EXPLORING EPIGENETIC DRUG DISCOVERY USING COMPUTATIONAL APPROACHES

by

EMESE SOMOGYVARI

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Doctorate of Philosophy

Queen’s University
Kingston, Ontario, Canada
March 2019

Copyright © Emese Somogyvari, 2019
Abstract

The misregulation of epigenetic mechanisms has been linked to disease. Current drugs that treat these dysfunctions have had some success, however many have variable potency, instability in vivo and lack target specificity. This may be due to the limited knowledge of epigenetic mechanisms, especially at the molecular level, and their association with gene expression and its link to disease. Computational approaches, specifically in molecular modeling, have begun to address these issues by complementing phases of drug discovery and development, however more research is needed on the relationship between genetic mutation and epigenetics and their roles in disease. Gene regulatory network models have been used to better understand diseases, however, inferring these networks poses several challenges. To address some of these issues, a multi-label classification technique to infer regulatory networks (MInR), supplemented by semi-supervised, learning is presented. MInR’s performance was found to be comparable to other methods that infer regulatory networks when evaluated on a benchmark E.coli dataset. In order to better understand the association of epigenetics with gene expression and its link with disease, MInR was used to infer a regulatory network from a Kidney Renal Clear Cell Carcinoma (KIRC) dataset and was supplemented with gene expression and methylation analysis. Gene expression and methylation analysis revealed a correlation between 5 differentially methylated
CpGs and their matched differentially expressed transcripts. Incorporating this information into a network allowed for the visualization of potential systems that may be involved in KIRC. Future analysis of these systems alongside the drug discovery and development process may lead to the discovery of novel therapeutics.
Co-Authorship

This research was conducted by Emese Somogyvari, under the supervision of Dr. Selim G. Akl and Dr. Louise M. Winn.
Acknowledgments

I would like to thank my friends, family, and supervisors for their support and guidance throughout this project.
Statement of Originality

I hereby certify that all of the work described within this thesis is the original work of the author. Any published (or unpublished) ideas and/or techniques from the work of others are fully acknowledged in accordance with the standard referencing practices.
Chapter 3: Related Work

3.1 Current Computational Approaches in Epigenetic Drug Discovery

3.1.1 Virtual Screening
3.1.2 Molecular Dynamics
3.1.3 Molecular Docking
3.1.4 Homology Modeling
3.1.5 Pharmacophore Modeling
3.1.6 Additional Methods

3.2 Biological Networks and Machine Learning in Target Identification

Chapter 4: Genome-wide Statistical Analyses Reveal Correlation Between DNA Methylation and Gene Expression in Kidney Renal Clear Cell Carcinoma

4.1 Introduction
4.2 Methods
4.2.1 Data Collection
4.3 Results and Discussion
4.3.1 B3GNTL1 is among Metabolic Genes Identified as Differentially Expressed in KIRC
4.3.2 HCN3 was found to be Significantly Correlated to Lower Survival Rates in Kidney Cancer
4.3.3 Tumour Suppressor Gene Epigenetically Silenced through HOXB3 in a Variety of Cancers
4.3.4 RHPN1 found to be a Key Regulator of the Podocyte - A Promising Target for Kidney Disease
4.3.5 ZC3H12A Encodes Protein which Contributes to KIRC Development
4.4 Conclusion

Chapter 5: Multi-label Support Vector Machines to Infer Regulatory Networks

5.1 Introduction
5.2 Methods
5.2.1 Data and Data Preprocessing
5.2.2 Semi-supervised Learning to Infer Labels
5.2.3 Classification Modeling
5.2.4 Model Evaluation
5.3 Results
5.3.1 Model Performance
5.3.2 Regulatory Network
List of Tables

4.1 Differentially methylated and expressed genes, and PANTHER classifications ........................................... 39
4.2 Genes with correlated DNA methylation and gene expression ........................................... 43
List of Figures

2.1 Epigenetic modifications ........................................ 7
2.2 Bringing a new drug to market .................................. 10
2.3 Drug discovery and development ............................... 13
2.4 Simplified representation of SVM .............................. 16

5.1 Precision/recall curves of various models on E. coli benchmark .. 54
5.2 ROC curves of MInR on E. coli benchmark .................... 54
5.3 SIRENE performance curves .................................... 56
5.4 Inferred E. coli regulatory network. Part 1/2 ................... 57
5.5 Inferred E. coli regulatory network. Part 2/2 ................... 58

6.1 ROC curve of MInR on healthy control dataset ............... 66
6.2 Inferred KIRC regulatory network ............................. 67
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>ARACNE</td>
<td>Algorithm for the Reconstruction of Accurate Cellular Networks</td>
</tr>
<tr>
<td>ATA</td>
<td>Aurintricarboxylic Acid</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine Tri-Phosphate</td>
</tr>
<tr>
<td>AUC-ROC</td>
<td>Area Under the Receiver Operating Characteristic</td>
</tr>
<tr>
<td>AZA</td>
<td>5-Azacytidine</td>
</tr>
<tr>
<td>CGI</td>
<td>CpG Islands</td>
</tr>
<tr>
<td>CLR</td>
<td>Context of Likelihood of Relatedness</td>
</tr>
<tr>
<td>CpG</td>
<td>Cytosine/Guanine</td>
</tr>
<tr>
<td>DAC</td>
<td>5-Aza-2’Deoxycytidine, or Decitabine</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DNMTs</td>
<td>DNA Methyltransferases</td>
</tr>
<tr>
<td>ENCODE</td>
<td>The Encyclopedia of DNA Elements Consortium</td>
</tr>
</tbody>
</table>
**FDR** False Discovery Rate

