeclamptic women compared to 164 healthy women [20]. A study over a population of 67 German patients with preeclampsia or superimposed pre-eclampsia and 100 controls with uncomplicated singleton pregnancies showed that "T235 of the angiotensinogen gene is a potent, independent risk factor for preeclampsia" [21]. Furthermore, "meta-analysis suggests that the factor V Leiden SNP is associated with an increased risk of preeclampsia" confirmed by an odds ratio of 1.81 (95% confidence interval [CI] 1.14-2.87) for all cases of PE and 2.24 (95% CI 1.28-3.94) for cases of severe preeclampsia [22].

But there have been also some experimental trials that can demonstrate the non-association between certain SNPs with the disease or some weak association. A study including 657 women affected by PE and their families revealed that "angiotensinogen, the angiotensin receptors, factor V Leiden variant, methylene tetrahydrofolate reductase, nitric oxide synthase and TNFa, tested in a large study of strictly defined pre-eclamptic pregnancies don't confer a high risk of disease" [26]. That "the K121Q polymorphism of the PC-1 gene is unlikely to be a major genetic factor predisposing to preeclampsia in Finnish women" was shown in a case-control study involved 133 women with preeclampsia and 115 healthy controls [27]. "MTHFR C677T polymorphism does not have a major role in the development of preeclampsia or placental abruption in the Finnish population" according to a study over 362 women (133 with preeclampsia, 117 with placental abruption, and 112 healthy controls) [28]. “The -56T HLA-G polymorphism is not associated with pre-eclampsia or eclampsia in population of 277 nulliparous females” [29]. “Polymorphisms of the adiponectin gene show a weak, but statistically significant, haplotype association with susceptibility to preeclampsia over 133 Finnish women with preeclampsia and 245 healthy control subjects” [30]. Finally, ‘results are not suggestive of an important contribution of the PAI1 genotype on preeclampsia across population of 115 healthy cases and 133 sick females” [31].

In all these experiments, different kinds of analyses have been used such as case control analysis, logistic regression methods, Chi Square analysis, etc., to assess genotype and allele frequency differences, and haplotype analysis has been performed using the expectation-maximization (EM) algorithm and the transmission/disequilibrium test.

Limitations on the Current Pre-eclampsia Research

From the results of these experiments a few pitfalls can be detected. For instance, the dataset size of these studies is usually limited. This is due to the fact that the most commonly used tools for disease association study often present limitations in term of scalability. In these analyses, the population size can reach a few hundred individuals, selected from a specific country, and the SNPs analysed may be selected a priori following some criteria which often is not very well justified. The size of the current datasets is quickly increasing thanks to the continuously improved techniques for gathering new data, including populations of millions of individuals and millions of SNPs. All these experiments might present relevant limitations in the analysis of such huge databases.

Additionally, there are some studies that have reported associations with pre-eclampsia, but sometimes attempts to replicate these findings have yielded inconsistent results. These studies need a more extensive analysis for validation purposes. Additionally, results from different types of approaches have provided contradictory outcomes, contradicting previous promising findings. Therefore, pre-eclampsia remains a complex disease, whose aetiology is still difficult to determine. More efficient tools, able to cope with large dataset sizes, need to be created in this medical field in order to provide more evidence for the genetic association of PE. According to these considerations, in this work medical datasets related to PE disease are analysed and used to test the proposed techniques.

2.3.6 The Medical Requirements

An important assumption of this work is that often the decisions made along the way have been driven by the practical medical needs. Data mining, statistics and intelligent analysis of genetic data are very important tools that have been able to help medical research to disclose the genetic association of several common diseases. Bioinformatics aims to provide the useful tools to perform this analysis but the continuous supervision, discussion and advice offered by the medical professionals is an essential part of the process. Interesting and efficient solutions from the computer science point of view may not have relevant meaning from the medical perspective; they may not be feasible, appreciable or competitive for medical applications. For this reason, in this work most of the choices made rely on medical advice or requests, even if they may not appear the most obvious approach from the computer science point of view. For instance, the analysis of the pre-eclampsia disease with decision trees algorithms (Chapter 4) has been performed constraining a continuous predictor into binary type in order to use decision trees instead of other solutions such as regression trees. This type of approach introduces more complication than using the regression trees, as there is a need to find the best threshold for the continuous to binary conversion. Although this could also affect the predictive accuracy, decision tree analysis was performed in response to a specific medical request to provide them an overview of the strengths and limitations of this approach.

2.4 Methods for Disease Association Studies

There are different kinds of methods that are widely used in database analysis and in this section a general overview of the statistical tools that have been used in this work is provided. In the context of genetic association studies, in order to detect genes responsible for complex diseases, two common techniques have become very popular: case-control analysis which performs analysis on populations of unrelated individuals and the 'Transmission Disequilibrium Test' which is methodology to analyse family-based populations. For this reason, these are the two methods that will be largely analysed and applied in this work. In terms of statistical power, there is not much difference between these two approaches both for rare and common disease analysis [32].

The genotyping process and the collection of data related to sick populations and their relatives in family-based association studies is usually more expensive and more time consuming as compared with the data gathering for population based studies. For these reasons, the case-control study has become more popular and common for genetic studies. Moreover, this method offers the possibility to include in the analysis variables that are not exclusively genetic. It is well known, and clearly demonstrated [33], that the environment plays an important role in the causes of an increased risk for certain diseases. Thus, being able to consider genetic factors together with life style ones, allows a stronger and more reliable overview in the extensive research of disease aetiology disclosure. On the other hand, family-based designs have a tendancy to be robust against population substructure, such as admixture, stratification or inbreeding, which may bias the distribution of the standard association studies, thus invalidating the test results.

2.4.1 Case Control Analysis

Porta, in 2008, defined a case control study as “The observational epidemiologic study of persons with a disease (or another outcome variable) of interest and suitable control group of persons without the disease (comparison group, reference group)” [34]. Case-control studies are a low cost and widely used type of epidemiological study that can be carried out by researchers in single facilities. The first step is the identification of an effect such as a single phenotype or a whole manifestation of a disease, for example Alzheimer’s, pre-eclampsia, or COPD. Subsequently, an investigation of any potential causative factors is performed in order to shape the typology of data to analyse.

The basic idea is to collect two different populations of relevant subjects, the people with a condition (‘cases’) and the people without the condition (‘controls’). The collection of the controls and cases groups, which contain individuals respectively healthy and affected from the disease under investigation, is a non trivial task as they can be affected by many irrelevant features. By means of medical records or interviews, researchers record the variables identified as risk factors, together with other non-risk variables, which can then be used to select matching controls, such as age, sex, race, geographic area of resi-

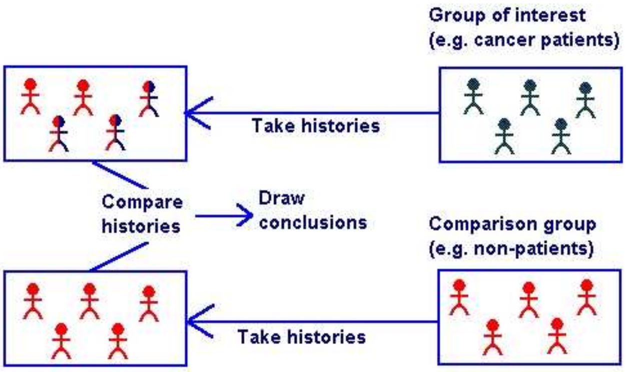


Figure 2.9: Case-Control Analysis

dence, etc. In terms of ascertainment of exposure, it is important to notice that long term recall of life style habits is probably not entirely reliable when the information comes from direct consultation with the subject due to the inevitable constraint of human memory. A better choice would be an accurate analysis of the general practice notes.

A number of healthy subjects (or controls) are then chosen who do not exhibit the outcome or effect under investigation — there may be one or more per case subject. In the best scenario, the non-risk variables should be similar in the two selected groups, allowing a reasonable elimination of these parameters in the analysis. The case and control groups are then compared on the proposed causal factors, and statistical analysis is used to detect and assess the degree of possible association between each one of these factors and the chosen phenotype, see Figure 2.9. The overall outcome of the analysis may be affected by some potential confounding features which require some statistical adjustment. In order to overcome this problem, it is advisable to choose the two groups by pairing each case with a control with the same value of some relevant parameters such as age or sex [35].

Hardy-Weinberg Equilibrium (HWE) and Case Control Analysis

Many association studies are based on the Hardy-Weinberg Equilibrium (HWE) test, a principle formulated independently by a British mathematician and a German physician in 1908. According to this principle, the frequencies of alleles and genetic heritage do not change through generations unless disturbing influences are introduced. These interfering events include mutations, natural selection, non-random mating, random genetic drift, gene flow, limited population size, etc.

In the real world, more than one of these factors is present to perturb the HWE, which therefore is considered to be only an ideal state. Genetic changes that occur in nature can be measured against this non-natural state, taken as a reference. The simplest case in nature is a single locus with two alleles, the dominant allele *A* and the recessive allele *a*, with respective allele frequencies *p* and *q*. According to this hypothesis:

$$freq(A) = p$$

$$freq(a) = q$$

$$p + q = 1$$

Supposing a population is in equilibrium ,the following applies for the *AA* homozygotes in the population:

$$freq(AA) = p^2$$

while for the *aa* homozygotes:

$$freq(aa) = q^2$$

and for the heterozygotes:

$$freq(Aa) = 2pq$$

Any state that differs from this equilibrium is caused by natural reasons and can be associated to a disease. The use of these equations provides a tool to work out the degree of association of a disease to a specific genotype, knowing the genetic mutation that causes the disease and the frequency of the disease. In order to test the deviation from HWE, the chi-squared test [36] is one of the most commonly used tools. Alternatively, there is a wide variety of different proposed approaches for the implementation of the exact test for HWE [37–40], providing that there are enough individuals present in the sample to adequately represent all genotype classes. If this is not the case, the Fisher's exact test [41] can be used as an alternative tool. An example of implementation of the exact test for HWE suitable for large scale studies of SNPs data was proposed by Janis et al. in 2005 [42]. Showing that the simple chi-squared test is affected by type 1 error (poor specificity), the authors demonstrated that their new approach controlled this type of error for large and small samples.

Within genetic association test for diseases, case control analysis is applied to detect any possible difference in allele frequency between cases and controls. Many factors can actually bias the allele frequency such as ethnicity and population history. However, although a detection of different allele frequency may not necessarily imply an association to the disease, these results can provide a basis for further analysis.

An alternative statistic to the chi-squared test which is commonly used for the case control analysis is the ‘Odds-Ratio’ which compares the probability of a certain event for two different groups, represented here by cases and controls. This parameter provides the size of effect from the comparison between cases and controls, that explains whether an exposure to a given risk factor increases the odds of developing a disease by two-fold, three-fold or higher.

Besides the chi-squared test and odds-ratio, there are different studies that test for disease association through the measure of the different HWE equilibrium for cases and controls.The ‘Trend Test’ [43] for instance is used when there is no HWE equilibrium between cases and controls. This method has been extended later using Bayesian methods to correct for multiple testing in cases of large numbers of SNPs [44]. Subsequently, a weighted average method has been proposed by Song and Elston, combining the strong point of the two above tests [45].

Case-Control Analysis: Decision Trees

Decision Trees are an example of a machine learning representation. This is a prediction modeling technique commonly used for classification, clustering and prediction tasks [46]. They are based on a realisation of appropriate tests at each step of the analysis which consists of mapping observations of inputs to conclusive target values. A decision tree has a hierarchical structure that attempts to classify initial instances based on a series of questions (or rules) about the attributes of the class. These attributes can belong to different type of variables from binary, nominal, ordinal and quantitative values, while the classes can be categorical, binary or ordinal. In other terms, given a dataset of attributes together with its classes, a decision tree produces a sequence of rules (or series of questions) that can be used to recognise the class.

Decision trees are sometimes known by two different names: a decision tree with a range of discrete (categorical or ordered) class labels is called a ‘classification tree’, whereas a decision tree with a range of continuous (numeric) output values is called a ‘regression tree’. The structure of a decision tree consists of one root and a variable number of nodes and leaves. All the nodes have exactly one incoming edge and may have more that one outgoing edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal nodes or decision nodes).

A decision tree is usually constructed recursively in a top-down manner. The first step of the construction algorithm begins with the reading of the entire dataset and the subsequent splitting of the data into two or more parts. It then repeatedly splits each subset into finer subsets until the split size reaches an appropriate level. In the simplest and most frequently case, for each node a test is performed considering only one attribute, in a way that the dataset is then split according to the attribute value. In the bottom of the tree there is the representation of the classes, each one corresponding to the most appropriate value of the output. Each class is assigned to a leaf. In the basic process, the classification of the data instance starts at the root node of the decision tree, by the test of the attribute belonging this node and moving down the tree branch corresponding to the value of the attribute. This process is then repeated at the node on this branch and so on. In the last step a leaf node is reached where the choice of class is made.

In many applications, decision structures which are smaller are definitely preferable to more complex ones, as they are more easily understood. Besides, the performance accuracy of a tree is strongly affected by its complexity which, in turn, is controlled by the stopping criteria used and the pruning method adopted. Usually, complexity is measured by one of the following metrics:

- total number of nodes,
- total number of leaves,
- tree depth,
- number of attribute used.

Decision tree induction respects the process of rule induction. Each path from root of a decision tree to one of its leaves can be transformed into a rule simply by conjoining the

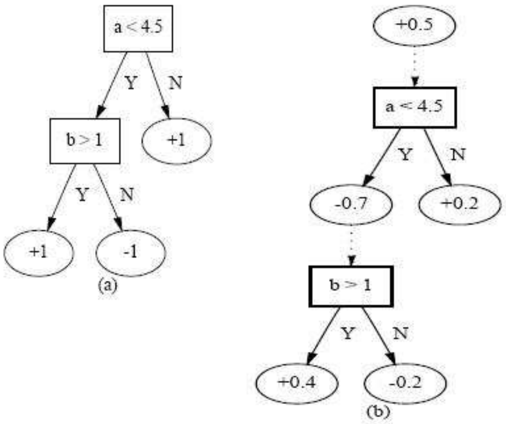


Figure 2.10: a) A decision Tree. b) The same decision tree rapresented as an alternating tree. Credit: [3]

tests along the path to form the antecedent part, and taking the leaf's class prediction as the output of class value.

Case-control Analysis: The Alternating Decision Tree — ADTree

In 1999, Freund and Mason proposed a new model of decision tree, the alternating decision tree, better known as ADTree. The ADTree is a generalization of decision trees, voted decision trees and voted decision stumps, which is relatively easy to interpret. It was created following successful demonstrations that applications of boosting procedures to decision tree produce very accurate classifiers. The standard classifiers were built following a majority vote over a number of decision trees. For this reason, these classifiers have the pitfall to be often large, complex and difficult to interpret [3].

In order to overcome these limitations, ADTree was created as a learning algorithm for alternating decision trees, based on boosting procedures. Experimental results have shown that it is competitive with other boosted decision tree algorithms and generates rules that are usually smaller in size and thus easier to interpret. In addition, these rules yield a natural measure of classification confidence, the classification margin, which can be used to improve the accuracy at the cost of abstaining from predicting examples that are hard to classify.

It is possible to derive an alternating decision tree from a simple decision tree. A simple decision tree, with two decision nodes and three prediction leaves, defines a binary classification rule which maps instances of the form of two attributes into one of two classes whose values are -1 and +1, see Figure 2.10. In an alternating decision tree, each decision node is replaced by two nodes: a prediction node and a splitter node. The decision node represents the same node of the simple decision tree, while the prediction node is characterised by a real number value. As in decision trees, an instance corresponds to a path along the tree from the root to one of the leaves. However, the difference from decision trees is due to the outcome result. The classification that is associated with the path is not the label of the leaf but the sign of the sum of the predictions along the path. The adjective 'alternating' of this new representation is indeed due to the presence of alternating layers of prediction nodes and splitter nodes.

Alternating decision trees give a much more flexible semantics for representing classifiers as compared with standard decision trees. The original and commonly used decision trees divide the instance space into disjoint regions. The majority of algorithms to learn decision trees apply an iterative procedure in order to split every instance space into different parts. Each part can be split at most once. In more simple terms, the only nodes that can be split are the leaves at each step. In general alternating decision trees, instead, the partition can be applied multiple times for each part.

The classification rules in the alternating tree are based on the definition of an 'instance' as a set of paths. In case of standard decision trees, at each decision node, the path follows the branch whose node value corresponds to the outcome of the decision. In the ADTree, from a prediction node, the path follows all of the branches present afterwards. More precisely, the path splits into a set of paths, each of which corresponds to one of the children of the prediction node. The union of all the paths reached in this way for a given instance is called the 'multi-path' associated with that instance. The classification outcome which is associated with a given instance corresponds to the sign of the sum of all the prediction nodes that are included in a multi-path [3].

Case-Control Analysis: Id3 and C4.5

ID3 (Iterative Dichotomiser 3) is a realization of a decision tree algorithm, proposed by Quinlan in 1986 [47]. The algorithm is based on the principle of Occam's Razor which is

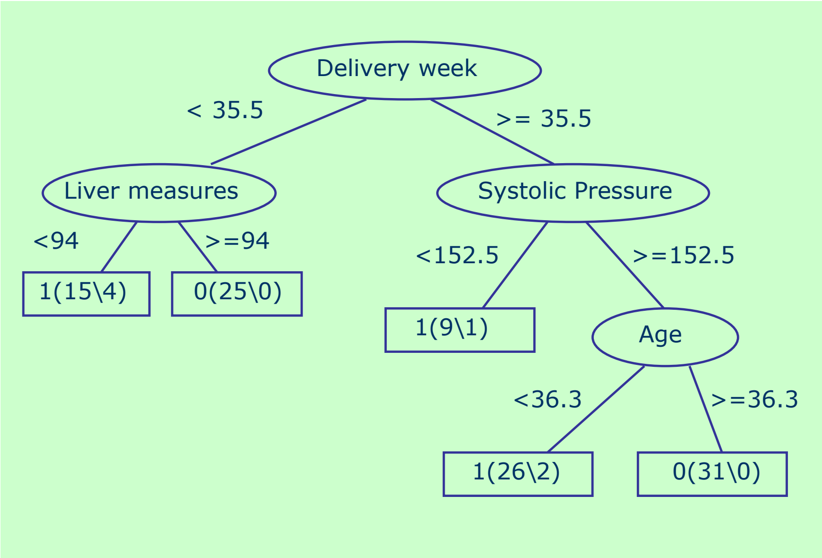


Figure 2.11: Example of an ID3 Tree.

formalised using the concept of information entropy. It is based on the principle that the world is inherently simple, and hence the smallest decision tree that is consistent with the samples is the one that is most likely to identify unknown objects correctly. According to this concept, small trees are preferred rather than bigger ones. However, the algorithm does not always produce the smallest tree, and is therefore a heuristic.

In practice, as a learning machine process, ID3 builds a decision tree from a fixed set of examples and the resulting tree is used to classify future samples. The example is composed by several attributes and is characterized by a Boolean class. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. An example of a tree created with the ID3 algorithm is shown in Figure 2.11.

Quinlan originally developed ID3 at the University of Sydney in order to improve the Concept Learning Systems by adding a feature selection heuristic [47]. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the *n* partitioned subsets (where *n* is the number of possible values of an attribute) to get their 'best' attribute. The algorithm uses a greedy search, which consists of determining the best attribute and never looking back to reconsider earlier choices. The ID3 algorithm can be summarized in the following steps:

(1) Take all unused attributes and calculate their entropy with respect to the test samples.

(2) Choose the attribute for which entropy is minimum.

(3) Make a node containing that attribute.

C4.5 is a decision tree generating algorithm, based on the ID3 algorithm and also developed by Quinlan in 1993 [48]. It contains several improvements, especially needed for software implementation, like choosing an appropriate attribute selection measure, handling training data with missing attribute values, and handling continuous attributes or those with differing costs. In building a decision tree it is possible to deal with training sets that have records with unknown attribute values by evaluating the gain, or the gain ratio, for an attribute by considering only the records where that attribute is defined. Furthermore, in using a decision tree, the classification of records that have unknown attribute values can be performed by estimating the probability of the various possible results.

Applications of Decision Trees in SNPs Analysis

In 2006, Jiang et al. highlighted that ensemble decision trees are promising algorithms for mining genetic markers for complex genetic diseases [49]. Microarray experiments have demonstrated how innovative technology can be used to classify biological types. The current analysis strategies based on learning algorithms can be divided into supervised and unsupervised methods. While the former is a useful tool for studying functional genomics, it is unable to relate different gene profiles to phenotypes. On the other hand, this task can be generally fulfilled by supervised learning, being a driven-target process where induction algorithms identify the genes responsible for a target. Moreover, microarrays provide massive parallel information that current statistical methods can hardly properly solve. Hence, machine learning solutions have been proposed in support of this aim.

Among these methods, decision trees algorithms appear to be one of the best choices for genetic analysis as they can partition the sample and feature gene space simultaneously. Thus, Jiang et al. proposed a novel tree-based ensemble method for microarray data, extracting first a subset of genes able to predict the rest, and then identifying genes that are relevant for a specific disease. Two years later, they extended this approach to sibling-pair analysis of pedigree, applying a decision approach to extract relevant SNPs for alcoholism [49].

In 2005, Kuang-Yu et al. applied boosting alternating decision trees to model disease trait information. They combined two boosting iterations (log and exponential loss functions) together with two decision trees algoritmhs (ADTree and classic boosting decision tree) and demonstrated that ADTree offers a more accurate representation of the disease status that allows for better detection of linkage evidence [50]. In 2007, Dong-Hoi et al. demonstrated that decision trees are a potential tool to predict the susceptibility to chronic hepatitis and cirrhosis from SNPs data. This was realised in an experimental result which showed the capability of C4.5 to distinguish cases from healthy people with an accuracy of 69.59% for chronic hepatitis and 76.2% for hepatitis [51]. Huang et al. in 2009 proposed a comparison of classification methods for predicting Chronic Fatigue Syndrome (CFS) based on SNPs data. In their experiment they compared and contrasted three different classifiers based on different representational models, such as probabilistic models for naive Bayes, regression models for SVM and decision tree models for the C4.5 algorithm. Their findings suggested that this type of analysis provides a plausible way to identify models in CFS [52].

This review of the literature shows the successful results that decision tree algorithms, in particular ADTree and C4.5, have achieved in the specific field of disease association study. However, the application of a single decision tree algorithm sometimes might have limitations due to the specific method employed for growing the tree. For instance, ADTree is based on a boosting procedure, while Id3 and C4.5 are based on information gain calculation for each attribute. This diversity might provide different results from the analysis of the same database. In order to overcome the limitations that a single technique may yield, in this Thesis a new framework, realised by a combined analysis of these three algorithms is proposed. In this way, only relevant solutions that are also confirmed and validated by more than one algorithm are taken in consideration for further analysis, providing a more robust solution to the problem analysed, see Chapter 4. C4.5 and Id3 are both based on the definition of information gain but they differ in the manner that the tree is grown. Whereas in Id3 solution, the stopping criteria are employed to regulate the size of the tree, C4.5 resolves the over-fitting problem by pruning the tree at the end of the analysis. The extent of pruning can be chosen by the user by setting an internal confidence parameter. These two different approaches might provide different shapes and especially different sizes for the trees. In order to test the performance of these two different approaches, both algorithms are employed in the analysis as, to our knowledge, this is a novel approach in a pre-eclampsia association study. C4.5 is supposed to be an improvement of the Id3 algorithm, but sometimes algorithms which are expected to be improved do not show their strength in some specific applications as well as in others. For this reason, Id3 is also included in the analysis, in order to provide further validation or rejection of the actual improvement. Moreover, once again, the choice of these specific tools has been driven by medical considerations, as they have been successfully used in previous analyses for different disease contexts.

2.4.2 Family Based Analysis

Family based studies have been the only option in genetic analysis for a long time and have led to the discovery of a large number of genes responsible for Mendelian diseases and traits. Nevertheless, they have not been as successful for more complex diseases such as diabetes, asthma or heart diseases. In these cases, a new approach based on population analysis has brought the best results, the case-control study. This provides better power and can deal with very large numbers of SNPs for mapping complex diseases or traits. The realisation of a case-control dataset is cheaper and faster than a family based one, which is also sensitive to genotyping errors, inflating the false-positive rate. Even if family based analysis produces more robust results for substructure and admixture problems, their software package implementation availability is still limited in comparison with the population-based commercial solutions.

The simplest design for family based analysis is the implementation of the Transmission Disequilibrium Test. Further extensions of this basic method provide solutions

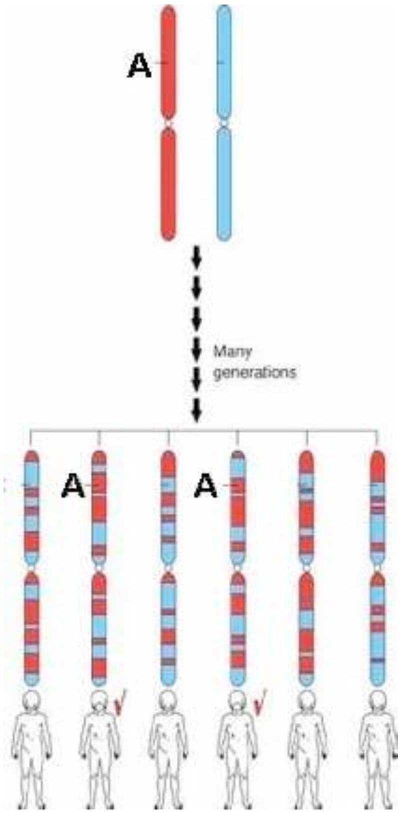


Figure 2.12: Haplotype transmition through generations. Credit: [4]

of problems such as missing parents, quantitative traits and use of additional siblings. The most successful extensions are based on non-parametric tests and they are referred as Family Based Association Tests (FBATs). They incorporate features such as general pedigrees, missing values, analysis of complex diseases and phenotypes, and they can handle datasets with multiple-comparison problems. Alternatively, likelihood based approaches are also popular in the field, as discussed later [32].

Family Based Analysis: Transmission Disequilibrium Test

Alternatively to case-control analysis, a different way of approaching the problem of disease association study is based on the Transmission Disequilibrium Test (TDT) which is a family based study, see Figure 2.12. It was developed by Spielman in 1993 [53]. In terms of statistical power, the differences between the TDT and the commonly used case-control study are generally small. The recruitment of a specific population and their relatives in family-based association studies usually requires more resources in terms of time and money than that of unrelated subjects in population based studies. Besides,

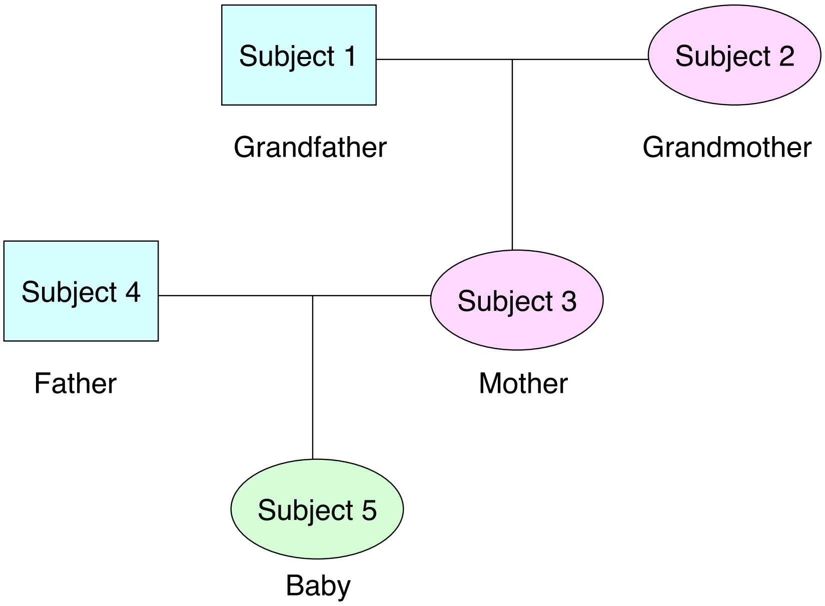


Figure 2.13: Family structure and identity number of the original database, in TDT analsyis.

more genotyping might be required for family-based studies, and together these factors have increased the choice of population designs over family-based studies. However, unlike population-based studies, family-based designs are robust against population substructure, a characteristic which might distort the distribution of the standard association statistic, leading to increased type 1 error (the probability that the null hypothesis is falsely rejected) or decreased power [32].

The Transmission Disequilibrium Test is one of a number of tests which aims to be robust against spurious associations due to population stratification by obtaining control alleles from relatives of cases. Population stratification consists of the presence in a population of distinct strata or groups that show limited inbreeding; they might have different disease rates and distinct allele frequency distributions. Failure to control for the stratification can invalidate tests of association. This technique, also known as the Transmission Distortion Test, is a simple family-based design for testing association which uses genotype data from trios. The trios consist of the two parents and their affected offspring (see Figure 2.13).

The idea behind the TDT is intuitive: under the null hypothesis, Mendel's laws determine which marker alleles (the genetic elements which can be detected by phenotype) are transmitted to the affected offspring. The TDT compares the observed number of alleles that are transmitted with those expected in Mendelian transmissions. The assumption of Mendelian transmissions is all that is needed to ensure valid results of the TDT. An excess of alleles of one type among the affected indicates that a disease-susceptibility locus (DSL) for a trait of interest is linked and associated with the marker locus. The presence of linkage implies a co-inheritance of genetic markers and phenotypes in families over several generations whereas the presence of association confirms the contribution of genetic variants to phenotypes.

If there is linkage but no association, the marker and the DSL will tend to be transmitted together, but different marker alleles will be transmitted with the DSL in different families. This results in no overall association of a particular allele that is transmitted with the trait. If there is association between the marker and a DSL but no linkage there is no tendency for the marker and the DSL to be transmitted together to offspring. In this case, one would not expect to see an excess of a particular allele transmitted in affected offspring. If an allele is transmitted to unrelated cases more often than would be expected by chance, this implies that it is linked and associated with the disease mutation. If the sample contains cases related to each other, coming from the same pedigree, then the TDT can become a test only of linkage rather than association.

If a parent is heterozygous for a marker, the chances for him to transmit both marker alleles to an affected case will be equal unless one of the allele markers is linked with the DSL and unless the marker and disease are associated. A sample of cases and their parents is genotyped and deviations from the expected 50-50 transmission are observed. Usually the sample consists of a set of trios, affected cases with their parents. However, pedigrees containing more than one affected case can also be used.

Originally, the TDT was used to test for linkage in the presence of association. However, because both linkage and association between the trait and the marker have to be present for the TDT to reject the null-hypothesis, the TDT is now typically used as a test for association. This dual-alternative hypothesis also means that the TDT avoids false positives that arise when association is present but linkage is not, as might happen in the presence of admixture and/or population stratification. (Admixture occurs when two or more subpopulations inbreed, so that two randomly chosen individuals in the population might have different degrees of genetic heritage from the original subpopulations.)

As the TDT is completely non-parametric, there is no requirement for a proper disease model neither assumptions about the distribution of the disease in the population. This makes this technique robust to potential mistakes of any features of the disease model or trait distribution setting However, the original TDT still present some drawbacks such as missing parents, general pedigrees, complex phenotypes and haplotypes with missing phase. These problems can be resolved with possible extensions of the basic method [32].

Family Based Analysis Test (FBAT)

The FBAT statistic of association is based on the definition of ‘covariance’ between genotype and phenotype. In this formula, different parameters can be set depending on different specifications of the problem. The trait and the allele in the population can have various distribution trends, unaffected offspring can be considered in the analysis together with alternative and multiple traits. Alternative genetic models for the offspring such as dominant or recessive genes, together with multiple allele analysis is also allowed in this extended technique. Moreover, the FBAT statistic approach can be generalized including arbitrary pedigree, missing parents, haplotypes, different null hypothesis and complex phenotypes (quantitative traits). It can happen that the genetic distribution of the offspring depends on unknown factors, called ‘nuisance’ parameters. The FBAT approach can handle with this limits creating a ‘sufficient statistics’ for them.

Whenever the genotypes are missing, information coming from grandparents can be useful as well as additional family members such as unaffected offspring can contribute information. Haplotypes can be studied in spite of single SNPs when the phase is known, in order to avoid multiple testing issues. In case the phase is unknown, a sufficient statistic can be applied to create a haplotypes distribution that do not rely on estimating the phase.

Family Based Analysis: Likelihood extensions

The likelihood methods are based on the definition of ‘probability density’ for the observed data as a function of genotype. In order to test the hypothesis of no association, either likelihood-ratio or score tests are used. For the former, different methods make use of different definitions of likelihood such as conditional logistic regression or multinomial likelihood. The latter is generally more popular and can be extended to account for multiple offspring and complex phenotypes. The Likelihood based approach offer more sophisticated tests, nested models and can be more efficient, but FBAT are more able to perform an easier analysis for screening and validation and they resolve better the problem of nuisance parameters.

Family Based Technique: Transmission Disequilibrium Test Implementations

The aim of this section is to provide the reader with an overview of the possible options that are available for the implementation of the TDT technique. Lots of work has been done with the attempt to improve this method and different tools have been developed in the past years based on different hypothesis. They include various input data formats, multiple statistical analysis, several population features and different genetic markers constraints.

- ETDT (Extended Transmission/Disequilibrium Test), proposed by Sham and Curtis in 1995, performs a TDT test on markers with more than two alleles using a logistic regression analysis [54]. Three different approaches are proposed. The first one considers one allele at a time and examines whether parents heterozygous for this allele transmit it to the affected offspring in more than 50% of occasions. The second approach addresses every heterozygous parental genotype separately whereas the last method attempts to establish a pattern of preferential transmission across genotypes. They prove that this test has a good power when LD is strong and if the disease is recessive.

- GASSOC (Genetic ASSOCiation). This is a statistical method for disease and genetic marker associations using cases and their parents and it was implemented in 1996 by Schaid et al [55]. It includes an extension of the transmission/disequilibrium test (TDT) for multiple marker alleles, as well as additional general tests sensitive to associations that depend on dominant or recessive genetic mechanisms.

- TDT/S-TDT. In 1998 Spielman and Ewens introduced the method ‘sib TDT’ or ‘S-TDT’ which provides a useful tool when the genetic information of the parents are missing or not complete. In this case, the method uses marker data from unaffected siblings instead of from parents, thus allowing application of the principle of the TDT to siblings without parental data. The overall analysis provides separate results for TDT, S-TDT, and the combined test, where necessary [53,56].

- RC-TDT (Reconstruction-Combined Transmission Disequilibrium Test) was implemented by M. Knapp in 1999. This is a family-based association method that allows testing for linkage in the presence of linkage disequilibrium between an autosomal marker and a disease even if there is only incomplete parental-marker information. Recently, Horvath et al. described a similar procedure (XRC-TDT) for X-linked markers. The distribution contains SAS macros that calculate the RC-TDT and XRC-TDT test statistics, as well as their respective exact p-values [57–59].

- QTDT (Quantitative (Trait) Transmission/Disequilibrium Test) was developed by Abecasis et al in 2000 to perform linkage disequilibrium (TDT) and association analysis for quantitative traits [60]. It includes supports for families of any size, with or without parental information and simple variance components modeling.

- ET-TDT (Evolutionary Tree - Transmission Disequilibrium Test) is a procedure based on combining the benefits of measured haplotype analysis (MHA) with TDT in order to find which haplotype or groups of haplotypes, as defined in an evolutionary tree, are responsible for increased (or decreased) relative risk for a genetic condition. A stringent requirement of the original TDT is that at least one parent must be heterozygous. Even in this case, transmission of allele may not be obvious when both parents are heterozygous for the same biallelic marker. Studying haplotypes instead of single SNPs have proved that parents and offspring are more informative. ET-TDT was written by Seltman et al. in 2001 and it is developed in two steps. Firstly, the haplotypes from the trios are inferred distinguishing between ‘ambiguous’ and ‘unambiguous’ ones. MHA uses the evolutionary relationships among haplotypes to produce a limited set of hypothesis tests. Then, the proper analysis is applied building an initial evolutionary tree with the haplotypes as nodes and step by step decreasing the size of the tree [61].

- FBAT (Family Based Association Test) is a user-friendly, well-documented software developed by Xin Xu et al. in 2001. It allows the user to test for association/linkage between disease phenotypes and haplotypes by utilizing family-based controls. It is robust to population admixture and population stratification, it can deal with not-known phase individuals by using weights, which are estimated from the sample. The method can handle any type of phenotype, including multiple phenotypes and missing parents, marker data, and/or phase, and provides both bi-allelic and multi-allelic tests, [62–64].

- TDT-AE (Transmission Disequilibrium Test Allowing for Errors). This program, written in 2004 by Gordon et al. computes a likelihood-based transmission disequilibrium test [65]. The data are genotypes on trios in which random genotyping errors leading to Mendelian inconsistencies may or may not have occurred. This program computes the TDT-AE statistic on all trios (whether Mendelian consistent or not) and thereby maintains a correct type-I error rate in the presence of random genotyping errors.

- SNP ASSISTANT has been created by Biodata Ltd. This is a software for SNP data managing that provides import and export from linkage format, data validation, pairwise LD calculation and visualization, case-control and TDT tests, visual comparison of two datasets, and relationships testing [66]. It is suitable for large projects and it does not depend on the genotyping method.

- TRANSMIT tests for association between genetic marker and disease by examining the transmission of markers from parents to affected offspring [67]. It was developed by Clayton in 1999 and the main features which differ from other similar programs are that it can deal with transmission of multi-locus haplotypes even if the phase is unknown, and that parental genotypes may be unknown. For a more detailed analysis of this software the reader is referred to chapter 3.

All these solutions have been developed from the basic approach of TDT in order to provide data analysts with useful tools in the SNPs study, trying to cover a wide range of constraints and different scenarios. The parameters involved in SNPs association analysis are different and can belong to different variables types depending on the specific application. The type of outcome can be a Boolean variable if the sick-healthy status is analysed but it can be also a categorical one if different degrees of disease severity are considered. Alternatively, if a phenotype is under study, the outcome could be continuous or it could consist of more than one physiological parameter. The disease model could change in different scenarios, varying from dominant to recessive or mixed. Different kinds of pedigree can be studied and it is already widely proved that often different populations present different susceptibility to specific diseases because of their genealogy. Sometime information can be missing especially when the diseases under analysis occur later in life and the genetic data from parents are not available anymore. On the other hand, information from siblings or other relatives of the family can become useful to recover this missing information.

Together with all these parameters that contribute to the general model for SNPs data analysis, one of the major problems that data analysts need to face in genetic association study is the possibility to deal with transmission multilocus-haplotypes from parents to offspring when the phase is unknown. There are not many methods that can deal with this problem and one of them that is widely used and accepted is the TRANSMIT software. One of the pitfalls of this software is that the haplotype length is fixed and set by the user in an initialization phase. The possibility to detect the minimum size of haplotypes responsible for an increased risk of disease would provide a better usage of this tool for further analysis. For a detailed analysis of this contribution, the reader is referred to Chapter 3

2.5 Methods for SNPs Dataset Size Reduction

This section provides an overview of the current techniques employed for dataset size reduction purposes. Feature selection is an active research area focused on selecting a subset of input variables by eliminating features with little or no predictive information. This finds applications in data mining, statistics and pattern recognition. The major benefits of this approach include facilitating visualisation and understanding of data, reducing time and storage requirements and defying the curse of dimensionality to improve prediction performance. Whereas there is a huge amount of work on feature selection in the general Machine Learning community, in this section the attention is focused exclusively on methods tailored for SNPs analysis, according to the main scope of this work.

2.5.1 SNPs and Haplotypes Tagging

Different methods can be found in the literature for selecting a small set of SNPs from the initial database for genome-wide association studies. This smaller group of informative SNPs, often referred as tag SNPs, are representative of the original SNP distributions in the genome. If they are chosen from haplotype data they are referred as haplotype tag SNPs (htSNPs).

The experimental determination of haplotypes realized with normal genotyping (based on PCR/sequencing) of an autosomal SNP is very expensive and time-consuming [68–70]. Alternative computational methods can provide very good tools for haplotyping in populations which have the genotypic information available for enough individuals (i.e. the alleles present for each SNP in a genetic locus). Several experiments have shown that these methods are reliable and feasible, as there are a relatively small number of haplotypes present in a given population which are maintained according to the rules of evolution.

In 2005, Zhang et al. gave five different ways to define tag SNPs [71]:

- A minimum set of SNPs able to distinguish a percentage of all the haplotypes [72].
- A minimum set of SNPs that can account for a certain percentage of overall haplotype diversity.
- A minimum set of SNPs that can account for a certain percentage of overall haplotype entropy [73].
- A minimum set of SNPs with a maximum overall haplotype prediction strength, defined as the measure of uncertainty in the prediction of haplotypes from genotypes data [74].
- A minimum set of SNPs with a maximum prediction power, which is based on the definition of LD [75].

According to the last definition, the measure of Linkage Disequilibrium between SNPs is a very common technique which is used to detect disease associations selecting only a subset of SNPs which are considered informative for this analysis.

For instance, Kruglyak in 1999 estimated the extent of LD surrounding common gene variants in the general human population, for mapping common diseases genes [76]. In 2001, Goncalo et al. showed the extent and distribution of LD in three genomic regions for association studies [77]. In 2002, Stacey et al. explained how the human genome can be objectively parsed into haplotype blocks and how this framework provides statistical power in association studies [78]. Tozaki, in 2005, demonstrated that LD measurement is useful for mapping genes in thoroughbred populations and also for complex traits, as LD is expected to be strong around the loci of diseases [79].

From the statistical point of view, the tag SNPs are required to statistically cover the non tagged SNPs with a quality measured by the squared correlation coefficient. Non tag SNPs are expected to be highly correlated with the tag ones with an $R^2 \geq 0.8$. Consequently, significant effort has been devoted to the optimisation point of view in order to minimize the number of tags in respect with a given prediction error. In other terms, the tag SNPs selection issue can be divided into three different aspects [80]:

- The informative SNPs selection problem (ISSP) — Detect a smaller amount of SNPs from the initial datasets, able to predict the rest of the SNPs for any given population.

- The SNP prediction problem — infer the rest of the SNPs from the tagged ones, minimizing the prediction error.

- The statistical covering problem — optimise the results, maximising the number of predicted SNPs from the chosen number of tag SNPs.

There are different ways for classifying tag SNPs and htSNP tools. In this section, two different classifications are shown, each one followed by a list of applicative solutions. The first is based on two different procedures: the block-based and the block-free methods. Another type of classification within the several algorithms for inferring the tag haplotypes from a population of genotypes, developed recently, is based on the difference between combinatorial methods (which focus on haplotype pairs for each individual) and statistical methods (which focus on the haplotype frequencies in the population). All these four distinct types of method are discussed in the following sections.

2.5.2 Block based methods

The block-based methods refer to the assumption that the human genome can be partitioned into different haplotype blocks of variable length [78]. In each of these blocks, most of the population share a small amount of common haplotypes [72, 81]. This means that different individuals have correspondent blocks containing mostly different haplotypes. Besides, the recombination of haplotypes belonging to the same block over time is rarely observed. Such haplotype framework provides a useful benchmark for association studies of genetic common variation between blocks. Following this definition, in many approaches the selection of the tag SNPs is performed after partitioning the genome into haplotype blocks. Example of block-based solutions proposed in the literature are illustrated in the following list.

