Analysing and predicting differences between methylated and unmethylated DNA sequence features

Isse Ali

Faculty of Technology

De Montfort University

A thesis submitted for the degree of

Doctor of Philosophy

2015

This thesis is primarily dedicated to: my family

Acknowledgements

An Exceptional thanks and sincere gratitude to my supervisors, Professor David Elizondo and Professor Martin Grootveld , who has supported me during the writing period and for offering to be my supervisors after Dr Huseyin seker left, despite having full number of PhD-students; who gave me full support I would like to give great thanks for smoothed out difficulties during the write up of the thesis and the constructive feedback, their input made this thesis worthwhile. I am truly grateful to their time in enhancing the structure and content of this thesis.

The work described in this thesis is the original work of the author except where specific reference or acknowledgement is made to the work or contribution of others. The following conference papers and journal articles have been produced from the PhD research within the thesis. All the authors have made valuable contributions to these publications.

- I. Ali and H. Seker. Detailed methylation prediction of CpG islands on human chromosome 21. Proceedings of the 10th WSEAS International Conference on MATHEMATICS and COMPUTERS in BIOLOGY and CHEMISTRY. ISSN: 1790-5125, PP : 147-152, 2009.

- I. Ali and H. Seker. A comparative study for characterisation and prediction of tissue-specific DNA methylation of CpG islands in chromosomes 6, 20 and 22.32nd Annual International Conference of the IEEE EMBS. ISSN: 1557-170X , pp : 1832 1835, 2010.

- Ali I, Seker H. 2010. "An identification and prediction methods for feature-subsets of CpG islands methylation based on human peripheral blood leukocytes of chromosome 21q . Conf. Proc. IEEE BIBE 2010:289-290.

Papers due to submit

- Isse Ali, David Elizondo and Martin Grootveld. DNA methylation display on Aging and gender differences based on unsupervised clustering.(Chapter-4)

- Isse Ali and David Elizondo and Martin Grootveld. Prediction of methylation classes of cpg islands on chromosomes 6, 20, 21 and 22" (Chapter-5).

- Isse Ali, David Elizondoand Martin Grootveld, Weighting methods towards severely imbalanced data.(Chapter-5)

- Isse Ali, Martin Grootveld and David Elizondo, Analysis of gender differences in DNA methylation. (Chapter-6)

Abstract

DNA methylation is involved in various biological phenomena, and its dysregulation has been demonstrated as being correlated with a number of human disease processes, including cancers, autism, and autoimmune, mental health and neuro-degenerative ones. It has become important and useful in characterising and modelling these biological phenomena in order to understand the mechanism of such occurrences, in relation to both health and disease. An attempt has previously been made to map DNA methylation across human tissues, however, the means of distinguishing between methylated, unmethylated and differentially-methylated groups using DNA sequence features remains unclear. The aim of this study is therefore to: firstly, investigate DNA methylation classes and predict these based on DNA sequence features; secondly, to further identify methylation-associated DNA sequence features, and distinguish methylation differences between males and females in relation to both healthy and diseased, statuses. This research is conducted in relation to three samples within nine biological feature sub-sets extracted from DNA sequence patterns (Human genome database). Two samples contain classes (methylated, unmethylated and differentially-methylated) within a total of 642 samples with 3,809 attributes driven from four human chromosomes, i.e. chromosomes 6, 20, 21 and 22, and the third sample contains all human chromosomes, which encompasses 1628 individuals, and then 1,505 CpG loci (features) were extracted by using Hierarchical clustering (a process Heatmap), along with pair correlation distance and then applied feature selection methods. From this analysis, author extract 47 features associated with gender and age, with 17 revealing significant methylation differences between males and females. Methylation classes prediction were applied a K-nearest Neighbour classifier, combined with a ten-fold cross- validation, since to some data were severely imbalanced (i.e., existed in sub-classes), and it has been established that direct analysis in machine-learning is biased towards the majority class. Hence, author propose a Modified- Leave-One-Out (MLOO) cross-validation and AdaBoost methods to tackle these issues, with the aim of compositing a balanced outcome and limiting the bias interference from inter-differences of the classes involved, which has provided

potential predictive accuracies between 75% and 100%, based on the DNA sequence context.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bioinformatics consists of biological data mining through the use of computational methods. This includes extracting meaningful and analysable features (or feature sets) from a large amount of experimental data, including microarray protein expression data, exome arrays, Ilumina arrays and DNA methylation arrays found in biomedical database resources. Bioinformatics employs statistical applications to determine and understand biological features in both health and disease, and groupings based on biological functionality, along with prediction association disease, i.e. DNA patterns associated with methylation and its relationship to DNA patterns, in order to predict DNA methylation loci and annotate their biological function and position within a particular gene. Bioinformatics uses machine-learning methods, along with statistical models and tools, to analyse DNA sequence features in relation to their biological functions.

1.1 Human Chromosomes and their physiological and biological functions

This section describes the existing information and theoretical background of epigenetics, along with its associated terminology and physiological and biological processes, including the terminology employed in the literature. It also explains the DNA sequence features used throughout the current study.

In an autosome cell, there is a balance of human chromosomes, comprising 23 pairs that make up the total of 46 chromosomes in the nucleus of each cell. Two pairs of these chromosomes are sex chromosomes, distinguishing females (XX) from males (XY). The other 22 pairs are identical in shape for both males and females, and are named autosomal chromosomes. Each chromosome consists of a long thread of DNA wrapped with proteins (i.e. the histones, see Figure 1.1) that make up a compact structure, allowing it to fit inside the cell nucleus. The chromosomes are passed through the generations from parents to offspring, and give specific instructions, thus

creating unique features in each offspring. Complex organisms, such as humans, inherit one copy of each chromosome from each parent, and this indicates that some diseases or defects are inherited from parents. An imbalance in chromosomes can also cause serious defects in development and growth.

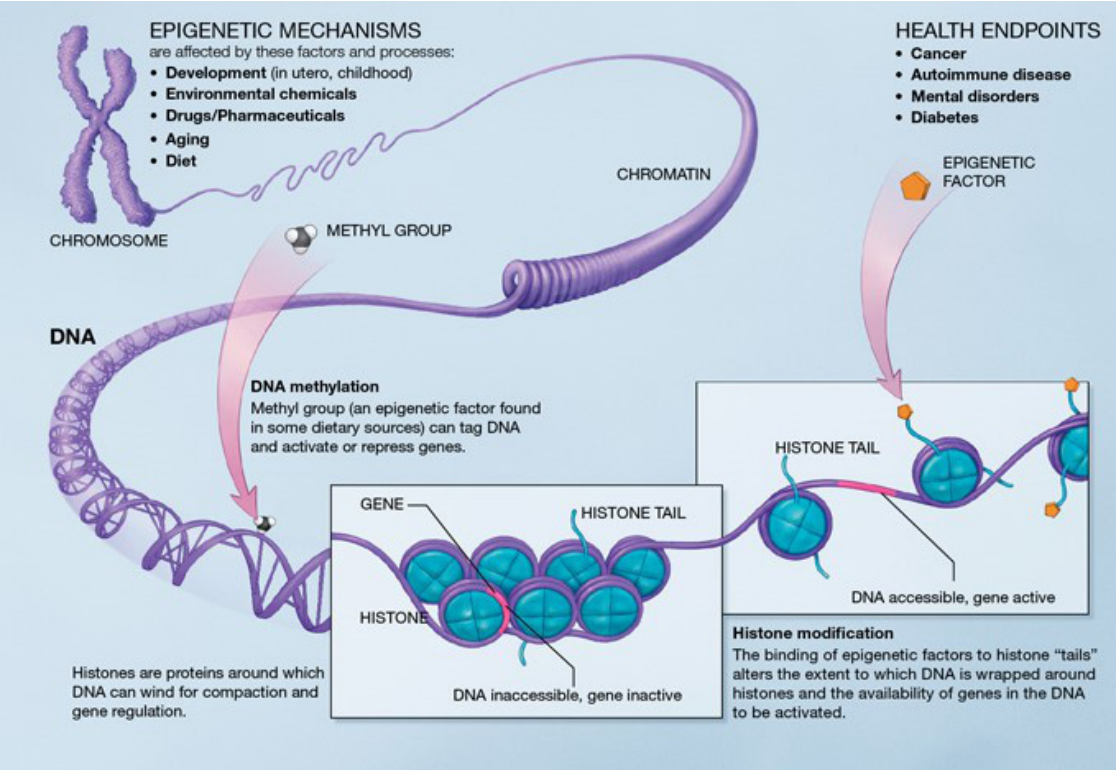


Figure 1.1 Representation of epigenetic mechanisms related to health and disease (adapted from [1]).

The cell divides from a single cell to multi-cellular tissue, organs and an organism. The growth of cells does not stop, but is economically controlled and only produced when it is required, i.e. skin cells dividing to repair damaged cells, replace old ones and repair faulty cells caused by cell division process itself. Chromosomes play an important role in this process, ensuring an even distribution of the copies of the chromosomes, and the correction of imbalanced distribution during cell growth. However, mistakes can take place during cell division, including the appearance of more, or fewer, chromosomes in a cell, along with structural changes, i.e. where components of chromosomes are either broken down or joined to other chromosomes. These changes cause serious issues, including cancer. This is mainly leukaemia, and other forms of cancer, found when the components of chromosomes are broken, deleted, or attached to other chromosomes.

More than one copy of some chromosomes per cell can cause developmental defects and mental retardation, an example of which is Downs syndrome, in which the cell

has one additional copy of chromosome 21. Downs syndrome is a common cause of mental retardation and health issues (resulting from simply having one additional copy of chromosome 21). This imbalance of chromosom disjoint is caused by DNA methylation, which affects up to 1 in 700 live births in the UK. It is also associated with a number of further disorders, including congenital heart disease, early-onset Alzheimers disease, and the risk of leukaemia (i.e. cancer) [7].

The 22 chromosomes are each given a number based on their size, being ordered from the largest to the smallest, i.e. chromosome 1 to chromosome 22. Chromosomes can be viewed during cell division through the use of a microscope, and as they compact with histone molecules. This compactness is reduced when cell division takes place. Indeed, they separate in the centromere position, which is between the two arms of the chromosome. For the terminology using Q and P arms, the Q-arm is generally greater than the P-arm (Figure 1.2). The arms are used to locate the position of both the chromosome and the gene, and indicate which part of the chromosome is displaced or deleted. In humans, each chromosome contains two long strands of the DNA chain, which is twisted and shaped into a double helix. Genes are located in the double helix strand of DNA, which is paired, and makes up only 1% of the genome. The functional genes are estimated to number between 20 to 25 thousand. Each gene carries specific genetic information concerning a specific role in relation to the body, i.e. growth, development, the defence system, and other millions (if not billions) of tasks. Each gene is in possession of a unique code controlling how and when it will be activated.

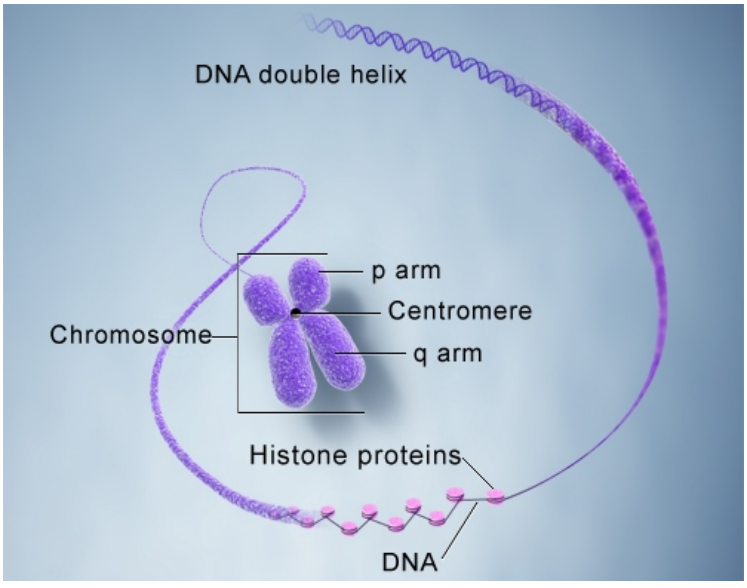


Figure 1.2 Schematic human chromosome packed with histone proteins and DNA (Adapted from [2]).

The code consists of DNA letters, with the combination of three of the four letters in a particular order being repeated. The combination (i.e. chemical bonding) of these three letters is known as a triplet for a particular order of producing a specific molecule (i.e. amino acid), which further builds up to create specific proteins required by the body to function physiologically. As discussed above, this leads to unique variations in each individual (with the exception of identical twins).

However, in some cases, the triplet combination is faulty, leading to the generated proteins being more or less active, or being produced in an inappropriate cell. These kinds of errors are known as gene mutations, and can cause defects in growth, development and other genetic conditions, i.e. tissue specificity: different cell types express the different proteins they require for their normal function in view of their differing roles in the body.

Therefore, identical genes in different cell types have diverse functions. For example, liver cells express lipase (an enzyme that digests fat particles), whereas in brain cells the same genes are completely switched off. One of the best known forms of gene control is DNA methylation: gene loci are attached to or removed from methyl groups (CH3), in a particular position on DNA, normally CpG, where the gene is switched on or off. However, an error can also cause imbalanced protein expression that causes genetic disorders, i.e. ageing, cancer, autism, auto-immune disorders, mental health issues, and neuro-degenerative disorders.

1.1.1 Epigenetics, CpG islands and DNA methylation

In epigenetic research, it is essential to investigate DNA sequence features, particularly DNA methylation for healthy and diseased samples, and also develop models as tools to determine (and understand) the manners in which DNA modification regulates gene activity, without changing the DNA sequence of that particular gene. This gene regulation is known as epigenetics, and is defined as "the study of the process by which genetic information is transferred into the substance and behaviour of an organism, i.e. the study of heritable changes that occur without any change in the DNA sequence" [8].

A further definition of epigenetics [8] is "the study of the chemical modification of specific genes, or the gene-associated proteins of an organism. The epigenetic mechanism can be defined as the way in which the cell information in genes is expressed and used by cells". The term 'epigenetics' came into general use in the early 1940s, when British embryologist, Conrad Waddington, used it to describe the interaction between genes, and gene products directing development, and which give rise to an organism phenotype (i.e. observable characteristics). Information subsequently revealed by epigenetic studies have revolutionised the field of genetics and developmental biology, and have, in particular led to the identification of a number of possible chemical modifications to DNA and proteins, i.e. the histones that are 'wrapped up' with the DNA

in the nucleus. These modifications can be determined when (or even if) a given gene is expressed in a cell or organism.

1.1.2 Epigenetics

The term, epigenetics is formed of the Greek word 'epi-', meaning 'over', 'above', and 'added to', and refers to the regulation of gene expression without changes in the underlying DNA sequences. Epigenetics, along with its related terms, have a number of meanings. Firstly, it is essential to define the genome, which is a complete set of haploid DNA and the functional component that it codes [9]. In the nucleus, DNA exists as a highly compressed structure consisting of DNA and proteins, otherwise known as chromatin. The epigenome is a sum of both the chromatin structure and patterns of DNA methylation, which is itself the result of an interaction between the genome and the environment.

Currently, three definitions are in use in the literature for the term 'epigenetic'. The main definition includes the transmission and maintenance of information through meiosis or mitosis (i.e., cell differentiation). This process is not limited to DNA-based transmission, but can also be protein-based (as broadly used in the literature on yeast [10; 11]). Meiotically- and mitotically-heritable changes in gene expression are not accompanied by changes in the DNA sequence. The altered patterns of gene expression can occur through a number of mechanisms based on DNA, RNA or proteins. This definition has been developed through developmental biology and by cancer researchers [12; 13; 14]. The other definition of epigenetics is a mechanism for the stable maintenance of gene expression involving the physical 'marking' of DNA or its associated proteins [11; 15].

Epigenetic processes are important for development and differentiation, in order to protect cellular function from being hijacked by abnormal processes, including gene expression deregulation in cancer [14; 15]. It is well reported that different levels of gene expression in different cellular states depend on mechanisms affecting the epigenetic process without changing the DNA sequences [9; 15; 16]. Such alterations can equally enhance, and repress gene expression most likely to influence DNA methylation and hence alter chromatin structures [9; 17].

1.1.3 CpG islands and DNA Methylation

DNA contains four bases, which refer to four letters of the alphabet: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G) [18]. These letters are linked by a phospodiester (p), which joins the two bases, i.e. cytosine-phospho-guanine (CpG). The probability of finding CpG dinucleotides in any given DNA sequence is 1/16; however, it has been found at much lower levels in the human genome [9]. The reason for this is that CG suppresses all genomes using cytosine methylation, and may refer to

the hypermutability of methylated cytosine. This CG suppression is found throughout the human genome [15], although small areas exist in which the density of CpGs is considerably higher than the expected values. The areas are approximately 300 to 3000 base-pairs long, and are known as CpG islands. They represent approximately 1% of the human genome sequence [18]. CpG islands have escaped the suppression of CG during the process of evolution, in view of the fact that they are not methylated, and therefore have escaped the above-noted mutational pressure [15; 18]. Over 60% of CpG islands are found in the promoter region, i.e. the 5 gene expression site [19]. The research community has a greater interest in promoters containing CpG islands, since when they are methylated they become permanently silenced, and therefore change a gene expression, and are inherited through mitosis (i.e. cell division) without any associated DNA sequence alterations [9].

DNA methylation is a chemical modification mediating an enzyme methyltransferase, which adds the methyl group at the CG DNA sequence site [18]. In the human genome, methylation of the cytosine molecule at CpG nucleotides in DNA is one of the major epigenetic alterations, and provides an important mechanism for distinguishing active genes from the inactive [9; 15]. Methylation in vertebrate DNA is limited to cytosine (C) nucleotides in the sequence CG, which is base-paired to precisely the same sequence, and in an opposite direction on the other strand of the DNA Helix [18]. Consequently, a single mechanism permits the existing pattern of DNA methylation to be directly inherited by the daughter DNA strands. An enzyme known as maintenance methylase (Figure 1.3) acts preferentially on those CG sequences that are base-paired with a CG sequence which is already methylated. Consequently, the pattern of DNA methylation on the parental DNA strand will act as a template for methylation of the daughter DNA strand, causing this pattern to be inherited directly following DNA replication [10; 15].

In addition, DNA methylation acts as a stimulant in the development of cancer, since it activates (or represses) certain cancer-associated genes [14]. CpG islands methylation, specifically in the promoter region, frequently leads to silencing of tumour suppressor genes. In different primary tumours, the cell cycles associated with inhibitors are hypermethylated, leading to the escape of cancer cells from apoptoses (i.e., cell death) and allowing them to continue proliferating. It has been reported that a major characteristic of human cancer and ageing consists of disruption of the epigenetic machinery and its features [11; 14; 17].

1.2 Aims and objectives of the thesis

DNA Methylation is primarily involved in a number of biological processes, including gene silencing, structural chromosomal stability, parental imprinting and suppressing the mobility of retrotransposons [20; 21; 22]. The disruption of DNA methylation

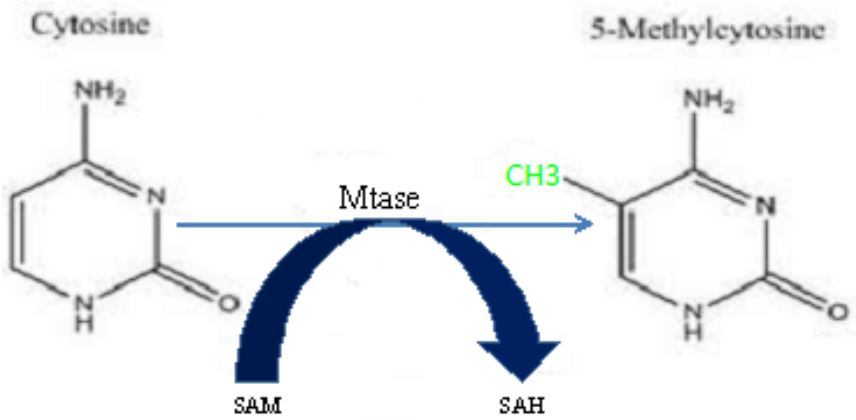


Figure 1.3 DNA methylation schematic process: DNA methylation means adding of methyl (CH3-group) to CG bases order. This reaction is activated by DNA-methyltransferase enzyme (Mtase) and uses S-adenosyl methionine (SAM) as a methyl donor.

has also been linked to various human diseases, including: cancer and human ageing [11; 12; 14; 21]; metabolic disorders [23]; complex age-related diseases (such as Alzheimers disease); developmentally linked illnesses (i.e. autism) and mental illness (i.e. depression) [20; 24]. It should be noted that, despite all advances to date, the analysis of DNA methylation status remains a challenging issue, particularly in relation to the human genome.

The research question has been established as: How can DNA methylation classes be distinguished by employing existing bioinformatics methods, and can DNA-sequence features or feature-subsets specific to the DNA methylation classes be extracted or predicted?

In response to this research question, the current study is focused on examining the differences between DNA methylation classes within extracted DNA-sequence features (see Chapter 5 for details), and CpG loci positions specific to gender, ageing and cancerous samples (see chapters 4 and 6 for details of the experiments). The researcher explores and develops a number of statistical models (methods) in order to predict and analyse issues relating to DNA methylation, including two novel methods: (1) Modified Leave-One-Out Cross Validation (MLOOCV); and (2) Cost-sensitive approach combined with Adaboost (C-Adaboost) (see Chapter 5 for details).

The aim of this work is to establish a background to DNA-methylation phenomena and issues related to epigenetics analyses, specifically in relation to small samples with large variable features. Synchronously, it will examine the work that has been undertaken, along with the issues faced during this aspect of the study, along with the

solutions. Throughout this document, an informed argument is established, leading the researcher to the appropriate direction and scope.

It is necessary for traditional statistics to adapt and be redesigned in order to manage these issues, although in a number of cases, this works effectively in creating a result that can be used in daily life. The work of the researcher reveals that this is not always the case, particularly when redesigning the traditional leave-one-out cross validation, which resulted in good predictive accuracy and greater reliability (Chapter 5).

Feature extraction and selection has been employed in many applications, and has also been shown to lead to an improved solution, including cancer research data from gene expression, and the prognosis of various drug treatments, together with genomic data from different species. It is considered that feature selection algorithms are overpowered when compared to other algorithms, giving the most effective solution regarding evaluation of biological data. However, it is necessary to adopt and redesign each study in order to apply to particular biological data, in view of its particular variation and the nature and the sources of this (see chapters 4 and 6 for details of the experimental design and its analysis). The outcome will assist in the design of target drugs, and also improve the manners in which patients are cured, a phenomenon based on the weight of patients and which does not include considerations of the impact of other factors (e.g. DNA methylation, tissue specificity and drug resistance or toxicity), a process potentially causing DNA methylation variation in different tissues.

1.3 Thesis contribution and outline

DNA methylation status in gender and ageing is the one of most important twenty first century issues in relation to epigenetic and personalised medication. 1505 CpG methylation positions of human chromosomes were investigated; 47 CpG loci positions specific to gender were extracted by the use of hierarchical clustering (Heatmap) with pair correlation distance methods. The CpG loci revealed significant methylation differences between male and female for healthy samples. An average linkage method was employed, and the subgroup further enlarged, in which each branch (dendrogram) represented a feature connected with a group of features, which have been graphically displayed. In addition, 11 CpG loci positions were identified, which were methylated in healthy samples, whereas the same CpG loci positions were unmethylated in cancerous samples (a Journal paper is in preparation [25]).

A comprehensive analysis was undertaken to identify and distinguish methylated, unmethylated and differentially methylation classes, based on extracted DNA sequence features. The extracted features were grouped according to their biological functions and further applied to K-nearest classifier (KNN) combined with tenfold cross validation. Due to the fact that a number of the datasets were severely imbalanced, it

has been established that direct analysis in machine learning is biased towards the majority class, i.e. the class with the most observations. This was required to design a reliable model to tackle the problems. The MLOO method was developed, which demonstrated improved predictive performance and consistent predictive accuracies in comparison to traditional KNN (conference paper are published [5; 6; 26] and Journal is due to submit [27]).

The researcher proposed the use of a combination of methods, weighting- and cost-sensitive, which resulted in improvements to the predictive performance of the minority class. The weighting assisted not only with the accuracy of the minority class, but also those classes not easy to separate. The results demonstrated an imbalance metrics improvement, including F-measure and G-mean. The values were judged on how well the model responded to the unseen data (test set). The combinatorial methods revealed an improved predictive performance for methylation classes in comparison to the single method (either weighting or prior probability (a Journal paper due to submit [28]).

A comprehensive feature selection was built, in order to distinguish DNA methylation differences between males and females by the investigation of specific features (loci positions) that play a role in DNA methylation for both health and cancer data. There was a further investigation of tissue specificity for male and female in DNA methylation differences in relation to the health of both genders. 10 features were identified in a colon tissue sample, which demonstrated DNA methylation differences in both male and female healthy samples. This is a major finding, as these features are not associated with gender specific tissues. This work has identified the most informative features, or feature subsets, distinguishing DNA methylation differences between male and female healthy samples. The soundness of the model has been demonstrated by the fact that, during the selection process, none of the features were ignored, and all went through in the search with a combination of pairings, and those with the lowest misclassification error rate were selected (A Journal paper is under corrections process [29]).

The rest of the report is organised as follows:

- Chapter **two** goes through the details of problems of epigenomic and genomic, informing the reader of the current state of these related technologies and giving details about important unsolved problems. It investigats imbalanced dataset prediction problems by using machine learning and looks into he current prediction methods and used DNA patterns that the methodology and evaluated various algorithms reported its prediction of fairness.

- Chapter **three** describes datasets that are used for the study and briefly evaluate techniques and models of learning algorithms. It gives details of author work in extending the traditional statistical methods to allow the processing imbalance

epigenetic data.

- Chapter **four** presents CpGs methylation gender differences display by employing unsupervised cluster analysis with average linkage method. It reported the extracted CpG loci positions that show significant methylation differences between male and female and also ageing methylation differences for both healthy and diseased sample (data).

- Chapter **five** presents outcome of the largest comprehensive data analysis of DNA sequence extracted features. This chapter looks into at what features subsets are associated with DNA methylation; it also compares DNA methylation classes employing designed predictive models and traditional predictive algorithms.

- Chapter **six** precedes onward the same dataset in Chapter 4 by employing features selection methodology to cope with noisy and large variable features and to select the most informative feature in relation of CpG loci methylation differences on gender. In this chapter is presented CpG loci position that are shown a significant CpG methylation differences on gender.

- Chapter **seven** gives an outline conclusion from the main body of the thesis reported in chapters 4, 5 and 6; a summary of the contribution of this work, strength and limitations of the research and states the future direction of research in this area.

Chapter 2

Literature review

Following the above brief introduction to the background of this study, and its aims and objectives, the current chapter contains three main sections (each with a subsection), exploring: (1) the existing information in relation to DNA methylation; (2) its associated functions; (3) its disease status; (4) the current picture in relation to bioinformatics introduced by researchers, and the methodology used in DNA methylation prediction. Furthermore, it will examine the issues raised by researchers, along with the misrepresentation of methods of analysis, and incomplete conclusions. The main aim of the current study is to correct such incomplete conclusions, and to design a representative predictive model to improve the current predictive methods in DNA-methylation classes and imbalanced data analysis, which has been previously ignored by the literature. In addition, this chapter will give the reasons behind such incomplete conclusive methods of prediction. It will, however, commence with a brief summary of each of the three sections.

2.1 Summary of the Literature review

DNA methylation is a biochemical modification of eukaryotic DNA, which generally occurs at the fifth (C5) position of cytosine residues in a 5-CG-3 position, known as CpG dinucleotide [10; 12; 15; 30]. In vertebrates, cytosine residue methylation in CpG nucleotides is an epigenetic marker necessary for physiological cell differentiation [19]. It has been demonstrated that over 60% of human genes promoters consist of unmethylated CpG islands [19]. Methylation of CpG islands are primarily involved in various biological processes, including gene silencing, structural chromosomal stability, parental imprinting and suppressing the mobility of retrotransposons [10; 15]. The disruption of DNA methylation has also been linked to various human diseases, such as cancer [10; 15; 31].

The prediction of DNA methylation classes is one of the most complex and challenging problems in bioinformatics, since the DNA sequence features characterising

methylation (in particular CpG islands) are dispersed throughout the human genome. However, advances in high-throughput technologies for computational genomics and epigenomics have assisted the ability to analyse large and variable amounts of data obtained from methylated, unmethylated and differentially-methylated DNA CpG islands. It should be noted that, despite the advances, the analysis of DNA methylation, and particularly for the human genome, remains a new and challenging issue.

One of the first studies of the Human Genome Project focussed on DNA methylation profiling of the Human Major Histocompatibility Complex, which has been shown to be the most gene-dense region in the human genome and containing genes with a diversity of functions (i.e., the immune system) on chromosome 6 (6p21.3) [32]. Yamada *et.al* [33] studied CpG islands methylation patterns on chromosome 21q, identifying 149 CpG islands, 103, 31 and 15, which were found to be unmethylated, methyalted and partially methylated, respectively. CpGs are calculated by computation through counting a number of CpGs in the sequence window [34]. An extended sequence window was added to obtain an unbiased predictive performance of the CpG islands-wide human genome [35]. However, any prediction approach depends on constraints applied, since CpG islands vary their distribution in the human genome, occurring 60% in the promoter region and housekeeping genes [19; 36; 37]. These CpG islands are mostly unmethyalted, while CpG islands in other regions are mostly methylated [19; 38]. Different cell lineage and tissues contain distinctive methylation patterns [30; 39; 40; 41]. This specific methylation is inherited through an unknown epigenetic process. However, not only is the diversity of tissues and cell lines methylation found to be a methylation profile, the disruption of methylation processes are suspected to cause a variety of methylation alterations [42], i.e. types of cancer demonstrate distinct methylation on tumour suppression genes [30; 43; 44]. It is clear that varieties of specific methylation profiles of tissues, cell lines and disease types require a development of a variety of predictive models and tools to specifically determine the methylation status of each condition. The attributes of CpG islands have been used to predict the methylated from the unmethylated [45]. Features extracted from DNA, its transcription binding site (TBS), and alu-, di- and trinucleotides were used to distinguish methylated and unmethylated classes[46]. In addition, DNA sequence features extracted repeat physio-chemical-properties and histone medication, which were used for methylation prediction [47; 48]. Other researchers have added extended features, including: CpG distance to transcription start site; CpG island frequencies in the methylated window; and methylated CpGs in a flanking sequence [38; 49; 50; 51].

A number of other researchers have also attempted to predict the DNA methylation of CpG islands [46; 52; 53; 54; 55]. However, their studies were limited, since they only considered the nucleotide sequence (CpGIs) and the Transcription Factor Binding Site (TFBS), which provides only an incomplete view of human DNA methylation. Bock *et.al.* [48] have recently extended Yamadas study by extracting DNA sequence

features associated with CpG islands, and analysing the data using statistical methods. However, the detail and consistent analysis of the features was not undertaken. In addition, the statistical approach used for the analysis was found to be insufficient, with potential for a misleading outcome, in view of the nature of such complex data. The aim of this study is therefore to develop a statistical strategy and undertake a detailed and comprehensive analysis of the features in order to establish a more accurate and reliable prediction of unmethylated, methylated and differentially-methylated CpG islands.

DNA methylation currently employs experiments using bisulphite DNA sequencing for a specific genomic region [49; 56]. The targeted region has been extended to the complete human genome, in order to employ predictive models to reduce experimental costs, and increase the speed of the methylation detection process [57; 58; 59]. However, it has been necessary to determine (sub-) features specific to DNA methylation, particularly in relation to health and disease status. These prediction approaches require numerical values representing selected features to distinguish between methylated and unmethylated CpGs. These features are the most studied DNA sequence patterns for the prediction of methylation status. However, the combination (or grouping) of features is essential. This study has grouped features and employed a fair approach (MLOOCV), while also further investigated interclass differences (imbalanced datasets) and a differentially methylated class. Furthermore, features were extended and grouped according to their biological function, in which sub-groups of four human chromosomes were studied in combination, along with their individual features, in order to interrogate features or feature subsets that are associated with DNA methylation based DNA sequence context.

A number of the methods solved very little in terms of analysis of epigenetic data and driven DNA sequence features. Some limitations in relation to DNA methylation classes predictions, and the analysis of the imbalance data were established through the use of direct machine learning. The features listed in Table 2.1 are the most important, and were calculated by various methods (referenced in Table 2.1). The features are driven from DNA sequence information, both directly and indirectly extracted from methylated, unmethylated and differentially-methylated samples. Furthermore, they are grouped into their biological associations and then listed into similar function groups.

In addition, methylation of the human genome is influenced by many factors, including age, gender and environment [15; 60]. A number of lifestyles will accelerate epigenetic deregulation (e.g., smoking, excessive use of alcohol, along with poor diet and stress), while the process is delayed by taking part in sport, healthy diet/lifestyles and physical fitness [10; 15]. This issue has led researchers to investigate the impact of age and gender on DNA methylation, particularly in CpG islands, which are mainly found to be free of methylation during normal physiological cell development. Here,

I have investigated the association between gender and DNA methylation, both in healthy and diseased samples, in order to gain increased knowledge concerning the pathophysiology of gender-related health outcomes. The designs of the experiment, along with the analytical details, are further reported in chapters 4 and 6.

Additionally, the author will discuss these issues in further detail during the chapters focussing on data analysis and experimental methodologies using machine learning. This will include the investigation and extension of existing methods to allow for the prediction of the imbalanced of methylated (sub-) classes or sub-set classes, and the extraction and selection of features related to methylation differences between male and female (both healthy and disease status), in addition to features associated with gender and age. Machine learning can be categorised into two main methods: (1) supervised (classification) and (2) unsupervised (clustering) algorithms. Both methods overlap in the manner in which they place objects into groups, but differ in the manner that groups are pre-defined for classification, whereas the classes are not pre-defined for cluster analysis.

The primary issue related to clustering concerns grouping a given collection of unlabelled patterns into meaningful clusters, and estimating the cluster structure, along with the number of clusters and cluster assignments. It has been assumed that clustering analysis has an unknown clustering structure, and that it is unique, with the aim being to identify a single partition or dendrogram. However, since the observation may cluster in more than one way, depending on the variable used, it is natural to allow for the existence of more than one clustering structure, and to identify multiple partitions or multiple dendrograms.

In the literature, two distinctive clustering approaches are in use, these being hierarchical and partition algorithms. Objects are placed into mutual groups, depending entirely on the protocols established prior to commencing the cluster [61]. However, classification consists of placing of objects into pre-defined groups. It contains two approaches: instance-based and rule-based classifiers. An instance classifier stores the training data, predicting the class of the stored data with respect to the nearest (distant) to the test data. However, a rule-based classifier attempts to generalise the rules of the test set. Moreover, in this section there is a review and brief discussion of a number of important machine-learning algorithms: this brief outline will investigate issues of classification design in relation to the imbalanced dataset, along with its possible solution.

2.2 Extracted DNA sequence features

This section investigates the most important biological feature sub-sets extracted from DNA sequence patterns. As noted in Table 2.1, there are 9 feature sub-sets extracted, which are described in more detail below:

- Sub-set 1 (tissue-specific CpGI DNA methylation) contains averaged sequence values calculated by using CpGcluster algorithms [62]. These are CGI-specific attributes (i.e. CG contents; CG%; number of CpGI; observed/expected ratio; CpGI distance; and CpGcluster p value).
- Sub-set 2 (DNA sequence properties and distribution) contains frequency average scores of all possible combination tetramers, and both specific and non-specific strands.
- Sub-set 3 (Dinucleotide-expected CG distribution) contains a score of 16 possible combinations of its observed/expected ratio.
- Sub-set 4 (CpG islands distribution) contains extracted attributes taking the distribution of CpG islands into consideration.
- Sub-set 5 (structural and physiochemical properties) contains predicted elements, including: rise, roll, tilt, twist and solvent-accessible surface area, as well as bending, curvature, stacking energy, turns, degree of twist, DNA cleavage, base per turn and six helical force constant. The calculations of these features were undertaken using DNAlive algorithms [62].
- Sub-set 6 (Exon and gene distribution) contains attributes extracted from the human genome, and high-confidence gene annotation from the consensus CCDS.
- Sub-set 7 (Evolutionary and conservation) contains attributes of phast conservation contents, calculated by the number of CpGI overlapping with elements of phastconcervation per CpCI, using a log-odds conservation score of 100 or more without repeat masking.
- Sub-set 8 (SNP) contain attributes based on the SNP features, and is calculated through the total number of SNPs in the window by counting number of SNPs from the UCSC genome browser.
- Sub-set 9 (Locus CpG islands methylation) contains a number of CpGs methylated values in the wide-human genome [63]. Further details can be found in chapters 3 and 6

The most important features are listed in Table 2.1, and have been calculated by various methods (referenced in Table 2.1). The features are driven from DNA sequence information, both directly and indirectly extracted from methylated, unmethylated and differentially-methylated samples. Furthermore, they are grouped for their biological association, followed by being listed in similar function groups. One of the most important biological sub-sets consists of tissue-specific CpG islands methylation features, in which the cell function is regulated and also makes the decision

Table 2.1 Details of the CpG islands feature-sets for Chromosomes feature-sets References

feature-sets	References
1. Tissue-specific CpGI methylation	[56; 64]
2. DNA sequence properties and distribution	[33; 48]
3. Dinucleotide (CG) distribution	[48]
4. CpGI distribution	[48]
5. DNA structure	[48]
6. Exon and gene distribution	[48]
7. Evolutionary and conservation	[48]
8. SNP	[33; 48]
9. Locus CpGs methylation(dinucleotides methylation)	[63]

whether a particular gene is switched on or off [15]. Furthermore, profiling 1.9 million CpG islands values from 43 samples of three human chromosomes were made as available resources [56]. It was reported that tissue-specific prediction [64] applied methylation and unmethylation from the resources by calculating DNA sequence properties [45; 48; 65; 66]. However, none of the papers have considered the manner in which their predictions influenced both inter-class differences, and also those of the differentially methylated class. Furthermore, different algorithms were employed for these features, depending on their biological function, and suitable statistical methods, including physio-chemical properties [67] calculated by the EMBOSS server developed specifically for the calculation of DNA and protein-associated features.

The other important algorithm to be developed is EMBOSS, which can calculate most biological properties (i.e., sequence information), and which is freely available to researchers. Phastcons are the most conserved features in DNA sequence distribution. Vertebrates have fewer than 2-3% phastcons in their exon, as compared to invertebrates [68]. The listed features are important biological features, and they become standard. However, the combination of these features was not undertaken with fair statistical methods, as CpG islands methylation may have not changed the backbone of DNA structure, hence, it is believed that DNA methylation is influenced by DNA-sequence context [69]. In order to understand de novo methylation, it is therefore necessary to investigate various features found in both single and comparative methods.

This formula is used for the calculation of DNA methylation, where β is the measurement of DNA methylation value of specific CpGs, on a scale of for 0 unmethylated and one for completely methylated forms

$$\beta = \frac{Max(M,0)}{Max(U,0)+Max(M,0)+100}$$

2.2.1 Tissue specificity

The somatic cells of individuals possess identical DNA sequence information (i.e. they are genetically-identical). However, the cells are shaped differently, based on their

biological functions. Therefore, despite the fact that the cells carry the same genetical information, their physiological function differs. These differences are linked by epigenetically-controlled DNA methylation with tissue-specific gene expression. As observed in the cellular phenotype, they originate from a stem-cell, through cell differentiation, without any change to the DNA sequence context [70]. Further details for tissue-specific CpG island methylation patterns can be found in the following references [56; 71]. It has been reported that cell-type specific methylation of blood cells has revealed methylation variation [72], and mono-allelic methylation differences have also been reported [33]. By contrast, tissue-specific genes have a low density of CpG islands [73]. Epigenetic modification has an impact on tissue-specific differentiation mechanisms, causing cancer and other illnesses related to methylation [74]. Moreover, tissue specificity and its gene expression demonstrates variation in individual samples [75]. The same study has reported that DNA methylation is influenced by tissue-specific factors, which are further dependent on the context of the position of the CpG islands. Although the references linked to these studies have improved our understanding of CpG island tissue-specific methylation, there are still important issues associated with DNA sequence features, DNA-context features and, more importantly, means of analysing these features. Therefore, extended experiments with consistent statistical methods are required to determine DNA methylation association features.

2.2.2 DNA sequence and distribution

Deoxyribonucleic acid (DNA) is the chemical compound which consists of in- formation that is required to develop and guide the activity of most living cells. DNA molecules are comprised of two twisting paired strands that are often referred to as a double helix. This contains four genetic alphabet chemical compounds, named nucleotide bases. These are Adenine (A), Thymine (T), Guanine (G) and Cyto- sine (C) [18].

The bases are complement-paired in two opposite strands, T paired with A, and G paired with C. The order of the bases determines the identity of information encoded in that region of the DNA molecule just as the order of letters determines the meaning of a word [18]. Sequencing involves determining the exact (meaning) order of the bases in a strand of DNA, by identifying one of the bases in the pair, and it then automatically determines the opposite strand (pair) [76]. Therefore, it is always reports just one base pair. Furthermore, assembling the sequences of all the bases in a large fragment of DNA, such as in genes, is necessary to read the sequence of overlapping segments. Large sequences can be assembled from short pieces just like putting together a puzzle which each of these segments read many times to ensure its accuracy. Knowing base pair resolution is necessary to construct maps of genome sequences. This allows annotating accurately all genomic features such as repeats,

SNPs, genes and their complete control elements. The human DNA sequence contains no gaps, and it has at least 99.9% correctness [76]. One such use is to seek for sequence features that increase a risk of specific diseases, and a type of genetic alteration that can frequently be found, for instance cancer, and also to understand methylation differences between healthy and disease features.

2.2.3 CG distribution

DNA contains four bases; CG units are one of the 16 possible combinations of nucleotides. CGs are mostly methylated, and they termed as Cytosine (C) methylation [18]. This methylation was adopted in vertebrates primarily as a means of maintaining DNA in a transcriptionally-inactive state. when C altered into T, this mutation can be transmitted to the next generation but only when they occur in the germ line, the cell line that gives rise to sperm or egg [15]. CG mutation is characteristic of all genomes that Cytosine is methylated and it is evidenced throughout the human genome, except in small clusters that are known as CpG islands, which remaine methylation-free [15; 19]. As shown in Figure 2.1, the number of CG-units can be counted for any given sequence that can be further analysed, as either methylated or unmethylated [77].

Figure 2.1 CpG distribution(Chromosome 21 (NT-002836.4 740746-742525)) sites are highlighted in yellow colour in the DNA sequence. These data can determine whether it is methylated or unmethylated.

2.2.4 CpG islands distribution

CpG islands seem to remain unmethylated in all cell types [15; 19], and they are found to surround the promoters of the so-called housekeeping genes are those that code for the many proteins, and which are essential for cell viability and therefore expressed in most cells [15; 18]. Moreover, many tissue-specific genes, which code for proteins, are only required in selective types of cells which are also found to be associated with CpG islands [15]. In the human genome, CpG contains less than 20% of the expected frequency [76]. However, CpG islands densities are significant higher than that of non-islands DNA [19]. About 60% of CpG islands are associated with genes, and approximately 58% of these human coding genes have CpG islands as their promoter; for this association, CpG islands can be used as potential gene markers [9; 12]. In addition, CpG island densities vary substantially across chromosomes, although their correlations with genes are reasonably well estimated on relative chromosomal gene densities [35; 76].

2.2.5 Sequence features/structure

DNA structure is very important for normal physiological cellular processes. Indeed, DNA exists in many possible conformations [78]. However, only three forms of DNA double-helix have been observed in organisms, i.e, A-DNA, B-DNA and Z-DNA forms, and their conformation depends on the sequence of the DNA, the amount and the direction of supercoiling, chemical modification (i.e, methylation) of the bases and also the solution conditions, such as concentration of metal ions and polyamines present of [79] these three forms, B-DNA is the most common form that is found in cells under physiological condition [80]. The other two double-helical forms of DNAs varies with regard to their geometry and dimensions. Z-DNA can be recognised by Z-binding proteins that are involved in the regulation of transcription [78]. A-DNA is found under non-physiological conditions, for example untwisted DNA and protein binding complexes such as that involved in RNA hybrid pairing processes [81]. The most important key question is therefore how DNA methylation influences DNA structure; this issue is briefly outlined in the next sub-section.

2.2.5.1 Bending flexibility, stiffness and untwisting

This sub-section briefly explores into how sequence order influences DNA structure. Sequence orders of AAA/TTT have rigid conformation, and also have a restricted range of roll and slide values whereas CA, CG, TA and TG have the weakest configuration [82; 83]. These show the best initiation sites of a double helix. In addition, TA step (B-form DNA) strands are the most flexible sequence features with respect to decreasing twist and increasing roll, but this structure property is highly context-dependent [79]. Analysis of human exon shows a preference for in-phase occurrence

of the three nucleotides CAG/TCG, with a weaker preference for AAG, GAG, ATG and GTG [84]. These sequences are flexible in the sense that they can take on a conformational feature of an out ward-facing minor groove on nucleosomal DNA [85]. Notwithstanding, the similar three nucleotides analysis did not show any preference occurrence in the opposite face matching to sequence, such as AAA/TTT. In contrast to these results, yeast genomic sequence analysis have shown strong signal for the dinucleotide AA/TT, which produced a peak at periodicity of circa 10.2bp [86] (these peaks are only found in eukaryotes rather than prokaryotes [86]). Most large DNA fragments have preference to B-DNA, which is less flexible [82]. Gardiner [82] reported that GGC and GCC sequences have more preference to confer bi-stability and they have a low stability and of favour the A-form of DNA. In contrast, the AA sequence steps are strongly in favour of B-form and restrains the A-structure.

A computational method has been developed for predicting the 3D structure of double helical conformation based on six bases step parameters that are named twist, roll, tilt, rise, slide and shift [82]. Consequently, all possible combinations of the structural properties of DNA oligomers were analysed as the length of the sequence increased from dinucleotide base-pairs to eight base-pairs. This concluded that the length of sequence increased, and the variation of conformational preference decreased, and also that the structure became less flexible and more consistent [82]. It is likely that DNA-methylation is linked to DNA structure, as this may prevent some conformational changes (without alteration of the DNA sequence), which indicates whether a specific gene is active or not.

In addition, DNA stability depends very much on base-pair composition for example, G-Cs are more stable than A-Ts, and this depends on the binding energy, since A-T requires less energy to dissociate. It also depends on the particular geometry of the relevant base order and its sequence context[79]. Indeed, it has been reported that DNA methylation give rise to an unusual tertiary DNA structure which protects the protein binding site from DNA-polymerase [48].

2.2.6 Exon and Genes/Genome

A genome is a complete set of DNAs organisms [9]. Normal human cells contain 23 chromosome pairs located in the nucleus [18]. The chromosomes contain three billion base pairs of DNA (letters), which carries genetic information essential to build the human body. The functional sections of the chromosomes are known as genes.

In the human genome, 20,000 to 25,000 genes are estimated, and each gene encodes an average of three proteins [76]. Genes have a complex structure which comprise two joined segments Exon and Intron, which are coding and non-coding segments respectively. Intron is removed from Exon by a ribonucleoprotein complex, splicesome [18]. Splicesome recognises sites at the 5 and 3 ends of introns then removes the introns from exons site. The retained segments, the exon from messenger RNA, are called

genes. As the splicesome recognises and cuts off different splice sites of introns, or the RNA sequence of transcribed gene, it generates isoform exoms that can encode for protein variants. For example, an enzyme recognise and copies the information in a genes DNA into the molecule Messenger ribonucleic acid (mRNA) [18].

The mRNA are transferred into a part of the cell, the cytoplasm, where mR- NAs are processed into amino acids, and then link them together in the correct order to form a specific protein. Proteins build up tissues and organ structures, as well as recycling control and chemical reactions, and give signals between the cells physiologically. However, if a cells DNA is influenced by epigenetic (environment) or genetic (DNA-mutation) events, it may produce abnormal proteins that can interrupt the bodys natural physiological processes, which lead to diseases such as heart disease, cancer and mental problems [10; 14; 16].

2.2.7 Evolution and conservation

This sub-section briefly provide an explanation of genome integrity, which is of much importance in normal life processes in biology. Some part of genome sequences alter more easily than others during evolution, specifically involving the non-coding DNA sequence, and it is more likely to change at a rate restricted by the frequency of random errors [18]. However, the conserved region that codes an important molecules such as ribosomal ribonucleic acid (rRNA) do not modify easily when it is mutated; errors are mostly repaired or removed [15]. Therefore, through evolution, many features of DNA patterns have been changed beyond all recognition, but the conserved regions of DNA sequences remain exactly recognisable in all living cells. It is also reported that the novo methylation of CpG islands are more likely to be stable during evolution [48]. These conserved regions (sequence features) are the ones to examine regarding the tracing of family relationships, hereditary disease histories, and studies the distance between organisms in the tree-of-life. More importantly, DNA methylation evolutionary and conserved region is an essential genome study. Indeed, it is most important to study phascon genomic regions in relation to other environmental impacts.

2.2.8 Single Nucleotide Polymorphisms (SNPs)

Here, Single Nucleotide Polymorphisms are briefly reviewed; indeed, these are the most important feature sub-set in molecular biology, which variant at a single base position between populations which is frequently found in different geographical or ethnic groups, and are also the most investigated DNA features. However, in relation to DNA methylation, these features are not studied.

SNPs are single base-pair substitutions in genomics at which different alleles exist with a frequency of at least 1% in one or more population [76**?**]. About ninety

percentage of sequence variants in a human are SNPs which occurs 100 to 300 bases along the three-billion size human genome [76]. For example, DNA sequences may change an Adenine (A) base to a Thymine (T) one ('AAGGCTAA to A**T**GGCTAA'). Normally, one base change does not affect the cell function. However, SNPs can occur in a protein-coding region, and it is believed that predisposes humans to disease and influence drug targets [87]. Furthermore, SNPs are influenced by methylation in a similar way, as noted in the DNA methylation section. SNPs are evolutionary stable from generation to generation [18], which makes them easier to study. However, less investigation has been done on DNA methylation with SNPs. Since SNPs are specific to individuals, it is mostly used in DNA fingerprint for forensic purposes, disease risks and treatment response assessments to specific populations or ethnic groups.

2.2.9 Locus CpGs methylation

Most genes have a high concentration of CpGs in their promoter (start site of protein initiation region). This means DNA methylation influences the promoter more than any other region of wide-genome; however, gene expression is dictated by DNA methylation whether a particular gene is active or inactive. DNA methylation is a trafficking and mechanism that regulates protein production, and which inform us on whether a particular gene is on or off. Imbalance or unfair trafficking can cause inappropriate body function such as immune system disturbances, speeding ageing, mental health problems (depression) and other serious diseases such as cancer, multiple sclerosis, Alzheimers disease and autisms.

In addition, genomic imprinting requires DNA methylation. Vertebrate cells are diploid, containing one set of genes inherited from the father and one set from the mother [18]. In a few cases, the expression of a gene has been found to depend on whether it is inherited from the father or the mother. This phenomenon is called genomic imprinting. Although the mechanism of imprinting is unknown, it has been experimentally show that DNA methylation is involved [10; 19]. In the human genome, more than 80% of CpGs are methylated, whereas of the remaining, less than a quarter is unmethylated and is also very unevenly distributed in the genome [19]. They are present at 10 to 20 times their average density in selected regions, known asCG islands, which are 300 to 3000 nucleotide pairs long [18] as discussed above.