HapBlock

In order to analyze the block-like LD patterns in the human genome and perform a tag SNPs selection, Zhang et al. in 2005 developed a set of dynamic and flexible algorithms based on different criteria [71]. These techniques can analyze both genotype and haplotype data coming from both unrelated and related individuals. This proposed software, called HapBlock, includes different programs for haplotype blocks partitioning and tag SNPs selection developed by the same authors in the previous years [82–84]. The problem can be approached in two different ways: a fixed section of genome partition with a minimum number of tag SNPs, or a fixed number of tag SNPs for predicting the maximum length of the genome. Both these approaches are used in the software [83]. Haplotypes are inferred for each block from a subset of SNPs included in the block, according to the high LD. Genotypes are instead inferred using the PL-EM algorithm [85], later shown, under the assumption of no recombination events.

Zhang et al. in 2005 gave three different definitions of haplotype block:

- A percentage of inferred or observed haplotypes must be common [72].

- A percentage of SNPs in the block must have a strong LD [83].
- There must not be historical recombination of the SNPs within the block.

Additionally, all the definitions for tag SNPs, previously given, are used in this software to find the optimal set of tag SNPs. The overall algorithm thus provides a unified platform to assess the power of association studies using tag SNPs, based on different methods.

htSNP2 and genassoc

In 2003, Chapman et al. discussed association studies and how the initial data used to select tag SNPs should be incorporated in the analysis. The genetic risk factors that are responsible for complex traits can be common SNPs with small effects, rare SNPs with large effects or (in the worst case) rare variants with small effects. Additionally, the definition of 'small effect' is often not clear. Focusing on the first hypothesis, they outline a formal statistical model for population-based association studies together with a power measurement based on LD function. This method, implemented in Stata [86], selects the tag SNPs capable of predicting the remaining common SNPs, defined as those with an allele frequency $>= 3\%$, with a minimum R^2 of 0.8 [87].

GERBIL

Kimmel et al. in 2005 proposed a stochastic model for genotype phasing in the presence of recombination. This is a blocks-based model and haplotypes are generated from a small number of core haplotypes, considering mutations, rare recombinations and errors. Within each block, an EM algorithm is used in order to redefine common haplotypes in a probabilistic setting and seek a solution that has maximum likelihood. This efficient and simple to use software package is called GERBIL (GEnotype Resolution and Block Identification using Likelihood) and it both reconstructs block partitioning and resolves the haplotypes [88].

2.5.3 Block-free methods

With the block-free methods, the tag SNPs are detected from the original set of SNPs without assuming prior block partitioning and the rest of the SNPs are predicted with a minimal error. These can be based on allelic bonds, better known as Linkage Disequilibrium, across one or a few gene regions. Some examples of software solutions available in the market for the block-free methods, also known as entropy methods, are listed below.

SVM/STSA

In 2007, Jing et al. highlighted how the choice of a tagging technique is very much affected by the SNPs prediction models used afterwards. They proposed a greedy ‘Stepwise Tag Selection Algorithm’ (STSA) and a ‘Local Minimization Tag Selection Algorithm’ (LMTA) version for tag SNPs selection, and suggested two different prediction models, multi linear regression (MLR) and support vector machines (SVMs).

The STSA detects the final set of tag SNPs adding one SNP per time, choosing the best one that minimises the prediction error. With the LMTA, the set of tags is chosen at the first step and each single tag is replaced subsequently with a best one until there is no significant improvement on the prediction quality. The MLR-tagging procedure was introduced in 2006 by He and Zelikovsky as a new approach for informative SNP prediction, based on multiple linear regression. In the MLR method, the predicted SNPs are just a linear combination of the tag SNPs, each one weighted with a different coefficient. The SVM instead is a learning system developed by Vapnik and Cortes in 1995 [89], which is very accurate and highly competitive with other solutions such as neural networks. The aim is to maximise the margin between solid hyperplanes that separate the two classes. This works also with data which is not linearly separable using non linear function to map data in the space.

Comparing the results with other different solutions in the fields ([90, 91]) through several experiments with the same datasets, it has been demonstrated that MRL prediction combined with stepwise tag selection (STSA) detects fewer tag SNPs, whereas the SVM solution uses fewer tags than the methods reviewed by Halladorsson in 2004 [92]. With MLR, both STSA and LMT give the same quality results, with LMT being faster. However, SVM combined with STSA is more accurate even if more time consuming [80].

STAMPA

In 2005 Haperin et al. proposed a new method to predict the rest of the SNPs given the tag SNPs set [90]. This is based on analysis of genotype information for unrelated individuals and does not rely on block partition of the genomic region. In order to select the tag SNPs, STAMPA uses the phase information in a reference training set and for predicting the rest of the SNPs the genotype information is used instead of the haplotype information of the tag SNPs, as previous techniques do. Additionally, to evaluate the average SNP prediction quality, the prediction accuracy measure has been proposed. Comparing the performance of this solution with other mentioned tools, STAMPA outperforms them consistently. STAMPA uses ten times less SNPs than IdSelect ([91]) and has a prediction accuracy 15% higher than HapBlock ([71]).

GA and KNN

In 2009, Chuang et al. proposed a method for selecting a small set of SNPs from the initial database, estimated to deal with dataset sizes of up to 10 million SNPs, for genome-wide association studies. A Genetic Algorithm (GA) to select informative SNPs is combined with the K-nearest neighbor (K-NN) method to evaluate the prediction accuracy of the tag SNPs set [93]. From the comparison of this method with the previously cited ones, this method was found to be more accurate, faster and provided a smaller number of tag SNPs for six different types of datasets.

Methods Based on R^2

Within the block-free methods, approaches based on using the square correlation coefficient R^2 between two different SNPs have been employed by different researchers obtaining good results [75, 87, 94]. In 2004, for instance, Carlson et al. proposed a method to select informative SNPs through the calculation of correlation with LD function, spotting high correlated SNPs with a value of $R^2 > 0.8$ [75].

PCA and Logistic Regression

In 2008, Zhang et al. proposed the use of Principal Component Analysis (PCA) and a regression model for genomic association data analysis in order to accommodate the presence of LD and reduce the multiple testing problem. This is based on statistical inferences being made simultaneously for a set of individuals and, in this case, for a set of SNPs. Adjustments need to be done when the analysis moves from an individual to populations, generally providing a stronger level of evidence to be observed in order to compensate for the number of inferences being made. A traditional example of a method proposed to overcome this problem is Bonferroni correction [95] but such corrections for multiple testing on a large SNPs dataset can lower the statistical power of the study, which in turn would require an increased sample size to retain enough power.

The tagging SNPs method is an alternative approach proposed to reduce multiple-testing problems, but tag SNPs may vary with the different haplotype construction methods and with different analyzed populations. Also, haplotype analysis has been proposed to resolve this problem but computational limitations may affect the choice. Further, ambiguous haplotypes may result from analysis of SNPs which are not in strong LD between each other. There is still some controversy as to the relative merits of the haplotype-based and genotype-based testing in terms of analysis power [82, 87]. Zhang et al. used PCA to detect a group of correlated SNPs which may or may not belong to the same DNA fragment. Subsequently, logistic regression of the studied disease was applied to each PC score to detect any possible association. They compared the correlation matrix obtained from PCA with the LD matrix based on the correlation coefficient R^2, showing a few advantages: SNPs do not require to be in HWE, nor in a contiguous DNA fragment if they belong to the same block [94].

Other Methods

Hierarchical Clustering, PCA and Multiple Regression are examples of approaches that have been proposed to improve the tagging SNPs or haplotype-based analysis. All of these methods aim to select a smaller number of SNPs that can be representative of the rest of the population. Also, a discrete Fourier Transformation based genotypic or haplotypic score to reduce the multiple-test issue was proposed by Wang and Elston in 2007 [96]. An alternative application of block free-method was proposed by Woosung in 2006 through the usage of the sequential t-test to determine the minimal required sample size [97]. He and Zelikovsky introduced a new approach for informative SNP prediction in 2006 based on multiple linear regression (MLR-tagging) [98].

2.5.4 Combinatorial Methods

The objective of tagging SNPs consists in choosing a small amount of SNPs that maximise a certain criterion, in this case the capability to predict the rest of the SNPs included in the initial dataset. The nature of this problem can be described as combinatorial as all the possible combinations of SNPs needs to be extracted from the original set and only the ones that maximise the criterion need to be selected. The main problem is that an exhaustive search through all subsets of SNPs is not a tractable process for even a moderate number of SNPs. Different approaches are thus proposed in the literature in order to bypass the complexity of a direct combinatorial search and detect the best selection of tagSNPs, addressing the problem for instance through iterative procedures.

HAPAR

Among other approaches proposed for extracting the smallest set of haplotypes explaining the genotypes in a population, in 2003 Wang et. al used a 'branch and bound' approach [99]. This makes use of a parsimony model suggested in several places and first proposed by Gusfield in 2001 [100]. Gusfield also proposed a linear programming formulation which considers the number of mutation events to generate the haplotype set [101].

Linear Reduction Method

In 2004, Jiungwu et al. suggested a combinatorial method based on a linear algebra approach for selecting tag SNPs, which can be combined with the LD measure. Basing the method on a 10% sample of the initial population, it is possble to predict the entire population from only 0.4% of it with 2% of accuracy [102].

2.5.5 Statistical Methods

In statistical models, the haplotypes are inferred through the frequency distribution among the population, allowing a higher degree of data complexity in terms of size, missing values or multiple allele problems. The EM algorithm, being one of the most popular approaches in this field, estimates the haplotype frequency through the maximisation of the sample likelihood under the assumption of HWE.

Partition Ligation (PL) algorithm

Alternative approaches rely on Bayesian statistics [103, 104] and infer haplotype frequency distribution from genotype of the sample and prior information about haplotype distribution. Niu et al., for instance, in 2002 proposed a new Monte Carlo strategy [103]. This strategy first breaks down all the marker loci into units and then uses Gibbs sampler to construct haplotype from the units and to rebuild the phase hierarchically. It uses a Dirichlet model to choose randomly haplotypes when the previously chosen haplotypes are not good for inferring others, using other software such as Phase or HAP tools. They showed that their algorithm is robust against violation of HWE, missing data and haplotype recombination.

PL-EM

The EM algorithm is a remarkably popular statistic approach for its interpretability and stability. One of the pitfalls of the EM algorithm alone is that the excessive number of possible loci in a single haplotype is limited by memory constraints. Therefore, an improvement has been suggested by Quin et al. in 2002, termed the PL-EM algorithm [85], which is based on partitioning the dataset into smaller sets. This solution is based on the application of the partition-ligation (PL) strategy, firstly introduced by Niu et al. in 2002 and coded in HAPLOTYPER [103], combined with the EM algorithm. Compared with the HAPLOTYPER, where the PL strategy is combined with the Gibbs sampling, this new approach is a deterministic procedure and much faster. It also includes a simple and robust approach for variance-estimation for the haplotypes selected at the final ligation stage.

BNTagger

In 2006, Lee et al. proposed a new method for tagging SNP selection called BNTagger [105]. This approach is based on conditional independence among SNPs and does not rely on strong assumptions such as prior block-partitioning, bi-allelic SNPs or a fixed number of haplotypes needed to predict a single tagged SNP. Moreover there is no assumption of fixing the neighborhood in which htSNPs are selected. Through the use of Bayesian networks, they try to select a small number of SNPs which are independent and highly predictive. Genotype data is the input of the system and the haplotype data containing tagging SNPs with maximal prediction accuracy is the output. Preserving a good performance over small datasets of SNPs, this method results in a better prediction accuracy as compared with other methods.

2.5.6 Current Limitations of SNPs Tagging Techniques

This literature review of feature selection techniques for SNPs dataset size reduction provides a general idea of how many different approaches have been proposed by different researchers in the recent past. Combinatorial methods provide good performance in terms of accuracy of the results but, in general terms, they experience problems when the size of the datasets increases, missing values are present and the final haplotypes set may not always be the smallest. Statistical methods usually overcome these problems, allowing analysis of very large sized datasets. However, the computational complexity associated with the large number of SNPs to be analysed often presents limitations from the hardware point of view. If on one hand the block-based solutions provide a reduction of the problem complexity, by focusing the analysis in smaller subsets of genetic information in so called blocks, on the other hand one of the main pitfalls of these methods is that the definition of block is not straightforward and there is no standard criteria for forming the blocks. Besides, each block ignores any inter-block correlations [106]. To overcome this problem block free solutions have sometimes been proposed at the expense of run time considerations. Each of these approaches thus presents differing strengths which makes them more fit for a given application and less for others.

The resdarch in this Thesis is focused on the linkage disequilibrium function because this is a tool that can be combined in all the possible types of methods discussed here, and for this reason it is a tool of relevant interest for all tasks of SNPs size reduction, regardless the basic approach employed. In particular, the study is performed upon the SNP association tool which was developed in 2007 by Gonzalez et al. in order to carry out common analysis in whole genome association studies. This specific implementation of linkage disequilibrium (LD) function has been chosen for possible improvement because it is implemented in the R language which, as later explained (2.6.2), is an open source, commonly used programming environment which is much appreciated by the research community for its flexibility and accessibility. Moreover, this software package provides comprehensive functionality for a variety of different genetic analysis purposes. It contains tools for data manipulation, exploratory data analysis with graphics and assessment of genetic association for both quantitative and binary traits. Different models of inheritance between dominant, recessive, over-dominant or log-additive can be chosen for the study. One of the main pitfalls of this software, as the authors state, is that detecting interaction between SNPs such as Linkage Disequilibrium become a critical issue when the size of the dataset reaches a large number of SNPs. For these reasons, in the research presented in this Thesis, an improvement of this software is proposed in order to overome the limitations that large datasets introduce in terms of computational complexity, and therefore time and memory constraints. A detailed analysis of the proposed new approach is presented in Chapter 6.

2.5.7 Clustering Techniques

In the new technique for redundancy elimination proposed in Chapter 6, different clustering techniques are employed. For this reason, in order to provide a grounding to these approaches, a brief overview of the clustering techniques that have been used in the experiments is now given. Clustering techniques aim to divide a population into natural subgroups where instances strongly resemble to each other. Nowadays there are several tools available that perform this task in different ways. These clusters can either be exclusive if the instance belongs in one single group or overlapping when the instances fall in several groups. They can also be probabilistic if the instances belong to each group with a certain probability. Finally they can be hierarchical when there are fixed groups at the top level, which are further refined in the lower levels even down to the individual instances.

Hierarchical Clustering

The Hierarchical Clustering is based on the use of the similarity matrix which gives a measure of the all pairwise relationships between the instances of the given dataset. These association measures are used to build a tree displaying the specified relationship between the entries. This technique realises a sequence of steps in which the dataset is partitioned in a first level categories which in turn contain the subsequent level of partition. There are therefore different output depending on the level chosen. The branches at the bottom of the tree represent an entry while the root of the tree represents the entire collection of entries.

There are two kind of hierarchical clustering techniques, the agglomerative and the divisive, which work in opposite direction. In the agglomerative method, the first step assigns a cluster to each instance and subsequently merge two cluster at each iteration until the final unique cluster is reached which consists of the whole dataset. There are different aggregation methods that can be used and each one is based on a different definition of similarity/ dissimilarity between two given groups of data [107, 108]. The following list includes some of the possible options available for similarity definition:

- Ward (recursively):

$$d(AB,C) = [(n_A + n_C)d(A,C) + (n_B + n_C)d(B,C) - n_C d(A,B)]/(n_A + n_B + n_C)$$

- Single: d(A,B) = minimum of all distances between A and B
- Complete: d(A,B) = maximum of all distances between A and B
- Average: d(A,B) = mean of all distances between A and B
- Mcquitty (recursively)

$$d(AB,C) = 0.5d(A,C) + 0.5d(B,C)$$

- Median or Gower's method (recursively)

$$d(AB,C) = 0.5d(A,C) + 0.5d(B,C) - 0.25d(A,B)$$

- Centroid can be calculated recursively:

$$d(AB,C) = [n_A d(A,C) + n_B d(B,C) - n_A n_B/(n_A + n_B)d(A,B)]/(n_A + n_B)$$

In the divisive method, the procedure starts from the root of the tree and the population is split up at each step until the leaves are reached and the amount of clusters is equal to the number of instances. This second approach provides advantages for users interested in the main structure of the data rather than in a detailed description of individual points. Nevertheless, this method brings computational problems due to the consideration of all the possible divisions of the data in two distinct subgroups, at the first step. For this reason this method has rarely been applied and it is commonly not included in the available clustering algorithms.

In order to visualise the results from the clustering, a special type of tree structure called a ‘dendrogram’ is provided by hierarchical clustering algorithms. This graphic consists of layers of nodes, one for each cluster. Lines connect nodes to represent clusters, nested into one other. Horizontal cuts of the tree detect different cluster solutions.

A typical feature of hierarchical clustering is that it can never correct a choice made at a certain step, as once the agglomerative technique has merged two clusters they cannot be separated any more. Equally, once that the divisive method has split up a cluster, this cannot be reunited any longer. This lack of flexibility, while on the one hand dropping the computational complexity, on the other hand precludes any possible correction [109–111].

K-Means

The ‘k-means’ algorithm is a simple, classic and straightforward technique which detects clusters by dividing the population into disjoint groups composed of numeric instances. The number of clusters, k, needs to be set beforehand so that k random points are chosen as cluster centers. The instances are assigned to each cluster according to a Euclidean distance matrix. Then the centroid of the instances is calculated for each group as the mean distance and this is considered the new cluster centre. Repeating the whole process until the same points are assigned to the same cluster brings the output to a stabilised value for each group. Once the iteration is stabilised, each point is assigned to its near cluster centre in order to minimise the Euclidean distance. Yet as this minimum is not global but only local, the final output is quite biased by the initial random choice. This means that different trials can easily give different results. For this reason this algorithm is usually run several times, with different initial choices and the result that is most consistent with the specific application (according to some determination) is chosen. There are then several variants of the basic ‘k-means’ technique which have been developed for different applications together with supporting analysis for choosing the best number of clusters [112].

EM-Algorithm

The ‘k-means’ technique shows some of the typical shortcomings of heuristic clustering such as the arbitrary division into *k* groups, the cut-off value to prevent each instance to become a single cluster, the possible influence of the ordering and the questionable choice of the final local minimum. An alternative method, that features a more principled statistical approach, can be used to overcome some of these drawbacks. The basic idea of an ‘Expectation Maximisation’ (EM) Algorithm is to assign each instances a probability to belong to each cluster instead of placing them categorically into one of them. In simple terms, the dataset can be seen as a group of clusters, each one with a different distribution pattern which gives a probability that a given instance would have a certain set of attribute values if it was a member of that cluster. Moreover, the clusters are not equally probable which means another distribution explains their relative populations. All these considerations necessitate a substantial number of parameters of the mixture model that need to be calculated. The basic idea is to iterate the same basic procedure used for ‘k-means’. After an initial estimation of the unknown parameters, the clusters probabilities for each instance is calculated and used to reestimate the parameters in an iterative process. The algorithm converges toward a fixed point without really reaching it and the process can be stopped when the goodness of the clustering is an acceptable one. This ‘goodness’ is measured by the overall likelihood that the data comes from the dataset, given the clusters found, and it increases at each iteration of the EM algorithm. As for the ‘k-means’ technique, as the maximum to which the algorithm converges is only local (and not global), the whole procedure should be repeated many times with different initial guesses and the largest maximum should be chosen [112].

2.5.8 Sampling Techniques

In this section an overview upon the sampling techniques is given in order to provide a background for the type of sampling method used in the RDsnp function, described in Chapter 6.

A description of the sampling solutions, provided by Levy et al., can be summarized as below [113]. Sampling is a relevant aspect related to the data collection task. It consists of the selection of a subset of observations from an initial population under analysis. It is a statistical approach used for extracting specific information from initial datasets especially for statistical inference tasks. The main reasons why a population is sampled for analysis is to reduce the costs and time for analysing every single individual and to overcome the problem of a dynamic population, in which individuals change with time. The sampling process can be divided into probabilistic and non-probabilistic types. In the former, every individual of the population has a chance of being selected, which is accurately determined. In the latter, there are some elements of the population that have no chance of being selected or this probability cannot be determined. Within these different types of approaches, various sampling methods can be applied, depending on the cost constraints, availability of information, expected quality of the results etc.

Simple random sampling consists of the selection of elements from the group with the same sampling probability. This means that all the subsets of the population, including every single element has the same probability of being selected, minimising bias and simplifying the analysis of results. A drawback of this method is that the subset sampled may not be representative of the population because of its random nature. Moreover, this approach does not provide sampling from different subsamples of the population. This can be overcome applying systematic and stratified techniques.

Systematic sampling consists of ordering the population in a given scheme and then selecting every n^{th} element where n is the ratio between the population size and the sample size. Providing that the starting point of sampling is random, this approach is a probabilistic one. In this case, all elements have the same probability of selection but different subsets of the same size have different selection probabilities. This approach presents limitations in presence of periodicities, losing the representativeness of the sample. Whenever the population is divided inro different subgroups, stratified sampling can be employed. Each stratum is sampled as an independent subgroup and therefore different sampling approaches can be applied. This method allows for inferences upon subgroups and more efficient statistical estimates. Some of the drawbacks are due to an increase of the cost and complexity of the sample selection, the bias related to the type of the specific stratification chosen and the possible requirement of a larger sample size, depending on the amount of strata.

Another type of sampling is based on clustering. These techniques selects individuals in specific area or cluster, with specific characteristics. Even if this reduces cost for the sampling phase, it could affect the accuracy of the results due to a potential bias in the cluster choice.

Sampling procedures are subject to different type of errors, in particular the sampling errors include selection bias when the selection probabilities are different from the expected ones and random sampling errors due to the elements in the sample being selected at random. In 2002, Daszykowski et al. gave an overview of different types of subset selection methods aimed at detecting the most representative elements of a dataset, for different applications [114]. Within the sampling methods, they distinguished two main groups, the cluster-based designs and the uniform designs. In the former the dataset is first clustered (for instance with a K-means approach) and then the representative objects are selected. In the latter the selection is applied uniformly within the dataset. The basic concept of uniform design techniques is to select the object which is the closest to the data mean as first component of the representative subset. Following this, in a recursive way, the dissimilarity between the object in the original dataset and the objects in the new subset created is assessed and the most dissimilar object to the ones already included in the subset is selected, until the final subset size is reached. Different algorithms based on this design can be found in the literature [115–117]. Regarding the cluster-based design, the selection is performed after the identification of different groups in the dataset. This requires the application of cluster techniques within the various options available [109].

Iteratively, the selection of one object from each cluster, e.g. the closest to the mean is repeated until a subset of desired size is obtained. If the data is originally clustered, this technique should be applied to each single cluster. In conclusion, as the uniform designs can deal with any type of data structure, they will be less computational expensive than the cluster-based methods when the original data is already composed of clusters [114].

In this approach presented in this Thesis, the random sampling technique is chosen for selecting representative samples from general groups created. This is due to the need to reduce the computational complexity and run time of the technique analysed. For a more detailed analysis of this issue the reader is referred to 6.7.

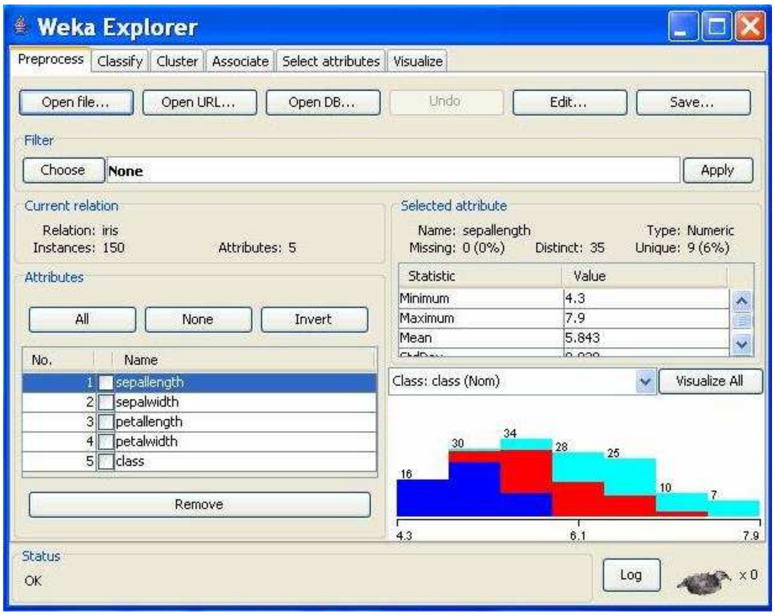
2.6 Software Packages

In the following section, the two major software packages used in this work are presented. The Weka tools are used for decision tree analysis, and the R language is used for the implementation redundancy detection technique, while the improvement of TRANSMIT software is realised in C++.

2.6.1 Weka

Weka is a software package which consists of a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from Java code via labraries. This software includes different types of data processing tools (e.g. pre-processing, classification, regression, clustering, association rules and visualization) but it is also well-suited for developing new machine learning schemes. An example of the friendly interface provided by Weka is shown in Figure 2.14.

The dataset is roughly equivalent to a two-dimensional spreadsheet or database table and it is a collection of examples which are usually known as ‘instances’. Each instance consists of a number of attributes, any of which can be nominal (one of a predefined list of values), numeric (a real or integer number) or a string (an arbitrary long list of characters, enclosed in ‘double quotes’). The external representation of an onstances class is an ARFF file, which consists of a header describing the attribute types and the data as comma-separated list.



Filter: None

Relation: iris
Instances: 150 Attributes: 5

Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843

No.	Name
1	□ sepallength
2	□ sepalwidth
3	□ petallength
4	□ petalwidth
5	□ class

Class: class (Nom)

Status: OK

Figure 2.14: Weka - exemple of friendly interface

In the pre-processing step, which which has a relevant role in machine learning, it is possible to apply different kind of filters to the dataset. The Weka filters package is composed of different tools that transform datasets by removing or adding attributes, re-sampling the dataset, removing examples and so on. All filters offer the options for specifying the input and output dataset and further different parameters, specific to each filter, can be set accordingly. The Weka filters package is organised into supervised and unsupervised filtering, which in turn are subdivided into instance and attribute filtering.

In the second step it is possible to choose within a large amount of classifiers which are the core of Weka. A classifier model is an arbitrary complex mapping from dataset attributes except one which represents the class attribute. Each classifier performs this mapping using different criteria, referring to different models of analysis. There are several different options for classifiers, most of which are related to evaluation purposes. These are learning algorithms composed by specific routines able to generates a classifier model from a training dataset (‘buildClassifier’), to evaluate the generated model on an unseen test dataset (‘classifyInstance’) and to generate a probability distribution for all classes (‘distributionForInstance’).

There are different ways to measure the classifier performance, for instance through the accuracy measurement. An example of this applications is the ‘hold-out estimate’ which is realised using a training set and a test set which are mutually independent. In order to estimate the variance in these performance estimates, hold-out estimates may be computed several times by creating different datasets, randomly sampling the original one in order to build the training and the test set. The average and the standard deviation of the accuracy is then computed from all the test datasets created.

An alternative method which is more elaborate is cross-validation. In this case the initial dataset is split up into *n* different subsets containing (approximately) the same amount of instances. At each step the testing dataset will be chosen as one of these subsets and the rest will represent the training set for the classification process. The cross-validation estimate of the accuracy is given by the average of the test results collected using every test fold. The subsets of data are created with the same class distribution present in the original dataset, just randomly or simply by small modifications from each other. In the latter case, the cross-validation can be stratified.

Finally, another feature of the Weka tool is the possibility to cluster the dataset grouping different individuals with the same features, extract possible rules between attributes and visualising every combination between the attributes and the class [118].

2.6.2 R Language

R is one of the languages that have been used in the course of this study. It is a general purpose informatics tool which has been created to perform calculations and graphics for statistical applications. It is a GNU project and it has been developed initially by Robert Gentleman and Ross Ihaka at the Statistics Department of the University of Auckland. Since 1997, a large group of people has been giving valuable contributions to the development of this useful tool. It has been inspired by the introduction of the S environment, created in 1988 by John Chambers and al. at Bell Laboratories [119] which explains many similarities and why much code written for S runs unaltered under R. It provides an Open Source route to statistical methodologies, creating good quality plots for publications, together with mathematical symbols and formulas. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including Linux, Windows and MacOS). Several packages are available for extensions of the basic software, through the CRAN family of Internet sites covering a very wide range of modern statistics. Moreover, as R, like S, is designed around a true computer language, it is possible for users to create new functions and procedures. For computationally intensive tasks, C, C++ and Fortran code can be linked and called at run time.

The reason for the selection of the R language in this Thesis is because it is an open source tool able to effectively handle and store data. It provides a range of operators for calculations on arrays (in particular matrices) and a large, coherent, integrated set of intermediate tools for data analysis and graphical display. It is finally a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

2.7 Summary

This review has aimed to provide a general overview of the core upon which this research work has been developed. Genetics is a field in continuous expansion which is therefore attracting the attention of an increasing number of researchers and a growing amount of money for studies. Thanks to the continuous optimisation of data gathering techniques, human beings are more and more interested in discovering the complex and intriguing rules hidden in their genetic heritage. Bioinformatics was born with the aim to provide researchers with original and versatile tools in order to achieve this challenging goal.

In the particular field of disease analysis for genetic susceptibility, lots of important contributions from all over the world have been published in relevant books and journals. Part of this Chapter is dedicated to the relevant achievements for an important disease, namely pre-eclampsia (PE). The main issue identified from the current literature review is the limitation of the tools employed for PE studies in terms of database size. Additionally more extensive research needs to be performed to validate or reject the current findings that appear sometimes contradictory or lacking in evidence. New improved tools need to be implemented in order to help research community in this task for coping with the continuously increasing size of genetic datasets. For this reason the majority of work in this Thesis is developed through experiments and tests applied to PE medical datasets.

Extensive analysis of the literature highlights the different research approaches carried out for disease association studies and the different tools which have been used for this purpose. In particular, the drawbacks of the current TDT tools solution are discussed and a possible improvement of the TRANSMIT software is proposed in order to provide the doctors with a tool able to select a smaller amount of SNPs associated with the disease under analysis. This improvement provides also additional information that can be used for further validation purposes, as described in Chapter 3. Furthermore, the limitations of the application of a single technique for case-control analysis with decision tree algorithm has led to the development of a new framework based on a combined analysis of multiple decision-tree algorithms in order to provide a more robust result, as shown in Chapter 4.

Moreover, within the specific feature selection techniques employed for the reduction in size of SNPs datasets, the application of LD function has resulted in a very flexible and versatile tool for detecting redundancy between SNPs. Unfortunately, however, this technique suffers from a serious limitation limitation for the analysis of large databases, as discussed. For this reason, in Chapter 6 a new proposed improvement of the LD function is shown and evaluated.

This material provides the reader with a solid base on which to understand the scope of this work and the reason for the choices that have been made along the way. In conclusion then, this Chapter provides an overview of the hypothesis that brought the realisation of the final comprehensive framework presented in Chapter 7.

Chapter 3

TDT : The TRANSMIT Software and the Proposed Optimization

3.1 Introduction

The growth of genomic information has increased the interest in gene-disease association studies. Within the different ways to approach this problem, the family based analysis is one of the most commonly used when the data is composed of individuals belonging to family groups. In particular, the Transmission Disequilibrium Test (TDT) is a successful technique for the analysis of genetic family based data.

In Chapter 2 an overview of the main tools used in the application of TDT is given. In this Chapter, attention is focussed on the TRANSMIT software because it differs from similar solutions in that it can deal with transmission of multi-locus haplotypes, even if the phase is unknown, and parental genotypes may be unknown. This tool, which has been widely used for genetic studies [120–122] is hereby analysed, discussed and assessed through experimental research. After highlighting the limitations of this algorithm, an optimisation of TRANSMIT is proposed through a multiple-test analysis of genetic information for the assessment of disease susceptibility. The results that emerge from the analysis of a medical dataset of pre-eclampsia are shown and discussed.

3.2 Epistasis Discovery

A topic of current interest in genetic data analysis for association studies is the non-linear interaction between genetic information. This phenomenon, better known as epistasis, is defined as a masking effect due to an allele at one locus preventing a second allele at a different locus from manifesting its effects [123]. Different methods for epistasis detection vary according to different types of analysis (association or linkage) and different type of trait (quantitative or qualitative).

Examples that have been successful in the literature are probabilistic graphical models, in which a graph denotes the conditional independence structure between random variables [124]. One of the most commonly known graphical representation of distributions is, for instance, the Bayesian network which has been used in linkage analysis [125, 126]. This is also known as a directed acyclic graphical model as it represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). There are different efficient algorithms that perform inference and learning in Bayesian networks. The use of probabilistic models together with the algorithms used to induce these models from data represents a relevant contribution in the application of evolutionary algorithms. Focused on the solution of the linkage learning problem, probabilistic models aim to explain the interactions between the components, taking support from the theory of graphical models. These models can represent a priori information about the problem structure, allowing a more efficient search of optimal solutions [127].

Evolutionary search algorithms for solving high-dimensional optimisation problems are an alternative and successful methodology able to cope with the current large amount of data available. They have become one of the best choices for many of the various Bioinformatics problems. An example of a successful evolutionary approach able to perform Bioinformatics optimisation tasks is the Estimation of Distribution Algorithms (EDAs). These represent a specific class of evolutionary optimisation algorithms, based on estimation of and sampling from probabilistic models and developed as a natural alternative to genetic algorithms in the last decade [128]. In the specific field of epistasis discovery, EDAs have been used to carry out a fine-grained stochastic search for optimisation purposes [129–131].

The TDT implementation, discussed in this Chapter, and relative extensions [132], are one of the various approaches that focus on the issue of association testing but can be used to detect and allow for epistasis in the specific field of family based studies [133]. Therefore, attention is focused on the TDT technique and its implementation with TRANSMIT software as it is a commonly used and appreciated tool in the specific field of family based analysis.

3.3 TRANSMIT Software

The TRANSMIT software, implemented by David Clayton at the University of Cambridge in 1999, is an algorithm which tests for association between genetic markers and diseases by examining the transmission of markers from parents to affected offspring. There are several tests that have been proposed for this kind of analysis but TRANSMIT differs from other similar programs because it can deal with transmission of multi-locus haplotypes, even if phase is unknown and parental genotypes may be unknown. The phase is defined as the arrangement of alleles at multiple loci on homologous chromosomes. For example, in a diploid individual with genotype Aa at one locus and genotype Bb at another locus, possible linkage phases are BA/ba or Ba/bA, where '/' separates the two homologous chromosomes [32].

A description of this software can be found in Clayton documentations [67] and is summarized in this section. This technique creates a score vector which is composed by two elements: observed transmissions of a certain haplotype to affected offspring and expected transmissions under Mendelian inheritance. When transmission is uncertain, this vector is averaged over all possible haplotype assignments to parents and offspring, using weights proportional to the probability of each assignment. Data from unaffected siblings (or siblings whose disease status is unknown) may be used to narrow down the range of possible parental genotypes which need to be considered. The program produces the two asymptotic chi-squared tests:

(1) For each haplotype or allele, a test on one degree of freedom (1-df) for excess transmission of that haplotype.

(2) A global test for association on $H-1$ degrees of freedom, where H is the number of haplotypes for which transmission data are available.

If there are rare haplotypes in the analysis, the approximate chi-squared distribution of test statistics will not hold. Two flags are available to protect against this. One causes aggregation and renumbering of alleles before the haplotype construction, the other one simply omits rare haplotypes from tests. The former approach inevitably results in some information loss but, when parents are missing, it may reduce the number of possible parental haplotypes that must be considered, and so reduce the computational time.

A good guideline to check how common a haplotype must be for the chi-squared tests to be used legitimately is to look at the table of 'observed' (O) and 'expected' (E) transmissions. If there are N heterozygous parents carrying a specific haplotype then, under the null hypothesis, the haplotype is expected to be transmitted $N/2$ times. The variance of $(O-E)$ will then be $N/4$. Therefore, an equivalent number of fully informative transmissions is given by multiplying by four the tabulated value for Var$(O-E)$ in the TRANSMIT output. A widely used guideline for the applicability of chi-squared tests is that they should only be used when all expected frequencies exceed five. This would correspond to ten fully informative transmissions and to a value of 2.5 for Var$(O-E)$. In the most recent version of the program a bootstrap test procedure is implemented, and this should be more accurate than the chi-squared approximations [67].

3.4 The Standard Experimental Approach

Following, an experiment with TRANSMIT software is shown and discussed in order to highlight the limits of this method and introduce thus a new and more efficient way to apply the TDT for SNPs analysis.

3.4.1 Experimental Data

In this study, the TRANSMIT software has been run with a pre-eclampsia dataset composed of 2,500 individuals coming from different clinics spread around the United Kingdom: Glasgow, Newcastle, Leeds, Nottingham, Leicester, Stoke, Birmingham, Oxford, Cambridge and London. Beside the clinical information, the original database contains genetic data on eight different SNPs that medical staff assumes to be potentially related to this disease. Each one of these SNPs can have two different kinds of allele, which are represented with the number 1 and 2 (see Table 3.1). As ‘C’ pairs with ‘G’, and ‘A’ with ‘T’, there are only two base-pair alternatives ‘CG’ or ‘AT’: hence a given SNP is *either* ‘CG’ or ‘AT’. Of these alternatives, one will naturally be more common than the other — 1 encodes the most common allele, while 2 encodes the rare allele.

The data input file for TRANSMIT should contain, for each person, the information displayed in Table 3.2, for this reason, all the clinical variable have been erased. In Table 3.2 the identify of the mother and the father must have the same family code. A sex attribute of 1 stands for male, a ‘affected disease status’ of 1 represents unaffected, 2 represents affected and 0 represents unknown; furthermore, ‘a’ and ‘b’ are the values of the alleles. In the input file, alleles must be coded as consecutive integers, with 0 representing unknown. Thus 0/0 represents completely missing data but, when each allele can have two value, 2/0 represents either 2/1 or 2/2. A particular observation is necessary for SNPs

Table 3.1: Description of the eight SNPs included in the initial dataset.

SNP	**Name**	**Allele 1**	**Allele 2**
SNP1	4072	T	C
SNP2	-1074	G	T
SNP3	3889	C	T
SNP4	172	C	T
SNP5	676	G	A
SNP6	1035	G	A
SNP7	6066	C	A
SNP8	11535	C	A

Table 3.2: Information contained in the input data for TRANSMIT.

Attributes	**Type**
family or pedigree code	*alphanumeric*
person’s identifier within family	*alphanumeric*
identity of father	*alphanumeric*
identity of mother	*alphanumeric*
sex	$1,2$
affected disease status	$0,1,2$
markers	a/b

which occur in the sex chromosome. Within the 23 pairs of chromosomes, there is one which is related to the sex of the individual. In the female population these chromosomes are the same (XX) whereas in the males population they are different (XY). For this reason, SNPs which come from the sexual chromosomes need a different codification. For instance if the SNP is X-related, which means it is present only in the X chromosome, the possible SNP values for a female are: 1/1, 1/2 and 2/2; whereas for a male the SNP values are 1/0 or 2/0. For markers on the X chromosome, males should have phenotypes coded a/0 or a/a, so that males and females have equal length records.

Although these data must appear in the order specified in Table 3.2, persons need not appear in the file in any particular order. Parents must be included in the data file even if no data concerning them are available; such entries are necessary to correctly identify relationships. Persons who appear on the data file only as parents do not need to have valid entries in the 'mother' and 'father' fields and their disease status may be coded as 0, as it is not used by the program [67].

3.4.2 Experimental Results

In the first experiment, the original database was pre-processed, in order to eliminate the clinical variables, and passed to the system. Interpreting the obtained output, it is quite clear that this kind of software can extract significant information from the database. In Figure 3.1, the first column shows the list of the haplotypes transmitted to the babies of mothers with pre-eclampsia and for each of them is shown the observed, expected occurrences, the variance and the χ^2 value. Each haplotype is composed by eight numbers, one for each SNP which in turn can take the value 1 or 2 (alleles). The first number of the haplotype sequence corresponds to the first SNP and the last number corresponds to the eighth SNP. The results show that there is an haplotype which seems to have a strict association with the disease as $\chi^2 = 4.20$ (df = 1). Following the interpretation of the results obtained in this manner, we decided to extend the analysis by introducing a multiple-test analysis of different combinations of SNPs, as this may provide new and interesting information about their association with the disease. In this way it is possible to provide the doctors with a table showing all the possible combinations of SNPs and their significance measured by the value of the χ^2 parameter which comes out from the statistical test applied to each haplotype transmitted from affected mother to affected child.

The original database is made by a list of attributes which includes the genetic information composed by a fixed sequence of different SNPs and in the first study performed these genetic data have been analyzed in a simultaneous manner. In the subsequent stage, additional data sets were formed by taking all the combinations of, initially, seven SNPs from the original eight SNPs, and then by iteratively reducing the number of SNPs. When each data set of SNPs had been extracted from the original database, they were then analyzed by the TRANSMIT software independently. In Figure 3.2, the results show that the significance of the analysis is altered depending on the selection of SNPs. This observation opens up a new method of experimentation.

3.5 The Improved TRANSMIT Approach

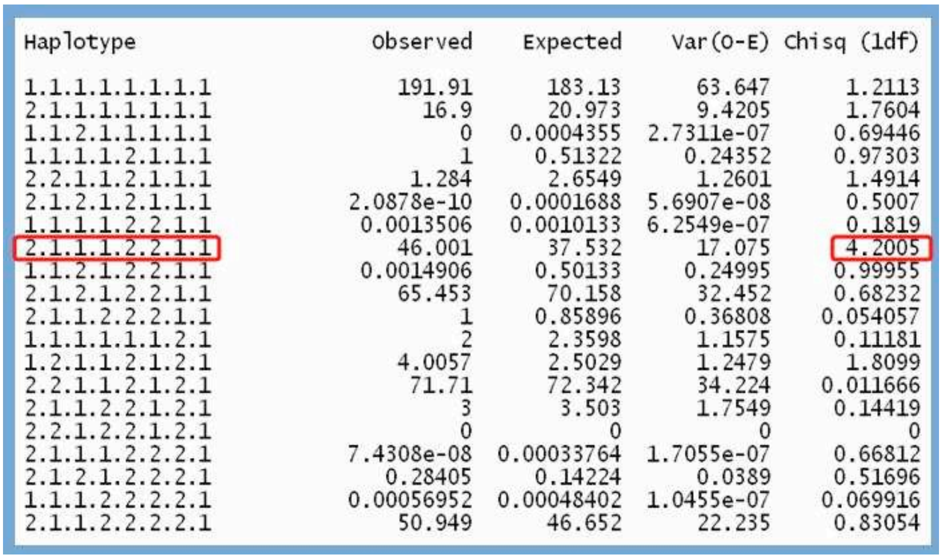
The original version of TRANSMIT was designed to analyze a string of SNPs in a single simultaneous analysis. In order to improve the performance of the software for a more extensive prospective assessment of the problem, a multiple-test analysis was implemented. The idea was to build different datasets from the original one and test the software with these new inputs in order to analyze different subsets of the available SNPs. The proposed methodology is to create an amount of databases subsets with all the possible combinations of 7 SNPs, 6 SNPs, 5 SNPs and so on.

3.5.1 Experimental Data

Given a set of n SNPs, the total number of all possible combinations of i SNPs ($i = 1 \ldots n$) taken from the set of n is

$$\sum_{i=2}^{n-1} {}^{i}C_n = \sum_{i=2}^{n-1} \frac{n!}{i!(n-i)!}.$$

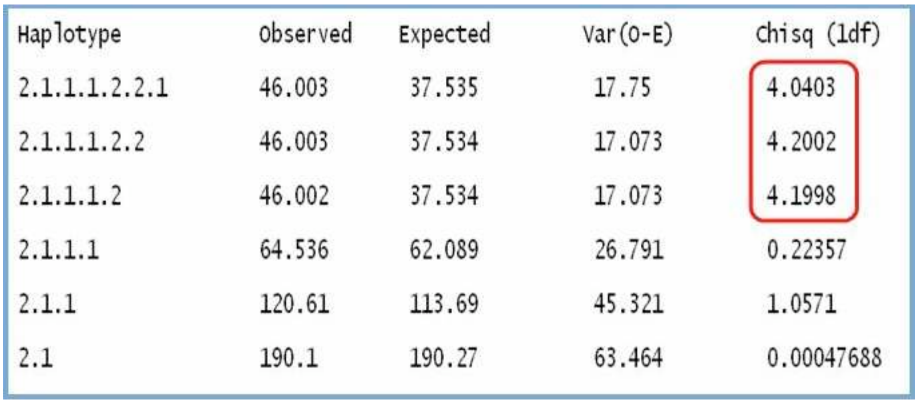
In our case, given $n = 8$, there are 246 combinations of data set to be analysed. The evaluation of the output provided from these 246 input sets may provide new information related to the association between the disease and a single or a set of SNPs.



```
Haplotype                    Observed     Expected      Var(O-E) Chisq (1df)

1.1.1.1.1.1.1.1                191.91       183.13        63.647      1.2113
2.1.1.1.1.1.1.1                  16.9       20.973        9.4205      1.7604
1.1.2.1.1.1.1.1                     0    0.0004355   2.7311e-07     0.69446
1.1.1.1.2.1.1.1                     1      0.51322       0.24352    0.97303
2.2.1.1.2.1.1.1                 1.284       2.6549        1.2601      1.4914
2.1.2.1.2.1.1.1            2.0878e-10    0.0001688   5.6907e-08      0.5007
1.1.1.1.2.2.1.1             0.0013506    0.0010133   6.2549e-07      0.1819
2.1.1.1.2.2.1.1                46.001       37.532        17.075      4.2005
1.1.2.1.2.2.1.1             0.0014906      0.50133       0.24995    0.99955
2.1.2.1.2.2.1.1                65.453       70.158        32.452    0.68232
2.1.1.2.2.2.1.1                     1      0.85896       0.36808   0.054057
1.1.1.1.1.1.2.1                     2       2.3598        1.1575    0.11181
1.2.1.1.2.1.2.1                4.0057       2.5029        1.2479      1.8099
2.2.1.1.2.1.2.1                 71.71       72.342        34.224   0.011666
2.1.1.2.2.1.2.1                     3        3.503        1.7549    0.14419
2.2.1.2.2.1.2.1                     0            0             0          0
2.1.1.1.2.2.2.1            7.4308e-08   0.00033764   1.7055e-07     0.66812
2.1.2.1.2.2.2.1               0.28405      0.14224        0.0389    0.51696
1.1.1.2.2.2.2.1            0.00056952   0.00048402   1.0455e-07   0.069916
2.1.1.2.2.2.2.1                50.949       46.652        22.235    0.83054
```

Figure 3.1: First result from TRANSMIT on sequences of eight SNPs.



```
Haplotype        Observed    Expected      Var(O-E)     Chisq (1df)

2.1.1.1.2.2.1    46.003      37.535        17.75          4.0403

2.1.1.1.2.2      46.003      37.534        17.073         4.2002

2.1.1.1.2        46.002      37.534        17.073         4.1998

2.1.1.1          64.536      62.089        26.791         0.22357

2.1.1            120.61      113.69        45.321         1.0571

2.1              190.1       190.27        63.464         0.00047688
```

Figure 3.2: Result from TRANSMIT on sequences of different amounts of SNPs.

3.5.2 Experimental Results

Table 3.4 and Table 3.5 show the results obtained from the analysis of all these 246 input files. The list contains all the sequences of SNPs that have been transmitted from the affected mothers to the affected babies and that have a significant χ^2 test within 1 degree of freedom. If we consider a p value smaller than 0.05, we need a χ^2 value greater than 3.84 in order to indicate an association between the haplotype and the disease. In each column the value of the correspondent SNP is displayed and this can be 1 or 2 as in these cases each SNP has two different alleles. The last column represents the maximum value of the global χ^2 with the respective degrees of freedom. At a first sight we can easily notice that for every single haplotype the χ^2 is greater than the minimum threshold,

Table 3.3: Rsults form TRANSMIT - Example of two haplotypes with the SNP 1, 3, 6 and 7.

Haplotypes	Observed	Expected	Var$(\mathbf{O}-\mathbf{E})$	χ^2
2.1.1.1	19.615	26.968	12.531	4.31
2.1.2.1	47.407	38.608	17.42	4.44

whereas the global χ^2 for each test is under the minimum value. From the biological point of view this means that there isn't enough evidence to reject the null hypothesis which claims no-association between the disease and the SNPs under analysis.

There is another interesting observation which comes from the analysis of these results. In the test performed with four SNPs there are two different versions of the same sequence of SNPs that seem to be significant: details are shown in Table 3.3.

The two haplotypes shown in Table 3.3 refer to the sequences of SNP1, SNP3, SNP6 and SNP7 (read from left to the right). If we focus our attention on SNP6 (the third one in the haplotype sequence) we notice that the test on both its alleles are significant. For allele 1, the observed occurrences are clearly less than the expected values, which means this allele is not associated to the disease. This is also highlighted by the results on allele 2, for which the occurrences are evidently higher than the expected ones. These two results support each other in the confidence to reject the null hypothesis.

3.5.3 Scalability

In terms of scalability, the maximum number of persons and families that are set by default for this software are 5000 and 1000, respectively. No explicit information is given by the TRANSMIT documentation on the maximum number of genetic markers to be analysed. However, the general TDT approach is acknowledged to have some limitations when the analysis is performed on large datasets of SNPs. Furthermore, the technique shown in this Chapter requires the creation of a number of subsets of a dataset that grows quickly as the number of SNPs to be analysed increases (see 3.5.1). Even a few hundred SNPs could require an extremely long time to be analysed. For this reason, the proposed technique is supposed to be applied at an advanced step of the genetic analysis work-flow, after the initial dataset has been pre-processed for elimination of SNPs that are not relevant to the analysis. A few dozen SNPs could be considered a dataset of reasonable size to provide

Table 3.4: Results from TRANSMIT on all possible combinations of sequences of SNPs.

SNP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	χ^2	**P-value**	**Max Global χ^2**
2 SNPs			1			2			4.30	0.04	7.17 2df
	2		1			2			4.54	0.03	12.61 12df
		1	1		2				3.98	0.05	
		1	1			2			4.28	0.04	
3 SNPs		1	1					1	4.13	0.04	
			1		2	2			4.38	0.04	
			1			2	1		4.14	0.04	
			1			2		1	4.31	0.04	
			1			2	1	1	4.63	0.03	
			1		2	2		1	4.23	0.04	
			1	1		2		1	4.05	0.04	
		1	1			2		1	4.29	0.04	
		1	1		2			1	3.97	0.05	
4 SNPs		1	1		2	2			4.38	0.04	
	2		1			2		1	4.17	0.04	
	2		1			1	1		4.31	0.04	19.71 12df
	2		1			2	1		4.44	0.03	19.71 12df
	2		1		2	2			4.55	0.03	
	2		1	1		2			4.19	0.04	
	2	1	1			2			4.54	0.03	
	2	1	1		2				4.18	0.04	
			1		2	2	1	1	3.99	0.04	
			1	1		2	1	1	4.04	0.04	
			1	1	2	2		1	4.08	0.04	
		1	1			2	1	1	3.92	0.05	
		1	1		2		1	1	3.89	0.05	22.44 15df
		1	1		2	2		1	4.35	0.04	
		1	1	1		2		1	4.03	0.04	
		1	1	1	2			1	3.98	0.05	
	2		1			2	1	1	4.49	0.03	
5 SNPs	2		1		2	2		1	4.17	0.04	
	2		1		2	2	1		4.05	0.04	
	2		1	1		2		1	4.20	0.04	
	2		1	1		2	1		4.08	0.04	
	2		1	1	2	2			4.20	0.04	
	2	1	1			2		1	4.18	0.04	
	2	1	1			2	1		4.10	0.04	
	2	1	1		2			1	3.82	0.05	
	2	1	1		2	2			4.54	0.03	
	2	1	1	1		2			4.20	0.04	
	2	1	1	1	2				4.20	0.04	

Table 3.5: Results from TRANSMIT on all possible combinations of sequences of SNPs.

SNP	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	χ^2	**P-value**	**Max Global** χ^2
			1	1	2	2	1	1	4.09	0.04	
		1	1	1	2	2		1	4.08	0.04	
	2		1		2	2	1	1	4.20	0.04	
	2		1	1		2	1	1	4.19	0.04	
	2		1	1	2	2		1	4.21	0.04	
	2		1	1	2	2	1		4.13	0.04	
6 SNPs	2	1	1			2	1	1	4.09	0.04	
	2	1	1		2		1	1	4.09	0.04	27.50 23df
	2	1	1		2	2		1	4.17	0.04	
	2	1	1		2	2	1		4.10	0.04	
	2	1	1	1		2		1	4.20	0.04	
	2	1	1	1		2	1		4.20	0.04	
	2	1	1	1	2			1	4.20	0.04	
	2	1	1	1	2		1		4.20	0.04	
	2	1	1	1	2	2			4.20	0.04	
		1	1	1	2	2	1	1	4.08	0.04	
	2		1	1	2	2	1	1	4.20	0.04	
7 SNPs	2	1	1		2	2	1	1	4.10	0.04	
	2	1	1	1		2	1	1	4.20	0.04	
	2	1	1	1	2		1	1	4.20	0.04	
	2	1	1	1	2	2	1		4.20	0.04	22.25 22df
8 SNPs	2	1	1	1	2	2	1	1	4.20	0.04	28.84 32 df

relevant benefits from the application of this technique in disease association studies. According to these considerations, this method is included within the final framework proposed in this Thesis, as shown in Chapter 7, following a pre-processing stage for the elmination of redundancy in the SNPs. Other solutions can be found in the literature, based for instance on probabilistic graphic models [134], able to detect epistasis between SNPs, without the requirement for feature selection methods in a pre-processing stage.

3.6 Summary

In this Chapter, an approach to the analysis of SNPs association with the pre-eclampsia disease for family based datasets is presented. Table 3.4 and Table 3.5 provide clinicians with an exhaustive summary of the χ^2 test. This method extracts the list of the haplotypes (of different length) that have a significance and displays the global χ^2 value for every test performed. The value of this parameter shows the degree of confidence that should be kept in accepting or rejecting the null hypothesis. Once the candidate genes have been identified, it is possible from such a table to detect whether there are SNPs that keep the same allele for all the haplotypes shown in the list or whether there is any SNP which is always present in all the different combinations of SNPs (such as SNP3 in this example). For this reason, this is a useful tool for understanding whether the genetic data under study should be considered for further studies or should be eliminated from consideration immediately.

In the analysis of complex diseases, clinicians often have to deal with huge databases containing very many SNPs coming from different genes. Scalability limitations of this technique entail the application of a pre-processing stage for size reduction when the number of SNPs is more than a few dozen. This issue is addressed in detail in Chapter 6. Once a first stage of redundancy detection and elimination has been carried out, the technique presented in this Chapter can then be used to identify those SNPs that do not show any association with the specific disease being considered. The result of this is a small number of SNPS that may be associated with a disease, requiring further medical investigation. Hence, the attention of the clinicians can then be focused only on the genetic targets that contain the most significant information for the disease under consideration.

Chapter 4

Case-Control Study: A New Framework for the Application of Decision Trees to SNPs Datasets Analysis

4.1 Introduction

In the previous Chapter an example of a family based study has been shown through the utilisation of the popular TDT technique as implemented with TRANSMIT software. This is feasible whenever the data under analysis is composed of individuals belonging to the same family. When this is not the case, different techniques need to be employed. This Chapter is thus focused on an alternative method that is available to use when the population of the dataset is completely unrelated, that is a case-control analysis. An overview of this method based on decision trees is shown in the first section.

Following this, a new proposed methodology for undertaking association studies with SNPs data is shown through the three main streams of action: (i) a pre-processing step, (ii) assessment of statistical significance, and (iii) actual analysis of the results. This technique is based on the assumption that a previous stage to eliminate redundancy has been applied to the dataset in order to remove the variables which contain superfluous information and thus could affect the quality of the final results. The actual analysis is based on the application of three of the most commonly used decision tree algorithms with the aim of comparing and contrasting them. Before applying these algorithms, a complex preprocessing stage of the initial database is performed in order to encode attributes (where necessary), explore the data, treat missing and unbalanced data, and set parameters. The application of this technique to an example of a medical database containing heterogeneous information about a list of patients affected by pre-eclampsia (PE) is described and the results are shown.

In the concluding part of this Chapter, several experiments have been performed with the available SNPs dataset for PE in order to give more examples of the application of this new methodology. Additionally, many considerations about data and medical implications are discussed to complete the overview of the association study for PE. All the operations and programming for the processing of the data is performed using the R language. The final goal of this Chapter is both to propose a valid method for SNPs analysis as an alternative to the TDT method previously discussed and, from the medical point of view, to discover any possible association, either genetic or phenotypic, with the specific disease of pre-eclampsia.

4.2 Case-Control Analysis and Decision Trees

There are different models that have been used by researchers for studying general genotype-phenotype associations depending on the kind of application. Population-based and family-based strategies with their numerous extensions are all widely used to detect genes associated with complex diseases. Association studies are defined as “a gene-discovery strategy that compares allele frequencies in cases and controls to assess the contribution of genetic variants to phenotypes in specific populations” [32]. In particular, Case-control study implies the creation of two different groups among the population, the cases and the controls group. In PE, for instance, a population of mothers can be considered and they can be split into sick mothers and healthy ones. However, the problem can also be studied by considering different prediction variables, like for instance a clinical feature of the disease.

Within the general data mining tools, there is a sub-class of algorithm widely used for case-control analysis in SNPs studies: the decision tree algorithms [51, 135, 136]. These are based on classification trees to predict membership of cases in the classes of a categorical dependent variable. In the study shown in this Chapter, three of these algorithms are taken in consideration: ID3, ADTree and C4.5 [47, 48, 118, 137, 138]. The aim of the new methodology here proposed is to detect the best and more reliable results by comparing or contrasting the outcomes obtained from a variety of decision tree algorithms, identifying commonality between trees.

4.3 Methodology

This is a kind of progressive analysis through which significant results are detected in the first stage then deepened and possibly confirmed in the subsequent steps. This section is divided in three main parts, according to the methodology structure: the pre-processing of the dataset, statistical significance evaluation of the results and results analysis. All these parts are discussed step by step in order to allow the methodology to be fully described.

4.3.1 Pre-processing Methodology

In this section, the analysis related only to a specific dataset is shown. In particular, in this context, a database (DB) will be referred as a specific subset of records obtained from the original (entire) set of records. In general, the original set of data consists of one or more attributes of SNPs and one or more attributes of phenotypic information. The pre-processing steps are shown in Figure 4.1.

Choice of Attributes

Every time a new DB is created there are attributes that may be deleted as considered useless or not informative for that specific population. In the first stage, all the remaining attributes can be kept in order to detect any possible feature that is significant in this analysis. The results which are further obtained will remove the less informative attributes. Considering the final DB, different kind of analysis can be performed in this study: some of them consider all the attributes, other are focused only on the SNPs or phenotypic attributes.

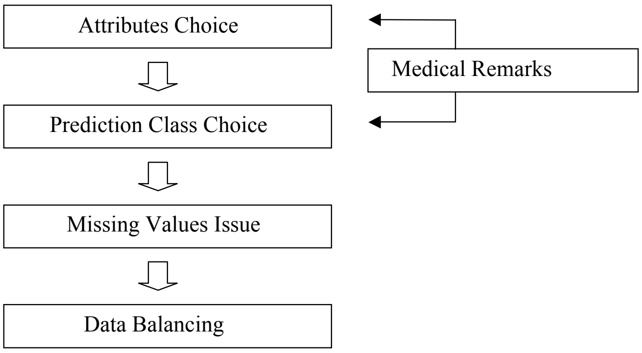


Figure 4.1: Sequence of steps to follow in the pre-processing of a general dataset composed of medical and SNPs attributes.

Sometimes within SNPs analysis, the set of data may include information about families. In this case, there can be some features coming from the relatives of the individuals under analysis (mothers or babies in the example of this study). This information can be easily transferred from the columns of one row to additional columns of a given individual, and hence analyzed as a new attribute. For instance, considering a dataset of only babies, new columns with the genetic information about their parents can be added.

Choice of Predictive Class

In general it is interesting to analyze the data considering different predictive variables for the same population. Many decision tree algorithms require the prediction class to be a categorical attribute and, more specifically, a Boolean one. If this is not the case, a threshold needs to be found in order to transform the variable into a Boolean one. As this analysis is a case-control one, only Boolean prediction variables are considered.

Consideration of Missing values

Some decision tree algorithms can deal directly with attributes containing missing values, whereas others cannot. It may be necessary to eliminate the missing values or to adopt another strategy such as imputation of missing data. It may not be appropriate to simply remove rows with missing values, in case there are many missing values and their deletion could therefore affect significantly the size of the DB. Attributes containing many missing values may still be dropped at a later stage. On the other hand, in the specific application in this study it is possible to keep the missing values as the algorithms chosen can deal with a codification for them.

Balancing of Data

The balancing of the DB is quite an important issue to be considered before performing the analysis. Often, the ideal situation is to have approximately half of the individuals belonging to the cases and half to the controls in order to have the least biased performance. If this is not the case, it is possible to create a new DB by selecting randomly a fixed number of people from both the groups.

Medical remarks

As the analyzed data is of clinical type it is useful to have feedback from the medical side throughout the whole process of analysis. There are attributes and prediction classes that have more relevance than others for the user, the doctor in this case. The final thresholds found for the prediction variable need to be considered from the medical point of view before defining them to be interesting. In general, any finding that can be significant from the statistical point of view it is not always meaningful for the medical community.

4.3.2 Statistical Significance Assessment

In this section an overview of the Kappa statistic is given as this is the parameter used for the assessment of the statistical significance of the results.

The Kappa statistic is a statistical measure of inter-annotators agreement for categorical items. It is typically used to assess the degree to which two or more judges, examining the same data, classify them in different exclusive categories. It is considered to be more robust than simple percent agreement calculation because it takes into account the agreement occurring by chance, even if there is some controversy upon the way chance can affect the judges [139]. A Kappa value equal to 1 indicates complete agreement and a Kappa value equal to 0 indicates agreement equivalent to chance. In general terms, the precision of a test is its ability to give results that do not rely on guess-work. Precision, as it pertains to agreement between observers, is often reported as a Kappa statistic [140].

Although there is some controversy upon the interpretation of Kappa values, a common reference scale defines a fair agreement the range of Kappa between 0.2 and 0.4 [141].

In this context, in order to assess the prediction accuracy of the algorithms employed, the Kappa statistic is calculated and used as the comparison between different algorithms and different experiments results, as it is the most commonly reported measure in medical literature [140]. However, the Kappa coefficient does not reflect sampling error and where it is intended to generalise the findings of a reliability study of a population of judges, the coefficient is frequently assessed for statistical significance through hypothesis tests.

Therefore in this work, in order to have a robust interpretation of the Kappa obtained from the empirical observations, appropriate tests are performed on the results. In particular in order to verify that the differences obtained in the Kappa value are statistically significant, tests for the analysis of the variance such as ANOVA or t-tests are applied to the results and the outcome is discussed. These tests are used to analyse the mean and variance of different populations in order to assess whether their means are equal. When only two populations are studied, a t-test is applied; when more than two populations are analysed, ANOVA can be used. In these experiments the setting of the critical p-value is kept to the standard value of 0.05.

4.3.3 Results Analysis Methodology

There are different kinds of algorithms for data mining that can be used but the essential idea is to perform repeated case-control analysis, each time defining the class and determining the subset of the original database to use. In this study, three decision tree algorithms which all work with a nominal class were used: ADTree [3], ID3 and C4.5.

The steps for the data analysis process are shown in Figure 4.2. For this study, the Weka software [118] was used to perform the analysis of the DB (note that in Weka the C4.5 algorithm is known as J48).

ADTree Analysis

In the first step the DB is analyzed with one of the three mentioned algorithms, arbitrarily chosen. In this study, for instance the ADTree algorithm is taken as first one. If the chosen predictive variable is a continuous one, it needs to be converted to a Boolean attribute.

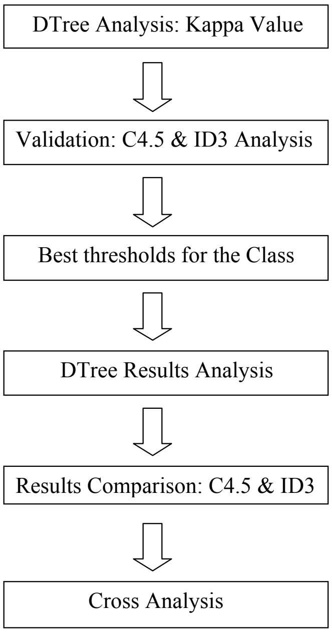

DTree Analysis: Kappa Value
Validation: C4.5 & ID3 Analysis
Best thresholds for the Class
DTree Results Analysis
Results Comparison: C4.5 & ID3
Cross Analysis

Figure 4.2: Sequence of steps to follow in the analysis of a genetic and clinical DB with the CBC as predictable variable. The applied algorithms are: ADTree, Id3 and C4.5.

Therefore, a range of thresholds has to be chosen for this class in order to detect the ones which give the most significant results. It is possible to select a fixed number of different thresholds, for instance 10, within the range of the variable and check for the reliability of the results.

All algorithm parameters are set to their default value (in Weka), as an exhaustive examination of the effects of varying parameters is outside the scope of this work. The validity of the analysis has been calculated through the use of the Kappa statistic and from the results it is easy to establish which thresholds provide a statistically significant result according to a Kappa value greater than 0.2. Then a further selection can be done by means of clinical feedback; there may be thresholds which don't have any particular medical meaning and other ones which correspond to medically accepted values.

In this analysis ten-fold cross-validation is used and repeated ten times with different seeds to create different random partitions of the data. The seeds can be chosen either arbitrarily or randomly. An average of the obtained results is then calculated.

Validation - C4.5 and ID3 analysis

In order to confirm (or not) these findings, the DB is processed with two other decision tree algorithms, C4.5 and ID3, with the same thresholds used in the previous analysis. These algorithms are also run with a ten-fold cross-validation using different seeds and the average and standard deviation of the Kappa value obtained from each result are calculated.

Determining Threshold for the Predictive Variable

In the next stage the focus is only on the data subsets whose thresholds give significant results (the ones for which Kappa is greater than 0.2) in order to check the Kappa trend. The algorithms that give the best results are run again and this time with the subsets of data whose class thresholds is included in the range previously found, choosing a number of different threshold values, for instance 10 or 20. From the results it is possible to detect the subset whose thresholds give a Kappa greater than 0.2.

Once that the thresholds have been chosen it is important to check the number of individuals involved in each test and the ratio of cases to controls in order to deal with a reliable test. If one of the final subsets has a case-control ratio above or below 50%, it could be not elegible for further analysis. Lots of studies have been published about the optimal case-control ratio and size of the dataset depending on the kind of application used in the analysis [142–145]. In this case, it is reasonable to consider a limit for the cases-controls ratio around one-third to be an acceptable one, as far as the amount of cases and controls don't fall below an arbitrary threshold of around 100 individuals.

ADTree Results Analysis

Focusing on those three versions of the dataset which give the best Kappa, it is now possible to analyze the results of the tests in order to determine whether they can be considered reliable. The first step consists of the comparison of the decision trees obtained from the DBs processed with ADTree, each one with a different class threshold. It can happen that, comparing the different trees, they have different shape and therefore different rules of classification but it is still possible to list the attributes which are present in all the trees obtained. The attributes can be detected from each node of the final classification tree.

Results Comparison - C4.5 and ID3

The same procedure is applied also to the second and the third algorithm, as far as they provide reliable results. For instance, for the set of trees obtained with ID3, each one with a different class threshold, a list of the attributes which are present in all the obtained trees is drawn up. In the end three lists of the attributes common to different class thresholds will be collected, each one from a different algorithm. Comparing these lists, a set of SNPs can be found, the ones detected from different algorithms and therefore which are present in all the results of this analysis.

Cross analysis

A cross analysis can now be performed between the decision trees obtained with the best algorithms, considering each time the same thresholds. For instance the ADTree trees obtained with specific class threshold can be selected and compared with the respective ones obtained from the ID3 algorithm. As before, a list can be drawn up with the attributes present in the results from different algorithms but with the same threshold. If there is a SNP which appears in all the lists created, it is more likely to have a reliable association with the predictable variable.

4.4 First Experiment: Babies Dataset

In this section an example of the application of this methodology to a real dataset related to PE is described.

4.4.1 Experimental Data

The DB under analysis contains 4529 instances and 105 attributes. The original dataset is composed of mothers, babies, fathers, grandparents and other relatives of the baby; there are fifty-two genetic attributes (SNPs) split across seven genes and fifty-three phenotypic (clinical) attributes, as follows:

(1) Genotype: 52 attributes:

- AGT gene: SNPs 1-8, alleles 1 and 2
- AGTR1 gene: SNPs 9-12, alleles 1 and 2
- TNF gene: SNPs 13-16, alleles 1 and 2
- F5 gene: SNP 17, alleles 1 and 2
- NOS3 gene: SNPs 18-22 and 24, alleles 1 and 2
- MTHFR gene: SNPs 25, 26, alleles 1 and 2
- AGTR2 gene: SNP 27

(2) Phenotype: 53 clinical attributes

- 5 concerning the individual's identity;
- 34 concerning maternal data, such as physical and physiological parameters, pregnancy details and current treatments;
- 6 concerning fetal data, such as the weight and gestational age at birth;
- 8 concerning the medical history of parents, partners or siblings of affected mothers.

The individuals of most interest for this disease are the mothers and the babies. There are actually four different conditions present in the original database: pre-eclampsia, eclampsia, other hypertensive diseases and normotensive (normal blood pressure). The only condition which is investigated in this study is pre-eclampsia.

4.4.2 Pre-processing Analysis

From the initial DB a subset is created containing only babies born from mothers with pre-eclampsia.

Attributes

In the first stage most of the attributes are kept. There are only a few attributes which are not meaningful when a database composed of babies is considered. These are the mothers features, such as blood pressure and blood test results.

Table 4.1: Prediction attributes for the babies

Attributes for the Babies	Type	Range
CBC	Percentage	$1-100$
Delivery gestation week	Integer	$22-42$ weeks

Predictive Class

The idea is to analyze the data considering different prediction variables as shown in Table 4.1 and in Table 4.2.

One of the most interesting variables listed in these tables is the ‘corrected birth-weight centile’ (CBC). This is the value of the weight of the baby at birth (as a percentage of the population) corrected for gestational age at birth, baby sex, ethnicity, mother’s height, mother’s weight and number of pregnancies. Hence, a CBC of 50 is the normal weight at birth, below this threshold it is considered underweight and a CBC exceeding this threshold is considered overweight. For each of these outputs different thresholds can be decided to define the cases and the controls in the dataset in order to perform a case control analysis. For instance the following values can be chosen: CBC $= 50$, Delivery gestation $= 35$, Systolic Pressure post partum ≤ 140, or Diastolic Pressure post partum ≥ 90

The results shown in this study are from a DB consisting only of babies, created from the original one by deleting the attributes considered not informative for a population of babies. The CBC attribute has been chosen as the predictive class and the final DB consists of 372 babies and 58 attributes. Beside the 53 SNPs listed above, there are six clinical variables for the babies: ‘Fetal disease status’, ‘Gestation at birth (weeks)’, ‘Gestation at birth (days)’, ‘Weight of the infant’, ‘Live at birth’ and CBC.

Missing values

Different trials were performed in order to understand if it is informative to retain the missing values or if their removal could have improved the study. The algorithms are applied to a dataset cleaned from the missing values and to a dataset with the missing values retained. The results obtained were the essentially unaltered, indicating that (for

Table 4.2: Prediction attributes for the mothers

Attributes for the mothers	Type	Range
CBC	Percentage	1 – 100
Delivery gestation week	Integer	22 – 42 weeks
Sys/Dias Pressure Post Partum	Integer	87 – 178
Highest Systolic	Integer	101 – 200
Highest Diastolic	Integer	65 – 150
Highest Proteinuria	Real	0.24 – 32.03
Highest ALT	Integer	2 – 875
Highest Urate	Integer	50 – 812
Highest Creatinine	Integer	49 – 990
Highest Urea	Real	1.6 – 33.8
Lowest Platelets	Integer	12 – 443

this data set) the missing values can be retained using the appropriate codification for the chosen algorithm.

Balancing of the data

As the CBC class is not Boolean, at this point it is not possible to balance the data because it is not yet clear the amount of cases and controls. Balancing of the data can be performed later, when a fixed CBC threshold is chosen and therefore the babies with a CBC greater than that threshold are considered as controls and these with CBC below that threshold considered as cases.

4.4.3 Statistical Significance Analysis: ADTree, C4.5 and Id3

As first step the DB is analyzed with the ADTree software from Weka. A range of thresholds has been chosen for the CBC class in order to detect the ones which give the most significant results. There are 9 different thresholds, from a CBC of 10 to a CBC of 90, and for each of them the Kappa value is calculated as shown in Table 4.3.

Table 4.3: Statistical results from ADTree: CBC=10-90

CBC Thresholds	10	20	30	40	50	60	70	80	90
Kappa	0.35	0.23	0.20	0.02	−0.03	0.02	−0.01	0	−0.01

From Table 4.3 it is clear that the first three thresholds (CBC of 10, 20, and 30) provide a statistically significant result (Kappa > 0.2) whereas the others have a quite low Kappa value. The ADTree algorithm is then run again with a set of 9 different seeds and the average and standard deviation of the results has been calculated as shown in Table 4.4.

Table 4.4: Kappa Average and Standard Deviation Over Nine Runs for ADtree, C4.5 and ID3 Algorithms

CBC Thresholds	10	20	30
ADTree Kappa	0.38(0.02)	0.32(0.03)	0.29(0.02)
C4.5 Kappa	0.27(0.03)	0.22(0.04)	0.28(0.05)
ID3 Kappa	0.15(0.04)	0.14(0.02)	0.18(0.04)

CBC Thresholds	40	50	60	70	80	90
ADTree Kappa	0.18(0.05)	< 0.07	< 0.04	< 0.04	< 0.05	< 0.04
C4.5 Kappa	0.18(0.04)	< 0.17	< 0.18	< 0.05	0	0
ID3 Kappa	0.03(0.02)	< 0.16	< 0.11	< 0.12	< 0.11	< 0.05

As a validation of these findings, the database was processed with the other two decision tree algorithms, C4.5 and Id3. The thresholds are the same used ADTree analysis. These algorithms have been run nine times with different seeds and the mean and standard deviation of the Kappa value has been calculated as shown in Table 4.4 and Figure 4.3.

In order to verify that the differences obtained from these experimental results are statistically significant, an ANOVA test has been performed on the results from the three algorithms, considering the CBC thresholds 30 and 40. Setting the critical p-value to 0.05, in all cases a very low p-value was obtained ($p = 1.86 \times 10^{-5}$ for ADTree, $p = 2.3 \times 10^{-4}$ for C4.5, $p = 3.06 \times 10^{-6}$ for Id3). From these results it is clear that the

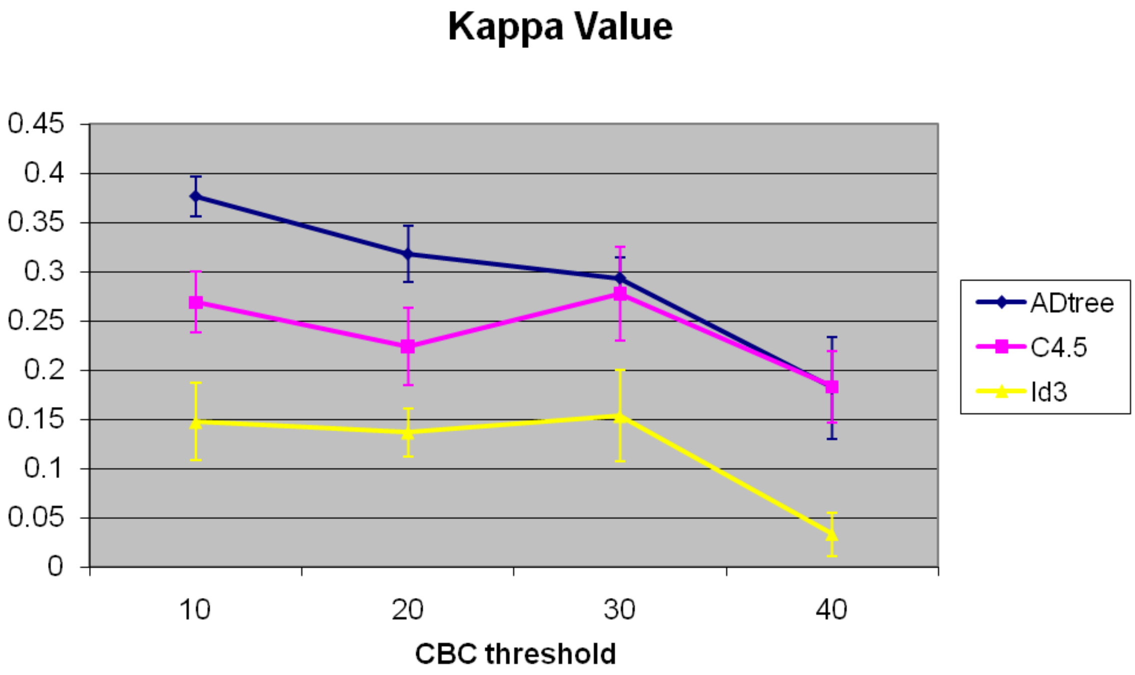


Figure 4.3: Kappa Values (means and standard deviations) of the three applied algorithms (ADTree C4.5 and Id3) versus weight of the baby expressed as CBC within the range CBC= 10-40.

null hypothesis can be rejected and the actual differences that can be noticed between different CBC threshold settings are statistically significant. In particular, the difference noticed between the CBC threshold 30 and 40 highlight a cut-off point for choosing a CBC threshold of significance for this analysis. Regarding ID3, no significant results have been obtained over the thresholds as shown in Table 4.4. Thus, ID3 does not appear to be able to detect relevant findings in this application, further investigation on this limitation is left for future studies.

As the CBC thresholds of 10, 20 and 30 have shown to be relevant, in the next stage interest is focused only on the CBC range between 4 and 30, in order to detect any other threshold with a high Kappa. The two algorithms that give the best results are run again, this time with 14 different thresholds of CBC in the range $4-30$.

In Figure 4.4, the mean and standard deviation of the Kappa value are shown for each different CBC threshold and for ADTree, C4.5 and the average between the two algorithms. This figure shows that the three thresholds with the best Kappa value are 6, 10 and 28. The fact that the trend in Kappa is not monotonic with CBC may be due to the

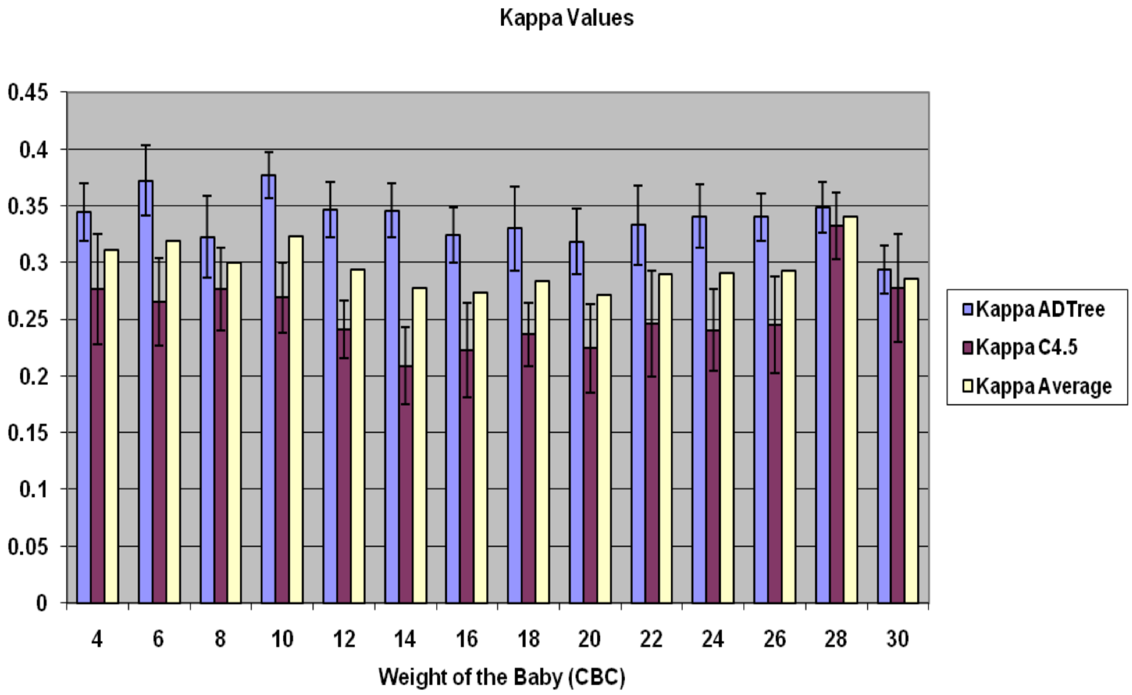


Figure 4.4: Kappa Values of the two applied algorithms (ADTree and C4.5) versus weight of the baby expressed as CBC within the range CBC= 4-30.

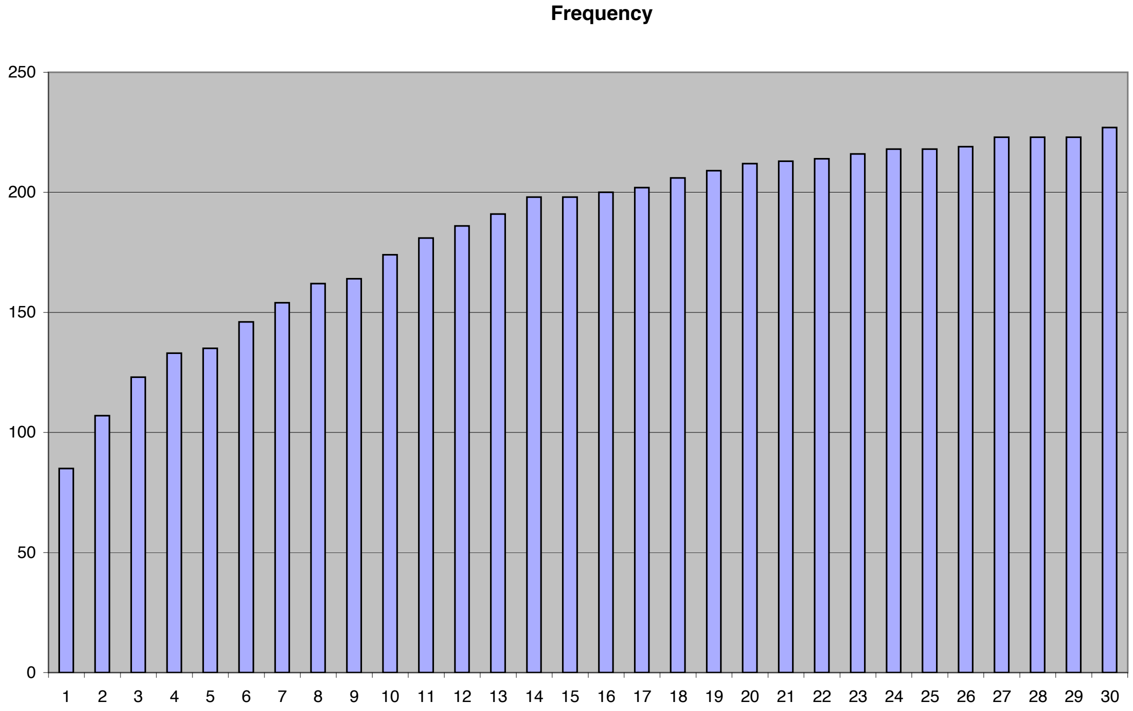


Figure 4.5: Number of cases versus the CBC of the babies.

presence of noise in the data but could also be due to the complex correlation between attributes such as CBC and week of delivery (later discovered). In order to verify whether the differences observed in this figure are statistically significant, an ANOVA test is performed for each of the two algorithms considering the 14 groups, each one with a different CBC threshold. The result of the ANOVA test with a critical value alpha = 0.05 shows that the null hypothesis of equal means between the 14 different groups can be rejected (p = 2.31×10^{-7} for the ADTree and p = 2.14×10^{-8} for C4.5). However, from this test it is not clear if all the distributions of Kappa are different from each other or if only some of them are different from the rest. In order to check if the differences between the three highest thresholds are significant, another ANOVA test has been performed between each of these Kappa distributions and the rest of the Kappa distributions obtained with the rest of the thresholds (df = 1). For CBC= 6, p = 0.001 was obtained for ADTree and p = 0.48 for C4.5. For CBC= 10, p = 0.0002 was obtained for ADTree and p = 0.34 for C4.5 whereas for CBC= 28, p = 0.32 was obtained for ADTree and p = 6.46×10^{-8} for C4.5. These results also appear evident from Figure 4.4. The CBC threshold of 6 and 8 are the highest for ADtree and only a few Kappa means from other CBC thresholds results fall in the range of ± one standard deviation. The same can be seen for C4.5 but this time for the threshold CBC = 28. In conclusion, ADTree indicates a CBC threshold equal to 6 and 10 as statistically different from the others, while C4.5 indicates a CBC threshold of 28. Although in this case there is not strong agreement between the two algorithms, the final three thresholds highlighed provide a significant Kappa for both of the algorithms. The analysis is carried out focusing on these three relevant thresholds in order to show an example of the application of the methodology proposed here. In general terms, the application of this statistical assessment methodology allows the validation or rejection of any type of result that is not revealed to be significant.