However, CpG islands are prevented from being readily methylated as noted above, although these areas are rich in the CpG dinucleotide. This is vital for promoter-associated CpG islands and its transcriptional machinery except for two important the novo methylation; (1) genes on the inactive X-chromosomes and (2) imprint genes [15]. These two situations are both methylated in the promoter at all sites, and this leads to the transcriptional repression of these genes [11; 15]. CpG island methylation is a natural selective process that mediates epigenetic inheritance [15; 18]. However, this is not always the case; some of the promoter methylation is linked to cancer

and ageing and also other syndromes [13] as noted above. Furthermore, leukemia and other hematological malignancies are found with most of the methylation in the promoter; the most common one is the acute myeloid leukemia, whereas non-malignant hematological diseases involved unmethylation [88]. Genes involved in cancer show hypermethylation in CpG islands promoter [12; 14].

Methylation on the human genome are influenced by many factors such as age, gender and environment [15; 60]. There are many lifestyles, including smoking, excessive alcohol use, diet and stress, which will accelerate epigenetic deregulation, whilst sports, healthy diet/lifestyle and physical fitness delays the same process [10; 15]. This problem has led researchers to investigate the impact of age and gender on DNA methylation, particularly in CpG islands, and this is mostly found to be free of methylation in normal physiological cell development. It is important to have an insight into the association between gender and DNA methylation both in healthy and diseased samples, in order to learn more about the pathophysiology of gender-related health outcomes.

Epigenetics is a dynamic process, and it is not fixed as previous thought, for example, The essence of DNA methylation is mostly reversible process and it is a continuous lifelong process [89; 90]. Even identical twins show DNA methylation differences with age [91]. However, gene expression is not regulated clearly from either genetic or environmental impact since they are both linked through the epigenome [92] which could also be possibly linked with gender. This dynamic status makes it very challenging in order to design, predict and model whether these epigene changes can determine both healthy and diseased state data, and also distinguish between genders with age range as a further consideration, and to link these DNA methylation differences to healthy and cancerous individuals. DNA methylation is associated with complex age-related diseases such as Alzheimers disease and cancer [10; 12; 15; 30]. Whilst DNA methylation changes are associated with age, and this has been studied to some degree, there is a shortage of reported studies on the relationship of gender and DNA methylation. Further information on this is available in Chapters 4 and 6.

2.3 Analysis of DNA methylation using bioinformatics Methods

2.3.1 Machine learning techniques

This sub-Chapter explains technical issues associated with DNA sequence analysis and corresponding research solutions. The techniques for analysing and extracting/selecting features from datasets can be grouped into two broad and overlapping categories: (1) clustering and (2)classification methods. Both methods place objects into classes, but the important difference is that the classes are not pre-defined in

cluster analysis. As an alternative, items are placed into mutual groups which entirely depend on the protocols that were established before starting the clustering process [61]. However, classification is the placing of objects into pre-defined groups. The two methods are also different with regards to the types of learning methods involved; clustering algorithms are unsupervised learning techniques, whereas classification methods are employed in supervised learning. Feature extraction computes new features from original feature-sets, whilst feature selection methods identify a sub-set of the available features for subsequent use [4].

2.3.1.1 Unsupervised learning

This sub-section will discuss unsupervised learning approaches, and further examines a range of applications of clustering methods demonstrating its potential for biological research usage. One of the approaches of these regularities is to class the similar objects into a set of groups. Grouping samples or objects according to their similarities is known as clustering. Clustering has been used excessively for biological data because it is beneficial towards the analysis of multi-dimensional data, which leads to a more descriptive and meaningful solutions. For example a good clustering process has predictive power. It is also one of best sub-feature selection methods available in bioinformatics.

2.3.1.2 Clustering

Clustering can be defined as an organisation or collection of patterns, usually represented as a vector of measurement, or a point in a multidimensional space based on similarity. The problem in clustering is to group a given collection of unlabelled patterns into meaningful clusters, and the aim of cluster analysis is to estimate this clustering structure, the number of clusters and cluster assignments. It is assumed in clustering analysis that clustering structure is unknown and unique, and the aim is to find a single partition or dendrogram. However, since the observation may cluster in more than one manner (depending on the variables used), it is natural to allow for the existence of more than one clustering structure and to find multiple partitions or multiple dendrogram. In the literature, two distinctive clustering approaches are often employed, specifically hierarchical and partition algorithms [4].

Clustering (unsupervised learning) has no class label present in the training features, so the classifier is left 'alone' to find its group. This leads to the discovery of similarities and differences amongst features which depend on the protocols that were established before commencing the clustering process [61]. Furthermore, there is no concept of accuracy for unsupervised methods, and the only means of evaluating the outcome is by its usefulness. For example, numerous attempts must be made to establish effective clustering methods which are based on fitting a specific

purpose or protocol. The advantage of these methods is the gain of information from knowledge-poor environments; particularly, when there is a large amount of unlabeled data available. Clustering can be used as a type of exploratory data analysis which may lead to evidence regarding the underlying structure in the data.

Clustering has been previously applied successfully to the analyzing of microarray data [93; 94; 95]. For example, gene expression profiles can be utilised to identify sub-features of co-expression genes [96]. However, it can also be applied to CpG island prediction [97], i.e, to identify CpG island clusters in a length sequence. In clustering, two metrics are predominantly employed to group objects; Euclidean dissimilarities [4] and correlation [98]. Both metrics are mostly employed for biological data, and recently a novel generation of algorithms have emerged, and which are used with one or more distance metrics. Euclidean distance K-means is one of the most used feature selection algorithms, and is based on squared Euclidean dissimilarities [4]. Some variants of K-means have been proposed in order to improve the efficiencies of algorithms, and to find the global optimum [99; 100; 101]. Furthermore, multiple clustering observations based on weighted distances (with weights determined by the cluster of variables) have been developed [102]. However, K-means remains one of the most successful algorithms used on high-dimensional datasets with filtering techniques, since it is easy to implement and has less computational complexity. It is stable and works well in practice, but Heatmap clustering is used mostly for the visualisation of biological features, particularly for microarray and Ilumina-array data, and also to extract meaningful features. This is further discussed in the literature review provided in Chapter 3 with further experimental details for feature extraction in Chapter 4.

2.3.1.3 Supervised Learning

Supervised learning applies to classifier development, which refers to the gradual reduction in error as training cases are presented to the classifier. An iterative process is involved with error diminishing through each iteration as the classifier learns from the training cases. Supervised learning also applies to algorithms in which a teacher continually evaluates the classifiers performance during training by making predictions either correct or incorrect. Unlike clustering, classification is the placing of objects into pre-defined groups. This assumes that each case has a valid class label. Additionally, a set of parameter values are adjusted to improve the classifiers performance. Subsequent sub-sections will further discussions regarding supervised machine learning.

2.3.1.4 models used for prediction

DNA methylation analysis is used for various predictive models. The three top predictive models are support vector machines, K-nearest Neighbour and Discriminant

Analysis; these classification methods are used to distinguish two or more DNA methylation classes. Classification is the grouping of objects into target classes. For example, DNA features are classified into methylated and unmethyalted classes in given samples. Samples are divided into training and test sets. The training set is used to build a predictive model, whereas the test set is used for evaluation, the predictive-model is judged by the misclassification error rate of the test set: the lower the error rate, the better the model. The most important classifiers that are used in DNA methylation prediction are the following:

- Support Vector Machines [103]
- K-nearest Neighbour [104]
- Decision Tree [105]
- AdaBoost [106]
- Linear (QDA) Discriminant Analysis [107]

Support Vector Machines is most popular in DNA methyaltion prediction [46; 47; 48; 66; 108; 109], K-nearest neighbour [5; 26; 110] and Decision Tree [45]. The K- nearest neighbor (KNN) classifier is one of the most popular non-parametric classifiers, and has been successfully applied to various problems in bioinformatics [52; 111]. It assigns to the point that the majority label amongst its nearest k in the training data point to x, and predicts the class-label of x based on the labels of those nearest k points. Increasing the k value has been shown to reduce the bias and decision boundaries becomes rather smooth and less sensitive to outliers [52]. It has been reported in some studies that applications of KNN resulted in higher predictive accuracies than those of Support Vector Machines, and is one of the most powerful methods [111]. It has also been reported that KNN is superior in comparative studies of seven classifiers [112]. However, it should be noted that in many cases, the success of a predictive method is mainly based on a characteristic of the dataset being analysed.

The K-nearest neighbour method finds every feature in the feature-set closest to its nearest neighbours. However, if there are more irrelevant features, the performance of the KNN is affected severely [113]. Blum and Largely [113] reported that samples without noisy or less irrelevant features show an increase in their performance via sub-optimal feature selection methods such as filtering, wrapper and embedded approaches. The points illustrated in Figure 2.2 donate to either of two classes where $\omega 1$ belongs to the black dots, and $\omega 2$ to the white dots. The red dot (p) is assigned to be classified into one of the two classes. This Figure shows that K=11 nearest neighbours from this class lie within a small area compared to the eleven neighbours from the other class. The circle shows the area surface within which the eleven nearest neighbours are positioned. Therefore, a given unknown feature with vectors $\omega 1$ and

distance nearest out of N training vectors identifies the KNN, regardless of class label, which is based on various distance measures, including the Euclidean and mahalanobis distance. Hence, the feature attribute p is assigned to the class of nearest neighbour.

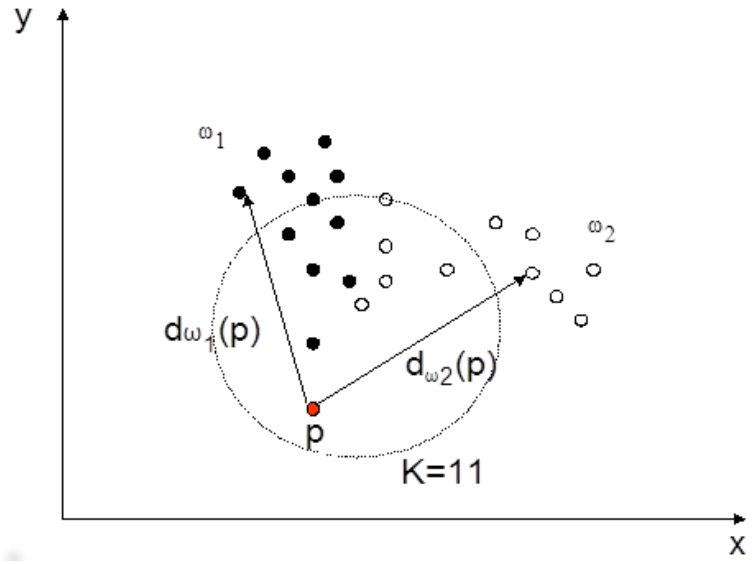


Figure 2.2 Model of points of two class labels, i.e. black and white dots. The red dot, p, is a new observation, which is to be classified into its nearest neighbour amongst these two classes [3].

Discriminant analysis is either linear or quadratic; linear discriminant analysis (LDA), is also termed Fisher discriminant analysis (FDA). It is the simplest classification method used for two-class prediction problems. LDA is the first choice for data analysis in view of its interpretability. However, LDA works only with data which has linear relationship. Indeed, quadratic discriminant analysis (QDA) is more favourable; QDA is a variant of LDA. Disadvantages of QDA is that it cannot be used for small samples with high dimensional data. It is also computationally expensive during data training, but the advantage of QDA is that it gives the best error estimation. Researchers try to reduce the parameters and shorten the model building. In general, LDA and QDA classifiers show good error estimation performance in different applications, when compared to other classifiers [114; 115].

The Boosting algorithm is used to improve the performance of weak classifier. It is widely applied for small samples and imbalanced data, and performs better than random guessing in view of a weak-classifier ensemble which becomes strong and improves error estimation [106; 116; 117]. Therefore, the weak classifier is iteratively repeated, and each cycle of it uses a different sub-set of training data. Each cycle of the training set or sample weighting scale are computed. The final cycle of the training set gives the weighted average outcome from all previous iterated training sub-sets [116], which shows improved predictive performance. AdaBoost is based on the modification

of the training set to build a strong classifier by iterating a weak classifier, a process which can be used to improve the predictive performance of small and imbalanced datasets [106; 118]. The advantage of AdaBoosting is that it has a minimum level of overfitting, it has less influence for large dimensions, and numbers of folds, and shows better predictive performance compared with other classifiers [3]. In addition, it has been experimentally reported that the error rate of the test decreases after the training set error becomes zero [116]. Other researchers reported that AdaBoosting have less overfitting, since parameters are determined in a step-wise fashion, where each iteration of a single parameter is computed. More details of method can be found in references [116; 118; 119]. However, the impact of the class differences and sample size has not been thoroughly investigated. These are further experimentally investigated in Chapter 5.

2.3.1.5 Feature Selection

The aim of this section is to explore methodology that relates to the feature selection of variables such as a multi-dimension and imbalanced data, and noisy features. It has been noted that major problems associated with large features, for instance DNA sequence distributions and patterns, has previously failed to pre- dict methylated and unmethylated features by using direct machine learning techniques without associated feature pre-processing [5].

Generally, researchers seek to design machine learning approach that can recognise features, speech recognition, fingerprint and iris identification, face recognition, reading text, DNA-sequence identification, and much more. It is therefore clear that reliable accurate feature selection via machine learning would be immeasurably useful to construct such systems in order to understand and select the most informative patterns of DNA sequence features from healthy and diseased human tissues. Building such methods are aimed to improve classification performance, and also for overcoming computational limitations.

Statistical feature selection has been used successfully in many applications, a number of feature selection methodologies have been developed [3; 4; 48; 111; 113]. For example, statistical pattern recognition, that represents a set of D features, is viewed as L-dimensional feature vector. The concept of decision theory is used to establish decision boundaries between two feature classes [3; 4]. Feature selection methods are operated in two approaches: training (learning) and testing (classification) (Figure 2.3) [4]. The role of the preprocessing stage is to select the informative features, and simultaneously deselect the noisy features by adding to any other operations that may contribute to defining a compact representation of the selected features. In the training method, the training data finds the appropriate features for representing the input features, and the classifier is trained to partition the feature space [4]. The feedback pathway shows a designer means to optimise the prerocessing and feature selection

strategies. The classifier design model allows the selected and trained features after optimisation in order consider and assign samples to one of the feature classes based on the measured features. It is also important to note that the classification (test set) does not play a role feature selection process in order for an unbiased estimate to be obtained [3; 4; 120].

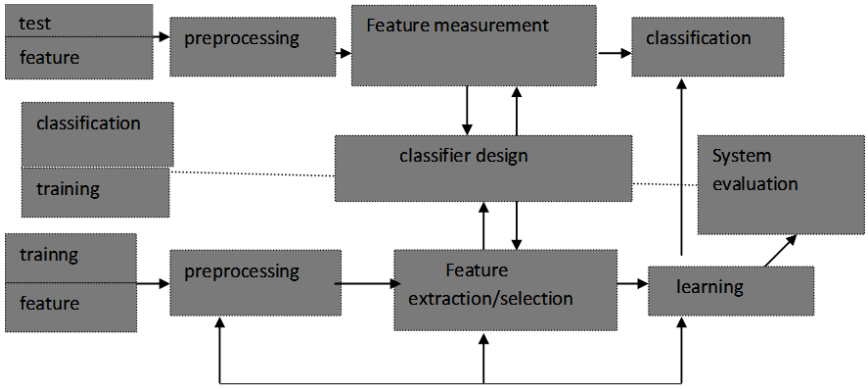


Figure 2.3 Feature selection statistical model; adapted from [4].

The classifiers role is to divide the feature space into regions that correspond to either class 1 or class 2 (as illustrated in Figure 2.2). For example, feature x, is an unknown feature, goes into the class 1 region, or otherwise into class 2, i.e, it is classified into one of the classes. However, this does not necessarily indicate that the decision is accurate; it can also suggest that a misclassification has occurred, for example, 1 feature may be incorrectly assigned to class 2 and vice-versa.

Generally, the feature selection method, classifier design, and classification error stages are combined to compute the best combination [121]. It is, however, important to choose the specific case study, in order to demonstrate that the various stages in the design of a classification system are not independent, but they can be closely interdependent. However, this may not be possible for a large dimension of features, and therefore the feature selection stage cannot be easily integrated with that of classifier design techniques such as those SVM, KNN and QDA, and thus must be included via feature sub-set selection methods. Ideally, it should be aimed to have a procedure to design the classifiers by minimising the error probability directly, and synchronously the method should be computationally-simple to allow also for a search for the optimal combination of features. The next section will focus on tackling the above mentioned multidimensional data problems.

2.3.1.6 Feature subset selection

Feature sub-set selection facilitates our understanding of the associations of particular patterns or sub-sets of features for large datasets, and it is necessary to pre-select

a small and informative number of features before applying machine-learning techniques [3]. This is known as filtering, and ranking these assignments to a feature of a numerical weight that is used to rank all features and then select the top ranked ones. In the filter model, the optimality rule for feature selection is independent of the classifier [96]. This can be used in the classifier stage (which is created for various feature combinations) and then select the best feature vector combination. This selection method is based on a feature separateness criterion introduced by used algorithms. However, in the wrapper filter approach, the selection method is not necessarily based on the values of an adopted class separateness criterion, but rather on the performance of the classifier itself [3; 120]. Other filters are reported in literature in order to produce better initial feature sub-sets; Kulback Lieber and Chi-square filters are two other common ones available [3].

The Kulback-Lieber filter uses the distance between histograms of feature values in order to compute the class separability of each feature, and chooses the sub-features with probability proportional to the Kulback-Lieber distance [3]. Subsequently, an applied thresholds is used to eliminate the noisy features, and then, it normalises each feature to probabilities of between 0.1 to 1, which ensures that no features will be completely ignored, and no features will be definitely accepted in the initialization process. The Chi-squared statistics is used for features or data in order to determine whether the tested data is statistically independent from one feature datapoint [3]. Both techniques are usually used for selective preprocessing in order to improve predictive accuracies, and simultaneously reduce noisy features that degrade class performance. More importantly, feature selection reductions in computation time are possible, for example, the wrapper method is the most commonly used for feature selection. In the wrapper approach, each feature is combined until all features are met, and then the classification error probability of the classifier is estimated [3]. The wrapper method selects the best sub-set feature, i.e,the one that shows minimum error probability. Further details can be obtained in Chapter 3 and 6.

2.3.2 Imbalanced inter-classes differences

Imbalance donates unequal numbers of observations within two or more classes; for example class 1 has more observations than the other. This situation is presented in terms of their ratio, for example that of methylated and unmethylated imbalance analysis are a common problem in the research community, where the ratio of some samples varies substantially.

2.3.2.1 Sample size and overlap of class prediction problems

In statistics, small samples are mostly imperfect for analysis and such samples are unreliable [26]; as the sample size decreases, the misclassification rate of imbalanced

data increases. Further, our experimental study (Chapter 5) indicates that cost sensitive weightings have less impact on medium and large samples, compared to a small sample size, and also showed less improved predictive performance.

Small samples cause the most difficult classification problems since they markedly overlap; in particular, the minority class cannot be separated from the majority one. However, when class separateness is not the issue, their prediction does not require a complicated form of modelling or design. However, classes overlap in different levels and within class it is hard to detect by direct machine-learning; such class-overlapped data have been previously studied [122; 123]. This work reported that class distributions were not only the problem, but when the class overlaps significantly, the minority class showed an increase of misclassification error rate [124]. It was also reported that overlapped classes failed to be distinguished by linear discriminant analysis, or were less sensitive to such linear analysis [122].

2.3.3 Research solution methodologies

In the literature, a number of imbalanced data problem have been reported with varieties of samples, in which some had severely imbalanced classes present. This research has developed/designed methods based on two approaches: (I) Data level and (II) Algorithms level. In the data approach, samples are re-balanced by dividing the majority sample into sub-sets which have equivalent minority class, i.e, each sub-set of the majority sample is combined with the minority class and applied into algorithm where outcome of the algorithms are averaged. The algorithm approach is to manipulate the classifier during training by weighting the minority class. Both methods attempted to reduce the classification error rate of the minority class.

2.3.3.1 Resampling approaches

Resampling strategies at the data level were employed in different techniques such as randomly over-sampling [117] the minority class and under-sampling [125] the majority class, as well as for combination of these two approaches [124] .These two techniques are used mostly imbalanced data. However, the question is how to decide what is the best for a particular dataset, and what factors are important to biological data, particularly when there is no control of the sample size. Hence, researchers are required to design a model continually in order to adapt to each individual case. Experimental studies have shown that imbalanced data depends on sample lengths, which found that sample size greater than 60 have less influence on imbalance, whereas more than 30 $\geq$ medium sample size have shown problems without re-balancing, whilst samples size less than 30 $\geq$ showed the worst performance. This study (Chapter5) was concluded with respect to the predictive performance on the skewed classes, depending on the severity of the imbalance and size of the sample. Small sample investigations are

required to be designed differently from those involving medium and large sample size.

Furthermore, class distribution and optimal resampling are other problems encountered, i.e, what is the best resampling procedure without losing any information, as well reliable predictive performance. Random sampling is the simplest method, but it does not best the suit some datasets. For example under-sampling without losing any information of majority class is ideal, since it also decreases computational time. Over-sampling methods were used to analyse imbalanced data and widely applied the minority class. Notwithstanding, over-sampling of the minority class was also reported as unreliable, and does not completely represent real data since it is added to synthetic generated data [126]. Moreover, it increases the computational time [127] , and overfits since the minority class repeated many times(many folds of synthetic dataset) [128]. However, under-sampling is not always the best particularly with small samples which has shown less performance and mostly overfits [129]. These small imbalanced data cannot be solved by under-sampling methods only. These problems are farther discussed with a number of statistical methods which can be used; however, each case or sample may only be suitable for a specific algorithm or methods.

2.3.3.2 Algorithmic approach

Imbalanced classes are a common problem encountered in data analysis, and the choice of algorithm employed. Modified leave-one-out incorporated KNN is used for imbalanced data [5], and probabilistic estimation is also reported in decision tree-based tree leaf methods [130; 131] although different penalties have been used for different classes in other analytical approaches [132]. It is important to understand why learning methods fail to predict imbalanced classes, since this requires knowledge of both learning algorithms and application domains in order to effective develop learning algorithm for imbalanced data analysis.

Two class classification problems are assessed in terms of their similarities or dissimilarities, and their target classes, where either numerical values or strings are introduced as target classes. Most classifiers work by only introducing two classes in the training dataset. Gentle AdaBoost, AdaBoost and KNN algorithms are co-operated with both methods in order to reduce the misclassification error rate of the imbalanced data, and both methods are further discussed in section 3.5 in detail.

2.3.3.3 Weighting and cost-sensitive approaches

In a normal data analysis application, data are analysed in a symmetrical way, and some datasets have one class which has more observations than the other, or one results in higher misclassification rate than the other. In such a case, two methods are in use in the literature. The misclassification error rate of the minority class depends on

inter-class differences or data separateability. Both problems use the weighting (cost matrix) as a composite for the outcome of imbalanced data. Where more attention has been paid to the minority class, high weightings is assigned to the minority class if the misclassification error rate increases. Suppose that $\beta(i, j)$ devotes the cost of predicting an instance from class i as class j. Let $\beta(i)$is the cost of misclassifying a methylated (i) and unmethylated (j) as a majority class, for example, then methylated (the minority) has less chance to be recognised than the unmethylated one since it has more observations than the former. A high cost is assigned to the minority class which adjusts the misclassification error and is continually updated as it is produced during a minimum misclassification. This update is mediated during data training in order to build models with the lowest misclassification error rate. When the model is left with an independent test rate, there are three main categories of cost-sensitive approaches reported in the literature [133]: (I) Adapting the classier to such cost-sensitive strategies applied in decision trees [134]; (II) Assigning each class to the lowest expected cost by using Bayes risk theory [131] (although this method requires membership probability outputs and cost estimation, but the cost is unknown during classification time); (III) modifying the training data distribution before being applied to machine-learning then cost-sensitive information is extracted from the learning method (this was termed as cost-sensitive learning by example weighting, and was reported to display a better performance compared to the other two methods; it is also easier to interpret as it does not require probability estimation from the classifier [133].

Prior probability and cost-sensitive approaches are combined during the training ensemble data space where the previous set $\beta(x_1, y_1)$; $x_1 \neq y_1$ with uniform probability of cost matrix [0 1;0 1].

- Fitensemle is a probability weighting that strengthens the individual learner and which converts the weighted mean square error approach to 1. Prior probability tried to adapt specific classifier learning algorithm during training, adapting the sample distribution into sample weights; training data sub-sets are weighted in each cycle of the weak learner. For example, the methylated DNA-class has less observations than the unmethylated class. However, they are mixed in different proportions; therefore, a prior probability approach was set up with values that represent fairly for both classes, and then the classifier is normalized, and hence prior probabilities of both the classes add up to 1 with a changing distributional outcome. This method is applied to standard algorithm where only the dataset is modified but not the learning algorithm.

- Modified data distribution approach by weighting data space; Fitensemble cost matrix data approach; training data are modified in respect of their misclassification rate. The majority class is fit to a data space, for example X, Y, β are also added into the cost matrix; x represents a training set of Y output or

target label, and β is the misclassification cost from training of the first step of the weak learned, i.e, training observations are changed with respect to the misclassification error rate. This method was reported in the translation theorem [125]. Therefore, the distribution of data space has been modified into a normal data distribution so that the misclassification rate of the training set is reduced and further updated. And the weighting cost with respect to its misclassification error rate. Our aim is to choose the hypothesis in order to reduce the misclassification error rate of the minority class. Disadvantages: the cost matrix is not available for some of the dataset, whereas the cost-sensitivity of the sample is estimated from training data. However, when both classes are equally represented, but are hardly separable, this can be applied to the cost matrix weighting in order to reduce the error misclassification rate.

2.3.4 Classifier assessment and evaluation metrics

In this section, I examine predictive models and their hypotheses in order to gain a minimum error for the designed model. For such methods, however, it is difficult to distinguish an optimal number of used parameters, since the number of attributes increases, and concomitantly find the best predictive model of new data (unseen data) and the generalisation performance. Classifier performance is assessed by cross-validation. Accuracy is the most traditionally used for assessing the performance of the classifier; however , imbalanced data classifications are no longer accepted-only the accuracy measurement of the minority class has less effect compared to the majority class [5; 122]. For example, total accuracy can be 98%, whereas that for the methylated (minority class) training set represents less than 5%; this accuracy is meaningless since it does not represent both classes equally. Two-class prediction samples can be driven for measurement processes which can be provided as a confusion matrix.

2.3.4.1 Cross-validation

Cross validation is a process required to assess model performance, for example how well the classifier behaves when unseen data is used. Unseen data is termed the test set which is left during the training of data, i.e,the unseen dataset (test set) is used in order to test the performance of the classifier (model). This process is known as cross-validation. There are four main cross-validation methods in use in the literature; (I) substitution, (II) Holdout, (III) M-fold and (VI) leave-one-out. Substitution [3; 135] is a process in which the same data is used for both training and testing of the predictive model; despite substitution being reported as a biased method, there are many researchers who is still using it. For example, as the sample to dimension ratio increases, the less bias (less overfitting) observed [3] but this procedure is not valid for small samples, large dimension to sample ratio and samples within imbalanced

classes [5; 136].

The simplest cross-validation is the holdout method. Data is divided into two parts: one is used for training, whilst the others for model testing so that the training process is the function of the test set where the output is evaluation of the misclassification error rate for the unseen data. The advantages of holdout is computational less expensive. However, holdout produces great variance, and it depends on the way and how the data is partitioned during training, since data are randomly divided, which give rise to differing results. In addition, when dealing with error, it is important to take account of the test-data variation and classification error caused by randomness, the selected size of test set, and how the error rate changes by modifying the test set size. Although randomness of internal partition of training data can be dealt with by running it many times, the used algorithm and averaging the results (accuracy with estimated deviation).

To overcome this disadvantage, k-fold cross validation is employed which improves the holdout method. The data is divided into k-subsets and rerun k-times, each fold of training one part of the k-sub-set is left and the remainder is trainined. The average of the error rate from training is computed, and it does not matter how the data is partitioned since each k-subset will be exactly in the test and the training sets k-1 times. As K increases, the variance of estimated values is reduced. However, it has also has disadvantages, i.e, data need to be re-run k-times, which is more expensive than the holdout method. As data divides randomly k- times, the variance will be increased despite its advantages, since data can be partitioned independently, and the researcher is free to choose how large is the training set as well as the test set.

Furthermore, leave-one-out cross validation (LOOCV) is an extended version of k-fold cross validation, where k = n. in this form of CV, data are assigned to the training set except one data point which left as a test set. The advantages of leave-one-out cross-validation is less bias. However, its variance is indirectly proportional to the bias, and therefore an increased number of folds (k) give rise to extended computational times. Other advantages of leave-one-out cross-validation is that it is mostly used on small samples and imbalance data (more details will be discussed in chapter 5). Leave-one-out cross validation shows superior model evaluation [136]. However, LOOCV has also limitations regarding its applications to small sample sizes and imbalanced data, which have been shown to display severely affect prediction performance and again bias towards to the majority class, i.e, those samples that have the most observations. This requires designing a fair predictive model, which was proposed as a modified leave-one-out cross-validation [5], a combinational method of k-fold and leave-one-out cross-validation.

The aim of cross-validation is to estimate how well the model fits to independent data during training data process, or with respect to the used training model. This misclassification error rate can be estimated, and this informs us about how well the

model fits a particular designed model. However, there are also other metrics to measure, such as positive predictive value [137], mean squared error [138] and root mean squared error [138]. In addition, cross-validation compares two classes with respect to their misclassification error rate. Furthermore, cross-validation can be combined with feature selection, and the aim of this process is to produce the best model that can select the informative features from large dataset. By using cross-validation in this manner, the model will fit the most informative of subset features (for further details are available in Chapter 6). Furthermore, cross-validation is useful only when both training and testing use the same sample (same population). For model selection evaluation involving cross-validation, there are four main steps of criteria that a such model needs to cover:

- Model selection must be simple and reproducible.
- Evaluate selected method with unseen data (cross-validation) to obtain a minimum estimated misclassification error rate.
- Use a training dataset to build the model.
- Evaluate the model by the unseen data (the test set)
- Repeat the last two points, and then average the overall estimated error rate (misclassification error rate). Subsequently, select the best model for classifiers which gives the highest predictive performance by comparing different predictors via estimations of their average error rate. Further details can be viewed in Chapter 3 (Methodology).

2.3.4.2 Prediction Performance assessments metrics

The outcome of prediction is the accuracy, which is assessed mainly by comparison of the following seven parameters which is driven from confusion matrix. We compared the predictive performance of both parameters with respect to seven metrics such as precision, recall, sensitivity, specificity, mean-weighted accuracy, F-measure (f-score) and Geometric mean (G-mean) by assessing the reliability of the used algorithms on the imbalanced dataset. These seven metrics are extracted from the confusion matrix, and are commonly used to evaluate prediction performances.

Table 2.2 Confusion Matrix

	Predicted as a positive class	Predicted as a negative class
Definite positive class	True positive (TP)	False negative (FN)
Definite Negative class	False positive (FP)	True negative(TN)

All possible combinations of the metrics are applied in order to assess the quality of the predictive model. Furthermore, predictive accuracies and class performances

are based on these four statistical approaches [139] in Table 2.2, which we adapted into our predictive model and assigned as follows:

- True positive $= 100 \times \frac{TP}{TP+FN}$
- False positive $= 100 \times \frac{FP}{TN+FP}$
- True negative $= 100 \times \frac{TN}{TN+FP}$
- False Negative $= 100 \times \frac{FN}{TP+FN}$
- Positive predictive Value $= 100 \times \frac{TP}{TP+FP}$
- Negative Predictive Value $= 100 \times \frac{TN}{TN+FN}$

Predictive accuracies and class performance are measured for four statistical measurements of TP, TN, FP and FN which is usually given as a percentage.

$$True\ positive = 100 \times \frac{TP}{TP + FN} \tag{2.1}$$

This part of formulae represents a prediction of the methylated DNA class, the sensitivity of which has been given as a percentage. This value does not give useful information, and it cannot be trusted without the combined prediction of the unmethylated class also, and hence is termed specificity. The specificity is calculated by following formulae.

$$True\ negative = 100 \times \frac{TN}{TN + FP} \tag{2.2}$$

Specificity is the correct proportion of the prediction of the unmethylated samples. Only two of these four measurements cannot be seen as an effective prediction or performance; indeed, it is highly biased, since it is dependent on the other two parameters.

$$Accuracy = 100 \times \frac{TP + TN}{TP + FP + TN + FN} \tag{2.3}$$

Accuracy is the proportion of the both correct identified results; true positive and true negative for methylated, unmethylated and differentially-methylated respectively. Furthermore, Mathews correlation has been assessed for the performance of two class problem predictive models for a single and all possible feature sub-set combinations. This formulae has been adapted [139], which is derived from [140].

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2.4}$$

this MCC calculation is used for four terms which are driven from the confusion matrix in order to assess the performance of two-class (binary) classifications. For the imbalanced data, it shows that use of only these four measurements are insufficiently precise to measure their predictive performance, and so it requires the inclusion of

other measurements driven from the confusion matrix. These are further assessed in imbalanced datasets.

2.3.4.3 F-measure

By examining only the performance of the positive class (methylated DNA), two metrics are used: the true positive rate and a positive predictive value. The true positive rate is denoted as recall (Re), which is given as a percentage, and which is employed as a retrieval of relevant objects. The other name of the positive predictive value is precision (Pr), which is also given as a percentage value and is used to identify the relevant objects of retrieval.

$$Recall = True\ positive = 100 \times \frac{TP}{TP + FN} \tag{2.5}$$

$$Precision = Positive\ Predictive\ Value = 100 \times \frac{TP}{TP + FP} \tag{2.6}$$

The combination of precision and recall is important when we are only interested in positive predictive classes, but integration of the two is termed the f-measure (details are available in [141]).

$$F - measure = 100 \times \frac{2RePr}{Re + Pr} \tag{2.7}$$

The F measure are also termed the harmonic mean, which is the combination of the precision and recall [142] since the harmonic mean value is increased, results in an assurance that the value of precision and recall are both high.

2.3.4.4 G-mean

G-mean is a square rooted product of true positive (TP) and true negative (TN) this will show high for balanced data whereas severely imbalanced data show low values of G-mean as described by Kubat[143].

$$G - mean = \sqrt{TP_{rate} \times TN_{rate}} \tag{2.8}$$

The G-mean is a measurement representing the two classes of performance involved, i.e, the true positive rate and true negative rate where both measurements show high values [143] the G-mean is used to evaluate the performance whether two classes are balanced or not. Comparing harmonic, geometric, and arithmetic means are explained in details in [142], this paper concluded that the harmonic mean is the most reliable to assess imbalanced data performance, (when compared to the efficiencies of geometric and arithmetic means).

2.3.4.5 ROC analysis

ROC is a visualised method which is performed by using the matched measurements of false positive rate and true positive rate. Where both measurements are plotted on the x axis and true positive on the y axis, this is termed received operating characteristic (ROC) curve. ROC curves are visualised as model performance.

Experimental procedures and imbalanced class predictions and evaluations of the above seven metrics are presented in Chapter 5 (Section 5.4).

2.4 Conclusions

DNA methylation-related topics presented in this Chapter are important to introduce the research reported in the thesis which is based on extracting, analysing and predicting DNA driven sequence features, i.e, those predicting methylation classes and DNA methylation level on CpGs loci positions in both healthy and diseased tissue samples.

In the second section (2.2) is given a brief review of an existing work and its relevant methodology on the DNA methylation status of CpG islands and nine extracted feature-sets that are used throughout the thesis. The third section (2.3) is described in more detail in machine-learning techniques and experimental methodologies, including investigation of and extending the existing methods to allow the prediction of imbalance methylated (sub-) classes or sub-set classes, extracting and selecting features related to methylation differences between males and females both healthy and disease states, and features associated with gender and age by using machine-learning techniques.

DNA methylation is a biological process in which C H3 functions are added to the site of 5-CpGs, without changes in the DNA sequence itself. This biological process has an impact on gene expression by adding or removing this small function mostly at promoter sites where gene activities commence. DNA methylation processes play an important roles in health and well-being in a protective manner, such as body defense and repair systems. However, DNA methylation disruption causes health problems such as depression, neural diseases, immune disorders and cancer.

Having introduced the problem and unsolved limitations in this literature review, Chapter three presents selective existing methods that have been used to predict DNA methylation classes, which leads to further design and selection of the most suitable algorithms for DNA methylation classes, together with selection of the most representative feature sub-set selection from large data samples. These selective and fair methods are used throughout this thesis.

Chapter 3

Materials and Methods

3.1 Introduction

Chapter 3 will present the selective models and classifiers that will be used throughout this thesis, as well as relevant literature in the field of biological data analysis problems.

3.1.1 Experimental Data

Three samples within nine datasets were studied. Samples 1 and 2 were the most challenging ones in view of severe imbalances; indeed, they had inter-class differences, which made it impossible to analyse them by applying machine learning strategies, without designing a 'fair' model to facilitate analysis. Sample 3 did not have the aforementioned problems, so suitable statistical methods were used.

Sample 1, the data is freely available in the public HEP database, which can be accessed at www.epigenome.org., containing three human chromosomes, i.e, 6, 20 and 22, extracted from 43 samples of twelve different tissues [56]; this will be discussed in more detail in Chapter 5. from the sample set containing 495 sub-samples of CpG islands, were extracted. These were averaged methylation changes between CpG pairs of identical samples, in order to minimise any bias produced by the length differences of sequence windows. This dataset was then extended, and 50 features extracted, which were analysed in this study[26].

Sample 2, which is also freely available at http://genome.cshlp.org/content/14/2/247/T1, was extracted from human chromosome 21q[33; 48]. It contained 147 CpG islands samples from the peripheral blood leukocytes or placenta of four healthy human individuals, of which CpG islands 103, 29, and 15 were found to be methylated, unmethylated and differentially methylated, respectively. In order to characterise the DNA sequences, a set of features was extracted across DNA sequences; these are summarised in Table 2.1 in the literature review section. These were represented as

averaged methylation changes between CpG pairs of identical samples, in order to minimise the bias produced by the length differences of sequence windows; in this data set, 3,759 features were extracted. The features were grouped according to their biological function, and subsequently further analyses was performed[5]. The experimental details of these experiments will be described in more detail in Chapter 5.

Sample 3, which is free available in NCBI-database gene expression amnibus under accession number GSE28094, containing a whole human chromosome, with data extracted from 1,628 human samples[63] that consisted of different healthy and diseased tissues. This experimental data contained a fluorescent signal from methylated and unmethylated alleles (CpG loci). DNA methylation and dinucleotide DNA spots were extracted from an Illlumina hybridisation spot array of CYS3 and CYS5, which are methylated and unmethylated respectively. In addition, a mixture of both hybridisations was designed for differentially-methylated rows, representing the ratios of spot values of individual arrays (loci), where the columns represent the individual samples. These spot arrays (features) contained 1,505 CpG sites drawn from 807 genes. These genes are included in oncogenes, tumour suppressor genes, differentially-methylated or expressed genes, imprinted genes, signal pathway genes, DNA repair and cell cycle control genes, and those responsible for metastasis, apoptosis and cell differentiation. These 1,505 features were extended to 1,506 features. (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28094). The experimental design and data analysis will be discussed further in Chapters 4 and 6.

3.2 Cross-validation and classifier technical evaluation (model selection)

Cross validation (CV) is a re-sampling process that was first used in the regression method, in comparison to the leave-one-out cross-validation process (LOOCV)[144], and holdout or M-fold cross validation[145]. Furthermore, LOOCV was extended[146]. Various cross-validation methods were compared, and LOOCV was found to be the most stable and powerful model selection approach[147; 148]. However, LOOCV does not produce consistent model selection with linear regression[149]. In addition, holdout cross-validation was found to be more effective when comparing deterministic penalties for model selection[150; 151]. Moreover, other researchers have argued that M-fold CV is the most promising method for model selection, since the method compensates for bias with a large variance during data partitioning[136; 152].

The theoretical background of cross-validation is discussed in[145]. The above three forms of cross-validation are different with regard to the manner that datasets are partitioned into the training and test sets; generally, in holdout CV, two thirds of the dataset is used as training to build the model, and one third for testing, which is a repetitive process and determines the overall average misclassification error rate

produced by partitioning these data. In M-fold cross-validation, one fold of the dataset is used for model testing, and the remainder is used for training. In leave-one-out cross-validation, all datasets undergo training, except for one data point, which is left to test the model. This split process is seen as universal and independent, where the training set is used to build the model and the test set for evaluating the model, which can be applied to most of the algorithms. Then, the outcome of the algorithms are judged based on how well the test-set performs with each one.

In general, cross-validation depends on two factors: (I) bias misclassification error, i.e, large samples with a small dimension will be less biased in comparison to small samples with a large dimension; and (II) variance caused by data partitioning- for example, random partitioning increases the error rate. The aim of cross validation is to arrive at a model that has both a lower bias and variance. For example, sample S contains X where X = $x_1, x_2, ..., x_N \in x_{1 \leq i \leq N}$ and each X corresponds to an additional membership, a target class Y. As described by[136], if $f(X_n) = \hat{S}f(X_n)$ where f is the algorithm, X_n is the sample, and $\hat{S}$ is the output of the algorithm.

Abbreviation	represents
L_j^t	training set
L^i	test set
$N_i - 1$	leave-one-out training
X_n	sample
f	algorithms
ζ_i	training size
s_m	number of folds
$\hat{S}$	algorithm output

Suppose that $G \geq 1$ is an integer and $L^t = 1, ..., L_G^i$ is a non-empty sequence subset of [1,...., n], the error estimation of $f(X_n)$ with training fraction $(L_j^t)1 \geq G$ is given by

$$\hat{L}^{cv}(f; X_n; (L_j^t)1 \leq j \geq G) = \frac{1}{G} \sum_{j=1}^{B} \hat{L}^{HO}(f; X_n; L_j^t). \tag{3.1}$$

This formula is used in general CV error estimation, where it can only differentiate $(L_j^i)1 \geq G$. This depends upon the choice of data partition. LOO is best used in a model selection where $L_j^t \equiv n_i$ and where the total sample is trained, except in one part of the sample or data point (test set), which is left $(N_i - 1)$ to evaluate the model[144; 145]. In this case, G is equivalent to n and $L_j^t = [j]^c$ for j = 1,..., n;

$$\hat{L}^{loo} = (f; X_n) = \frac{1}{n} \sum_{j=1}^{n} \gamma(f(X_n^{-j}; \zeta_i) \tag{3.2}$$

Where X_n^{-j} is equivalent to $\zeta_{i,i \neq j}$. Some literature refers to this as the 'delete-

one' method, rather than the leave-one-out (LOO) method[153]. Furthermore, M-fold CV has also been proposed[105; 145] in order to reduce the computational barrier of LOOCV. As described by Breiman[105], data is partitioned into subsets of equal $\frac{n}{m}$ where each undergoes training and testing in turn. Suppose that $s_1, ..., s_m$ is the subset of [1, ..., n] with $\forall_j f(S_j) \approx \frac{n}{m}$. Therefore, M-fold CV is an estimator of error of $f(X_n)$ as is described in equation (3.2), for g is equivalent to m and $L_j^t = S_j^c$

$$\hat{L}^{mf}\left(f; X_n; (S_j^t), 1 \leq j \geq m\right) = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{S_j}\sum_{i\in S_j}\gamma(\hat{s}(X_n^{-S_j}); \zeta_i)\right] \tag{3.3}$$

When $X_n^{-S_j} = (\zeta_{i,i\in S_j})$, this takes less computational time than LOOCV, i,e, S training is $\frac{1}{M}\frac{n-n_i}{m}$, in which its bias and variance are the same as $n - n_i$ regardless of the different methods. The M-fold CV variance is given as $\hat{L}^{mf}(f; X_n; L_j^t 1 \leq j)$, which captures most of the information regarding the model performance and assesses model selection. The combination of n_i, n_m and G are modified in LOOCV[5]. Thus, the larger m (for example $m \geq n$), the smaller the bias is $\sum_m^n = \frac{1}{M}\sum_{i=1}^{M}\sum_m^n(i)$. This is very sensitive, since it has large variants when $m \leq 3$, which is known to be a large bias and a lower variant. However, $m \geq 10$ is recommended experimentally. Therefore, $m \geq 10$ is preferable, since increasing the number of the M-fold also increases the computational time. The modified leave-one-out (MLOOCV), however, shows a significantly improved predictive performance, where each independent sample is split into 10 sub-samples, which produce ten independent predictive models, from which their output can be averaged. More detail about this process will be provided in the next sub-section, and the associated experimental details will be reported in Chapter 5.

3.2.1 Modified Leave-One-Out Cross Validation

The modified leave-one-out cross-validation (MLOOCV) method is a combination of M-fold and leave-one-out cross-validation, which is based on the sub-sampling cascade, and incorporating the LOO method to overcome the prediction problems caused by small and imbalanced classes within the dataset. Various cross validation methods are proposed for the assessment of predictive models[105; 145]. The leave-one-out and M-fold methods are widely used, and they have been found to be satisfactory for various dataset sizes. However, the traditional leave one out cross validation method is not productive for small and imbalanced datasets, or classes within datasets. The M-fold cross-validation works very well in practice; however, both methods have been shown to be less productive with imbalanced datasets, and result in a bias towards the majority class[48; 139]. Therefore, they are less able to predict the minority class, which leads to inconclusive results and invalid interpretations. Thus, a combination of

both methods was employed for this study, rather than any individual method, such as those employed many researchers in past studies by [64].

$$\hat{L}^{mloo} = (f; X_n, S_j^i) = \frac{1}{S}\frac{1}{M}\sum_{s=1}^{s_n}\gamma(f(M_s^{-\hat{s}}; \zeta_i)(3.4)$$

Where $\frac{1}{M}$ represents the average of ten independent balanced sub-sets, $\frac{1}{S}$ is a subset of the balanced M sample, M^{-s} is the training set of the balanced M sample using leave-one-out cross validation. Figure 3.1 shows the general M-fold cross validation, whereas Figure 3.2 represents the proposed modified leave-one-out cross validation model, in which the majority class is divided into ten independent sub-sets, and each subset is then composite to the minority class, which results in ten balanced M sub-sets. The leave-one-out cross-validation is then applied to the balanced subsets with a KNN classifier. Cross-validation is used to assess a model evaluation; for example, how well a classifier behaves when unseen data is used to test the performance of the classifier. The unseen data is known as the test set, which is not included during training data, since this unseen data (test set) is used to test the performance of the classifier (model). However, it should be noted that model selection is still an option for the researcher, bearing in mind that each case of the dataset may require a specific CV procedure to evaluate the overall model performance.

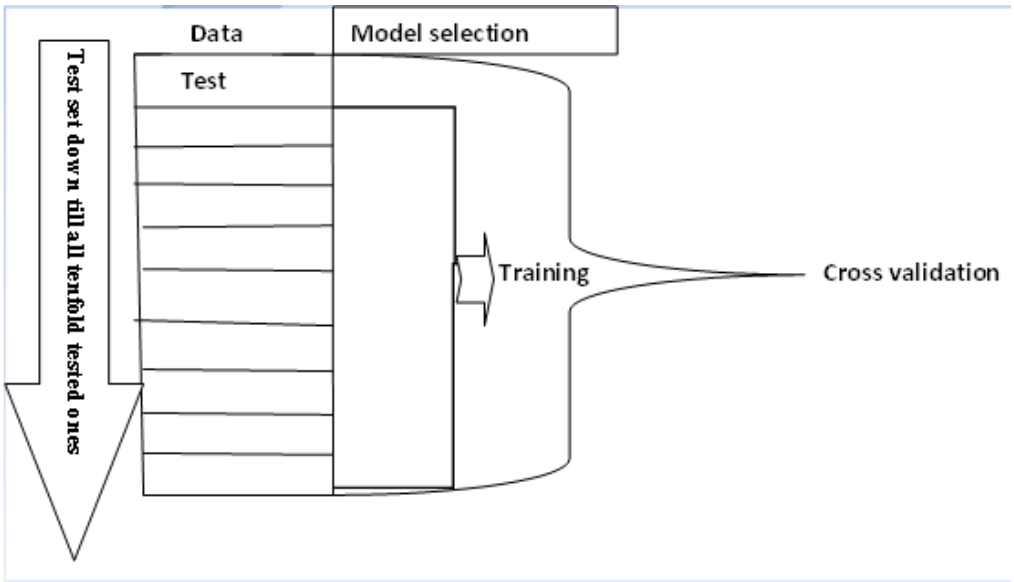


Figure 3.1 M-fold cross validation statistical model.

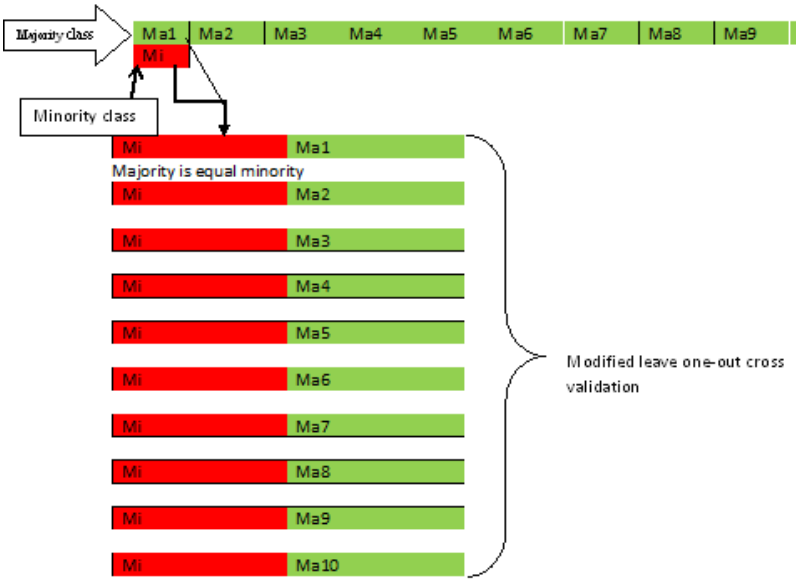


Figure 3.2 Modified leave-one-out cross validation statistical model.

3.3 Predictive methods

3.3.1 General introduction of classification

In statistics and machine-learning, classification is assigning a new set of data (data points) into classes of origin, on the basis of training data which contains observations that are predefined for targets or data labels. The labels can be categorical or numerical, for example methylated and unmethyalted are categorical, but this parameter can also be replaced a numerical value of [0,1]. An algorithm can be classified as either methylated or unmethylated, but it cannot select the most informative classes or subset features. This is based solely on the knowledge of the researcher or investigator; it is up to the researcher to identify a suitable means to design a specific model in order to distinguish or discriminate the interested classes or features. Each new data point can be predicted, as can its class membership of the closest training set.

3.3.1.1 K-nearest neighbour classifier (K-NN)

K-nearest neighbour classification is the simplest non-parametric learning algorithm. It belongs to the top ten classifiers, and is used for predicting DNA methylation classes. K-NN is the most suitable algorithm for studying data with an unknown probability

density, or that is not easy to calculate from the sample. This technique was first applied with poor knowledge of data distribution[104]. The non-parametric rule has since been developed (and is based on the K-nearest neighbour method), to tackle the prediction problem caused by poorly distributed data that cannot be used for other discriminate analysis methods, since it is not able to estimate the probability density from the sample. Other researchers have extended the K-NN properties and produced a number of K-nearest neighbour variances (between k=[1,n])[3; 104]. Furthermore, K-NN has been investigated in many studies, and has been compared with Bayesian error properties[154; 155; 156; 157].

3.3.1.2 Basic principles of K-NN

K-NN uses predominantly Euclidean distance as a measurement metric, an index which is based on the distance between the test and the training sets. K-NN identifies the training instances that are closest to the test-set, and then the distances (similarity) between them are computed.