Case-Control Ratio Checking

Once the CBC values are fixed, the balancing of the data can be checked. The number of cases and controls involved in each test results are shown in Figure 4.5. In particular:

- for CBC = 6: 147 cases (39.5%) and 225 controls

- for CBC = 10: 177 cases (47.6%) and 195 controls
- for CBC = 28: 243 cases (65.3%) and 129 controls

These results are acceptable regarding both the absolute size of the population (372) and the proportions of cases and controls, as the case-control ratio is above 0.33 and there are more than 100 individuals for each group.

4.4.4 Results Analysis

ADTree Results

Focusing on these three versions of the DB, the next step consists of the comparison of the three decision trees obtained from the three DBs processed with ADTree. Comparing the three different trees it is clear that they have different shape and therefore different rules of classification but it is possible to list the attributes common to all of them. Besides 'gestational week at birth'(which is always present), the attributes found for CBC equal to 6,10 and 28 are shown in Table 4.5. In the last column, the attributes common to the first three columns are shown. It is important to notice that final solutions appeared to be very stable as every (common) attribute to the three different CBC thresholds appeared 100% of the time in every run.

Results Comparison — C4.5 and ID3

The same procedure is applied also to the C4.5 algorithm as it provided similar results. The list of the SNPs are shown in Table 4.6 and Table 4.7. Concerning the clinical variables, there is still the attribute 'Gestational week at birth' which is common to the algorithm results and the variable 'sex' which is present only in the first two thresholds (CBC = 6 and CBC = 10) . Once again, all the (common) attributes to the three different CBC thresholds appeared 100% of the time in every run.

Cross Analysis

A cross analysis can now be performed between two decision trees obtained with the two algorithms (ADTree and C4.5), considering each time the same thresholds (CBC = 6, 10,

Table 4.5: Common attributes to the three CBC thresholds 6,10 and 28 for the ADTree algorithm results.

			CBC			
Gene	SNP	Allele	6	10	28	All
AGT	1	1	*y*			
AGT	3	2			*y*	
AGT	6	2			*y*	
AGTR1	10	2		*y*		
AGTR1	11	2			*y*	
AGTR1	12	2	*y*			
F5	17	2	*y*	*y*		
NOS3	19	2		*y*		
NOS3	21	2	*y*	*y*	*y*	*y*
NOS3	24	2	*y*	*y*		
MTHFR	26	2			*y*	
AGTR2	27	2	*y*	*y*	*y*	*y*

Table 4.6: Common attributes to the three CBC thresholds 6,10 and 28 for the C4.5 algorithm results.

			CBC			
Gene	SNP	Allele	6	10	28	All
AGT	1	1		*y*	*y*	
AGT	1	2		*y*	*y*	
AGT	3	2	*y*	*y*	*y*	*y*
AGT	4	2	*y*			
AGT	6	1			*y*	
AGT	7	2			*y*	
AGT	8	1	*y*			
AGT	8	2	*y*	*y*	*y*	*y*
AGTR1	9	1	*y*	*y*	*y*	*y*
AGTR1	9	2	*y*	*y*		
AGTR1	10	1		*y*	*y*	
AGTR1	10	2	*y*			
AGTR1	11	1	*y*			
AGTR1	11	2	*y*		*y*	
AGTR1	12	1		*y*	*y*	
AGTR1	12	2	*y*	*y*	*y*	*y*
TNF	13	1	*y*	*y*		
TNF	13	2	*y*	*y*		
TNF	14	2	*y*	*y*	*y*	*y*
TNF	15	2	*y*			
TNF	16	1	*y*	*y*		
TNF	16	2			*y*	

Table 4.7: Common attributes to the three CBC thresholds 6,10 and 28 for the C4.5 algorithm results.

			CBC			
Gene	SNP	Allele	6	10	28	All
F5	17	2		*y*	*y*	
NOS3	18	2		*y*		
NOS3	19	1	*y*			
NOS3	19	2	*y*	*y*		
NOS3	20	1	*y*	*y*	*y*	*y*
NOS3	20	2	*y*			
NOS3	21	1			*y*	
NOS3	21	2	*y*		*y*	
NOS3	22	1			*y*	
NOS3	22	2	*y*	*y*	*y*	*y*
NOS3	24	1	*y*	*y*	*y*	*y*
NOS3	24	2	*y*			
MTHFR	25	1	*y*	*y*	*y*	*y*
MTHFR	25	2	*y*	*y*	*y*	*y*
MTHFR	26	2	*y*	*y*		
AGTR2	27	1		*y*	*y*	
AGTR2	27	2	*y*	*y*		

28). The results are shown in Table 4.8 and each attribute appears 100% of the time in every run.

Furthermore, if the attention is focused on the results when the CBC is 28 a new dataset composed of the common attributes found in the results from both the two algorithms can be created. These attributes are: ‘sex’ , ‘Gestational week at birth’, AGT SNP3, AGTR1 SNP11 and NOS3 SNP 21. Processing this new dataset with both the ADTree and C4.5 algorithm, two interesting rules which are common to the two final trees are found, with a statistical significance of $k = 0.38$ for C4.5 and $k = 0.41$ for ADTree. The first rule claims that male babies, born after the 35th week of gestation and with an AGT SNP3 allele2 of 1 have a good probability to have a normal weight ($CBC > 28$). The confidence of the C4.5 algorithm is measured by the ratio between the corrected classified instances over the uncorrected ones, which is 84/24. The ADTree measure of confidence is instead made by the ‘classification margin’, analyzed on prior work [146] and it has a absolute value of 1.29. The second finding shows that male babies, born after the 35th week of gestation and with an AGT SNP3 allele2 of 2 and an AGTR1 SNP11 allele2 of 1 have a good probability to be under weight ($CBC < 28$). For the C4.5 the confidence parameters measures 21/5 and for the ADTree the classification margin has an absolute value of 0.76.

Table 4.8: Common attributes to the three CBC thresholds 6,10 and 28 for the two algorithms: ADTree and C4.5.

Gene	SNP	Allele	CBC 6	CBC 10	CBC 28
AGT	3	2			*y*
AGTR1	11	2			*y*
AGTR1	12	2	*y*		
F5	17	2		*y*	
NOS3	19	2		*y*	
NOS3	21	2			*y*
AGTR2	27	2	*y*	*y*	

Following these results, the analysis is performed with only one attribute, the ‘delivery gestation week', and the CBC predictive variable. An association between these two parameters is found with a good significance as shown by the Kappa value of 0.4212 for both the ADTree and C4.5 analysis. The ADTree algorithm detects an interesting threshold for the 'GestationatBirthw' equal to 35.5 to discriminate the small babies (cases) from the normal one (CBC $>$ 10) and in C4.5 the threshold is set at 35 weeks of pregnancy. This means that babies delivered before 35 or 35.5 week of gestation are likely to have a CBC $<$ 10.

4.4.5 Discussion

The methodology shown in this study provides researchers with a guideline for data mining in the specific application of case-control analysis for SNPs. This technique may find an association between the SNPs and the disease or its phenotypes. However, it is also possible that the results don't show a significant direct connection between the SNPs and the disease as found in this study. In this case it is still possible to detect a reduced number of SNPs that may play an important role in the genetic association, as for example in this specific experiment.

From the methodological point of view, it is possible to conclude that thanks to this strategy, some attributes are rejected as not relevant for the analysis, the number of the instances are decreased and a set of attributes, clinical or genetic, are found to be correlated to the predictive variable, as show in the lists of the common attributes in the example described in this study, see Table 4.5, Table 4.6 and Table 4.7 and Table 4.8. From a comparison of these Tables it is also clear that there are SNPs such as AGT 2 and AGT 5 that never appear in the results; these SNPs can thus be ignored in further analysis.

From the clinical perspective, there are at least two important findings which emerge from this methodology. The first is the significance of the threshold CBC of 10. From the study on the validity of the thresholds three different values have been found: i.e. 30, 20 and 10. The feedback from the medical point of view confirmed the clinical importance of a CBC of 10 for babies affected by pre-eclampsia, as it is a clinically accepted threshold used to identify growth restricted babies, which have then a higher risk of problems in the neonatal period. The second finding is the dependency of the CBC on the 'week of delivery' parameter. In the formula for calculating the CBC, the birth weight is adjusted considering parameters including the ‘week of delivery’. This means that there shouldn’t be any association between these two attributes. From the results of this analysis on PE disease, an association between these two parameters has been found: women with pre-eclampsia who deliver before 35 weeks of pregnancy are more likely to give birth to babies with a CBC under the value of 10.

The proposed methodology provides (besides the opportunity to find new and challenging results) a useful tool for the screening stage where a reduction in the number of cases is the main goal. However, it is important to remark that a relevant assumption of this proposed technique is that the dataset is previously processed in order to remove all the possible attributes which appear to be redundant. If two different variables are very similar *and* they are both kept in the analysis, one of these variables may appear in one of the resulting trees and the other variable may appear in another resulting tree, because the algorithms might pick up only one them. In this case, as they do not belong in the common attributes list between different algorithms or between different CBC thresholds, they will never appear in the final list of relevant variables for the study. For this reason, in the final framework proposed in this Thesis, this technique is applied in a second stage, after the pre-processing of the data for redundancy elimination, see Chapter 7. In the following sections, the analysis is continued further in order to explore different ways of approaching the problem, still based on the proposed methodology. The considerations and remarks that emerge from this further analysis are of course highlighted for the study of the specific medical dataset for PE. However, they can be generalized for other disease studies.

4.5 Second Experiment: Mothers Dataset

Following the previous analysis, an important observation arises concerning the significance of considering the genotype of the mothers rather that the babies. It is still difficult and risky to collect information related to the DNA of babies in pregnancy, as this requires an invasive test. On the other hand such information can easily be collected from the mothers. For this reason, in this new research, the analysis will be focused only on the mothers, using the CBC of the baby as the predicted class. The clinical condition of the grandmothers when their mothers were born is another significant information which would be interesting to analyze. A heritable trend can be detected across the two generations validating the genetic association of this disease. Unfortunately, the current database does not contain sufficient information to perform this kind of analysis.

In this section another example of the application of this methodology to a real example dataset is described, still related to pre-eclampsia.

4.5.1 Experimental Data

At this stage the analysis is applied to a DB composed by only mothers with PE, representing 568 individuals with 54 clinical attributes (the previous ones together with the mother age) and 52 genetic ones (26 SNPs).

4.5.2 Pre-processing Analysis

Attributes

Considering an analysis of the mothers, there are a few attributes that can be dropped from the original DB, as not relevant for the study. These are the clinical information of the parents, partners or siblings of the affected women. From the 54 initial clinical variables, therefore a DB of 37 clinical attributes and 52 SNPs is built. The clinical variables are:

- 33 physiological attributes about the mothers: number of pregnancies, blood pressure, urine and blood tests, delivery and post-partum features and current treatment.
- 4 baby information as weight and gestation week at birth.

Predictive Class

In Table 4.2 all the potential predictive variables for the mothers are shown with their range of values. At this stage the study is focused only on the CBC, which has been chosen also previously for the babies analysis.

Missing values

Also in this case, different trials have been performed in order to understand if it is informative to retain the missing values or if their removal could have improved the study.

At the beginning the idea was to eliminate all the instances (individuals) containing any missing value. Under this hypothesis, the attention is focused on the attributes which are mostly composed by missing value: 'DiagnosisProteinuria', ' HighestProteinuria' and 'Complication'. In order to avoid a drastic reduction of the size of the DB these attributes have been deleted from the DB. Two more attributes have been under analysis for the large number of missing values: 'HighestAlt' and 'HighestUrea'. If the analysis is performed keeping these two attributes, a Kappa value equal to 0.31 (s.d. 0.02) is obtained for ADTree and equal to 0.22 (s.d. 0.04) for C4.5. If the analysis is preformed without these attributes, a Kappa equal to 0.27 (s.d. 0.04) is obtained for ADTree and a Kappa equal to 0.20 (s.d. 0.03) is obtained for C4.5, see Table 4.9 and Table 4.10 for ADTree and Table 4.11 and Table 4.12 for C4.5. In order to assess whether the differences observed between the Kappa obtained with and without attributes are statistically significant, a paired t-test is applied to the results for both ADTree and C4.5, after checking for normality using the Shapiro test [147]. The one-tailed t-test shows a p-value = 0.001 for ADTree and p-value = 0.01 for C4.5. The two-tailed test shows a p-value = 0.003 for ADTree and p-value = 0.03 for C4.5. Having set the critical p-value to 0.05 (df = 8), these results confirm that we can reject the null hypothesis, stating that the inclusion of the two attributes affects the statistical significance of the results, decreasing the Kappa for both ADTree and C4.5.

Table 4.9: Statistical results from ADTree run without 'Highest ALT' and 'Highest Urea' and with CBC=10

Seeds	1	10	20	30	40	50	60	70	80
Kappa	0.22	0.33	0.23	0.30	0.32	0.28	0.28	0.24	0.23

In a second stage, the idea is to keep all the attributes and all the instances with missing values, as they are coded by the algorithm. The algorithms are applied to the dataset cleaned from the missing values and to the dataset with the missing values retained. The

Table 4.10: Statistical results from ADTree run with ‘Highest ALT’ and ‘Highest Urea’ and with CBC=10

Seeds	1	10	20	30	40	50	60	70	80
Kappa	0.29	0.35	0.31	0.34	0.31	0.30	0.33	0.28	0.31

Table 4.11: Statistical results from C4.5 run without ‘Highest ALT’ (HA) and ‘Highest Urea’ (HU) and with CBC=10

Seeds	1	10	20	30	40	50	60	70	80
Kappa	0.14	0.21	0.16	0.21	0.22	0.21	0.19	0.22	0.17

Table 4.12: Statistical results from C4.5 run with ‘Highest ALT’ and ‘Highest Urea’ and with CBC=10

Seeds	1	10	20	30	40	50	60	70	80
Kappa	0.26	0.22	0.13	0.28	0.22	0.17	0.23	0.20	0.24

results obtained were once again unaltered, confirming that the missing values can be retained as the codification used by the chosen algorithm was appropriate. At this point then the DB is composed by 89 attributes (52 SNPs + 37 clinical variables).

Balancing of the data

In the first stage, after eliminating all the individuals which contain any missing value the balance of the data can be performed. Considering the CBC equal to 10, the number of cases and controls involved in the tests are:

- without HA and HU: 161 cases (47%) and 178 controls over 339 individuals.
- with HA and HU: 154 cases (48%) and 167 controls over 321 individuals.

These results are acceptable regarding both the absolute size of the population and the proportions of cases and controls, as the case-control ratio is above 0.33 and there are more than 100 individuals for each group.

In the second stage, all the attributes are kept, with the codification for the missing values made by the algorithm. In this case the initial DB was composed by 568 instances but as not all the babies have their clinical or genetic information, the final obtained DB is composed by 339 mothers each one with the information about the CBC of the respective baby. The balance of the DB results then the same as shown before in the option ‘without the HA and HU’

Recoding of SNPs

The re-codification of the SNPs is an important issue that needs to be considered with the genetic analysis. This consists of replacing the two numbers that characterize the information on the genetic marker specified by the given SNP with a single number, losing the information contained in every single allele.

This step is necessary whenever the relative position of the allele(phase) is unknown. In more simple terms it would be relevant to know whether a certain SNP has the value 1/2 or 2/1. If this information is not available, as in most cases, the SNPs need to be encoded considering a SNP 1/2 to be the same as a SNP 2/1. The Tables 4.13 and 4.14 show the way the SNPs need to be recoded in order to overcome this missing information.

It is clear that the the amount of genetic variable included in the analysis will be therefore halved.

Table 4.13: SNPs recoding for all the SNPS except sexual SNPs

SNPs	Allele1	Allele2	Coding
Same Allele	1	1	2
Different Allele	1	2	3
Different Allele	2	1	3
Same Allele	2	2	4

Table 4.14: SNPs recoding for sexual SNPs

Sex SNPs	Allele1 (X chr)	Allele2 (Y chr)	Coding
Male: Common Allele	1	0	2
Male: Rare Allele	2	0	4
Female: Same Allele	1	1	2
Female: Dif Allele	1	2	3
Female: Dif Allele	2	1	3
Female: Same Allele	2	2	4

4.5.3 Statistical Significance Analysis

Best Threshold(s) for the Predictive Variable

From the clinical point of view the threshold CBC = 10 has been found to have a relevant significance when a population affected by Pre-eclampsia is considered. For this reason the analysis will be focused only on this threshold, leaving aside the other 2 significant thresholds previously found (CBC = 6 and 28).

ADTree Analysis

In the first stage the DB is analyzed with the ADTree software from Weka. The ADTree algorithm is run with a set of 10 different seeds and the average of the Kappa has been calculated as shown in Table 4.15.

Table 4.15: Mothers Dataset - Statistical results from ADTree: CBC = 10 over 10 different seeds

Seeds	1	10	20	30	40	50	60	70	80	90
Kappa	0.32	0.27	0.32	0.30	0.28	0.28	0.32	0.37	0.32	0.30

Validation — C4.5 Analysis

In order to have a validation of these findings, the DB has been processed also with the C4.5 algorithm. The analysis with ID3 is not considered at this stage because this algorithm works with all the nominal attributes. As in the mother DB there are plenty of numeric attributes this software will not be applied.

The C4.5 algorithm has been run ten times with different seeds and the average of the Kappa value has been calculated as shown in Table 4.16. From these results of Table 4.16 it is clear that with C4.5 a result similar to ADTree has been obtained with a significant Kappa.

Table 4.16: Mothers Dataset - Statistical results from C4.5: CBC = 10 over 10 different seeds

Seeds	1	10	20	30	40	50	60	70	80	90
Kappa	0.19	0.14	0.20	0.23	0.22	0.15	0.15	0.20	0.20	0.15

4.5.4 Results Analysis

Results Comparison — C4.5 and ADTree

Comparing the results obtained with the two software, it is clear that C4.5 gives a long list of attributes, whereas ADTree produces a simple tree, with a reduced number of nodes. There are actually no SNPs in common to these two trees and the clinical attributes which are common are: 'Week of Delivery', 'HighestALT', 'HighestUrea', 'HighestCreatinine', 'MaternalHeight'.

Medical Remarks

One of the most interesting information which it would be worth to analyze is the 'Onset of Disease'. The database could be split into two different sets of populations: the early onset (EOS) disease mothers, which are the ones that fell ill in an early stage of pregnancy and the late onset (LOS) disease mothers which are the one who fell ill later. The interesting remark about these two different groups is that PE that happens EOS is more related to the baby's genotype whereas the LOS disease is more due to the mother's genotype [148–150].

Moreover, when the tree is analyzed it is always important to check that the rules extracted from the tree have a consistency with the medical meaning.

About the SNP code another remark needs to be done. The Threshold SNP = 3 means heterozygote which corresponds to SNP = 1,2 or SNP = 2,1. The SNP $\neq$ 3 instead is not a very useful information as it can mean either SNP = 2 (1,1) or SNP = 4 (2,2). Only if the rare allele (2) is very rare, then SNP = 1 can be considered the most likely solution.

4.6 Mothers Dataset with Attributes Reduction

In this stage the second experiment has been repeated with a new DB cleaned from a few attributes which are not considered relevant for the analysis and could therefore affect the results. All the Post Partum attributes are eliminated and the information on the blood pressure is reduced to only two attributes out of eight (Booking Systolic and Diastolic, Diagnosis systolic/diastolic 1 and 2, Highest systolic and diastolic). Performing the analysis with the missing values with both the algorithms, a good significance of the tests is still preserved, considering the Kappa values averaged over 10 different seeds, see Table 4.17. In this table, from the comparison of the means and standard deviations, it is clear that there is no significant difference between the dataset with and without the missing values.

Table 4.17: Mothers Dataset - Kappa mean and standard deviation over 10 different seeds for ADTree and C4.5 test with and without missing values

Algorithms	ADTree	C4.5
without MV	0.32(0.03)	0.26(0.01)
with MV	0.33(0.03)	0.23(0.02)

In the Table 4.18 the common clinical attributes from the results obtained with ADTree and C4.5 with and without missing values (MV) are listed and these appeared 100% of the time in every run.In the ADTree analysis, the SNPs obtained without missing value are SNP 9 and SNP 19 whereas with missing value only SNP 24 is obtained. Regarding C4.5, the SNPs obtained without missing value are: 3, 4, 7, 10, 12, 14, 15, 16, 18, 19, 20, 22, 26, 27. The SNPs obtained with missing value are: 2, 4, 11, 13, 14, 15, 16, 17, 18, 19, 26. From all these results it is clear that SNP 19 is mostly present in the output, which could highlight the significant association to the PE in mothers datasets.

Following these results, different trials have been performed in order to understand better the association between the SNPs in the dataset with PE. In particular an analysis with only SNPs and one with SNPs together with gestational week parameter are performed. While the analysis with a DB of only SNPs does not give significant results as shown in Table 4.19, analyzing a DB of only SNPs together with delivery gestation week, a significant Kappa is obtained, see Table 4.20.

One interesting aspect of the C4.5 algorithm is the problem of over-fitting which may affect the prediction accuracy. This means that a decision tree, able to classify every single instance from the training set, is not necessarily better than a smaller tree that does not fit all the training data. In order to avoid this problem different solutions have been proposed in the literature, such as a stopping criteria for the tree growth or pruning from the less relevant branches. In C4.5, the pruning method is based on estimating the error

Table 4.18: Mothers Dataset - Common attributes to the ADTree and C4.5 algorithms with and without missing values for CBC = 10

	ADTree		C4.5	
Attributes	no MV	MV	no MV	MV
DeliveryGestation	*y*	*y*	*y*	*y*
HighestALT	*y*	*y*	*y*	*y*
HighestUrea	*y*	*y*	*y*	*y*
HighestDiastolic	*y*	*y*		
HighestCreatinine	*y*	*y*	*y*	*y*
AGE	*y*			
Parity	*y*			*y*
LowestPlatelets	*y*		*y*	*y*
HighestProteinuria		*y*		
FetalDiseaseStatus			*y*	*y*
Parity			*y*	
HighestSystolic			*y*	*y*
MaternalHeight		*y*		*y*

rate of every sub-tree and replacing the sub-tree with a leaf whenever the estimated error is considered not relevant. In other words, after deleting the less informative branches, the prediction accuracy of the model is expected to increase [48]. The threshold for defining an error rate as ‘not relevant’ can be set by the user through the internal confidence parameter. Tuning this parameter, in particular decreasing the confidence value, decreases in turn the size of the tree. While, on one hand, a reduction in the tree size improves the interpretability of the classification rules, on the other hand, it can delete the contribution that an attribute provides to the classification, deleting this attribute from the tree. In these experiments, this would have affected also the probability that the trees resulting from the C4.5 could contain the same attributes, both genetic and clinical, included in the ADTree solutions. The results shown here have been obtained by setting the parameters of the algorithms to their default values, while exploratory experiments have been performed to check the variability of the trees size with the changing of the confidence parameter for pruning. Nevertheless it would be interesting as future work to analyse the different results that can be obtained from C4.5 with different confidence values in comparison with the results obtained with ADTree algorithm, for every significant CBC threshold.

Table 4.19: Statistical results from ADTree run with only SNPs and with CBC=10

Seeds	0	10	20	30	40	50	60	70	80
Kappa	−0.03	−0.02	−0.12	−0.01	−0.06	−0.04	0.01	−0.05	−0.01

Table 4.20: Statistical results from ADTree run with SNPs and ‘Delivery gestation week’ and CBC=10

Seeds	0	10	20	30	40	50	60	70	80
Kappa	0.34	0.36	0.32	0.34	0.37	0.38	0.35	0.37	0.33

4.7 Mothers Dataset and Phenotype Analysis

The further step is to perform an analysis of the Phenotype of patients with PE. Thus, the second experiment has been repeated including only the phenotypic information. From the clinical point of view, it is useful to have an easy to use interface between the doctor and the patients data. The usual request of the doctor is to be provided with a device able to analyse the data from the patient with a computer and ultimately receive the probability to have a small baby or a mother with complications or with blood pressure problem post partum. The genetic data are more difficult to be collected by the doctor because it requires a laboratory analysis.

From the biological point of view it is interesting analyzing all the attributes but from the medical point of view it is better to start from the most available ones (clinical) and then add the genetic data. In the medical study, in order to provide a diagnosis, it is also important to choose the right predictive variables to be clearly distinguishable from the outcome one.

4.7.1 Pre-processing Analysis

Attributes

In the analysis of only clinical attributes it is important to add some information to the original DB of mothers. These features are related to the parents and partner of the mother. In particular there are 3 attributes that have to be recovered from the initial DB: systolic and diastolic blood pressure and essential hypertension requiring (EHT) medication (for both parents and the partner). This will add nine more attribute to the original DB.

Regarding the information about the blood pressure, as before, only two over eight attributes are kept as the remaining information is very much related to each other. In the first analysis the post natal attributes as ‘current oral contraceptive pill’ (OCP) and ‘anti hypertensive treatment’ (antiHT) are not included as they have no sense from the predictive point of view as they happen after the delivery. They can be used for further research as outcome(class).

Predictive Class

Concerning the predictive class of the Phenotype analysis, there are actually four different attributes that can be used as an outcome: CBC, complications (convulsions), blood pressure measure (with Sys >= 140 as a case or Dias <= 90 as a case) or ‘on antiHT treatment’. In the first analysis the CBC will be considered. In particular the CBC = 10 will be still the threshold for the case-control study.

Missing Values

The missing value will be kept in this analysis and coded according to the algorithm rules.

Data Balancing

Considering the list of the mothers without the genetic information, a DB of 364 mothers, 27 attributes and the CBC class is obtained. The attributes are:

- 18 mother information as number of pregnancies, age, hight, blood and urine tests.
- 9 parents and partner information about blood pressure.

The dataset is composed by 174 cases (48%), 190 controls (52%), which is an acceptable case-control ratio.

4.7.2 Statistical Significance Analysis

ADTree Analysis

Processing the new DB with ADTree, the results of Kappa shown in Table 4.21 are obtained, with a Kappa mean of 0.30. The attributes present in the final results are the following: DeliveryGestation, HighestALT, SystolicGranpa, HighestCreatinine, HighestUrea, DiastolicGranma, MaternalWeight.

Medical remarks

An important remark is about the blood pressure (BP) problems. If the mother of the mother has low BP, the mother should have it as well. In this case if the mother has PE

Table 4.21: Statistical results from ADTree run with CBC=10 for Phenoptype analysis

Seeds	0	10	20	30	40	50	60	70	80
Kappa	0.35	0.35	0.29	0.26	0.27	0.32	0.30	0.26	0.33

it means that there is a strong element triggering PE which is not due to high BP. Indeed usually the early PE is due to placenta problems. These problems are more related to the baby genome than the mother one. The late PE is more likely to be due to BP problem which is something more related to genetics of the mother.

4.8 Mothers Dataset and Alternative Classes

In this last section, further experiments are performed using alternative class choices in order to give an overview of how the class may affect the final results of the analysis in PE datasets.

4.8.1 'Week of delivery' as class

According to the medical remarks, a new dataset has been created containing only a small amount of attributes from the original one. These are:

- 7 Mothers features - Parity, Number of Pregnancies, Smoker, Highest Systolic and Diastolic Pressure, BMI(Kg/m2) and Age.
- 3 Grandmother features - EHTRequiringMedication, Systolic and Diastolic Pressure.
- 3 Grandfather features - EHTRequiringMedication, Systolic and Diastolic Pressure.
- 3 Partner features - EHTRequiringMedication, Systolic and Diastolic Pressure.

Within these attributes there is a new one, BMI, which comes from the combination of two previous one, 'Maternal Weight'(MW) in Kg and 'Maternal Height' (MH) in meter. More precisely the relation is shown in Equation 4.1.

$$BMI = MW/MH^2 \tag{4.1}$$

The class is then set as ‘Week of Delivery’. As the class needs to be Boolean, according to the medical advise, the threshold week of delivery is chosen equal to 34 to distinguish between the cases and the controls. Performing the analysis with this new DB doesn’t give any significant statistical result. Further experiments on new DBs, each composed by one of the listed attributes (in particular: BP of the parents, booking BP of the mother, BMI and Parity), do not provide any relevant result, both applying the ADTree algorithm and the C 4.5 one.

4.8.2 ‘Blood Pressure after delivery’ as class

In the final experiment the ‘Post natal systolic’ or ‘Post natal diastolic’ is considered as a predictive class. An analysis is then carried out to show how this attribute is related with one of the following:

- Booking blood pressure of the mother
- BMI of the mother
- Parents blood pressure
- Week of delivery
- Creatinine level

Post natal systolic or diastolic attribute need first to be converted in Boolean variables and then the BMI value needs to be calculated. Considering the ‘Systolic blood pressure post partum’ as a class, the threshold has to be set equal to 140. This means that the individuals with a ‘Systolic blood pressure post partum’ $<=$ 140 are controls and the rest are cases. In the DB the cases amount to 110 and the controls amount to 896. Over the total number of individuals equal to 1006, there are only 11% of cases which are not enough to perform a reliable case-controls analysis. Regarding the ‘Diastolic blood pressure post partum’, setting the threshold to 90, the cases are referred to every individual with a ‘Diastolic blood pressure post partum’ $>$ 90. At this point, the amount of cases is 31 which over the total of 1006 individuals represents just a 3% of the population.

This kind of analysis needs therefore to be referred as future study when more complete and extended databases will be available.

4.9 Summary

The aim of this Chapter is to describe the general framework that has been adopted in the application of decision tree algorithms to the analysis of SNPs data related to cases of pre-eclampsia. As previously discussed (Section 2.3.6), the choice of the decision tree algorithms has been driven by the medical request to have an overview of the limitations and strength that this type of approach provides.The results show the validity of this methodology to detect a subset of attributes associated with the predictable variable, providing a reduction in the size of the dataset. This is realised comparing and contrasting the solutions obtained from three different algorithms. Additionally, an extended analysis of the statistical significance of the results provides the user with a reliable tool for the selection of the best CBC threshold in this pre-eclampisa association study. Moreover, from the clinical point of view, this study confirmed that the medical interpretation of the 'corrected birth-weight centile' (CBC) value of 10 is indeed a meaningful cut-off, and confirmed association between an infant's CBC and the 'week of delivery' parameter.

In the second part of this Chapter several experiments are carried out with different populations and different variables settings. This is to highlight that factors such as changing attributes and the class choice, keeping or eliminating the missing values, and knowing some medical aspects of the disease under analysis can help improve the likelihood of the study providing statistical significance to the analysis.

In conclusion, this study provides researchers with a generic framework to be used for further research analysis of such data, following a pre-processing stage of redundancy elimination.

Chapter 5

Decision Trees and Artificial SNPs Datasets

5.1 Introduction

In the previous Chapter, three decision tree algorithms were examined in the context of association studies. The next stage in the analysis would be to investigate in depth these algorithms under different initial conditions. In order to permit this analysis, the concept of an artificial dataset is illustrated and discussed. The aim of this Chapter is to present an overview of the issues and considerations related to the creation of synthesised sets of SNPs data, together with a study of the pitfalls of the three algorithms employed. Knowing the rules that a database is based on allows the possibility to test for the performance and the limitations of a given procedure. In this Thesis more than one real dataset is used to perform different analysis but the use of artificial datasets is also included in order to have a clearer and deeper analysis of the methodologies and algorithms used and proposed. Building such a dataset provides the possibility, for instance, to detect the ability of a specific technique to classify the data in the correct way. In this scenario, the most significant steps to be followed for this purpose will be shown and discussed, analysing all aspects of this multi-faceted problem.

This chapter is meant to give an initial review of a problem that is complex and extensively researched. This overview provides future researchers with a good starting point for further consideration and realisation of more extensive experiments.

5.2 SNPs Datasets Rules: An Overview

It is nowadays feasible to collect a huge amount of genetic data in reasonable time thanks to the continuously improved techniques in biological fields. However it is not always possible to have access to specific datasets shaped by the fittest features for the kind of analysis that needs to be performed. In these cases the best solution is to create the source of data in an artificial way, so that any method or technique which is proposed can be tested in a easily repeatable way.

In this context, the attention is addressed to genetic dataset for diseases association study. A generic medical dataset can be composed by either exclusively genetic information that in the specific case are SNPs, or also clinical attributes. In this case the attention will be focused only on a dataset composed by SNPs data, according to the aim of this research. The rules used to build these new DBs are discussed and agreed with a medical support in order to create the most reliable set of data. Different operations can be performed for each DB created. It is possible to vary the size (which is the number of instances or patients), change the probability of occurrence of a certain allele or a couple of allele, increase the number of columns of the DB, which correspond to the number of SNPs analyzed. Finally, one of the most relevant issue, the probability of contracting a disease for each value of the SNP can be set and changed every time. From now on, in order to give a concrete overview of the dataset creation the pre-eclampsia (PE) disorder will be taken as example.

5.2.1 Phenotypic Background

The probability to contract a specific disease is always related to some clinical features such as clinical history or physical and physiological parameters of the individual. For this reason, before starting a genetic analysis a sort of initial condition for each person needs to be set and this is different depending on previous and current health state. In synthesis, the susceptibility to a certain disease can be expressed by the general formula 5.1.

$$Pdisease = P(SNPs) + P(Phen) \tag{5.1}$$

which takes in consideration both the genetic and clinical data. In case of healthy

Table 5.1: Added risk to contract PE due to clinical conditions.

Clinical conditions	Added Risk of PE	CI (95%)
Reference: Healthy	3%	–
Age >40	∗80%	20 – 260%
Obesity: BMI > 29	∗280%	175 – 459%
Twin Pregnancy	∗400%	230 – 760%
Assisted Reproduction	∗330%	–
Previous Hypertension	∗700%	–
Diabetes	∗200%	–
Previous PE	∗700%	570 – 870%
Not Previous PE	< 100%	–

women, for instance, the probability to contract PE is estimated to the value of 3%. As soon as one clinical condition is added this probability is increased of a certain factor. The relative increment of probability conditionate to the healthy state, for each clinical circustance is shown in Table 5.1, in accordance with [151]. The last column shows the 95% confidence interval (CI) for the relative risk. The susceptibility to PE due to a clinical condition is then calculated by the formula 5.2.

$$P(Phen) = P(Healthy) * (1 + P(cond)) \tag{5.2}$$

The added risk for more than one condition requires more complicated analysis, as there are correlations between the risk factors. For instance, women who have previous hypertension are also more likely to be obese. Therefore the risk associated with obesity partly contributes to the relative risk associated with hypertension, and vice versa.

5.2.2 One SNP Analysis

In the first instance, a DB with a single column of SNP is created. In order to realize this, there are a few steps to consider that can be summarized in the following issues:

- SNP code (see 4.5.2)

Table 5.2: Example of allele frequencies (*AlleleCondition* 1) and probability to contract PE disease for each SNP value, according to the medical advise.

SNP value	Allele frequency	Disease Risk
2	36%	3%
3	48%	4.5%
4	16%	6.8%

- DB size
- Allele Frequency
- Disease Risk
- Disease Model

Setting the allele frequencies

Considering the recoding of the SNP, the possible values that belong to this new variable are just three: 2 and 4 which correspond to the homozygous (1,1) and (2,2) and the 3 value which corresponds to the heterozygous (1,2) and (2,1). For each of these values the occurrence frequency needs to be set in a way that reflects the real case, being the allele 1 the common one and the allele 2 the rare one. There is no precise answer to this problem as different SNPs can have different frequencies of their values. The criterion is to collect the information from the huge genetic DBs created from the human genome analysis and choose a threshold that is an estimation around the average. According to the medical advice, these probabilities have been set to 48% for SNP equal to 3, 16% for SNP equal to 4 and 36% for SNP equal to 2. These hypotesis are referred with the name *AlleleCondition* 1 as shown in Table 5.2.

Setting the disease probabilities for each allele

The third step concerns the choice of the probability to contract a disease given a certain value for the SNP. This choice is not strict and constant for any case. There are different values of the probability depending on the disease and, for each allele value, depending on the SNP. That means that SNP8 equal to 2 can give 80% of probability to contract a disease whereas SNP6 equal to 2 can give 30% of the probability. For this reason, it is important once more the medical expertise to select the reasonable threshold for each SNP and for each allele value. Regarding the case of PE disease, a real example of PE risk can be calculated according to the medical advice. For instance, referring to the *AlleleCondition* 1, the probability to contract PE for the artificial SNP = 3 is set to 4.5 % and for SNP = 4 to 6.8%, Table 5.2. The probability related to SNP = 2, which means homozygous SNP (1,1), is not taken in consideration for added risk of disease as the allele 1 is the most common and usually not related to the disease. The allele 2 is the most rare and the one who has more probability to be linked to the disease either in a positive or in a negative way. For this reason, the risk to contract a disease for SNP equal to 2 is set at the healthy rate of 3%.

Unfortunately, in the study of PE there are still lots of uncertainty about the probability of contract the disease under a certain genotype hypothesis. One example of the findings is related to the SNP ‘Factor V Leden’ which shows an occurrence of 5% for the SNP value of 3 with a risk factor of 1.49 as shown in Table 5.3, [152].

It is important to notice that even if the contribution from the genetic side to the disease susceptibility sometimes is not very impressive, the added risk due to clinical background can considerably increase the range of the probability.

Setting the size of the DB

In the last step, the change of the DB size is considered. This consists of creating a flexible number of instances that corresponds to the number of patients analysed. The size of the DB is a quite important issue in genetic analysis for instance for case control studies. Indeed lots of research has been done in order to detect the minimum size of a dataset for a reliable analysis, [142], [143], [144], [145]. Together with the case control ratio, this parameter plays an important role in the performance evaluation of the technique as increasing the size of the DB usually increases the statistical significance of the test.

Table 5.3: Set of values for three different candidate genes.

Param set	Allele Frequency	Odd Ratio	95% Conf Int
Factor V Leiden	0.05	1.49	1.13 – 1.96
STOX1 T	0.35	–	–
STOX1 C	0.66	–	–
STOX1 CC	0.43	1	–
STOX1 CT	0.46	1.2	–
STOX1 TT	0.12	6.86	–
TCF7L2	0.43	1.6	–

5.2.3 Multiple SNPs Analysis

The further step is the addition of a new column to the DB which means the introduction a new SNP in the analysis. Beside the previous steps this new attribute introduces more rules for setting the existing parameters. The occurrences frequencies have to be set either in the same way of the first SNP or with different values, according to the medical advice. Regarding the probability to contract a disease, this SNP can be completely independent from the previous one which means the rules for its parameters can be chosen apart. There are anyway cases in which the values of this SNP are somehow related to the value of the previous SNP and in this case new complex rules need to be created considering the interaction between these two SNPs. A similar analysis can be done with three SNPs, four SNPs and so on. The more SNPs are considered in the analysis the more complex the rules can become. The size of the DB can be affected from the amount of SNPs that are considered in the analysis as the algorithm can present some difficulties in dealing with huge size of dataset.

5.2.4 Family SNP Analysis

Considering a family-based analysis, once that the number of SNPs has been fixed for an individual, the attention can be focused on the other family members in order to detect any trend. In the case of a population of babies for instance, the further columns that can be added to the DB are the genetic information from the mother and the father of the subjects. There are of course different rules that regulate the value of the SNPs of the baby, given the parents genotype: the inheritance low together with mutations and recombinations. Following all these considerations, the problem can be analyzed under family point of view. Information from siblings and other relatives can also be considered useful.

5.3 Artificial Dataset: One SNP Analysis

In this section an example of the application of this theory proposed for the creation of an artificial dataset is described. The analysis is still related to pre-eclampsia disease. The synthetic datasets created are processed with decision tree algorithm previously used in order to test their performance with different input data.

5.3.1 Experimental Data

The first test for the algorithms is performed with a DB composed only by one SNP and the class, which is Boolean and is equal to 1 for a case and 0 for a control. The values of the SNP can be 2, 3 and 4 according to the previous codification.

Database Size

In this section, the experiments are carried out with DBs of different size to detect the different performance of the algorithms. The analyzed DBs are composed by 100, 1,000, 10,000 and 100,000 individuals. The respective occurrences frequencies and the probabilities of contracting the PE disease are shown in Table 5.2.

Table 5.4 shows the Kappa values obtained from the analysis of the DB with ADTree, C4.5 and ID3 with different size of the input dataset. It is clear that there is no significant results for any of the algorithm analyzed. Moreover when the size of the DB reaches the level of 100,000 individuals, the software is not able to perform the analysis. Therefore a population between 10,000 and 100,000 individuals is the maximum size allowed when the analysis is carried out on a single SNP.

Table 5.4: Kappa value under *AlleleCondition* 1 with ADtree, C4.5 and ID3 for different size of the 1 SNP DB.

DB Size	ADTree	C4.5	ID3
100	−0.08	0	−0.014
1,000	0.06	0.06	0.06
10,000	0	0	0
100,000	–	–	–

Allele Frequency

In this step the size of the DB is fixed to a reasonable value allowed by the software, 10,000 individuals. The probability to contract a disease for each value of the SNP is considered constant so that the parameter that is now flexible is the allele frequency. As stated before, the allele 1 is usually considered as the common one and the allele 2 as the rare one. The allele frequency setting is reduced to only one parameter problem, for instance the occurrence of allele 1. A set of allele frequencies for the SNP values of 1 can be chosen and the occurrences for allele 2 are just the complementary probabilities to 1. Then the joined probabilities for the respective combination of allele (1/1, 1/2-2/1 and 2/2) recoded as (2,3,4) can be calculated. Considering for instance an occurrence p of 25% for the allele 2 and an occurrence q = 1-q of 75% for the allele 1, under the HWE hypothesis, the joined probability for SNP equal to 1/2 or 2/1 is double the product of the two probabilities ($2 \times p \times q$), whereas for SNP equal to 1/1 or 2/2 the probabilities are respectively p^2 and q^2. In Table 5.5 are shown three examples of allele frequency for allele 1 and the consequent frequency for the new codification.

Probability to contract a disease

In this section the size of the DB is fixed to the same amount previously chosen. Keeping the allele frequency constant as defined in *AlleleCondition* 1, the software is tested with different datasets, each one with a different value of ‘Probability to contract the disease’. The idea is to explore the cases in which the probability belongs to the range 3-90 %. This

Table 5.5: Three Examples of Allele Frequencies and Joined Probabilities for the new codification

SNP value(s)	All Freq 1	All Freq 2	All Freq 3
1	0.75	0.50	0.99
2	0.25	0.50	0.01
1/1 (2)	0.56	0.25	0.98
1/2 - 2/1 (3)	0.37	0.50	0.02
22 (4)	0.06	0.25	≈ 0

is because, referring to PE, the probability to contract PE for healthy women is 3% and considering any possible added risk due to clinical conditions, the risk could resonably rises up to around 90%. An array of probability is then created initially composed by ten elements. Each element corresponds to the following value of risk: 3, 10, 20, 30, 40, 50, 60, 70, 80 and 90 %. All these occurrences are referred to the SNP value of 4 which correspond to the rare-rare combination of allele.

In order to have a more general overview of the problem, the disease risk for the SNP allele equal to 2 and 3 are set in a similar way around the value of 50. In this way the disease risk for the most rare case of allele (4) can span in a range that is smaller and greater than this threshold. This means that both the ‘protective’ and ‘adverse’ effect of the rare allele can be considered. The summary of this new condition for the risk to contract a disease are shown in Table 5.6 and it is referred as *RiskCondition* 1. Thus, ten input files have been created to be processed with the three different algorithms.