Let's suppose that $(x_i, y_i), .., (x_n, y_n) \equiv D$ Where $X_i \in R^d, y \in \Re^d$

$$D_x(x_i, x_j) = \sqrt{\sum_{k=1}^{d}(X_{ik} - X_{jk})} \tag{3.5}$$

D is the distance between two data points, which can be a minimum of $D_x = 0$, i.e, the two points are classified according to the nearest neighbour. There are factors that influence the KNN performance, such as the choice of number of k-nearest neighbours. For example, K=1 is sensitive to outliers, and the classification performance is degraded, whereas with a large K, the classification is too 'smooth', as a larger number of neighbours may take account of the data-points from other classes.

K-NN properties :

- K-NN classifier is simple, easily interpretable and performs very well in a broad application.
- K-NN is a 'lazy' classifier (zero effort at training time and full effort at prediction time).
- The experimental error rate (training set) is approximately zero. This means that the error rate between training and test set approaches asymptote.

Disadvantages:

- A decision boundary is sometimes a very approximate indecision boundary, particularly when k<3.

- High dimensional data classification performance is degraded, and requires high computational time.
- A number of k-nearest neighbours are unavailable during data training, i.e, single prediction (k=1) is not flexible (too much bias), while multiple prediction (k=n) is also too 'smooth' and very expensive for computing.

For this study, in view of its flexibility, effectiveness and power, K-NN was adapted alongside the modified leave-one-out cross-validation method.

3.3.1.3 Quadratic discriminant analysis

Discriminant analysis refers to a set of vectors of observations (x) of an event, each of which corresponds to the assigned target y; one set of data is denoted as the training set. The classification is to determine a newly given observational vector, which is assigned to a class membership, a process that in some literature is known as prediction. A discriminant method is used when the data is large and contains enough information to be grouped into classes in terms of the origins of samples. Literature available reveals different discriminant methods that can be used for DNA methylation analysis. In general, however, observation requires two classes, which are required to have at least two-dimensional surfaces to separate the classes; for example, Quadratic Discriminant Analysis (QDA) can be used to separate the two classes. QDA is a generalised version of the linear model that can be used for more complex separating surfaces. Linear discriminant analysis (LDA) and QDA are closely related. However, the two classes of dataset are assumed to be normally distributed in LDA, and therefore their class covariances are equivalent, whereas in QDA it is assumed that there is a different covariance for each class. Hence, the covariance matrix is estimated for each class c$\in C$(c = 1, 2, ... , c_n) of QDA, while LDA uses a pooled covariance matrix. The classification rule of QDA is noted as in [158; 159]

$$\underbrace{Cr(x_i)}_{1\leq c\leq C} = (x_i - \mu_c)^T \Sigma_c^{-1}(x_i - \mu_c) + \ln|\Sigma_c| - 2\ln\Pi_k \tag{3.6}$$

where Σ_c is the class covariance matrix of class c, μ_c is the mean factor of class c, and Π_c is the prior probability of class C, which is further estimated by

$$\hat{\mu}_c = \frac{1}{(n_c)}\Sigma_{i=1}^{(n_c)} x_i, \tag{3.7}$$

$$\hat{\Sigma}_c = \frac{1}{(n_c)}\Sigma_{i=1}^{(n_c)} (x_i - \mu_c)^T, \tag{3.8}$$

$$\hat{\Pi}_c = \frac{n_c}{n}, \tag{3.9}$$

where n_c is the size of the class c, and n is the size of the training dataset. x_i is set as the lowest classification score, which is based on the QDA classification rule as formulated in equation (3.6). In this equation, the first right-hand term represents the Mahalanobis distance. Since it assumes that LDA has an equal covariance matrix of the two classes, a combined covariance matrix is formulated as:$\hat{\Sigma_c} = \Sigma$ for $1 \leq c \leq C$

$$\Sigma_{combined} = \frac{1}{n}\Sigma_{c=1} n_c \Sigma_c \tag{3.10}$$

This is substituted for class covariance in the first equation without constants, and driven by the LDA classification rule:

$$Cr(x_i) = (x_i - \mu_c)^T \Sigma^{-1}_{combined}(x_i - \mu_c) - 2\ln \Pi_c \tag{3.11}$$

The Mahalanobis distance is equivalent to the first term of equation 3.11 when the prior probability (Π_c) is constant. QDA analysis has a poor predictive performance as the dimension to sample ratio increases[114]. In addition, QDA terminates during model building (data training process), particularly for small samples with high dimensions. This termination is caused since the covariance matrix (Σ_c and $\Sigma_{combined}$) becomes singular. However, feature selection can be used to overcome these problems. This will be further discussed in the present chapter (section 3.4), and the experimental details are explained in Chapter 6.

3.3.1.4 Decision Tree

A decision tree is a non-linear classifier that uses a single display to group data into classes. Such a tree's branches represent feature values, the corners of the branches represent the possible values of the features, and the leaf represents the class label. Test sample prediction follows this hierarchical order, from the tree node to the root via the leaf. The most popular tree algorithms are ID3 ID3[160], C4.5[160] and CART[119]. A decision tree algorithm is designed in two ways: (I) tree-building, and (II) 'pruning'. The tree-building method is based on splitting training recursively until it assigns into its class of origin. The partition of the training set is followed by pruning, which minimises possible overfitting. 'Pruning' involves a generalising of the decision tree, and 'smoothing' of the branches of the initial tree. This improves the classification error rate; therefore, pruning from tree root node to the sub-branches and leaves will lead to a reduced error rate.

Imbalanced data require the building of a decision tree, since the leaf represents the class label. However, this approach is more likely to select the majority class. This bias can be reduced by taking many test samples, in order to distinguish the minority class from the majority one. However, there are some limitations to this method; for example, the algorithm could terminate prematurely during the splitting (training set)

process before the classifier finds the minority class, a consequence of which is that the minority class would not contribute to overall error reduction. Hence, it is more likely that some branches representing the minority class will be ignored or removed during the training split and replaced with majority ones. To tackle this problem, the C4.5 decision was employed for use with the imbalanced datasets in this study [126; 161]. In the literature, the following steps have been used to undertake decision tree prediction[3]:

- Each stage of splitting the training dataset into sub-sets follows a specific rule: suppose X_t is a subset node t where x_t is the sub-set of training set X , further x_t becomes two disjoint subsets x_{t_m} and x_{t_u} the label m is methylated and the label u is unmethylated. The top node represents the training sub-set of class X, and each further split follows this rule.

- The criteria for splitting must follow the best chosen sub-set from X classes of possible candidates.

- It is important to control the stop-split rule of a growing tree, and the leaf node is terminated as it reaches the target threshold.

- The stop-splitting rule is required to assign each leaf to a particular class.

It should be noted that these steps are not specific to one approach, and each step can be used in more than one method. These steps can be adapted using the AdaBoost algorithm, which will be explained in detail in the next section.

3.3.1.5 Fit ensemble algorithms (AdaBoost) are used for imbalanced data analysis

For the last two decades, ensemble methods have mostly been used in pattern recognition[162; 163]. Ensemble involves the construction of a number of classifiers under one 'umbrella' in order to solve complex pattern recognition problems[164]. This enables the building up of a set of weak classifiers from the training set, which are classified by averaging the prediction outcome of each weak classifier. The essence of the ensemble method is building a strong classifier from the original training data by initiating a single classifier, i.e, the output of the classifier is aggregated. Figure 3.3 shows a number of parameters and factors that can be modified to create ensemble methods. Data matrix and training data are frequently employed as a basic architecture with a base classifier[162]; the output of the ensemble and the most effective methods have been reported in [165]. The majority vote method is the most common in current literature[166]. The aim of aggregating the classifier in the ensemble is to generalise the misclassification error rate. The error that the base classifier produces in each iteration is not necessarily the same, which enhances the recognition ability of the

minority class. Boosting is an iterative method of base classification learning algorithm; it generates a sequence classifier and updates repeatedly. Each update of the training cycle assigns a weight, and adaptively changes the weight in each boosting cycle. Misclassified instances of each iteration are assigned according to weight in the next learning cycle. There are two methods for this: (I) class probabilities and (II) modelling of the base classifier. There are also two important questions requiring answers: (I) how can each iteration of weighted training samples be updated?, and (II) How can the hypotheses of the outcome be measured? To address these two problems, the use of AdaBoost is proposed, where a parameter (α) is used in response to both problems[116], in order to update each data space and weigh the class differences. Adaboost has a property that makes it a strong classifier, in that it contains several combined classifiers[118]. The α parameter leads to a significant improvement in the base classifier, and a coupled reduction in variance and bias.

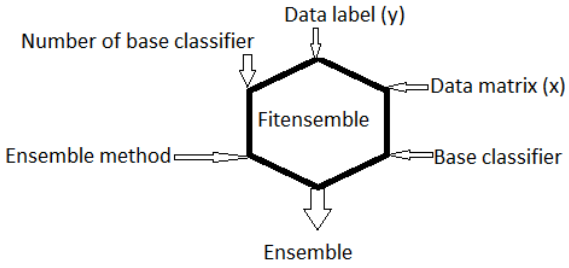


Figure 3.3 Ensemble scheme model.

Adaboost was designed to address two class problems, for example input of training data is $(x_1, y_1, ., x_n, y_n)$ where the i^{th} instance (x_i,y_i) x_i is a function of data set X and class label y_i represents a value in Y with a bi-class $Y = \pm 1$. The given base-learning classifier is iteratively run, such that t =1, ...,T. W_i^t represents the weight of the i^{th} training instance of the iteration, where all weights are set as equivalent at the beginning, and w_t is updated in each iteration, as shown in algorithm 3.1. The task of the base learner is to find a base classifier for $h_t : X \Rightarrow Y$. As the base classifier h_t is trained, AdaBoost finds a parameter $\alpha_t \in R$ in order to measure the predictive performance of the classification h_t. Subsequently, w_t is updated in each iteration. H(x) represents the final classification criterion for a weighted majority vote of the number T ($\sum_{t=1}^{T}$) of base classifiers, and α_t is the weight parameter given to each of the base classifiers (h_t) . E_w represents the initial error calculation, with $w_i = w_1, w_2, w_n$, and I a multiplicative indicator for each iteration, where weight increases alongside misclassification errors, as the sample x_i is misclassified by h_t which is dependent on E_w. $W^{t+1}(i)$ is the weight of the second round iteration. E_w requires

Algorithm 3.1 AdaBoost.M1 algorithm introduced by Freund and Schapire [106]

1: **procedure** SUPPOSE THAT X $=[x_1, x_2, ..., x_i]$ AND Y $= [y_1, y_2..., y_i]($ ▷)where i= 1, 2, ..., n; $x_i \in X, y_i \in Y =$ [-1, +1]
2: Initialise $w^1(i)$ ▷ $w_i \in X$ where w_i is a distribution drawn from X.
3: $W_i = \frac{1}{n}$ where i = 1,2,...,n; ▷ normalise W_i to probability distribution.
4: choose a number of weak learners t = 1,...,T.
5: Train the base learner h_t: $X \rightarrow Y$ by using distribution w^t.
6: Estimate the error so that $E_t = Pr_i\ w^i(h_t(x_i) \neq y_i)$.
7: E_t>1/2 terminate, or else calculate α_t. ▷ Where $\alpha_t = 1/2log(\frac{1-E_t}{E_t})$.
8: choose α_t as a weight updating parameter.
9: Update the weight; $W_i \leftarrow W_i exp(\alpha_t I(y_i \neq h_t(x_i, y_1))$ ▷ where i = 1,2,.., N. This should be normalised so that $\sum_i W_i = 1$.
10: $W^{t+1}(i) = \frac{W^t(i)\exp(-\alpha_t h_t(x_i)y_i)}{Z_t}$ ▷ Z_t represents a normalising factor, and the final output is the AdaBoost.m1 algorithms: $H(x) = (\sum_{t=1}^{T} \alpha_t h_t(x_i))$.
11: **end procedure**

a training error with W_i^N; i = 1,2,..,n, so that W_1, W_2, W_n , and i represent the test set. Hence, AdaBoost increases the weight of misclassified sets at each round by a factor of $W^{t+1}(i)$. With respect to weighted training errors, this will not depend on either α_t nor on h_t. Furthermore, $W^{t+1}(i)$ can be incorporated within the final step of the base classifier without affecting the optimisation step. $H(t) = sign \sum_t^T a_t h_t(x)$ - represents an additive model, since it fits a stepwise classifier, which performs better than a single classifier. Specifically, the single classifier is repeated to alter the original distribution of data. AdaBoost with a tree is reported to be "the best 'off-the-shelf' classifier" [119]. However, AdaBoost.M1 does not perform well with categorical data, showing an unstable performance. In this study, Gentle AdaBoost was employed; the details of the experiment will be provided in Chapter 5 (Section 5.4). Both algorithms have similarities, but the main difference between them is that Gentle Adaboost uses class probability to update weights, whereas Adaboost.m1 uses the 1/2 log ratio [118]. $H_t(x) = \sum_{t=1}^{T} a_t h_t(x)$ where $h_t(X) = P_w(y = 1|x) = P_w(y = -1|x)$ rather than $(a_t = 1/2log(1-err_t$. Thus far, author have identified two class problems, and in the next step author will explore multi-class classification problems. Several researchers have examined this issue[116; 118]. Freund and Schapire[116] proposed a similar approach to the aforementioned two-class problems, where AdaBoost.m2 is extended to solve multi-class problems[106]. In this case, Y represents k = 1, 2, K, where K rpresents the number of possible classes.

AdaBoost.m2 can be generalised in the same manner as AdaBoost.m1. Similarly, it fails to predict when the overall accuracy is lower than 50%. However, when k=2, the probability of 0.5 is a random guess, but with $1/k \ln 0.5$, $\alpha_t \ln 0.5$ still gives a strong prediction; therefore, the base classifier performs more effectively than pure guess-prediction. AdaBoost.m2 uses more complex error measurement methods than

Algorithm 3.2 Gentle AdaBoost algorithm by Friedman and Robert [118]

1: **procedure** SUPPOSE THAT X $=[x_1, x_2, ..., x_i]$ AND Y $= [y_1, y_2..., y_i]$(▷)where i= 1, 2, ..., n; $x_i \in X, y_i \in Y = [-1, +1]$
2: *Initialise* $w^1(i)$ ▷ $w_i \in X$ where w is a distribution drawn from X.
3: $W_i = \frac{1}{n}$ where i = 1,2,...,n; ▷ normalise W_i to probability distribution.
4: choose a number of weak learners t = 1,...,T.
5: fit the regression function to the base learner h_t: $X \rightarrow Y$ by using distribution w^t and weighting the least-square of y_i to x_i.
6: $H(x) \leftarrow H(x) + h_t(x)$ and update.
7: update $W_i \leftarrow \mathrm{W}_i exp(-y_i h_t(x))$ this will be normalised.
8: Estimate the error so that $E_t = Pr_i\ w^i(h_t(x_i) \neq y_i)$.
9: The combined final output $H(x) = sign \sum_{m=1}^{M} h_t(x)$.
10: **end procedure**

m1, namely pseudo-loss; each iteration of the classifier is fed with a pseudo-loss function which varies from sample-to-sample and one round to the next. For pseudo-loss modification, AdaBoost focuses on the base classifier that has misclassified the target classes. AdaBoost (M2) was subsequently developed, and shows improved multi-class prediction with an additional base learning classifier design[116]. Similarly, the α_t parameter is used to reduce the training misclassification error rate of the combined base classifier, and h_t is the algorithm output of the original dataset (X) and its class target (Y) of $X \rightarrow Y$, for K response y_i for k class problems, where each class k has a value of 1 or -1 for each training instance (x_i, y_i), and where mislabelled $y \in Y - (y_i)$, Q(i,y) represents weights from the correctly-labelled y, extracted from y_i for $y \in Y(-y_i)$. Each tree represents class k where $W^t_{1,y}$. This reduces one large tree to sub-trees, where class k fits the disjointed additive model. Hence, one class opposes all others.

For the representation of AdaBoost.M2, (x_i, y_i) represents training instances and each mislabelled $y \in Y - (y_i)$ is defined $\tilde{x}_{i,y}$ = (i,y) associate with $(y_{i,y}) = 0$. Therefore, $n = n(k-1)$ is a set of samples, and each sample contains a pair of (i,y), which corresponds to distribution D (i,y) and this is equivalent to D(i)/k-1. Therefore, the i^{th} round of the base classifier (h_t) finds the smallest error rate of the next round, until it reaches the iteration number threshold. Freund[106] defines this as follows:

$$\hbar_t(i, y) = 1/2(1 - h_t(x_i, y_i) + h_t(x_i, y)) \tag{3.12}$$

Suppose that sample i of X_i is misclassified for $h(x_i) \neq y_i$ so that $h(i, h(x_i)) = 1$, hence $pr_i\ D[h(x_i \neq y_i] \leq pr_i\ D[\exists y \neq y_i\colon\ \hat{h}(i, y) = 1]$. Since one versus all base classifiers has an output of one, and the rest are zeros, the error $\hat{h}$ will be:

$Pr(i, y)\ D[\hat{h}(i, y) = 1] \geq \frac{1}{k-1} Pr\ D[\exists y \neq y_i\colon\ \hat{h}(i, y) = 1]$.

In this study, AdaBoost was applied to the dataset, since it performs better than other experimental models. AdaBoost has stepwise optimisation properties, and it also has overfitting immunity, as described in[118]. Moreover, it has an overall error

Algorithm 3.3 Freund and Schapire[106] describe AdaBoost.M2 algorithm

procedure INPUT DATA $((x_i, 1), y_i, 1), ..(x_i, k), (y_{ik})(\triangleright$)where i = 1,2,..,n and y_{ik} labels of value -1 and 1 is target for class k and observation i.
    Specify a number iterations of weak learner: t = 1, 2, 3, T.
    start with the weight factor: $W^t_{1,y} = \frac{w_i}{k-1}$ for i = 1, 2, 3,..., n and $y \in y - (y_i)$ w^t is a sample distribution drawn from X.
    Initialise $W^t_1 = \sum_y \neq y_i \; W^t_{i,y}$.
    $Q_t(i, y) = \frac{W^t_{i,y}}{W^t_i}$ for $y \neq y_i$.
    Set $w_t(i) = \frac{W^t_i}{\sum_{i=1}^n W^t_i}$.
    Train the base classifier with sample distribution w^t combined with weights (Q_t); retrieve $h_t X \times Y$ corresponds [0,1].
    Calculate $E_t = 1/2 \sum_{i=1}^n W_t(i)\; 1 - h_t(x_i, y_i) + \sum_{y \neq y_i} Q_t(i, y) h_t(x_i, y)\Big)$.
    Calculate $\alpha_t = \frac{E_t}{1-E_t}$.
    update the first round of base classifier and start the next round: $W^{t+1}_{i,y} = W^t_{i,y} \alpha_t^{(1/2)(1+h_t(x_i,y_i)-h_t(x_i,y))}$; where i = 1, 2, 3, n, $y \in Y - (y_i)$.
    Final output of Adaboost.M2 will be: H(X) = $armax_{y \in Y} \sum_{t=1}^T \left(Log \frac{1}{\alpha_t}\right) h_t(x, y)$.
end procedure

reduction ability for each iteration of learning classifier, via the following:

$$\sum_{i, y_i \neq h_t(x_i)} w^t(i) = \frac{1 - E_t}{2} \tag{3.13}$$

Increasing E_t corresponds to minimising the training error for each iteration. α_t is used to reduce the training error of the combined base classifier, and also reduces the training error at each training step. Despite the fact that the overall predictive accuracy of the data is improved, an imbalanced dataset demonstrates a less predictive performance, particularly in relation to the minority class.

3.3.1.6 Aims of Weighting Efficacy

Weighting efficacy is a parameter for which each iteration of base classifier reduces the weights of correctly-classified training instances, whilst increasing the weights of those that are incorrectly classified. α_t must be a positive value, as training error should be less than that of randomly guessing, 50%, by considering that

$$\sum_{i, y_i = h_t(x_i)} W^t(i) > \sum_{i, y_i \neq h_t(x_i)} W^t(i) \tag{3.14}$$

Where correctly predicted, the left-hand equation 3.14 is greater than the right one, i.e, E_t is smaller than 50% and the final output of predictive accuracy is greater than 50% (by chance). However, sometimes Adoboost does not perform any better than

pure guesswork where the number of misclassified samples is equal to the correctly-classified ones:

$$\sum_{i,h_t(x_i)=y_i} w^{t+1}(i) = \sum_{i,h_t(x_i)\neq y_i} w^{t+1}(i) \tag{3.15}$$

Therefore, when the weight is updated, the misclassified and correctly-classified samples are balanced. However, this form of weighting increases misclassification instances at an equal ratio, and decreases the correctly-classified instances, also at an equivalent ratio, until the sample distribution becomes equally weighted and the predicted classification output is either correct or incorrect. However, imbalanced data within unequal class distributions is prone to a misclassification of the minority class. Although AdaBoost improves the predictive performance of the minority class, the majority class may have more samples misclassified than the minority class. This may arise because the minority class has less opportunity to be identified, which mostly shows a lower classification predictive performance than the majority class. In order to tackle this problem, a cost-sensitive method is combined with the AdaBoost. This will not affect the final output of the classification when multiplied by a constant positive number at each iteration of the weighted step; however, it does modify the cost distribution of training data. The cost-sensitive method finds the exact ratio between minority and majority classes. The essence of this method is to improve the classification performance of the minority class. The experimental details of this method will be provided in Chapter 5.

3.4 Feature extraction and selection

The aim of this section is to explore methods that can extract useful or informative (sub-)features from multi dimension datasets, imbalanced data and noisy features. It has been reported that major problems are associated with large features, for example

DNA sequence distribution and patterns

[5], failed to predict methylated from unmethylated features when applied to direct machine-learning without feature preprocessing.

3.4.1 Feature selection

It is important to demonstrate that the various stages of classification design are closely interdependent of feature selection. However, it may not be possible for a large dimension of features to be easily integrated with classifier design techniques such QDA and KNN. The aim of feature selection is to combine the classifier and feature selection methods by minimising the error probability directly. The combined method should be computationally simple, in order to allow a search for the optimal

feature combination. The next sub-sections will discuss further how to tackle multi dimensional data problems, as noted above.

3.4.1.1 t-test Approach

The t-Test is used to determine whether or not two samples have statistically significant differences. These differences are evaluated via the sample size, the standard deviation of the two groups and the mean difference between the two samples. Suppose the mean difference of two classes, C_1 and C_2 , is given by: $C_1 - C_2 = 0$ where $\mu_{c_1} - \mu_{C_2} = 0$, with the assumption that the features $(x_1, , x_m)$ we assume that the variances of the two classes are not different where $\delta^2_{c_1} = \delta^2_{c_2} = \delta^2$. The closeness of the classes can be identified by the hypothesis:

$$H_0 : \delta_\mu = \mu_{c_1} - \mu_{c_2} = 0 \tag{3.16}$$

$$H_1 : \delta_\mu = \mu_{c_1} - \mu_{c_2} \neq 0 \tag{3.17}$$

Where H_0 is either rejected or not; if it is rejected, an alternative hypothesis is accepted (H_1). This is denoted Z = X-Y, where X and Y represent random variables of the data points of the two classes, C_1 and C_2 respectively. The data points (features) are assumed to be statistically independent, where estimated $z = \mu_{c_1} - \mu_{c_2}$ and independent variables $\delta^2_{c_1,c_2} = 2\delta^2$, which can be noted as:

$$\bar{Z} = \frac{1}{m}\sum^{m}(x_i - y_i) = \bar{x} - \bar{y} \tag{3.18}$$

Clearly, the variance $\bar{Z}$ is a part of a normal distribution

$$u(\mu_{c_1} - \mu_{c_2}, \frac{2\delta^2}{m}) \tag{3.19}$$

At this point, the sample is large enough (n). However, in most cases, the variance is unknown; the density function is then replaced in the z equation, which is noted as:

$$Q_D = \frac{(\bar{x} - \bar{y}) - (\mu_{c_1} - \mu_{c_2})}{s_z\sqrt{\frac{z}{n}}} \tag{3.20}$$

This leads to a standard deviation notation,

$$S^2_z = \frac{1}{2n-2}\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(y_i - \bar{y}) \tag{3.21}$$

where X_{c_1} and Y_{c_2} are normally-distributed variables and have the same variance δ^2. Therefore, equation 3.23 becomes a t-distribution (Q_D) with 2N-2 degrees of freedom, as noted in equation 3.24. These equations approximate a t-test of the

given null hypothesis, either a one-tailed or two-tailed test, and each is given with appropriate degrees of freedom. The P-value can be calculated from the T-table and then a statistical significance threshold set based on the preference of the investigator. The significance level for the P-value is usually 99.99, 95 or 90%, whereby a null hypothesis is either accepted or rejected in favour of an alternative hypothesis based on these significance values.

3.4.1.2 Feature selection wrapper methods

For a multi dimensional dataset, it is necessary to pre-select a small and informative number of features before applying machine-learning techniques [3]. This is known as filtering and ranking, and features a numerical weight that is employed to rank all the features; the top-ranked features are then selected. Filtering methods are independent in the classification rule, for example, the t-test can be used at any stage of the classification process, which is based on testing whether two features are statistically different or not [96], and then the best feature-vector combinations are selected. This selection method is based on the feature separateness and its selection criterion.

However, the wrapper approach is not necessarily based on the values of an adopted class separateness criterion, but rather on the performance of the classifier(s) itself [3; 111]. Several filters are therefore required in order to produce an improved initial feature sub-set. The Kulback Lieber and Chi-square filters are the two most commonly found in the literature [3], as noted in section 2.4.1.6 of the literature review, which describe how the Kulback-Lieber method uses the distance between histograms of feature values to compute the class separability of each feature, and selects the sub-features with a probability proportional to the Kulback-Lieber distance. However, the Chi-squared statistic tests whether the data in question is statistically independent from one feature data point to another [3]. Both techniques typically use feature preprocessing in order to improve predictive accuracy, and concomitantly reduce noisy features that degrade the class performance. Although these two methods are widely applied to feature selection, the wrapper method is the most popular.

There are two wrapper methods discussed in the literature[167]: forward sequential feature selection (FSFS), and backward sequential feature selection (BSFS). FSFS starts with one feature f_{x_1} , by identifying an optimum feature or feature sub-set. Specifically, FSFS finds the best feature sub-set by combining a pair of features. BSFS begins with a number of features and eliminates the least informative ones sequentially (f_{x-1}). There are four main steps to the selection process:

- Searching feature or feature-sub-set from feature space by generating the search approach for features assessment. For example, FSFS starts with zero features, whereas BSFS begins with the total features, and features are then added and

removed iteratively, respectively.

- In feature sub-set evaluation, it is determined whether the added or removed features result in a better or worse predictive performance, i.e, comparison of a pair of features, after which a decision can be made according to their performance; thus, if a feature is better than the previous one, then it is replaced. Otherwise it is ignored, moving on to the next feature until all the features have been evaluated.

- The feature search must have a limit so that the search must stop once a suitable stopping criterion is reached. For example, this criterion could be a number of iterations, finding an optimum feature sub-set, or that the training error rate does not improve.

- The validation assessment (the test set) is independent and does not include the aforementioned three steps. For example, the test set judges the performance of the selected feature (sub-set) when the training reaches the assigned stopping criteria.

Suppose D is the total number of features, X is a sample size, K is a selected feature and C represents the class variables. $D_0 = D_1$, D_k represents the K features that are already selected; suppose that $d_1 \in (X_1, ..., X_n - D_k)$ is given as $f[D_k \bigcup (d, j) \geq f(A_k \bigcup (d_1)) \forall d \in (x_1, ..., x_n)$ -D_k where D_k is given as $D_{k+1} = D_k \bigcup d_1$, and the latter is added to each run of the data for one feature of T time, which is how the process is named FSFS. However, FSFS does not guarantee optimal feature selection, as some of the literature claims[168; 169], since it cannot guarantee that the selected feature is the optimum one. Notwithstanding, SBFS is exactly the opposite of FSFS, where the $\hat{D}_k - D_1$ sub-set contains (N-K) features; hence, the choice is $_0 \in \hat{D}_k f(D_k - (_{0j} \geq f(\hat{D}_k - ()_s \in \hat{D}_k \hat{D}_{k+1} = \hat{D}_k(Z_0)bn/Z$, which is computationally-expensive. It also requires more interdependence features than FSFS; however, it may yield an improved level of accuracy, although it is not guaranteed that BSFS will perform more effectively than FSFS. The main steps for identifying the best sub-set feature use a QDA classifier, which computes the feature vector datapoints and adds in one feature increments each time by combining feature pairs for all possible combinations. The process continues until the best sub-set (feature) is selected.

3.5 Clustering

Clustering involves analysing and extracting features from data that has unknown base distributions. It groups data into sub-sets based on similarities or dissimilarities. In general, the clustering process consists of the following steps [4]:

- Representing data

- Deciding on the metric (distance) that should be used
- Clustering using algorithms
- Cluster definition
- Evaluation of results

Data representation involves indicating the nature of the data such as scale, size, dimension and number of clusters. The choice of metrics also depends on data structure and the inter-relationship of features or datapoints, which leads to a model design that fits the data structure. In more technical clustering sites, features are represented as feature vectors, where features are given as $X = (x_1, .., x_n)$, and n represents the number of dimensions. Some of the clustering criteria steps are described as follows[3]: (I) Clustering must be stable and should not be changed when an extra group is added (growing) or some objects are removed, and (II) there should also be independent initial object ordering.

Clustering has been applied successfully in the field of bioinformatics, such as for microarray data, specifically Illiumina arrays microarray data, for example in gene expression profiles to identify sub-features of co-expression genes[99]. Clustering has also been applied to CpG islands in order to identify the length of CpG island clusters [64]. Moreover, it has been utilised in CpGs methylation to determine unknown cancer types with in various tissues[63]. The most popular clustering methods are hierarchical and K-means clustering. K-means is one of the most used feature extraction algorithms, and is based on squared Euclidean dissimilarities. Some variants of K-means have been proposed in order to improve the efficiency of algorithms, and also to find the global optimum[100; 101; 102]. K-means is one of the most successful algorithms used on multi dimensional datasets with filtering techniques, since it is easy to implement and is less computationally-complex. However, heatmap clustering is the most powerful visualisstion technique in bioinformatics.

3.5.1 Hierarchical clustering

Hierarchical clustering is one of the most powerful visualisation tools available for use with genomic data. It applies feature extraction and visualisation in DNA methylation, and enables the grouping of similar objects into one cluster, which categorises them according to their natural origins. Heatmap clustering results in a dendrogam, where the branched objects can be visualised based on their similarities, since clusters transform into one level group. Hierarchical clustering groups according to three main categories: single-linkage [170], complete-linkage [171; 172] and average linkage [173; 174] algorithms.

Single-link is the minimum distance between two clusters; complete-link is the maximum distance between all pairwise points in two clusters; and the average linkage

is the average distance between any two closest clusters to the all-pairs points of clusters. The difference between complete-link and single-link is that complete-link gives a compact cluster, whereas single-link results in a chaining effect, which tends to produce cluster-elongation [175]. Hierarchical data is grouped one level at a time, followed by the next level until culminating in a single branch or tree. The two closest branches are the same cluster that initiates from multilevel trees to form a single dendrogram (cut-off points). There are three main steps required to perform data clustering: (I) identifying the dissimilarity and similarity datapoints between pairs; (II) calculating cluster datapoints of pair distance, using measurements (Euclidean distance); and (III) determining the dendrogram linking all datapoints to the single tree connection, from either top to bottom, or vice-versa. This represents hierarchical clustering; the most common used metric being the Euclidean distance which is given as:

$$X_{a,b} = \sqrt{\sum_{i=1}^{p}(x_{a_i} - x_{b_i})^2}), \tag{3.22}$$

where p is dimensional variables, since the average measurements change from one group of samples to the next, and the dissimilarities are removed for each growth of the tree-nodes (branches), i.e, these cluster groups are joined from one level to the next based on the similarity of the datapoints, which is measured by Euclidean distance, the distance between points. The bottom level of the dendrograms is the final group of objects, and these are the selected features. However, selection of the features depends on the interest of the investigator, whether this is the top-level (a group of features) or the bottom, a single dedrogram. This is known as a agglomerative and divisive hierarchical algorithm in the literature[4]. In the present study, divisive clustering was applied and adopted with an average linkage. However, empirical methods have shown that a scaled dataset gave rise to improved results using correlation distance rather than Euclidean measurement; more details regarding this will be provided in Chapter 4. The three divisive methods that were adopted are:

- Single-linkage as the minimum distance between two Clusters:

$$D(A,B) = \underbrace{Min^{L}}_{a\in A, b\in B}\, d(a,b) \tag{3.23}$$

- Complete-linkage as the maximum distance between all pairwise points in two clusters:

$$D(A,B) = \underbrace{Max^{L}}_{a\in A, b\in B}\, d(a,b) \tag{3.24}$$

- Average linkage as the average distance between any two closest clusters to the all-pairs points of clusters:

$$D(A,B) = 1/A1/B \underbrace{\sum}_{a\in A, b\in B} d(a,b), \tag{3.25}$$

where d(a,b) is the distance between objects $a \in A, b \in B$, and A and B are two sets of objects (clusters).

3.6 Conclusions

This chapter presents the selective materials and methods that are important within DNA methylation and wide genomic data. It also justifies the use of the selected methods, explaining why they are important for this work. A suitable study design for imbalanced datasets, small sample sizes and multidimensional data are essential for fair statistical analysis, which will yield results comparable with other studies. Developing a fair predictive model is very important, since some of the studied datasets were severely imbalanced; for example the chromosome 6 dataset had a ratio of methylated to unmethylated of 1:20, which cannot be analysed directly via machine-learning techniques, which require the design of a modified leave-one-out cross validation and the combination of two methods (weighting and cost-sensitive), to be applied to imbalanced datasets. These methods were experimentally applied to DNA methylation classes, and the results obtained are presented in Chapter 5.

Learning algorithms are another factor that must be considered. The learning method must be fair for all data-sets and should not be biased towards any particular methods or dataset, but must also produce easily interpretable results. However, the aim of this thesis is not to develop new algorithms, but to design an intelligent method that 'teaches' the classifier the behaviour of the data in order to produce representative results, for example data within imbalanced classes. Modified Leave-One-Out cross validation (MLOOCV), combined with K-nearest neighbours, is a simple and fast method compared to that of most learning algorithms, and produces better results for imbalanced data when compared to traditional leave-one-out cross validation methods.

We investigated the weakness of DNA methylation classes prediction, particularly imbalanced datasets within the sample, and their impact on direct usage of machine-learning and prediction procedures. It is noted that for sub-sampling of the minority class, and over-sampling of the majority class, some data is deleted and synthetic data is added to the minority class, respectively. Both methods have disadvantages; deleting part of the samples means that important information is ignored, whereas adding synthetic data is computationally-demanding. This thesis will employ Modified Leave-One-Out cross-validation, combined weighting and a cost-sensitive method incorporated within AdaBoost. Furthermore, Heatmap clustering and feature selection with Quadratic Discriminant Analysis will be considered for CpGs methylation prediction. In the next three chapters (4, 5 and 6) experimentally investigated and

discussed unsupervised cluster analysis was performed on gender and ageing methylation differences, the prediction of DNA methylation classes, and feature selection analysis for CpG loci positions that are specific to gender (via employment of the selected reported methods).

Chapter 4

DNA methylation dependence on Ageing and gender: investigation based on unsupervised clustering

This chapter presents the DNA methylation genome (detailed previous chapter section 3.1.1), employing a wide analysis by adapting a supervised clustering algorithm that uses a correlation metric to group CpG loci that are specific to gender and age. We have extracted and predicted CpG loci position methylation which is specific to gender. This point of methylation shows a significant difference between genders. Furthermore, author identified 47 CpG loci, denoting a specific gender, and which are further graphically displayed.

4.1 Introduction

Epigenetic regulation on the human genome is influenced by many factors, such as age, gender (tissue specificity) and environmental influences [176; 177]. DNA methylation is one of the most studied fields in the human epigenome. Genomic methylation is caused by environmental exposure and ageing, which changes gene regulation, without changing the DNA sequence itself [20]. This alteration is critical for normal cellular function and development; however, many lifestyle variables can affect CpGs methylation, including smoking, excessive alcohol use, diet and stress. These increase the rate of the epigenetic deregulation, whilst sport, a healthy diet, and physical fitness delay the epigenetic process [20; 21; 176]. It has also been reported that DNA methylation increases with age [63]. Epigenetics is a dynamic process, and it is not as fixed as previously thought but largely reversible. DNA methylation is a continuous process, which takes place over the long period of an individuals lifespan [89; 90; 176]. Even identical twins show variations in DNA methylation as they age [91]. It is well documented that women live longer than men, on average [60]. CpG methylations have

been associated with age-related illnesses such as metabolic disorders [23] and cancer [43]. Furthermore, DNA methylation disruption has also been linked with other complex age-related diseases, such as Alzheimers disease, developmentally-linked illnesses (autism), and mental illness (depression) [20; 24].

This has led researchers to investigate more closely the impact of age and gender on DNA methylation; particularly on that in CpG islands, and which is mostly found in the form of free methylation in normal physiological cell development; however, DNA methylation disruption causes disease. It is important to analyse, predict, and visualise the association between age and gender and DNA methylation, in both healthy and diseased samples. Other researchers have identified CpG island features, such as DNA sequence patterns, DNA structure and DNA physico-chemical properties [5; 48; 64]. However, these experiments failed to address CpG features specific to gender, and there are also limited reports on available age-related methylation features. It is not possible to determine if gene expression is regulated predominantly by genetic or environmental impact, since both are linked through the epigenome [92]. The alteration process, as it affects DNA methylation, is little known, although two main factors have been reported in the literature [91; 178]. These are: (I) environmental exposure that has taken place over a long period may cause cellular triggering with methylation changes, for instance, stress-related cases showing gene expression caused by methylation changes [24]; and (2) DNA methylation may occur spontaneously with or without an environmental influence, for example, an inherited genotype during cell replication. Both these cases make it difficult to determine whether methylation changes are associated with either age or gender. The process leading to methylation variation as affected by gender, as well as ageing-related methylation, suggests that an investigation of the wider CpG loci position could help locate the biomarker associated with methylation pathways.

This dynamic process makes changes very challenging to design, predict and model; in particular, attempting to understand whether epigenetic changes can determine the behaviour of both healthy and diseased tissue, and to distinguish differences between gender and ageing methylation in a variety of age-ranges, and to link these DNA methylation differences to healthy and cancerous individuals. While epigenetic changes associated with age have been studied to some degree, there is a shortage of reported studies on the relationship between gender and DNA methylation. It is also necessary to determine how nucleotides are associated with age and gender, and to understand whether a DNA methylation fingerprint can be correlated with ageing and gender. Author therefore compared genome-wide DNA methylation fingerprints from a large cohort of 963 samples across a wide age range for our study. Author designed unsupervised clustering algorithms, combining average linkage with a correlation coefficient, in order to group and extract CpG loci associated with males and females, in both healthy and diseased samples. The extracted CpG loci serve to provide further

understanding of epigenetic differences, and suggest that age and gender are associated with DNA methylation, and this explains how changes in gene expression over a longer timescale can be replicated in a further cohort with a large investigation.

4.1.1 Material and methods

Author extracted raw-data comprising 1628 [179] healthy and diseased human samples, containing 1505 CpG loci positions of extracted values from Illlumina hybridisation spot array of CYS3 and CYS5 methylated and unmethylated respectively, where the rows represent the ratios of the spot values of individual arrays (loci), and the columns represent the individual samples. These samples were analysed in 1505 spot arrays.

Before analysing the dataset, author excluded biological and technical bias, since these could be either tissue- or gender-specific. Author also removed data for which gender or age information was not stated and unavailable. Author included samples for individuals whose age and gender were both known, and samples that were not gender-specific. The remaining samples comprised 963 samples, of which 328 were healthy and 635 diseased. The samples were divided into two groups of healthy (control) and cancer samples. Each group contained samples from both genders. Author then analysed these groups separately using unsupervised heat map clustering for measuring the correlation coefficients, since these gave improved results over the Euclidean distance one. Comparing both methods, the correlation shows a clear data grouping. Table 1 shows the age range of the healthy samples.

Table 4.1 The age range for healthy samples

Age ranges	0-2	3-9	10-19	20-29	30-39	40-49	50-69	70-99	100+	Total
Male	16	0	3	9	13	13	69	41	4	168
Female	17	0	4	20	12	17	45	31	14	160

Table 4.2 The age range for cancer samples

Age ranges	0-2	3-9	10-19	20-29	30-39	40-49	50-69	70-99	100+	Total
Male	10	14	13	7	17	42	189	112	0	404
Female	7	7	6	6	14	11	111	69	0	231

Divisive clustering is applied and adopted with the average linkage combined with the correlation distance, since the latter shows empirically better results than Euclidean measurements. The Euclidean distance is measured by the distance between two CpG loci where A $= (a_1, a_2, ..., a_n)$, B $= (b_1, b_2, , b_n)$, n represents CpGs, and d is the distance between pair features (CpGs) which are adopted. In the following

formula adopted from [95; 98]:

$$d(A,B) = \sqrt{\sum_{n=1}^{n}(a,b)^2} \tag{4.1}$$

The Euclidean distance is combined with the average linkage as shown below:

$$D(A,B) = 1/A1/B \underbrace{\sum}_{a\in A, b\in B} d(a,b) \tag{4.2}$$

Correlation distance is sometimes used to judge how well two clusters align. This differs from the Euclidean distance, which can be calculated using the following formula:

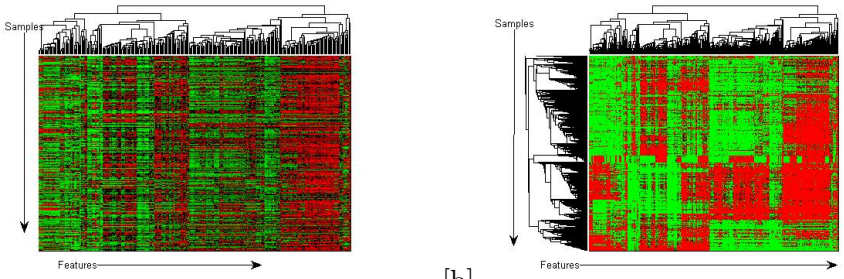
$$corr = 1 - \frac{(x_a - \bar{x}_a)(x_b - \bar{x}_b)}{\sqrt{(x_a - \bar{x}_a) - (x_a - \bar{x}_a)'}\sqrt{(x_b - \bar{x}_b) - (x_b - \bar{x}_b)'}} \tag{4.3}$$

$\bar{x}_{a,b}$ is set to the mean of the observation, and d($x_{a,b}$) represents the distance between two CpG loci. Then these two are plugged into the correlation formula, which becomes 1-correlation distance, which is generalised between 0 and 1. Furthermore, hierarchical clustering is applied based on average linkage. The correlation coefficient matrix data is computed, resulting in a dendrogram (a grouping of all elements into a single tree for any set of n_i loci). A similarities matrix is computed using a correlation coefficient as described above. This is based on the similarity score for all pairs of loci (arrays), in order to identify the highest values of these pairs. A node is assembled by joining these pairs of loci, and each profile of the node is computed by averaging the observation for joining features (locus). Similarities are then further updated with a new node replacing the two joined features. This process is repeated 2^{N-1} times until a single feature (dendrogram) is formed.This is a linear ordering through each node or tree to maximise the similarities of CpG loci methylation positions. Author used the male gender labelled as target data, then divided data into two sub-sets, the first one being ages between 0 and 50, and the second 51 to 106 years. This order is displayed by colouring the intensity of the methylation, so that the dendrogram relationship is assessed by colour intensity; red is used highly-methylated, green unmethylated, and colour intensities between 0.25 and 0.74 are differentially-methylated.

4.1.2 Results

DNA methylation variation associated with age and gender was displayed using unsupervised hierarchical clustering, with correlation distance and average linkage for the 1505 CpG loci positions. These loci positions correspond to the 963 human samples, comprising 328 and 635 healthy and cancerous samples respectively. Dividing the data into male and female, and age-groups between 0 to 50, and 51 to 100+ (this age-

division has given the best proportion of the sample for the analysis) for each sample and analysing comprehensive findings resulted in a significance of 47 CpG loci, which are specific to gender and age. Some of the CpGs relate to gender, whereas the others relate to age. Heatmap clustering renders it possible to group methylation differences between age groups in addition to between males and females. There is a significant association amongst these groups, and unsupervised heatmap clustering returned one group, and a single dendrogram labelled with both gender and age, based on a pair correlation of 1505 features. All the samples are effectively grouped, and the interested ones were focused on for further investigation. Figure (4.1) displays the healthy male and female samples, which were labelled with gender and age using two metrics: Euclidean and correlation distance. Both showed separated patterns (dendrograms), where features are grouped into red for methylated and green for unmethylated respectively, for age and gender.



[a] [b]

Figure 4.1 Comparison clustering metrics using a heatmap clustering display of CpG loci, where the rows represent nucleotide arrays (CpGs loci positions) of 1505 CpGs, and the columns represent individual healthy samples (328), [a] illustrates Euclidean distance, and [b] is a correlation distance display which shows clearer patterns than that in [a]

.

Since the correlation measurement was clearer in Figure 4.1(b) for the methylation display, author applied a correlation approach and further selected CpG loci patterns that showed clear separation, and further focused on and identified the CpG loci positions that contribute to the patterns. The importance of the method is that it groups the methylation similarities of CpG loci positions (features) associated with gender and also identifies which CpGs are methylated in particular CpG loci positions. Furthermore, Figure 4.2, an enlargement of Figure 4.1b, in which 22 features were displayed, shows that methylated and unmethylated features are clearly separated. These two distinct groups represent methylated (red) and unmethylated (green) classifications.

To investigate this more closely, the grouped features in Figures 4.3[b] and 4.3[c] represent the enlarged section of figure 4.3 [a], which displays clear feature clusters (heatmap) and 54 features extracted (CpG loci positions) from the total number of

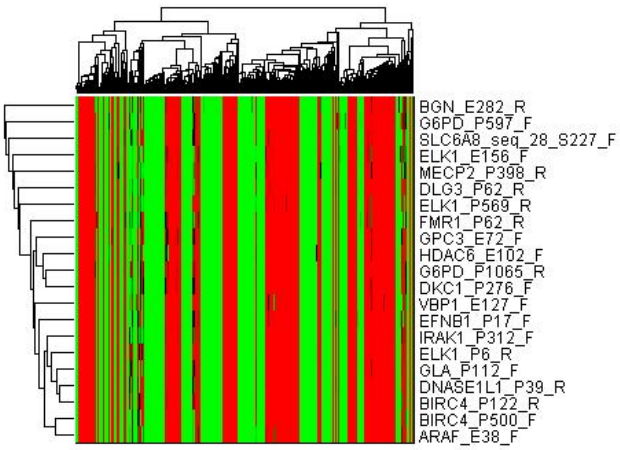


Figure 4.2 Cluster display of 21 CpG loci, that clearly shows methylation differences based on gender in 328 healthy samples. The methylation level is represented by colour intensity, where green represents unmethyalted, and red methylation groups; lower intensity colours represent the differential methylation classification.

samples. As noted in Figure 4.3[b] these features show a clear separation of patterns. This provide scope for further investigations to determine which age groups relate to the clusters; therefore, two clusters with four sub-clusters were selected, alongside the 54 features illustrated in Figure 4.3[c].

Moreover, the Figure 4.3[c] was reduced into two clusters, containing 17 samples. The features were also reduced from 54 to 47, giving two clearly-separated clusters, which are specific to gender in Figure 4.3[d]. These features (CpG loci position) have a mean methylation average level of 0.09 for males, and 0.65 for females; this represents a significant difference in average methylation values. For these age-groups, CpG loci are highly methylated in females, whereas males show a lower level of methylation. The methylation variation on gender is evident for very young ages (Figure 4.3[d]), the variation explored being the first visually displayed, particularly for healthy samples. Some of the features (CpG loci positions) were associated with genes responsible for growth and immune response. These 47 features indicate significant methylation differences between males and females, but the clustering method also grouped the samples by age-group. Although the features showed a different level of methylation between females, and males there were no distinguishable methylation differences relating to the different age-groups.

The results of these methylation differences and CpG loci for the 47 features are reported in Table 3. Furthermore, the median methylation averages are assessed with boxplot wiskers plots, where the two age-groups (n=17) show a compact distribution

methylation level in each cluster or age-group in Figure 4.4. This result was compared with another research reported on age-increase related methylation [176]. We found that the extracted features are less correlated with age; this confirms that the features are specific to gender.

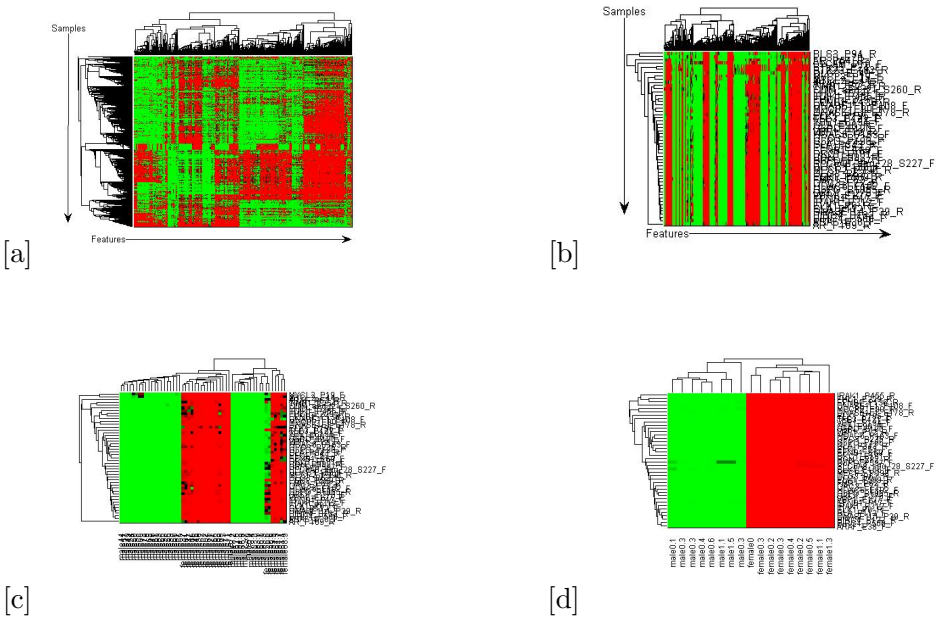


Figure 4.3 [a] CpG loci methylation patterns of 1505 features, where the columns represent complete samples of 328 healthy individuals. [b], an enlarged section from [a] of 47 CpG loci with all samples. [b] emphasises the CpG loci that shows clear pattern separation. [c] represents an enlarged section of two compact clusters from [b] of the 47 CpC loci, whereas [d] illustrates two compact clusters of 17 samples, which are extracted from [c], and in which the 47 features of males are unmethylated, whilst the female samples are highly methylated.

Subsequently, author extracted the same features from both genders, comparing the same sex feature in the control and cancer samples Figure 4.5 displays heatmap clustering of features for representation of cancer and healthy samples from same sex individuals these were shown to have clear pattern variations, as expected (Figure 4.5); healthy samples showed methylation variation (Figure 4.3). In addition, the plotted feature values for males show a low level of methylation, whereas the same features values from females show moderate-to-high methylation values (Figure 4.5). Whilst the methylation variations may not be strongly associated with wide de novo methylation on the X-chromosome, this result reveals an association of methylation differences with gender, age and cancer-related methylation.