Table 5.6: *RiskCondition* 1: Allele frequencies and probability to contract a disease for each SNP value.

SNP value	Allele frequency	Disease Risk
2	36%	50%
3	48%	54.5%
4	16%	$3, 10, 20..90\%$

In Table 5.7 is shown the trend of Kappa value for the results obtained processing the 10 files with ADTree, C4.5 and ID3. It is clear that the algorithms provide significant result when the probability to contract a disease exceeds the threshold of 60%, which is also related to a DB with an ammount of cases grater than 30%. This results are also exactly confirmed by each of the algorithm used.

Table 5.7: Kappa value under *RiskCondition* 1 with ADtree, C4.5 and ID3 for different Disease Risk of the 1 SNP DB.

Disease Risk %	ADTree	C4.5	ID3	Cases %
3	0	0	0	3.93
10	0	0	0	6.86
20	0	0	0	10.53
30	0	0	0	15.91
40	0	0	0	20.85
50	0	0	0	26.18
60	0.60	0.60	0.60	32.14
70	0.70	0.70	0.70	39.35
80	0.78	0.78	0.78	46.37
90	0.87	0.87	0.87	54.27

Disease Model

In order to have a deeper analysis of the problem, a new model for the influence of the disease risk is introduced. If rare allele (2) is the responsible for the disease it could be possible that the allele combination 1/1 gives risk disease equal to 0%, whereas 2/2 give a risk of 100% and 1/2 and 2/1 50%. In general terms, there can be these extreme cases to analyze as shown in table 5.8 together with all the intermediate rates. Some examples are: 1:99, 5:95, 10:90.. 85:15 and so on.

Analysing more in depth the problem, the disease model depends actually on the type of allele in terms of dominance or recessiveness. The disease is usually associated to the most rare allele (2) and it can either be dominant or recessive. In the former the presence

Table 5.8: Three Examples of Disease Risks for each SNP value

SNP value(s)	Disease Risk 1	Disease Risk 2	Disease Risk 3
2	0 : 100	0 : 100	0 : 100
3	50 : 50	100 : 0	0 : 100
4	100 : 0	100 : 0	100 : 0

of allele 2 is enough to cause the disease, in the latter only the 2/2 allele value implies the disease. A simple example is shown in Table 5.9.

Table 5.9: Disease Models

SNP values	Recessive	Mixed	Dominant
2	$control$	$control$	$control$
3	$control$	$half-case$	$case$
4	$case$	$case$	$case$

In the Recessive model, the disease risk for SNP value equal to 2 or 3 is the same one and it represents the healthy condition which therefore could be reasonably set to 3%. The allele value of 4 represents instead the disease condition and for this value different values of probability can be set as shown in Table 5.10.

In the Dominant model, the disease risk for SNP value equal to 3 or 4 is the same one and it represents the disease condition. For this SNP values once again different amounts of probability are set as shown in Table 5.10. The allele value of 2 represents instead the health condition which, as before, can be set to 3%.

Regarding the Mixed model, the allele 2 represents still the healthy condition (disease risk of 3%), the allele 4 represents the disease condition and can be set with 9 different values between 10% and 90%. The allele 3 affects the disease probability only half the way of allele 4, as shown in Table 5.10.

Table 5.10: Disease Risk for each Disease Model

SNP values	Recessive	Mixed	Dominant
2	3%	3%	3%
3	3%	5,10..45%	10,20..90%
4	10,20..90%	10,20..90%	10,20..90%

5.4 Experimental Results for Disease Models: One SNP Dataset

A population of 10,000 individuals will be considered for every experiment performed as this is a reasonable size for a good performance of the algorithms.

5.4.1 The Recessive Model

Variable Allele Frequency

At the this stage, the disease risk is kept fixed for each SNP values and only the allele frequency is changed to check the performance of the three algorithms under analysis. The chosen risk disease is 3% for allele 2 and 3, as for the recessive model, and 10% for the allele 4. According to the definition, the allele 1 is the most common one, therefore the allele 1 frequency is expected to be greater than the allele 2 frequency. For this reason five different DBs are created, each one with five different combinations of allele frequencies for each SNP value 1 and 2 (p and q) as shown in Table 5.11.

Variable Allele Frequency and Variable Disease Risk

This same procedure is made for 9 different disease risks in the range from 20 to 90 for the allele value of 4 and a disease risk of 3% for the alelle value of 2 and 3. In this way 45 (5×9) different datasets are created, each one with all the possible combinations between allele frequency and risk of disease. The Kappa results from the three algorithms are shown in the Table 5.12 and Table 5.13.

The number of cases changes for every experiment because the SNP column is created

Table 5.11: Allele Frequencies (AF) setting for the nine datasets

SNP value(s)	Recode	AF	AF 1	AF 2	AF 3	AF 4	AF 5
1	–	p	0.50	0.60	0.70	0.80	0.90
2	–	q	0.50	0.40	0.30	0.20	0.10
1/1	2	p^2	0.25	0.36	0.49	0.64	0.81
1/2 - 2/1	3	$2 \times p \times q$	0.50	0.48	0.42	0.32	0.18
2/2	4	q^2	0.25	0.16	0.09	0.04	0.01

randomly every time but it is still limited to a narrow range due to the same rules used. An example of percentage of cases for each of the 45 files created is shown in Table 5.14.

Table 5.12: Recessive Model: Kappa Value over 5 different Allele Frequencies datasets with disease risk between 10 and 40%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,3,10)	Cases %	4.60	3.83	3.42	2.77	3.04
	κ	0	0	0	0	0
(3,3,20)	Cases %	7.42	6.09	4.19	3.53	3.39
	κ	0	0	0	0	0
(3,3,30)	Cases %	9.88	7.42	5.51	4.41	2.89
	κ	0	0	0	0	0
(3,3,40)	Cases %	12.60	9.04	6.03	4.22	3.23
	κ	0	0	0	0	0

5.4.2 The Dominant Model

The previous procedure is now repeated for the dominant disease model as described in Table 5.10.

Table 5.13: Recessive Model: Kappa Value over 5 different Allele Frequencies datasets with disease risk between 50 and 90%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,3,50)	Cases %	15.53	10.87	7.52	5.16	3.37
	κ	0	0	0	0.44	0.23
(3,3,60)	Cases %	17.75	11.96	7.86	5.25	3.42
	κ	0.65	0.63	0.57	0.51	0.28
(3,3,70)	Cases %	18.94	13.72	8.92	5.76	3.58
	κ	0.70	0.72	0.67	0.57	0.29
(3,3,80)	Cases %	22.25	15.39	9.33	5.92	3.68
	κ	0.80	0.78	0.75	0.64	0.31
(3,3,90)	Cases %	24.88	17.94	10.91	6.27	3.74
	κ	0.86	0.86	0.81	0.67	0.38

Table 5.14: Recessive Model: Cases over different Allele Frequencies datasets with a variable disease risk for the allele (2,3,4)

Disease Risk	AF 1	AF 2	AF 3	AF 4	AF 5
10	4.60	3.83	3.42	2.77	3.04
20	7.42	6.09	4.19	3.53	3.39
30	9.88	7.42	5.51	4.41	2.89
40	12.60	9.04	6.03	4.22	3.23
50	15.53	10.87	7.52	5.16	3.37
60	17.75	11.96	7.86	5.25	3.42
70	18.94	13.72	8.92	5.76	3.58
80	22.25	15.39	9.33	5.92	3.68
90	24.88	17.94	10.91	6.27	3.74

Variable Allele Frequency and Variable Disease Risk

This time, the disease risk is fixed to 3% for allele 2, as for the dominant model, and an identic value is fixed for the allele 3 and 4, but variable from 10% to 90%. As before, the allele frequency is then changed, creating five files each one with a different combination of allele frequencies for each SNP value 1 and 2 (p and q) as shown in Table 5.11.

In this way 45 (5 × 9) different datasets are created, each one with all the possible combinations between allele frequency and risk of disease. The Kappa results from the three algorithms are shown in Tables 5.15 and 5.16. As before, an example of percentage of cases for each of the 45 files created is shown in Table 5.17.

Table 5.15: Dominant Model: Kappa Value over 5 different Allele Frequencies datasets with disease risk between 10 and 40%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,10,10)	Cases %	7.93	7.85	6.62	5.53	4.67
	κ	0	0	0	0	0
(3,20,20)	Cases %	15.93	14.60	10.89	9.26	6.15
	κ	0	0	0	0	0
(3,30,30)	Cases %	23.30	20.06	16.60	13.03	7.73
	κ	0	0	0	0	0
(3,40,40)	Cases %	30.69	26.21	21.55	16.80	9.49
	κ	0	0	0	0	0

5.4.3 The Mixed Model

In this last section the same procedure is repeated for the mixed disease model still for 10,000 patients.

Variable Allele Frequency and Variable Disease Risk

This time the risk disease is set to 3% for allele 2, to the range 5%-45% for allele 3 and to the range 10%-90% for the allele 4. We create then the 45 files with every different

Table 5.16: Dominant Model: Kappa Value over 5 different Allele Frequencies datasets with disease risk between 50 and 90%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,50,50)	Cases %	38.74	33.10	26.58	19.99	12.27
	κ	0.31	0.39	0.11	0.21	0
(3,60,60)	Cases %	45.94	39.45	32.26	23.51	13.56
	κ	0.42	0.49	0.57	0.61	0.64
(3,70,70)	Cases %	52.94	46.57	37.72	27.07	15.73
	κ	0.52	0.61	0.66	0.70	0.72
(3,80,80)	Cases %	61.32	51.60	42.55	31.17	17.89
	κ	0.66	0.72	0.77	0.79	0.80
(3,90,90)	Cases %	68.50	58.19	47.09	34.60	20.39
	κ	0.80	0.84	0.87	0.88	0.86

Table 5.17: Dominant Model: Cases over different Allele Frequencies datasets with a variable disease risk for the allele (2,3,4)

Disease Risk	AF 1	AF 2	AF 3	AF 4	AF 5
10	7.93	7.85	6.62	5.53	4.67
20	15.93	14.60	10.89	9.26	6.15
30	23.30	20.06	16.60	13.03	7.73
40	30.69	26.21	21.55	16.80	9.49
50	38.74	33.10	26.58	19.99	12.27
60	45.94	39.45	32.26	23.51	13.56
70	52.94	46.57	37.72	27.07	15.73
80	61.32	51.60	42.55	31.17	17.89
90	68.50	58.19	47.09	34.60	20.39

combination of allele frequencies and disease risk and the Kappa results from the three algorithms are shown in the Table 5.18 and 5.19.

The number of cases for each of the 45 files created is shown in Table 5.20.

Table 5.18: Mixed Model: Kappa Value over 5 different Allele Frequencies datasets with disease risk between 10 and 40%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,5,10)	Cases %	8.30	7.12	6.63	5.33	4.79
	κ	0	0	0	0	0
(3,10,20)	Cases %	15.68	13.19	12.01	8.25	6.58
	κ	0	0	0	0	0
(3,15,30)	Cases %	23.19	20.07	16.52	12.83	7.93
	κ	0	0	0	0	0
(3,20,40)	Cases %	31.32	26.30	21.19	16.70	9.94
	κ	0	0	0	0	0

Table 5.19: Mixed Model: Kappa Value over 5 different Allele Frequencies datasets with disease risk between 50 and 90%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,25,50)	Cases %	38.31	32.77	26.73	19.95	11.59
	κ	0.32	0	0.33	0	0.49
(3,30,60)	Cases %	45.95	38.39	31.44	24.35	13.85
	κ	0.40	0.48	0.57	0.62	0.65
(3,35,70)	Cases %	53.58	45.82	37.88	27.62	15.89
	κ	0.52	0.59	0.67	0.70	0.71
(3,40,80)	Cases %	60.95	51.45	41.97	29.92	18.19
	κ	0.64	0.72	0.76	0.80	0.80
(3,45,90)	Cases %	68.24	59.06	47.26	34.37	19.49
	κ	0.80	0.85	0.87	0.87	0.86

Table 5.20: Mixed Model: Cases over different Allele Frequencies datasets with a variable disease risk for the allele (2,3,4)

Disease Risk	AF 1	AF 2	AF 3	AF 4	AF 5
10	8.30	7.12	6.63	5.33	4.79
20	15.68	13.19	12.01	8.25	6.58
30	23.19	20.07	16.52	12.83	7.93
40	31.32	26.30	21.19	16.70	9.94
50	38.31	32.77	26.73	19.95	11.59
60	45.95	38.39	31.44	24.35	13.85
70	53.58	45.82	37.88	27.62	15.89
80	60.95	51.45	41.97	29.92	18.19
90	68.24	59.06	47.26	34.37	19.49

5.4.4 Disease Risk from 50% to 60% for the 3 models

Following the previous results, the attention is addressed to a deeper insight within the disease risk range of 50 to 60% for the three disease models. This analysis is realized in the attempt to understand the reasons of the unexpected Kappa trend obtained under the specific conditions. Another observation arises from the stability of the results obtained. Every time that a test is performed, a new random dataset is created. The new input provides in turn a different set of results. Running the algorithms more than once for every test provides a validation of the stability of the results.

We then repeat the experiment for all the allele frequency set but setting the disease risk to 10 different value between 50 and 60%, i.e. DR= 50, 51..59%. The results obtained for the three models are shown respectively in Tables 5.21, 5.22, 5.23.

5.4.5 Discussion

Before discussing the outcomes of the experiments, a multi-variate ANOVA test has been performed on the results, in order to examine the strength of different effects on the value of Kappa, considering the two factors: disease risk (DR) and allele frequency (AF). Set-

Table 5.21: Recessive Model: Kappa Value over 5 different Allele Frequencies datasets with 9 different disease risks between 50 and 60%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,3,50)	Cases %	14.74	10.24	7.36	4.86	3.38
	All	0	0.55	0	0	0
(3,3,51)	Cases %	15.15	10.27	7.15	4.67	3.54
	ADTree κ	0.56	0.57	0.54	0.15	0.22
	ID3 κ	0.56	0.57	0.54	0.10	0.22
	C4.5 κ	0.56	0.57	0.50	0.05	0.05
(3,3,52)	Cases %	15.32	10.47	7.44	4.78	3.39
	ADTree κ	0.31	0.55	0.1	0.44	0.08
	ID3 κ	0.31	0.55	0.1	0.44	0.08
	C4.5 κ	0.31	0.55	0	0.44	0
(3,3,53)	Cases %	15.34	10.69	7.47	4.81	3.39
	All	0.59	0.55	0.54	0.46	0
(3,3,54)	Cases %	16.04	11.26	7.55	4.92	3.25
	All	0.56	0.56	0.55	0.45	0.07
(3,3,55)	Cases %	15.99	11.33	8.06	5.26	3.45
	All	0.60	0.59	0.48	0.48	0.24
(3,3,56)	Cases %	16.53	11.33	7.81	5.23	3.43
	All	0.60	0.60	0.55	< 0.45	0.25
(3,3,57)	Cases %	16.55	11.76	7.85	5.13	3.45
	All	0.62	0.62	0.56	0.46	< 0.22
(3,3,58)	Cases %	17.38	11.66	7.73	4.81	3.64
	All	0.63	0.62	0.57	0.49	0.22
(3,3,59)	Cases %	17.18	12.03	8.29	5.25	3.20
	All	0.65	0.60	0.59	0.49	0

Table 5.22: Dominant Model: Kappa Value over 5 different Allele Frequencies datasets with 9 different disease risks between 50 and 60%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,50,50)	Cases %	17.18	12.03	8.29	5.25	3.20
	ADTree κ	0.40	0.05	0.42	0	0.1
	ID3 κ	0.40	0.05	0.42	0	0.1
	C4.5 κ	0.07	0	0.39	0	0
(3,51,51)	Cases %	38.31	33.34	26.83	19.63	11.82
	ADTree κ	0.32	0.41	0.48	0.34	0.55
	ID3 κ	0.32	0.41	0.48	0.30	0.55
	C4.5 κ	0.32	0.41	0.48	0.16	0.55
(3,52,52)	Cases %	39.19	33.93	28.66	20.96	12.00
	All	0.32	0.41	0.49	0.54	0.57
(3,53,53)	Cases %	40.27	35.29	28.49	21.94	12.49
	All	0.34	0.42	0.48	0.55	0.58
(3,54,54)	Cases %	41.65	35.25	28.90	22.03	12.36
	All	0.35	0.43	0.51	0.57	0.57
(3,55,55)	Cases %	41.63	37.09	29.75	21.80	13.41
	All	0.37	0.44	0.52	0.58	0.59
(3,56,56)	Cases %	43.50	37.09	30.18	21.78	13.05
	All	0.37	0.46	0.53	0.60	0.61
(3,57,57)	Cases %	43.50	37.56	31.08	22.48	13.13
	All	0.38	0.46	0.52	0.59	0.59
(3,58,58)	Cases %	43.14	38.83	30.69	23.02	13.59
	All	0.38	0.47	0.54	0.59	0.62
(3,59,59)	Cases %	44.94	38.66	31.24	24.09	13.79
	All	0.40	0.47	0.56	0.60	0.62

Table 5.23: Mixed Model: Kappa Value over 5 different Allele Frequencies datasets with 9 different disease risks between 50 and 60%.

DR(%)	SNP values	AF 1	AF 2	AF 3	AF 4	AF 5
(3,25,50)	Cases %	38.34	32.77	27.62	19.65	11.78
	All	< 0.12	0	0.44	0.53	< 0.07
(3,25.5,51)	Cases %	39.03	33.21	26.87	20.14	12.14
	All	0.30	0.34	0.48	0.53	0.55
(3,26,52)	Cases %	39.83	33.63	28.90	20.85	12.54
	All	0.34	0.40	0.45	0.54	0.58
(3,26.5,53)	Cases %	40.61	34.71	28.17	20.57	11.79
	All	0.34	0.41	0.49	0.54	0.57
(3,27,54)	Cases %	41.53	35.61	29.16	21.14	12.58
	All	0.36	0.43	0.50	0.52	0.58
(3,27.5,55)	Cases %	41.80	36.11	29.38	21.44	12.39
	All	0.36	0.45	0.50	0.57	0.59
(3,28,56)	Cases %	43.38	37.00	29.00	21.84	13.19
	All	0.37	0.47	0.52	0.57	0.61
(3,28.5,57)	Cases %	43.84	36.84	30.47	22.84	13.40
	All	0.38	0.46	0.54	0.60	0.57
(3,29,58)	Cases %	43.50	38.18	31.37	22.44	14.47
	All	0.38	0.46	0.54	0.60	0.62
(3,29.5,59)	Cases %	46.16	38.66	30.99	23.38	13.94
	All	0.39	0.48	0.56	0.61	0.62

ting the critical p-value to 0.05, for all the models, the test provides good evidence that both disease risk and allele frequency affect the value of Kappa when the disease risk is grater than 50. The p-values obtained from this analysis are reported in table Table 5.24.

Table 5.24: Multi-variate ANOVA Test: p values from the test for recessive, dominant and mixed model

Model (%)	p-val (DR)	p-val (AF)
Recessive (DR = 60-90)	1.7×10^{-6} $(df = 3)$	2.7×10^{-10} $(df = 4)$
Recessive (DR = 50-60)	1.0×10^{-4} $(df = 9)$	3×10^{-9} $(df = 4)$
Dominant (DR = 60-90)	6×10^{-9} $(df = 3)$	2.6×10^{-5} $(df = 4)$
Dominant (DR = 50-60)	2.2×10^{-5} $(df = 9)$	2×10^{-4} $(df = 4)$
Mixed (DR = 60-90)	4.3×10^{-8} $(df = 3)$	7.5×10^{-5} $(df = 4)$
Mixed (DR = 50-60)	3.1×10^{-6} $(df = 9)$	1.9×10^{-8} $(df = 4)$

It is clear from the results shown in Tables 5.12, 5.13, 5.15, 5.16, 5.18 and 5.19 that as far as the DR is under 40% in all the disease models, the Kappa value is zero. As the DR rises, up to the value of 60%, a good range of Kappa value is obtained for all three models and for most of the cases the Kappa value is the same for the 3 algorithms. As expected, the Kappa value rises as the DR increases for each allele frequency setting. Concerning the trend with the allele frequency, it is clear that for the Recessive model, an increase of the frequency of the allele one decreases the value of Kappa, whereas for the Dominant and the Mixed model the trend in Kappa is opposite.

There is then a no clear change in the Kappa value when the disease risk is set to 50%, which is easily understandable considering that in this case the probability to be sick and healthy are the same. That implies the dataset is obtained with extreme cases. In order to check the stability of the Kappa values obtained in the result, the same test is performed nine different times (every time a new dataset is built randomly) and the Kappa values obtained are compared at each step. An example of this analysis is shown in Table 5.25 and Table 5.26 for the allele frequency 25%-50%-25% (AF1) and for the disease risk equal to 40, 50 and 60. The results show that the DR equal to 40 and 60 provides quite stable results, whereas the disease risk equal to 50 gives a clearly random answer.

Table 5.25: Recessive Model: κ stability over 9 tests with AF (25,50,25) and DR = 40,50,60

Disease Risk(%)	Algorithm	K1	K2	K3	K4
(3,3,40)	ADTree κ	0	0	0	0
	ID3 κ	0	0	0	0
	C4.5 κ	0	0	0	0
(3,3,50)	ADTree κ	0.48	0	0.10	0
	ID3 κ	0.48	0	0.10	0
	C4.5 κ	0.40	0	0	0
(3,3,60)	ADTree κ	0.64	0.64	0.64	0.62
	ID3 κ	0.64	0.64	0.64	0.62
	C4.5 κ	0.64	0.64	0.64	0.62

Table 5.26: Recessive Model: κ stability over 9 tests with AF (25,50,25) and DR = 40,50,60

Disease Risk(%)	Algorithm	K5	K6	K7	K8	K9
(3,3,40)	ADTree κ	0	0	0	0	0
	ID3 κ	0	0	0	0	0
	C4.5 κ	0	0	0	0	0
(3,3,50)	ADTree κ	0.31	0	0.54	0.35	0.55
	ID3 κ	0.31	0	0.54	0.35	0.55
	C4.5 κ	0.10	0	0.51	0.35	0.52
(3,3,60)	ADTree κ	0.64	0.62	0.63	0.64	0.62
	ID3 κ	0.64	0.62	0.63	0.64	0.62
	C4.5 κ	0.64	0.62	0.63	0.64	0.62

5.5 Future work - Multiple SNPs Datasets

Following the previous analysis, a new SNP can be included in the study, increasing the number of the dataset columns by one. Introducing new SNPs, arises a few issues that can affect badly the results of the process. In this section, in order to avoid confusion between numbers, the SNPs values (2,3,4) are recoded as SNPs new values (A,B,C). When running the algorithm, in general terms, it is suggested not to use numeric codification for the SNPs but categorical, in order to have a better interpretation of the results.

5.5.1 Redundancy

The first potential problem is the redundancy of information that the new SNP can add to the dataset. If there are two attributes of the dataset which contain roughly the same information, that is they are either mostly the same or close to a combination, the algorithm may skip one of the two SNPs following some criteria which may be not acceptable. In order to check how the algorithm cope with this problem, a dataset should be created starting with two SNPs having very similar values (or being very much correlated between each other).

In order to give an example, a correlation between two SNPs is realized, creating two columns which are very similar between them, but this would not be the only solution. This means that 90% of the values of the two SNPs are AA, BB or CC. There are of course many possible combinations of 2 SNPs taken from the 3 disease models. In this context, a simple example is shown in the case of 2 SNPs according to the Recessive Model. Under this hypothesis, there are still lots of situations where these two SNPs, similar at 90%, have different combination of the disease risk for SNP equal to 4. In one of the possible scenario, the disease susceptibility could be affected by either both of them or only one of them. Althoght, in both cases, the aim is to check how the algorithm copes with the two SNPs. In this scenario, the analysis could be addressed to the case where only one SNP gives a substantial contribution to the disease and the other one is only redundant. This means to set disease risk for example of SNP1 equal to (3,3,80) and SNP2 equal to (3,3,10) (as shown in Table 5.27). Subsequently the case of SNP1 equal to (3,3,80) and SNP2 equal to (3,3,20), (3,3,30) and so on could be considered.

In general term, in order to fix the redundancy problem, a pre-processing stage needs to be introduced where all the attributes are checked for their surplus of information and erased if necessary. A better analysis of the general problem of redundancy in SNPs dataset is extensively discussed in Chapter 6.

Table 5.27: $SNP-SNP\ Condition$: Allele frequencies and probability to contract a disease for each SNP value of the 2 recessive SNPs.

SNP value	DR SNP1	DR SNP2
A	3%	3%
B	3%	3%
C	80%	10%

5.5.2 Non linear interaction between attributes

The second issue that may arise when more than one SNP is considered, is the non linear interaction between different attributes. There may be cases when every SNP alone does not contribute to the susceptibility to contract a disease but considering for instance two SNPs together, there may be a particular combination of their values that is responsible for an increased risk of the disease. In order to give an idea of this problem, an example of dataset can be built, composed by two SNPs with this kind of non linear interaction between them, as shown in Table 5.28. It is important to make sure that the case-control ratio of the dataset is always set at a reasonable value over the threshold of 33%. In this specific example for instance, setting the allele frequency and the probability to contract a disease as shown in column $1-DRSNPs$ of Table 5.28, a dataset with 6.41% of cases over 10000 individuals is obtained. In order to build a dataset with an acceptable cases-controls ratio, one possibility is to set the probability to contract a disease to the ones shown in column $2-DRSNPs$ of Table 5.28. In this way a dataset with 49.87% of cases over 10000 individuals is obtained. There will be of course different solutions with different combinations of the disease risk and allele frequencies that could provide an acceptable case-control ratio. An extensive analysis of the problem of non linear interaction could be performed as a future work, considering the possibility to glue together the information related to 2 different SNPs in a single column of the dataset, as shown in Table 5.29.

Table 5.28: Example of non-linear interaction: Allele frequencies and probability to contract a disease for each values of the two SNPs.

SNP value	Allele frequency	1-DR SNPs	2-DR SNPs
A	36%	3%	3%
B	48%	4.5%	60%
C	16%	6.8%	80%
AA	–	3%	3%
BC	–	70%	70%
CB	–	90%	90%
number of cases	–	6.41%	49.87%

Table 5.29: Example of possible SNPs recoding for the non linear interaction issue

SNP1	SNP2	Coding
A	*A*	*AA*
A	*B*	*AB*
A	*C*	*AC*
B	*A*	*BA*
B	*B*	*BB*
B	*C*	*BC*
C	*A*	*CA*
C	*B*	*CB*
C	*C*	*CC*

5.6 Summary

This Chapter provides the reader with an overview of the problems that need to be faced in the creation of artificial SNPs datasets, with particular relevance to the case of the PE disease association study. The final aim of this work is to test the algorithms that have been used in the previous chapter for studying disease association in order to highlight their strengths and weaknesses.

There are many different aspects that need to be taken into consideration in this kind of task. Different diseases may arise in different ways, are caused by different genetic markers and genes, and have different probabilities to occur. More than one disease model can be employed in the analysis depending on the dominance or recessiveness of the genetic information. Allele frequencies for each single SNP differ for every marker and for each disease model chosen. SNPs can be similar to each other, creating redundancy in the initial dataset which in turn affects the performance of the analysis and the quality of the results. Having the possibility to build such an artificial dataset that is tunable by the user, allows a wide range of analysis of the pitfalls of a given technique.

In particular, the experiments performed in this work showed that the algorithms under analysis present a limitation in term of database size, as a population of 100,000 patients with one SNP cannot be analysed by the software. A combined analysis is also performed to check how disease risk and allele frequency can affect the statistical significance of the results obtained. The experiments show that for a disease risk lower than 50%, the algorithms are not able to provide statistical significance of the results. The general trend of Kappa value increases with the disease risk, as expected. An increased frequency of the common allele (1) causes a drop of the Kappa value for the recessive model and an increase of the Kappa value for the dominant and the mixed model. This is also an expected result as, when the rare allele (2) is the one responsible for an increased disease risk, if its frequency is very low, its effect will not be detected. All these results are proved to be significant through a multi-variate ANOVA test. In conclusion, the experiments show that the disease association with a single SNP needs to be evident (>50% of disease risk), in order to be detected by the algorithms and in this context, the recessive model is less likely to provide results which are statistically significant.

This review, is supposed to give a general idea of how complex the problem is and how much work needs to be done in order to provide a useful tool for facing this task. The final aim is to raise discussions and open questions in order to suggest possibilities for further extensive research.

Chapter 6

The RDsnp Method - A fast approximation to the LD algorithm.

6.1 Introduction

One of the most common and challenging problems for the analysis of large datasets is the detection and elimination of redundancy. This issue is becoming very pressing due to the continuously growing size of the biological datasets nowadays available in genetic studies. The current research is focused on decreasing the computational complexity of methods used for elimination of redundancy.

Within the study of Single Nucleotide Polimorphisms (SNPs), the concept of redundant information is directly associated with the definition of linkage disequilibrium. This function measures the association of two alleles at two different loci, revealing inheritance of genetic markers over generations. The current methods used to measure the redundancy observed between SNPs computes the pairwise linkage disequilibrium between genetic markers, providing the square correlation coefficient (R^2) as output value. These methods, in spite of a good accuracy, present big limitations in terms of computational complexity due to combinatorial explosion.

In this Chapter a new method for redundancy detection, called RDsnp, is presented. This makes use of an existing linkage disequilibrium tool, proposing an optimisation of its application. An overview of the problem is illustrated in the first paragraph, whereas in the Methods section the RDsnp method is explained in detail.

Following that, in order to test the proposed method, new databases need to be provided with different redundancy values. Therefore three new techniques are introduced to create an artificial dataset with a given redundancy, called ‘Copy’, ‘Permutation’ and ‘Rules’. One of these methoda, the Copy, is used in the experiments and a second one, Permutation, is employed to provide a validation of the results obtained with the previous one.

The Results section shows the characteristic of artificial datasets created with the proposed Copy method. Subsequently, several experiments using the new technique with different clustering techniques are shown in order to motivate the final choice of the best clustering algorithm, EM. The RDsnp method is tested with artificial datasets with a given redundancy as well as with examples of a real datasets. The performance and the computational time of the newly proposed approach is studied in order to demonstrate its benefits. Finally, the findings are summarised and new suggestions are made about some potentially interesting future improvements of RDsnp.

6.2 Redundancy Elimination in SNPs datasets

Nowadays the analysis of genetic data for different medical purposes is becoming increasingly important [15, 16]. It has been widely shown and confirmed that many human diseases are somehow related to the information contained in human genome [17]. If doctors could decode the information enclosed in the DNA chain, it would be possible to detect to a certain extent the susceptibility of every one of us to contract a specific disease [19, 153–156]. It could be feasible also to understand how every one of us reacts to different drugs and therefore it may be possible to provide every patient with a personalized treatment.

Due to the continuously improvement and optimization of the gathering data techniques in the genetic field, it is now possible to collect a huge amount of data in reasonable time. This has focused researchers’ attention on the increasing datasets size, which is becoming a top priority issue [157, 158].

Many tools and techniques for data mining, such as SAS [159], STATA [86], SPSS [160], Weka [112] or MySQL [161], are currently available for applications in every possible field. A common problem for these techniques is the necessity to reduce the time spent to perform a study. In order to achieve this goal, many studies have focused on decreasing the size of datasets without losing important information.

A SNP is usually coded with two different numbers, referred as allele: one number refers to the genetic information located in one chromosome and the other number refers to the information located in the paired chromosome. The number of different values for each allele can be 2 or 3, when an allele can have only 2 different values, the respective SNP is called bi-allelic [8].

Since sequencing the entire human genome was completed on 2003, large amounts of data have been available for medical studies. The current SNPs datasets can reach sizes of hundreds of thousands of SNPs, creating lots of problems for the pre-processing and analysis approaches. The idea of reducing the size of datasets by elimination of redundant information is related to the specific definition of redundancy commonly applied to SNPs studies. This definition arises from the correlation, or lack of independence, between SNPs in close proximity, known as linkage disequilibrium [162].

This Chapter presents an approach to reduce the SNPs database size by the elimination of information which is redundant. In general terms the presence of redundancy implies a surplus of information in the database which can be inferred from other subgroups of data. This superfluous data is therefore not relevant for further analysis and can be eliminated without compromising the content of the original data. In this specific case the dataset can be seen as a matrix composed by rows representing patients and columns consisting of SNPs. The aim of this new method is to decrease the size of the matrix by reducing the number of columns. This goal is achieved by eliminating sets of SNPs whose values can be inferred by others.

The problem of SNPs tagging, which consists of selecting a smaller amount of informative SNPs from the original dataset, is essentially a feature selection problem from the machine-learning point of view [163]. A huge amount of work can be found in the literature on feature selection in the machine learning community. However, in this context, attention is focused on the specific applications for SNPs analysis. There are several papers in the literature which show different approaches to the redundancy elimination problem [75, 164–168]. In Chapter 2, an extensive analysis can be found on the feature selection problem tailored to SNPs.

The new method proposed is called RDsnp and makes use of the linkage disequilibrium function, currently implemented in more than one language. This function computes the pair wise linkage disequilibrium between all the genetic markers included in the input matrix through the correlation coefficient (R^2). The new RDsnp method makes a different use of this function from the common one. Instead of calculating the redundancy between every pair of SNPs, it measures the redundancy between each SNP and few random columns, comparing then the set of R^2 values obtained for each SNP.

In order to group different sets of SNPs, characterized by different redundancy values R^2, an appropriate clustering method needs to be applied. Within this study the most common clustering techniques used for genetic applications have been analyzed and tested. These include agnes, daisy, pam, clara [169], hierarchical clustering, K-Means [112] etc. Following an assessment of the different results obtained, as shown in the Result section, the evidence of the best final results has been obtained with the application of the Expectation Maximization algorithm.

This technique is somehow similar to the better known K-Means technique. In Chapter 2 a detailed analysis of the used clustering techniques is shown.

6.3 RDsnp - Redundancy Detection for SNPs datastes

In the first part of this section the steps of the RDsnp method are shown. The second section explains the creation criteria for the synthetic datasets which will be used to test the new technique.

6.3.1 Method

There are different tools that can be considered for redundancy measurements such as MRMR in Matlab [170] or Duplicate Remover in SQL [171]. The one that has been chosen in this study is built for SNPs application and is therefore widely used and known in genetic analysis. It is called LD and in R language it is a function implemented by Gregory R. Warnes [172], contained in the package called ‘SNPassoc’, released on April 2009 [173]. The LD function computes the pair wise linkage disequilibrium between genetic markers. One of the output values of LD function is the correlation coefficient (R^2) which is used in the proposed method for testing redundancy between SNPs.

The squared correlation coefficient R^2 is a formal measure of linkage disequilibrium. Considering two bi-allelic (i.e. SNP) loci with allele frequencies p_1, p_2, q_1, q_2 and haplotype frequencies h_{11}, h_{22}, h_{12}, h_{21} a measure of the magnitude of linkage disequilibrium is given by the following definition [174]:

$$R^2 = (h_{11}h_{22} - h_{12}h_{21})/p_1p_2q_1q_2$$

In selecting SNPs for genotyping studies avoiding redundancy, it is necessary to detect the variation within a region of the genome by ensuring that the SNPs chosen are adequate proxies for the other SNPs in the region. In this scenario, $R^2 \geq 0.8$ is often used as a threshold for selecting redundant SNPs [175, 176].

The LD procedure, which has been chosen in this study, analyses a matrix of SNPs and, by performing a test for linkage disequilibrium, it calculates the redundancy through the assessment of the parameter R^2. In this current method, the redundancy is calculated considering every possible combinations of SNPs pairs within the data matrix. Given a set of n SNPs, the total number of all possible combinations of k SNPs ($i = 1 \dots k$) taken from the set of n is

$$C_{k,n} = \frac{n!}{k!(n-k)!}$$

In this specific case, as the measure the redundancy is always between two SNPs, $k = 2$. Therefore the combinations of SNPs become:

$$C_{2,n} = \frac{n(n-1)}{2} \cong \frac{n^2}{2}$$

For instance, in an input dataset of 1000 SNPs ($n = 1000$) there are almost $5 \cdot 10^5$ combinations of SNPs to be measured for redundancy. If the dataset is composed of 10000 SNPs, the combinations to be measured rise to $5 \cdot 10^7$ pair of SNPs. Considering that with the new technologies a reasonable number of 500k SNPs can be gathered, the amount of needed measurements can reach more than 100G. This enormous set of operations may affect the performance of the standard machines nowadays available for these analysis.

The new idea is to decrease the computational complexity of the current method, avoiding the redundancy measurement for every possible combination of SNPs pairs taken from the dataset. The new approach is developed starting from the following steps:

(1) Every single SNP from the initial dataset is compared with a random SNP column created artificially: the redundancy is measured between each SNPs of the dataset and the random column applying the LD function.

(2) All the SNPs which show the different redundancy value R^2 measured with the random column are grouped in different clusters. Every SNPs that belong to the same cluster has the same redundancy against the same random column which implies it is likely to be also redundant against the SNPs belonging to the same group. The degree of redundancy between SNPs of the same cluster is probably different from the one obtained with the given random column.

(3) In order to measure the real value of redundancy between SNPs belonging to the same cluster it is sufficient to select two SNPs, from the same cluster, that can be representative of the cluster. The LD function can be applied to these two representative SNPs to calculate their real redundancy. The sampling of the representative two SNPs in this procedure is made randomly; for further discussion on this, the reader is referred to Section 6.7.

(4) Point 3 is applied to each cluster, in order to calculate the redundancy between the SNPs belonging to each cluster.

(5) Any cluster that contains SNPs with a value of redundancy greater than 0.8 is considered a group of redundant SNPs. Only one SNPs is kept for each of these clusters and the remaining SNPs are eliminated. This final SNP, representative of the cluster, is selected, once again randomly.

By applying this mew methodology it is possible to reduce drastically the amount of computational complexity. In this case indeed, if the input dataset is composed by n SNPs and a random column is considered in the analysis, the number of measurements to be performed is equal to n.

A drawback of this first approach is the bias that a random column can introduce, as the value of redundancy between a SNP and the given random column depends also on the random column values. For this reason it is more reasonable to extend this technique to a greater number of experiments, each one with a different random column. Every time a random column is chosen, different value of redundancy is obtained against the same SNP and against the same SNPs belonging to the same cluster created. In this way, with n SNPs and c random columns, the amount of measurements to be performed is equal to:

$$C_{c,n} = n \cdot c$$

Applying this formula to 1000 SNPs and 5 random columns, only 5000 comparisons are needed to test for redundancy and with a dataset of 500k SNPs and 5 random columns the amount of measurements results only 1M instead of 100G with the original method. In the end, all the results from each experiment are combined in one single output. The final RDsnp method is therefore developed through the following steps:

(1) Every single SNP from the initial dataset is compared with c SNP random columns created artificially : the redundancy is measured between each SNPs of the dataset and each random column applying the LD function.

(2) A vector for each SNP has been created. This vector dimension is equal to the number of random columns used and each of its component shows the redundancy value of the given SNP against each used random column.

(3) A clustering technique is then applied to this population of vectors in order to group the SNPs with the most similar value of redundancy. This procedure may require the calculation of the distance matrix of the whole population of SNPs. For the distance matrix calculation a measurement to high combinatorial number of SNPs needs to be applied. This may create disadvantages in terms of computational complexity making RDsnp comparable to the original technique. In fact, this population of vectors have a very short length (c random columns). The input matrix used in the original method, instead, is composed by SNPs columns whose length corresponds to the number of patients analyzed, usually consisting of few thousands units.

(4) The real value of redundancy of each group of SNPs is calculated, as before, randomly picking two SNPs up from the same group and applying the LD function.

(5) The groups of SNPs (each one for each vector), that show a redundancy value $R^2 \geq 0.8$ are considered redundant. Only one representative SNPs of these clusters is kept, randomly chosen.

A framework of the new methodology is shown in Figure 6.1, 6.2 and 6.3.

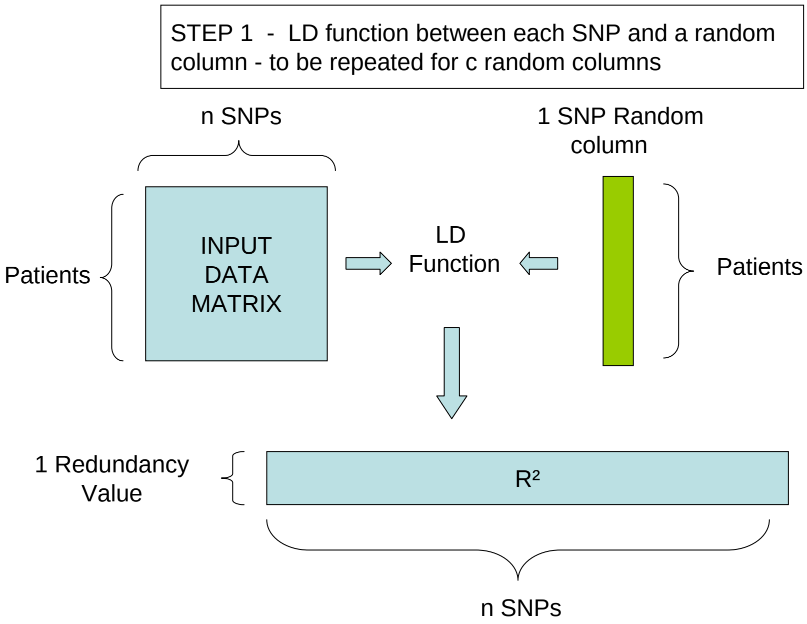

STEP 1 - LD function between each SNP and a random column - to be repeated for c random columns
n SNPs
1 SNP Random column
Patients
INPUT DATA MATRIX
LD Function
Patients
1 Redundancy Value
R²
n SNPs

Figure 6.1: The flow chart model of the RDsnp method - In the first step the redundancy between n SNPs and c random columns is calculated

As there are different techniques that can be used for clustering, a variety of clustering algorithms have been examined in order to find the best algorithm to be used in this application.

The Hierarchical Clustering is based on the definition of similarity or dissimilarity between instances of the datasets. The elements that are more similar are grouped together in a procedure composed by sequential steps. At each step two clusters are grouped together depending on their similarity. Different solutions can be chosen in the end depending on

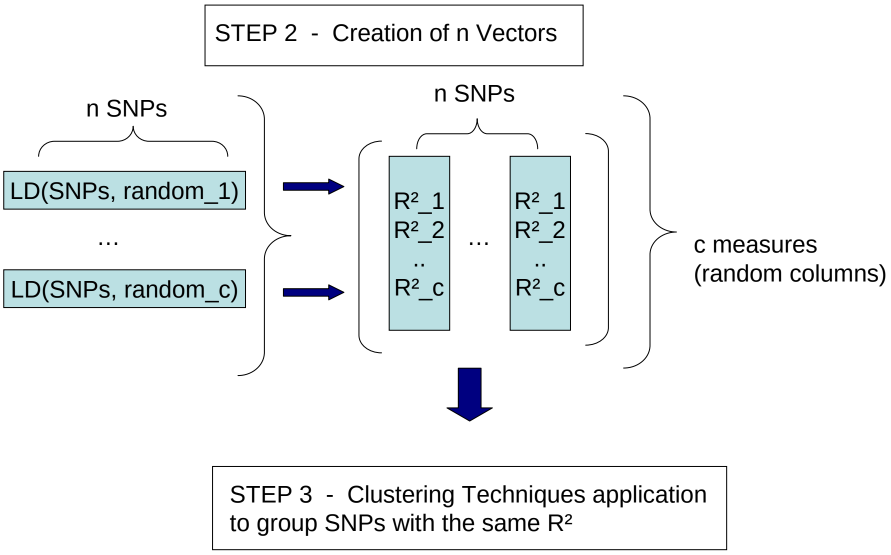


Figure 6.2: The flow chart model of the RDsnp method - In the second step, *n* vectors are created out of *n* SNPs. In the third step the Clustering techniques are applied

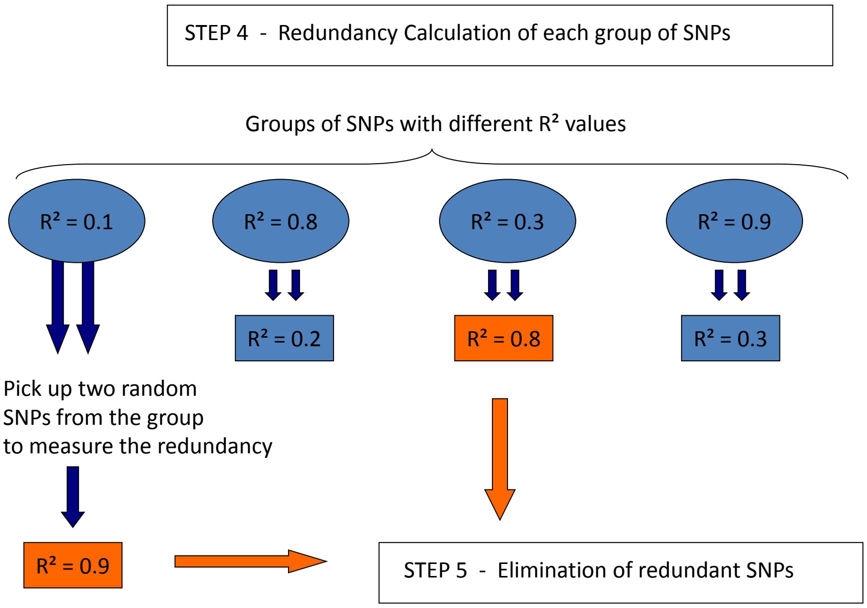

STEP 4 - Redundancy Calculation of each group of SNPs
Groups of SNPs with different R² values
R² = 0.1
R² = 0.8
R² = 0.3
R² = 0.9
R² = 0.2
R² = 0.8
R² = 0.3
Pick up two random SNPs from the group to measure the redundancy
R² = 0.9
STEP 5 - Elimination of redundant SNPs

Figure 6.3: The flow chart model of the RDsnp method - In the forth step the real redundancy between SNPs belonging to the same cluster is calculated. In the fifth step SNPs with $R^2 \geq 0.8$ are detected for elimination.

the number of clusters that are required by the user. The 'k-means' algorithm is based on a creation of different clusters defined initially by random centroids. Following an optimization of the first random choice is performed until the results stabilized into an optimal center for each group. As the final solution is biased by the initial choice, the software is usually run several time and the best result is chosen. The EM algorithm is a statistical approach that assigns each instances with a probability to belong to a given cluster instead of placing them directly into one of them. In the R language, the function that performs the EM clustering technique is called 'Mclust' and it is included in the package 'mclust' [177] published on July 2009. For a better analysis of these three techniques, the reader is referred to the Chapter 2.

In this study, it is important to remember that the number of clusters present in the dataset is not known beforehand as the number of different redundancies in the input matrix is unknown. For this reason, it would be reasonable not to consider any clustering technique which requires the number of clusters as an input parameter. Nevertheless, in the results section an example of K-means clustering technique is applied to the RDsnp method in order to show how this technique perform in this specific application. This can be achieved building an artificial dataset with a known number of redundant clusters. Moreover, there are several variants of the basic 'k-means' technique which have been developed for different applications together with supporting analysis for choosing the best number of clusters.

6.3.2 Artificial Dataset Creation

In order to test the main method explained in the previous section, a good source of datasets needs to be provided. As sometimes it is not immediately available the kind of dataset needed for this study, an analysis upon synthetic datasets creation has been carried out also in this Chapter for this specific application. In this way it is possible to create a database with given features necessary to test specific limitations of the method.

There are different ways that can be used to create an artificial dataset. In this section, three different approaches to build a redundant database are shown in order to give a confirmation and validation of the results obtained with each one of them.

As the target dataset is composed by SNPs, the first issue that needs to be faced is the allele frequency setting. In order to build a realistic dataset, the allele frequency needs to be set to a value which is reasonable and commonly accepted. Every SNP usually has a different frequency of the two alleles (1 and 2). If the allele 1 is the most common and the allele 2 the most rare one, a realistic occurrence of the allele 1 can be set as 80%. Consequently, the frequency of the allele 2 can fairly be set to 20%. A dataset composed by a fixed number of SNPs with a biological distribution of allele can then be created. The second step is the creation of a given redundancy between SNPs.

Artificial Dataset - Copy Method

Using the first method, called 'Copy Method', two different SNPs columns are created with the same allele frequency. Every component of a SNP column is composed by the value of the given SNP for each patient in the dataset.

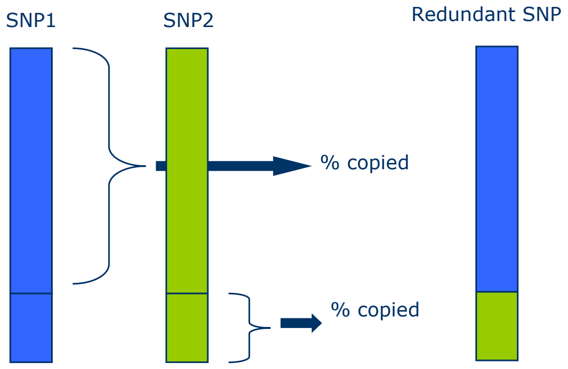


Figure 6.4: Copy Method for artificial dataset creation

In order to create a new SNP, redundant to the first SNP, a percentage of the first SNP has been copied and the rest has been copied from the second SNP. In this way, provided that the length of the SNPs are big enough, the distribution of alleles has been preserved, Figure 6.4. The length of the SNP is given by the number of patients included in the dataset and a population of 1000 individuals can be considered a reasonable choice for

Table 6.1: Percentage of SNP to be copied or permuted for a given redundancy

Redundancy R2	% of SNP
0.9	6.8
0.8	12.5
0.7	18.5
0.6	25.8

preserving the alleles distribution. Following a sequence of simulations steps, the right percentage of SNP, to be copied from the first SNP, has been detected in order to obtain a given redundancy in a population of patients, as shown in Table 6.1.

As mentioned before, the redundancy value is given by the squared correlation coefficient R^2 in the range 0-1. As shown in Table 6.1 in order to create two SNPs for instance with a redundancy of $R^2 = 0.9$, 6.8% of the first SNP needs to be copied in to the second one. These results are plotted in Figure 6.5, showing the linear increasing of the percentage of SNP column copied with the decreasing trend of the R^2.

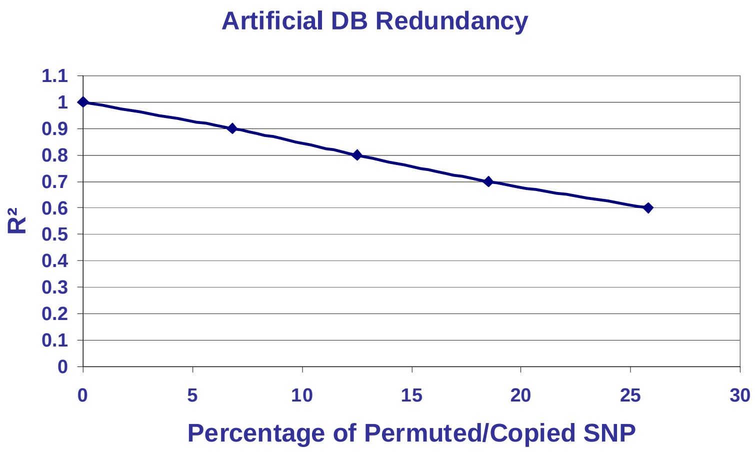


Figure 6.5: Percentage of SNP column to be copied or permuted

Artificial Dataset - Permutation Method

In order to have a validation of the first method a different approach has been used to create a similar artificial dataset and compare then the results. In Figure 6.6 the second method called 'Permutation Method' is shown.

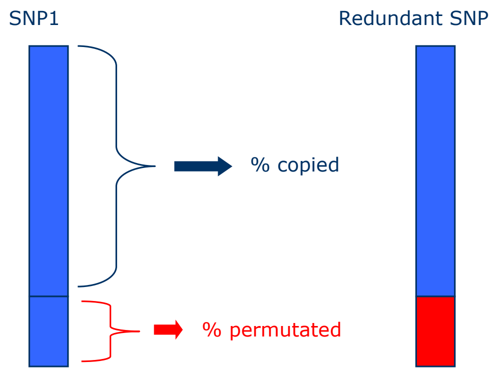


Figure 6.6: Permutation Method for artificial dataset creation

A SNP column is created with the agreed distribution of allele as a reference. A second SNP is then created, redundant to the first just by coping a certain percentage of the previous SNP column and permuting the rest of it. In this way the allele distribution of the patients population is preserved and at the same time a SNP with a given value of redundancy, depending on the permutation ratio, is created. In order to determine the percentage of SNP column to be permuted for each value of R^2, a sequence of simulation tests were performed until the best approximation was found. These amounts resulted to be the same found for the 'Copy Method' as shown in Table 6.1 and Figure 6.5, providing a validation of the previous method for the chosen size of 1000 patients.

Artificial Dataset - Rules Method

In the last method, a new artificial dataset is built following some rules that provide a given amount of redundancy. The basic idea is to the following: If the first SNP has an

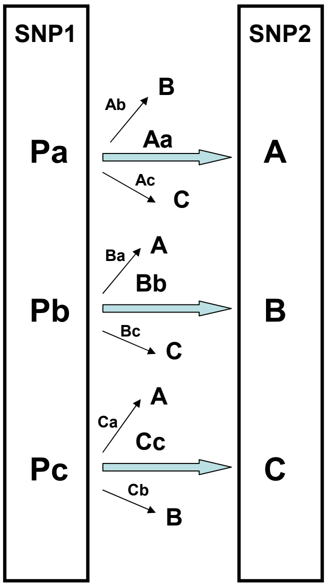


Figure 6.7: Rules set for redundancy realization

'A' then the second SNP has a 'A' with a probability equal to 'r', where 'r' is the chosen redundancy. Otherwise, the second SNP has a 'B' or 'C' with (100-'r')/2 probability in either cases. The same rule is extended to the case of the first SNP having a 'B' or 'C' as shown in Figure 6.7. In this way the percentage of redundancy, that was initially set, is preserved but the original distribution of the SNP allele (occurrences of 'A', 'B', 'C') is lost. In Figure 6.8 an example of rules set is shown for redundancy equal to 90%. Performing several simulations, the best results for simulating datasets is obtained with R^2 equal to 0.9, 0.8, 0.7 and 0.6, as shown in Table 6.2.

6.4 Experimental Results

This section shows the results obtained from creating an artificial dataset with different degrees of redundancy applying the methods previously discussed. In order to synthesize the analysis carried out in this Chapter, only two of these methods are used in the study, as this is enough for providing a validation of the results. The Copy method is employed

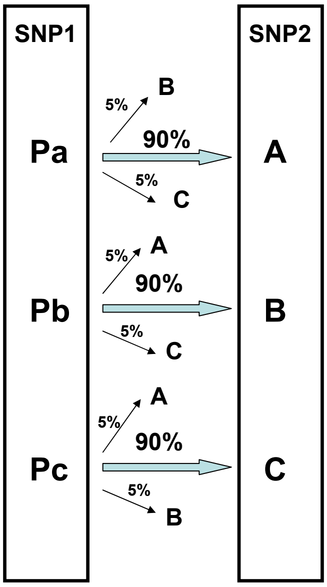


Figure 6.8: Example of rules for redundancy equal to 90%

at the first stage and the Permutation method is subsequently applied in order to have a validation of the results from the first method.

In order to find the best sequence of steps to follow, the appropriate settings of the parameters used and the most successful clustering technique, different ways to test the RDsnp technique are shown. In the first step an artificial dataset is created in order to know exactly the amount of redundancy present between each SNP. In a second stage the technique is applied to a real dataset to possibly detect any degree of redundancy between

Table 6.2: Allele Frequencies setting for different value of redundancy for a dataset built with the Rules Method

Allele Frequency Set	$R^2 = 0.9$	$R^2 = 0.8$	$R^2 = 0.7$	$R^2 = 0.6$
(Aa, Ab, Ac)	$(98,1,1)$	$(95,3,2)$	$(92,4,4)$	$(88,8,4)$
(Ba, Bb, Bc)	$(1,98,1)$	$(3,95,2)$	$(4,92,4)$	$(8,88,4)$
(Ca, Cb, Cc)	$(1,1,98)$	$(3,2,95)$	$(4,4,92)$	$(8,4,88)$

Table 6.3: Degrees of Redundancy present in the artificial dataset

SNPs Names	Redundancy (R2)
1 – 20	Identical(1)
21 – 40	0.9
41 – 60	0.8
61 – 80	0.7
81 – 100	0.6
101 – 120	Random

real SNPs.

6.4.1 Results: LD and the Artificial Datasets

With the Copy Method a new dataset has been created, composed by 1000 patients and n 120 SNPs, whose redundancy is set as shown in Table 6.3. Applying the original LD function to the new dataset, it is clear that the six different group of SNPs can be easily detected within the initial dataset, see Figure 6.9. The first 20 are identical and therefore have a redundancy value of one, the second 20 are supposed to be redundant of $R^2 = 0.9$ and this is confirmed by the LD function as shown in Figure 6.9. The same validation applies for the other groups of SNPs.

Also with the Permutation Method a new dataset composed by 1000 patients and 120 SNPs is created, with the same redundancy setting as before (Table 6.3). Then, applying to original LD function to this dataset, it is again shown that the Permutation Method provides a good accuracy for the redundancy setting (Figure 6.10).

A similar experiment is repeated also for the dataset created with the Rules method and the LD function is applied to confirm the validity of the method, as shown in Figure 6.11.

6.4.2 RDsnp Results from Artificial Dataset

As validated by the LD function, the proposed Copy technique has provided good results in terms of creating a dataset with a given redundancy and this has been confirmed by the

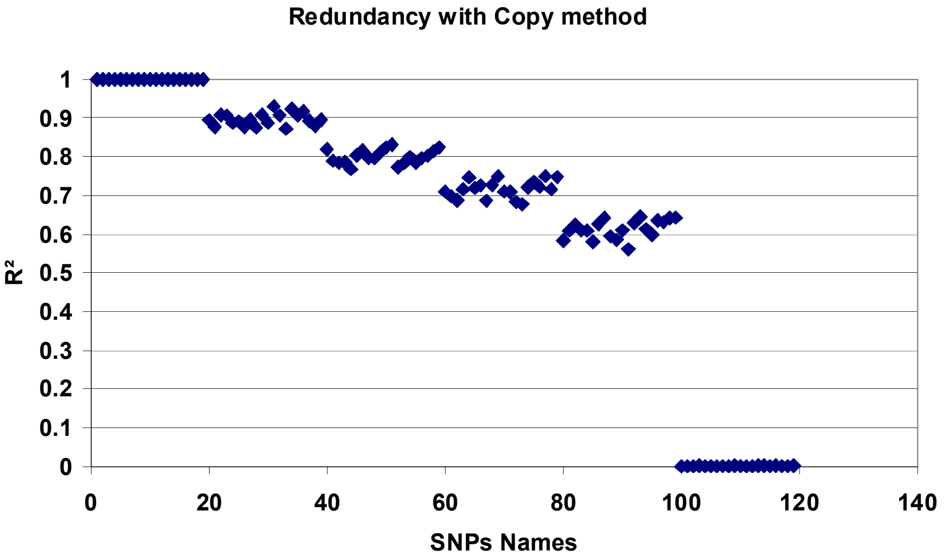


Figure 6.9: Results from LD applied to an artificial DB composed by 1000 patients and 120 SNPs. The first 20 SNPs are identical, the second 20 have $R^2 = 0.9$, the following 20 have $R^2 = 0.8$ and so on. The last 20 are just random.

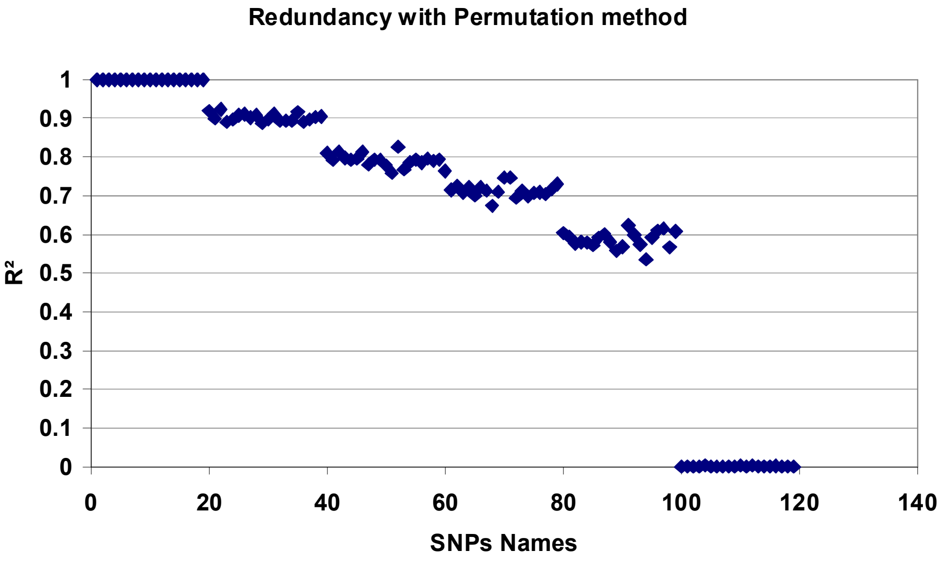


Figure 6.10: Results from LD applied to the artificial DB

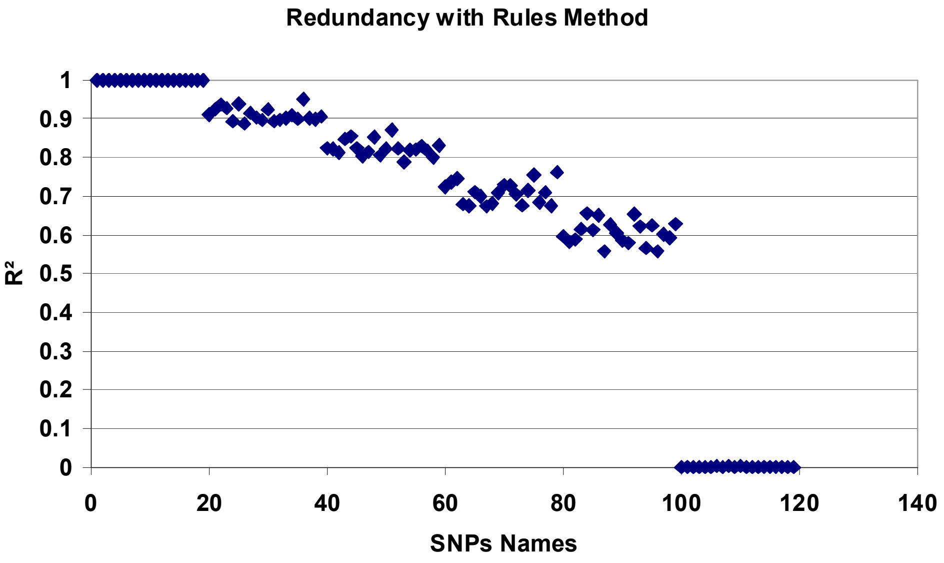


Figure 6.11: Results form LD applied to the artificial DB built with Rules Method

alternative Permutation method. Under this validation, the Copy method will be used to built the datasets for the further experiments.

There are different parameters that can be set in order to check for the performance trend of the RDsnp technique: number of SNPs n and patients in the dataset, number of random columns c to be used in the analysis and different clustering techniques. For this reason, in the course of the analysis different databases have been created to perform various kind of tests in order to have a wide overview of the proposed methodology. For each different parameters setting, discussed in each of the following sections, only the techniques which provided the best results are mentioned and shown.

Dataset(1000x60) and Hierarchical Clustering

Initially a DB of 1000 patients and 60 SNPs has been created, whose the first 20 SNPs are redundant of 0.9, the next 20 are redundant of 0.8 and the last 20 are just random. After applying the new method with three different random columns, the results are plotted as distribution of R^2 with its mean and variance in Figure 6.12.

The first column of this Figure shows the R^2 value for each random column (represented by each row of the picture) and it is clear that it is not easy to distinguish between the redundant SNPs (the first 40) and the random ones (the last 20). The variance graphics, which are plotted in the second column of the Figure 6.12 show that the last 20 SNPs are more spread that the previous ones, but it is still difficult to set a fixed threshold to identify them. For this reason five different measurements of the redundancy against five different random columns have been performed and, this time, the results have been put together in a single matrix. This output is then compose by 60 columns (or vectors), as the number of SNPs and c=5 rows as the number of different random columns. As the next step, in order to detect the two groups in this new population (the redundant and the random SNPs), another experiment has been performed. If SNP A and SNP B are redundant and therefore similar, they should have a similar redundancy value against the same reference SNP (in this case the random column created). If c random columns are used, SNP A should have very similar redundancy values to SNP B when compared to each of the c random columns. Plotting these values in a c dimensional space will result in locating SNP A and SNP B in a position close to each other. This is due to the similar

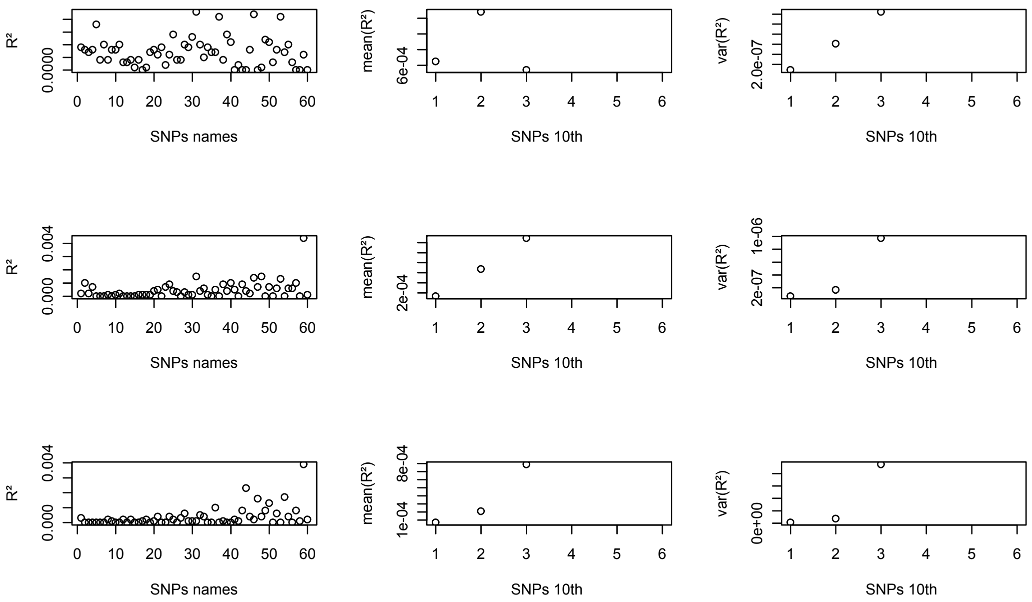


Figure 6.12: Results from LD applied to the artificial DB with three different random columns. Each row of the picture represent a result from each random column. The columns of the picture represent, from left to right, the R^2 value, the R^2 mean and the R^2 variance, for each SNP represented on the x axis.

values that A and B have for each of their correspondent *c* coordinates. For this reason, if the points obtained in the space are clustered, SNP A and B are expected to belong to the same cluster. This type of analysis requires the calculation of the distance of every point between each other. In order to avoid this high computational process, an alternative solution has been explored. The distances between the points and the origin of the metric system are plotted in order to see if any patterns could be detected to distinguish the first 40 SNPs (redundant) from the last 20 ones (random). The results are shown in Figure 6.13 and once again it is difficult to detect the first 40 redundant SNPs from the last 20. This is probably due to the fact that even if redundant SNPs are close in the space and grouped in a cluster, they may have the same distance to the origin as another group of SNPs, still redundant but with a different degree of redundancy.

Consequently, the distance matrix is calculated and a clustering technique is then applied to detect the groups of SNPs (represented by the vectors) with the same redundancy value. The clustering algorithm that gave the best results is Hierarchical Clustering with 'McQuitty' method of Aggregation. As shown in the Dendrogram of Figure 6.14, this technique is able to detect two different groups: the first one includes all the redundant SNPs except for one. Moreover, it includes two random SNPs. The second cluster includes all the random except for two and beside, it includes also one redundant SNP. In more precise terms, this procedure gives 2 false positive SNPs over 40 and 1 false negative SNP over 20.

Dataset(1000x6000) and Kmeans

As next step, in order to evaluate the influence of the number of SNPs to the performance, a database composed by 1000 patients and 6000 SNPs is created, whose SNPs, the first 2000 are 0.9 redundant, the second 2000 are 0.8 redundant and the last 2000 are just random. After applying the new method with *c*=5 random columns, the distance matrix of the SNPs vectors is calculated and then different clustering techniques are applied. The best results are obtained by the Kmeans approach setting 2 target clusters. Over the redundant SNPs cluster, 68 are random SNPs which means a false positive rate of 1.7% over 4000 SNPs. Whereas within the non-redundant cluster there are 1162 false negative SNPs, making the false negative rate equal to 58.8%. If on one hand these results are very

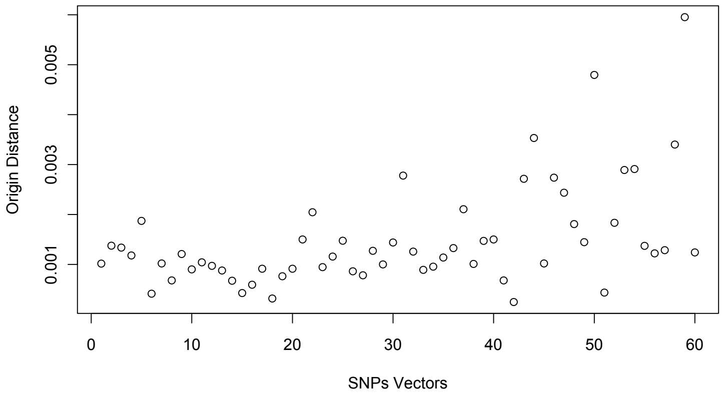


Figure 6.13: Distance from the origin of the SNP vectors created. The vector dimension is equal to the number of random columns used and each of its component shows the redundancy value of the given SNP against each used random column.

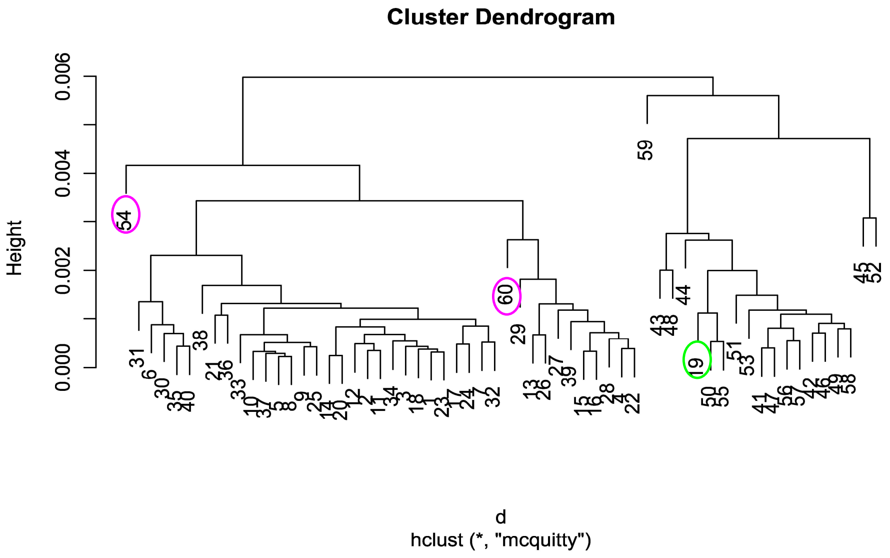


Figure 6.14: Results for the Hierarchical Clustering technique with McQuitty method of aggregation. The red circles highlight the 2 false positive results whereas the green circle detects the false negative SNP found. The numbers at the bottom of the Dendrogram represents the name of the SNPs

Table 6.4: Redundancy versus clusters detected by EM algorithm on 1000x1000 input matrix

Clusters	1	2	3
Redundant SNPS	332	168	0
Random SNPS	0	0	500

promising, on the other hand, there are still a relevant mistake to be amended. For this reason in the next step other clustering techniques are used to detect the best one for the purpose of this study.

Dataset(1000x1000) and EM algorithm

In this specific experiment, a dataset composed by 1000 patients and 1000 SNPs was created, whose SNPs, the first 250 are 0.9 redundant, the second 250 are 0.8 redundant and the last 500 are just random. Applying the Expectation Maximization algorithm to the results from 5 set of random columns an excellent result is obtained as shown in Table 6.4.

The EM algorithm detects three clusters, whose the first 2 include exactly all the redundant SNPs, whereas the third cluster includes all the random SNPs, providing an accuracy value of 1.

6.4.3 RDsnp Results from Real Dataset

The real DB that has been used in these experiments is composed by a group of mothers checked for the pregnancy condition of Pre-eclampsia and it has been created from data taken at different European clinics. The dataset is characterized by 339 patients representing the rows of the input matrix and 26 SNPs, each one with two different allele encoded with the number 1 or 2. The genotype attributes are therefore the following:

- AGT gene: SNPs 1-8, alleles 1 and 2
- AGTR1 gene: SNPs 9-12, alleles 1 and 2
- TNF gene: SNPs 13-16, alleles 1 and 2

- F5 gene: SNP 17, alleles 1 and 2
- NOS3 gene: SNPs 18-22 and 24, alleles 1 and 2
- MTHFR gene: SNPs 25, 26, alleles 1 and 2
- AGTR2 gene: SNP 27

Both the original method and the proposed one detect a redundancy of 0.879 between the first and the fifth column of the input dataset which correspond to SNP1 and SNP5. Due to the small number of SNPs the performance comparison give better results for the original method than for the new one. The results were obtained in 6.38 seconds using the LD function against 8 seconds spent for performing the new method with three random columns setting.

In order to check how the new technique performs with a larger real database, the original database has been replicated two to six times and the run time and accuracy of the results have been checked. Their average and standard deviation over 20 runs are plotted in Figure 6.15 and Figure 6.16, respectively. It is clear from Figure 6.15 that the speed of the RDsnp becomes competitive when the number of SNPs exceeds 60 (in this specific case). Also, the accuracy shows improvements with the increased number of SNPs considered in the analysis.

6.5 Performance Analysis

In this section a performance analysis is carried out, in order to show the benefits of the new technique. Considering the results from a general method divided by True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN), the study is based on the following definitions:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Sensitivity = \frac{TP}{(TP+FN)}$$

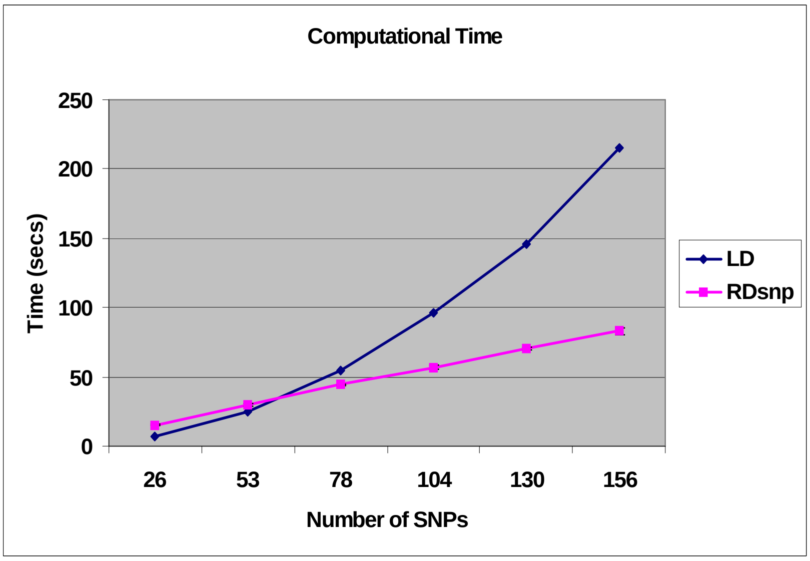


Figure 6.15: Run Time comparison for Real Database: Time in minutes versus number of SNPs for both RDsnp with 5 columns and the original LD function

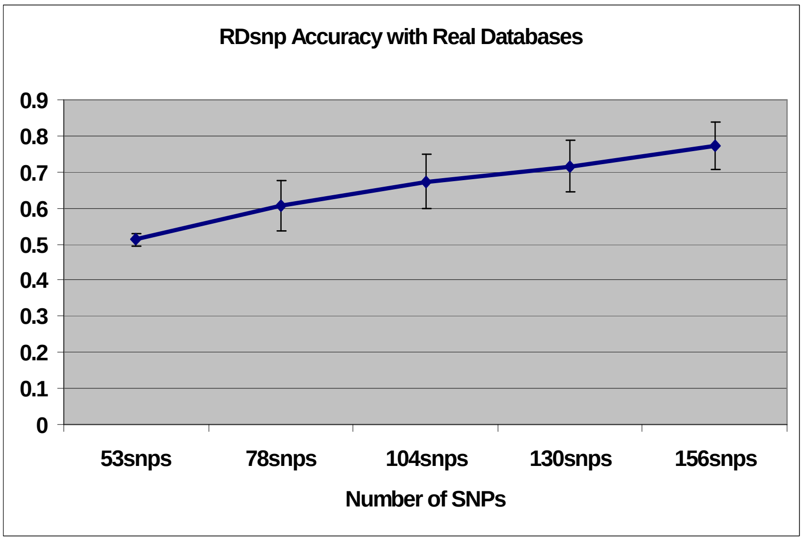


Figure 6.16: Accuracy Real Database versus number of SNPs for RDsnp with 5 columns

$$Specificity = \frac{TN}{(TN+FP)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

Several groups of experiments have been performed, setting every time a different value for the number of random columns c and a different value for the number of SNPs n used in the artificial dataset. The different number of SNPs considered for each trial are: n = 200, 400, 600, 800 and 1000 SNPs. The different sets of random columns used for each SNPs dataset are: c = 3, 4, 5 and 6. All these datasets are created in a way that half of the SNPs are random and half are redundant - in particular one quarter has a redundancy of 0.8% and one quarter has a redundancy of 0.9%. For each number of random column setting (3,4,5 and 6), 20 different experiments have been run and the averages and standard deviations of the results performance have been calculated and displayed in different graphics. In Figure 6.17 and Table 6.5 the Specificity of the new method applied to the described datasets is shown. With the same specification explained above, Figure 6.18 and Table 6.6 show the Sensitivity, Figures 6.20 and Table 6.7 show the Accuracy and Figures 6.19 and Table 6.8 show the Precision of the new method. While the Sensitivity appears to have wider fluctuations, the other three parameters are more stable. In particular, the Sensitivity, Accuracy and Precision tend to increase with a higher number of random columns used, without showing particular improvement with the number of SNPs analysed. Moreover, the level of these three parameter is reasonably high for any size of dataset analysed.

In conclusion, it is clear that when the number of random columns is fixed to the value of five, a good performance is preserved, obtaining an Accuracy of 0.75. This means that three quarter of the SNPs included in the dataset are correctly detected as redundant or not redundant. Additionally, with a Precision of 0.8, the results obtained are confirmed by repeated experiments for 80 times over 100 trials.

Table 6.5: Specificity comparison versus number of SNPs. Four different number of random columns *c* (3,4,5,6) have been chosen to run the RDsnp(c) method

Number of SNPs	200	400	600	800	1000
RDsnp(3)	0.83	0.80	0.82	0.86	0.85
RDsnp(4)	0.80	0.82	0.84	0.86	0.84
RDsnp(5)	0.88	0.86	0.83	0.85	0.84
RDsnp(6)	0.87	0.90	0.90	0.89	0.87

Table 6.6: Sensitivity comparison versus number of SNPs. Four different number of random columns *c* (3,4,5,6) have been chosen to run the RDsnp(c) method

Number of SNPs	200	400	600	800	1000
RDsnp(3)	0.55	0.63	0.64	0.53	0.51
RDsnp(4)	0.69	0.64	0.66	0.51	0.68
RDsnp(5)	0.73	0.73	0.74	0.75	0.65
RDsnp(6)	0.81	0.75	0.73	0.68	0.78

Table 6.7: Accuracy comparison versus number of SNPs. Four different number of random columns *c* (3,4,5,6) have been chosen to run the RDsnp(c) method

Number of SNPs	200	400	600	800	1000
RDsnp(3)	0.68	0.71	0.73	0.70	0.68
RDsnp(4)	0.74	0.75	0.75	0.68	0.76
RDsnp(5)	0.80	0.79	0.79	0.80	0.75
RDsnp(6)	0.84	0.82	0.81	0.78	0.83

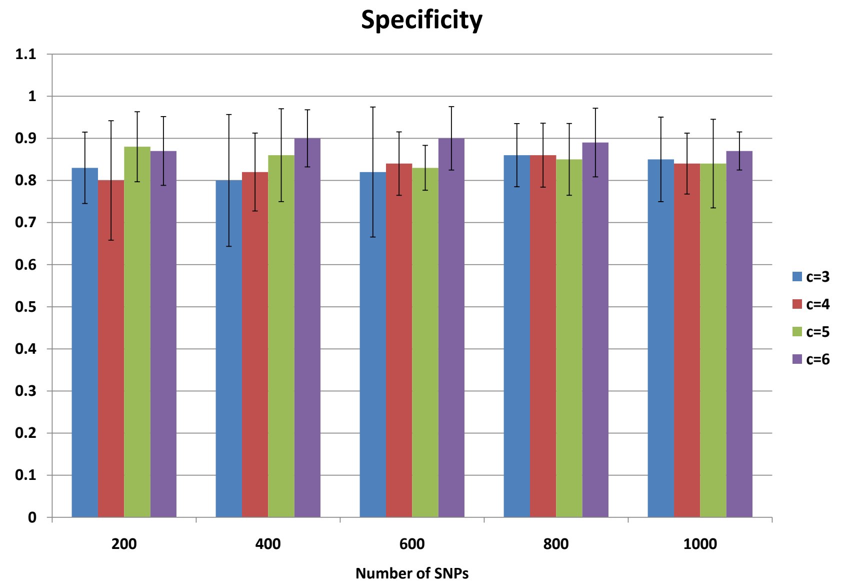


Figure 6.17: Specificity comparison: Time in minutes versus number of SNPs for different numbers of random columns(c)

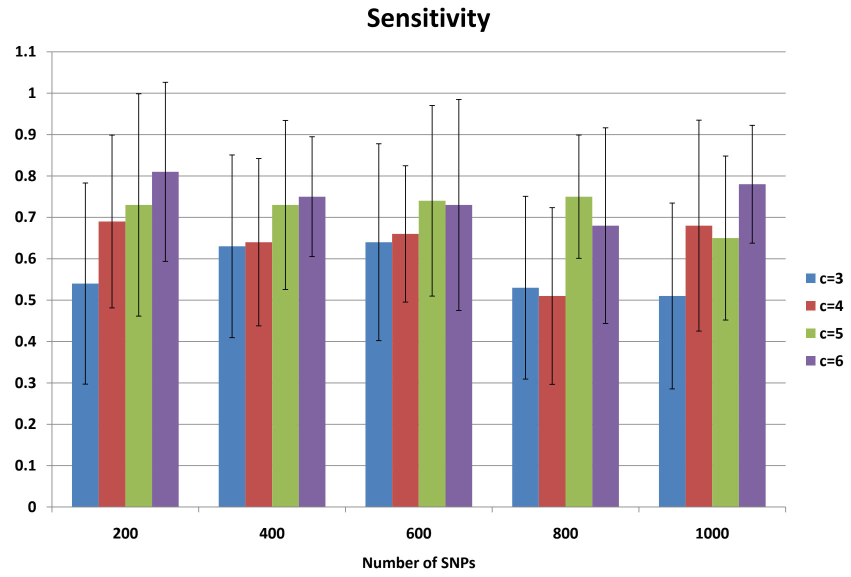


Figure 6.18: Sensitivity comparison: Time in minutes versus number of SNPs for different numbers of random columns (c)

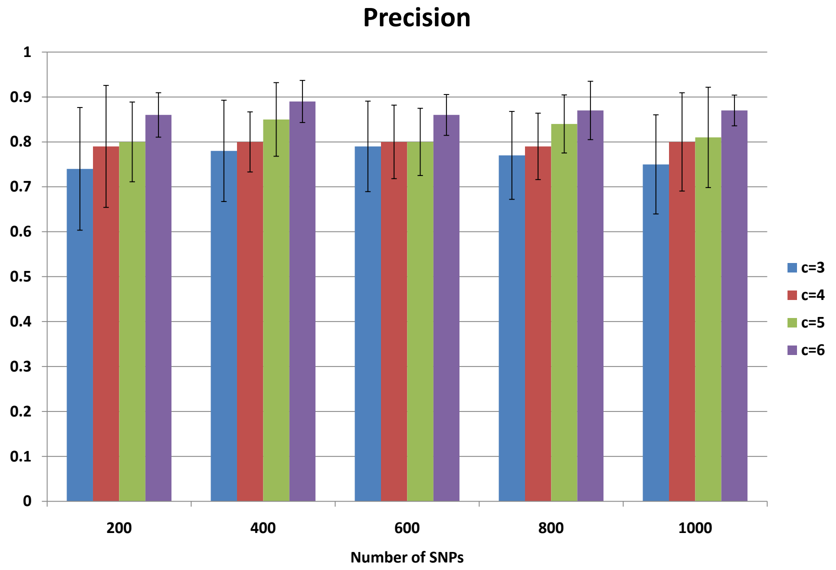


Figure 6.19: Precision comparison: Time in minutes versus number of SNPs for different numbers of random columns (c)

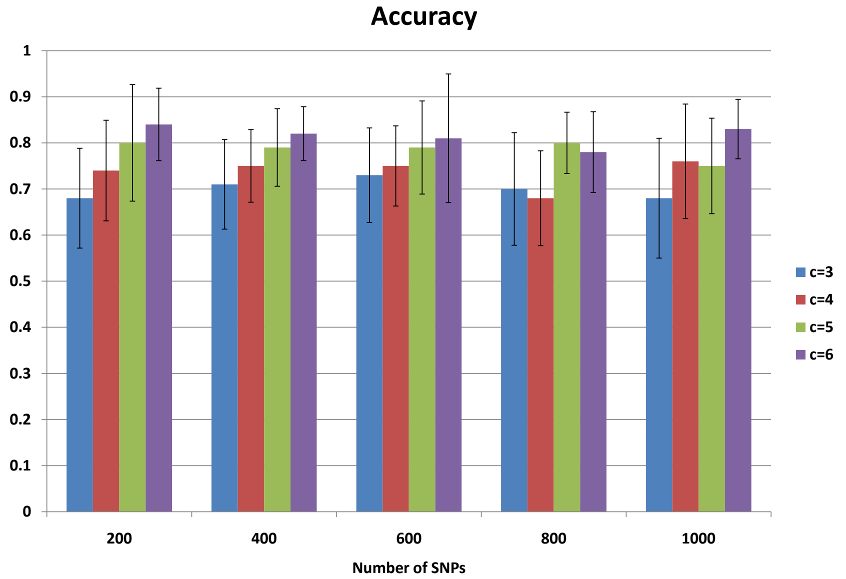


Figure 6.20: Accuracy comparison: Time in minutes versus number of SNPs for different numbers of random columns (c)

Table 6.8: Precision comparison versus number of SNPs. Four different number of random columns c (3,4,5,6) have been chosen to run the RDsnp(c) method

Number of SNPs	200	400	600	800	1000
RDsnp(3)	0.74	0.78	0.79	0.77	0.75
RDsnp(4)	0.79	0.80	0.80	0.79	0.80
RDsnp(5)	0.80	0.85	0.80	0.84	0.81
RDsnp(6)	0.86	0.89	0.86	0.87	0.87

6.6 Computational Time Analysis

Together with the analysis of the performance, a research on the computational time of the results obtained is carried out. In this way the efficiency of the two techniques can be compared.