To gain more information regarding whether the selected features were associated with either gender or age, author compared the results with previously reported results [176; 180; 181]. Author found that three-quarters of the CpG loci showed a slight methylation increase for those of older age, for both males and females. In addition, DLG3E340F was highly methylated in the cancer samples and unmethylated control

Table 4.3 47 extracted CpG loci position, which show significant methylation differences between males and females in healthy samples

Extracted features–ID	met-Av-mal	Met-Av-fem	met-dif	fem/mal-ratio
AR-P189-R	0.1	0.37	0.27	4
ARAF-E38-F	0.08	0.81	0.73	11
BIRC4-P500- F	0.11	0.76	0.66	7
BIRC4-P122-R	0.05	0.77	0.72	14
DNASE1L1-P39-R	0.1	0.76	0.66	7
GLA-P112-F	0.08	0.83	0.75	11
ELK1-P6-R	0.08	0.77	0.68	9
IRAK1-P312-F	0.08	0.69	0.61	8
EFNB1-P17-F	0.24	0.77	0.53	3
VBP1-E127-F	0.08	0.78	0.7	10
DKC1-P276-F	0.07	0.71	0.65	10
G6PD-P1065-R	0.15	0.8	0.65	5
HDAC6-E102-F	0.1	0.79	0.69	8
GPC3-E72-F	0.09	0.69	0.6	8
FMR1-P62-R	0.04	0.65	0.61	17
ELK1-P569-R	0.07	0.61	0.54	9
DLG3-P62-R	0.18	0.67	0.5	4
MECP2-P398-R	0.09	0.52	0.43	6
ELK1-E156-F	0.13	0.48	0.34	4
SLC6A8-seq-28-S227-F	0.05	0.56	0.51	11
G6PD-P597-F	0.36	0.72	0.36	2
BGN-282-R	0.16	0.71	0.56	5
DKC1-E101-F	0.06	0.48	0.43	8
EFNB1-E69-F	0.08	0.71	0.63	9
ELK1-E53-F	0.1	0.73	0.63	8
GLA-P343-R	0.06	0.7	0.64	11
G6PD-P196-F	0.06	0.56	0.51	10
GPC3-P235-R	0.06	0.71	0.65	12
HDAC6-P153-F	0.05	0.58	0.53	11
VBP1-P12-R	0.07	0.48	0.41	7
G6PD-E190-F	0.07	0.56	0.49	8
GLA-E98-R	0.04	0.4	0.36	9
VBP1-P194-F	0.06	0.61	0.55	10
TFE3-P421-F	0.05	0.49	0.44	10
ELK1-P195-R	0.04	0.45	0.41	12
DNASE1L1-E178-R	0.15	0.53	0.39	4
MECP2-E90-R	0.15	0.44	0.29	3
DNASE1L1-P108-F	0.06	0.61	0.56	11
EFNB1-P136-R	0.08	0.53	0.45	7
FHL1-E229-R	0.05	0.5	0.45	10
IRAK1-P455-R	0.04	0.55	0.51	14
FHL1-P768-F	0.09	0.42	0.34	5
CDM-seq-21-S260$_R$	0.05	0.59	0.54	11
TIMP1-E254-R	0.11	0.63	0.51	5
ARAF-P63-R	0.04	0.32	0.28	8
MYCL2-E44-R	0.05	0.47	0.42	9
MYCL2$_P$19$_F$	0.08	0.55	0.47	7
DLG3-E340-F	0.07	0.55	0.48	7
STK23-E182-R	0.8	0.94	0.14	1
STK23-P24-F	0.63	0.89	0.26	1
L1CAM-P19-F	0.07	0.53	0.47	8
AR-P54-R	0.04	0.16	0.12	4
PLS3-E70-F	0.09	0.18	0.09	2
PLS3-P94-R	0.11	0.27	0.16	2

Abbreviations:met-Av-mal (methylation average for male), Met-Av-fem (methylation average for female),met-dif (methylation differences) and met-ratios (methylation ratios for male to female).

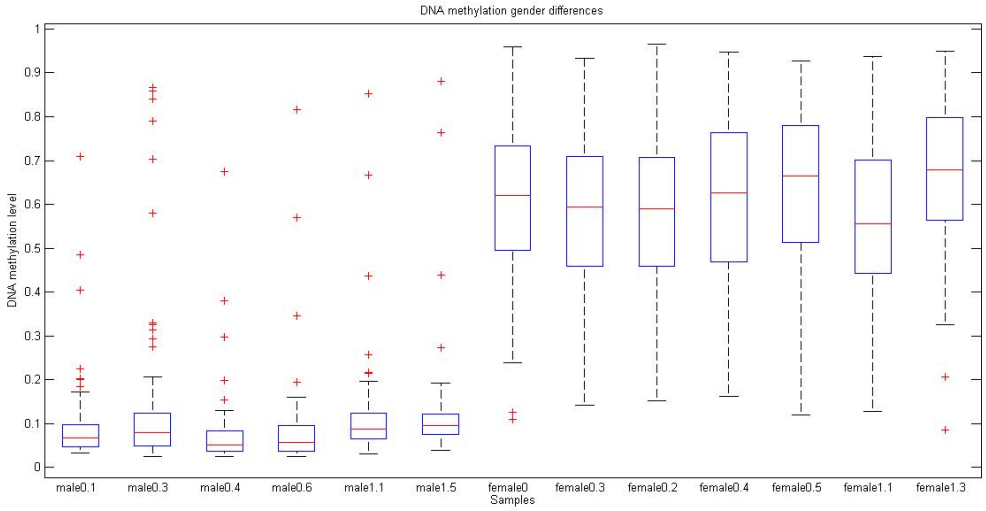


Figure 4.4 Boxplots displaying the methylation distribution of the 47 CpG loci positions of 17 healthy samples, which are extracted from Figure 4.3[d]. This methylation distribution indicates that females are more highly methylated than males.

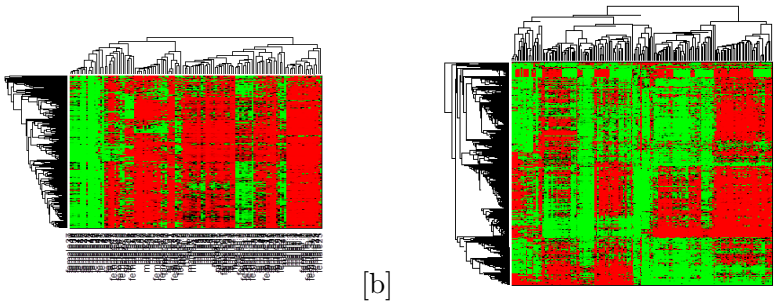
[a] [b]

Figure 4.5 Unsupervised cluster analysis comparison of healthy and cancerous samples, and their methylation differences. This cluster display represents the same features of cancer versus control for the same gender of CpG loci positions. [a] is cancer *versus* healthy male comparison, and [b] represents cancer *versus* healthy female one. Males show a lower CpG methylation status compared to females.

ones, whereas SLC6A8 showed three times more methylation in cancer samples, particularly older samples; however, SLC6A8 methylation is associated more with gender than age. Moreover, RAD54BP227F showed a small decrease in methylation level with age for both males and females.

4.1.3 CpGs methylation level for different age-groups

This sub-section investigates whether the age-group in the previous section overlaps; data was divided into two age-groups between 0 to 50 and 51 to 100+ in each sample.



[a] [b]

Figure 4.6 CpGs methylation comparison of two age-groups using heatmap clustering algorithms combined with average linkage and correlation as the distance metrics. Each row represents nucleotide arrays (CpGs loci positions), and the columns represent individual healthy samples. [a] represents age-groups between 0 and 50, and [b] between 51 and 100+. Green represents unmethylated, red represents methylated; colour is ordered by intensity based on the correlation coefficient, where less intensity colour is assigned as the differentially-methylated form.

The analysis was conducted following the same approach as that of the previous section, and the results are displayed in Figures 4.6 to 4.10. Figure 4.6 illustrates the results of the healthy samples between 0 to 50 years for [a], and 51 to 100+ years for [b].

All the features were clearly separated, and after multiple sub-group comparisons, two sections from the 0 to 50 and 51 to 100+ year age groups were selected. These sections were further enlarged and displayed in Figures 4.7 and 4.8 respectively. 22 CpG loci were extracted; these are the most separated features, showing significant methylation differences in both gender and age groups. There is a substantial correlation between ageing and methylation, as illustrated in Figure 4.7[c] and [d]. The 22 CpG loci positions were unmethylated for the middle age group with the average age distribution of the samples being 10 years (figure 4.7 [c]), whereas the same features were highly methylated for the very young age-group (Figure 4.7[d]). These results revealed that there were some patterns distinguishing the genders for the middle age-group, whilst the younger age-group revealed a separation between males and females, in which 14 out of 18 (70%) of the group were males. The middle age group has a slightly larger age distribution than the younger ones; the three younger age ranges are mis-grouped into the middle age-group. Hence, author can identify methylated and unmethylated CpG loci positions (features) using heatmap clustering, where features were grouped with respect to age and gender. This has been reported to a limited extent in a previous study [63]. The 22 selected loci will be further explored, to determine their biological usefulness. To characterise the features which play an important role, we enlarged the dendrogram in order to display their correlations.

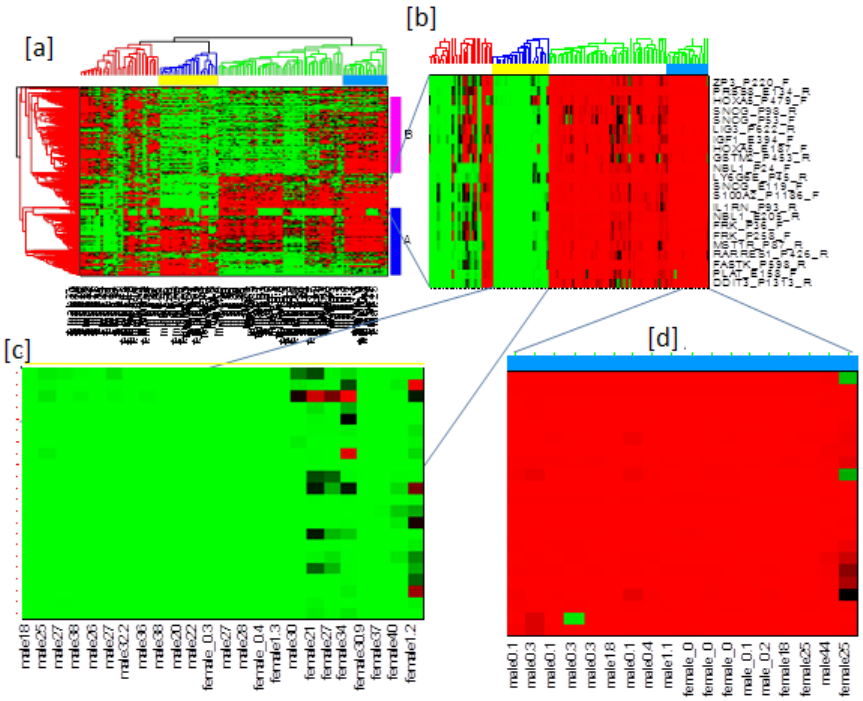


Figure 4.7 Two-dimensional representation of an unsupervised learning cluster analysis of 125 healthy samples of both genders (age-group 0 to 50 years). [b] illustrates 22 extracted features (CpG Loci), and [c] and [d] show enlarged sections of methylated and umethylated CpG positions, respectively; this shows clear separation of methylation patterns between both age-groups and genders.

Furthermore, author compared these in the same manner noted above for samples of the age range between 51 and 106. Overall, the results showed that increases with ageing methylation. 33 features were extracted, which shows a clear separation of three methylated, unmethylated and differentially methylated classes (Figure 4.8). Furthermore, three age-groups were selected, in which the 33 loci were unmethylated (Figure 4.8[b]), differentially methylated (Figure 4.8[c]) and highly methylated (Figure 4.8[d]); these have an average age distribution of 59.8, 64.9 and 77 year, respectively. The sub-selected features show a clear natural grouping. Therefore, correlation and unsupervised learning are able to distinguish methylated from unmethylated features; these also show gender and age differences, where 22 loci are consistently methylated in a younger age-group compared to the older one (Figure 4.7[c] and [d]). The number of CpG loci that show methylation increase with ageing; these are NBL1, LY6, S100A2, Il1NR, FRK,MST1R and FASTK, while HOXA5, SNCG, LIG3, IGF1 and GSTM2 show methylation decreases as age increases. This demonstrates that author can identify methylated and unmethylated CpG loci positions (features) using heatmap clustering based on grouping with respect to age and gender.

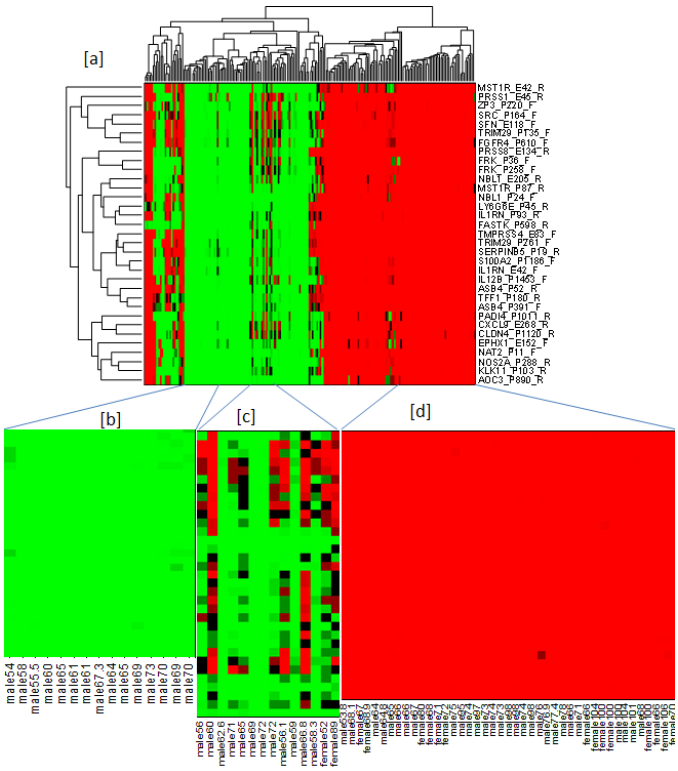


Figure 4.8 Two-dimensional representation of unsupervised learning cluster analysis of 33 features (CpG Loci) of healthy samples of both genders (age-group 51 to 100+ years). [b], [c] and [d] show enlarged sections of umethylated, differentially-methylated and methylated samples, respectively. This provides evidence that CpG methylation increases with ageing.

To gain more information regarding the cancer samples, author used the same

approach employed in the previous section; the two age groups are illustrated in Figures 4.9 and 4.10. As notable, the feature clustering for the cancer samples, and their methylation patterns are greatly changed, i.e., they are disorganised when expressed relative to the healthy samples (Figure 4.6). In addition, the cancer samples showed no clear groupings with clustering analysis (Figure 4.10), unlike the healthy samples (Figure 4.8). 21 methylated and 23 unmethylated features were selected from the two sections that contain 15 cancer samples. Furthermore, author investigated age range in relation to differences in DNA methylation and found that some of the features were either entirely unmethylated or unselected (unbound hybridisation) in cancer sample alleles, whereas the healthy samples were completely methylated. However, cancer samples show differential methylation regardless of age, whereas the methylation pattern is scattered throughout the heatmap (clustering space), as illustrated in (Figure 4.10), which shows the same features as those methylated in cancer samples and unmethylated as reported in Table 3.

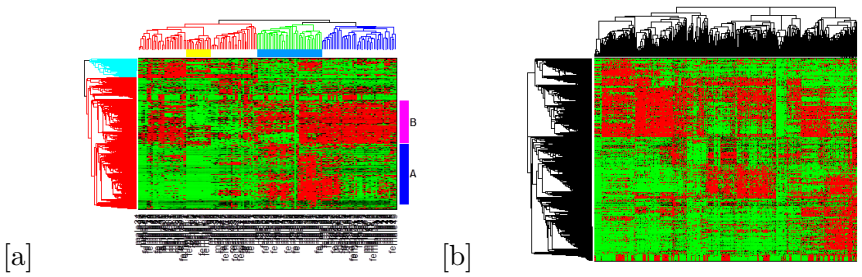


Figure 4.9 Representation of two-dimensional unsupervised learning cluster analysis of 635 cancer samples. Rows represent 1505 CpG loci, and columns for cancer samples (635). [a] is a representation of age between 0 and 50, and [b] that between 51 and 100+ years
.

On average, CpGs methylations are reduced in cancer samples. However, these features are specific to gender, which show methylation level differences. This indicates that current medication, such as chemotherapy, may not exert the same toxicological effect for males and females; existing medication, medical doctors give cancer patients who have similar symptoms (the same type of cancer) the same treatment without considering methylation gender differences, which may exert drug resistance effect or a greater toxicity impact.

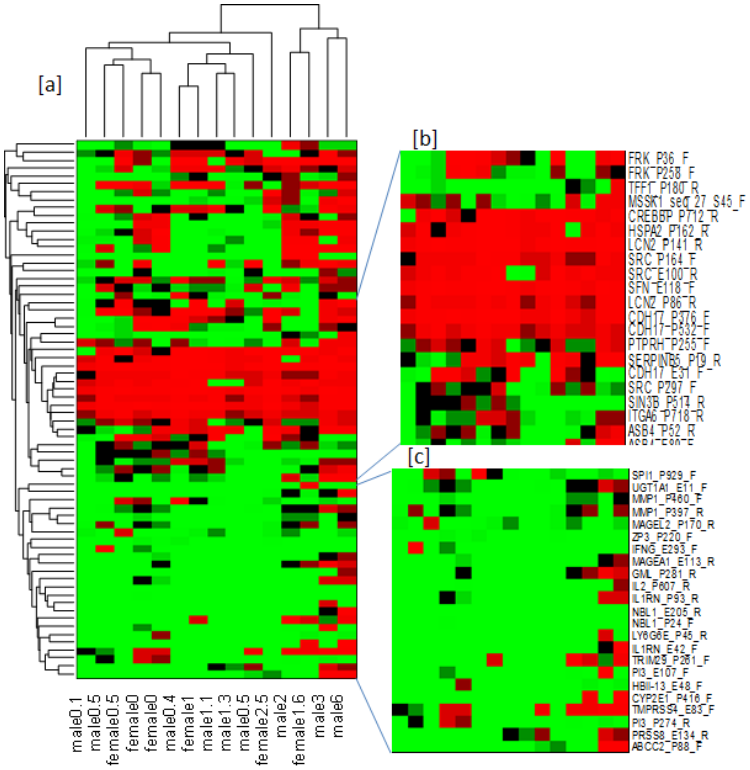


Figure 4.10 Two-dimensional unsupervised cluster analysis of 15 cancer samples. Individual samples were labelled gender and age (between 0 to 50 years); [b] is an enlarged section from [a], and which contains 21 methylated CpGs; [c] is the enlarged section from [a], which contains 23 methylated CpGs positions.

Table 4.4 CpG loci position: methylated in healthy samples and unmethylated cancer samples

features	Healthy samples	Cancer samples
FRK–P36F	M	Un
FRK–P258F	M	Un
ZP3–P200F	M	Un
IL1RN–P93–R	M	Un
NBL1–E205–R	M	Un
NBL1–P24–R	M	Un
LY6G6E–P45–R	M	Un
PRSS8–E134–R	M	Un
TRIM29–P261–F	M	Um
CDH17–P376–F	M	Un
ABCC2–P88–F	M	DM

methylated (M) and unmethyalted (Um)

4.2 Discussions and conclusions

CpG methylation classification and profiling are essential for understanding natural cellular development and differentiation. Therefore, the study of a wide normal human epigenome with a variety of age ranges is an important venture. However, this work is more challenging than employing data in a machine-learning context, in view of the nature of epigenetic data, its tissue-specificity and methylation dynamic nature. Hence, a correct understanding of what entails normal DNA methylation, and how DNA methylation differs based on gender and ageing, has the potential to advance our understanding of CpG methylation changes in disease intensely. Hence, author identified methylation differences between males and females, and CpG loci positions are specific to ageing together with those possibly associated with cancer. Annotating the CpG loci methylation position in normal human DNA sequences will increase elucidation of the DNA methylation risk development in disease states. Employing heatmap clustering, average linkage and combined correlation metrics, author identified CpG loci positions associated with gender, age and cancer. These features were validated with other recent studies [56; 63; 71; 72; 176]. Indeed, these resourced advanced and high-resolution data for DNA methylation profiling, but the researchers did not address gender methylation variations in the normal human epigenome, and a comprehensive analysis of ageing-related methylation was not conducted.

Factors reported as contributing to DNA methylation include ageing, types of cancer, inflammation, and carcinogen exposure such as tobacco, arsenic, diet, alcohol and asbestos [176; 177; 182]. It is found that methylation variation can be suggested by gender; thus, throughout the lifespan of males and females, methylation modification occurs, which can exert an impact on the risk of disease, prevention and medication. In addition, environmental exposure induces CpG loci methylation alterations to normal appearing tissues (genes), which leads to a modification in physiological function. This work suggests that methylation changes do not occur only in specified and normal tissues, but are associated with ageing and gender, which is of interest to our future study. Methylation variation related to gender was the initial research finding related to normal tissue, which motivated this study. Another motivation was that ageing-related methylation occurs in normal tissues since cancer is associated with ageing [183; 184]. The CpG islands loci methylation increases were reported in normal prostate and colon tissues [180; 185]. Moreover, methylation variations linked to gender were reported[177]. Our findings confirm these results, and in addition, they show that CpG methylation differences between males and females depend on age, and confirm that there is age-related methylation variation in the T lymphocytes of new-born, elderly and middle aged human [186]. Heatmap clustering on CpG loci shows a clear grouping of methylation differences between males and females, as well as methylation age-related features. Author reported that CpG loci position methylation are asso-

ciated with gender and ageing. Author also identified CpG loci methylation states in normal human tissues specific to gender and ageing (Table 3.3), a phenomenon which has also been confirmed in previous studies [176; 186]. The comparison of CpG loci in with that from a previous study (Table 3.3) showed that this work had more features associated with ageing, and our results confirm this finding; author also extended the number of samples and illustrated them. Non-tissue-specific CpG loci methylations are found to be highly associated with ageing, suggesting the existence of a common mechanism to elucidate methylation changes. This common mechanism could be explained statistically, since methylation increases with age, whilst a lack of methyl-transferase (enzyme),which maintains CpGs methylation, decreases with ageing. This unstable methylation status may significantly risk genomic instability which may cause cancer and other illnesses.

Author identified and patterned CpG loci (features) methylations associated with gender differences that increase with age and cancer across 1505 CpG loci positions of 963 samples. Author illustrated and annotated unidentified CpG methylation differences on gender, and age-related CpG methylation, which were found to be non-tissue specific. Author contributed to understanding of the CpGs methylation differences associated with gender, and also methylation changes relating to ageing in normal tissues. Unreported methylation differences between males and females, and age-related methylation modifications represent a substantial contribution to our fundamental understanding of methylation process for both normal and diseased tissues, since these CpGs methylations are linked to age-related illnesses, such as cancer and Alzheimers diseases, as well as mental-related problems such as depression and Autism, which provide an initial pursuit of biomarkers or disease susceptible ones, which may lead to useful clinical detection and prevention processes.

DNA methylation is essential for biological processes; and for the representation of methylation differences between genders, and this is very important for personalised drug design, since the methylation differences shown in Figure 4.5 are not just grouping features (CpG loci position), but comprehensively represent one of the most important features associated with the large human methylation fingerprint. Furthermore, methylation experiments on the human genome will assist in a comprehensive state-of-the-art study of genome DNA methylation, in order to combine extracted features.

Chapter 5

Analysis and prediction for DNA methylation sequence driven features

This chapter contains two sections, and hence two objectives. The first objective was to predict and explore the methylation classes based on DNA extracted feature subsets of four feature subsets from chromosomes 6, 20 and 22 (as detailed in chapter 3, section 3.1.1), seven feature subsets from chromosome 21, and four feature subsets extended from chromosome 21. A novel method (modified leave-one-out cross validation) was employed, which generated an improvement in results that is reported on further in the first section. The feature subsets were grouped according to their biological meaning, and hence were combined in all possible combinations during analysis. The aim of the investigation was not only to select a few feature subsets, but to further investigate the biological feature usability for methylation classes of extracted subsets, leading to a comprehensive feature subset analysis for the prediction of DNA methylation classes (although these analyses identified the most informative feature subsets by employing Modified Leave-One-Out cross validation). This model was based on preparing the dataset in order to adapt to machine-learning, where the imbalanced data were rebalanced and then analysed.

The second section investigates the same dataset, which provide a fair predictive model for imbalanced classes, by employing Adaboost combined with a cost-sensitive method (Section 5.4); this is the extended method of the current predictive ones based on sub-sampling and oversampling datasets, and which are disadvantaged in data analysis.

5.1 Introduction

DNA methylation is an inheritable biochemical modification of eukaryotic DNA, which generally occurs at the fifth [C5] position of cytosine's phospho-guanine residue in a 5-CG-3 biomolecule known as CpG dinucleotide [20; 21; 30]. This modification represses an activity of transcription site [22; 30]. In vertebrates, cytosine residue methylation in CpG nucleotides is an epigenetic marker that is necessary for physiological cell differentiation [20; 30]. Indeed, more than 60% of DNA sequence composition in promoters are CpG islands, which are generally unmethylated CpG islands [19]. Recent genomic analyses have shown that more than 80% of CpGs are methylated [72]. However, the remaining small fraction - the CpG islands at cell differentiation [20], are unmethylated. It has been reported that DNA sequence composition and length act as the backbone of DNA methylation regulation [187]. This is, however, in contrast to some extent by other researchers [188], who argue that active chromatin, the transcription start site, and also environmental influences are the most important factors in methylation regulation.

The prediction of DNA methylation is the one of the most complex and challenging problems in bioinformatics, since DNA sequence features that characterise methylation, in particular CpG islands, are dispersed throughout the human genome, and are mostly concentrated in the promoter area of most genes [20; 189; 190]. However, advances in technology in computational genomics and epigenomics has helped analyse a large amount of data obtained from methylated, unmethylated and differentially-methylated DNA of CpG islands. Methylation of CpG islands is mainly involved in various biological processes, such as gene silencing, structural chromosomal stability, parental imprinting, and suppression of the mobility of retrotransposons [20; 21; 22]. The disruption of DNA methylation has also been linked to various human diseases, such as cancer [11; 12; 21]. It should also be noted that, despite all advances to date, the analysis of DNA methylation, particularly for the human genome, remains a challenging problem.

DNA methylation profiling and comparative analyses are very important for understanding this process in relation to DNA composition (context). Many researchers have profiled epigenetic data and made them available in biological databases for further computational study [33; 56; 63]. Yamada [33] studied the methylation patterns of CpG islands on chromosome 21q and identified 147 CpG islands. Of these, 103, 29, and 15 were found to be methylated, unmethylated, and differentially-methylated, respectively. In addition, CpG islands of three human chromosomes (6, 20 and 22) have been profiled from 43 samples of 12 different tissues [56]. This is discussed further in the material and methods section of this study. These data were extended and 50 features were extracted and analysed [64]. Other researchers have also made CpG island predictions [41; 46; 191; 192]. However, these studies are limited, since

they examined only small sets of features, which provide only an incomplete view of human DNA methylation. Bock [48] extended the Yamadas[33] study by extracting DNA-sequence features associated with CpG islands, and analysed the data using statistical methods. Both studies [48; 64] have improved our current understanding of DNA methylation features that are based on sequence-driven features. However, the amount of information that can be obtained from both analyses is limited. In addition, the statistical approaches used for the analysis are required to be improved in view of the nature of such complex data. It was decided to extend these predictive models in order to extract meaningful biological information from these four human chromosomes. The aim of this study is therefore to develop a statistical strategy and to carry out a detailed and comprehensive analysis of the features for more accurate and reliable predictions of methylated, unmethylated and differentially-methylated CpG islands.

DNA sequence-specific factors and epigenetic modification have been reported to play a major role in the DNA methylation process [63]. However, these studies have not produced a comprehensive predictive model that would enhance understanding of the regulation of DNA methylation. In order to gain more systematic information from DNA sequence patterns, a comprehensive predictive model was developed to distinguish methylated from unmethylated and differentially-methylated classes. The feature subsets that are more closely associated with DNA methylation than the other two subclasses (unmethylated and differentially-methylated) were also determined. It was shown that tissue-specific CpG islands, DNA sequence properties and distribution, and exon and gene distribution are significantly associated with DNA methylation. These findings, are further reported in the following sections.

5.2 Materials and method

5.2.1 CpG islands Data

DNA sequence data were collected from publicly available literature [48; 64]. These data contain 642 samples from four human chromosomes. DNA features were then calculated using different methods and prepared for analysis. After data were filtered, data used throughout this chapter comprised 470, 113 and 59 samples of unmethylated, methylated and differentially-methylated DNA respectively, which are summarised in Table 5.1. In order to characterise the DNA sequences, a set of features was extracted from DNA sequences. The extracted features are also summarised in Tables 5.2 and 5.3. Data shown in Table 5.2 contain 495 samples of CpG islands.

These were obtained from three human chromosomes (6, 20 and 22) derived from 43 samples from 12 tissues from healthy individuals [56]. The averaged methylation changes between CpG pairs of identical samples were calculated in order to minimise

Table 5.1 Details of the CpG island samples

Chromosomes	Methylated samples	Unmethylated samples	Differentially-methylated samples
6	12	125	8
20	6	23	12
21	29	103	15
22	66	219	24
Total	113	470	59

any bias produced by the differences in the length of sequence windows.

Table 5.2 Details of the CpG islands feature-sets for chromosomes 6, 20 and 22

Extracted features	No of features
1. Tissue-specific CpGI methylation	12
2. Evolutionary and conservation	9
3. sequence distribution (Dinucleotide)	16
4. DNA structure and properties	13
All the feature sets listed above	50

As shown in Table 5.2, 50 features were extracted and further divided into four subsets. These can be described as follows: subset 1 (1: CpGI-specific DNA methylation) contains 12 attributes and averaged sequence values calculated using CpGcluster algorithms [97]. These are CGI-specific attributes (CG contents, CG%, number of CpGI, observed/expected ratio, CpGI distance, and CpGcluster-pvalue). Subset 2 (2: Evolution and conservation) contains nine attributes of phase conservation contents, calculated by the number of CpGI overlapping with elements of phase conservation per CpCI using a log-odds conservation score of 100 or more without repeat masking. Subset 3 (3: CG distribution) contains 16 attributes and represents a score of 16 possible combinations of its observed/expected ratio. Subset 4 (4: structural and physiochemical properties) contains 13 attributes, and includes predicted elements, such as rise, roll, tilt, twist and solvent-accessible surface area, as well as bending, curvature, stacking energy, turns, degree of twist, DNA cleavage, bases per turn and six helical force constants. Calculations of the features were performed using DNA live algorithms [62].

These data (Table 5.3) were extracted from 147 samples of CpG islands derived from chromosome 21 of peripheral blood leukocytes, or the placenta of four human healthy individuals.These methylation changes between CpG pairs of identical samples were also averaged in order to minimise any bias produced by differences in the length of sequence windows.

As shown in Table 5.3, 3,759 features were extracted and further sub-divided into seven subsets, which can be described as follows: Subset 1 (DNA sequence proper-

Table 5.3 Details of the CpG island feature-sets for Chromosome 21

feature-sets	No of features
1. DNA sequence properties and distribution	2870
2. CG distribution	112
3. CpGI/tissue-specific distribution	91
4. DNA structure	196
5. Exon and gene distribution	364
6. Evolutionary and conservation	70
7. SNP	56

ties and distribution) comprises 2,870 attributes and is of the highest dimension. It contains a frequency average score of all possible combination tetramers (both specific and non-specific strands); Subset 2 (CG distribution) contains 112 attributes and represents a score of Cs, Gs and CpGs (its observed/expected ratio); subset 3 (the distribution of CpG islands) contains 91 attributes extracted taking the distribution of CpG islands into consideration; subset 4 (predicted DNA structure) contains 196 attributes and includes predicted elements, such as rise, roll, tilt, twist and solvent accessible surface area; subset 5 (exon and gene distribution) contains 364 attributes and was selected from human genome, high-confidence gene annotation from the consensus CCDS; subset 6 (evolution and conservation) contains 70 attributes and is a calculated average of $CG \geq 50\%$ and a length greater than 400bp of observed /expected ratio without repeat masking; and subset 7 (Single Nucleotide Polymorphism SNP) contains 56 attributes, which was calculated from the UCSC genome browser.

5.2.1.1 Predictive method: K-Nearest Neighbour Classifier (K-NN)

The K-nearest neighbour (K-NN) classifier is one of the most popular non-parametric classifiers, and has been successfully applied to various problems in bioinformatics [52; 53; 111]. It assigns the point at which the majority label among its nearest k-neighbours in the training data points to x, and predicts the class-label of x based on the label that k points to. Increasing the k value to show reduced bias and decision boundaries becomes rather smooth and less sensitive to outliers [52; 53]. It has been reported in some studies that KNN resulted in a higher predictive accuracy than that of Support Vector Machine, which is one of the most powerful methods available [111]. However, it should be noted that in many cases, the success of a predictive method is mainly based on a characteristic of a dataset being analysed. For this study, in view of its flexibility, effectiveness and power, KNN is adapted, together with the modified leave-one-out cross validation method, which was previously used successfully to address the imbalance in intraclass problems, and also to give better predictive accuracy [5; 26].

5.2.1.2 Modified Leave-One-Out Cross Validation

Cross-validation has been assessed for predictive models of machine learning. For small datasets, which is the case in this study, leave-one-out has been widely used. The M-fold cross validation method is also found to be satisfactory for various data sample sizes. However, when the dataset is quite unbalanced, which is the case in this study, these two methods have been found to be biased towards the class with the highest number of samples, and this could lead to a misleading interpretation [48; 64]. Therefore, in this study, a modified leave-one-out cross-validation was employed, incorporating the KNN-classifier. To clarify, small samples (66 samples in the methylated group and 24 differentially-methylated) were kept constant, whereas the unmethylated data were randomly divided into 20 folds of equal size of both small samples, the 20 different models and predictive accuracies then obtained. These 20 divisions were then analysed in a single feature-sub-set (each feature subset consisted of either two or three classes), and all possible combinations were calculated using a modified LOO cross validation with KNN classifier experimentally selected for k values between 1 and 11.

5.2.1.3 Predictive and technical evaluation methods

It was subsequently intended to examine whether or not these combinations of features can be predicted when two or more features are combined, a process an examination of which feature subsets are associated with DNA methylation in terms of its sequence patterns.

- Divide unmethylated samples into M equal parts, for example unmethylated X_u and methylated X_m.
- Divide X_u into 20 equal parts, so that each of these parts are equal to that of X_m, where X_m is a constant.
- Composite the X_u and X_m groups to build the new M-fold predictive models, where the number of methylated and unmethyalted fraction/groups are equal.

After this step, each X_u was added to X_m (constant). The composite sub-set of balanced data were then applied to KNN with leave one out cross-validation, which is based on similarity according to their distance in order to determine out the K nearest neighbours which predict the outcome, specifically building M independent predictive models, which create different M averages and standard errors. The highest predictive model was then selected whilst eliminating the lowest ones.

All possible combinations of the subset models were applied in order to select the best predictive model and its combination. Furthermore, predictive accuracies and class performance were based on four statistical terminologies [139], which were adapted into the predictive model and assigned as follows:

- TP = number of times X_i is methylated, and D_i is estimated (true positive).
- TN = number of times X_i is unmethylated, and D_i is estimated (true negative).
- FP = number of times X_i is unmethylated, and D_i is predicted as methyalted (false positive).
- FN = number of times X_i is methylated, and D_i is predicted as unmethyalted (false negative).

Predictive accuracies and class performance were measured for four statistical measurements of TP, TN, FP and FN, which is usually given as a percentage.TP value alone does not give useful information and cannot be trusted without combining TN. Only a combination of two of these four measurements can be viewed as an effective prediction or performance for causing bias, since they are dependent on the other two.

Accuracy represents the proportion of both correctly identified results; a true positive and true negative for methylated, unmethylated and differentially-methylated. Furthermore, the Matthews correlation coefficient (MCC) assesses the performance of two-class problem predictive models for a single and all possible feature sub-set combinations calculated by formula 2.4 adapted from [139] from [140].The MCC calculation uses four terms driven from the confusion matrix (Table 2.2) in order to assess the performance of the two-class (binary) classification.

5.3 Results and Discussion

This result compromises all possible combinations (4x120) of the four biological feature subsets obtained from direct or indirect DNA sequences of three human chromosomes and seven-feature subsets from chromosome 21, as listed in Tables 5.2 and 5.3. A total of 2,000 analyses were carried out, and the predictive accuracies obtained are summarised and presented in Tables 5.4 and 5.21.

5.3.1 Data analysis for chromosome 6

5.3.1.1 Methylated and unmethylated fractions of chromosome 6

The association between methylated and unmethylated classes based on DNA sequence features were examined as noted in Table 5.2. The single feature set prediction and tissue specificity subset showed the highest accuracy of 100%, as well as a Matthews correlation coefficient value of 1.00. The other three individual subsets indicated a low total predictive accuracy of between 37.50% to 62.50%. The accuracies of these three individual feature subsets, evolution and conservation, dinucleotide distribution, DNA structure and physicochemical properties, are 37.50%,

62.50% and 62.50%, and their standard errors are 17.60%, 5.89% and 17.68% respectively. The standard error was calculated from a 20-fold 'modified leave-one-out cross-validation' of both the sensitivity and specificity of each subset, as well as all combinations throughout the analysis, as noted in Table 5.4. Both sensitivity and specificity fluctuated, despite the DNA structure subset showing improved sensitivity.

Table 5.4 Results (% correctly classified) for the analysis of methylated and unmethylated classes of chromosome 6.

combined feature-subsets	No of features	acc(M)	acc(uM)	total acc	st-error	MCC
1	**12**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
2	9	50.00	25.00	37.50	±17.68	-0.17
3	16	58.333	66.67	62.50	±5.89	0.25
4	13	75.00	50.00	62.5	±17.68	0.26
{1,2}	21	100.00	87.50	95.00	±8.84	0.85
{1,3}	28	100.00	100.00	100.00	0.00	1.00
{1,4}	25	100.00	100.00	100.00	0.00	1.00
{2,3}	25	57.50	85.00	71.25	±20.04	0.43
{2,4}	22	65.42	75.42	68.65	±7.66	0.42
{3,4}	29	64.58	69.17	66.88	±9.72	0.12
{1,2,3}	37	100.00	100.00	100.00	0.00	1.00
{1,2,4}	34	100.00	100.00	100.00	0.00	1.00
{1,3,4}	41	100.00	100.00	100.00	0.00	1.00
{2,3,4}	38	100.00	100.00	100.00	0.00	1.00
{1,2,3,4}	50	100.00	99.58	99.79	±0.29	0.99

Abbreviations: acc(M) = average predictive accuracy for methylated class; acc(uM) = average predictive accuracy for unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold are those showing the highest predictive accuracies and correlation coefficients.

However, these three individual subsets show a low correlation coefficient, particularly the evolutionary and conservation subset. Of the two subset combinations, subset (1) showed predictive power when it was paired with other individual subsets, which gave a predictive accuracy of 95% to 100%. However, when subset (1) was excluded from the analysis, the other pairs showed a reduced accuracy and MCC. However, the accuracy and MCC value were slightly better than the single subset results. It should also be noted that 100% results of tissue-specific (1) feature subset may be overfit; however, it is not likely in this case since it randomly repeated the analysis and also shows combinations with other subsets a significant increase of predictive accuracy; hence the tissue-specific feature subset is the most informative one for DNA methylation class prediction. The total accuracy of the subsets (2,3; 2,4 and 3,4) was 71.25%, 68.65%, and 68.88% respectively. Furthermore, subsets (2,4) and (3,4) showed the least accuracy amongst the pairs, while subset (3,4) gave the lowest correlation compared to the other two subset combinations. Three subset combina-

tions showed the highest predictive accuracies and a Matthews correlation coefficient of 1.00. Furthermore, all subsets showed a total predictive accuracy and MCC values of 99.79% and 0.99, respectively. This was slightly lower than that achieved with the three subset combinations.

5.3.1.2 Differentially-Methylated versus unmethylated analysis for chromosome 6

Differentially methylated CpGI has been observed at different stages of cancer cells, while 55% of CpGI were differentially methylated in tumour cells [193]. Subset (1) showed a better predictive accuracy than the other three individual subsets, followed by subset (4). The accuracies of the individual subsets were 81.25%, 43.75%, 62.50% and 68.75% for subsets (1), (2), (3) and (4) respectively. Subset (1) shows a good correlation coefficient, while subset showed no correlation between the two classes. Subsets (3) and (4) gave approximately the same accuracy values, although each subset showed zero correlation coefficients (MCC). This indicates that the numbers of correctly classified and misclassified samples were exactly equal in cases in which the MCC value became zero (numerator = 0). Subset (1) incorporated the most predictors when compared to the other three subsets. The accuracy of two subset combinations (1,2; 1,3 and 1,4) was 97.96%, 95.00% and 79.90% , with standard errors of 2.49%, 5.89% and 8.10% respectively. For the pairs without subset (1), the accuracy and correlation coefficient values were decreased, although they showed a better class performance compared with single subset prediction, as noted in Table 5.5. The three subset combination (1,2,3) showed the highest predictive accuracy as well as class performance followed by subsets (1,2,4). Furthermore, three other subset combinations (1,3,4) and (2,3,4) showed a reduction both in accuracies and a correlation coefficient of accuracy of 65.50% and 62.60%, and coefficients of 0.37 and 0.16, respectively. The combination of all the subsets showed both the second highest predictive accuracy as well as class performance. The accuracy and correlation coefficient were 95.63% and 0.91 respectively.

5.3.1.3 Methylated and differentially-methylated for chromosome 6

This section presents the predictive accuracy and Matthews correlation coefficient values of methylated *versus* differentially-methylated results. For single sub-set prediction, the tissue-specific subset gave both the highest predictive accuracy and Matthews correlation coefficient values when compared to the other three individual subsets, which gave an accuracy and correlation coefficient of both 100% and 1.00, respectively. The other individual sub-sets showed lower accuracies and class performance. The accuracies of these three individual subsets (2,3 and 4) were 55.00%, 55.00% and 68.75% respectively. Two-subset (1,3) and (1,4) combinations showed 100% accuracy

Table 5.5 Results (% correct classification) for the analysis of differentially-methylated and unmethylated classes of chromosome 6.

combined feature-subsets	No of features	acc(DM)	acc(uM)	total accuracy	st-error	MCC
1	12	75.00	87.50	81.25	±8.84	0.63
2	9	37.50	50.00	43.75	±8.84	-0.13
3	16	62.50	62.50	62.50	0.00	0.00
4	13	62.50	75.00	68.75	±8.84	0.38
{1,2}	**21**	**96.25**	**99.67**	97.96	±2.42	**0.98**
{1,3}	28	100.00	91.67	95.00	±5.89	0.92
{1,4}	25	76.88	82.92	79.896	±8.10	0.63
{2,3}	25	33.75	62.50	48.13	±21.21	0.52
{2,4}	22	40.625	98.14	64.01	±40.67	0.38
{3,4}	29	67.50	64.17	65.83	±2.36	0.25
{1,2,3}	37	98.13	99.71	98.92	±1.38	0.97
{1,2,4}	34	65.63	87.08	76.35	±18.41	0.52
{1,3,4}	41	67.00	64.00	65.50	±9.19	0.37
{2,3,4}	38	26.88	98.32	62.60	±50.52	0.16
{1,2,3,4}	50	95.63	95.63	95.63	3.54	0.91

Abbreviations: acc(DM) = average predicve accuracy for differentially-methylated class; acc(uM) = average predictive accuracy for unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficients.

and MCC values only when a tissue-specific feature subset was included in the analysis. However, these were reduced when they were excluded from the analysis, as shown in more detail in Table 5.6. The pairs (2,3; 2,4; and 3,4) and three sub-sets (2,3,4), and a without tissue-specific CpGI subset (1) showed accuracies of 65%, 56.25%, 65.00% and 70%, with a mean standard deviation of 17.68%, 8.84%, 2.36% and 8.84% respectively. This analysis clearly distinguished methylated and differentially-methylated forms based on tissue-specific methylation, an observation which has been validated by previous studies[193; 194].

5.3.1.4 Three class prediction of chromosome 6

An attempt was made to predict methylated from unmethylated and differentially-methylated forms based on three-class problems. The accuracy of three-class predictions revealed a lower predictive accuracy overall for the single subset and each subset combination. Once again, the tissue-specific CpGI subset had a better predictive accuracy in single subset analysis. The accuracy of the tissue-specific CpGI sub-set was 77.30%, followed by dinucleotide distribution, which gave a value of 61.30%. Evolution and conservation and DNA structure subsets showed the lowest predictive accuracy, with large variability in standard error values when compared to the other individual subsets. For two subset combinations, tissue-specific CpGI was shown to have a good predictive accuracy when combined with one of the other three individual subsets.

Table 5.6 Results (% correct classification) for the analysis of Methylated and Differentially methylated classes of chromosome 6.

combined feature-sets	No of features	acc(M)	acc(DM)	total acc	st-error	MCC
1	**12**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
2	9	66.67	37.50	55.00	±20.62	0.04
3	16	66.67	37.50	55.00	±20.62	0.04
4	13	75.00	62.50	68.75	±8.84	0.38
{1,2}	21	100.00	87.50	95.00	±8.84	0.88
{1,3}	28	100.00	100.00	100.00	0.00	1.00
{1,4}	25	100.00	100.00	100.00	0.00	1.00
{2,3}	25	75.00	50.00	65.00	±17.68	0.26
{2,4}	22	62.50	50.00	56.25	±8.84	0.14
{3,4}	29	67.50	64.17	65.83	±2.36	0.32
{1,2,3}	37	100.00	100.00	100.00	0.00	1.00
{1,2,4}	34	100.00	100.00	100.00	0.00	1.00
{1,3,4}	41	100.00	100.00	100.00	0.00	1.00
{2,3,4}	38	75.00	62.50	70.00	±8.84	0.38
{1,2,3,4}	50	100.00	100.00	100.00	0.00	1.00

Abbreviations: acc(M) = average predictive accuracy for methylated class; acc(DM) = average predictive accuracy for differentially-methylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The bolded figures are those showed highest predictive accuracies and correlation coefficients.

However, when subset (1) was excluded, the predictive accuracy decreased. Furthermore, three-feature subset combinations also showed an improved predictive accuracy when compared to both the paired subset combination and the single subset analyses. The predictive accuracies of the three-subset combinations (1,2 and 3; 1,2 and 4; and 1,3 and 4) was 84.09%, 78.02% and 80.83%, with mean standard errors of 18.92%, 2.50% and 2.89% respectively. In addition, the feature subsets (1,2 and 3) gave a slightly higher predictive accuracy than that of the (1,3 and 4)subsets, despite showing a much higher standard error. It may therefore be concluded that feature subset (2) affects the overall predictive performance. Further details are shown in Table 5.7.

5.3.2 Data analysis for chromosome 20

5.3.2.1 Methaylated and unmethaylated chromosome 20

Chromosome 20 had the smallest sample set and the smallest sub-classes when compared to the four studied chromosomes (samples). The results of the analyses are reported in Table 5.8. For single subsets, i.e., (1) and (2), gave effective predictive accuracies and class performance, whereas subsets (3) and (4) showed lower predictive accuracies and no correlation between the two methylated and unmethylated classes. For pairs of the subsets, subset (1) gave the best predictive accuracy when combined with the other three individual subsets. The accuracies of these pairs (1,2 and 2,4)

Table 5.7 Results (% correctly classified) for the analyses of methylated, unmethylated and differentially-methylated classes (for the three class-prediction) of chromosome 6.

Combined feature-sets	No of features	acc(M)	acc(Unm)	acc(DM)	total acc	st-error
1	12	72.73	66.67	75.00	77.30	±4.31
2	9	62.50	80.00	25.00	54.06	±28.10
3	16	46.15	50.00	75.00	61.31	±15.67
4	13	54.55	66.67	50.00	58.13	±8.62
{1,2}	21	70.00	60.00	75.00	74.17	±7.64
{1,3}	28	70.00	63.64	75.00	75.00	±5.70
{1,4}	25	72.73	66.67	75.00	77.31	±4.31
{2,3}	25	50.00	57.14	37.15	48.54	±9.94
{2,4}	22	55.56	66.67	28.57	53.33	±19.60
{3,4}	29	40.00	46.15	85.71	63.40	±24.60
{1,2,3}	**37**	**72.73**	**63.64**	**100.00**	**84.09**	±18.92
{1,2,4}	34	72.73	70.00	75.00	78.02	±2.50
{1,3,4}	**41**	**75.00**	**75.00**	**80.00**	**80.83**	±2.89
{2,3,4}	38	55.56	66.67	28.57	51.88	±19.59
{1,2,3,4}	50	75.00	75.00	80.00	80.83	±2.89

Abbreviation: acc(M) = average predictive accuracy for the methylated class; acc(uM) = average predictive accuracy for the unmethylated class; acc(DM) = average predicve accuracy for differentially-methylated class; total acc.= total average predictive accuracy; st-error = standard error. The figures in bold correspond to those showing the highest predictive accuracies and correlation coefficients.

gave higher class performances and also had good predictive accuracies since no misclassification occurred. In contrast, subsets (2,3), (2,4) and (3,4) showed the worst predictive accuracies, and negative correlations. Concordantly, the standard deviation increased for these three subsets.

The three-subset combination (1,2,3) showed 100% sensitivity, whereas specificity was reduced to 66.67% with a standard error of 23.17%, and a Matthews correlation coefficient of 0.70. The subsets (1, 3, 4) gave an accuracy of 83.33%, and an MCC value of 0.60, which is lower than the other combinations of three and two subsets. Whilst the subset combination (2,3,4) showed less predictive accuracy than the other three-subset combinations, it gave a 75% accuracy and a 0.50 value for the Mathews correlation coefficient. All combined subsets showed 93.10% and 0.67 accuracy and Mathews correlation coefficient values respectively.

5.3.2.2 Differentially-methylated versus unmethylated classes for chromosome 20

This part of the study examined individual subsets as well as all possible subset combinations (the results are presented in Table 5.9). For single feature subsets, subset (1) again had the highest predictive accuracy (88.50% compared to 71.43%, 68.57% and 74.25% for the other three subsets). Single subsets showed higher sensitivity than

Table 5.8 Results for the analyses of methylated and unmethylated classes for chromosome 20.

Combined feature-sets	No of features	acc(M)	acc(uM)	total acc	st-error	MCC
1	**12**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
2	**9**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
3	16	50.00	66.67	58.33	±11.79	0.17
4	13	33.33	66.67	50.00	±23.57	0.00
{1,2}	21	100.00	100.00	100.00	0.00	1.00
{1,3}	28	83.33	66.67	75.00	±11.79	0.31
{1,4}	25	100.00	100.00	100.00	0.00	1.00
{2,3}	25	16.67	50.00	33.33	±23.57	-0.35
{2,4}	22	33.33	66.67	50.00	±23.57	-0.10
{3,4}	29	33.33	50.00	41.67	±11.79	0.17
{1,2,3}	37	100.00	66.67	83.33	±23.57	0.71
{1,2,4}	34	100.00	100.00	100.00	0.00	1.00
{1,3,4}	41	83.33	83.33	83.33	0.00	0.60
{2,3,4}	38	83.33	66.67	75.00	±11.79	0.50
{1,2,3,4}	50	83.33	95.65	93.10	±8.71	0.67

Abbreviations: acc(M) = average predictive accuracy for the methylated class; acc(uM) = average predictive accuracy for the unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold are those showing the highest predictive accuracies and correlation coefficients.

specificity. For all the two-subset combinations, prediction accuracy was improved over that noted with single feature subsets, although subsets (3,4) resulted in the lowest predictive accuracy when compared to all other pairs. Its specificity was very low (Table 5.9). The three-subset combinations had both a high predictive accuracy and $Matthew^s$ correlation coefficient, with the exception of subsets (2,3,4), which showed a low-specificity and classification performance. It was also the least effective when compared to the other three combinations of subsets. In addition, all subsets showed good predictive accuracies with high sensitivity and classification performance.

5.3.2.3 Methylated and differentially methylated classes for chromosome 20

For single subset predictions, subset (1) had the highest predictive accuracy of 91.67%, whilst the other three individual subsets had both lower accuracies and Matthews correlation coefficients. The subsets (2,4) had the same sensitivity, whilst subset (3) had a slightly higher sensitivity of 66.67%. For two-subset combinations, subset (1,4) showed the highest predictive accuracy when compared to the three other three pair sets, but only when subset (1) was present in the pairs. However, when subset (1) was excluded, the combined remaining two showed both the lowest correlation and accuracy values, as detailed in Table 5.10.

Table 5.9 Results for the analysis of differentially-methylated and unmethylated classes of chromosome 20.

Combined feature-sets	No of features	acc(DM)	acc(uM)	total acc	st-error	MCC
1	12	91.30	83.33	88.57	±5.64	0.71
2	9	86.96	41.67	71.43	±32.02	0.27
3	16	91.30	25.00	68.57	±46.88	0.20
4	13	82.61	58.33	74.29	±17.17	0.32
{1,2}	**21**	**95.65**	**91.67**	**94.29**	±2.8	**0.82**
{1,3}	28	91.30	66.67	82.86	±17.42	0.54
{1,4}	25	91.30	83.33	88.57	±5.64	0.71
{2,3}	25	86.96	75.00	82.86	±8.45	0.55
{2,4}	22	91.30	83.33	88.57	±5.64	0.71
{3,4}	29	95.65	25.00	71.43	±49.96	0.25
{1,2,3}	37	91.30	75.00	85.71	±11.53	0.62
{1,2,4}	**34**	**91.30**	**91.67**	**91.43**	±0.26	**0.82**
{1,3,4}	41	91.30	75.00	85.71	±11.53	62.20
{2,3,4}	38	78.26	33.33	62.86	±31.77	10.00
{1,2,3,4}	**50**	**91.30**	**85.71**	**93.33**	±11.53	**0.89**

Abbreviations: acc(DM) = average predictive accuracy for the differentially-methylated class; acc(uM) = average predictive accuracy for the unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold are those showing the highest predictive accuracies and correlation coefficients.

Three-feature subset combinations showed an overall good predictive accuracy, apart from the (2,3,4) subset combination, for which the predictive accuracy and Matthews correlation coefficient decreased since feature subset (1) was excluded from the analysis. In general, the methylated and differentially-methylated classes showed a lower predictive accuracy. This is the smallest sample that generally caused lower predictive power. The predictive accuracy of the three-subset combination was 72.27%, 83.33% and 72.22%; for (1,2) (1,3) and (1,4) respectively (All subsets had an accuracy of 72.22%). This is approximately the same as that found for the three subset combinations, although all subsets had a lower Matthews correlation coefficient than the three subset combinations.

5.3.2.4 Three class prediction of Chromosome 20

Three-class analysis for the tissue-specific CpGI subset demonstrated the highest accuracy for single subset analysis, whereas the evolution and conservation subset, dinucleotide distribution and DNA structure properties showed a lower predictive accuracy. Their predictive accuracies were in the range of 40.28% to 76.90%. Furthermore, the two and three-feature subset combinations showed a reliable predictive accuracy, particularly when tissue-specific CpGI was included in the analysis. In contrast, combined subsets without the tissue-specific CpGI subset showed a decrease in predictive accu-

Table 5.10 Results for the analyses of Methylated and Differentially-methylated classes of chromosome 20.

Combined feature-sets	No of features	acc(M)	acc(DM)	total acc	st-error	MCC
1	**12**	**100.00**	**83.33**	**91.67**	±11.79	**0.79**
2	9	66.67	16.67	41.67	±35.36	-0.19
3	16	50.00	66.67	58.33	±11.79	0.16
4	13	50.00	83.33	66.67	±23.57	0.35
{1,2}	21	66.67	83.33	77.78	±11.79	0.50
{1,3}	28	50.00	91.67	77.78	±29.46	0.47
{1,4}	25	66.67	91.67	83.33	±17.68	0.61
{2,3}	25	33.33	58.33	50.00	±17.68	-0.81
{2,4}	22	66.67	33.33	50.00	±23.57	0.00
{3,4}	29	16.67	75.00	55.56	±41.25	-0.10
{1,2,3}	37	50.00	83.33	72.22	±23.57	0.35
{1,2,4}	34	66.67	91.67	83.33	±17.68	0.57
{1,3,4}	41	50.00	83.33	72.22	±23.57	0.35
{2,3,4}	38	33.33	58.33	50.00	±17.68	-0.81
{1,2,3,4}	50	83.33	50.00	72.22	±23.57	0.32

Abbreviations: acc(M) = average predictive accuracy for the methylated class; acc(DM) = average predictive accuracy for the differentially-methylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold correspond to those showing the highest predictive accuracies and correlation coefficients.

racy, in addition to class performance. The tissue-specific CpGI subset (1) was found to be the best predictor when compared to the other feature-subsets, as reported in Table 5.11.

5.3.3 Data analysis for Chromosome 22

5.3.3.1 Methylated and unmethylated class prediction

Single subset analysis showed that the tissue-specific CpGI subset had the highest predictive accuracy and Matthews correlation coefficient (MCC) of 98.48% and 0.96 respectively. The other three individual sub-sets (2,3,4) gave predictive accuracies of 75.75%, 65.91% and 54.55% respectively. These three subsets, however, had a lower correlation coefficient when compared to that of subset (1), whereas subset(2) gave an improved predictive accuracy when compared to other two individual subsets (3,4). For the two-subset combination, the predictive accuracy improved overall from 64.02% to 100%, and the correlation coefficient (MCC) also increased from 0.29 to 1.00 when compared with the single $subset^s$ value, which increased from 9.11% to95.50%. However, most showed a lower class prediction and correlation coefficients (MCC), particularly when subset (1) was excluded. The three-subset combinations also showed reliable predictive accuracies as well as good correlation coefficients (MCC), although

Table 5.11 Results for the analyses of methylated, unmethylated and differentially-methylated classes (for the three class-prediction) of chromosome 20.

Combined feature-sets	No of features	acc(M)	acc(Unm)	acc(DM)	total acc	st-error
1	12	60.00	87.50	66.67	76.19	±14.35
2	9	25.00	60.00	87.50	61.36	±31.42
3	16	40.00	75.00	25.00	40.28	±25.65
4	13	50.00	87.50	63.64	66.67	±18.98
{1,2}	21	80.00	85.72	83.33	82.58	±2.87
{1,3}	28	80.00	87.50	83.33	83.22	±3.76
{1,4}	25	80.00	87.50	85.71	83.97	±3.92
{2,3}	25	33.33	63.64	87.50	63.94	±27.15
{2,4}	22	25.00	55.56	87.50	59.44	±31.25
{3,4}	29	55.00	55.56	87.50	64.43	±26.35
{1,2,3}	37	66.67	83.33	85.71	79.17	±10.17
{1,2,4}	34	66.67	71.43	100.00	84.61	±18.00
{1,3,4}	**41**	**80.00**	**85.71**	**87.50**	**87.50**	±3.92
{2,3,4}	38	50.00	70.00	80.00	70.00	±15.81
{1,2,3,4}	50	66.67	83.33	85.71	79.17	±10.38

Abbreviations: acc(M) = average predictive accuracy for methylated class; acc(uM) = average predictive accuracy for unmethylated class; acc(DM) = average predicve accuracy for differentially-methylated class; total acc. = total average predictive accuracy; st-error = standard error.The figures in bold are those showing the highest predictive accuracies and correlation coefficients.

these were slightly lower than that attained with the two-subset combination. Furthermore, combinations of all sub-sets gave approximately the same predictive accuracy as the three subsets. Further details are shown in Table 5.12.

5.3.3.2 Differentially-methylated and unmethylated classes for chromosome 22

Single subset prediction showed that subset (1) dominated the classification potential when compared to the other three single subsets (2,3,4), which gave an accuracy of 75% and a correlation (MCC) of 0.51. These three subsets (2,3,4) gave accuracies of 68.75%, 43.75% and 47.92% respectively. The subset (3,4) had a negative correlation (MCC), whereas subset(2) showed a weak correlation between the differentially-methylated and unmethylated groups. Two subset combinations (1,2; 1,4; 1,3) gave high Matthews correlation coefficient (MCC) values of 0.88, 0.80 and 0.76 respectively. In addition, the subsets (2,3; 2,4 and 3,4) showed reduced accuracies of 67.73%, 68.72% and 64.02% respectively. There was also a reduction of the Matthews correlation coefficient when the tissue-specific CpGI subset was excluded from the analysis. Further details are shown in Table 5.13.

Of the three subset combinations, subset (1,2,4) had the best class predictions between methylated and differentially-methylated ones, followed by subsets (1,3,4)

Table 5.12 Results for the analyses of methylated and unmethylated classes of chromosome 22.

combined feature-sets	No of features	acc(M)	acc(uM)	total-acc	st-error	MCC
1	12	100.00	96.97	98.48	±2.14	0.96
2	9	78.79	72.73	75.76	±4.29	0.52
3	16	50.00	81.82	65.91	±22.50	0.33
4	13	51.52	57.58	54.55	±4.29	0.09
{1,2}	**21**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
{1,3}	**28**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
{1,4}	**25**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
{2,3}	25	78.79	89.39	84.09	±7.50	0.69
{2,4}	22	65.42	75.42	70.42	±7.66	0.41
{3,4}	29	68.94	59.09	64.02	±6.96	0.30
{1,2,3}	37	100.00	99.32	99.66	±0.48	0.99
{1,2,4}	34	100.00	98.94	99.47	±0.75	0.98
{1,3,4}	41	100.00	99.85	99.92	±0.107	0.99
{2,3,4}	38	67.27	79.47	73.37	±8.62	0.46
{1,2,3,4}	50	99.09	100.00	99.55	±0.64	0.99

Abbreviations: acc(M) = average predictive accuracy for the methylated class; acc(uM) = average predictive accuracy for the unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold are those showing the highest predictive accuracies and correlation coefficients.

and (1,2,3). The accuracies of these three-subset combinations were 94.64%, 93.67% and 67.87% respectively. It should also be noted that the three subset combinations gave an improved predictive accuracy than the two-subset combination. The pairs of subsets (2,3) also showed a lower predictive accuracy than the combination of three subsets (2,3,4). Furthermore, all subset analyses showed the highest total predictive accuracy, in addition to class performance ability.

5.3.3.3 Methylated and differentially-methylated for chromosome 22

For tissue-specific CpGIs, subset (1) showed the highest predictive accuracy and class performance when compared to the other individual subsets, followed by subset (3). The accuracies of these subsets are 93.75% and 75%, with Matthews correlation coefficients of 0.88 and 0.50 respectively. The other two individual subsets (2,4) showed lower predictive accuracies and weak Matthews correlation coefficients. The subsets (3,4) had the lowest accuracy and the lowest class correlation coefficients when compared to combined pairs, as shown in Table 5.14.

Whilst the three-subset combination revealed an improved overall accuracy than the two-subset combination, a combination of both combinations showed approximately the same Matthews correlation coefficient, whereas the combination of subsets (2,3 and 4) revealed a marked improvement in both accuracy and Matthews correla-

Table 5.13 Results for the analyses of differentially-methylated and unmethylated classes of chromosome 22.

Combined feature-sets	No of features	acc(DM)	acc(uM)	total accuracy	st-error	MCC
1	12	83.33	66.67	75.00	±11.79	0.51
2	9	66.67	70.83	68.75	±2.95	0.38
3	16	45.83	41.67	43.75	±2.95	-0.13
4	13	45.83	50.00	47.92	±2.95	-0.042
{1,2}	21	90.00	98.73	94.37	±6.17	0.88
{1,3}	28	79.38	99.48	89.43	±14.22	0.80
{1,4}	25	83.96	99.10	91.53	±10.71	0.80
{2,3}	25	49.79	85.67	67.73	±25.37	0.35
{2,4}	22	50.21	87.24	68.72	±26.18	0.40
{3,4}	29	68.94	59.09	64.02	±6.96	0.30
{1,2,3}	37	50.21	85.52	67.87	±24.97	0.31
{1,2,4}	34	90.63	198.66	94.64	±5.68	0.88
{1,3,4}	41	87.71	99.63	93.67	±8.43	0.84
{2,3,4}	38	53.54	85.75	69.64	±22.77	0.39
{1,2,3,4}	**50**	**99.09**	**100.00**	**99.55**	±0.64	**0.96**

Abbreviation: acc(DM) = average predictive accuracy for differentially-methylated class; acc(uM) = average predictive accuracy for unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold correspond to those showing the highest predictive accuracies and correlation coefficients.

tion coefficient value. In addition, all subset combinations gave a reliable predictive accuracy and class performance.

5.3.3.4 Three-class predictions of Chromosome 22

The three-class problem (one class compared to all the other classes, specifically methylated *versus* unmethylated and differentially-methylated groups) was examined (results are presented in Table 5.15). Tissue-specific CpGI was the best predictor subset when compared to the three other individual feature subsets. The accuracies of the four individual subsets were 85.55%, 53.70, 54.05 and 54.77%, and their mean standard errors were 8.65%, 5.18%, 10.38% and 3.47% for tissue-specific CpGI, evolution and conservation, dinucleotide distribution and DNA structural properties, respectively. Combination of two to three subsets showed the highest sensitivity when the tissue-specific CpGI subset was included in the analysis, while in its absence the sensitivity was reduced. Moreover, all subset combinations demonstrated effective predictive accuracy, although this was somewhat lower than that of the two-feature subset combination.

Table 5.14 Results for the analyses of methylated and differentially-methylated classes of chromosome 22.

Combined feature-sets	No of features	acc(M)	acc(DM)	total accuracy	st-error	MCC
1	12	100.00	87.50	93.75	±8.84	0.88
2	9	58.33	41.67	50.00	±11.79	0.050
3	16	75.00	75.00	75.00	0.00	0.50
4	13	66.67	54.17	60.42	±8.84	0.21
{1,2}	21	90.21	100.00	95.10	±6.92	0.92
{1,3}	28	100.00	79.17	89.58	±14.73	0.81
{1,4}	25	95.83	83.33	89.58	±8.84	0.80
{2,3}	25	71.88	88.66	80.27	±12.00	0.59
{2,4}	22	62.92	91.64	77.28	±20.31	0.57
{3,4}	29	68.94	59.09	64.02	±6.96	0.29
{1,2,3}	37	78.96	98.88	88.92	±14.09	0.77
{1,2,4}	34	100.00	91.67	95.83	±5.89	0.92
{1,3,4}	41	100.00	83.33	91.67	±11.79	0.85
{2,3,4}	38	71.67	90.15	80.91	±13.07	0.64
{1,2,3,4}	**50**	**99.09**	**100.00**	**99.55**	±0.64	**0.96**

Abbreviations: acc(M) = average predictive accuracy for the methylated class; acc(DM) = average predictive accuracy for the differentilally-methylated class; total acc.= total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The figures in bold are those showing the highest predictive accuracies and correlation coefficients.

5.3.4 Data analysis for Chromosome 21

5.3.4.1 Methylated *versus* unmethylated classes for chromosome 21

Table 5.16 represents the simulation results of individual feature-subsets by comparing the two predictive models: Modified Leave-One-Out cross validation and traditional leave-one-out cross validation. This shows that the MLOOCV approach has more effective predictive performance when compared to that of the LOOCV technique, particularly for the methylated (minority) class. For example, the single feature-set DNA structure shows inconsistent class performance (imbalanced features), where the methylated, unmethylated and total predictive accuracies were 27.59%, 83.50% and 71.21% respectively, compared to balanced model results of 55.17%, 52.41.50% and 53.28% . Exon and gene distribution (balanced features) gave the best predictive class performance when compared to both balanced and imbalanced single feature-sets. DNA sequence properties and distribution for both (balanced and imbalanced featuresets) showed the lowest class performance when compared to that achieved with six featuresets. This could be attributable to the imbalance between the sample size and their large attributes, which may contain some noisy features. However, the other five feature sets showed very similar performances.

Table 5.15 Results for the analyses methylated, unmethylated and differentially-methylated classes (for the three class-prediction) of chromosome 22.

Combined feature-sets	No of features	acc(M)	acc(unm)	acc(DM)	total accuracy	st-error
1	**12**	**90.00**	**85.71**	**73.33**	**85.55**	±8.65
2	9	50.00	60.00	52.63	53.70	±5.18
3	16	47.81	50.00	66.67	54.05	±10.38
4	13	53.33	58.33	60.00	54.77	±3.47
{1,2}	21	89.47	84.61	68.62	82.64	±10.84
{1,3}	28	90.00	86.67	71.43	84.29	±9.90
{1,4}	25	89.47	85.71	66.67	82.51	±12.23
{2,3}	25	50.00	60.00	50.00	51.47	±5.77
{2,4}	22	50.00	56.52	55.56	52.14	±3.52
{3,4}	29	53.03	55.56	54.29	54.05	±3.45
{1,2,3}	37	90.00	85.71	73.33	84.40	±8.65
{1,2,4}	34	88.89	85.71	64.71	82.32	±13.14
{1,3,4}	41	90.48	87.50	80.00	87.65	±5.40
{2,3,4}	38	60.00	64.70	55.00	59.82	±4.85
{1,2,3,4}	**50**	**90.47**	**85.71**	**78.57**	**85.95**	±5.99

Abbreviations: acc(M) = average predictive accuracy for the methylated class; acc(uM) = average predictive accuracy for the unmethylated class; acc(DM) = average predictive accuracy for differentially-methylated class; total acc. = total average predictive accuracy; st-error = standard error. The bolded figures are those showed highest predictive accuracies and correlation coefficients.

Table 5.16 Percentage of predicted accuracy of individual featuresets, and comparison of the balanced and imbalance feature sub-sets[5].

feature-sets	predictive accuracy(M		predictive accuracy(unM)		total predictive accuracy	
	MLOOCV	LOOCV	MLOOCV	LOOCV	MLOOCV	LOOCV
f1	67.07	24.14	76.53	73.79	71.72	62.88
f2	52.07	24.14	55.52	77.67	55.34	65.91
f3	46.90	20.69	64.83	75.73	55.73	63.64
f4	55.17	27.59	52.41	83.50	53.28	71.21
f5	54.48	20.69	71.72	80.58	63.14	67.42
f6	47.64	20.69	64.14	80.58	56.55	67.42
f7	44.14	20.69	92.76	97.09	66.21	80.30

Furthermore, the remaider of the analysis used modified leave-one-out cross validation. These results are reported in Table 5.17. Single-subset analysis, sequence patterns and exon and gene distribution showed the highest predictive accuracy, followed by single nucleotide polymorphism. Subsets (2,3 and 4) showed a negative correlation, but also gave a lower predictive accuracy. In addition, evolution and conservation attributes showed a total accuracy of 53.45%, with a weak correlation

coefficient.

For the two subset combinations, the total predictive accuracy was increased when at least two subsets were combined, which showed an improved predictive accuracy and class performance for both sequence pattern and exon and gene distribution subsets. Their accuracy and Matthews correlation coefficient values were 77.59% and 0.55 respectively, and the mean standard deviation was 2.44%. The next highest predictive accuracy is CpG island distribution and exon and gene distribution, which had an accuracy of 71.21% and a correlation coefficient (MCC) of 0.42. As shown in Table 5.18, exon and gene distribution was the most important subset.

For the three-subset combinations, subsets (1,3,6) gave the highest predictive accuracy and MCC values (79.31% and 0.59 respectively). For the three-subset combination, the evolution and conservation subset is dominant, and present for all four selected subsets. In addition, the three-subset combinations gave a slightly higher predictive accuracy when compared to that attained with two-subset combination. For the four-subset combinations, the method in this study only selected a single combination-subset (1,2,3,5) from all possible combinations of four subsets, which gives an accuracy of 70.69% and correlation coefficient of 0.41. For five-subset combinations, subsets (2,3,4,5,7) had the highest predictive accuracy (79.31%) and a correlation coefficient of 0.59, followed by feature subset (1,2,5,6,7), which gave a total accuracy of 77.56% and a Matthews correlation coefficient of 0.60. Subsets 2 and 5 were the best predictors in the five subset combinations.

For the six-subset combination, subsets (1,2,3,4,5, and 7) gave the highest predictive accuracy, as well as class performance; this was followed by subsets (2,3,4,5,6 and 7). The total accuracies of these subsets were 79.31% and 75.86%, and their correlation coefficients were 0.59 and 0.52 respectively. The best subsets that appeared in all selected subset results were subsets (2,3,6 and 7). All-subset combinations showed slightly more effective predictive accuracy than the single-subset prediction, despite the standard deviation of all subsets being much lower than that of the single subset prediction. For the latter, subset (1) gave 100% of total predictive accuracy, whereas the other individual subsets showed lower predictive accuracies and a weak correlation. For example, subsets (4 and 7) showed a very weak negative correlation, whilst exon and gene distribution were no better than random guessing. For the two-subset combination, subsets (1,2; 1,4 and 1,7) gave 100% total predictive accuracy, and a class performance without any misclassification, whereas subsets (1,3) gave a total accuracy of 70.67% and a correlation coefficient of 0.42. When subset (1) was excluded from the pairs, the total accuracies decreased, and the class performance was also very low (data not shown).

Table 5.17 Results for the analyses of methylated and unmethylated classes of chromosome 21.

Combined feature-sets	No of features	acc(M)	acc(unM)	total acc	st.error	MCC
1	2870	51.72	93.10	72.41	±29.26	0.49
2	112	53.33	46.67	50.00	±4.71	-0.03
3	91	26.67	46.67	36.67	±14.14	-0.21
4	196	55.17	34.48	44.83	±14.63	-0.11
5	364	65.52	82.76	74.14	±12.19	0.49
6	70	55.17	51.72	53.45	±2.44	0.07
7	56	51.72	79.31034	65.52	±19.51	0.32
{1,5}	3234	79.31	75.86	77.59	±2.44	0.55
{3,5}	455	68.97	73.45	71.21	±9.02	0.42
{5,6}	434	67.93	72.41	70.17	±6.58	0.41
{5,7}	420	67.59	73.10	70.34	±9.75	0.41
{1,3,6}	**3031**	**82.76**	**75.86**	**79.31**	±4.88	**0.59**
{3,6,7}	217	67.93	72.41	70.17	±6.58	0.41
{4,5,6}	630	71.38	73.45	72.41	±5.85	0.47
{5,6,7}	490	67.93	72.41	70.17	±6.58	0.41
{1,2,3,5}	3437	67.93	73.45	70.69	±8.29	0.41
{1,2,3,5,6}	3507	64.48	74.12	69.31	±8.29	0.39
{1,2,3,4,5}	3633	67.93	73.45	70.69	±8.29	0.41
{2,3,4,5,7}	**819**	**82.76**	**75.86**	**79.31**	±4.88	**0.59**
{1,2,5,6,7}	**3472**	**72.41**	**82.76**	**77.56**	±7.31	**0.59**
{1,2,3,4,5,7}	**3703**	**82.76**	**75.86**	**79.31**	±4.88	**0.59**
{1,2,3,4,5,7}	3689	67.93	73.45	70.69	±8.29	0.41
{1,2,3,5,6,7}	3563	68.28	73.45	70.86	±8.05	0.41
{2,3,4,5,6,7}	889	72.41	79.31	75.86	±4.88	0.52
{1,.....,7}	3759	72.41	79.31	75.86	±4.88	0.52

Abbreviations: acc(M) = average predictive accuracy for the methylated class; acc(uM) = average predictive accuracy for the unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The bolded figures are those showed highest predictive accuracies and correlation coefficients.

Table 5.18 Results for the analyses of differentially-methylated and unmethylated classes of chromosome 21.

Combined feature-sets	No of features	acc(DM)	acc(unM)	total accu.	st.error	MCC
1	**2870**	**100.00**	**100.00**	**100.00**	**0.00**	**1.00**
2	112	53.33	60.00	56.67	±4.71	0.13
3	91	73.33	46.67	60.00	±18.86	0.21
4	196	40.00	80.00	60.00	±28.28	-0.22
5	364	60.00	40.00	50.00	±14.14	0.00
6	70	53.33	73.33	63.33	±14.14	0.27
7	56	46.67	46.67	46.67	0.00	-0.03
{1,2}	2982	100.00	100.00	100.00	0.00	1.00
{1,3}	2961	68.67	72.67	70.67	±6.60	0.42
{1,4}	3066	100.00	100.00	100.00	0.00	1.00
{1,7}	2926	100.00	100.00	100.00	0.00	1.00
{1,2,4}	3178	100.00	100.00	100.00	0.00	1.00
{1,2,5}	3346	61.33	71.33	66.33	±8.96	0.35
{1,2,7}	3038	100.00	100.00	100.00	0.00	1.00
{1,3,4}	3157	68.67	72.67	70.67	±6.60	0.42
{1,2,6}	3052	100.00	100.00	100.00	0.00	1.00
{1,2,4,5}	3542	57.33	68.67	63.00	±9.90	0.25
{1,3,4,6}	3227	66.67	73.33	70.00	±4.71	0.40
{1,3,5,7}	3381	54.67	70.00	62.33	±11.79	0.25
{1,3,6,7}	3087	66.67	73.33	70.00	±4.71	0.40
{1,2,3,4,5}	3633	56.00	69.33	62.67	±10.37	0.24
{1,3,4,5,6}	3591	66.67	73.33	70.00	±4.71	0.40
{1,2,3,4,5,6}	3703	51.33	68.00	59.67	±14.61	0.20
{1,2,3,4,5,7}	3689	66.67	73.33	70.00	±4.71	0.40
{1,.....,7}	3759	46.67	60.00	53.33	±9.43	0.07

Abreviations: acc(DM) = average predictive accuracy for the differentially-methylated class; acc(uM) = average predictive accuracy for the unmethylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The bolded figures are those showed highest predictive accuracies and correlation coefficients.

5.3.4.2 Differentially methylated and unmethylated 21

The three-subset combination of subset (1,2,4; 1,2,7 and 1,2,6) gave 100% total accuracy in addition to class performance, whilst subsets (1,2,5 and 1,3,4) had lower predictive accuracies than the other three subsets. Subsets 1 and 2 were the best predictors in the combinations. In addition, subset (1) was present in all selected results of the three-subset combinations, while subset (2) contained only four of the selected subset results. For the four-subset combination, the overall accuracies and class performance were reduced. This also showed low sensitivity when compared with that of the three subset-combination (as detailed in Table 5.18). Although the total accuracy was lower than that for the three-subset combination, subset(1) was the best predictor since it appeared in all four selected combined subsets. The five- and six-subset combinations showed approximately the same total accuracies as the three subsets, whilst their class performance was slightly reduced when compared with that of the three-subset combinations. Furthermore, all subsets showed a further decrease in accuracy and correlation, particularly when subset (1) was removed from these combinations.

5.3.4.3 Methylated and differentially-methylated results for chromosomes 21

For single-subset prediction (for this approach must be noted that each subset contains two subclasses), subset (1) revealed the highest predictive accuracy and the highest classification performance with a total accuracy of 96.67%. Its correlation coefficient (MCC) value was 0.94, whereas the other individual subsets (2,3,4,5,6 and 7) showed only a weak correlation.

For the two-subset combination, subsets (1,2), (1,4) and (1,7) gave a total accuracy of 96.67%. However, when subset (1) was excluded from the analysis, the total accuracy was considerably reduced (further details are shown in Table 5.19). The best predictor of the paired subsets is DNA sequence patterns for subset (1). For the three-subset combinations, the total accuracies were reduced overall, with the exception of subsets (1,2,4), which did not change when compared with that of the two-subset combination.

For the four-subset combination, subsets (1,2,4,7) showed a total accuracy of 96.67%, which is the same as the single subset (1) result. Furthermore, the total accuracy of the other four-subset combinations was further reduced when compared to the two- and three-subset combinations. In addition, five- and six-subset combinations also showed a further decrease in accuracy, as well as a weak positive correlation. All subsets showed a reduced accuracy.

Table 5.19 Results for the analyses of methylated and differentially-methylated classes of chromosome 21.

Combined feature-sets	No of features	acc(M)	acc(DM)	total acc	st.error	MCC
1	**2870**	**93.33**	**100.00**	**96.67**	±4.71	**0.94**
2	112	46.67	46.67	46.67	0.00	-0.07
3	91	53.33	46.67	50.00	±4.71	0.00
4	196	26.67	80.00	53.33	±37.71	0.08
5	364	60.00	60.00	60.00	0.00	0.20
6	70	33.33	46.67	40.00	±9.43	-0.27
7	56	26.67	46.67	36.67	±14.14	-0.25
{1,2}	2982	93.33	100.00	96.67	±4.71	0.93
{1,4}	3066	93.33	100.00	96.67	±4.71	0.94
{1,7}	2926	93.33	100.00	96.67	±4.71	0.94
{2,5}	476	46.67	73.33	60.00	±18.86	0.21
{1,2,4}	3178	93.33	100.00	96.67	±4.71	0.94
{1,5,7}	3290	53.33	66.67	60.00	±9.43	0.20
{2,3,4}	399	53.33	46.67	50.00	±4.71	0.00
{2,5,7}	532	46.67	66.67	56.67	±14.14	0.06
{5,6,7}	490	53.33	66.67	60.00	±9.43	0.20
{1,2,3,4}	3269	53.33	46.67	50.00	±4.71	0.00
{1,2,4,6}	3248	33.33	60.00	46.67	±18.86	-0.04
{1,2,4,7}	3234	93.33	100.00	96.67	±4.71	0.94
{1,2,3,4,5}	3633	60.00	60.00	60.00	0.00	0.20
{1,2,3,4,7}	3325	53.33	46.67	50.00	±4.71	0.00
{1,4,5,6,7}	3556	53.33	66.67	60.00	±9.43	0.20
{1,2,3,4,5,6}	3703	53.33	60.00	56.67	±4.71	0.13
{1,2,3,4,5,7}	3689	60.00	60.00	60.00	0.00	0.20
{1,.....,7}	53.33	55.17	3759	54.54	±1.30	0.08

Abbreviation: acc(M) = average predictive accuracy for methylated class; acc(DM) = average predictive accuracy for differentially-methylated class; total acc. = total average predictive accuracy; st-error = standard error; MCC = Matthews Correlation Coefficient.The bolded figures are those showed highest predictive accuracies and correlation coefficients.

Table 5.20 Results for the analyses of methylated, unmethylated and differentially-methylated classes (for the three class-prediction) of chromosome 21.

Combined feature-sets	No of features	acc(M)	acc(DM)	acc(Unm)	total acc	st.error
1	2870	84.61	88.88	72.72	82.08	±2.35
2	112	61.54	50.00	77.78	63.10	±13.98
3	91	66.67	85.71	11.11	54.50	±38.76
4	196	71.43	66.67	75.00	71.03	±4.18
5	**364**	**100.00**	**95.00**	**95.00**	**96.67**	±2.85
6	70	80.00	88.88	60.00	72.92	±14.80
7	56	81.82	72.73	62.50	71.97	±9.67
{1,2}	2982	84.62	81.82	77.78	76.28	±4.71
{1,3}	2961	83.33	75.00	77.78	75.00	±4.24
{1,4}	3066	83.33	81.82	70.00	74.00	±7.30
{1,5}	3234	83.33	81.82	70.00	74.00	±7.30
{1,7}	2926	76.92	81.82	70.00	72.28	±5.93
{2,5}	476	63.63	50.00	70.00	63.57	±10.21
{5,6}	434	100	90.00	100.00	94.23	±5.78
{5,7}	420	100.00	90.00	100.00	94.23	±5.78
{1,2,4}	3178	83.33	80.00	72.73	74.24	±5.43
{1,2,5}	3346	83.33	80.00	70.00	73.60	±6.94
{1,2,6}	3052	84.61	88.89	62.50	75.18	±14.16
{1,2,7}	3038	84.61	90.00	70.00	76.81	±10.35
{1,3,4}	3157	84.61	81.82	77.78	76.28	±3.49
{1,3,6}	3031	83.33	81.82	70.00	74.00	±6.94
{1,5,7}	3290	85.71	90.00	77.78	78.94	±6.20
{2,3,4}	399	70.00	54.56	66.67	63.45	±8.13
{2,5,7}	532	72.73	62.50	55.56	63.16	±8.64
{5,6,7}	490	100.00	90.00	95.00	94.44	±2.56
{1,2,3,4}	3269	85.71	81.82	87.50	80.00	±2.91
{1,2,4,6}	3248	83.33	75.00	77.78	75.00	±4.25
{1,2,4,7}	3234	84.61	80.00	77.78	77.30	±3.49
{1,2,3,4,5}	3633	83.33	81.82	72.73	75.92	±5.74
{1,2,3,5,6}	3507	84.62	81.82	80.00	76.85	±2.45
{1,4,5,6,7}	3556	83.33	78.57	71.43	74.48	±5.99
{1,2,3,4,5,6}	3703	84.62	83.33	80.00	80.00	±2.38
{1,2,3,4,5,7}	3689	84.62	81.82	80.00	78.21	±2.32
{1,.....,7}	3759	80.00	80.00	58.33	68.85	±12.50

Abbreviations: acc(M) =average predictive accuracy for methylated class; acc(DM) = average predictive accuracy for differentially-methylated class;acc(uM) = average predictive accuracy for unmethylated class; total acc. = total average predictive accuracy; st-error = standard error.The bolded figures are those showed highest predictive accuracies and correlation coefficient.

5.3.4.4 Three class prediction of Chromosome 21

For single-subset analysis, exon and gene distribution, evolution and conservation and sequence distribution showed the highest predictive accuracies of 96.63%, 82.08% and 72.92%, with standard errors of 2.85%, 2.35% and 14.80% respectively. DNA structure and properties gave the lowest predictive accuracy, whilst exon and gene distribution subset was the most effective predictor for both single and paired subset analysis. Moreover, for the three, four, five and six-subset combinations, sequence distribution was the best subset, and was present in almost all the selected combined subsets despite showing a decrease in predictive accuracy when compared to that of other subset combinations, as detailed in Table 5.20.

5.3.4.5 Methylation sub-classes prediction for Chromosome 21

A total of 44 features were extracted from chromosome 21, which were further grouped into four subsets, as detailed in Table 5.2. A total of 440 analyses were performed, and the predictive accuracies obtained through these analyses are summarised and presented in Table 5.21. Single featuresets, dinucleotide distribution (f3) and evolutionary and conservation subset (f2) showed the highest class performance, in addition to predictive class accuracy, with a total accuracy of 77.41% and 70.34%, and standard errors of 6.10 and 3.41 respectively. Evolution and conservation (f2) gave effective predictive class performance, which confirms the results for chromosomes 6 and 22 described above. However, the two other features (f1 and f4) showed fluctuations in class performance; these may represent 'noisy' features.

The association between at least two combined feature subsets was then investigated. This shows that the accuracy steadily increased, whilst the class performance remained approximately equivalent to the single subset. The best class performance yield when tissue-specific features (f1) and dinucleotide distribution (f3) were combined was f1 combined with f2. The total predictive accuracies of both classes (methylated and unmethylated) were 69.66% and 80% respectively. Featuresets/values showed in bold are those with the highest class performance in Table 5.21. Combining paired featuresets (f2,f3; f2,4; or f3,f4) shows overall higher predictive accuracies despite a decrease in predictive performance. The three featuresets combinations (tissue-specific feature (f1), Evolution and conservation (f2) and dinucleotide distribution (f3)) achieved the highest accuracy, in addition to an effective class performance when compared to those of any of the other three combinations. Furthermore, combining f4 with other two feature subsets resulted a lower class predictive performance, although the total accuracy did not change substantially. This may have been caused by some of the physiological and chemical properties not complementing other feature subsets. In general, combinations of four feature subsets slightly reduced class performance, whereas three feature subset combinations showed consis-

Table 5.21 Highest mean predictive accuracies (%) and standard error for combinations and individual of the feature sub-sets arsing from the M-LOO-based analysis of chromosome 21[6].

No of features	combined feature-sets	pred-Acc(M)	pred.Acc.(uM)	Tot-pred-Acc	standard-error
single feature	f1	65.52	79.31	72.41	9.75
	f2*	**58.62**	**68.97**	**63.79**	**7.31**
	f3	68.97	79.31	74.14	7.31
	f4	51.72	62.07	56.90	7.31
2 features	{f1,f2}	72.41	75.86	74.14	2.44
	{f1,f3}*	**65.52**	**79.31**	**72.41**	**9.75**
	{f1,f4}	51.72	62.07	56.90	7.32
	{f2,f3}	51.72	68.97	60.35	12.19
	{f2,f4}	51.72	62.07	56.90	7.31
	{f3,f4}	66.20	86.55	76.38	14.39
3 features	**{f1,f2,f3}***	**72.41**	**75.86**	**74.14**	**2.44**
	{f1,f2,f4}	55.17	62.07	58.62	4.88
	{f1,f3,f4}	51.72	62.07	56.90	7.31
	{f2,f3,f4}	51.72	62.07	56.90	7.32
4 features	{f1,f2,f3,f4}	67.59	88.97	78.28	15.12

* bolded feature sub-sets/values shown are those which exhibit the highest class performance.

tent class performance, particularly when feature subset (f4) was excluded. However, any combination that excluded the tissue-specific feature (f1) exhibited the lowest class performance.

5.3.5 overall summary of four chromosome analysis/discussion

For single feature-sets, throughout the analysis of the three chromosomes (6, 20 and 22), tissue-specific CpGI methylation showed the highest total predictive accuracy when compared to that of the other three individual feature subsets, followed by dinucleotide distribution and DNA structure. These showed approximately the same predictive accuracies, whilst the evolution and conservation subset had the lowest class performance level, as well as predictive class accuracy throughout these three chromosomes. A combination of two or more feature subsets revealed how the predictive accuracy changed, specifically which subsets had the highest accuracy and class performance throughout the four chromosomes. The results showed that tissue-specific CpGI had the highest predictive accuracy when combined with either evolutionary and conservation, or dinucleotide distribution and DNA structure throughout the analysis. This confirmed the recent finding that tissue specificity is the most influential factor for DNA methylation status [193].

Chromosome 21 showed a lower accuracy when compared to the three other chromosomes. This study investigated whether it is possible to predict the three

datasets and the combinations of methylated *versus* unmethylated, methylated *versus* differentially-methylated, and methylated *versus* differentially-methylated. The analysis showed that predictive accuracy was reduced when differentially-methylated and unmethylated classes were combined for all chromosomes. It also showed a reduced class performance and some fluctuations that may have caused the 20-fold cycle, in which some samples have a reduced accuracy. For small samples, these 'noisy' samples may have reduced the accuracy, since the mean average of those of 20-fold predictive models was employed. However, the differentially methylated and methylated showed consistent results for all chromosomes. In terms of combinations of three subsets, the tissue-specific CpGI feature one was also the most predictive subset when combined with the other three feature ones. This also showed a reduction in total accuracy, with a small increase in standard error value. In contrast, when the tissue-specific CpGI subset was excluded from the analysis, almost all combinations had reduced accuracies as well as lower sensitivities or a positive predictive rate. Differentially-methylated and unmethylated combinations were found to have lower predictive accuracies when compared to either differentially-methylated *versus* methylated, and methylated *versus* unmethylated classes. Chromosome 21 showed a higher predictive accuracy when the differentially-methylated and unmethylated classes were combined, particularly in terms of sequence properties and distribution (f1). The combination of methylated, and either unmethylated or differentially-methylated subsets without subset (f1) gave a lower predictive accuracy, and sensitivity.

Furthermore, exon and gene distribution showed the highest accuracy when compared to the other three feature-subsets for two classes (methylated and unmethylated), whilst differentially-methylated and unmethylated gave the highest predictive accuracy for DNA sequence properties and distribution with a good predictive performance level. In addition, the combination of methylated and differentially-methylated classes also showed a good predictive accuracy with 93% sensitivity. Sequence properties and distribution of methylated and unmethylated classes had an accuracy of 74.41%, with a small decrease in sensitivity and class performance, which skewed to the specificity side. The other three single subsets showed some fluctuation in their accuracies and predictive performances. In addition, two-subset combinations in the three-class prediction model showed an overall increase in both predictive accuracy and class performance. This was clear when DNA sequence properties and distribution were included in the analysis. Furthermore, the feature-subset combination showed both a better predictive accuracy and class performance, with a smaller standard deviation than the individual subset analysis.

The three-class analysis feature combination showed a higher accuracy, particularly when the tissue-specific CpGI (chromosomes 6, 20 and 22) and sequence properties and distribution (Chromosome 21) subsets were present in the combination. Furthermore, exon and gene distribution gave the highest accuracy compared to the other single

subset, followed by DNA sequence properties and distribution.

The two-feature combination incorporating DNA sequence properties and distribution, and exon and gene distribution, provided an improved predictive performance than the other pairs in the feature subsets. When three-features subsets were combined, exon and gene distribution, evolution and conservation, and SNPs, showed a higher accuracy of 94.44%, with a good class performance. It should also be noted that class performance was measured, with a mean standard deviation of 2.56%. Performances increased by 32.80% when feature subsets were combined through all subsets from two- to six-feature ones. This indicates that a combination of features gave an improved predictive power when compared to single feature subsets. Biological functionality thus depends not only on one factor, but rather on a multifactorial systems that can be linked to DNA methylation. However, DNA methylation depends on DNA sequence contexts, such as sequence density, composition, length and environmental influences. The model used in this study was effective in distinguishing between three classes and their feature subsets when compared to those used in previous studies.

5.3.6 Conclusion

CpG islands of four human chromosomes (6, 20, 21 and 22) were studied to distinguish methylated, unmethylated and differentially-methylated samples based on DNA sequence features. Data were extracted from 642 DNA samples and a prediction was carried out to determine whether the extracted features were associated with the methylated, unmethyalted and differentially-methylated classes. The combinations of feature subsets that were more correlated with these three classes were also determined. DNA methylation prediction is based on sequence-driven features, which have been widely studied using machine learning techniques. This study presented a new method that examine the features of CpG $islands^{'}$ sequences. Here I have analysed three classes, namely methylated, unmethylated and differentially-methylated ones. Each analysis examined individual instances, as well as all possible combinations of the three classes. This approach distinguishes not only methylated, unmethylated and differentially-methylated classes, but in contrast to previous literature available, also dealt with unbalanced data. This method (M-LOO) can be used for the two-class problem in binary classification approaches. indeed, it was possible to select the best feature subsets when two or more feature subsets were combined for very unbalanced class data. This method improved predictive accuracy as well as class performance. It showed that sequence properties and distribution, tissue-specific CpGI and DNA structure feature subsets had the highest predictive accuracies. The method used in this study (Modified Leave-One-out cross validation) was able to predict methylated from both unmethylated and differentially-methylated classes, which gave an improved predictive accuracy and class performance when compared to traditional methods. The tissue-specific (f1), dinucleotide distribution (f3) and DNA structure

feature sets (f4) showed the highest performance when expressed to relative to all other three featureset combinations, followed by the two feature-set combination of tissue-specific features (f1) and DNA structure (f4).

When comparing the predictive accuracies of single- and two-feature sets, all combinations of feature-sets showed an increase in predictive performance of approximately 30% occurred. The results showed a significant correlation between tissue-specificity and sequence patterns, as well as DNA physiocochemical structure. The results obtained through a comprehensive analysis of all possible combinations of the feature subsets show that both methylated and differentially-methylated CpG islands can be distinguished from unmethylated CpG islands by using the novel method described above. This potentially identified a predictive accuracy of between 75% and 100% based on a DNA sequence context. It was found that DNA methylation tends to be associated with tissue-specific CpG islands, whereas the differentially-methylated forms are more correlated with DNA sequence distribution. Notably, the modified LOOCV shows high reducibility and is more robustly identified when the feature subsets were combined. Another interesting observation is that the modified-LOO-based analysis showed that the tissue-specific CpGI feature-set achieved the highest predictive accuracy when combined with the other feature sets. This also further supports the robustness of the modified-LOO cross-validation approach, since tissue-specific CpGI and DNA sequence properties feature-sets are one of the most important and effective attributes demonstrated in previous studies.

The methods used in this study could be compared with feature selection methods in order to determine whether the same feature subsets could be selected from the dataset. Further investigations will be required to evaluate the results obtained from sample clustering in order to select the same subset that was selected for the predictive model in this study. Nevertheless, the DNA features described in this study, and their predictive accuracy based on the three classes described in this chapter serve as the initial basis for understanding DNA methylation patterns throughout a range of healthy tissues.

5.4 Weighting methods towards severely imbalanced data

This section provides an effective predictive statistical model for severely imbalanced DNA methylation classes, and aims to create a balanced outcome, hence limiting the bias interference from differences inherent within the classes. Furthermore, performances of the balanced outcome were evaluated using six metrics: precision, sensitivity, specificity, accuracy, F-measure and the G-mean. This empirical method shows some improvements, particularly in terms of the predictive performance of the minority classes when compared to those achieved with unweighted algorithms.

5.4.1 Introduction

Misclassification of skewed data is common in many fields in which the number of samples in one class exceeds those in the other (minority class) by a large amount. This problem represents a major challenge for the data-mining community, since the performances of machine-learning outcomes are severely diminished.

The Boost algorithm is generally employed to reduce the classification error rate by adding a small weighting vote to each step of the training process prior to proceeding to the next stage [116]. The Ensemble learning method has been applied to various classification problems, and has been found to perform better when compared to results acquired from the most sophisticated learning algorithms. For example, AdaBoost is a linear combined single classifier that can produce small error rates in the training datasets. Indeed, the use of AdaBoost has markedly increased in recent years [124; 195; 196]. Since it serves as a combined classification system, it produces excellent predictive accuracy, as documented in many studied cases. However, two forms of misclassification can occur: (I) unequal classes or observations and (II) class separate-ability, where classes overlap and cannot be discriminated with an additive weighting rule.

The technique of cost-proportional rejection has been proposed with the aim of reducing misclassification rates by weighting training data through partitioning (under-sampling) into subsets, and averaging the outputs arising from this[125]. The stacking method, in which the majority class is divided into subsets in a non-overlapping manner during the training phase, and the outputs averaged, has also been proposed. However, this process has been found to be computationally- expensive and overfits the minority class[129]. Moreover, it has the limitation of not making adequate allowances for small samples. In addition, Liu [128] reported a balance cascade and ease ensemble approach relating to sampling when incorporating the AdaBoost algorithm. Other studies have reported imbalance and data analysis problems, including both under- and oversampling [117; 197; 198; 199; 200]. Under- sampling techniques offer the advantage of being balanced and faster in training, although they suffer from ignoring selected useful information. Furthermore, Ali [5] proposed a modified leave-one-out method as reported in Chapter (3.2.1), which is similar to under-sampling methods [128]. This method does not use probability or density interpolation; more importantly, all data samples are considered. In this method, a minority sample is sequentially added to the equally-portioned majority class during training throughout the entire dataset. Again, all samples are considered; this method also averages the outputs. In this manner, the output predictive accuracy represents the balance in which both the majority and minority classes are considered equivalently during the analysis. This method has less bias than the other approaches, such as extrapolation, density and prior probability, and operates by manipulating the original dataset,

whereas the modified leave-one-out approach uses the original dataset.

Seiffert *et.al* [201] described an under-sampling technique that claims to provide improved results when compared to those achieved using over-sampling methods such as SmoteBoost [126]. Seiffert *et.al* [201] compared the predictive performances of the AdaBoost, SmoteBoost and RusBoost techniques, and these algorithms showed similar levels of generalisation. AdaBoost was described as one of the best algorithms for small imbalanced datasets [116; 124]. However, the authors did not test this approach on multidimensional datasets, leading to severe imbalances in their model. In order to limit these effects, two methods were combined in this study: prior probability (empirical weighting) and cost sensitivity. This combination showed improvements in the predictive performance. The combined method also leads to a reduction of the misclassification rate of minority classes during the training phase when compared to a single weighing method in which the minority is added to high cost in order to balance the sample proportions of the two classes. The weighted cost method was therefore tested against the number of false negatives and false positives in order to scale the error rate during training. This approach can be employed with any machine-learning algorithm. The choice will also depend on prior knowledge of the learning algorithm used. For example, in the case of AdaBoost, the method is used with a prior probability consideration, which involves the addition of the weighted classes in the maximal probability to one. During this procedure, the training error is calculated and scaled. The highest misclassified class (minority) is hence weighted heavily, whilst the majority class is weighted lower in response to the misclassification rate, so that the learning algorithm continues to be fed with a scaled misclassification proportion.

5.4.2 Material and methods

The dataset used in this study comprised 642 samples from four human chromosomes that were partitioned into three sub-classes as presented in (previous section,Table 5.1). The Gentle AdaBoost method was applied for two-class problems, and the AdaBoostM2 method for three-class ones. AdaBoost minimises the exponential loss. The numeric optimisation is, however, set up differently. The AdaBoost method provided a good level of classification success. However, it is very expensive computationally, and the modified leave-one-out approach has a slightly better predictive accuracy with lower computational complexity. Despite AdaBoost being computationally-expensive, it remains a good candidate for small sample datasets ($20 \leq$). This particularly applies to binary classification problems with imbalanced classes. In addition to incorporating a weight and cost matrix, weighting was also applied in the first iteration, with further updates within the Adaboost rule at the level of weight, as illustrated in Figure 5.1. This leads to adjustments in the weight during the training phase of the process in response to the misclassification rate of the minority class.

As shown in Table 5.22, the unmethylated class has more observations than the

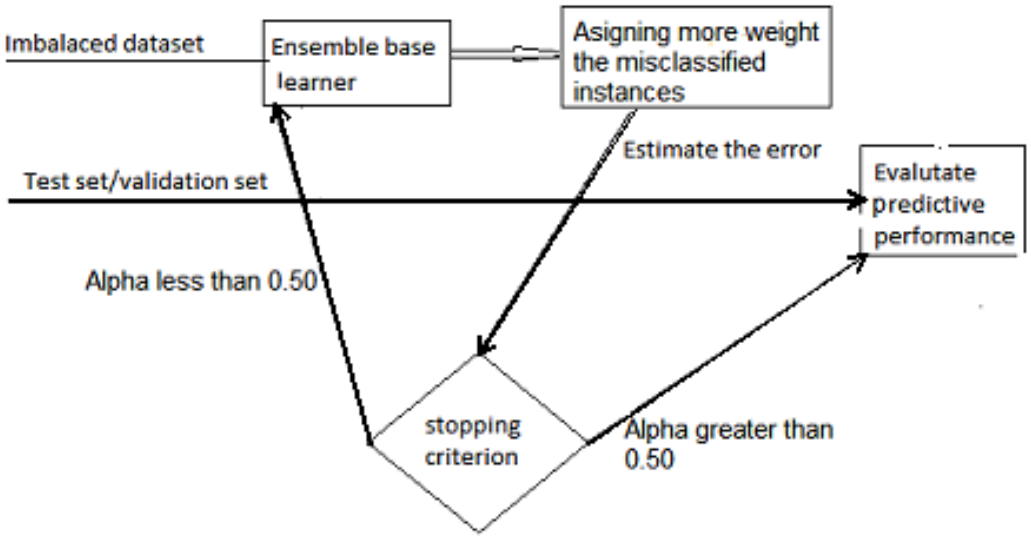


Figure 5.1 Imbalanced dataset classification process with weighting criterion.

differentially-methylated one and is thus more expensive. Consequently, fewer observations were generated for the majority class than for the minority ones (methylated and differentially-methylated). It was assumed that the classes (methylated, unmethylated and differentially-methylated) were mixed in different proportions; a prior probability was therefore set up for the three classes, and that was comparable to the values that we would expect to observe in a symmetrical dataset. Furthermore, the fit ensemble method was applied, which normalises prior probability and adds up to a value of one. Hence, it is expected that the outcome will be more realistic than if it were deduced when randomly based on class proportions. In addition, the majority class (unmethylated) is asymmetric and unequally represented in the training set. A cost parameter for assigning a high cost to the most misclassified class (minority), and a lower one to the majority class (unmethylated) in the dataset was applied. This led to a reduction in the misclassification level of either the methylated or the differentially-methylated (i.e, the minority) classes, since it placed more weight on their training. Furthermore, misclassification is passed as a square matrix in the training set. For example, β(i, j) represents the true class of i as observation j; β(i, j) corresponds to the cost matrix of the classified observation into class j, since i is the true class, whereas the diagonal matrix (i, i) of the cost matrix will be zero. In addition, both the methylated and the unmethylated classes were incorporated into the cost matrix:

$$\begin{pmatrix} 0 & \beta & \text{minority class(methylated)} \\ 1 & 0 & \text{majority class(unmethylated)} \end{pmatrix}$$

In this example, the cost of misclassification of the methylated class is $\beta 1$, where two classes are adjusted, the prior probability (p_i) of the two classes is acquired by using $p_i = \beta_i \times P_i$ for class i = 1,2, and $j \neq i$.These prior probabilities are either passed

into fit ensemble, or alternatively computed from class frequencies in the training dataset, and subsequently the ensemble fits the default cost matrix [202], as described below:

$$\begin{pmatrix} 0 & \beta \\ 1 & 0 \end{pmatrix}$$

Fit ensemble uses the adjusted probabilities and cost matrix for training its weak learners. Altering the cost matrix manipulates the prior probabilities where the probability to be selected for the minority class increases. Furthermore, the minority class is weighted heavily against the majority one. For example, when the dataset is in the form of a 1:10 proportion, it should supply the misclassification-based scaling so that the cost matrix reduces the misclassification rate of the minority class as shown in Table 5.22.