The degree of complexity is calculated through the measure of the time spent by both the techniques to perform the analysis. In order to compare the performance, the two methods need to be tested considering the same input and the same output. The input therefore will be a matrix with a number of patients set to 1000 and a number of n SNPs variable to detect a trend of the function. Five different experiments have been performed, respectively with n = 200, 400, 600, 800, and 1000 SNPs, each one with a distribution of redundancy set in the same proportions as before. The number of random columns c is also variable and 4 different trials have been carried out. For every trial different numbers of random columns are fixed, setting once again c = 3, 4, 5 and 6. The clustering technique used is the one based on the EM algorithm and the output of the system is the list of SNPs to be either eliminated or kept. The results of the complexity analysis are shown in Table 6.9 and they are an average over 20 trials.

In Figure 6.21 the results from the original method together with the 4 results obtained from the new technique applied with four different number of random columns are displayed.

This picture shows how the RDsnp function follows a linear trend with time, compared with the polynomial trend for the original LD function. This is also confirmed by anal-

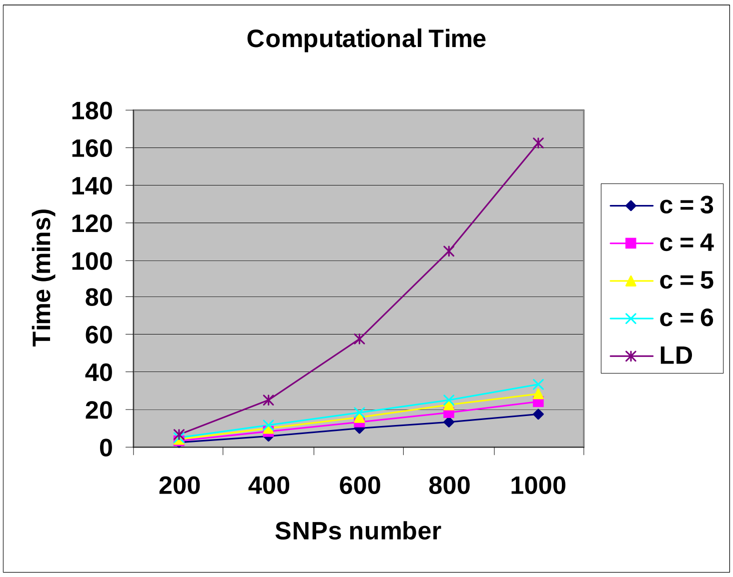


Figure 6.21: RDsnp Computational Time comparison between different numbers of random columns (c) together with the LD function trend Time in minutes versus number of SNPs

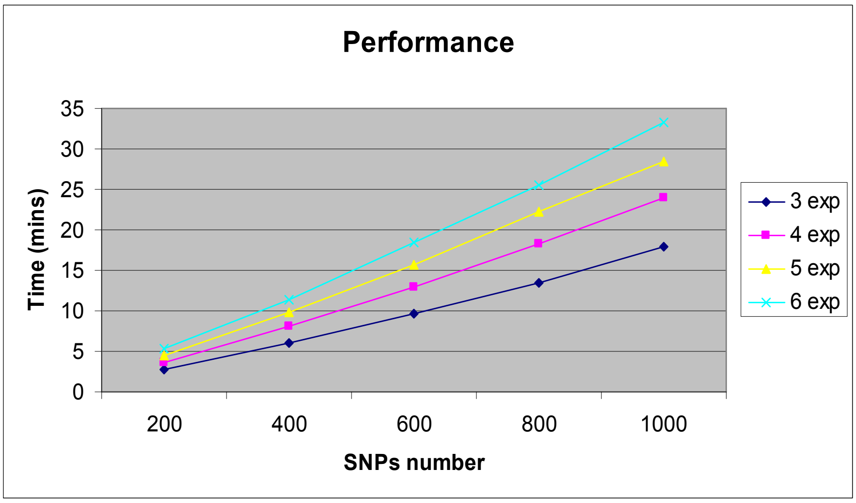


Figure 6.22: RDsnp Computational Time comparison between different numbers of internal cycles (c): Time in minutes versus number of SNPs

Table 6.9: Computational time comparison: Time in minutes versus number of SNPs. Four different number of random columns c (3,4,5,6) have been chosen to run the RDsnp(c) method

Number of SNPs	200	400	600	800	1000
LD Function	6.33	24.94	58.06	104.47	162.11
RDsnp(1)	2.78	6.02	9.74	13.45	18.00
RDsnp(2)	3.60	8.02	12.98	18.32	24.04
RDsnp(3)	4.43	9.77	15.71	22.26	28.47
RDsnp(4)	5.26	11.35	18.46	25.53	33.25

ysis of the theoretical computational complexity of the new method, compared with the original method. The LD complexity for a dataset composed by n SNPs is O(n^2) (6.3.1). The RDsnp is the combination of the application of the LD function between n SNPs and c random columns (O(cn)) together with the clustering technique applied to n vectors of c components. In the case of the application of K-means clustering, the complexity results are O(n) [178], whereas for EM algorithm the complexity is O(cn) [179–181]. In conclusion, the RDsnp presents less computational complexity than the original LD (O(cn) against O(n^2)) when $c < n$.

In Figure 6.22 only the results from the new technique are highlighted in order to show the monotonic rise of the performance with the number of random columns.

Considering a number of SNPs n = 1000 and a number of random columns c = 3, the RDsnp technique takes 28 minutes to obtain the results whereas the original method requires 162 minutes. Making an extrapolation of the function which gives the time vs. the number of SNPs an estimation of the time necessary for very large datasets can be provided. Considering for instance a dataset of 5,000,000 SNPs it will take 103.99 days to perform the original analysis compared with the 7.56 days necessary for the new method, using 5 random columns.

6.7 Random Sampling Discussion

In this specific technique, random sampling has been chosen to select two representative components from the clusters created for the SNPs redundancy detection (see point 3 in Section 6.3.1). For consistency, the same random sampling is applied also to select the final SNPs, representative of each cluster containing redundant SNPs, to be kept in the dataset.

There are different possible sampling solutions that can be found in the literature and an overview is given in Chapter 2. Random selection has the advantage that it is free from bias due to its random nature but on the other hand it may present difficulties to obtain two results that are representative for the whole population. One way to overcome this problem would be the application of stratified sampling which yields more accurate results than simple random sampling. However, one of the major drawbacks of this approach is an increase of the cost and complexity of the sample selection and the bias related to the type of the specific stratification chosen [114]. Alternatively, cluster sampling could be employed, breaking down the population into many different clusters, selecting the number of clusters that can be representative of the whole and then for each cluster selecting a random sample. All these alternative solutions provide a more accurate and possibly more reliable solution with the aim of selecting more representative components of the groups. The reason why the random sampling has been chosen from among the possible solutions is due to the aim of reducing the computational complexity and consequently minimising the time spent to perform the analysis. The strength of the new technique proposed here is the improved speed of analysis that can be provided when the dataset under analysis reaches enormous size. For this reason, the optimisation obtained through reduction of the computational complexity has been the major goal. In this case applying a cluster based sampling, for instance, would have increased the computational complexity and thus the time of processing. Nevertheless, as future work, it would be of relevant interest to apply different types of sampling techniques and check if these solutions could provide better performance while still preserving a competitive run time as compared with the original LD function.

6.8 Summary

This Chapter describes a new technique, RDsnp, for detecting groups of redundant SNPs within medical datasets in order to eliminate this redundancy. The analysis is based on the usage of the LD function, for linkage disequilibrium analysis, implemented in the R language. In the first instance, three different techniques are proposed for creating artificial datasets with a given redundancy, testing also for their accuracy. Subsequently, two of these techniques have been used to create new groups of datasets to test the proposed method.

Different sets of experiments have been performed in order to check how the new method performs under different conditions. As there are various internal parameters that can be set, such as the number of SNPs, patients and random columns, different databases have been created and analysed in order to give a wide overview of the proposed methodology. The best results obtained provided motivation for choosing the EM algorithm as the best clustering technique for this application. Considering an initial dataset of 1000 patients and 60 SNPs, the results of these experiments showed that Hierarchical Clustering provided good resulta with two false positive SNPs detected on 40 SNPs and one false negative SNP detected on 20 SNPs. Increasing the number of SNPs to 6000 in the dataset provided better results if the clustering technique applied was K-Means, which provided a false positive rate of 1.7% over 4000 SNPs. However, the false negative rate obtained was then 58.8% over 2000 SNPs. In order to overcome this problem, other clustering techniques have been applied and the EM algorithm provided the best results, being able to distinguish the 500 SNPs with a redundancy of 0.8 and 0.9 between each other from the rest of the SNPs, in a dataset composed of 1000 SNPs. For this reason, the EM algorithm was chosen as the best clustering technique for this application.

A comparison analysis of the RDsnp method against the original LD function has been carried out in order to investigate whether there is any substantial improvement in performance. These trials provided very positive results in respect of reducing the computational complexity of the current method, without losing significant information. Performing a real dataset analysis, the experiments showed that for more than 60 SNPs, the RDsnp becomes much faster than the LD function, still preserving a good accuracy while also increasing with the number of SNPs analysed. Regarding the analysis of the

RDsnp performance with artificial datasets, the results showed that for large datasets the performance of the new method is much better that the original. Setting the number of random columns $c = 4$, on a population of 1000 patients and 1000 SNPs, 760 SNPs (Table 6.7) are correctly detected as redundant or not redundant in 33 minutes (Table 6.9), against 162 minutes of the current method.

In terms of accuracy, a very good result of 0.75 is achieved for this dataset. This means that over a dataset composed of 100 SNPs all redundant between each other, 75 SNPs are correctly detected as redundant SNPs so that they may be deleted from the dataset for further analysis. In more general terms, three quarter of the total number of SNPs in a dataset is correctly assessed as redundant or not redundant.

In conclusion, this study provides researchers with a new technique to detect the redundancy in large datasets of SNPs data which is more efficient than the commonly used one thanks to an overall reduction in computational complexity. This study therefore proves to be an important achievement in the increasingly popular problem the reduction in size of large SNPs datasets.

Chapter 7

New Guideline for Large SNPs Database Analysis

7.1 Introduction

The improvement in genotyping technologies allowed scientists to collect large amount of data in relatively short time, still preserving an high accuracy at a cheap prize. The continuously growing of databases size nowadays available for genetic studies has in turn brought the researchers attention into the development of new methods for data mining. Several different analysis have been performed in the past in order to provide suitable tools for extracting the relevant information hidden inside this huge amount of data.

One of the main issues that needs to be faced nowadays by data miners is the reduction of the time spent for performing analysis with such large databases. Together with this, also the computational complexity has grown exponentially with the amount of data gathered. These relevant difficulties have brought limitations and problems to the tools currently used in this field. Lots of new different approaches have been so far proposed in order to overcome these obstacles and in this work an overview of the problem has been presented.

In this final Chapter a global idea of the work that has been carried out in the past three years of research is shown. Different solutions focused on resolving the problem of dataset size reduction have been analyzed and improved. Thus, the ultimate aim is to provide researchers with a guideline for future research work in genetic data analysis.

7.2 New Methodology

The methodology that is proposed in this work is composed by two main steps. In the first instance an extensive analysis of redundancy possibly present in the dataset is applied in order to eliminate the superfluous information which affects the performance of the subsequent analysis slowing down any ulterior process. In this way, only a small amount of SNPs can be selected to be considered for association studies. In second instance the proper SNPs analysis is carried out to detect the minimum amount of SNPs which show association with a specific disease. This second step, in turn, is divided in two different main paths as depending on the initial available data, different processes need to be followed.

In case the original dataset is composed by family members, as it is supposed to be analyzed with a family-based method, the 'Improved TRANSMIT Method', which is illustrated in Chapter 3, is applied. On the contrary, if the dataset is composed by completely unrelated individuals, as required for Case-Control studies, the 'Combined Decision Trees Analysis', discussed in Chapter 4, is carried out. The whole process is shown in Figure 7.1.

7.2.1 Step 1 - Redundancy Elimination

Various kind of data sources can be available for association studies, in this work two different datasets types are covered : family-based and case-control datasets.

In the first one, the population is composed by trios, characterized by a mother, a father and at least one of their children. These kind of data are generally used to study how genetic information is passed through generations from parents to offspring. In this case the elimination of redundancy should be carried out in separate populations extracted from the original one. Within each of these subsets, individuals must be independent between each other. This can be achieved by building a population of unrelated parents for instance. These subpopulation can thus be processed with the 'RDsnp' method for redundancy elimination.

The second type of dataset considered in this study is the one for case-control analysis that is composed by individuals which do not have a parental link with each other to avoid

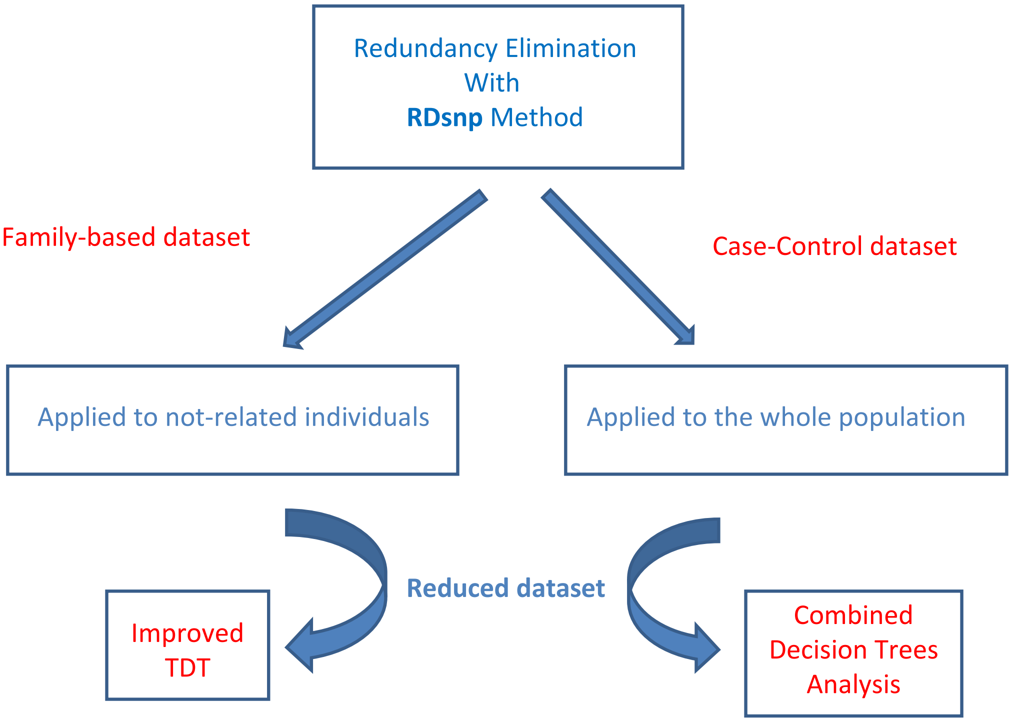


Figure 7.1: Flow chart of the RDsnp method for SNPs analysis.

a biased result. If this condition is fulfilled, the dataset can be directly processed with the 'RDsnp' technique, that, as extensively shown in Chapter 6, is based on the popular LD function, used to calculate the linkage disequilibrium between genetic markers. This function, providing the measure of the squared correlation coefficient R^2, detects how much two different SNPs are linked together. In more simple terms, the probability for SNP to be present, given the presence of a second SNP, is calculated and highlighted by LD function. Any SNPs which then present an $R^2 \geq 0.8$ is defined redundant and therefore eligible for subsequent elimination. In order to avoid an excessive computational complexity of the problem, the new proposed RDsnp method can be applied.

The basic flow chart in Figure 7.2 shows the main steps included in this procedure. Instead of applying the LD function directly to the initial dataset, a few random columns are created with a realistic biological distribution of the allele. The LD function is then

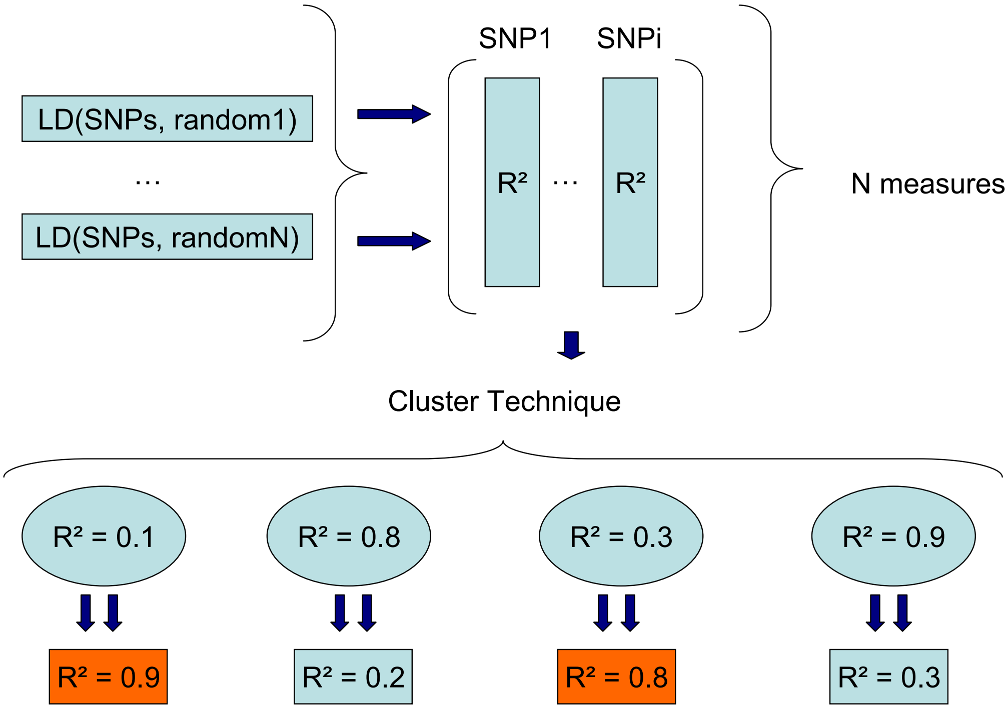


Figure 7.2: Flow chart of the RDsnp method for SNPs analysis.

applied separately to the dataset and one of these created columns, in turn. The R^2 results obtained from each random column, are joined and displayed in a space whose dimensions are equal to the number of random columns. Clustering this population of vectors containing the R^2 for each SNP, provides a selection of SNPs showing a $R^2 \geq 0.8$ and therefore defined as redundant. These SNPs can be removed from the dataset expediting further analysis. For a detailed analysis of this method the reader is referred to the Chapter 6.

Different parameters can be set in this method and the choice is dependent on the specifications of the study that needs to be performed, such as amount of SNPs and patients. Further parameters setting includes the amount of random columns c and the choice of a proper clustering technique. As shown in Chapter 6, for instance, if the dataset is composed by around 100 SNPs, using $c = 3$ random columns and the EM Clustering technique, 75 SNP are correctly detected as redundant or not redundant in only 28 minutes. Compared with the 168 minutes taken from the original LD function, this new method results very promising as first step for data mining tasks.

7.2.2 Step 2 (a) - Improved TRANSMIT Analysis

Assuming that genetic information from trios are available to the data miner, a family based study can be performed. In this case a Transmission Disequilibrium Test can be considered. As described in Chapter 5 the samples used consist of a set of trios, two parents and their affected offspring. Collecting this kind of data is generally more expensive and more time consuming. Nevertheless, whenever this data is available it is advisable that TDT is used for association studies as preliminary analysis. In order to apply a TDT, TRANSMIT software is employed in the study.

An important distinction is made in the following two paragraphs in order to separate the case between Boolean and continuous outcome of the study. In general terms, if a disease is studied, the outcome is usually binary variable as it can have a value for instance equal to 1, representing the disease and a value such as 0, representing the healthy status. If a phenotype is to be studied, then the outcome variable can be continuous or categorical. In this case thus, a separate analysis needs to be carried out.

Boolean Outcome

The most common case is when a set of SNPs needs to be detected in association with a specific disease. This means that the outcome of the analysis is a Boolean variable, i.e. disease or not disease. In this case, applying TRANSMIT provides information upon the transmission of SNPs from affected parents to their children. With the original version of TRANSMIT a list of haplotypes transmitted from affected parents to offspring is obtained. The list of SNPs for each haplotypes is fixed and in the specific dataset hereby analysed it includes 8 different SNPs.

The new idea proposed in this methodology is to extract a subset of data from the original database and apply TRANSMIT to these datasets. The subsets of data are created by selecting every possible sub-combination of the original set of SNPs. In the specific case of the given dataset, the procedure is applied to select 7 SNPs taken from the original database, then 6 SNPs, 5 SNPs and so on. Checking the value of the χ^2 for each haplotype of different length, it is possible to detect the minimum set of SNPs that show a strict association with the disease. When the analysis is focused on complex conditions, often, more than one gene is involved in the contribution to the predisposition for the disease. For each gene there is then a considerable amount of SNPs that is usually detected as possibly responsible of the disease, in the first stage through biological and molecular processes.

The new approach that is described in this section, provides clinicians with a useful technique to select a small amount of SNPs, significant for the specific study, eliminating the majority of genetic information that doesn't show any association with the disease under analysis. In Figure 7.3 and 7.4 the flow chart of this method is shown.

Continuous Outcome

Supposing that a symptom or phenotype of the disease represents the variable under analysis, this quantity might not Boolean but continuous or categorical. In this case, TRANSMIT software cannot be applied directly without a pre-processing of the dataset.

The idea is analogue to the one used for the case-control analysis. A fixed number of thresholds is set for the outcome in order to be converted in a Boolean variable. Referring to the example shown in Chapter 4 for case-control analysis, the CBC as measure of disease severity degree can be used in the case of Pre-eclampsia. Fixing 10 different values of CBC and building 10 different datasets accordingly, provides the set of inputs needed for the analysis. Each created dataset is processed with the method used for the Boolean outcome explained in the previous paragraph. The result showing the greatest χ^2 is the one with the highest degree of significance and therefore can be chosen as the final outcome. The threshold set for this result is the best threshold to be chosen for association study for that specific disease.

7.2.3 Step 2 (b) - Combined Decision Trees Analysis

Alternatively to the TDT, one of the most commonly used technique for association studies is the case-control analysis. Plenty of different tools have been proposed by the research community in order to perform this kind of approach and the Decision Trees al-

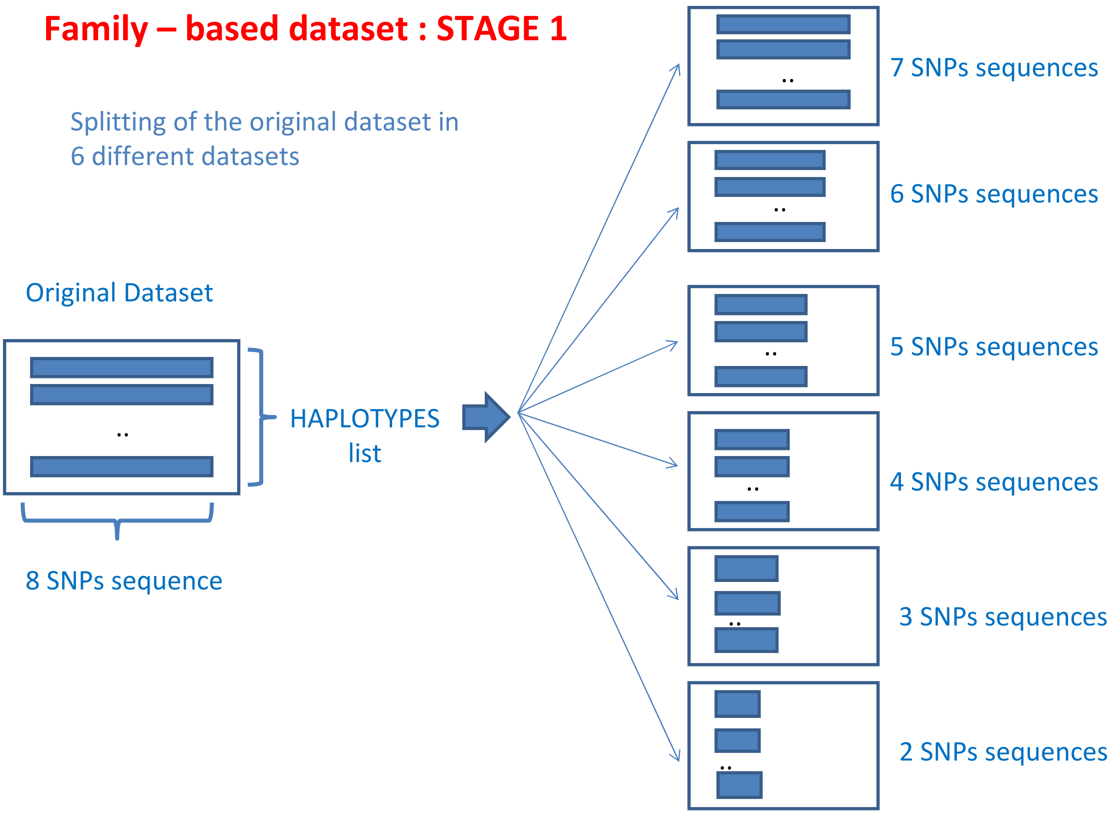


Figure 7.3: Flow chart of the new proposed method for SNPs analysis in family based datasets with the TDT technique. In the first step several datasets are built from the original one, wiht different amount of SNPs.

gorithms have resulted one of the most popular and often successful technique. Similarly to the TDT analysis the outcome of the study can be either Boolean or continuous type. Both of these cases are treated in the following two paragraphs in order to give a complete overview of the general problem.

Boolean Outcome

In the most simple case scenario of Boolean variable, the combined analysis of Decision Trees can be carried out processing the original dataset with the three different software that have been presented in this work, namely ADTree, C4.5 and ID3.

For every result that shows an acceptable significance with a K value greater than 0.2,

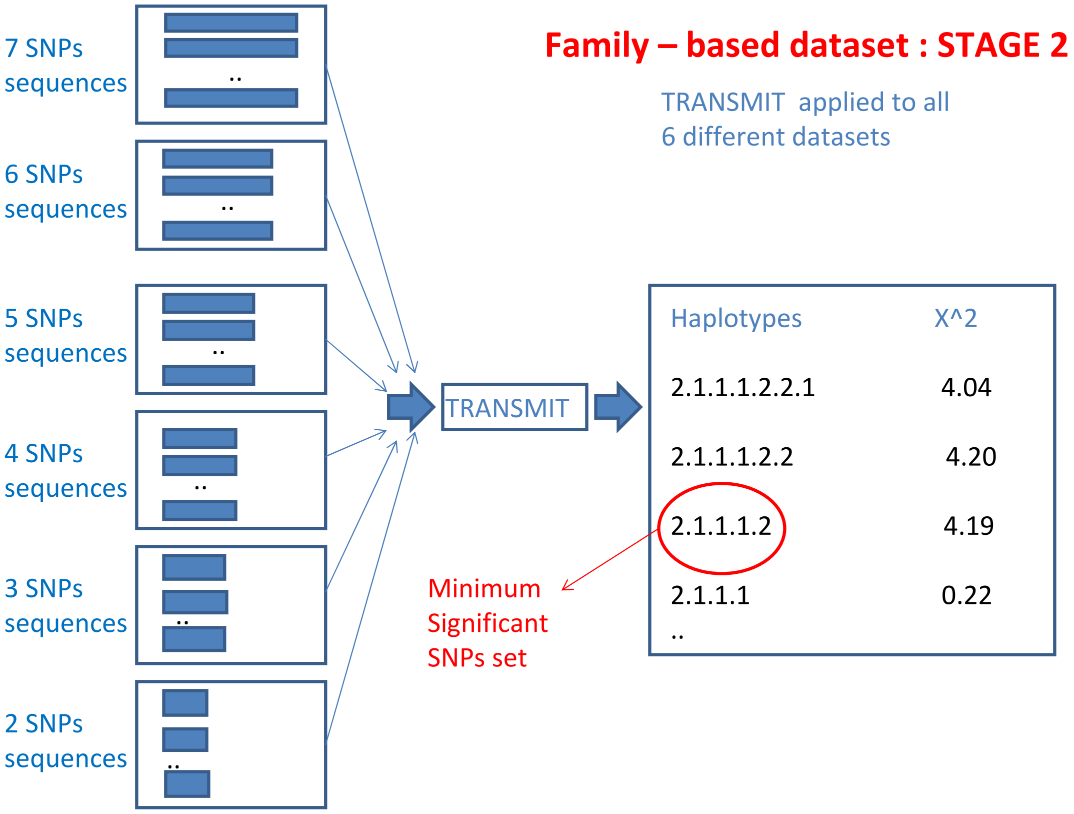


Figure 7.4: Flow chart of the new proposed method for SNPs analysis in family based datasets with the TDT technique. in the second step TRANSMIT is applied to select a smaller amount of SNPs associated with the disease.

the list of SNPs obtained from each of the three solutions can be compared and contrasted between each other. Selecting the largest amount of SNPs that is common to the three solutions from each algorithm, decreases the size of the original database, focusing the study upon the SNPs associated with the disease. A narrowed analysis can be performed then applying the algorithms to the sub set of SNPs detected in the first stage.

Continuous Outcome

A less common scenario, but still very useful especially in case of lack of information, is characterized by a continuous outcome of the analysis. This is the case for instance when the disease does not have a perfectly defined profile, able to distinguish between

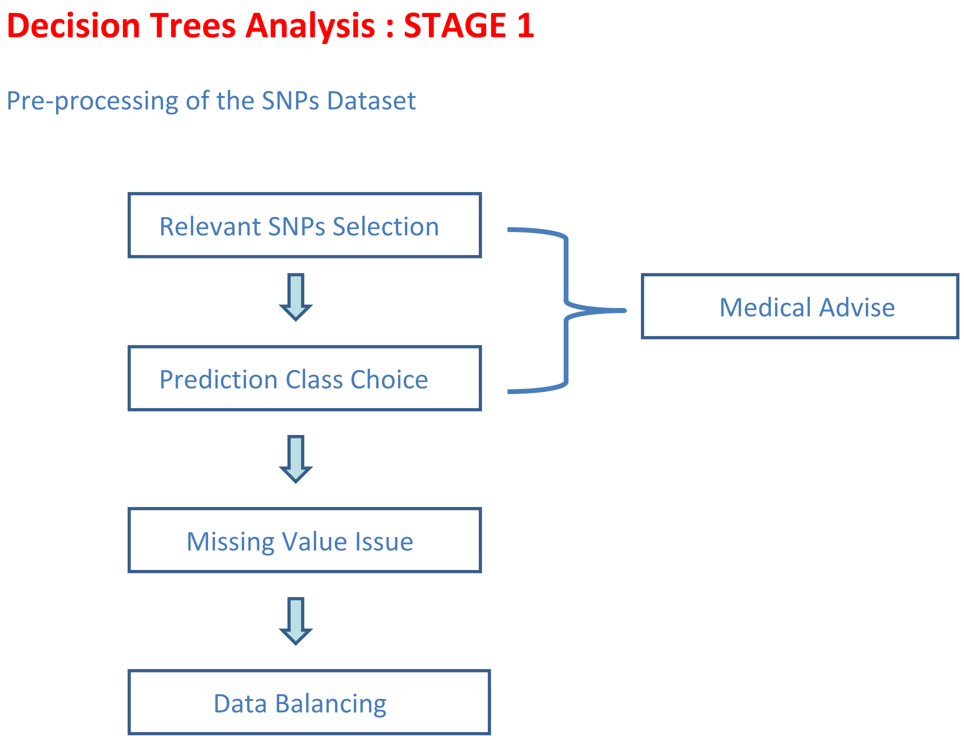


Figure 7.5: Flow chart of the new combined Decision Tree method for SNPs analysis .

the two evident conditions of confirmed disease and healthy status. There are for instance diseases characterized by different stages of severity such as cancers, diabetes, Alzheimer or other progressive disorders. In this case the outcome is represented by a categorical parameter.

Alternatively, the disease can be clearly described by a binary variable but the study can be focused only on a specific symptoms or phenotype involved. In all these cases, a more articulate analysis needs to be carried out. The process is divided in two main parts, the pre-processing of the initial dataset and the proper analysis. The pre-processing stage includes all the most relevant steps that are usually faced in data mining tasks as shown in Figure 7.5:

1. Selection of elective SNPs. This operation is often driven by medical staff, geneticists or biologists as complex phenomena regulate the cause-effect system of

genotype-phenotype link. Within the whole human genome, only a subset of SNPs is identified by professionals as supposed to be involved in the studied diseases. This together with logistic and financial factors contributes to the final choice of SNPs for the initial dataset.

2. Prediction Class Choice. Similarly, for this step the medical advice is needed to select the most relevant variable that can be expected to have a close connection with the severity of the disease.

3. Missing value issue. This problem needs to be considered whenever there is a lack of data. If the algorithm does not account of this, the patients with missing values need to be removed from the dataset. Alternatively, if the algorithm can deal with missing values it is advised to perform the analysis with and without the patients with missing values in order to check for the impact that this lack can introduce.

4. Data Balancing. Any time a case-control analysis is carried out it is important to check for the case control ratio in order to limit the bias that an excess of one of the two class can bring to the analysis. The ideal scenario is 50% cases and 50% controls. If this is not the case, two subsets of population from cases and controls can be extracted with the same size.

The proper analysis is shown in Figure 7.6. Alike previously shown, a number of thresholds are set for the continuous outcome in order to convert it in a Boolean variable. The same number of datasets are accordingly created each one with a different threshold of the outcome. The three algorithms presented in this work, without any specific priority, are applied to the dataset subject to the previous processing. The Kappa value for each trial is calculated and checked.

The result with the highest significance is chosen for further analysis. The threshold related to the chosen outcome is therefore the best choice for obtaining a relevant result. Once that the threshold is fixed, the case-control ratio needs to be checked as part of pre-processing analysis and the dataset can require to be balanced. Listing the SNPs obtained with the best solution allows to perform narrowed studies on the subset of data selected. Moreover, carrying out cross analysis between these three tools upon the small

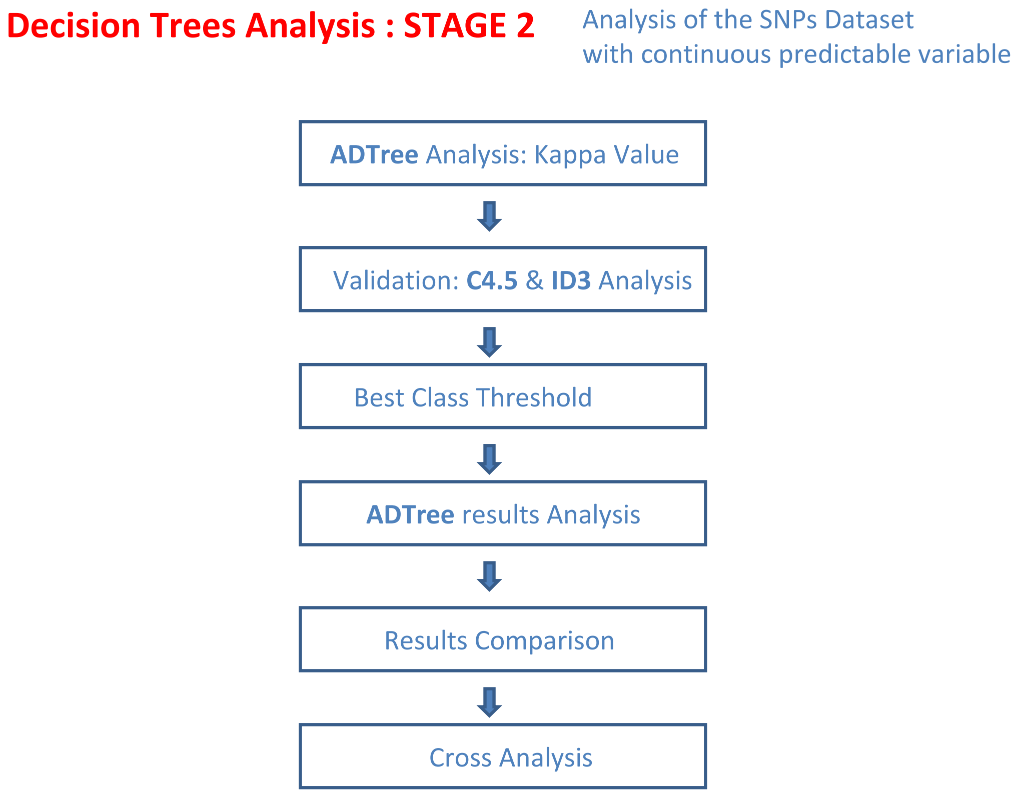

Decision Trees Analysis : STAGE 2
Analysis of the SNPs Dataset
with continuous predictable variable
ADTree Analysis: Kappa Value
Validation: C4.5 & ID3 Analysis
Best Class Threshold
ADTree results Analysis
Results Comparison
Cross Analysis

Figure 7.6: Flow chart of the new combined Decision Tree method for SNPs analysis .

final dataset, a better significant result can be obtained and validated within the three algorithms.

A final important remarks needs to be made about the flexibility of this technique. Conversely to the TDT, this kind of disease association study can be extended also to an analysis of life style factors together with any possible initial health status of the patients under study. This would provide a wider scenario for the disease aetiology analysis as it is remarkably proved that most diseases are due to a combined effects of genetic and environmental factors.

7.3 Summary

A complete overview of the new methodology proposed in this work is explained in this final Chapter. In the current literature, many different solutions can be found in the field of genetic data analysis for the specific task of dataset size reduction. Two of the main types of approach are discussed in this work, the family based analysis and the case control study. The choice between these two options is restricted by the initial dataset type. For family based analysis, the TDT is the proposed method and for case-control study, the decision tree algorithms are the alternative path to follow. Both of these basic techniques are suggested in this work in a new and improved overarching framework.

The transmission disequilibrium test is performed through the application of the TRANSMIT software. An optimisation of the performance of this software is proposed as a novel approach for family based analysis. Instead of using the fixed length of haplotypes (composed for instance by 8 SNPs), every possible combination of SNPs, taken in groups of different size from 7 to 2, is calculated. The TRANSMIT is then applied to all these subsets of data in order to select the best sequence of SNPs that shows association with the disease.

The case control analysis is performed for the dataset containing independent individuals, applying a combined analysis with three different algorithms commonly used for significant SNPs detection in association studies. The comparison of these three techniques in a sequential cross-analysis method allows researchers to choose the best option for every different application.

For both of these novel approaches a consideration on the outcome type is also included. Generally the disease is the discriminative variable representing the Boolean class. However, in some cases, different continuous variables such as phenotypes can be taken in consideration. In case of non Boolean class, a new approach is proposed in this methodology, for both family based and case control analysis. The best threshold to be chosen for converting the class from continuous to Boolean variable is found through a detailed analysis of the statistical significance of results. The optimal result is the one that selects the subset of SNPs with the highest statistical significance.

This articulated analysis is completed by a pre-processing of the dataset for the elimination of any possible redundancy present in the original dataset between SNPs. This method is based on the commonly used LD function for measuring linkage disequilibrium between genetic markers. The optimisation of the this tool hereby proposed provides a significant improvement of the current techniques used for redundancy detection in large datasets, thanks to the reduction in computational complexity.

Chapter 8

Conclusions

One of the most challenging goals for the Bioinformatics community is the resolution of problems related to the rapdily increasing size of datasets. The reduction in size of data source(s) is particularly needed to permit the use of genetic data in association studies. This is due to the continuously growing size of genetic data that is currently available in a cheaper and faster manner. This aspect represents the first problem tackled in this Thesis. The most commonly used approach is based on the application of the linkage disequilibrium (LD) function, implemented in different software languages, for measuring the square correlation coefficient between pair-wise SNPs taken from the dataset. The aim is to detect SNPs which are highly correlated in order to remove the genetic markers that do not bring relevant information for further analysis. Selecting the target SNPs is enabled by predicting the remainder of the SNPs population which is *not* required for the study. Although the current approach provides results with a high accuracy, it also reveals considerable limitations in terms of computational complexity when the analyzed database reaches enormous size such as one million SNPs per set. The selection of the eligible SNPs is an important stage of the analysis, as the massive number of genetic markers that represent the attributes included in the input matrix inevitably affects the performance of any subsequent analysis methods applied. In particular, this can affect the time and the computational complexity of the process, restricting the choice of the tools and machines eligible for this kind of analysis. In order to overcome this limitation, a new fast scalable tool for approximation of the LD function has been developed, namely RDsnp, and its performance has been tested through several experiments. The results of this analysis have shown how the RDsnp provides a substantial improvement in terms of run time, due to a significant reduction of the computational complexity compared to the original version. The experiments have shown also that the accuracy of the results has been maintained at a reasonable level. However, a possible further improvement of the accuracy may possibly be obtained in future tuning different parameters involved in the new technique such as number of random columns created.