Fit ensemble uses the adjusted probabilities and cost matrix for training its weak learners. Altering the cost matrix manipulates the prior probabilities where the probability to be selected for the minority class increases. Furthermore, the minority class is weighted heavily against the majority one, for instance when the dataset is the in the form of a 1:10 proportion, it should supply the misclassification-based scaling so that the cost matrix reduces the misclassification rate of the minority class as shows in Table 5.22.

Table 5.22 shows weighted and unweighted overview of the unmethylated *versus* differentially-methylated classes of chromosome 6, both unweighted and weighted

Table 5.22 Weighted and unweighted overview of unmethylated *versus* differentially-methylated classifications of chromosome 6

unweighted		weighted		
differentially-methylated	2	6	5	3
unmethylated	1	124	1	124

The two-class prediction of the Gentle AdaBoost algorithm is described as follows: It is assumed that training data is$(x_1, y_1, \beta_1, , (x_N, y_N, \beta_N)$ where x_i is represents a value of attributes and y_i represents the class label with $Y = \pm 1$. If the function $G(x) = \sum_{t=1}^{T} \beta_t C_t g_t(x)$ is the one iteration of weighted base classifier (where its output has the values of ± 1), $g_t(x)$ is the base classifier for which β_t is the cost matrix, and C_t represents a constant value. The final output is the sum signG(x). The base classifier (gt(x)) is trained by weighting those samples that have the most misclassification numbers in an additive manner. This is performed by weighting each round of the base classifier, where the final learner, signG(x), is from the sum of a linear combination of the base classifier ($g_t(x)$). This linear additive model is used as a basis to build a strong classifier that overpowers the single method, as shown in Box 5.1. The summary of Gentle AdaBoost is described in [118]. Gentle AdaBoost uses an estimated class probability to update the weighting function of a training iteration,

where the update is $g_t(x) = p_d(y = 1/x - p_d)(y = -1/x)$, instead of the half log ratio, as utilised in some AdaBoost versions [118]. The same paper found that half-log ratios were unstable during the update process, which resulted in a large variations in the boundary.

Box 5.1 Gentle AdaBoost algorithm combined with cost-matrix

Gentle AdaBoost

- Initialise weights for D (i) = 1/n where i = 1, 2.., N and G(x) = 0.
- Repeat t-times for t = 1,.,T: where each round D(i) is updated
- Use the fitensemble classification function of $(g_t(x))$, the base classifier weighted with least-squares of y_i to x_i with weights D (i).
- Update $G(x) \leftarrow G(x) + g_t(x)$.
- Update $D(i) \leftarrow D(i)exp(-y_i(g_t(x_i))$ and then normalise.
- The final output of the classifier is sign $G(x) = [\sum_{t=1}^{T} \beta_t C_t g_t(x)]$.

Additionally, the predictive performance of both parameters were compared with respect to seven metrics, specifically precision, recall, sensitivity, specificity, mean-weighted accuracy, F-measure (f-score) and Geometric mean (G-mean) by assessing the reliability of the algorithms used on imbalanced datasets. The seven metrics are extracted from a confusion matrix (Table 5.23), and are commonly used to evaluate predictive performances. It has been reported that a single metric performance assessment is of no value since provides biased information [203]. Indeed, the author of this work compared various metrics for assessing predictive performance.

These results were compared with those previously published by the researchers. All the analyses were repeated ten times to assess model stability, and iterated (t=100) for each weak learner. In this case, a classification tree was used as a base classifier incorporated with five-fold cross validation, since some of the classes contained less than ten samples. Furthermore, averages of a standard deviation and the six metrics values from Table 5.23 were calculated in each repeated analysis.

Table 5.23 Confusion Matrix

	Predicted as a positive class	Predicted as a negative class
Definite positive class	True positive (TP)	False negative (FN)
Definite negative class	False positive (FP)	True negative(TN)

- Precision = $\frac{TP}{TP+FP}$
- Sensitivity = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$

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

$$G - mean = \sqrt{Sensitivity.Specificity} \tag{5.1}$$

$$F - measure = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{5.2}$$

5.4.3 Results and Discussions

DNA methylation has recently been successful predicted, particularly for two-class classification problems, via the use of DNA sequence features in order to distinguish differences between methylated and unmethylated classes. However, imbalanced classes have not been taken into account in situations where these predictions were not representative of data.

The results contain DNA sequence feature data obtained from four human chromosomes: 6, 20, 21 and 22. These datasets are quite imbalanced, and for this reason, AdaBoost algorithms were used in combination with weighting models in order to overcome the bias outcome that is usually used for imbalanced data analysis. Two models of data weighting were employed, specifically (i) an empirical method: a prior-probability was set for the training in order to combine the two classes into single one, (ii) assigning a high cost to the site of the minority class based on its misclassification rate. The dataset represents samples of three classes: methylated, unmethylated and differentially-methylated. Both the methylated and differentially-methylated groups represent the minority classes; a high cost was placed on the methylated class (minority) samples in order to compensate for the outcome of the predictive accuracy, and hence reduce the misclassification error rate of this class. The results are presented in Tables 5.24 to 5.31.

5.4.4 Experimental results of two classes analysis

In order to evaluate the predictive performance of methylated versus unmethylated classification successes, the prior probabilities of both classes were added up to a total of one. This showed some improvement for imbalanced datasets in order to increase the predictive accuracy, and also to diminish the minority class misclassification error rate when prior probability and cost-sensitive approaches were combined. As is notable in Table 5.24, both the methylated and unmethylated classes showed a significant classification improvement when combining these two methods. This indicates that the sensitivity and specificity levels showed significant improvements, since the predictive accuracies did not show much change (as expected). Therefore, the sensitivity levels of the minority classes increase, whereas the error rate of their misclassification decreases. Indeed, the sensitivity increases from 75±14.0 to 77±9.0%; that is, the results show a reduction in standard error value from 14 to 9%. In addition, the methylated

versus differentially-methylated class comparisons showed an a significant improvement of sensitivity (predictive accuracy of the minority class), namely an increase from $71 \pm 10.0\%$ to $75 \pm 6.0\%$, and a reduction in the standard error from 10 to 6% respectively, as shown in Table 5.24. However, there was no significant improvement when the differentially-methylated group was compared to the unmethylated ones, although combining Adaboost with the cost-sensitive approach showed some improvement, particularly after the 30^{th} round of the base classifier. This is consistent with previous reports in which it is barely possible to distinguish differentially-methylated and unmethylated DNA [29] forms from each other. Furthermore, as indicated in Table 5.24, unmethylated versus differentially-methylated DNA class comparisons provided a more favourable improvement, in which the misclassification error rate of the minority class decreased, and concomitantly, the sensitivity increased from $82 \pm 5.0\%$ to $93 \pm 8.0\%$. The methylated *versus* differentially-methylated classification comparisons also show an increase from $55 \pm 18.0\%$ to $60 \pm 9.0\%$, with 50% reductions in standard error. Although the standard error value is relatively high, the error is reduced by 50% when the two methods are combined. Furthermore, results for the methylated and unmethylated classification comparison showed a small sensitivity increase from $75 \pm 9.0\%$ to $78 \pm 9.0\%$. The error rate of the majority (unmethylated) class was reduced. In addition, the G- and f-mean values indicate a consistent increase, whereas the overall predictive accuracies revealed small improvements.

Table 5.24 Representative predictive accuracy and performance assessment of the methylated and differentially-methylated samples for chromosome 6

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
D-met vs unmet	1,1.5	65.0 ± 12.0	49.0 ± 9.0	98.0 ± 1.0	95.0 ± 1.0	55.0 ± 8.0	69.0 ± 6.0	unweighted
D-met vs unmet	1,15.6	68.0 ± 6.0	48.0 ± 11.0	99.0 ± 1.0	95.0 ± 1.0	55.0 ± 9.0	68.0 ± 9.0	weighted
Met vs D-met	1.5,1	83.0 ± 6.0	90.0 ± 9.0	71.0 ± 10.0	83.0 ± 8.0	86.0 ± 7.0	80.0 ± 8.0	unweighted
Met vs D-met	1.5,1	85.0 ± 3.0	92.0 ± 6.0	75.0 ± 6.0	85.0 ± 4.0	88.0 ± 3.0	83.0 ± 4.0	weighted
Met vs unmet	1,10	88.0 ± 13.0	75.0 ± 14.0	97.0 ± 4.0	92.0 ± 4.0	80.0 ± 12.0	0.85 ± 0.08	unweighted
Met vs unmet	1,10	84.0 ± 16.0	77.0 ± 9.0	96.0 ± 5.0	$92. \pm 5.0$	80.0 ± 11.0	86.0 ± 6.0	weighted

Abbreviations: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

Table 5.25 represents the results obtained using both weighted and unweighted samples of three classes of chromosome 20. Clearly, the combined method significantly improved the predictive accuracies significantly. This chromosome contained the smallest sample size when compared to the datasets from the other three chromosomes. Inter-class differences, and also the methylation and unmethylation classes, were severely imbalanced. The proposed method gives rise to a significant predictive per-

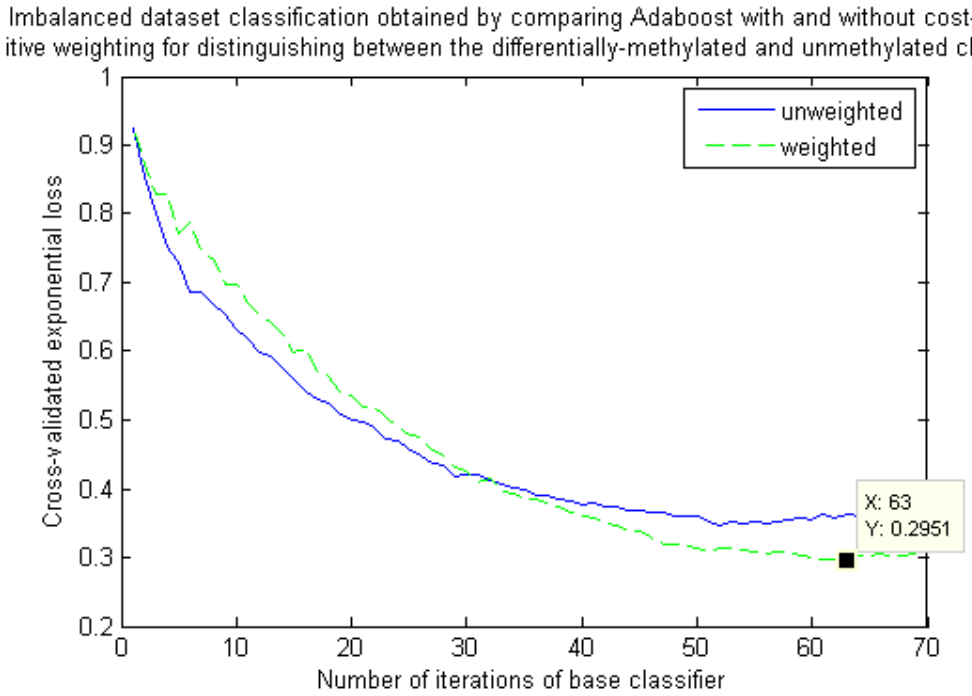


Figure 5.2 Comparison of Adaboost with the cost-sensitive approaches for distinguishing between differentially-methylated *versus* unmethylated DNA sequence classes. This is a two-class problem in which the error rate (cross validation exponential loss) was plotted as a function of number of base classifier iterations.

formance improvement when expressed relative to results obtained with moderately-imbalanced classes. Moreover, the unmethylated and differentially-methylated classes also showed improvements, where both G-means (89.0 ± 4.0 to $95.0 \pm 5.0\%$) and f-measures ($94.0 \pm 2.0\%$ to $97.0 \pm 3.0\%$) gave higher performance values when the weighting and cost-sensitive strategies were combined. The methylated and differentially-methylated classification comparisons provided the highest predictive accuracy improvement of the minority class, and displayed a standard error decrease of 50% (from 18% to 9%). The G-means and f-measure showed improved results when the two methods were combined. Moreover, the overall predictive accuracy showed a considerable improvement for the combined approach.

Table 5.25 Representative two-class predictive accuracies and performance assessment for comparisons of the methylated, unmethylated and differentially-methylated samples for chromosome 20.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
unmet vs D-met	1.9,1	91.0 ± 2.0	98.0 ± 3.0	82.0 ± 5.0	92.0 ± 3.0	94.0 ± 2.0	89.0 ± 4.0	unweighted
unmet vs D-met	1.9,1	96.0 ± 4.0	99.0 ± 2.0	93.0 ± 8.0	97.0 ± 4.0	97.0 ± 3.0	95.0 ± 5.0	weighted
Met vs D-met	1,2	54.0 ± 11.0	55.0 ± 18.0	78.0 ± 7.0	70.0 ± 7.0	54.0 ± 14.0	64.0 ± 12.0	unweighted
Met vs D-met	1,2	61.0 ± 8.0	60.0 ± 9.0	80.0 ± 6.0	73.0 ± 4.0	60.0 ± 7.0	69.0 ± 5.0	weighted
Met vs unmet	1,3.8	82.0 ± 12.0	75.0 ± 9.0	95.0 ± 4.0	91.0 ± 4.0	78.0 ± 9.0	84.0 ± 5.0	unweighted
Met vs unmet	1,3.8	84.0 ± 15.0	78.0 ± 11.0	96.0 ± 5.0	92.0 ± 5.0	81.0 ± 11.0	86.0 ± 7.0	weighted

Abbreviation: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

Moreover, as shown in Table 5.26, the chromosome 21 samples were examined in which the classes were severely imbalanced. Performance levels for the differentially-methylated and unmethylated classes improved significantly (from $81 \pm 4.0\%$ to $88 \pm 5.0\%$), and the overall accuracy remained unmodified (as expected), whilst the G-means values showed small improvements. However, the error rate did not indicate any significant improvement in the results obtained. Furthermore, the methylated and unmethylated classes showed no significant improvement at the level of predictive accuracy. In addition, methylated *versus* differentially-methylated classification comparisons failed to show significant improvements, despite the fact that the G-means value showed a small improvement.

As shown in Table 5.27, analyses of chromosome 22 (methylated *versus* methylated classes) showed no significant changes. This is attributable to the larger sample size of this chromosome (when compared to the other three chromosomes). Prior probability was therefore sufficient to provide an effective level of predictive accuracy. There were no improved outcomes when combined with the cost-matrix (which weighted the

Table 5.26 Representative two-class predictive accuracies and performance assessments for distinguishing between the methylated, unmethylated and differentially-methylated classes of chromosome 21.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
unmet vs D-met	6.9,1	97.0 ± 1.0	98.0 ± 1.0	81.0 ± 4.0	96.0 ± 1.0	98.0 ± 0.0	89.0 ± 2.0	unweighted
unmet vs D-met	6.9,1	98.0 ± 1.0	98.0 ± 1.0	88.0	96.0 ± 1.0	98.0 ± 0.0	93.0 ± 3.0	weighted
Met vs D-met	2,1	96.0 ± 3.0	92.0 ± 3.0	93.0 ± 7.0	93.0 ± 3.0	94.0 ± 2.0	92.0 ± 4.0	unweighted
Met vs D-met	2,1	91.0 ± 4.0	92.0 ± 5.0	82.0 ± 8.0	89.0 ± 3.0	92.0 ± 2.0	87.0 ± 4.0	weighted
Met vs unmet	1,3.6	89.0 ± 3.0	63.0 ± 7.0	98.0 ± 1.0	90.0 ± 1.0	74.0 ± 4.0	79.0 ± 4.0	unweighted
Met vs unmet	1,3.6	92.0 ± 5.0	61.0 ± 7.0	98 ± 1.0	90.0 ± 2.0	73.0 ± 5.0	77.0 ± 4.0	weighted

Abbreviations: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

minority class). This suggests that larger samples are less influenced by imbalanced data analysis than are small samples; chromosomes 6, 20 and 21 were severely affected when compared to chromosome 22, where both weighting methods showed the lowest level of improvement in the results acquired.

Table 5.27 Representative two-class predictive accuracies and performance for comparative assessments of the methylated, unmethylated and differentially-methylated classes of chromosome 22.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
unmet vs D-met	9,1	96.0 ± 0.0	97.0 ± 1.0	64.0 ± 4.0	94.0 ± 1.0	97.0 ± 0.0	79.0 ± 3.0	Un-weighted
unmet vs D-met	9,1	97.0 ± 1.0	97.0 ± 1.0	70.0 ± 7.0	94.0 ± 1.0	97.0 ± 0.0	82.0 ± 4.0	weighted
Methylated vs D-met	2.75,1	92.0 ± 2.0	96.0 ± 1.0	78.0 ± 6.0	91.0 ± 2.0	94.0 ± 1.0	86.0 ± 3.0	Un-weighted
Methylated vs D-met	2.75,1	92.0 ± 1.0	96.0 ± 1.0	78.0 ± 4.0	91.0 ± 1.0	94.0 ± 1.0	87.0 ± 2.0	weighted
Methylated vs unmet	1,3.3	98.0 ± 1.0	97.0 ± 2.0	100.0 ± 0.0	99.0 ± 1.0	98.0 ± 1.0	98.0 ± 1.0	Un-weighted
Methylated vs unmet	1,3.3	99.0 ± 1.0	97.0 ± 2.0	100.0 ± 0.0	99.0 ± 1.0	98.0 ± 1.0	99.0 ± 1.0	weighted

Abbreviations: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

5.4.5 Experimental results arising from three-class analysis

Additionally, three-class problems were also investigated, and one class was examined *versus* the others. For chromosome 6, the methylated classification versus the remainder yielded a high predictive accuracy when compared to other combinations, which

gave predictive accuracies of 92.3 ± 2.0% (positive = sensitivity) and 100.0 ± 0.0% (negative predictive accuracy). The unmethylated class *versus* all the others showed an improved predictive accuracy. Prior probability combined with the cost-sensitive model showed a marginally improved predictive accuracy (77.6±10.0% to 78.9±4.0%) when compared to that achievable with a single model. Additionally, the differentially-methylated group *versus* the remainder gave only a lower predictive accuracy (predictive positive rate); however, the overall predictive accuracy indicates an increased positive predictive rate from 46.3 ± 5.7% to 53.6 ± 1.9%, with a reduced standard error from 5.7 to 1.9%). However, whilst chromosome 6 was studied in [64], only the overall accuracy was examined, which may not be reliable. Moreover, results do not equivalently represent those from the two classes involved, since the dataset is severely imbalanced.

Table 5.28 The results of three class predictive accuracies and performance assessment for comparisons of the methylated, unmethylated and differentially-methylated classes of chromosome 6.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
Methylated vs all	1,11	100.0 ± 0.0	92.3 ± 0.0	100.0 ± 0.0	99.3 ± 0.0	96.1 ± 0.0	96.0 ± 0.0	unweighted
Methylated vs all	1,11	100.0 ± 0.0	92.3 ± 0.0	100.0 ± 0.0	99.3 ± 0.0	96.1 ± 0.0	96.0 ± 0.0	weighted
unmet Vs all	6.3,1	96.0 ± 0.0	98.0 ± 0.1	77.6 ± 0.1	94.7 ± 0.7	87.2 ± 0.9	96.9 ± 0.4	unweighted
unmet Vs all	6.3,1	96.0 ± 0.0	99.0 ± 0.3	78.9 ± 0.4	95.7 ± 0.3	88.4 ± 0.4	97.5 ± 0.2	weighted
D-met vs all	1,17	55.0 ± 12.0	46.3 ± 5.7	97.4 ± 0.7	94.1 ± 0.7	67.0 ± 0.4	50.2 ± 8.4	unweighted
D-met vs all	1,17	72.5 ± 5.3	53.6 ± 1.9	98.4 ± 0.3	95.0 ± 0.3	72.6 ± 1.4	61.6 ± 3.2	weighted

Abbreviations: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

Chromosome 20 contains small samples, and an imbalance between the classes, in which the ratio is (1, 1.5, 0.5 for the methylated, unmethylated and differentially-methylated classifications respectively). Comparison of the methylated classification with the remainder resulted in significant improvements in predictive accuracies, as well as in the misclassification rate of the methylated ones (minority class); both gave an increase in predictive accuracy, and a reduced standard error of 61.0 ± 23.4% to 64.1±18.1%, and 23.7% to 18.1% respectively. For the methylated classification *versus* the remainder, an improved predictive accuracy was obtained, where the positive prediction of the minority class increased from 63.4 ± 4.2% to 70.5 ± 7.5%, although the associated error rates showed no modification. Other researchers have tried to classify methylated- and unmethylated-based DNA sequences. However, these studies did not consider imbalanced datasets, and their predictive assessments are limited to overall accuracies and correlation coefficients (MCC) [64].

Table 5.29 Representative results acquired for comparisons of three-class predictive accuracies and performance assessments of unmethylated and differentially-methylated classes for chromosome 20.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
Methylated vs all	1,5.8	31.7 ± 5.3	61.0 ± 23.4	89.0 ± 0.6	85.9 ± 2.8	72.5 ± 13.7	39.7 ± 6.2	Unweighted
Methylated vs all	1,5.8	56.6 ± 11.7	64.1 ± 18.1	92.7 ± 1.6	87.8 ± 3.8	76.4 ± 10.6	57.8 ± 8.6	weighted
unmet Vs all	1.3,1	95.6 ± 4.1	88.9 ± 4.1	94.1 ± 5.4	90.7 ± 3.2	91.4 ± 3.3	92.1 ± 2.7	Unweighted
unmet Vs all	1.3,1	99.1 ± 2.8	84.2 ± 0.03	98.8 ± 4.0	89.0 ± 2.6	91.2 ± 2.5	91.0 ± 2.1	weighted
D-met vs all	1,2.4	66.7 ± 10.4	63.4 ± 4.2	86.1 ± 3.7	79.0 ± 3.1	73.8 ± 3.8	64.8 ± 6.7	Unweighted
D-met vs all	1,2.4	47.5 ± 11.2	70.5 ± 7.5	81.0 ± 3.1	78.8 ± 3.1	75.5 ± 4.8	56.1 ± 9.4	weighted

Abbreviations: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

As shown in Table 5.30, when assessing results obtained using methylated chromosome 21 *versus* the other two classes (unmethylated and differentially-methylated), the level of accuracy increased to $92.8 \pm 4.8\%$, with no signs of improvement in the predictive accuracy of the minority class when combining both methods. In contrast, the G-means and f-measure values showed small improvements. For weighting, however, there was no observed improvement for the misclassification error rate for the minority class. Furthermore, for the unmethylated versus all the remaining group comparisons, no significant improvements were indicated after weighting, whereas the differentially-methylated classification versus the remainder all showed significant improvements in the predictive performance of the minority classes of $100 \pm 0.0\%$ and $99.2 \pm 2.8\%$ respectively. Although there was no significant improvement in the predictive performance of the minority class, the standard error values was reduced when both models were combined. An increase in the misclassification error rate for the majority class was also shown, together with a reduction in this parameter for the minority class.

A reliable predictive accuracy, as well as a reduction in the minority misclassification error was obtained when comparing the chromosome 22 two-group classification (methylated *versus* the remainder). In contrast, a comparison of differentially-methylated classification versus the remainder did not improve the predictive accuracy, where the results showed a very unreliable value for this index and a high misclassification error, particularly with regards to the differentially-methylated class. Furthermore, the G-means and f-measure values improved significantly, ei.e, from $21.4 \pm 5.3\%$ to $47.8 \pm 2.5\%$, and $5.9 \pm 2.2\%$ to $33.2 \pm 3.4\%$, respectively. Although the total accuracy was high, their individual specificity (type II error) was higher than the sensitivity (type I error); this is consistent with other experiments previous section for this chap-

Table 5.30 Representative results of predictive accuracies and performance assessment for comparisons of the methylated, unmethylated and differentially-methylated classes of chromosome 21.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
Methylated vs all	1,4.1	49.0 ± 3.2	92.8 ± 4.8	88.8 ± 0.7	89.2 ± 1.0	90.75 ± 2.6	64.1 ± 3.6	Unweighted
Methylated vs all	1,4.1	51.4 ± 6.2	92.5 ± 5.9	89.2 ± 1.2	89.6 ± 1.5	90.8 ± 3.3	65.9 ± 6.1	weighted
unmet Vs all	2.3,1	99.0 ± 0.7	87.1 ± 0.9	96.7 ± 2.2	89.1 ± 1.0	91.8 ± 1.3	92.7 ± 0.7	Unweighted
unmet Vs all	2.3,1	99.0 ± 0.8	87.3 ± 1.2	96.7 ± 2.6	89.2 ± 1.3	91.9 ± 1.6	92.8 ± 0.8	weighted
D-met vs all	1,8.8	97.3 ± 4.7	100.0 ± 0.0	99.7 ± 0.52	99.7 ± 0.48	99.9 ± 0.3	98.6 ± 2.5	Unweighted
D-met vs all	1,8.8	93.3 ± 3.1	100.0 ± 0.0	99.3 ± 0.35	99.3 ± 0.32	99.6 ± 0.2	96.5 ± 1.7	weighted

Abbreviation: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

ter [27](MLOOCV obtained results).

Table 5.31 Representative results of three-class predictive accuracies and performance assessment for comparisons of methylated, unmethylated and differentially-methylated classes of chromosome 22.

sample	S-ratio	M-Prec	M-sen	M-Sp	M-acc	M-fmeas	M-Gmean	w-cost
Methylated vs all	1,3.9	53.2 ± 9.6	83.5 ± 2.6	88.5 ± 2.1	87.8 ± 2.0	85.9 ± 2.3	64.6 ± 8.2	Unweighted
Methylated vs all	1,3.9	89.8 ± 5.6	89.3 ± 1.6	97.3 ± 1.3	95.5 ± 1.2	93.2 ± 1.2	89.5 ± 3.2	Unweighted
unmet Vs all	2.4,1	99.6 ± 0.3	92.9 ± 0.4	98.8 ± 0.77	94.3 ± 0.5	95.8 ± 0.5	96.1 ± 0.3	Unweighted
unmet Vs all	2.4,1	83.7 ± 1.5	97.6 ± 0.7	70.5 ± 1.7	87.0 ± 0.9	83.0 ± 0.9	90.1 ± 0.7	weighted
D-met vs all	1,11.9	7.1 ± 3.4	5.1±1.7	91.9 ± 0.2	82.9 ± 1.9	21.4 ± 5.3	5.9 ± 2.2	Unweighted
D-met vs all	1,11.9	54.6 ± 7.2	23.9 ± 2.6	95.7 ± 0.6	82.9 ± 1.5	47.8 ± 2.7	33.2 ± 3.4	weighted

Abreviations: S-ratio = sample ratio, M-Prec (precision average), M-sen (sensitivity average), M-Sp (specificity average), M-acc (averaged accuracy), M-fmeeas (averaged F-measure), M-gmean (averaged G-mean) and W-cost (weighting)

5.4.6 Conclusions

In many practical applications, asymmetric datasets contain many observations for one class, and comparatively less for the other classes. This study presented two existing models that can be employed to tackle the bias problems introduced by such severely imbalanced datasets. It was also shown that small samples severely affect classifier performance when compared to medium-imbalanced datasets. It was shown experimentally that the combination of prior probability and cost-weighting is more

effective than a single weighting method, and it was therefore possible to differentiate methylated class samples from both unmethylated and differentially-methylated ones, with an excellent predictive performance. This approach revealed a high predictive performance for the distinction of methylated and differentially-methylated classes (the minority classes), without losing predictive performance of the majority class (the unmethylated class). Furthermore, the performance of algorithms based on these combinatorial methods, is assessed with six metrics, which represent a wide range of predictive measurements, particularly when compared to previous studies that only report overall accuracies and correlation coefficients [48; 64]. Ten cycles were set within five-fold cross-validation in order to assess the stability of the model (in which the weighting cost of each repeated analysis is assessed by measuring the misclassification error rate), and the average predictive accuracy with its associated standard deviation value was calculated. This method can also be applied to imbalanced data problems in other domains, and can also be extended to multi-class prediction models.

Chapter 6

Analysis of gender differences in DNA methylation

6.1 An analysis of the relationship between DNA methylation and gender

The current chapter presents differences in CpG methylation between males and females. Author carried out a comprehensive supervised data analysis of 1506 CpG methylation positions (as detailed in chapter 3, section 3.1.1). This analysis was based on 963 samples, of which 328 were healthy samples (168 taken from males and 160 from females), and 635 were cancerous samples (404 from males and 231 from females). Author employed the sequential forward method of feature selection, which was combined with Quadratic Discriminant Analysis (QDA), in order to identify the variation in CpG methylation positions between males and females. This was followed by an assessment of the predictive accuracy and performance of the selected individual and feature sub-sets.

Author identified 6 features from healthy samples and 18 from cancerous ones which are significantly affected by gender differences. In addition, we extended the investigation by examining differences in tissue-specific methylation according to gender; this enabled us to characterise 25 CpG positions from leukocytes, and 10 from healthy colon samples.

6.2 Introduction

Gene expressions are regulated by DNA methylation without changing the DNA sequence itself [20]. DNA methylations are influenced by environmental and lifestyle factors, including smoking, excessive alcohol intake, diet, age, gender, and stress. All of these factors accelerate epigenetic deregulation, whereas taking part in sports,

having a healthy diet and lifestyle, and maintaining a good level of physical fitness all appear to delay this process [20; 21; 60]. These methylation changes are heritable through the process of cell replication, and can affect DNA stability and normal cellular function; for this reason, author and others have suggested that DNA methylation may play an important role in human diseases and common risk factors [204]. Methylations in CpG islands have been observed in many types of tumours [205]. Although the CpG positions in genomic regions which are responsible for the changes (methylation) have not been completely identified, perturbing gender-specific differences in methylation have led the author to investigate the impact of gender on DNA methylation, particularly in CpG islands, which are mostly found to be free of methylation in normal physiological cell development. DNA methylation disruption is linked to stress [24] and complex age-related diseases such as Alzheimers disease and cancer [20; 43]. It is essential to have an insight into the association between gender and DNA methylation both in healthy and diseased samples in order to learn more about the pathophysiology of gender-related health outcomes. Researcher and others have identified features of CpG islands, such as DNA sequence patterns, DNA structure, and DNA physiochemical properties [5; 48; 64]. However, these researchers did not address gender-specific features of DNA methylation. The essence of DNA methylation is a predominantly reversible process. CpG methylation is a continuous process of ageing [89; 90] in which identical twins have shown differences in DNA methylation with increasing age [91]. It is impossible to say with certainty that gene expressions are regulated by either genetic or environmental factors, since they can also be inherited through cell division [92]. For the same reason, it is difficult to establish the effect of gender on this process. These problems make it very challenging to design, predict and model whether these epigenetic changes can be determined in both healthy and diseased data, in order to distinguish gender differences across a broad age range, and to link these DNA methylation differences to healthy and cancerous individuals. While epigenetic changes associated with age have been studied to some degree [206; 207; 208], there is a shortage of reported studies on the relationship between gender and DNA methylation. To determine how nucleotide methylation is associated with gender and to understand whether DNA methylation fingerprints can be correlated with it, author compared genome-wide DNA methylation fingerprints from 963 samples across a wide age range. Author designed feature selection algorithms with filtering methods in order to determine which features are important, and also to establish their predictive accuracy of gender-based DNA methylation for both healthy and cancerous samples. The preliminary results show the existence of gender-related DNA methylation differences; these findings may be followed up by larger and more sophisticated studies in the future.

6.2.1 Material and Methods

6.2.1.1 Data extraction and experimental proceedings

Author initially extracted data from 1628 human samples [179]; these samples consisted of various types of both healthy and diseased tissues. The experimental datapoints contain a fluorescent signal from methylated and unmethylated alleles. DNA methylation and dinucleotide DNA spots were extracted from an Illumina hybridisation spot array of CYS3 and CYS5 methylated and unmethylated respectively, and also a mixture of both hybridisations have been designed for extraction of the differentially-methylated forms [63]. A ratio of each spot array represents the methylation value of loci position (CpG) of individual samples. These spot arrays (features) were based on 1505 CpG sites from 807 genes. These genes include oncogenes, tumour suppressor genes, differentially-methylated or expressed genes, imprinted genes, signal pathway genes, DNA repair genes, and cell-cycle control genes, in addition to those responsible for metastasis, apoptosis and cell differentiation. A generalised age distribution was added to this, bringing the total number of features to 1506 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28094).

Author excluded samples that did not have information on the age or gender of the participants, and author also excluded any samples which did not provide information about the source of the biological sample. 936 samples remained; this was made up of 328 samples of healthy and 635 samples of diseased tissue. Details of the final studied samples are given in Table 6.1. In addition, author extracted leukocytes (157) and colon samples (96) from healthy samples.

6.2.1.2 Data analysis

The studied dataset was divided into two categories: male and female. Corresponding data were then further separated into the two classes of healthy and cancerous tissue. Furthermore, data were split into training and test set data, whereby a quarter of the data was used for testing, and the remainder was used for training purposes. In addition, t-statistics were used for P-value estimation in order to compare the separability of individual features, and also how well the features are separated for the two classes. As with the t-test, the filtering method of feature selection is based on the feature separateness criterion and does not consider interactions between features [111; 167; 209], which may contain no biologically important features. Consequently, author used the wrapper method in order to select the feature[4; 113; 167] or feature sub-set that is most important for the purpose of distinguishing gender differences. This method searches sequentially through the training set for misclassification errors of the quadratic discriminant analysis (QDA) in order to select the optimal feature sub-sets. In addition, we applied ten-fold cross-validation integrated with ten times repeated cycles. Each repeated analysis stops when it has found the

first minimum of the cross-validation misclassification error in order to test stability of the classifier (QDA). Then the performances of the selected sub-sets are evaluated by plotting the misclassification error (MCE) of the test set on each repeated cycle during the feature selection process. Furthermore, author used Matthews Correlation Coefficient (MCC) to calculate predictive accuracies and assess predictive power. The results were averaged and their standard deviation was estimated.

Table 6.1 Summary of studied samples

Sample status	Male	Female	Total sample
Health sample	168	160	328
Leukocyte healthy	70	87	157
Colon healthy	56	40	96
Cancer sample	404	231	635

6.2.2 Results

The results are based on the analysis of 1506 features from 963 samples; these samples consisted of 328 healthy samples (control) and 635 cancerous samples. Author applied two statistical filtering methods: the t-test and wrapper methods. For the purpose of estimating P values, author used the absolute t-test. This method showed individual separateness of features according to their P-values without considering interaction between the features. In order to explore the interaction between features or feature sub-sets, we applied the forward sequential (wrapper) feature selection method and computed a misclassification error rate. As the misclassification error rate of the test-set is reached at a minimum, the terminated selection will gives the best feature sub-set. Furthermore, the performance of the selected feature sub-set is evaluated by computing the misclassification error rate of the test set. The best selected features and their biological functions are presented in Tables 6.2, 6.5 and 6.8 for healthy samples, and Table 6.11 for cancerous samples. The optimal selected feature sub-sets with their predictive accuracies and MCC were calculated for predictive performance assessments; this will be discussed in more detail later in this section.

6.2.2.1 Results arising from the healthy samples

The six features most associated with gender were selected from 1506 features. These included ELK1-P6-R, ELK1-E156-F, DNASE1L1-E178-R, Ripk4-E166-F, ROR2-E112-F and Fgf12-E61-R, which were the features that displayed the most significant differences between males and females. In three of the CpG positions (ELK1-P6-R, ELK1-E156-F and DNASE1L1-E178-R), there were significant methylation differences between the two genders, with P-values of $P \leq 1.1E-68, P \leq 1.5E-50$ and

$P \leq 2.2E-46$ respectively. In addition, the biological functions of the genes associated with these positions are annotated in Table 2 and detailed in the discussion. The

Table 6.2 Selected features from healthy samples

F-ID	P-values	CpG Associated genes	function
373–ELK1–P6-R	$1.09 \times E^{-68}$	ELK1: ETS domain-containing protein Elk-1	Bind AT-rich sequence in order to activate transcription complex.
369–ELK1–E156-F	$1.50 \times E^{-50}$	ELK1: ETS domain-containing protein Elk-1	Bind AT-rich sequence in order to activate transcription complex
331–DNASE1L1–E178-R	$2.24 \times E^{-46}$	DNASE1L1: Deoxyribonuclease-1-like 1	DNA and cell repair.
1190–RIPK4–E166-F	0.32	RIPK4: Receptor-interacting serine/threonine-protein kinase 4	Involved in epithelial development and cell growth.
1193–ROR2–E112-F	0.71	ROR2: Tyrosine-protein kinase transmembrane receptor ROR2	Cell growth and bone morrow development.
460–FGF12–E61-R	0.85	FGF12: Fibroblast growth factor 12	Involved in nervous system development and cell repair.

ten times repeated analysis performed within ten-fold cross-validation showed that ELK1-P6-R was the best individual feature, as it was repeated in the first position of each fold of the selected sub-sets, followed by ELK1-E156-F . DNASE1L1-E178-R is the third most important feature, and may determine the distribution of gender differences as presented in Table 4, which details the order of importance selected. In addition, ELK1-P6-R, DNASE1L1-E178-R and ROR2-E112 were the best selected feature sub-sets, since they had the highest predictive accuracy and the lowest misclassification error rate of $94.1 \pm 1.2\%$, and a MCC of 0.88 ± 0.02 (Table 6.3). The misclassification error rates of the best one hundred selected features are plotted in Figure 1, which shows that after the 29^{th} feature, the misclassification error rate is minimal. This rate remains stable until feature 90, after which point the curve rises, i.e. over-fitting took place from this point on. This indicates that our selection method is reliable, since the selected features are shown to be stable, with low error rates in each iteration of the data. However, ELK1-P6-R found X-chromosomes which may originated from the female copy as it compensates for X-linked dosage, which reported one of five house-keeping genes methylated in the female [210].

As the selection of subsets is able to predict gender-associated feature subsets with high predictive performance, it is also equally important to investigate whether individual features (CpG loci) make an equivalent contribution to gender-related differences. Furthermore, individual features were measured for their predictive accuracies and their predictive performance was also assessed using MCC, incorporating a QDA classifier. The results of this analysis are summarised in Table 6.4. Three out of the six features, ELK1-P6-R, ELK1-E156-F and DNASE1L1-E178-R, show high predictive accuracies and performances; this was subsequently validated by the t-test. As

Table 6.3 The best selected feature sub-sets with 10-fold cross-validation for healthy samples

Feature sub-sets	sen-std	sp-std	Acc-std	MCC-std
ELK1, ELK2	93.9 ± 2.8	91.2 ± 1.7	92.5 ± 0.8	85.2 ± 1.8
ELK1, ELK2, Ripk4, Fgf12	93.2 ± 2.6	92.4 ± 1.0	92.8 ± 1.5	85.6 ± 3.1
ELK1, DNASE, ROR2	94.5 ± 1.1	93.7 ± 1.7	94.1 ± 1.2	88.2 ± 2.3
ELK1, ELK2	93.3 ± 1.8	91.3 ± 1.9	92.3 ± 1.5	84.6 ± 3.0
ELK1, ELK2, Ripk4	91.4 ± 2.1	92.3 ± 2.3	91.9 ± 0.7	83.8 ± 1.5
ELK1, Ripk4	92.4 ± 2.2	90.6 ± 2.0	91.5 ± 1.8	82.9 ± 3.6
ELK1	94.7 ± 3.0	91.2 ± 1.6	92.9 ± 2.1	85.9 ± 4.3
ELK1, ELK2, Ripk4	90.3 ± 2.8	92.6 ± 1.7	91.5 ± 0.7	83.0 ± 1.4
ELK1, ELK2	92.7 ± 1.0	94.0 ± 2.1	93.4 ± 1.0	86.8 ± 2.1
ELK1, ELK2	94.6 ± 1.4	91.6 ± 1.6	93.0 ± 0.7	86.1 ± 1.4

Abbreviations: Sensitivity (Sen) represent male and specificity (SP) represents female, standard deviation (std) predictive accuracy (Acc) and Matthews Correlation Coefficient (MCC).

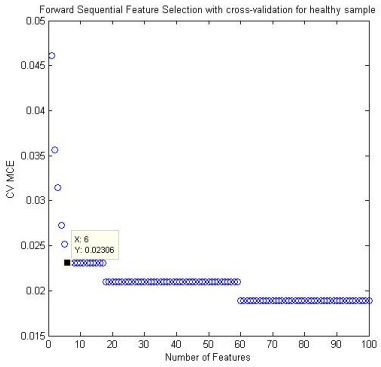


Figure 6.1 10-fold misclassification error and the sequential selected first 100 features from healthy samples.

shown in Table 2, the wrapper method revealed that these CpG loci position have a significant discriminatory power. The accuracies of the loci positions are $92.2 \pm 2.0\%$, $89.2 \pm 1.7\%$ and $88.7 \pm 1.8\%$ respectively. The other three loci positions (Ripk4-E166-F, ROR2-E112-F and Fgf12-E61-R) resulted in lower predictive accuracy scores when individual features were analysed, despite showing evidence of better performance when combined (feature sub-set selection). Comparing the predictive performance of sensitivity or specificity led the author to discover that Ripk4-E166-F and ROR2-E112-F have a higher specificity of $89.1 \pm 2.1\%$ and $83.9 \pm 6.9\%$ respectively for females, while FGF12 has a sensitivity of only $70 \pm 5.5\%$ for males (Table 6.4). Despite the fact that individual features gave lower predictive accuracies, three out of the six selected features demonstrated a better performance and also high correlation coefficients. This suggests that feature interaction increases the predictive accuracy of the sub-sets rather than that of individual features. In addition, to further narrow down whether the same features contribute to the tissues, we analysed data relating to leukocytes and colon samples which may provide evidence of tissue-specific methylation. As a result, we were able to identify ten feature sub-sets (CpG loci positions) in which significant differences between the two genders were evident.

Table 6.4 Individual CpG predictive performance of selected features over 10-fold cross-validation.

individual features	sen-std	sp-std	Acc-std	MCC-std
ELK1–P6-R	94.2 ± 2.1	90.3 ± 2.7	92.2 ± 2.0	84.6 ± 4.1
ELK1–E156-F	86.8 ± 2.5	91.4 ± 1.6	89.2 ± 1.7	78.4 ± 3.4
DNASE1–E178-R	84.5 ± 2.8	92.7 ± 1.8	88.7 ± 1.8	77.7 ± 3.5
Ripk4–E166-F	19.9 ± 3.7	89.1 ± 2.1	55.3 ± 1.2	12.4 ± 3.1
ROR2–E112-F	28.5 ± 3.0	83.9 ± 6.9	56.7 ± 4.2	15.4 ± 11.9
Fgf12–E61-R	70.9 ± 5.5	23.9 ± 5.5	46.8 ± 1.4	-6.1 ± 3.6

Abbreviations: Sensitivity (Sen) represent male and specificity (SP) represents female, standard deviation (std), predictive accuracy (Acc) and Matthews Correlation Coefficient (MCC).

6.2.3 Data analysis of normal leukocytes

Using the methods explained in the previous section, author investigated whether or not the selected features are also tissue-specific.

Ten optimal feature sub-sets, containing a total of 25 individual features, were selected from 1506 features of normal leukocyte samples. From this analysis, two out of the 25 features were frequently selected in the first position by ten-times repeated analysis incorporated with ten-fold cross-validation. ELK1-P6-R was selected six times, followed by BIRC4-P122-R, which was selected four times; these are both considered to be the most important features. The CpG loci of the selected optimal features are plotted in Figure 6.2, which shows the number of selected features and their individual frequencies during the selection process.

Furthermore, author compared the predictive accuracies and correlation coefficients of the selected sub-sets. The sub-set of BIRC4-P122-R, HDAC6-E102-F and

IL10-P348-F has the highest level of predictive accuracy, and the highest MCC; these were $85.2 \pm 1.2\%$ and 0.70 ± 0.03 respectively. This was followed by the sub-set of ELK1-P6-R, DLG3-P62-R, ELK1-E53-F, P2RX7-E323-R and TGFBI-P31-R, which showed a good level of predictive accuracy at $84.6 \pm 3.4\%$ and an MCC of 0.69 ± 0.07. From this analysis, author identified nine out of 25 CpG loci in which there was evidence of significant DNA methylation differences between males and females. These results were validated by a two-sided t-test ($P<0.05$), the results of which can be viewed in Table 6.5. BIRC4-P122-R was found to be the most significant contributor to gender differences, with a P-value of $8.8 \times E^{-}23$. The selected sub-sets have been grouped into two clusters, with each group beginning with either ELK1-P6-R or BIRC4-P122-R. Hence, these individual features belong to those two clusters, with the exception of two features, i.e. P2RX7-E323-R and TNFRSF10D-P70-F. The two loci have a common function, namely transmembrane signal activation, which play a key role in normal cellular response [211; 212]. Furthermore, the methylation distribution plot displayed revealed that there are significant methylation differences between males and females for the loci positions BIRC4-P122-R and HDAC6-E102-F, whereas the locus position of IL10-P348-F indicates no significant ($P \leq 7.5E-02$) methylation differences between the two genders (Figure 6.3). However, feature sub-set selection showed a high predictive accuracy; this is presented in more detail in Table 6.6.

Table 6.5 Selected features of normal leukocytes with their biological functions.

Features-ID	P-values	CpG Associated genes	function
ELK1-P6-R	$9.1 \times E^{-21}$	ELK1(ETS domain-containing protein Elk-1)	Binds AT-rich sequence,and activates transcription complex.
ZNF264-P397-F	$5.3 \times E^{-02}$	ZNF264(Zinc finger protein 264)	Plays a role in transcriptional activation, DNA and zinc ion binding.
SEMA3B-E96-F	6.6^{-02}	SEMA3B(SEMA3B protein)	Involved in growth and neural development.
TNFRSF10A-P171-F	$5.6 \times E^{-01}$	TNFRSF10A (Tumor necrosis factor receptor superfamily member 10A)	Involved in cell-death signal and induces cell apoptosis.
PCGF4-P760-R	$7.8 \times E^{-01}$	PCGF4(Polycomb copmlex protein BMI-1)	Essential for Signal transduction activity and sequence-specific DNA binding.
BIRC4-P122-R	$8.8 \times E^{-23}$	BIRC4: E3 ubiquitin-protein ligase XIAP	Cell-death inhibitor/upregulates various types of tumour cells.
TNFRSF10D-P70-F	$2.2 \times E^{-01}$	TNFRSF10D (Tumour necrosis factor receptor superfamily member 10D)	Essential for the transmembrane signalling receptor activity and TRAIL binding.
BCR-P422-F	$7.0 \times E^{-01}$	BCR (Breakpoint cluster region protein)	Required for signal transduction and GTPase activity regulation.
EPHA1-P119-R	$7.4 \times E^{-02}$	EPHA1(Ephrin type-A receptor 1)	Regulates transmembrane-ephrin receptor activity and protein kinase binding.
ABL2-P459-R	$1.1 \times E^{-01}$	ABL2 (Abelson tyrosine-protein kinase 2)	Regulates the protein tyrosine kinase activity.
GSTM2-P109-R	$4.6 \times E^{-01}$	GSTM2(Glutathione S-transferase Mu 2)	Required for the metabolism of glutathione and Phase II conjugation.
HPN-P823-F	$7.2 \times E^{-02}$	HPN (Serine protease hepsin)	Essential for cell growth and maintenance of cell morphology.
DLG3-P62-R	$6.2 \times E^{-19}$	DLG3(Disks large homolog 3)	Essential for neurotransmitter receptor binding and downstream transmission in the postsynaptic cell and axon guidance.
ELK1-E53-F	$3.0 \times E^{-17}$	ELK1 (ETS domain-containing protein Elk-1)	Binds AT-rich sequence in order to activate transcription complex.
P2RX7-E323-R	$4.8 \times E^{-01}$	P2RX7 (P2X purinoceptor 7)	Required for both fast synaptic transmission and the ATP-mediated lysis of antigen-presenting cells.
TGFBI-P31-R	$6.4 \times E^{-01}$	TGFBI (Transforming growth factor-beta-induced protein igh3)	Essential for cell-collagen interactions and cartilage endochondral bone formation.
HDAC6-E102-F	$6.8 \times E^{-21}$	HDAC6(Histone deacetylase 6)	Essential for transcriptional regulation, cell-cycle progression, and developmental events.
IL10-P348-F	$7.5 \times E^{-02}$	IL10(Interleukin-10)	Inhibits the synthesis of a number of cytokines, activated macrophages and helper T-cells.
ETS1-P559-R	$2.9 \times E^{-03}$	ETS1(Protein C-ets-1)	Regulates transcription factor binding and sequence-specific DNA binding activity.
PLS3-E70-F	$9.0 \times E^{-03}$	PLS3(Plastin-3)	Essential for actin and calcium ion binding.
MYBL2-P354-F	$3.2 \times E^{-02}$	MYBL2(Myb-related protein B)	Essential for cell survival, differentiation and activation of transcription factor.
IGF2AS-P203-F	$2.9 \times E^{-01}$	IGF2AS (Putative insulin-like growth factor 2 antisense gene protein)	This gene expresses a paternally imprinted antisense and is overexpressed in Wilms' tumour.
PYCARD-E87-F	8.7^{-01}	PYCARD (Apoptosis-associated speck-like protein containing a CARD)	Involved in the apoptotic process and protein homodimerization activity.
EPHB3-P569-R	$1.3 \times E^{-02}$	EPHB3(Ephrin type-B receptor 3)	Controls ephrin-receptor activity and axon guidance receptor activity.
KCNQ1-E349-R	$1.9 \times E^{-01}$	KCNQ1(Potassium voltage-gated channel subfamily KQT member 1)	plays a role in tissue-specificity, with preferential expression from the maternal allele in some tissues, and biallelic expression in others.

Moreover, the analysis of individual features revealed that BIRC4-P122-R has the highest predictive accuracy and MCC of $88.2 \pm 2.0\%$ and 0.77 ± 0.04 respectively, compared to the other 24 individual features. This is followed by HDAC6-E102 of accuracy ($86.9 \pm 2.4\%$) and MCC (0.74 ± 0.05.) Although ELK1-P6-R was the most frequently selected feature, it had the third highest predictive accuracy and MCC ($86.3 \pm 2.1\%$ and 0.72 ± 0.04, respectively); this suggests that ELK1-P6-R interacts with features other than BIRC4-P122-R and HDAC6-E102-F. Overall, individual features had a lower predictive accuracy than the feature sub-sets. Four out of 25 individual features have shown high predictive accuracies, whereas the others have lower predictive ones than those achieved via employment of combined features (Table 6.6). In addition, six out of the 25 selected features gave higher sensitivity values (correctly predicted for male) and seven out of these 25 individual features have also shown good specificity (correctly predicted for female), with predictive accuracy values between $71.5 \pm 1.8\%$ and $89.0 \pm 5.0\%$ for males, and between $72.4 \pm 5.4\%$ and $91.5 \pm 3.5\%$ for females. Furthermore, author found that these individual features have a slightly higher methylation range for females than for males.

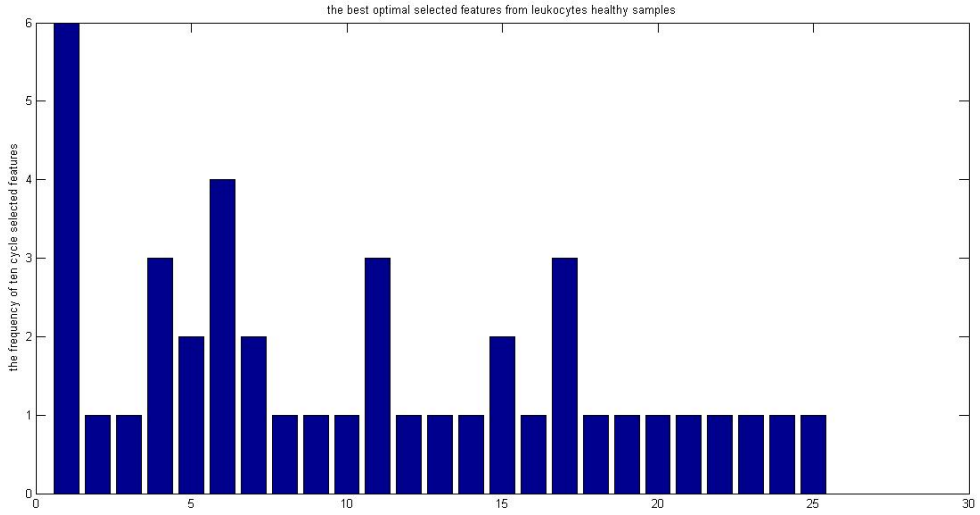


Figure 6.2 Number of selected features in ten-times repeated analysis.

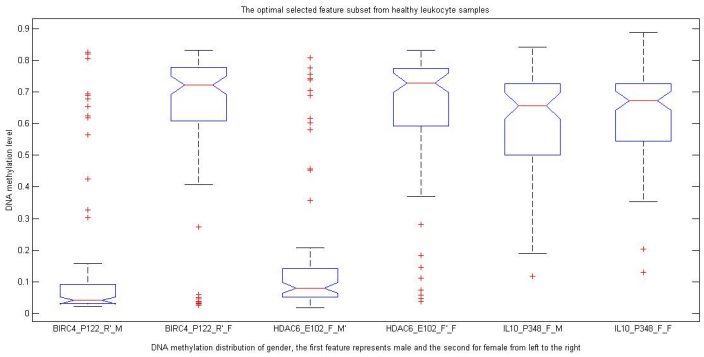


Figure 6.3 DNA methylation distribution of gender, the first feature represents male and the second for female from the left to the right

Additionally, ELK1-P6-R and BIRC4-P122-R were assigned to different sub-sets in all repeated analyses; one possible explanation for this is that they perform different biological functions. Although they both show significant DNA methylation differences between males and females, their associated genes may have antagonistic functions. ELK1 activates transcription factors for binding, specifically at AT-rich DNA binding sites [213], whereas BIRC4 is associated with the genetic suppression of the transcription factor and further activates the process of cell death [204]. Moreover, in six out of the 25 selected individual features, there were significant differences between males and females, P<0.05. However, for the majority of features, the differences were insignificant (P0.05). By increasing the p-value to $P \leq 0.1$, eleven of the selected features, and 14 of which were selected with wrapper methods, show significant differences in DNA methylation between males and females. The significant features are mostly associated with DNA binding and cell-cycle control, with more than half being associated with DNA binding. The selected sub-sets are presented in Table 6.6. As ELK1 regulates transcriptional factor and is expressed in almost all tissues, author identified that the locus positions of ELK1-P6-R are highly methylated in females but hypomethylated in males. Furthermore, ELK1 has three CpG positions, only two of which were selected; these were more highly methylated in females than in males. Two loci positions for BIRC4 were selected, in which the female samples were hypermethylated whilst the male samples were hypomethylated. In addition, the locus region BIRC4-P122-R was shown to be the best feature for discriminating between males and females. However, the ELK1-P6-R region was the most frequently selected feature of the ten-times repeated analysis incorporating ten-fold cross-validation. Furthermore, in the HDAC6-E102-F position, there was also a significant difference between males and females, with $P \leq 6.8E-21$. As demonstrated in Table 6.5, the selected feature sub-sets revealed higher predictive accuracies than the individual features; this suggests that these selected sub-sets interact. These results

are consistent with further analysis of individual features using the same method, which shows that BIRC4-P122-R has the highest predictive accuracy and MCC values when compared with the other selected features in Table 6.7. In both leukocyte and colon samples, both loci position were present in both features, and there were significant methylation differences between the two genders.

Table 6.7 Predictive accuracies of individual features of leukocytes.

individual features	sen-std	sp-std	Acc-std	MCC-std	P-values
ELK1-P6-R	80.2 ± 3.4	91.3 ± 2.0	86.3 ± 2.1	72.4 ± 4.4	$9.1 \times E^{-21}$
ZNF264-P397-F	39.6 ± 5.3	57.1 ± 8.9	49.3 ± 6.2	-3.2 ± 12.0	0.05
SEMA3B-E96-F	71.5 ± 1.8	38.1 ± 7.4	52.8 ± 4.3	10.0 ± 7.9	0.07
TNFRSF10A-P171-F	16.1 ± 4.4	91.5 ± 3.5	57.6 ± 1.6	11.9 ± 7.0	0.6
PCGF4-P760-R	31.0 ± 6.8	76.8 ± 8.6	56.4 ± 2.6	9.0 ± 5.0	0.8
BIRC4-P122-R	85.8 ± 5.6	90.1 ± 5.6	88.2 ± 2.0	76.5 ± 3.8	$8.8 \times E^{-23}$
TNFRSF10D-P70-F	24.2 ± 9.2	86.5 ± 2.8	58.4 ± 5.0	12.5 ± 15.0	0.22
BCR-P422-F	76.1 ± 8.2	14.5 ± 5.3	41.6 ± 1.8	-11.9 ± 5.3	0.70
EPHA1-P119-R	80.1 ± 3.3	38.8 ± 4.5	57.0 ± 2.5	20.4 ± 4.9	0.07
ABL2-P459-R	89.0 ± 5.0	23.5 ± 4.8	52.6 ± 1.9	16.8 ± 5.2	0.1
GSTM2-P109-R	27.7 ± 4.1	82.3 ± 4.3	57.7 ± 3.7	12.0 ± 9.3	0.5
HPN-P823-F	50.1 ± 12.8	61.8 ± 4.7	56.5 ± 6.9	11.9 ± 14.9	0.07
DLG3-P62-R	84.5 ± 2.4	81.2 ± 6.6	82.6 ± 3.6	65.5 ± 6.4	$6.2 \times E^{-19}$
ELK1-E53-F	86.3 ± 3.8	84.5 ± 3.8	85.3 ± 3.2	70.6 ± 6.3	$3.0 \times E^{-17}$
P2RX7-E323-R	53.9 ± 11.2	50.1 ± 3.4	51.9 ± 5.7	4.0 ± 12.2	0.5
TGFBI-P31-R	15.7 ± 4.9	86.7 ± 4.8	55.0 ± 2.0	3.5 ± 5.1	0.6
HDAC6-E102-F	88.4 ± 3.6	85.7 ± 3.4	86.9 ± 2.4	73.8 ± 4.7	$6.8 \times E^{-21}$
IL10-P348-F	31.6 ± 2.9	74.8 ± 4.3	55.6 ± 1.7	7.2 ± 3.3	0.07
ETS1-P559-R	57.1 ± 4.5	72.4 ± 5.4	65.5 ± 4.7	29.9 ± 9.3	0.003
PLS3-E70-F	78.1 ± 3.9	40.0 ± 10.6	57.0 ± 4.7	19.3 ± 7.6	0.009
MYBL2-P354-F	89.9 ± 7.1	30.5 ± 6.5	57.1 ± 3.6	25.5 ± 9.1	0.03
IGF2AS-P203-F	78.6 ± 7.9	25.9 ± 7.3	49.2 ± 2.8	6.0 ± 8.2	0.3
PYCARD-E87-F	54.4 ± 3.2	43.3 ± 5.6	48.1 ± 2.1	-2.4 ± 3.3	0.9
EPHB3-P569-R	69.8 ± 4.7	25.8 ± 4.9	45.2 ± 2.6	-5.0 ± 5.4	0.01
KCNQ1-E349-R	65.6 ± 9.1	41.6 ± 2.3	52.3 ± 3.6	7.5 ± 8.3	0.2

6.2.3.1 Normal colon tissue samples

Author investigated whether selected features (CpG loci) are also tissue specific by the examination of healthy colon samples. Author was able to identify CpG loci methylation differences between male and female samples. The first ten selected features showed significant gender differences with low p-values of $P \leq E - 20$, and further analysis of individual features revealed high predictive accuracies of between 96.7% and 100%, and MCCs of between 0.93 and 1.00 (Table 6.9). ELK1-P6-R is the most important feature since it repeated in the first position for all ten-fold cross-validation with ten times repeated analysis, followed by STK23-E182-R and VBP1-E127-F. The gene associated with STK23 plays an important role in the activation of the protein serine/threonine kinase pathway and in ATP binding, while VBP1 is essential for post-translational protein modification and the binding of unfolded proteins [214]. Furthermore, observing the results for the best selected sub-set confirms the existence

of significant methylation differences between the two genders; with the DNA of female samples tending to be much more highly methylated than was the case for males (Figure 6.4). STK23-E182-R shows differential methylation for males compared with an observation hypermethylation in females (Figure 6.4). STK23-E182-R shows differential methylation for males consistent with the threshold of unmethylated CpG loci as reported by experiments [63].

Moreover, the selected features are annotated in the protein and gene-card database, which indicates that seven out of ten gene-protein interactions have a DNA binding function. This suggests that gender differences in DNA methylation can be linked to transcription regulation, and also indicates that gender-related differences in DNA methylation are correlated with the manner in which genes are regulated parallel with genomic imprinting. This confirms that genes that are essential for the particular function for males are on, whereas the same gene in females is off; there is also an influence from the environment, such as the impact of their jobs, behaviour, lifestyle and lifetime processes, as expressed in epigenetics [89].

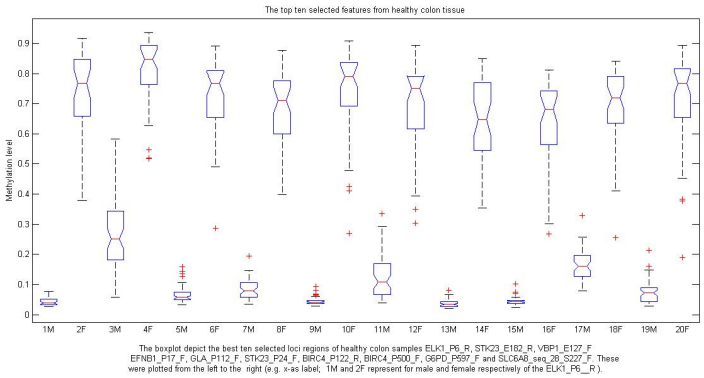


Figure 6.4 The box-plots depict the best ten selected loci regions of healthy colon samples ELK1-P6-R, STK23-E182-R, VBP1-E127-F EFNB1-P17-F, GLA-P112-F, STK23-P24-F, BIRC4-P122-R, BIRC4-P500-F, G6PD-P597-F and SLC6A8-seq-28-S227-F. These were plotted from the left to the right (e.g. x-as label; 1M and 2F represent for male and female respectively of the ELK1-P6-R).

6.2.4 Data analysis results acquired on cancer samples

To obtain more information about the influence of gender on the methylation of CpG loci, author used the same methods to investigate gender-related differences in DNA methylation.

Using forward sequential feature selection integrated with ten-fold cross-validation, author identified the 18 features most associated with gender from a total of 1506 features. Author also calculated the predictive accuracy and MCC of feature sub-sets and individual features, and estimated the P-values using the two-tailed t-test. The selected features comprise ten sub-sets obtained from ten-time repeated analysis.

GLA-P112-F is the most frequent selected feature, and is considered to be the most important feature since it was selected in the first position of the all ten repeated cycles. The best selected sub-set is GLA-P112-F, BIRC4-P122-R, HDAC6-E102-F, GLA-E98-R, NQO1-E74-R, and MAP2K6-E297-F; this sub-set shows the highest predictive accuracy and MCC values of $97.5 \pm 1.5\%$ and 0.95 ± 0.03 respectively. This sub-set yields the lowest misclassification error rate, which is displayed in Figure 6.5. Overall, the performance of sub-sets was effective; details of this are listed in Table 6.10.

Table 6.10 The predictive performance of statistically selected sub-set features from cancer samples.

Feature subset	sen-std	sp-std	Acc-std	MCC-std
GLA-P112-F, DKC1-P276-F, HDAC6-E102-F, G6PD-P196-F, NASE1L1-P108-F, NQO1-E74-R, HIF1A-P488-F	96.6 ± 2.4	97.4 ± 1.0	97.1 ± 1.2	93.7 ± 2.5
GLA-P112-F, BIRC4-P122-R, GLA-E98-R, NQO1-E74-R, MAP2K6-E297-F	94.0 ± 2.0	98.1 ± 0.9	96.6 ± 1.0	92.6 ± 2.1
GLA-P112-F, NQO1-E74-R, SYBL1-P349-F, HIF1A-P488-F,DNMT1-P100-R	95.0 ± 1.1	97.4 ± 0.6	96.5 ± 0.6	92.5 ± 1.4
GLA-P112-F, NQO1-E74-R,HIF1A-P488-F, DNMT1-P100-R	95.9 ± 1.8	97.4 ± 1.0	96.8 ± 1.2	93.2 ± 2.6
GLA-P112-F, HDAC6-E102-F,IRAK1-P312-F, GLA-E98-R,NQO1-E74-R,MAP2K6-E297-F	95.4 ± 2.2	97.8 ± 0.8	97.0 ± 1.2	93.4 ± 2.7
GLA-P112-F, BIRC4-P122-R,HDAC6-E102-F, GLA-E98-R, NQO1-E74-R,MAP2K6-E297-F	95.9 ± 2.4	98.5 ± 1.1	97.5 ± 1.5	94.7 ± 3.3
GLA-P112-F,ARAF-E38-F,HDAC6-E102-F, G6PD-P196-F,NQO1-E74-R, HOXA11-P92-R	93.4 ± 1.3	97.7 ± 0.4	96.1 ± 0.4	91.6 ± 1.0
GLA-P112-F, HDAC6-E102-F,IRAK1-P312-F,NQO1-E74-R,MAP2K6-E297-F, HIF1A-P488-F, SYBL1-E23-R	94.3 ± 1.1	98.3 ± 0.8	96.9 ± 0.8	93.2 ± 1.7
GLA-P112-F, DKC1-E101-F,G6PD-P196-F, GLA-E98-R, NQO1-E74-R, MAP2K6-E297-F, CCNC-P132-R	95.5 ± 1.8	97.4 ± 0.4	96.7 ± 0.7	92.9 ± 1.6
GLA-P112-F, BIRC4-P122-R, GLA-E98-R,NQO1-E74-R, MAP2K6-E297-F, HIF1A-P488-F	93.8 ± 1.5	97.4 ± 0.2	96.1 ± 0.6	91.6 ± 1.3

Sensitivity (Sen) represent male and specificity (SP) represents female, standard deviation (std), predictive accuracy (Acc) and Matthews Correlation Coefficient (MCC).

Table 6.11 Summary of selected features from cancerous samples

547-GLA-P122-F	$1.70 \times E^{-99}$	GLA (Alpha-galactosidase A)	A variety of mutations in this gene affect the synthesis, processing, and stability of this enzyme, which causes Fabry disease.
109-BIRC4-P122-R	$1.01 \times E^{-87}$	BIRC4 (E3 ubiquitin-protein ligase XIAP)	Cell death inhibitor/up regulate varieties tumour cell types.
998-NQO1-E74-R	0.0037	NQO1(NAD(P)H dehydrogenase [quinone] 1)	Regulates enzymatic activity gene disruption risk for different forms of cancer and is linked to Alzheimer's disease (AD).
601-HDAC6-E102-F	$2.24 \times E^{-77}$	HDAC6(Histone deacetylase 6)	HDAC6 plays a critical role in transcriptional regulation, cell-cycle progression, and developmental events.
316-DKC1-P276-F	$9.23 \times E^{-78}$	DKC1(H/ACA ribonucleoprotein copmlex subunit 4)	RNA binding protein which regulates synthase activity and telomerase activity.
69-ARAF-E38-F	$1.69 \times E^{-81}$	ARAF(Serine/threonine-protein kinase A-Raf)	Involved in cell development and growth.
315-DKC1-E101-F	$6.07 \times E^{-65}$	DKC1(H/ACA ribonucleoprotein complex subunit 4)	RNA structure stabilisation and maintenance of ribosome biogenesis and telomere.
546-GLA-E98-R	$2.86 \times E^{-61}$	GLA (Alpha-galactosidase A)	Developmental protein
748-IRAK1-P312-F	$1.93 \times E^{-63}$	IRAK1(Interleukin-1 receptor-associated kinase 1)	Activates innate immune response against foreign pathogens and cell growth.
1314-SYBL1-P349-F	0.02	VApm7, SYBL1 (Vesicle-associated membrane protein 7)	Involved in cell fusion and transportation of vesicles to their transmembranes and is essential for exocytosis during immune response.
617-HIF1A-P488-F	0.04	HIF1A (Hypoxia-inducible factor 1-alpha)	Controls homeostatic response, activates transcription of many genes, e.g. metabolism angiogenesis and apoptosis, and facilitates cellular oxygen intake.
518-G6PD-P196-F	$2.31 \times E{-}61$	G6PD (Glucose-6-phosphate 1-dehydrogenase)	Energy transfer and biosynthetic reactions.
334-DNMT1-P100-R	0.47	DNMT1 DNA(cytosine-5)-methyltransferase 1)	Regulates CpG methylation and maintains the DNA methylation pattern in the newly synthesized strands; this is important for epigenetic inheritance.
332-NASE1L1-P108-F	$3.36 \times E^{-39}$	DNASE1L1 (Deoxyribonuclease-1-like 1)	Cellular repair and production of 5'-phosphomonoesters and deoxyribonuclease activity.
847-MAP2K6-E297-F	0.04	MAP2K6 (mitogen-activated protein kinase kinase 6)	Regulates extracellular signal pathway, immune and transcription activation, and apoptosis.
636-HOXA11-P92-R	0.03	HOXA11(Homeobox protein(HOXA11))	Involved in female fertility and is active during embryonic development. Gene mutation causes radio-ulnar synostosis with amegakaryocytic thrombocytopenia.
1313-SYBL1-E23-R	0.13	VApm7,SYBL1 (Vesicle-associated membrane protein 7)	Involved in the transportation of protein and the activation of immune responses and natural killer cells.
158-CCNC-P132-R	0.58	CCNC (Cyclin-C)	Involved in protein phosphorylation and inhibits RNA polymerase initiation complex.

Author then investigated individual features and discovered that 12 out of the 18 features have a higher predictive accuracy of between 51.8% and 94.5% and, of the 18 identified CpG loci, 15 showed significant DNA methylation differences between the two genders, with high overall predictive accuracies. 10 out of the 18 features have as MCC value of between 0.75 ± 0.02 and 0.90 ± 0.02. GLA-P112-F was the highest selected feature; the associated gene plays an important part in enzymatic activity and protein synthesis [215]. The second most important feature was ARAF-E38-F; its associated gene plays a key role in cell differentiation and growth. The third feature was IRAK1-P312-F, which is associated with immune response. These features were significantly associated with gender and have P-values of 1.70E-99. In comparison with the individual features, ARAF-E38-F had the second highest predictive accuracy with a small standard error, and hence this feature may make a significant contribution to gender-related differences in dinucleotide patterns. ARAF-E38-F is the second most important feature, followed by IRAK1-P312-F. These two features have predictive accuracies of 94.5 ± 1.3 and 94.8 ± 1.9, and MCC values of 0.89 ± 0.03 and 0.89 ± 0.04 respectively. 15 out of 18 selected features have P-values which are extremely small, i.e. they have statistical discrimination power (Table 6.11). With regard to SYBL1-E23-R and CCNC-P132-R, the predictive accuracy was lower in the individual predictive model than in the feature sub-set, as expected. They both show insignificant P-values ($P5\%$) in the absolute t-test, but both were selected by the wrapper filtering method. This suggests that these two features interact with other optimal sub-sets, thus giving them a higher predictive accuracy in feature sub-sets than they do when evaluated as individual features. Further details of this are available in Table 6.12.

In addition, the performance of feature sub-sets demonstrated a higher level of predictive accuracy than that of individual features. This provides further confirmation of the interaction of the selected sub-sets, and leads to the tentative hypothesis of a biological link between the sub-sets. However, this is required to be proven using biological experiments. Overall, the performance of feature sub-sets was effective and the rate of misclassification errors was small, as noted in Figure 6.5, which shows that after 31 features, the minimum rate of misclassification error was attained, and the graph remains level from that point onwards.

Author then proceeded to investigate the differences between pairs of samples (control versus cancer) for both males and females. Using the methods described in the previous section, we discovered a significant association between HS3ST2-P171-f and cancer in both the male and female sample pairs, with a predictive accuracy of $99 \pm 1.02\%$ and $99.3 \pm 0.7\%$ respectively. The DNA methylation level of the best selected feature is plotted in Figure 6.6. This shows that there are significant methylation differences between the control (healthy) and cancerous samples: the median of the cancerous samples is higher, but there were no significant methylation differences between the healthy male and female samples. With regard to the cancerous samples

for males and females, the methylation distribution for females is slightly more skewed. The absolute t-statistics revealed significant differences between the two genders, with $P \leq 2.0E-62$ for males and $P \leq 7.03E-32$ for females. Figure 6.6 shows

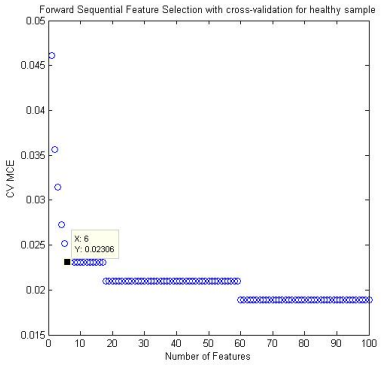


Figure 6.5 10-fold CV misclassification error rates and the sequential selected first 100 features from cancer samples.

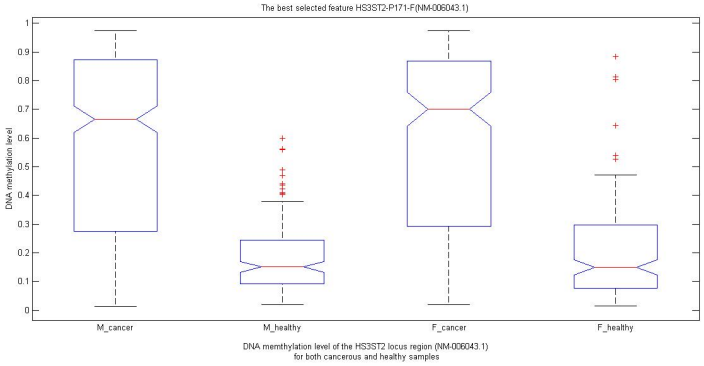


Figure 6.6 DNA memthylation level of the HS3ST2 locus region (NM-006043.1) for both cancerous and healthy samples.

the methylation distribution of this selected feature. This indicates that there are significant differences between the control (healthy) and cancerous samples, while the locus position (HS3ST2-P171-1) is highly methylated in both male and female cancerous samples. HS3ST2 could therefore serve as a biomarker which was found to be hypermethylated in various tumour types [216].

6.2.5 Discussion

This work attempts to predict differences in the methylation status of DNA which are related to gender. Author used feature selection methods and empirical predictive models in order to distinguish and select the important features of methylation loci based on 1506 features, including age distribution, from 328 healthy and 635 cancer-

ous samples of different tissues. As far as author is aware, this study is the most comprehensive to date. Author used the filtering method with statistical assessment, in addition to applying wrapper forward feature selection in order to investigate the interaction between features. Researcher assumes that sequential forward selection searched through the features and selected associated feature sub-sets. These selected sub-sets may have biological similarities, whereas t-statistics explored individual features separately, assuming that there is no interaction between the features.

6.2.5.1 Data analysis of the results from healthy (control) samples

Six loci positions (CpG) were selected from healthy samples, of which three were significantly associated with gender methylation; ELK1-P6-R, ELK1-E156-F and DNASE1L1-E178-R ($P \leq E - 45$). ELK1 binds purine-rich sequences and activates transcriptional factor binding sites. This leads to another gene, HDAC2, being recruited; this gene supresses histone acetylation and reduces protein expression [217]. In addition, preferential binding of ELK1 to the rs3122605 allele (G) was reported. This causes upregulation of the IL-10 protein, which has been found to be a risk biomarker for systemic lupus erythematosus (SLE), particularly in European Americans [218]. They found that T cells, B cells, and monocytes are highly expressed in samples from SLE sufferers and the severity of the disease increases with the activity of B cells. In addition, ELK1 was highly expressed in rheumatoid arthritis tissue, which suggests that it is linked with inflammation [219]. Furthermore, ELK1 was highly expressed in a blood sample [218] and rheumatoid arthritis tissues [219] of a patient with this auto-immune disorder. This protein binds to specific oligodeoxynucleotides of the IL-10 gene, which leads to it being expressed during the immune response. A high level of ELK1 gene concentration was also found in samples from patients with auto-immune diseases [218]. In addition, the Il-10 gene is found to be associated with famine exposure, which leads to an increase in DNA methylation in which female DNA is more highly methylated than that of males [220].

DNASE1L1, the second selected feature, repairs faulty DNA, controls the cell cycle, and is expressed in the heart and skeletal muscles [221]. An association between DNASE1L1 and SLE has been reported, in which the activity of this feature reduces with increasing severity of the disease. It has been assigned as one of the five biomarkers associated with SLE [222]. In animal models, females were found to have more antibodies against dsDNA than males [223]. Furthermore, patients with the DNASE1L1 mutation showed reduced activity [224]. However, other studies failed to find this mutation in SLE patients [225]. Since none of the above mentioned studies examined DNA methylation, it is reasonable to suggest that DNA methylation may activate the development of SLE. Other studies reported that the dinucleotide methylation in males increases with age, whereas age had less influence on methylation in females [226; 227].

6.2.5.2 Tissue-specific analysis of leukocytes

The next stage of this study was to investigate whether the selected features are tissue specific. Using the wrapper method incorporated with ten-fold cross-validation and a QDA classifier, 25 out of a total of 1506 features were selected from normal leukocytes. These features contain ten feature sub-sets which were obtained by ten times repeated analysis, as detailed in the Results section.

Five out of the selected 25 features show significant methylation differences between males and females, with good predictive accuracies and MCC values. BIRC4-P122-R, ELK1-P6-R, HDAC6-E102-F, ELK1-E53-F and DLG3-P62-R have shown the best predictive accuracies for individual feature analysis (Table 6.7). Furthermore, visualising the best selected sub-set confirms that there is a significant difference in methylation between the two genders: BIRC4-P122-R and HDAC6-E102-F revealed that female samples were highly methylated, whilst male samples showed evidence of hypomethylation. In the same sub-set, the selected feature IL10-P348-F showed no significant methylation differences between the two genders (Figure 6.3). Moreover, the sub-set of BIRC4-P122-R, HDAC6-E102-F and IL10-P348-F had the highest predictive accuracy and Matthews correlation coefficient (MCC); these were $85.2 \pm 1.2\%$ and 0.70 ± 0.03 respectively. The gene associated with X-linked lymphoproliferative syndrome (also termed BIRC4), is required for homeostatic immune response regulation during the cell-division and apoptosis processes. The gene associated with DLG3-P62-R plays an important role in binding neurotransmitter receptors and is expressed in the brain during early development. Mutations of the DLG3 gene cause mental retardation and learning difficulties where the gene-mutation indicated severe depression [228]. However, this study reported that whole genome and CDNA sequencing of the affected male who presented with these symptoms did not reveal the existence of a DLG3-mutation; this suggests that DNA methylation might be the cause of gene disruption of the studied males. The gene associated with HDAC6-E102-F is essential for cell recovery and stress response, and it was reported that disruption of the HDAC6 gene leads to cell apoptosis [229]. In addition, HDAC6 regulates the presentation of T cells to lymphocytes by the formation of immune synapses [230]. VBP1 acts as a molecular chaperone protein, and is assumed to be essential for transporting the Von Hippel-Lindau protein from the perinuclear granules to the cytoplasm. VBP1-E127-F was found to be highly methylated in females, but hypomethylated in males. It has been reported that VBP1is primarily expressed in the brain [214].

6.2.5.3 Analysis of the data from normal colon samples

Author was able to use DNA methylation features to distinguish the males from females. The first ten selected features showed significant gender-related differences with small P-values ($P \leq E - 20$); further analysis of individual features has resulted

in higher predictive accuracies and MCC of between $96.7 \pm 1.4\%$ and $100 \pm 0.00\%$, and 0.93 ± 0.06 and 1.00 ± 0.00 respectively. Further details of this are available in Table 6.9. At the SLC6A8 locus position, DNA was found to be highly methylated in female samples but hypomethylated in males. In another study, this was reported to be tissue-specific, affecting only the testes [231] However, since our data does not relate to gender-specific tissues this indicates that SLC6A8 is not only specific to the testes. Furthermore, the same study reported that all promoters of three tissues were unmethylated in their CpG islands. In addition, SLC6A8 was expressed in the skeleton muscle, kidneys, testes, colon, heart, brain, small intestine, and prostate. The prediction and feature selection confirms a significant difference in methylation distribution between males and females (Figure 6.3). This suggests that SLC6A8 is tissue specific [232]. However, our results have shown that SLC6A8 is gender-specific, as can be noted by the significant difference in methylation levels between the two genders. Author expect that if the SLC6A8 locus position is tissue-specific, it will be selected from leukocyte samples, which are contained with almost all tissues, rather than from colon tissue. An experiment is required in order to identify the real function of SLC6A8; it has been suggested that its function is to regulate the second copy of the X-chromosome, which guides production of a methylation signal that may be hidden in the promoter region. With the line of DNA methylation minimised, the complexity of the genetic reproduction is guided by the memory of cell fate [20]. Moreover, EFNB1-P17-F was one of the top ten selected features from normal colon tissue in which a significant difference in methylation status between males and females was identified. The gene associated with EFNB1 was reported to be linked with X-chromosome inactivation, and has shown higher levels of methylation in females than in males [210]. Furthermore, mutation of this gene affects females more severely than males [233], and 92% of the mutations were found to have a paternal origin. Our study indicates that DNA methylation may contribute to severe defects which mostly affect females, since some patients showed symptoms even though no mutations were identified in the studied samples [234].

6.2.6 Data analysis of cancer classification samples

Eighteen features were selected from cancer samples dataset, 14 of which have shown discriminatory power with P-values of less than 5%. GLA-P112-F was the best selected feature; the gene associated with this feature has an important role in enzymatic activity: it regulates protein synthesis, processing, and protein stability. In addition, the sub-set of GLA-P112-F, DKC1-P276-F, HDAC6-E102-F, G6PD-P196-F, NASE1L1-P108-F, NQO1-E74-R and HIF1A-P488-F was found to be the optimal feature sub-set with the highest performance. In this sub-set, significant gender-related differences were observed ($P<5\%$). Furthermore, HOXA is heritable and gender-specific [226], with higher levels of methylation found in females than in males of

the same age. DNMT1 was found to be insignificant in individual feature selection, with $P5\%$. However, it was selected by forward feature sub-set selection as a result of feature interaction, and higher levels of methylation were found in males than in females [235]. ELK1 and DNMT1 have not been statistically selected within the same feature sub-set. However, they are included in the optimal feature sub-sets for healthy and cancerous samples respectively. Consequently, we can assume that there is no direct interaction between these two features. However, they have both been reported to influence the epigenetic process. For example, ELK1 activates HDAC2, which leads to a reduction of histone acetylation, and ERK, which leads to a reduction of histone proteins and phosphorylation [236]. The phosphorylated HDAC2 and Akt1 prevent methylation and maintain the stability of DNMT1 [217]. DNMT1 is mostly responsible for the methylation of cancerous cells and the repression of tumour suppressor genes [237]. In addition, DNMT1 was differentially methylated in various tissues, with higher levels found being in females than in males [238].

For the purpose of validation, we calculated the predictive accuracy and Matthews correlation coefficient values of individual features. As summarised in Table 6.2, we annotated their biological functions in healthy tissue. Tables 6.5, 6.8, and 6.11 give the functions of these features in leukocytes, colon tissue, and cancerous samples respectively. ELK1-P6-R and GLA-P112-F were the two selected features which were found to be the most significantly associated with gender-related DNA methylation differences, with P-values of $P \leq 7.03E-68$. These two features also had high predictive accuracies with small standard errors, i.e. of $92.2 \pm 2.0\%$ and $93.1 \pm 1.1\%$ respectively. Moreover, the same feature was obtained from the analysis of pairs of cancerous and healthy samples obtained from both males and females. This feature is HS3ST2, a locus region (NM-006043.1) that is significantly associated with cancer. For both males and females, the P-values obtained were small: $P \leq 2.0E-62$ for males and $P \leq 7.03E-32$ for females. They both yielded a high predictive accuracy of $99 \pm 1.02\%$ and $99.3 \pm 0.7\%$ respectively. Other studies reported that HS3ST2 was hypermethylated in samples from tissues affected by breast, colon, pancreatic and lung cancer [216]. Furthermore, it was also found that HS3ST is repressed in gastrointestinal tumours [239], suggesting that a DNA methylation process is responsible.

These results confirm that there are differences between males and females which can be distinguished on the basis of DNA methylation fingerprints. These differences may play a very important role on the effects of drug administration, disease prevention, and the environmental impact of individuals (i.e., behaviour). These effects can be explained by factors such as alcohol influence, drug dosage and diet intake; it has already been confirmed that dietary intake affects females more than males [240]. However, other factors may also have an influence, such as age, weight and environmental influence; further investigations of these would therefore be worthwhile. The findings of this study may be limited by the fact that healthy data contains pooled

tissues which make it difficult to compare the outcomes. However, author used feature selection in order to select the informative features, and the results were also validated with leukocytes and colon tissue as noted in the Results section. Another limitation of the data source is the fact that it represents 807 genes out of 25000 of the human genome, i.e., the results cannot be completely conclusive unless the study is extended to the whole human genome.

6.2.7 Conclusion

Biological characteristics which aid the detection of DNA methylation have been found to be different in different diseases and conditions such as tumours, auto-immune diseases, and mental disorders. However, few studies demonstrate the existence of any specific molecular difference that can help distinguish between male and female subjects. This could help us understand the molecular differences between males and females; this would, in turn, lead to better modelling and management of the disease processes. It is therefore important to identify gender-based differences in DNA methylation, and for this purpose, a bioinformatics study was conducted in order to analyse the data collected from different tissues of 963 samples, consisting of 328 healthy (168 male and 160 female) and 635 cancerous samples (404 male and 231 female). We investigated the CpG methylation positions of 1506 features from the 963 samples in order to determine whether these contribute additional information regarding variations in methylation according to gender. Author employed sequential forward feature selection combined with QDA in order to select the most informative features or feature sub-sets, and then assess the predictive accuracies and performance of both these models.

From the analysis of 1506 features, 6 biological ones were identified to be highly associated with gender-related differences in healthy samples, and 18 in cancerous samples. Among these features, the loci positions ELK1, DNASE1L1 and ROR2 were the best selected feature sub-set for the healthy samples, yielding the highest predictive accuracy of $94.1 \pm 1.2\%$ and Matthews correlation coefficient (MCC) values of 0.88 ± 0.02, whereas the best selected feature subset for the cancerous samples included the loci positions GLA-P112-F, BIRC4-P122-R, HDAC6-E102-F, GLA-E98-R, NQO1-E74-R and MAP2K6-E297-F, which resulted in the highest predictive accuracy and (MCC) values of $97.5 \pm 1.5\%$ and 0.95 ± 0.03, respectively. Furthermore, comparison of pairs of samples from subjects of the same sex, the locus position HS3ST2-P171-f was found to be a specific to cancer only with a predictive accuracy of $99 \pm 1.02\%$and $99.3 \pm 0.7\%$ for males and females respectively.

In addition, we also characterised 25 features (CpG loci positions) from normal leukocytes and 10 from colon tissues. These features revealed significant ($P \leq 6.8E-21$) methylation differences between males and females. It was also observed that ELK1, BIRC4-P122 and GLA are frequently repeated in the sub-sets and therefore

can be regarded as the best individual gender discriminator for healthy and cancerous samples. The respective predictive accuracies for these features were 92.2±2.0%, 85.2± 1.2% and 93.1±1.1%. To date, the selected features have not been reported [63], which indicates that they have potential to serve as biomarkers in our studies.

Author are able to predict gender-associated features from the whole-genome fingerprint of the DNA methylome in a sample of healthy and cancerous individuals. We observed that ELK1-P6-R, BIRC4-P122-R and GLA have a strong association with gender. Higher levels of activation of ELK1 in the immune system were found in females rather than in males [218]. The most significant DNA methylation differences between males and females were identified in the loci positions ELK1-P6-R and BIRC4-P122-R, although their associated genes have a different function. ELK1 activates a transcription factor for binding, especially at AT-rich DNA binding sites, whereas the gene associated with BIRC4 suppresses the transcription factor and further activates the cell-death process. Furthermore, the locus region HS3HT2 was the best selected feature when comparing cancerous samples with the control (healthy) samples, thus confirming that this feature is highly hyper-methylated in various tumour types [216]. Researcher retrieved the features which were most highly associated with gender, annotated them with their biological functions from the human genome database, and presented their theoretical and biological processes. Author were able to distinguish male- and female-associated features with a predictive accuracy of between 74.5% and 100%. At these loci points, methylation appears to be different for each gender, and gender differences widely influence methylation, and so may play a key role in associated diseases and mental problems, behavioural changes and immune disorders.

An identification of gender-associated loci where DNA methylation is under genetic control will facilitate future investigations into dinucleotide methylation and its association with disease. However, this investigation does not examine the impact of age and environmental factors on DNA methylation; further research in this area is therefore recommended.

Chapter 7

Thesis conclusions and Future Study

This section summarises the different approaches to the problem that have been developed across this thesis. One of the most difficult problems to analyse consisted of the genomic, epigenetic and systems biology data, whilst the imbalanced data was imprecise and biased towards the prediction of direct machine learning. This was demonstrated during the analysis of all data in which the imbalance responded badly. As stated by [5]: 'Regular leave-one-out with KNN cannot predict imbalance feature-sets.' Redesigning leave-one-out with KNN could well predict imbalance feature-sets.

This study, focused on the CpGs sequence features of human chromosomes, has encompassed the following: the calculation of DNA sequence patterns; grouping the sequence features to their biological functions; predicting and analysing DNA methylation classes; clustering individual methylation differences in ageing and gender; selecting CpG loci positions specific to gender; and developing fair and suitable predictive models. There now follows a summary of the significance of this thesis and its contribution to the field, along with a discussion of the strengths and limitations of the work, with recommendations for the future direction of research in this area.

7.1 Brief summary of the work

This study has developed a fair predictive model and established that the methods employed serve as an effective improvement on traditional methods. There has been an identification of DNA sequence features that interplay methylation classes, a process providing the ability to extract and predict CpG loci positions specific to differences in relation to gender and ageing.

DNA methylation forms one of the most important and remarkable phenomena processes of gene expression regulation. In view of their importance, DNA patterns have been investigated in great depth, particularly DNA sequence features, since DNA

letters have been annotated [241]. However, there has been little previous focus on the study of DNA methylation classes. Experimental analysis has attempted to identify DNA methylation patterns, along with their biological functions and the manner in which they control gene expression, and which features (or feature sets) plays an important role. Two methods are currently in use for the investigation of DNA methylation, including computational prediction and experimental laboratory (i.e. microarrays and illumina-array). A number of studies have investigated specific tissues in order to identify methylation patterns [56; 242]. A number of tools of bioinformatics have been developed to extract DNA sequence features, and various predictive models have been used to distinguish methylated and unmethylated classes which can be applied directly to machine learning algorithms [48; 64]. However, these have not produced reliable prediction methods, due to a number of the datasets were severely imbalanced, and a failure to group the feature sub-sets into their biological functions. In order to tackle these issues, it has been necessary to investigate DNA sequence patterns, and also to apply the most effective predictive model, as well as grouping methylation classes based on their biological associations.

Therefore, this current study has developed a Modified Leave-One-Out cross validation (MLOOCV) strategy, and also grouped the extracted features to their biological functions, which were then applied to the MLOOCV method (Chapters 3 and 5). Comparative studies of CpGs methylation gave the ability to distinguish methylated classes from unmethylated ones, and also differentially- methylated ones. The study revealed that the following were significantly associated with DNA methylation: tissue-specific feature subsets; DNA sequence properties; exon and gene distribution sub-sets; single polymorphism sub-sets; and DNA structure subsets. It has also been demonstrated experimentally that a small sample size severely affects predictive performance, when compared to that achievable with medium-imbalanced datasets. It was therefore proposed that methods (i.e. cost-sensitive and weighting approaches) were adopted, with an Adoboost algorithm combined with a decision tree classifier. This proposed method revealed a high predictive performance for the distinction of methylated, differentially-methylated and unmethylated classifications for imbalanced datasets, particularly the minority class (methylated class), without losing predictive performance in relation to the majority class (unmethylated class).

This work has considered various learning algorithms for the analysis of DNA methylation classes based on DNA sequence features, i.e. KNN, Decision tree and QDA . These have produced improved predictive performances, further enhanced by grouping the features into their biological functions, along with all possible combinations of feature sub-sets. The research has shed light on DNA methylation classes and predicts features associated with this classification criterion, as well as having further identified that the loci position has a different methylation status in males and females in both healthy and cancerous samples.

There now follows a summary of the contribution of this thesis, based on the chapters under discussion. The following points give an outline of the original contribution of this research:

DNA methylation status in relation to gender and ageing is the one of most important 21^{st} century issues in relation to epigenetic and personalised medication. Chapter 4 investigated a set of loci (features) extracted from 1505 DNA sequence methylation positions. The 47 CpG loci positions specific to gender were extracted by the use of hierarchical clustering (Heatmap) with pair correlation distance methods. The CpG loci revealed significant methylation differences between genders for healthy samples. Heatmap algorithms group both gender and age associated loci (features), while disregarding others. An average linkage method was employed, and the sub-group further enlarged, in which each branch (dendrogram) represented a feature connected with a group of features, which have been graphically displayed. 11 CpG loci positions were identified, which were methylated in healthy samples, whereas the same CpG loci positions were unmethylated in cancerous samples (Table 4.4 provides details).

In Chapter 5, a comprehensive analysis was undertaken to identify and distinguish methylated, unmethylated and differentially-methylated classes, based on extracted DNA sequence features. The extracted features were grouped according to their biological functions, as reviewed in Chapter 2 and summarised in Table 2.2.

A comprehensive analysis has been undertaken involving KNN, combined with ten-fold cross validation. Since a number of the datasets were severely imbalanced, it has been established that direct analysis in machine learning is biased towards the majority class, i.e. the class with the most observations. This was required to design a reliable model to tackle such problems. The MLOO method was developed, which demonstrated improved predictive performance and consistent predictive accuracies in comparison to traditional KNN. It has been demonstrated that LOOCV with KNN reveals a bias towards the majority class, whereas MLOOCV demonstrates a consistent predictive performance, which predicts the following: SNPs sub-set; tissue-specific feature sub-sets; DNA sequence properties; exon and gene distribution sub-set; single polymorphism sub-set; and DNA structure sub-set. MLOOCV can also be used for any imbalanced datasets impossible to predict directly into machine learning, as it reveals fluctuated and inconsistent results.

The same dataset was comprehensively analysed, with a decision tree as a base classifier AdaBoost method. Adaboost is designed for small and imbalanced data, unlike MLOOCV methods. It reveals good predictive accuracies, however, since it is more computationally-expensive than MLOOCV. The researcher proposed the use of a combination of methods, i.e., weighting- and cost-sensitive ones, which resulted in improvements to the predictive performance of the minority class. The weighting assisted not only with the accuracy of the minority class, but also those classes not easy to separate. The results demonstrated an imbalance metrics improvement,

including the consideration of F-measure and G-mean parameters. The values were judged on how well the model responded to the unseen data (test set). The combinatorial methods revealed an improved predictive performance for methylation classes in comparison to the single method (either weighting or prior probability approaches).

In Chapter 6, a comprehensive feature selection method was built, in order to distinguish DNA methylation differences between males and females by the investigation of specific features (loci positions) that play a role in DNA methylation for healthy and cancerous data. There was a further investigation of tissue specificity for males and females in DNA methylation differences in relation to the health of both genders.

This study has aimed to distinguish DNA methylation differences between genders, in order to assist in the design of a target drug, and also improve the manner in which patients are treated, which has been based on the weight of patients without considering the impact of other factors, such as DNA methylation, tissue specificity, and drug resistance or toxicity that may cause DNA methylation variation in differing tissues. These factors form the future direction of DNA methylation studies.

In this Chapter, 10 features were identified in a colon tissue sample, which demonstrated DNA methylation differences in both healthy male and female samples. This is a major finding, since these features are not associated with gender specific-tissues. A classifier was built, combined with feature selection methods (t-test and wrapper). The gender difference features revealed a high predictive accuracy for both individual selected features used by the t-test, as well as the sub-set-selected features also used by the wrapper method. This work has identified the most informative features, or feature sub-sets, distinguishing DNA methylation differences between healthy male and female samples. The soundness of the model has been demonstrated by the consideration that, during the selection process, none of the features were ignored, and all went through in the search with a combination of pairing; those with the lowest misclassification error rate were selected.

7.1.1 Strengths and limitations

A strength of this thesis consists of the development of fair predictive models (MLOOCV). Traditional methods (i.e. hold-out and m-fold cross validation) can only be valid with a large sample size and balanced dataset, and cannot handle as imbalanced one [48; 64]. A further strength consists of teaching machine-learning in accordance with the behaviour of the dataset, including pre-processing, feature selection and data-mining. This study has utilised and adapted the most intelligent means of overcoming bias towards an imbalanced dataset. The predictive model involves the prediction of feature sub-sets associated with DNA methylation, establishing that DNA methylation tends to be associated with tissue specificity, whereas the differentially-methylated class are correlated with DNA sequence distributions [27].

In this work, the researcher has been able to predict DNA methylation classes, in addition to identifying and profiling CpG loci methylation position differences between healthy and also cancerous individuals of both genders, from 963 samples across a wide age range. Feature selection methods have been designed, and interrogated CpG loci methylation positions specific to both healthy and cancerous individuals of both genders. Taken together, these comprehensive analyses have provided an exploration of CpG methylation classes, CpG methylation differences in relation to gender, and results on DNA methylation changes results in cancer samples. This has made possible the comprehensive predictive analysis and investigation of DNA methylation variability in this work.

However, this work indicated a number of limitations. The prediction model (MLOOCV) is utilised with a K-nearest neighbour classifier. Hence, a further classifier is required to compare the results acquired, and it would also be beneficial to extend the predictive model with a Support Vector Machine (SVM) approach.

The number of CpGs used in the study was limited. They were not represented in all CpGs loci positions of the genome, as they only represented a small fraction of human genes (less than 3.2%), which requires further extension in order to reduce the limitations of the study. This can be investigated further by employing a greater proportion of the entire human genome arrays in order to ensure a complete conclusion.

The potentially most critical component missing in this work is the tissue specificity CpG methylation differences according to gender. Nevertheless, it has been reported that methylation is a life-long continuous process, which is not always the case, since CpG methylation is specific to one tissue. Hence, targeting tissue-specific methylation in order to compare methylation variation in ageing and gender may not produce useful information. Despite this issue, the current study has investigated CpG methylation differences between males and females in Leukocytes and Colon tissues collected from a healthy population, which is self-referencing and hence eliminates such bias.

7.1.2 Future direction of this work

DNA sequence features have been analysed and briefly reviewed in previous chapters. Therefore, this sub-section will further discuss the direction of future research.

In Chapter 3, three independent samples were presented in order to study DNA methylation features. These features were grouped into their biological functions, as listed in Table 2.2. This is required to be extended and further sub-grouped, in order to ensure that it contains an improved representation of the complete human DNA features (should future technology allow), since the cell contains three billion base pairs of DNA letters.

In Chapter 4, hierarchical clustering was used to identify CpG methylation differences according to gender and age. Further methods are required in order to compare the results, and further investigation is needed into geographical methylation differences in relation to environment and ethnicities.

Chapter 5 contained an investigation of methylation differences between three classes (i.e. methylated, unmethylated and differentially methylated), based on calculated CpG island sequence features. In this chapter, various classification models were developed and reported in the thesis, including MLOOCV and Combinatorial Adaboost. Both methods demonstrate improvements in predictive performance in comparison to traditional LOOCV and Adaboost. Both methods employed three classes of prediction problems, which could be extended into multi-classes prediction methods for imbalanced data.

In Chapter 6, a forward feature selection method was created and combined with a quadratic discriminate analysis (QDA) classifier which identified a number of gender difference features (loci position). The method produced a high predictive performance for individual and sub-set features. The studied dataset in Chapter 6 contained only 807 gene arrays, which were screened in relation to the entire human genome. Thus, the arrays comprised 3.2% of the human genome required to further extend the total human genome, in order to attain a complete conclusion. In addition, samples for both control (healthy) and diseased (cancerous) samples could be extended in order to produce a comprehensive analysis and representation of a specific population.

7.1.3 Final conclusions

In this report, comprehensive CpGs methylation predictions for both healthy and diseased human tissues (data) have been established. Such predictions contribute to a DNA methylation predictive model, including a severely imbalanced dataset, which modifies the manner that data can be analysed with a fair predictive model, i.e. by modelling the data before applying machine-learning. This work also leads to new directions in this field, illuminating the means in which the current diagnostic and medication of cancer patients is questionable, since it has not considered influence of gender on the DNA methylation status. Hence, this work should eventually lead to improvements of current cancer diagnostics, prevention and medication, particularly in relation to personalised drug design.