The second important issue that genetic data miners need to face is disease association studies. This is based on the detection of a link between DNA components and disease risk. From one perspective, if the dataset under analysis has a family based structure, one of the most successful techniques is based on the Transmission Disequilibrium Test which detects the genetic markers responsible for an increased disease risk from the analysis of the information transmitted through generations. In this Thesis, attention is focused on the TRANSMIT software, as among the available tools it presents some important strengths such as dealing with transmission of multi-locus haplotypes, even if both phase and parental genotypes may be unknown. Nevertheless, the original version of this tool can only select a fixed amount of SNPs resulted in association with the disorder. Hence, this provides a subset of genetic markers which is not necessarily minimal in size. In order to solve this problem, an optimisation of the TRANSMIT software based on multiple-test analysis is proposed, providing a previous elimination of redundant SNPs. The results that emerge from the application of this new technique to a medical dataset illustrates the type of added information that this tool provides as compared with the original version of the software. A smaller number of SNPs that are statistically significant are selected in the disease association analysis. By checking the observed occurrences versus those expected, for the SNPs present in the final haplotypes, it is possible to identify any validation of the positive or negative effect of a SNP in the risk of disease. This technique therefore provides a clinician with a wider selection of information for selecting a smaller number of relevant genetic markers on which to perform further analysis.

Case control analysis performed by the decision tree algorithms is a widely established approach for datasets composed of unrelated individuals. However, every individual decision tree algorithm implemented in the literature presents limitations for different aspects of the problem. Alternatively a combination of different tools in a cross-analysis of the data, validating or rejecting the results obtained from one algorithm, can provide a robust solution to overcome the weakness of a single analysis method. Under this scenario, a new framework for selection of relevant SNPs in disease association studies is proposed. Experiments have shown that with this tool it is possible to identify a smaller number of SNPs that may be relevant in the analysis, using a combination of ADTree and C4.5 algorithms. This refinement, besides the validation provided but two algorithms, could not be achieved with the application of a single algorithm. Several experiments have been performed in order to show how different results can be obtained applying different variable settings. Selecting different attributes, changing the class choice, keeping or eliminating missing values or having relevant feedback from the medical side can help improve the final outcome of the analysis, consequently increasing the statistical significance of the results. This study has also validated the important medical meaning of the CBC threshold of 10 as relevant cut-off for distinguish small babies born from pre-eclamptic pregnancies.

It is important to remark that, from scalability limitations, the two proposed approaches for selecting a smaller amount of relevant SNPs in the disease association study (TDT and case-control analysis) assume a pre-processing stage where the redundant SNPs are eliminated in order to reduce the dataset size, when necessary. For this reason, in conclusion, a comprehensive framework is proposed for genetic data analysis. This includes the redundancy elimination process, followed by relevant SNPs selection stage in disease association studies for both family and population based datasets.

8.1 Main contributions to Bioinformatics

This Thesis illustrates a proposal of a new guideline aimed to provide a significant tool for supporting new scientific analysis for medical and genetic applications, in light of the current pressing needs of the genetic research community following the disclosure of the entire human genetic heritage. One of the most relevant goals of this work is appropriate dataset size reduction without losing information relevant to the scope of the analysis. The main contributions to the field of Bioinformatics, achieved in the past three years of research are stated as follows:

(1) In terms of redundancy elimination, a new effective tool has been created and dis-

cussed in this Thesis namely the RDsnp method. This is based on the linkage disequilibrium (LD) function for measuring the degree of bond between SNPs. The original version of this tool is designed to measure the linkage disequilibrium between every possible pair of SNPs taken from the original input matrix. The new improved version of the LD function avoids the requirement to calculate such an extensive amount of values. This is appealing to a confrontation of the LD value for each SNPs with one or more random SNPs created. Through a more comprehensive approach to the problem, RDsnp provides a drastic reduction of the computational complexity, a common problem frequently encountered while attempting to fulfil this task. In a pre-processing stage, which precedes the targeted investigation, this technique represents a useful and efficient method for SNPs data size cuts.

(2) Within the family base studies the Transmission Disequilibrium Test (TDT) is employed for detecting the SNPs set associated to the disease. One of the most common solutions for the implementation of the TDT is TRANSMIT software. The original version of TRANSMIT is based on the analysis of a fixed length sequence of SNPs. A pre-defined amount of different genetic markers are analyzed through the χ^2 statistical test in their transmission from parents to offspring. This approach, which is based on the Mendelian law of inheritance, selects the set haplotypes inherited by affected offspring whose observed occurrence overreaches the expected one. This excess of SNPs present in children with a given condition is a proof of the genetic disease risk association. In the novel method proposed every possible combination of SNPs subsets, taken from the original fixed-components sequence are created. The TRANSMIT software is then applied to the all subsets of data generated. Considering a sequence of SNPs resulted in high association with a given disease, whenever an irrelevant SNP is removed, the significance degree of the reduced set is not expected to drop. This rationale provides a rule for selecting the genetic markers that present an evident association to the disease, allowing a dataset size reduction through the deletion of irrelevant SNPs.

(3) For case-control analysis, decision tree algorithms are one of the most commonly applicable tools for evaluation of risk disease association. The new method illus-

trated proposes the combined exploitation of three different decision tree algorithms as a more reliable tool for analysis. Lists of effective SNPs are compared between the three solutions in order to reject or validate the obtained results. Additionally an extensive method of analysis is suggested in the case of a continuous variable as the predictive class. The best choice of the threshold for the Boolean conversion of the outcome variable is achieved through a detailed statistical study of the results significance.

(4) A final combination of the three previous points provides the new comprehensive framework for further genetic research. This allows analysis to be performed of different dataset structures such as population of related and unrelated individuals, or different types of outcome studied. This includes such variables as Boolean, categorical or continuous classes. Additionally, an extensive analysis focused on redundancy elimination depends on the pre-processing stage of the entire framework. This work presents a novel methodology, adaptable for different needs and data constraints, for the accomplishment of the increasingly demanding tasks that the genetic community must face in the complex research area of disease aetiology.

8.2 Potential medical implications

In the analysis of complex diseases, clinicians are often required to deal with a huge amount of genetic information often derived from different genes. The possibility to remove the surplus information from the constantly growing genetic datasets and select the target SNPs that present an evident association to the disease under study provides a new hope for the medical application of genetic disorder analysis.

Through the optimization of the TRANSMIT software application, in family based studies the doctor is provided with an exploratory test able to exclude the sequence of SNPs that do not show any association with the specified disease. It is possible to select the SNPs that always appear in different haplotypes and to detect whether the same allele is always present for a given SNP. Such detailed analysis provides a compendium of helpful tools for clinicians to outline the cause of diseases.

Similarly, for case-control studies, the new combined analysis based on decision trees, provides the doctor with a list of elective SNPs which play a role on either the disease risk or the disclosure of a single aspect of the condition, such as phenotype.

Furthermore, the introduction of the RDsnp method for redundancy elimination allows data miners to keep their attention focused on the manageable amount of SNPs that can be considered as independent attributes of the input matrix, represented by the medical dataset. The remainder of the SNPs population can easily be predicted with different statistical tools. This is a significant achievement for the dataset screening stage of the analysis.

If the scientists are able to detect the part of genome responsible for a given disease, lots of research could be performed in a more accurate way, with the aim to develop both a targeted therapy and a more precise diagnosis. This would allow the prevention of certain diseases and to delay the progression of the disease at its onset. Individuals which are positively tested for high predisposition to certain conditions can find a support in following a treatment or life style choice in order to prevent or reduce this threat.

Additionally, breakthroughs are continuously made by the genetic community in order to establish a new challenging discipline namely gene therapy. This is realized through the insertion of genes into an individual's cell and biological tissues to treat disease, such as cancer where destructive mutant alleles are replaced with functional ones.

Many diseases are already proved to be associated with genetic information coming from one or more genes such as pre-eclampisa, sickle cell anaemia, cystic fibrosis, Aicardi Syndrome, Huntington's disease, Alzheimer, diabetes, obesity, arthritis and various cancers. Despite this, many of these diseases require further extensive research in order to validate previously discovered findings. Moreover, many genetic components have still an unknown and undiscovered function and meaning.

Many years will be spent in this amazingly attractive research area of biological function understanding that is continuously bringing a considerable support to human life in priority areas such as healthcare and wellbeing applications.

8.3 Suggestions for further research

Considerable effort has been spent to provide a relevant contribution to the Bioinformatics community through the new optimizations proposed in this work, applied to different aspects of the important problem of DNA analysis. However, there are still plenty of unknown processes and functions in the genetic field that require extensive further investigation.

Regarding the redundancy study, although the proposed technique provides a significant advance on what is currently available in the field of SNPs datasets size reduction, there are still numerous aspects of this analysis which can be extended and potentially improved. For example, it would be interesting to perform more experiments with even larger datasets both in width and length to check how this method performs. Even if 1000 patients is a reasonable amount of instances for a common dataset, in the future there will be an increased availability of databases containing several thousand even millions of patients. The amount of SNP information is also rapidly increasing and in the near future there will be hundreds of millions of available SNPs to be analyzed. Moreover, the number of used random columns affects the accuracy and obviously the performance of the method. Therefore it would be interesting to assess the maximum number of random columns that gives the best accuracy while the technique retains its competitive computational performance. Within all these suggested improvement, it is potentially worthwhile to investigate different clustering techniques that can results more efficient processes with the new parameters settings. Moreover, in order to give a wider overview of the studied problem, an analysis should be performed on artificial datasets with more than one type of redundancy, each one still with different subsets of intensity.

Concerning the family based studies, the optimization of the TRANSMIT software can be applied in a progressive analysis. This candidate gene filter may be used by clinicians starting with the analysis of one single SNP, in order to detect the interesting ones. In the second step it may be possible to add the second SNP and check which ones may be significant. In this way experimental work may be reduced and improving the analysis process.

New ideas can be developed also around the case control studies through decision tree algorithms. It would be remarkable to develop a new solution of analysis which is also based on the case control procedure but makes use of the genetic data from relatives of the initial population. In the TDT techniques, the genetic information from the parents and relatives is used to be stored as new individuals in the dataset (parents, siblings, etc.) and are analyzed through a family based method. In the case control method, one of the main constraint of the population is that it is represented by unrelated individuals. Trying to include in the dataset the heredity aspect of genetic data, would require the transposition of this information from the rows of the input matrix to the columns. New attributes can be created in the dataset, storing the SNP value of the mother or grandmothers' of babies under analysis. The genetic information between generations is of course related due to inheritance rules but new findings can be revealed when the study is focused on their association with a given disease. Additionally, this concept can be extended to the fathers' and grandfathers' genetic information, as research has already proved the influence of fathers genes in PE risk [23].

Finally, it would be pertinent to explore in more detail the possibility to generate suitable artificial datasets in order to test different tools available for data miming. In the field of genetics, with reference to the SNPs, a general overview is given in Chapter 5 but still many more considerations are needed for quality artificial datasets to be created. There is more than one factor that affects the link between a disease and human DNA. Frequently the causes are due to the presence of mutations located in different genes and different SNPs can have various allele frequencies. The risk of disease changes in turn both with genetic and environmental features. All these considerations, together with the possible presence of different linkage disequilibrium values between genetic markers, in different degrees of intensity, provides the background for further development. This may lead to series of experiments designed to achieve the best realization of a synthesized dataset which can reasonably resemble and approximate actual genetic data.

8.4 Dissemination

8.4.1 Publications

The three major contributions of this work have been the foundation of content for three different publications:

- Within the analysis over the Transmission Disequilibrium Test, the novel approach based on TRANSMIT software for family based datasets in SNPs association studies has been published in the following book chapter:

  L. Fiaschi, J. M. Garibaldi and N. Krasnogor, *Multiple-Test Analysis of Sequences of SNPs for Determining Susceptibility to Pre-eclampsia*, *Computational Intelligence and Bioengineering* of the series *Frontiers in Artificial Intelligence and Applications*, IOS PRess, volume 196 in 2009, in the proceeding *Computational Intelligence and Bioengineering - Essays in Memory of Antonina Starita* edited by Francesco Masulli, Alessio Micheli and Alessandro Sperduti.

- The study carried out upon the Combined Decision Trees analysis for case control study of SNPs data has been published in the following:

  Fiaschi Linda, Garibaldi Jonathan M. and Krasnogor Natalio, *A framework for the application of decision trees to the analysis of SNPs data*, CIBCB'09: *Proceedings of the 6th Annual IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE Press, 2009, pages 106-113.

- The RDsnp method proposed for redundancy elimination in large size SNPs databases is under preparation for submission to a further journal paper:

  Fiaschi Linda, Garibaldi Jonathan M. and Krasnogor Natalio, *Redundancy Detection in Biallelic Single Nucleotide Polymorphism Datasets*.

8.4.2 Conferences, International Workshops and Seminars Presentations

The following events have contributed to the dissemination of the results obtained with this research:

- **L. Fiaschi, J. M. Garibaldi and N. Krasnogor** *SNPs Redundancy Analysis* - at the conference EURO XXIII in Bonn, Germany, 6th of July 2009.

- **L. Fiaschi**, *New methodology for SNPs analysis* - at BIOPTRAIN Workshop Florence, Italy, 10th June 2009.

- **L. Fiaschi, J. M. Garibaldi and N. Krasnogor** *A Framework for the Application of Decision Trees to the Analysis of SNPs Data* - at the conference IEEE CIBCB 2009 Nashville (TN), USA, 31st March 2009.

- **L. Fiaschi**, *Standards for SNPs Analysis with Decision Trees Tools.* - at IMA group seminar, Nottingham, UK, 24th February 2009.

- **L. Fiaschi**, *Diseases Association Studies for SNPs Data.* - at BIOPTRAIN Workshop, Innsbruck, Austria, 12th January 2009.

- **L. Fiaschi**, *Decision Tree Algorithms in Pre-eclampsia analysis.* - at ASAP group seminar, Nottingham, 15th November 2007.

- **L. Fiaschi, J. M. Garibaldi and N. Krasnogor** *SNPs Analysis in Pre-eclampsia.* - at EURO-CBBM Workshop, Prague, Czech Republic, 8th July 2007.

- **L. Fiaschi**, *Genetic Data Analysis.* - at ASAP group seminar, Nottingham 22nd November 2006.

- **L. Fiaschi**, *Alzheimer Analysis and Weka.* - at Queens Medical Center, Nottingham 4th August 2006.

- **L. Fiaschi**, *An overview on Decision Trees.* - at Queens Medical Center, Nottingham 17th March 2006.

- **L. Fiaschi**, *Learning classifier Systems for genetic data* - at BIOPTRAIN internal meeting, School of Computer Science, Nottingham 9th January 2007.

- **L. Fiaschi**, *Genetic association study: an overview* - at BIOPTRAIN internal meeting, School of Computer Science, Nottingham 2nd March 2006.

Bibliography

[1] Genome news network. visited on 05-09.

[2] Genome.gov. visited on 05-09.

[3] Llew Mason Yoav Freund. The alternating decision tree learning algorithm. *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 124–133, 1999.

[4] Dr. peter d adamo. visited on 05-09.

[5] Justin Petrone. Illumina says new humanomni chip 'important flagship' for next round of gwas. *BioArray News*, 2009.

[6] Biology online.

[7] Willian S. Klug, Michael R. Cummings, Charlotte A. Spencer, and Maichael A. Palladino. *Concepts of Genetics*. Pearson Benjamin Cummings, 2009.

[8] Genetics concepts - site www.ncbi.nlm.nih.gov. visited on 05-09.

[9] Genetic overview - www.genetics.gsk.com. . visited on 05-09.

[10] Schuler GD, Boguski MS, Stewart EA, and et al. A gene map of the human genome. *Science*, 274:540–546, 1996.

[11] L. D. Stein. Human genome: End of the beginning. *Nature*, 431:915–916, October 2004.

[12] E. Pennisi. Gene counters struggle to get the right answer. *Science*, 301:1040–1041, 2003.

[13] T. Hollon. Human genes: How many? *The Scientist*, 15, 2001.

[14] Bruce R. Korf. *human Genetics and Genomics*. Balckwell, 2007.

[15] Ian C. Gary, David A. Campbell, and Nigel K. Spurr. Advances in knowledge discovery and data mining. *Human Molecolar Genetics*, 9(16):2403–2408, 2000.

[16] N.J. Schork, D. Fallin, and J.S. Lancbury. Single nucleotide polymorphism and the future of genetic epidemiology. *Clinical genetics*, 58:250–264, 2000.

[17] K.E. Lohmueller, C.L. Pearce, M. Pike, E.S. Lander, and J.N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.*, 33:177–182, 2003.

[18] S. Daher, N. Sass, LG Oliveira, and R. Mattar. Cytokine genotyping in preeclampsia. *Am J Reprod Immunol.*, 55:130–135, 2006.

[19] A. Fekete, A. Ver, K. Boegi, A. Treszl, and J. Rigo J. Is preeclampsia associated with higher frequency of hsp70 gene polymorphisms? *Eur J Obstet Gynecol Reprod Biol*, 126:197–200, 2006.

[20] E. Kamali-Sarvestani, S. Kiany, B. Gharesi-Fard, and M. Robati. Association study of il-10 and ifn-gamma gene polymorphisms in iranian women with preeclampsia. *J Reprod Immunol.*, 72:118–126, 2006.

[21] G. Kobashi, K. Shido, A. Hata, H. Yamada EH Kato, M. Kanamori, S. Fujimoto, and K. Kondo. Multivariate analysis of genetic and acquired factors; t235 variant of the angiotensinogen gene is a potent independent risk factor for preeclampsia. *Semin Thromb Hemost.*, 27:143–147, 2001.

[22] J. Lin and P. August. Genetic thrombophilias and preeclampsia: a meta-analysis. *Obstet Gynecol.*, 105:182–192, 2005.

[23] Rolv Skjaerven, Lars J Vatten, Allen J Wilcox, Thorbjrn Ronning, Lorentz M Irgens, and Rolv Terje Lie. Recurrence of pre-eclampsia across generations: exploring fetal and maternal genetic components in a population based cohort. *BMJ*, 331, 2005.

[24] Pre-eclampsia - site www.preeclampsia.org. visited on 05-09.

[25] Fiona Lyall and Michael Belfort. *Pre-eclampsia: Etiology and Clinical Practice*. Cambridge University Press., 2007.

[26] The GOPEC Consortium. Disentangling fetal and maternal susceptibility for pre-eclampsia: A british multicenter candidate-gene study. *American journal. Human. Genetic*, 77:127–131, 2005.

[27] T. Saarela, M. Hiltunen, S. Helisalmi, S. Heinonen, and M. Laakso. Plasma cell membrane glycoprotein-1 k121q polymorphism in preeclampsia. *Gynecol Obstet Invest*, 31:124–127, 2006.

[28] E. Jaaskelainen, L. Keski-Nisula, S. Toivonen, EL. Romppanen, S. Helisalmi, K. Punnonen, and S. Heinonen S. Mthfr c677t polymorphism is not associated with placental abruption or preeclampsia in finnish women. *Hypertens Pregnancy*, 25:73–80, 2006.

[29] VL Doherty, AN Rush, SP Brennecke, and EK Moses. The -56t hla-g promoter polymorphism is not associated with pre-eclampsia/eclampsia in australian and new zealand women. *Hypertens Pregnancy*, 25:63–71, 2006.

[30] T. Saarela, M. Hiltunen, S. Helisalmi, S. Heinonen, and M. Laakso. Adiponectin gene haplotype is associated with preeclampsia. *Genet Test*, 10:1090–6576, 2006.

[31] T. Haekli, EL Romppanen, M. Hiltunen, S. Helisalmi, K. Punnonen, and S. Heinonen. Plasminogen activator inhibitor-1 polymorphism in women with pre-eclampsia. *Genet Test*, 7:265–268, 2003.

[32] Christoph Lange Nan M. Laird. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, pages 385–394, 2006.

[33] M. Manuguerra, G. Matullo, F. Veglia, H. Autrup, A.M. Dunning, S. Garte, E. Gormally, C. Malaveille, S. Guarrera, S. Polidoro, F. Saletta, M. Peluso, L. Airoldi, K. Overvad, O. Raaschou-Nielsen, F. Clavel-Chapelon, J. Linseisen, H. Boeing, D. Trichopoulos, A. Kalandidi, D. Palli, V. Krogh, R. Tumino, S. Panico, H.B.

Bueno-De-Mesquita, P.H. Peeters, E. Lund, G. Pera, C. Martinez, P. Amiano, A. Barricarte, M.J. Tormo, J.R. Quiros, G. Berglund, L. Janzon, B. Jarvholm, N.E. Day, N.E. Allen, R. Saracci, R. Kaaks, P. Ferrari, E. Riboli, and P. Vineis. Multi-factor dimensionality reduction applied to a large prospective investigation on gene-gene and gene-environment interactions. *Carcinogenesis*, 28(2):414–422, 2007.

[34] Miquel Porta. *Dictionary of Epidemiology*. Paperback, 2008.

[35] Case-control analysis - www.bmj.com/epidem/epid.8.html. visited on 05-09.

[36] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arised from random sampling. *Philosophical Magazine*, 5:157–175, 1990.

[37] Hardane JBS. An exact test for randomness of mating. *J. Genet*, 52:157–175, 1954.

[38] Wellek S. Test for establishing compatibility of an observed genotype distribution with hardy-weinberg equilibrium in the case of a biallelic locus,. *Biometrics*, 60:694–703, 2004.

[39] Lavene H. On a matching problems arising in genetic. *An. Mat. Stat.*, 20:91–94, 1949.

[40] Hernandez JL and Weir BS. A disequilibrium coefficient approach to hardy-weinberg equilibrium testing. *Biometrics*, 45:53–70, 1989.

[41] R.A. Fisher. On the interpretation of chi-square from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85:87–94, 1922.

[42] Janis EW, David JC, and Goncalo RA. A note on exact test of hardy-weinberg equilibrium. *Am. J.Hum. Genet.*, 76:887–893, 2005.

[43] Freidlin B, Zheng G, Li Z, and Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Him. Hered.*, 53:146–152, 2002.

[44] Christel Faes, Marc Aerts, Helena Geys, Geert Molenberghs, and Lieven Declerck. Bayesian testing for trend in a power model for clustered binary data. *Environmental and Ecological Statistics*, 11:305–322, 2004.

[45] Kijoung Song, Mohammed Orloff, Qing Lu, and Robert Elston. Fine-mapping using the weighted average method for a case-control study. *BMC Genetics*, 6(Suppl 1):S67, 2005.

[46] Margaret H. Dunham. *Data mining - Introductory and advanced topics*. Pearson Education Inc., 2003.

[47] J. Ross Quinlan. Induction of decision tree. *Machine Learning*, 1(1):81–106, 1986.

[48] J. Ross Quinlan. C4.5: Programs for machine learning. *Machine Learning*, 16(3):235–240, 1994.

[49] Yang Jiang, Qingpu Zhang, Xia Li, Lei Du, Wei Jiang, Ruijie Zhang, Jing Li, and Shaoqi Rao. Analysis of sib-pair ibd profiles using ensemble decision tree approach: Application to alcoholism. *Computational Intelligence and Bioinformatics*.

[50] Kuang-Yu Liu, Jennifer Lin, Xiaobo Zhou, and Stephen Wong. Boosting alternating decision trees modeling of disease trait information. *BMC Genetics*, 6(Suppl 1):S132, 2005.

[51] Dong-Hoi Kim, Saangyong Uhmn, Young-Woong Ko, Sung Cho, Jae Cheong, and Jin Kim. Chronic hepatitis and cirrhosis classification using snp data, decision tree and decision rule. *Computational Science and Its Applications ICCSA 2007*, 4707:585–596.

[52] Lung-Cheng Huang, Sen-Yen Hsu, and Eugene Lin. A comparison of classification methods for predicting chronic fatigue syndrome based on genetic data. *Journal of Translational Medicine*, 7(1):81, 2009.

[53] Spielman RS, McGinnis RE, and Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52:506–516, 1993.

[54] Sham PC and Curtis D. An extended transmission disequilibrium test (tdt) for multi-allele marker loci. *Ann Hum Genet*, 59:323–336, 1995.

[55] Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genetic Epidemiology*, 13:423–449, 1996.

[56] Spielman RS and Ewens WJ. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet*, 62:450–458, 1998.

[57] Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet*, 64:861–870, 1999.

[58] Knapp M. Using exact p values to compare the power between the reconstruction-combined transmission/disequilibrium test and the sib transmission/disequilibrium test. *Am J Hum Genet*, 65:1208–1210, 1999.

[59] Horvath S, Laird NM, and Knapp M. The transmission/disequilibrium test and parental-genotype reconstruction for x-chromosomal markers. *Am J Hum Genet*, 66:1161–1167, 2000.

[60] Abecasis GR, Cardon L, and Cookson WOC. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, 66:279–292, 2000.

[61] Seltman H, Roeder K, and Devlin B. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet*, 68:1250–1263, 2001.

[62] Horvath S, Xu X, and Laird NM. The family based association test method: strategies for studying general genotype-phenotype associations. *European J. Hum. Genet.*, 9:301–306, 2001.

[63] Laird NM. Rabinowitz. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Am J Hum Genet*, 50:211–223, 2000.

[64] Laird NM, Horvath S, and Xu X. Implementing a unified approach to family based tests of association. *Genetic Epi.*, 1:36–42, 2000.

[65] Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, and Ott J. A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur J Hum Genet.*, 2004.

[66] Snp assistant - site www.biodata.ee. visited on 05-09.

[67] David Clayton. Transmit - site www-gene.cimr.cam.ac.uk. visited on 05-09, 1999.

[68] K Zhang, J Zhu, J Shendure, GJ Porreca, JD Aach, RD Mitra, and GM Church. Long-range polony haplotyping of individual human chromosome molecules. *Nature genetics*, 38(3):382–387, 2006.

[69] C Burgtorf, P Kepper, M Hoehe, C Schmitt, R Reinhardt, H Lehrach, and S Sauer. Clone-based systematic haplotyping (csh): a procedure for physical haplotyping of whole genomes. *Genome research*, 13(12):2717–2724, 2003.

[70] C Ding and CR Cantor. Direct molecular haplotyping of long-range genomic dna with m1-pcr. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7449–7453, 2003.

[71] Kui Zhang, Zhaohui Qin, Ting Chen, Jun S. Liu, Michael S. Waterman, and Fengzhu Sun. HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics*, 21(1):131–134, 2005.

[72] N. Patil, A. J. Berno, D. A. Hinds andW. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–23, 2001.

[73] Nothnagel M, Frst R, and Rohde K. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered*, 54:186–198, 2002.

[74] Daniel O. Strama, Christopher A. Haimana, Joel N. Hirschhornb, David Altshulerb, Laurence N. Kolonelg, Brian E. Hendersona, and Malcolm C. Pikea. Choosing haplotype-tagging snps based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Human Heredity*, 55:27–36, 2003.

[75] C. Carlson, M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.,*, 74:106–120, 2004.

[76] Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet*, 22:139 – 144, 1999.

[77] Goncalo R. et al. Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet*, 68:191 – 197, 2001.

[78] Stacey B. Gabriel, Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S. Lander, Mark J. Daly, and David Altshuler. The Structure of Haplotype Blocks in the Human Genome. *Science*, 296(5576):2225–2229, 2002.

[79] Teruaki Tozaki, Kei ichi Hirota, Telhisa Hasegawa, Motowo Tomita, and Masahiko Kurosawa. Prospects for whole genome linkage disequilibrium mapping in thoroughbreds. *Gene*, 346:127 – 132, 2005.

[80] Jingwu He and Alexander Zelikovsky. Informative snp selection methods based on snp prediction. *IEEE TRANSACTIONS ON NANOBIOSCIENCE,*, 6:60–67, 2007.

[81] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, , and E. S. Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 294:1719–23, 2001.

[82] Kui Zhang, Minghua Deng, Ting Chen, Michael S. Waterman, and Fengzhu Sun. A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America*, 99(11):7335–7339, 2002.

[83] Kun Zhang and Li Jin. HaploBlockFinder: haplotype block analyses. *Bioinformatics*, 19(10):1300–1301, 2003.

[84] Peisen Zhang, Huitao Sheng, and Ryuhei Uehara. A double classification tree search algorithm for index snp selection. *BMC Bioinformatics*, 5(1):89, 2004.

[85] ZS Qin, T Niu, and JS Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *American journal of human genetics*, 71(5):1242–1247, 2002.

[86] Stata Technical Support. *Stata Statistical Software: Release 11.* Stata Press, 2009.

[87] Chapman JM, Cooper JD, Todd JA, and Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered*, 56:18–31, 2003.

[88] G Kimmel and R Shamir. Gerbil: Genotype resolution and block identification using likelihood. *Proceedings of the National Academy of Sciences of the United States of America*, 102(1):158–162, 2005.

[89] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[90] E Halperin, G Kimmel, and R Shamir. Tag snp selection in genotype data for maximizing snp prediction accuracy. *Bioinformatics*, 21(1):195–203, Jun 2005.

[91] Christopher S. Carlson, Michael A. Eberle, Mark J. Rieder, Joshua D. Smith1, Leonid Kruglyak, , and Deborah A. Nickerson. Additional snps and linkage-

disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genetics*, 33:518–521, 2003.

[92] Halldrsson BV, Istrail S, and De La Vega FM. Optimal selection of snp markers for disease association studies. *Hum Hered*, 58:190–202, 2004.

[93] Chuang Li-Yeh, Jr. Yu-Jen Hou, and Cheng-Hong Yang. A novel prediction method for tag snp selection using genetic algorithm based on knn. *World Academy of Science, Engineering and Technology*, 53:1325–1330, 2009.

[94] Fengyu Zhang and Diane Wagener. An approach to incorporate linkage disequilibrium structure into genomic association analysis. *J Genet Genomics*, 35:381–385, 2008.

[95] H. Abdi. *Bonferroni and Sidak corrections for multiple comparisons*. Sage, 2007.

[96] Improved power by use of a weighted score test for linkage disequilibrium mapping. *The American Journal of Human Genetics*, 80(2):353 – 360, 2007.

[97] Woosung Yang and Jun Nakaya. Statistical applications for snps analysis. *Chem-Bio Informatics Journal*, 6:55–68, 2006.

[98] Jingwu He and Alexander Zelikovsky. MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics*, 22(20):2558–2561, 2006.

[99] L Wang and Y Xu. Haplotype inference by maximum parsimony. *Bioinformatics (Oxford, England)*, 19(14):1773–1780, 2003.

[100] D Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J Comput Biol*, 8(3):305–323, 2001.

[101] D Gusfield. A practical algorithm for optimal inference of haplotypes from diploid populations. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB*, 8:183–189, 2000.

[102] Jingwu He and A. Zelikovsky. Linear reduction methods for tag snp selection. volume 2, pages 2840 –2843, sept. 2004.

[103] T Niu, ZS Qin, X Xu, and JS Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American journal of human genetics*, 70(1):157–169, 2002.

[104] M Stephens, NJ Smith, and P Donnelly. A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, 68(4):978–989, 2001.

[105] Phil Hyoun Lee and Hagit Shatkay. BNTagger: improved tagging SNP selection using Bayesian networks. *Bioinformatics*, 22(14):e211–219, 2006.

[106] Phuong TM, Lin Z, and Altman RB. Choosing snps using feature selection. *BMC Bioinformatics*, 4(2):241–57, 2006.

[107] B. Everitt. *Cluster Analysis*. Heinemann Educational Books, 1974.

[108] J. A. Hartigan. *Clustering Algorithms*. New York, Wiley, 1975.

[109] Leonard Kauffman and Peter J. Rousseeuv. *Finding groups in data - An introduction to cluster analysis*. John Wiley and Sons, Inc., 1990.

[110] Anil K. Jain and Richard C. Dubes. *Algorithm for clustering Data*. Prentice-Hall, Inc., 1988.

[111] Michael R. Anderberg. *Cluster analysis for Applications*. Academic Press new Tork and London, 1973.

[112] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[113] Paul S. Levy and Stanley Lemeshow. *Sampling of Populations: Methods and Applications*. John WIley and son, inc., 3rd edition, 1999.

[114] M. Daszykowski, B. Walczak, and D. L. Massart. Representative subset selection. *Analytica Chimica Acta*, 468(1):91 – 103, 2002.

[115] R. W. Kennard and L. A. Stone. Computer aided design of experiments. *Technometrics*, 11(1):pp. 137–148, 1969.

[116] Robert D. Clark. Optisim:? an extended dissimilarity selection method for finding diverse representative subsets. *Journal of Chemical Information and Computer Sciences*, 37(6):1181–1188, 1997.

[117] Brian D. Hudson, Richard M. Hyde, Elizabeth Rahr, John Wood, and Julian Osman. Parameter based methods for compound selection from chemical databases. *Quantitative Structure-Activity Relationships*, 15(4):289–289, 1996.

[118] Weka - site www.cs.waikato.ac.nz. visited on 05-09.

[119] Richard A. Becker, John M. Chambers, and Allan R. Wilks. *The new S language: a programming environment for data analysis and graphics*. Wadsworth and Brooks/Cole Advanced Books & Software, Monterey, CA, USA, 1988.

[120] Sapna Chadha, Katie Miller, Lisa Farwell, Liz B Lightstone, Mark J Daly, John D Rioux, , and Timothy J Vyse. Haplotype structure of tnfrsf5-tnfsf5 (cd40cd40l) and association analysis in systemic lupus erythematosus, 2005.

[121] G. R. Kazeem and M. Farrall. Integrating case-control and tdt studies, 2005.

[122] S Wu, C Fann, Y Jou, J Chen, and W Pan. Association between markers in chromosomal region 17q23 and young onset hypertension: a tdt study. *J Med Genet.*, 39:42–44, 2002.

[123] Bateson W. *Mendel's principles of Heredity*. Cambridge University Press, 1909.

[124] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[125] Suh YJ, Finch SJ, and Mendell NR. Application of a bayesian method for optimal subset regression to linkage analysis of q1 and q2. *Genet Epidemiol*, 21:706–711, 2001.

[126] Oh C, Ye KQ, He Q, and Mendell NR. Locating disease genes using bayesian variable selection with the haseman-elston method. *BMC Genet*, 4:S69, 2003.

[127] R. Santana, P. Larraaga, and J. A. Lozano. Challenges and open problems in discrete edas. Technical report, University of the Basque Country, 2007.

[128] Ruben Armananzas, Inaki Inza, Roberto Santana, Yvan Saeys, Jose Flores, Jose Lozano, Yves Peer, Rosa Blanco, Victor Robles, Concha Bielza, and Pedro Larranaga. A review of estimation of distribution algorithms in bioinformatics. *BioData Mining*, 1(1):6, 2008.

[129] Moore JH, Barney N, Tsai C-T, Chiang F-T, and andWhite BC Gui J. Symbolic modeling of epistasis. *Epistasis. Hum Hered*, 63:120–133, 2007.

[130] Tatsuya Okabe, Yaochu Jin, Bernhard Sendhoff, and Markus Olhofer. Voronoi-based estimation of distribution algorithm for multi-objective optimization. pages 1594–1601, 2004.

[131] Alden H. Wright and Sandeep Pulavarty. On the convergence of an estimation of distribution algorithm based on linkage discovery and factorization. pages 695–702, 2005.

[132] Kathryn L. Lunetta, Stephen V Faraone, Joseph Biederman, and Nan M. Laird. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *American journal of human genetics*, 66(2):605–614, 2000.

[133] Heather J. Cordell. Epistasis: what is means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[134] Bing Han, Meeyoung Park, and Xue-wen Chen. A markov blanket-based method for detecting causal snps in gwas. *BMC Bioinformatics*, 11(Suppl 3):S5, 2010.

[135] Kuang Yu Liu, Jennifer Lin, Xiaobo Zhou, and Stephen TC Wong. Boosting alternating decision trees modeling of disease trait information. *BMC Genet.*, 6, 2005.

[136] Qian Xie, Luke D Ratnasinghe, Huixiao Hong, Roger Perkins, Ze-Zhong Tang, Nan Hu, Philip R Taylor, and Weida Tong. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. *BMC Bioinformatics*, 6, 2005.

[137] Computer and information science and engineering - site www.cise.ufl.edu. visited on 05-09.

[138] Building classification models: Id3 and c4.5 - site www.cis.temple.edu. visited on 05-09.

[139] The myth of chance-corrected agreement - site visited on 11-10.

[140] Viera AJ and Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med.*, 37(5):360–3, 2005.

[141] Douglas G. Altman. *Practical Statistics for Medical Research.* CRC Press, 1991.

[142] Sean Hennessy, Warren B. Bilker, Jesse A. Berlin, and Brian L. Stromu. Factors influencing the optimal control-to-case ratio in matched case-control studies. *American Journal of Epidemiology*, 149(1):195–197, 1999.

[143] W.D. Dupont. Power calculations for matched case-control studies. *Biometrics*, 44:1157–1168, 1988.

[144] Wonkuk Kim, Derek Gordon, Jonathan Sebat, Kenny Q. Ye, and Stephen J. Finch. Computing power and sample size for case-control association studies with copy number polymorphism: Application of mixture-based likelihood ratio test. *PLoS ONE*, 3:3475, Oct 2008.

[145] Fulvio De Santis, Marco Perone Pacifico, and Valeria Sambucini. Optimal predictive sample size for case-control studies. *Journal Of The Royal Statistical Society Series C*, 53:427–441, 2004.

[146] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26:1651–1686, 1998.

[147] P. Royston. As 181: The w test for normality. *Applied Statistics*, 31:176–180, 1982.

[148] M Vatish, N.J Sebire, C Allgood, C McKeown, H.C Rees, and S.D Keay. Triploid/diploid mosaicism (69xxy/46xx) presenting as severe early onset

preeclampsia with a live birth: placental and cytogenetic features. *European journal of obstetrics, gynecology, and reproductive biology*, 112:233–235, 2004.

[149] SA Obed and Aniteye Patience. Birth weight and ponderal index in pre-eclampsia: A comparative study. *Ghana Med J*, 40:8–13, 2006.

[150] Xiao Yan Zhong, Stefan Gebhardt, Renate Hillermann, Kashefa Carelse Tofa, Wolfgang Holzgreve, and Sinuhe Hahn. Parallel Assessment of Circulatory Fetal DNA and Corticotropin-Releasing Hormone mRNA in Early- and Late-Onset Preeclampsia. *Clin Chem*, 51(9):1730–1733, 2005.

[151] C. Nelson-Piercy. *Pre-eclampsia*, chapter Pre-eclampsia: the women at risk. RCOG Press, 2003.

[152] James A. Thorp, Marjorie L. Zucker, Fred V. Plapp, Jane M. Rachel, and Craig Hinkle. Factor v leiden gene mutation and preeclampsia? *Journal of Maternal-Fetal Investigation*, 7(1):19–20, 1997.

[153] A. Goate, M.C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, L. James, R. Mant, P. Newton, K. Rooke, P. Roques, C. Talbot, M. Pericak-Vance, A. Roses, R.Williamson, M. Rossor, M. Owen, and J. Hardy. Segregation of a missense mutation in the amyloid precursor protein gene with familial alzheimer's disease. *Nature*, 349:704–706, 1991.

[154] E.I. Rogaev, R. Sherrington, E.A. Rogaeva, G. Levesque, M. Ikeda amd Y. Liang, H. Chi, C. Lin, K. Holman, T. Tsuda, L. Mar, S. Sorbi, B. Nacmias, S. Piacentini, L. Amaducci, I. Chumakov, D. Cohen, L. Lannfelt, P.E. Fraser, J.M. Rommens, and P.H. St George Hyslop. Familial alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the alzheimer's disease type 3 gene. *Nature*, 376:775–778, 1995.

[155] CP Hersh, DT Miller, DJ Kwiatkowski, and EK Silverman. Genetic determinants of c-reactive protein in copd. *Eur Respir J*, 28:1156–1162, 2006.

[156] Nathaniel H. Robin, Paul B. Tabereaux, Raymond Benza, and Bruce R. Korf. Genetic Testing in Cardiovascular Disease. *J Am Coll Cardiol*, 50(8):727–737, 2007.

[157] Jonathan L. Haines and Margaret A. PericakVance. *Genetic Analysis of Complex Disease*. WileyBlackwell, 2nd edition edition, 2006.

[158] Jinhwa Kim, Chaehwan Won, and Hyeonsu Byeon. A genetic multi-agent rule induction system for stream data. *Networked Computing and Advanced Information Management, International Conference on*, 2:54–58, 2008.

[159] Anthony J Barr and James Howard Goodnight. *Statistical analysis system*. Raleigh, 1971.

[160] Julie Pallant. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using SPSS for Windows Version 15*. Open University Press, Milton Keynes, UK, USA, 2007.

[161] Michael Widenius and Davis Axmark. *Mysql Reference Manual*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2002.

[162] Bruce S. Weir. *Genetic data analysis II*. Sinauer Associated, Inc., Sunderland, Massachuttes, second edition, 1996.

[163] K. Hao. Genome-wide selection of tag snps using multiple-marker correlation. *Bioinformatics*, 23(23):3178–3184, 2007.

[164] Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, and Jason H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, 69:138–147, 2001.

[165] Yujin Chung, Seung Yeoun Lee, Robert C. Elston, and Taesung Park. Odds ratio based multifactor-dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics*, 23:71–76, 2007.

[166] D Brassat, AA Motsinger, SJ Caillier, HA Erlich, K Walker, LL Steiner, BA Cree, LF Barcellos, MA Pericak-Vance, S Schmidt, S Gregory, SL Hauser, JL Haines, JR Oksenberg, and MD Ritchie. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in african americans. *Genes Immun.*, 7:310–315, 2006.

[167] M. Byng, J.C. Whittaker, A.P. Cuthbert, C.G. Mathew, and C.M Lewis. Snp subset selection for genetic association studies. *Ann. Hum. Genet*, 67:543–556, 2003.

[168] G. Johnson, L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dubridge, and et al. Haplotype tagging for the identification of common disease genes. *Nat. Genet*, 29:233–237, 2001.

[169] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[170] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. In *In Proceedings of the Computational Systems Bioinformatics Conference*, pages 523–529, 2003.

[171] Remove duplicate entries software.

[172] G. Warnes. The genetics package. *R News*, 3, 2003.

[173] Juan R. Gonzalez, Lluis Armengol, Xavier Sole, Elisabet Guino, Josep M. Mercader, Xavier Estivill, and Victor Moreno. Snpassoc: an r package to perform whole genome association studies. *Bioinformatics*, 23(5):654–a–655, 2007.

[174] Pak Sham. *Statistics in Human Genetics.* Arnold, 1998.

[175] Clare Constantine, Lyle Gurrin, Christine McLaren, Melanie Bahlo, Gregory Anderson, Chris Vulpe, Susan Forrest, Katrina Allen, Dorota Gertig, and the HealthIron Investigators. Snp selection for genes of iron metabolism in a study of genetic modifiers of hemochromatosis. *BMC Medical Genetics*, 9(1):18, 2008.

[176] Christopher S. Carlson, Michael A. Eberle, Mark J. Rieder, Qian Yi, Leonid Kruglyak, and Deborah A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74:106–120, 2004.

[177] Chris Fraley and Adrian E. Raftery. Mclust version 3 for r: Normal mixture modeling and model-based clustering. Technical report, 2006.

[178] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001. 10.1023/A:1012801612483.

[179] Jarkko Salojaervi, Kai Puolaemaki, and Samuel Kaski. Expectation maximization algorithms for conditional likelihoods. pages 752–759, 2005.

[180] Sam Roweis. Em algorithms for pca and spca. pages 626–632, 1998.

[181] Jong-Hoon Ahn and Jong-Hoon Oh. A constrained em algorithm for principal component analysis. *Neural Comput.*, 15:57–65, 2003.