References

[1] nihroadmap.nih.gov. accessed 20/08/2015. xi, 2

[2] ghr.nlm.nih.gov/handbook/basics/chromosome. accessed on 25/08/2015. xi, 3

[3] S. Theodoridis and K. Kouttroumbas, *Pattern Recognition*. Elsevier, P(261-586), (2009). xi, 27, 28, 29, 30, 34, 47, 50, 57, 59

[4] A. K. Jain, R. P. Duin, and J. Mao, "Statistical pattern recognition: A review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 1, pp. 4–37, 2000. xi, 24, 25, 28, 29, 58, 60, 127

[5] I. Ali and H. Seker, "Detailed methylation prediction of cpg islands on human chromosome 21," in *Proceedings of the 10th WSEAS international conference on Mathematics and computers in biology and chemistry*, (Stevens Point, Wisconsin, USA), pp. 147–152, World Scientific and Engineering Academy and Society (WSEAS), 2009. xv, 9, 26, 28, 32, 34, 35, 42, 44, 55, 64, 83, 98, 110, 126, 149

[6] I. Ali *et al.*, "An identification and prediction methods for feature-subsets of cpg islands methylation based on human peripheral blood leukocytes of chromosome 21q," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pp. 3233–3236, IEEE, 2011. xv, 9, 106

[7] M. Hattori, A. Fujiyama, T. D. Taylor, H. Watanabe, T. Yada, H.-S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.-K. Choi, E. Soeda, M. Ohki, T. Takagi, Y. Sakaki, S. Taudien, K. Blechschmidt, A. Polley, U. Menzel, J. Delabar, K. Kumpf, R. Lehmann, D. Patterson, K. Reichwald, A. Rump, M. Schillhabel, A. Schudy, W. Zimmermann, A. Rosenthal, J. Kudoh, K. Shibuya, K. Kawasaki, S. Asakawa, A. Shintani, T. Sasaki, K. Nagamine, S. Mitsuyama, S. E. Antonarakis, S. Minoshima, N. Shimizu, G. Nordsiek, K. Hornischer, P. Brandt, M. Scharfe, O. Schon, A. Desario, J. Reichelt, G. Kauer, H. Blocker, J. Ramser, A. Beck, S. Klages, S. Hennig, L. Riesselmann, E. Dagand, T. Haaf, S. Wehrmeyer, K. Borzym, K. Gardiner, D. Nizetic, F. Francis, H. Lehrach,

R. Reinhardt, and M.-L. Yaspo, "The dna sequence of human chromosome 21," *Nature*, vol. 405, pp. 311–319, May 2000. 3

[8] C. E. Dictionary-Complete, "Unabridged 10th edition 2009© william collins sons & co," *Ltd. URL: http://dictionary. reference. com/browse/(accessed August 20, 2015)*, 1979. 4

[9] J. M. Levenson and J. D. Sweatt, "Epigenetic mechanisms in memory formation," *Nat Rev Neurosci*, vol. 6, pp. 108–118, Feb. 2005. 5, 6, 19, 20

[10] A. P. Feinberg, "Phenotypic plasticity and the epigenetics of human disease," *Nature*, vol. 447, pp. 433–440, May 2007. 5, 6, 11, 13, 21, 22, 23

[11] J. Mill, T. Tang, Z. Kaminsky, T. Khare, S. Yazdanpanah, L. Bouchard, P. Jia, A. Assadzadeh, J. Flanagan, A. Schumacher, *et al.*, "Epigenomic profiling reveals dna-methylation changes associated with major psychosis," *The American Journal of Human Genetics*, vol. 82, no. 3, pp. 696–711, 2008. 5, 6, 7, 22, 80

[12] I. Keshet, Y. Schlesinger, S. Farkash, E. Rand, M. Hecht, E. Segal, E. Pikarski, R. A. Young, A. Niveleau, H. Cedar, and I. Simon, "Evidence for an instructive mechanism of de novo methylation in cancer cells," *Nat Genet*, vol. 38, pp. 149–153, Feb. 2006. 5, 7, 11, 19, 23, 80

[13] C. B. Yoo and P. A. Jones, "Epigenetic therapy of cancer: past, present and future," *Nat Rev Drug Discov*, vol. 5, pp. 37–50, Jan. 2006. 5, 23

[14] N. Stransky, C. Vallot, F. Reyal, I. Bernard-Pierrot, S. G. D. de Medina, R. Segraves, Y. de Rycke, P. Elvin, A. Cassidy, C. Spraggon, A. Graham, J. Southgate, B. Asselain, Y. Allory, C. C. Abbou, D. G. Albertson, J. P. Thiery, D. K. Chopin, D. Pinkel, and F. Radvanyi, "Regional copy number-independent deregulation of transcription in cancer," *Nat Genet*, vol. 38, pp. 1386–1396, Dec. 2006. 5, 6, 7, 21, 23

[15] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & development*, vol. 16, no. 1, pp. 6–21, 2002. 5, 6, 11, 13, 16, 18, 19, 21, 22, 23

[16] D. Rodenhiser and M. Mann, "Epigenetics and human disease: translating basic biology into clinical applications," *Canadian Medical Association Journal*, vol. 174, no. 3, pp. 341–348, 2006. 5, 21

[17] J. Bradbury., "Human epigenome project-up and running," *Plos Biology*, vol. 1(3), pp. 316–319, 2003. 5, 6

[18] A. Bruce, B. Dennis, L. Julian, R. Martin, R. Keith, and W. James D, *Molecular Biology of The Cell*. New York: Garland Science, 1994. 5, 6, 17, 18, 19, 20, 21, 22

[19] F. Antequera and A. Bird, "Number of cpg islands and genes in human and mouse," *Proceedings of the National Academy of Sciences*, vol. 90, no. 24, pp. 11995–11999, 1993. 6, 11, 12, 18, 19, 22, 80

[20] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & Development*, vol. 16, no. 1, pp. 6–21, 2002. 6, 7, 63, 64, 80, 125, 126, 144

[21] A. P. Feinberg, "Phenotypic plasticity and the epigenetics of human disease," *Nature*, vol. 447, pp. 433–440, May 2007. 6, 7, 63, 80, 126

[22] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, 2007. 6, 80

[23] R. Barres and J. R. Zierath, "Dna methylation in metabolic disorders," *The American journal of clinical nutrition*, vol. 93, no. 4, pp. 897S–900S, 2011. 7, 64

[24] C. Murgatroyd, A. V. Patchev, Y. Wu, V. Micale, Y. Bockmühl, D. Fischer, F. Holsboer, C. T. Wotjak, O. F. Almeida, and D. Spengler, "Dynamic dna methylation programs persistent adverse effects of early-life stress," *Nature neuroscience*, vol. 12, no. 12, pp. 1559–1566, 2009. 7, 64, 126

[25] I. Ali, D. Elizondo, and M. Grootveld, "Dna methylation display on aging and gender differences based on unsupervised clustering.(chapter-4)." In preparation. 8

[26] I. Ali and H. Seker, "A comparative study for characterisation and prediction of tissue-specific dna methylation of cpg islands in chromosomes 6, 20 and 22," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 1832–1835, IEEE, 2010. 9, 26, 30, 41, 83

[27] I. Ali, D. Elizondo, and M. Grootveld, "Prediction of methylation classes of cpg islands on chromosomes 6, 20, 21 and 22.." Unpublished experimental results, pp.1-20,2015. 9, 122, 152

[28] I. Ali, D. Elizondo, and M. Grootveld, "Weighting methods towards severely imbalanced data (chapter-5).." due to submit. 9

[29] I. Ali, M. Grootveld, and D. Elizondo, "Analysis of gender differences in dna methylation. (chapter-6)." due to submit. 9, 116

[30] R. M. Brena, T. H.-M. Huang, and C. Plass, "Toward a human epigenome," *Nat Genet*, vol. 38, pp. 1359–1360, Dec. 2006. 11, 12, 23, 80

[31] R. J. Weeks and I. M. Morison, "Detailed methylation analysis of cpg islands on human chromosome region 9p21," *Genes, Chromosomes and Cancer*, vol. 45, no. 4, pp. 357–364, 2006. 11

[32] V. K. Rakyan, T. Hildmann, K. L. Novik, J. Lewin, J. Tost, A. V. Cox, T. D. Andrews, K. L. Howe, T. Otto, A. Olek, J. Fischer, I. G. Gut, K. Berlin, and S. Beck, "Dna methylation profiling of the human major histocompatibility complex: A pilot study for the human epigenome project," *PLoS Biol*, vol. 2, p. e405, 11 2004. 12

[33] Y. Yamada, H. Watanabe, F. Miura, H. Soejima, M. Uchiyama, T. Iwasaka, T. Mukai, Y. Sakaki, and T. Ito, "A comprehensive analysis of allelic methylation status of cpg islands on human chromosome 21q," *Genome Research*, vol. 14, no. 2, pp. 247–266, 2004. 12, 16, 17, 41, 80, 81

[34] M. Gardiner-Garden and M. Frommer, "Cpg islands in vertebrate genomes," *Journal of molecular biology*, vol. 196, no. 2, pp. 261–282, 1987. 12

[35] D. Takai and P. A. Jones, "Comprehensive analysis of cpg islands in human chromosomes 21 and 22," *Proceedings of the National Academy of Sciences*, vol. 99, no. 6, pp. 3740–3745, 2002. 12, 19

[36] F. Antequera, "Structure, function and evolution of cpg island promoters," *Cellular and Molecular Life Sciences CMLS*, vol. 60, no. 8, pp. 1647–1658, 2003. 12

[37] Y. Wang and F. C. Leung, "An evaluation of new criteria for cpg islands in the human genome as gene markers," *Bioinformatics*, vol. 20, no. 7, pp. 1170–1177, 2004. 12

[38] H. Cedar and Y. Bergman, "Linking dna methylation and histone modification-patterns and paradigms," *Nature Reviews Genetics*, vol. 10, no. 5, pp. 295–304, 2009. 12

[39] C. De Bustos, E. Ramos, J. M. Young, R. K. Tran, U. Menzel, C. F. Langford, E. E. Eichler, L. Hsu, S. Henikoff, J. P. Dumanski, *et al.*, "Tissue-specific variation in dna methylation levels along human chromosome 1," *Epigenetics & chromatin*, vol. 2, no. 1, p. 7, 2009. 12

[40] R. A. Rollins, F. Haghighi, J. R. Edwards, R. Das, M. Q. Zhang, J. Ju, and T. H. Bestor, "Large-scale structure of genomic methylation patterns," *Genome Research*, vol. 16, no. 2, pp. 157–163, 2006. 12

[41] E. Schilling and M. Rehli, "Global, comparative analysis of tissue-specific promoter cpg methylation," *Genomics*, vol. 90, no. 3, pp. 314 – 323, 2007. 12, 80

[42] J. F. Costello, M. C. Frühwald, D. J. Smiraglia, L. J. Rush, G. P. Robertson, X. Gao, F. A. Wright, J. D. Feramisco, P. Peltomäki, J. C. Lang, *et al.*, "Aberrant cpg-island methylation has non-random and tumour-type–specific patterns," *Nature genetics*, vol. 24, no. 2, pp. 132–138, 2000. 12

[43] M. Esteller, "Epigenetics in cancer," *New England Journal of Medicine*, vol. 358, no. 11, pp. 1148–1159, 2008. 12, 64, 126

[44] M. Esteller, "Cancer epigenomics- dna methylomes and histone-modification maps," *Nature Reviews Genetics*, vol. 8, no. 4, pp. 286–298, 2007. 12

[45] F. A. Feltus, E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino, "Predicting aberrant cpg island methylation," *Proceedings of the National Academy of Sciences*, vol. 100(21),, pp. 12253–12258,, 2003,. 12, 16, 26

[46] F. Fang, S. Fan, X. Zhang, and M. Q. Zhang, "Predicting methylation status of cpg islands in the human brain," *Bioinformatics*, vol. 22, no. 18, pp. 2204–2209, 2006. 12, 26, 80

[47] S. Fan, M. Q. Zhang, and X. Zhang, "Histone methylation marks play important roles in predicting the methylation status of cpg islands," *Biochemical and biophysical research communications*, vol. 374, no. 3, pp. 559–564, 2008. 12, 26

[48] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter, "Cpg island methylation in human lymphocytes is highly correlated with dna sequence, repeats, and predicted dna structure," *PLoS Genet*, vol. 2, p. e26, 03 2006. 12, 16, 20, 21, 26, 28, 41, 44, 64, 81, 84, 123, 126, 150, 152

[49] Y. Zhang, C. Rohde, S. Tierling, T. P. Jurkowski, C. Bock, D. Santacruz, S. Ragozin, R. Reinhardt, M. Groth, J. Walter, and A. Jeltsch, "Dna methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution," *PLoS Genet*, vol. 5, p. e1000438, 03 2009. 12, 13

[50] D. Jia, R. Z. Jurkowska, X. Zhang, A. Jeltsch, and X. Cheng, "Structure of dnmt3a bound to dnmt3l suggests a model for de novo dna methylation," *Nature*, vol. 449, no. 7159, pp. 248–251, 2007. 12

[51] V. Handa and A. Jeltsch, "Profound flanking sequence preference of dnmt3a and dnmt3b mammalian dna methyltransferases shape the human epigenome," *Journal of molecular biology*, vol. 348, no. 5, pp. 1103–1112, 2005. 12

[52] E. R. Doughtery, H. Jianping, and M. L. Bittner, "Validation of computational methods in genomics," *Current Genomics*, vol. 8, no. 1, p. 1, 2007. 12, 26, 83

[53] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004. 12, 83

[54] M. R. Segal, "Validation in genomics: Cpg island methylation revisited," *Statistical applications in genetics and molecular biology*, vol. 5, no. 1, p. Article 29, 2006. 12

[55] E. Schilling and M. Rehli, "Global, comparative analysis of tissue-specific promoter cpg methylation," *Genomics*, vol. 90, no. 3, pp. 314 – 323, 2007. 12

[56] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck, "Dna methylation profiling of human chromosomes 6, 20 and 22," *Nat Genet*, vol. 38, pp. 1378–1385, Dec. 2006. 13, 16, 17, 41, 77, 80, 81, 150

[57] D. Zilberman and S. Henikoff, "Genome-wide analysis of dna methylation patterns," *Development*, vol. 134, no. 22, pp. 3959–3965, 2007. 13

[58] P. Dehan, G. Kustermans, S. Guenin, J. Horion, J. Boniver, and P. Delvenne, "Dna methylation and cancer diagnosis: new methods and applications," *Expert Rev. Mol. Diagn.*, vol. 9(7), pp. 651–657, 2009. 13

[59] K. L. Thu, L. A. Pikor, J. Y. Kennett, C. E. Alvarez, and W. L. Lam, "Methylation analysis by dna immunoprecipitation," *Journal of cellular physiology*, vol. 222, no. 3, pp. 522–531, 2010. 13

[60] S. N. Austad, "Why women live longer than men: sex differences in longevity," *Gender medicine*, vol. 3, no. 2, pp. 79–92, 2006. 13, 23, 63, 126

[61] A. H. Field, *cluster and classification techniques for biosciences.* Cambridge University Press., 2007. 14, 24

[62] J. R. Goñi, C. Fenollosa, A. Pérez, D. Torrents, and M. Orozco, "Dnalive: a tool for the physical analysis of dna at the genomic scale," *Bioinformatics*, vol. 24, no. 15, pp. 1731–1732, 2008. 15, 82

[63] A. F. Fernandez, Y. Assenov, J. I. Martin-Subero, B. Balint, R. Siebert, H. Taniguchi, H. Yamamoto, M. Hidalgo, A.-C. Tan, O. Galm, *et al.*, "A dna methylation fingerprint of 1628 human samples," *Genome research*, vol. 22, no. 2, pp. 407–419, 2012. 15, 16, 42, 59, 63, 72, 77, 80, 81, 127, 137, 147

[64] C. Previti, O. Harari, I. Zwir, and C. del Val, "Profile analysis and prediction of tissue-specific cpg island methylation classes," *BMC Bioinformatics*, vol. 10, no. 1, p. 116, 2009. 16, 45, 59, 64, 80, 81, 84, 120, 123, 126, 150, 152

[65] F. Fang, S. Fan, X. Zhang, and M. Q. Zhang, "Predicting methylation status of cpg islands in the human brain," *Bioinformatics*, vol. 22, no. 18, pp. 2204–2209, 2006. 16

[66] C. Bock, J. Walter, M. Paulsen, and T. Lengauer, "Cpg island mapping by epigenome prediction," *PLoS Comput Biol*, vol. 3, no. 6, p. e110, 2007. 16, 26

[67] J. R. Goñi, A. Pérez, D. Torrents, and M. Orozco, "Determining promoter location based on dna structure first-principles calculations," *Genome Biol*, vol. 8, no. 12, p. R263, 2007. 16

[68] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15(8), pp. 1034–1050, 2005. 16

[69] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker, "Human dna methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, pp. 315–322, Nov. 2009. 16

[70] R. Jaenisch and A. Bird, "Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals," *Nature genetics*, vol. 33, pp. 245–254, 2003. 17

[71] V. K. Rakyan, T. A. Down, N. P. Thorne, P. Flicek, E. Kulesha, S. Gräf, E. M. Tomazou, L. Bäckdahl, N. Johnson, M. Herberth, *et al.*, "An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tdmrs)," *Genome research*, vol. 18, no. 9, pp. 1518–1529, 2008. 17, 77

[72] R. Illingworth, A. Kerr, D. DeSousa, H. Jrgensen, P. Ellis, J. Stalker, D. Jackson, C. Clee, R. Plumb, J. Rogers, S. Humphray, T. Cox, C. Langford, and A. Bird, "A novel cpg island set identifies tissue-specific methylation at developmental gene loci," *PLoS Bio*, vol. 6, p. e22, 01 2008. 17, 77, 80

[73] A. P. Bird, "Cpg-rich islands and the function of dna methylation," *Nature*, vol. 321, pp. 209–213, May 1986. 17

[74] R. A. Irizarry, C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan, and A. P. Feinberg, "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific cpg island shores," *Nat Genet*, vol. 41, pp. 178–186, Feb. 2009. 17

[75] B. C. Christensen, E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, J. L. Wiemels, H. H. Nelson, M. R. Karagas, J. F. Padbury, R. Bueno, D. J. Sugarbaker, R.-F. Yeh, J. K. Wiencke, and K. T. Kelsey, "Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context," *PLoS Genet*, vol. 5, p. e1000602, 08 2009. 17

[76] H. G. project, "Human genome project (accessed 13-03-2009)," accessed 13-03-2009 2009. 17, 18, 19, 20, 21, 22

[77] P. H. Kim S., Li M. and N. K, "Predicting dna methylation susceptibility using cpg flaking sequence," *Pacific symposium on biocomputing*, vol. 13, pp. 315–326, 2008. 18

[78] P. C. Champ, S. Maurice, J. M. Vargason, T. Camp, and P. S. Ho, "Distributions of z-dna and nuclear factor i in human chromosome 22: a model for coupled transcriptional regulation," *Nucleic acids research*, vol. 32, no. 22, pp. 6501–6510, 2004. 19

[79] A. Travers, "The structural basis of dna flexibility," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 362, no. 1820, pp. 1423–1438, 2004. 19, 20

[80] A. Leslie, S. Arnott, R. Chandrasekaran, and R. Ratliff, "Polymorphism of dna double helices," *Journal of molecular biology*, vol. 143, no. 1, pp. 49–72, 1980. 19

[81] M. C. Wahl and M. Sundaralingam, "Crystal structures of a-dna duplexes," *Biopolymers*, vol. 44, no. 1, pp. 45–63, 1997. 19

[82] E. J. Gardiner, C. A. Hunter, M. J. Packer, D. S. Palmer, and P. Willett, "Sequence-dependent dna structure: a database of octamer structural parameters," *Journal of molecular biology*, vol. 332, no. 5, pp. 1025–1035, 2003. 19, 20

[83] W. Li and P. Miramontes, "Large-scale oscillation of structure-related dna sequence features in human chromosome 21," *Phys. Rev. E*, vol. 74, p. 021912, Aug 2006. 19

[84] P. Baldi, S. Brunak, Y. Chauvin, and A. Krogh, "Naturally occurring nucleosome positioning signals in human exons and introns," *Journal of molecular biology*, vol. 263, no. 4, pp. 503–510, 1996. 20

[85] I. Brukner, R. Sanchez, D. Suck, and S. Pongor, "Sequence-dependent bending propensity of dna as revealed by dnase i: parameters for trinucleotides.," *The EMBO journal*, vol. 14, no. 8, p. 1812, 1995. 20

[86] J. Widom, "Short range order of in two eukaryotic genomes: relation to chromosome structure.," *J. Mol. Biol.*, vol. 259, pp. 579–588., 1996. 20

[87] K. Downes, B. J. Barratt, P. Akan, S. J. Bumpstead, S. D. Taylor, D. G. Clayton, and P. Deloukas, "Snp allele frequency estimation in dna pools and variance components analysis," *Biotechniques*, vol. 36, no. 5, pp. 840–845, 2004. 22

[88] S. Agrawal, M. Unterberg, S. Koschmieder, U. zur Stadt, U. Brunnberg, W. Verbeek, T. Büchner, W. E. Berdel, H. Serve, and C. Müller-Tidow, "Dna methylation of tumor suppressor genes in clinical remission predicts the relapse risk in acute myeloid leukemia," *Cancer research*, vol. 67, no. 3, pp. 1370–1377, 2007. 23

[89] R. S. Alisch, B. G. Barwick, P. Chopra, L. K. Myrick, G. A. Satten, K. N. Conneely, and S. T. Warren, "Age-associated dna methylation in pediatric populations," *Genome research*, vol. 22, no. 4, pp. 623–632, 2012. 23, 63, 126, 137

[90] M. F. Fraga and M. Esteller, "Epigenetics and aging: the targets and the marks," *Trends in Genetics*, vol. 23, no. 8, pp. 413 – 418, 2007. 23, 63, 126

[91] M. F. Fraga, E. Ballestar, M. F. Paz, S. Ropero, F. Setien, M. L. Ballestar, D. Heine-Suñer, J. C. Cigudosa, M. Urioste, J. Benitez, *et al.*, "Epigenetic differences arise during the lifetime of monozygotic twins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 30, pp. 10604–10609, 2005. 23, 63, 64, 126

[92] D. Dolinoy, "The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome.," *Nutrition Reviews,*, vol. Suppl 1, pp. 7–11, 2008. 23, 64, 126

[93] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, vol. 17 suppl 1, pp. S22–S29, 2001. 25

[94] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278(5338), pp. 680–686, 1997. 25

[95] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530–536, Jan. 2002. 25, 66

[96] C. Ambroise and J. M. Geoffrey, "Selection bias in gene extraction on the basis of microarray gene-expression data.," *Pnas.*, vol. 99 (10), pp. 6562–6566, 2002. 25, 30, 57

[97] M. Hackenberg, C. Previti, P. Luque-Escamilla, P. Carpena, J. Martinez-Aroza, and J. Oliver, "Cpgcluster: a distance-based algorithm for cpg-island detection," *BMC Bioinformatics*, vol. 7, no. 1, p. 446, 2006. 25, 82

[98] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95(25), pp. 14863–14868, 1998,. 25, 66

[99] K. Marchal, F. De Smet, K. Engelen, and B. De Moor, "Computational biology and toxicogenomics," *Predictive toxicology*, pp. 37–85, 2004. 25, 59

[100] Y. Kim, W. N. Street, and F. Menczer, "Evolutionary model selection in unsupervised learning.," *Intelligent data analysis*, vol. 6, no. 6, pp. 531–556, 2002. 25, 59

[101] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999. 25, 59

[102] B. L. Ilana, "A generalized clustering problem, with application to dna microarrays," *Statistical Applications in Genetics and Molecular Biology*, vol. 5, pp. 1–24, 2006. 25, 59

[103] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. 26

[104] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13(1), pp. 21–27, 1967. 26, 47

[105] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984. 26, 44

[106] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997. 26, 27, 28, 52, 53, 54

[107] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005. 26

[108] H. Zheng, H. Wu, J. Li, and S.-W. Jiang, "Cpgimethpred: computational model for predicting methylation status of cpg islands in human genome," *BMC medical genomics*, vol. 6, no. Suppl 1, p. S13, 2013. 26

[109] M. Bhasin and G. Raghava, "Pcleavage: an svm based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences," *Nucleic acids research*, vol. 33, no. suppl 2, pp. W202–W207, 2005. 26

[110] L. Lu, K. Lin, Z. Qian, H. Li, Y. Cai, and Y. Li, "Predicting dna methylation status using word composition," *Journal of Biomedical Science and Engineering*, vol. 3, no. 07, p. 672, 2010. 26

[111] I. I. Saeys Y and Larra, "A review of feature selection techniques in bioinformatics.," *Bioinformatics*, vol. 23 (19), pp. 2507–2517, 2007. 26, 28, 57, 83, 127

[112] Y. Tominaga, "Comparative study of class data analysis with pca-lda, simca, pls, anns, and k-nn," *Chemometrics and Intelligent Laboratory Systems*, vol. 49, p. 105115, 1999. 26

[113] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning.," *Artificial Intelligence*, vol. 97, pp. 245–271, 1997. 26, 28, 127

[114] S. J. Dixon and R. G. Brereton, "Comparison of performance of five common classifiers represented as boundary methods: Euclidean distance to centroids, linear discriminant analysis, quadratic discriminant analysis, learning vector quantization and support vector machines, as dependent on data structure," *Chemometrics and Intelligent Laboratory Systems*, vol. 95, no. 1, pp. 1–17, 2009. 27, 49

[115] H. Trevor, T. Robert, and F. Jerome, "The elements of statistical learning: data mining, inference and prediction," *New York: Springer-Verlag*, vol. 1, no. 8, pp. 371–406, 2001. 27

[116] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Ann. Statist.*, vol. 26, pp. 1651–1686, 10 1998. 27, 28, 51, 52, 53, 110, 111

[117] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: Improving classification performance when training data is skewed," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, IEEE, 2008. 27, 31, 110

[118] J. Friedman, T. Hastie, R. Tibshirani, *et al.*, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000. 28, 51, 52, 53, 113, 114

[119] L. Breiman *et al.*, "Arcing classifier (with discussion and a rejoinder by the author)," *The annals of statistics*, vol. 26, no. 3, pp. 801–849, 1998. 28, 49, 52

[120] S. Degroeve, Y. Saeys, B. De Baets, P. Rouzé, and Y. Van De Peer, "Splicemachine: predicting splice sites from high-dimensional local context representations," *Bioinformatics*, vol. 21, no. 8, pp. 1332–1338, 2005. 29, 30

[121] B. Yoshua and G. Yves., "No unbiased estimator of the variance of k-fold cross-validation.," *Journal of machine learning research.*, vol. 5, pp. 1089–1105, 2004. 29

[122] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007. 31, 34

[123] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *SIGKDD Explor. Newsl.*, vol. 6(1), pp. 40–49, 2004. 31

[124] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009. 31, 110, 111

[125] B. Zadrozny, J. Langford, and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 435–442, IEEE, 2003. 31, 34, 110

[126] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. 32, 50, 111

[127] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Workshop on Learning from Imbalanced Datasets II, ICML , Washington DC, 2003*, 2003 pp., 1-8. 32

[128] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 2, pp. 539–550, 2009. 32, 110

[129] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992. 32, 110

[130] J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C. E. Brodley, "Pruning decision trees with misclassification costs," in *Machine Learning: ECML-98*, pp. 131–136, Springer, 1998. 32

[131] B. Zadrozny and C. Elkan., "Learning and making decisiions when costs and probabilities are both unknown.," in *In Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, pages 204-213*, San Francisco, CA, August 2001. 32, 33

[132] Y. Lin, Y. Lee, and G. Wahba, "Support vector machines for classification in nonstandard situations," *Machine learning*, vol. 46, no. 1-3, pp. 191–202, 2002. 32

[133] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, pp. 973–978, Citeseer, 2001. 33

[134] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004. 33

[135] D. Foley, "Considerations of sample and feature size," *Information Theory, IEEE Transactions on*, vol. 18(5), pp. 618–626, 1972. 34

[136] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010. 35, 42, 43

[137] A. H. Fielding and J. F. Bell, "A review of methods for the assessment of prediction errors in conservation presence/absence models," *Environmental conservation*, vol. 24, no. 01, pp. 38–49, 1997. 36

[138] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679 – 688, 2006. 36

[139] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000. 37, 44, 84, 85

[140] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochim. Biophys. Acta.*, vol. 405, pp. 442–451, 1975. 37, 85

[141] D. Lewis and W. Gale., "Training text classiers by uncertainty sampling.," in *In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information, pages 73-79,*, New York, NY, August 1998. 38

[142] P.-N. Tan, M. Steinbach, V. Kumar, *et al.*, *Introduction to data mining*, vol. 1. Pearson Addison Wesley Boston, 2006. 38

[143] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning*, vol. 30, no. 2-3, pp. 195–215, 1998. 38

[144] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974. 42, 43

[145] S. Geisser, "The predictive sample reuse method with applications.," *Journal of the American Statistical Association,*, vol. 70(350),, pp. 320–328., (1975). 42, 43, 44

[146] A. Celisse and S. Robin., "Nonparametric density estimation by exact leave-p-out cross-validation.," *Computational Statistics and Data Analysis*, vol. 52(5), pp. 2350–2368, 2008. 42

[147] P. S. Leo Breiman, "Submodel selection and evaluation in regression. the x-random case," *International Statistical Review / Revue Internationale de Statistique*, vol. 60, no. 3, pp. 291–319, 1992. 42

[148] E. C. Burman, Prabir and D. Nolan., "A cross-validatory method for dependent data," *Biometrika*, vol. 81(2), pp. 351–358., 1994. 42

[149] J. Shao., "Linear model selection by cross-validation.," *Journal of the American Statistical Association*, vol. 88(422), pp. 486–494., 1993. 42

[150] M. Kearns, "A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split.," *Neural Computation*, vol. 9(5), pp. 1143–1161., 1997. 42

[151] P. Blanchard, Gilles; Massart, "Discussion: Local rademacher complexities and oracle inequalities in risk minimization.," *Ann. Statist.*, vol. 34(6), pp. 2664–2671, 2006. 42

[152] P. Zhang, "Model selection via multifold cross validation," *The Annals of Statistics*, vol. 21, pp. 299–313, 1993. 42

[153] K.-C. Li, "Asymptotic optimality for cp,cl, cross-validation and generalized cross-validation: Discrete index set.," *Ann. Statist. 5.*, vol. 15(3), pp. 958–975, 1987. 44

[154] M. Hellman, "The nearest neighbor classification rule with a reject option," *Systems Science and Cybernetics, IEEE Transactions on*, vol. 6(3), pp. 179–185, 1970. 47

[155] D. Fukunaga, Keinosuke; Hummels, "Bayes error estimation using parzen and k-nn procedures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-9(5), pp. 634–643, 1987. 47

[156] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule,," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. vol.SMC-6, no.4, pp. 325–327, 1976. 47

[157] T. Bailey and A. Jain, "A note on distance-weighted k-nearest neighbor rules," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 4, pp. 311–313, 1978. 47

[158] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989. 48

[159] W. Wu, Y. Mallet, B. Walczak, W. Penninckx, D. Massart, S. Heuerding, and F. Erni, "Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to nir data," *Analytica Chimica Acta*, vol. 329, no. 3, pp. 257–265, 1996. 48

[160] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014. 49

[161] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002. 50

[162] N. M. Wanas, R. A. Dara, and M. S. Kamel, "Adaptive fusion and co-operative training for classifier ensembles," *Pattern Recognition*, vol. 39, no. 9, pp. 1781–1794, 2006. 50

[163] L. Rokach, "Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography," *Computational Statistics & Data Analysis*, vol. 53, no. 12, pp. 4046–4072, 2009. 50

[164] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 3, pp. 226–239, 1998. 50

[165] L. I. Kuncheva, J. C. Bezdek, and R. P. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern recognition*, vol. 34, no. 2, pp. 299–314, 2001. 50

[166] N. Wanas, *Feature based architecture for decision fusion*. PhD thesis, University of Waterloo, 2003. 50

[167] R. Kohavi and G. H. John, "Wrappers feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–3224, 1997. 57, 127

[168] D. KOLLER, "Toward optimal feature selection," *Proc. 13th International Conference on Machine Learning*, pp. 284–292, 1996. 58

[169] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 7, pp. 900–912, 2004. 58

[170] P. H. Sneath, R. R. Sokal, *et al.*, *Numerical taxonomy. The principles and practice of numerical classification*. No. ISBN-10: 0716706970, W.H. Freeman AND Company, first edition edition ed., 1973. 59

[171] B. King, "Step-wise clustering procedures," *Journal of the American Statistical Association*, vol. 62, no. 317, pp. 86–101, 1967. 59

[172] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 515–524, ACM, 2002. 59

[173] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963. 59

[174] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 26, no. 4, pp. 354–359, 1983. 59

[175] G. Nagy, *State of the art in pattern recognition*, vol. 56. IEEE, 1968. 60

[176] B. C. Christensen, E. A. Houseman, C. J. Marsit, S. Zheng, M. R. Wrensch, J. L. Wiemels, H. H. Nelson, M. R. Karagas, J. F. Padbury, R. Bueno, *et al.*, "Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context," *PLoS Genet*, vol. 5, no. 8, p. e1000602, 2009. 63, 69, 77, 78

[177] C. J. Marsit, E. A. Houseman, A. R. Schned, M. R. Karagas, and K. T. Kelsey, "Promoter hypermethylation is associated with current smoking, age, gender

and survival in bladder cancer," *Carcinogenesis*, vol. 28, no. 8, pp. 1745–1751, 2007. 63, 77

[178] J. Vijg and J. Campisi, "Puzzles, promises and a cure for ageing," *Nature*, vol. 454, no. 7208, p. 1065, 2008. 64

[179] A. F. Fernandez, "A dna methylation fingerprint of 1,628 human samples." http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28094, 22 March 2011. Accessed 1 May 2013. 65, 127

[180] L. Shen, Y. Kondo, G. L. Rosner, L. Xiao, N. S. Hernandez, J. Vilaythong, P. S. Houlihan, R. S. Krouse, A. R. Prasad, J. G. Einspahr, *et al.*, "Mgmt promoter methylation and field defect in sporadic colorectal cancer," *Journal of the National Cancer Institute*, vol. 97, no. 18, pp. 1330–1338, 2005. 69, 77

[181] J.-P. Issa, "Age-related epigenetic changes and the immune system," *Clinical Immunology*, vol. 109, no. 1, pp. 103–108, 2003. 69

[182] B. C. Christensen, J. J. Godleski, C. J. Marsit, E. Houseman, C. Y. Lopez-Fagundo, J. L. Longacker, R. Bueno, D. J. Sugarbaker, H. H. Nelson, and K. T. Kelsey, "Asbestos exposure predicts cell cycle control gene promoter methylation in pleural mesothelioma," *Carcinogenesis*, vol. 29, no. 8, pp. 1555–1559, 2008. 77

[183] J.-P. J. Issa, Y. L. Ottaviano, P. Celano, S. R. Hamilton, N. E. Davidson, and S. B. Baylin, "Methylation of the oestrogen receptor cpg island links ageing and neoplasia in human colon," *Nature genetics*, vol. 7, no. 4, pp. 536–540, 1994. 77

[184] B. Richardson, "Impact of aging on dna methylation," *Ageing research reviews*, vol. 2, no. 3, pp. 245–261, 2003. 77

[185] B. Kwabi-Addo, W. Chung, L. Shen, M. Ittmann, T. Wheeler, J. Jelinek, and J.-P. J. Issa, "Age-related dna methylation changes in normal human prostate tissues," *Clinical cancer research*, vol. 13, no. 13, pp. 3796–3802, 2007. 77

[186] J. Tra, T. Kondo, Q. Lu, R. Kuick, S. Hanash, and B. Richardson, "Infrequent occurrence of age-dependent changes in cpg island methylation as detected by restriction landmark genome scanning," *Mechanisms of ageing and development*, vol. 123, no. 11, pp. 1487–1503, 2002. 77, 78

[187] F. Lienert, C. Wirbelauer, I. Som, A. Dean, F. Mohn, and D. Schübeler, "Identification of genetic elements that autonomously determine dna methylation states," *Nature genetics*, vol. 43, no. 11, pp. 1091–1097, 2011. 80

[188] A. Bird., "Putting the dna back into dna methylation.," *Nat. genet.*, vol. 43 (11), p. 10501051, 2011. 80

[189] M. Weber, J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schuebeler, "Chromosome-wide and promoter-specific analyses identify sites of differential dna methylation in normal and transformed human cells," *Nature genetics*, vol. 37, no. 8, pp. 853–862, 2005. 80

[190] A. M. Deaton, S. Webb, A. R. Kerr, R. S. Illingworth, J. Guy, R. Andrews, and A. Bird, "Cell typespecific dna methylation at intragenic cpg islands in the immune system," *Genome Research*, vol. 21, no. 7, pp. 1074–1086, 2011. 80

[191] R. J. Weeks and I. M. Morison, "Detailed methylation analysis of cpg islands on human chromosome region 9p21," *Genes, Chromosomes and Cancer*, vol. 45, no. 4, pp. 357–364, 2006. 80

[192] S. Veerla, I. Panagopoulos, Y. Jin, D. Lindgren, and M. Höglund, "Promoter analysis of epigenetically controlled genes in bladder cancer," *Genes, Chromosomes and Cancer*, vol. 47, no. 5, pp. 368–378, 2008. 80

[193] F. Simmer, A. B. Brinkman, Y. Assenov, F. Matarese, A. Kaan, L. Sabatino, A. Villanueva, D. Huertas, M. Esteller, T. Lengauer, *et al.*, "Comparative genome-wide dna methylation analysis of colorectal tumor and matched normal tissues," *Epigenetics*, vol. 7, no. 12, pp. 1355–1367, 2012. 87, 88, 106

[194] K. S. J. Ohgane, S. Yagi, "Epigenetics: The dna methylation profile of tissue-dependent and differentially methylated regions in cells," *Placenta*, vol. 29, pp. 29–35, 2008. 88

[195] E. A. Cortés, M. G. Martínez, and N. G. Rubio, "A boosting approach for corporate failure prediction," *Applied Intelligence*, vol. 27, no. 1, pp. 29–37, 2007. 110

[196] H. He, E. Garcia, *et al.*, "Learning from imbalanced data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 21, no. 9, pp. 1263–1284, 2009. 110

[197] G. H. Nguyen, A. Bouzerdoum, and S. L. Phung, *Learning pattern classification tasks with imbalanced data sets*. INTECH Open Access Publisher, 2009. 110

[198] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley*, 2004. 110

[199] M. Wasikowski and X.-w. Chen, "Combating the small sample class imbalance problem using feature selection," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1388–1400, 2010. 110

[200] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Computational Intelligence*, vol. 26, no. 3, pp. 232–257, 2010. 110

[201] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 40, no. 1, pp. 185–197, 2010. 111

[202] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 1, pp. 63–77, 2006. 113

[203] F. Liu, E. TAstesen, J. K. Sundet, T.-K. Jenssen, C. Bock, G. I. Jerstad, W. G. Thilly, and E. Hovig, "The human genomic melting map," *PLoS Comput Biol*, vol. 3, p. e93, 05 2007. 114

[204] H. T. Bjornsson, M. I. Sigurdsson, M. D. Fallin, R. A. Irizarry, T. Aspelund, H. Cui, W. Yu, M. A. Rongione, T. J. Ekström, T. B. Harris, *et al.*, "Intra-individual change over time in dna methylation with familial clustering," *Jama*, vol. 299, no. 24, pp. 2877–2883, 2008. 126, 135

[205] S. Sinha, D. Thomas, L. Yu, A. J. Gentles, N. Jung, M. R. Corces-Zimmerman, S. M. Chan, A. Reinisch, A. P. Feinberg, D. L. Dill, *et al.*, "Mutant wt1 is associated with dna hypermethylation of prc2 targets in aml and responds to ezh2 inhibition," *Blood*, vol. 125, no. 2, pp. 316–326, 2015. 126

[206] S. Mani, K. Szymańska, C. Cuenin, D. Zaridze, K. Balassiano, S. C. Lima, E. Matos, A. Daudt, S. Koifman, V. W. Filho, *et al.*, "Dna methylation changes associated with risk factors in tumors of the upper aerodigestive tract," *Epigenetics*, vol. 7, no. 3, pp. 270–277, 2012. 126

[207] H. Huang, Z. Chen, and X. Huang, "Age-adjusted nonparametric detection of differential dna methylation with case–control designs," *BMC bioinformatics*, vol. 14, no. 1, p. 86, 2013. 126

[208] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, *et al.*, "Genome-wide methylation profiles reveal quantitative views of human aging rates," *Molecular cell*, vol. 49, no. 2, pp. 359–367, 2013. 126

[209] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005. 127

[210] M. Bibikova, E. Chudin, B. Wu, L. Zhou, E. W. Garcia, Y. Liu, S. Shin, T. W. Plaia, J. M. Auerbach, D. E. Arking, *et al.*, "Human embryonic stem cells have

a unique epigenetic signature," *Genome research*, vol. 16, no. 9, pp. 1075–1083, 2006. 129, 144

[211] F. Rassendren, G. N. Buell, C. Virginio, G. Collo, R. A. North, and A. Surprenant, "The permeabilizing atp receptor, p2x7 cloning and expression of a human cdna," *Journal of Biological Chemistry*, vol. 272, no. 9, pp. 5482–5486, 1997. 132

[212] M. A. Degli-Esposti, W. C. Dougall, P. J. Smolak, J. Y. Waugh, C. A. Smith, and R. G. Goodwin, "The novel receptor trail-r4 induces nf-κb and protects against trail-mediated apoptosis, yet retains an incomplete death domain," *Immunity*, vol. 7, no. 6, pp. 813–820, 1997. 132

[213] R. Janknecht and A. Nordheim, "Elk-1 protein domains required for direct and srf-assisted dna-binding," *Nucleic acids research*, vol. 20, no. 13, pp. 3317–3324, 1992. 135

[214] M. Hemberger, H. Himmelbauer, H. P. Neumann, K. H. Plate, G. Schwarzkopf, and R. Fundele, "Expression of the von hippel-lindau-binding protein-1 (vbp1) in fetal and adult mouse tissues," *Human molecular genetics*, vol. 8, no. 2, pp. 229–236, 1999. 136, 143

[215] C. M. Shanahan, D. Proudfoot, A. Farzaneh-Far, and P. L. Weissberg, "The role of gla proteins in vascular calcification," *Critical Reviews in Eukaryotic Gene Expression*, vol. 8, no. 3-4, 1998. 140

[216] K. Miyamoto, K. Asada, T. Fukutomi, E. Okochi, Y. Yagi, T. Hasegawa, T. Asahara, T. Sugimura, and T. Ushijima, "Methylation-associated silencing of heparan sulfate d-glucosaminyl 3-o-sulfotransferase-2 (3-ost-2) in human breast, colon, lung and pancreatic cancers," *Oncogene*, vol. 22, no. 2, pp. 274–280, 2003. 141, 145, 147

[217] P.-O. Estève, Y. Chang, M. Samaranayake, A. K. Upadhyay, J. R. Horton, G. R. Feehery, X. Cheng, and S. Pradhan, "A methylation and phosphorylation switch between an adjacent lysine and serine determines human dnmt1 stability," *Nature structural & molecular biology*, vol. 18, no. 1, pp. 42–48, 2011. 142, 145

[218] D. Sakurai, J. Zhao, Y. Deng, J. A. Kelly, E. E. Brown, J. B. Harley, S.-C. Bae, M. E. Alarcón-Riquelme, J. C. Edberg, R. P. Kimberly, *et al.*, "Preferential binding to elk-1 by sle-associated il10 risk allele upregulates il10 expression," *PLoS Genet*, vol. 9, p. e1003870, 2013. 142, 147

[219] C. Seemayer, S. Kuchen, P. Kuenzler, V. Rihosková, J. Schedel, M. Neidhart, B. Michel, R. Gay, and S. Gay, "Expression of focal adhesion kinase, akt/pkb,

elk-1 and p90rsk in rheumatoid arthritis tissues but not at sites of cartilage invasion," *Arthritis Res Ther*, vol. 5, no. Suppl 3, pp. 1–2, 2003. 142

[220] E. W. Tobi, L. Lumey, R. P. Talens, D. Kremer, H. Putter, A. D. Stein, P. E. Slagboom, and B. T. Heijmans, "Dna methylation differences after exposure to prenatal famine are common and timing-and sex-specific," *Human molecular genetics*, vol. 18, no. 21, pp. 4046–4053, 2009. 142

[221] G. Malferrari, U. Mazza, C. Tresoldi, E. Rovida, M. Nissim, M. Mirabella, S. Servidei, and I. Biunno, "Molecular characterization of a novel endonuclease (xib) and possible involvement in lysosomal glycogen storage disorders," *Experimental and molecular pathology*, vol. 66, no. 2, pp. 123–130, 1999. 142

[222] Y. Zhang, M. Zhao, A. H. Sawalha, B. Richardson, and Q. Lu, "Impaired dna methylation and its mechanisms in cd4+ t cells of systemic lupus erythematosus," *Journal of autoimmunity*, vol. 41, pp. 92–99, 2013. 142

[223] M. Napirei, H. Karsunky, B. Zevnik, H. Stephan, H. G. Mannherz, and T. Moroy, "Features of systemic lupus erythematosus in dnase1-deficient mice," *Nat Genet*, vol. 25, pp. 177–181, June 2000. 142

[224] K. Yasutomo, T. Horiuchi, S. Kagami, H. Tsukamoto, C. Hashimura, M. Urushihara, and Y. Kuroda, "Mutation of dnase1 in people with systemic lupus erythematosus," *Nat Genet*, vol. 28, pp. 313–314, Aug. 2001. 142

[225] E. Balada, J. Ordi-Ros, S. Hernanz, J. Villarreal, F. Corts, M. Vilardell-Tarrs, and M. Labrador, "Dnase i mutation and systemic lupus erythematosus in a spanish population: Comment on the article by tew et al," *Arthritis & Rheumatism*, vol. 46(7), no. 7, pp. 1974–1976, 2002. 142

[226] M. P. Boks, E. M. Derks, D. J. Weisenberger, E. Strengman, E. Janson, I. E. Sommer, R. S. Kahn, R. A. Ophoff, *et al.*, "The relationship of dna methylation with age, gender and genotype in twins and healthy controls," *PloS one*, vol. 4, no. 8, p. e6767, 2009. 142, 144

[227] O. El-Maarri, T. Becker, J. Junen, S. S. Manzoor, A. Diaz-Lacava, R. Schwaab, T. Wienker, and J. Oldenburg, "Gender specific differences in levels of dna methylation at selected loci from human total blood: a tendency toward higher methylation levels in males," *Human genetics*, vol. 122, no. 5, pp. 505–514, 2007. 142

[228] P. Tarpey, J. Parnau, M. Blow, H. Woffendin, G. Bignell, C. Cox, J. Cox, H. Davies, S. Edkins, S. Holden, *et al.*, "Mutations in the dlg3 gene cause nonsyndromic x-linked mental retardation," *The American Journal of Human Genetics*, vol. 75, no. 2, pp. 318–324, 2004. 143

[229] S. Kwon, Y. Zhang, and P. Matthias, "The deacetylase hdac6 is a novel critical component of stress granules involved in the stress response," *Genes & development*, vol. 21, no. 24, pp. 3381–3394, 2007. 143

[230] J. M. Serrador, J. R. Cabrero, D. Sancho, M. Mittelbrunn, A. Urzainqui, and F. Sánchez-Madrid, "Hdac6 deacetylase activity links the tubulin cytoskeleton with immune synapse organization," *Immunity*, vol. 20, no. 4, pp. 417–428, 2004. 143

[231] F. Iqbal, C. B. Item, R. Ratschmann, M. Ali, E. Plas, and O. Bodamer, "Molecular analysis of guanidinoacetate-n-methyltransferase (gamt) and creatine transporter (slc6a8) gene by using denaturing high pressure liquid chromatography (dhplc) as a possible source of human male infertility," *Pak. J. Pharm. Sci*, vol. 24, no. 1, pp. 75–79, 2011. 144

[232] C. Grunau, W. Hindermann, and A. Rosenthal, "Large-scale methylation analysis of human genomic dna reveals tissue-specific differences between the methylation profiles of genes and pseudogenes," *Human molecular genetics*, vol. 9, no. 18, pp. 2651–2663, 2000. 144

[233] S. R. Twigg, K. Matsumoto, A. M. Kidd, A. Goriely, I. B. Taylor, R. B. Fisher, A. J. M. Hoogeboom, I. M. Mathijssen, M. T. Lourenço, J. E. Morton, *et al.*, "The origin of efnb1 mutations in craniofrontonasal syndrome: frequent somatic mosaicism and explanation of the paucity of carrier males," *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 999–1010, 2006. 144

[234] M. B. Tew, F. C. Arnett, J. D. Reveille, and F. K. Tan, "Mutations of bone morphogenetic protein receptor type ii are not found in patients with pulmonary hypertension and underlying connective tissue diseases," *Arthritis & Rheumatism*, vol. 46, no. 10, pp. 2829–2830, 2002. 144

[235] H. S. Tapp, D. M. Commane, D. M. Bradburn, R. Arasaradnam, J. C. Mathers, I. T. Johnson, and N. J. Belshaw, "Nutritional factors and gender influence age-related dna methylation in the human rectal mucosa," *Aging cell*, vol. 12, pp. 148–155, February 2013. 145

[236] S.-H. Yang and A. D. Sharrocks, "Piasxα differentially regulates the amplitudes of transcriptional responses following activation of the erk and p38 mapk pathways," *Molecular cell*, vol. 22, no. 4, pp. 477–487, 2006. 145

[237] M. Pradhan, P.-O. Esteve, H. G. Chin, M. Samaranayke, G.-D. Kim, and S. Pradhan, "Cxxc domain of human dnmt1 is essential for enzymatic activity," *Biochemistry*, vol. 47, no. 38, pp. 10000–10009, 2008. 145

[238] R. Yuen and W. Robinson, "Review: a high capacity of the human placenta for genetic and epigenetic variation: implications for assessing pregnancy outcome," *Placenta*, vol. 32, pp. S136–S141, 2011. 145

[239] P. W. Ang, M. Loh, N. Liem, P. L. Lim, F. Grieu, A. Vaithilingam, C. Platell, W. P. Yong, B. Iacopetta, and R. Soong, "Comprehensive profiling of dna methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features," *BMC cancer*, vol. 10, no. 1, p. 227, 2010. 145

[240] G. F. Combs, M. I. Jackson, J. C. Watts, L. K. Johnson, H. Zeng, J. Idso, L. Schomburg, A. Hoeg, C. S. Hoefig, E. C. Chiang, *et al.*, "Differential responses to selenomethionine supplementation by sex and genotype in healthy adults," *British journal of nutrition*, vol. 107, no. 10, pp. 1514–1525, 2012. 145

[241] J. D. Watson, F. H. Crick, *et al.*, "Molecular structure of nucleic acids," *Nature*, vol. 171, no. 4356, pp. 737–738, 1953. 150

[242] R. S. Illingworth, U. Gruenewald-Schneider, S. Webb, A. R. W. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews, and A. P. Bird, "Orphan cpg islands identify numerous conserved promoters in the mammalian genome," *PLoS Genet*, vol. 6, p. e1001134, 09 2010. 150

Appdx A

- I. Ali and H. Seker. Detailed methylation prediction of CpG islands on human chromosome 21. Proceedings of the 10th WSEAS International Conference on MATHEMATICS and COMPUTERS in BIOLOGY and CHEMISTRY. ISSN: 1790-5125, PP : 147-152, 2009.
- I. Ali and H. Seker. A comparative study for characterisation and prediction of tissue-specific DNA methylation of CpG islands in chromosomes 6, 20 and 22.32nd Annual International Conference of the IEEE EMBS. ISSN: 1557-170X , pp : 1832 1835, 2010.
- Ali I, Seker H. 2010. "An identification and prediction methods for feature-subsets of CpG islands methylation based on human peripheral blood leukocytes of chromosome 21q . Conf. Proc. IEEE BIBE 2010:289-290.

Papers due to submit

- Isse Ali, David Elizondo and Martin Grootveld. DNA methylation display on Aging and gender differences based on unsupervised clustering.(Chapter-4)
- Isse Ali and David Elizondo and Martin Grootveld. Prediction of methylation classes of cpg islands on chromosomes 6, 20, 21 and 22" (Chapter-5).
- Isse Ali, David Elizondoand Martin Grootveld, Weighting methods towards severely imbalanced data.(Chapter-5)
- Isse Ali, Martin Grootveld and David Elizondo, Analysis of gender differences in DNA methylation. (Chapter-6)

Published papers are deleted due to copy rights.