**GO** Gene Ontology

**GRN** Genetic Regulatory Networks

**HATs** Histone Acetyltransferases

**HCN** Hyperpolarization-activated Cyclic Nucleotide-Gated

**HDACs** Histone Deacetylases

**HTS** High Throughput Screening

**IL7R** Interleukin Receptor 7

**IUPAC** International Union of Pure and Applied Chemistry

**KIRC** Kidney Renal Clear Cell Carcinoma

**MBDs** Methyl-CpG-Binding Domain Proteins

**MI** Mutual Information

**MInR** Multi-label classification to Infer Regulatory networks

**mRNA** Messenger RNA

**MRNET** Minimum Redundancy Networks

**NCBI** The National Centre for Biotechnology Information

**NCI** National Cancer Institute
OvR One-vs.-Rest

PANTHER Protein Analysis Through Evolutionary Relationships

PAZAR A Public Database of Transcription Factor and Regulatory Sequence Annotation

RBF Radial Basis Function

RCC Renal Cell Carcinoma

RPKM Reads Per Kilobase of transcript per Million mapped reads

SAH S-Adenosyl Homocysteine

SAHA Suberoylanilide Hydroxamic Acid

SAM S-Adenosyl Methionine

SAR Seasonal Allergic Rhinitis

SIRENE Supervised Inference of Regulatory Networks

SVM Support Vector Machines

TCGA The Cancer Genome Atlas

TF Transcription Factor

TRRUST Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining

TSS Transcription Start Site
Chapter 1

Introduction

1.1 Problem and Motivation

The term epigenetics describes the regulation of genomic functions leading to heritable changes in gene expression that are outside of the deoxyribonucleic acid (DNA) sequence, and is thought to be the link between environmental factors and gene expression. An epigenetic property can be defined as a cell property, which is mediated by genomic regulators, which allow a cell to remember a past event [61]. Epigenetic modifications include DNA methylation, histone modifications (acetylation, methylation, phosphorylation) and adenosine tri-phosphate (ATP)-dependent chromatin re-modeling. The misregulation of these components, regardless of the DNA sequence, has been shown to lead to an increased incidence in several diseases including Type II diabetes, cancer and Alzheimers, with a particular focus on DNA methylation and its link with cancer [12, 13, 92].

The development of drugs that treat these dysfunctions has been gaining interest in recent years due to the potential to reverse inappropriate epigenetic modifications [43]. Current epigenetic drugs have had some success, however many have variable
1.1. PROBLEM AND MOTIVATION

potency, instability \textit{in vivo} and lack target specificity. This may be due to the limited knowledge on epigenetic mechanisms, especially at the molecular level, which restricts the development and discovery of novel therapeutics and the optimization of existing drugs. Computational approaches, specifically in molecular modeling, have begun to address these issues by complementing phases of drug discovery and development.

Although these studies, which are outlined in Chapter 3, have contributed many successes to the field of epigenetic drug discovery, more research is needed on the link between genetics, epigenetics and disease; in many cases, the relationship between genetic mutations and epigenetics and their roles in the pathogenesis of a disease is not clear [85]. Such studies have become an area of particular interest because they may elucidate new epigenetic targets and provide novel scaffolds for the treatment of disease. Additionally, there is an increased amount of methylation and matched gene expression data, yet there exist few tools for their integrative analysis [55].

Integration of multiple -omics data such as transcriptomics (the study of transcriptomes - messenger RNA expressed by genes, and their function), metabolomics (the study of metabolites - molecules involved in metabolism), and proteomics (the study of proteins expressed by an organism, cells, tissues), has the potential to provide insight on the underlying architecture of complex diseases such as cancer, especially in a network context. Networks are able to represent important properties of biological systems, and have the potential to provide insight on complex diseases [107]. In this regard, a machine learning approach to infer transcriptional regulatory networks, supplemented by differential methylation and differential expression analysis, is proposed in this thesis to identify potential drug targets involved in epigenetic diseases, with a specific focus on cancer.
1.2 Hypothesis

Computational techniques have been used to infer regulatory networks from gene expression data [11, 84]. However, a characteristic of biological data is the lack of negative examples. This is the root of many of the limitations of current techniques. Networks have also been used to identify disease-associated genes [83]. Currently, the correlation between DNA methylation and gene expression has not been thoroughly explored in epigenetic diseases. Therefore, applying these techniques to an epigenetic disease and integrating DNA methylation and gene expression analysis into such networks may provide insights on the epigenetic mechanisms behind the disease. The following hypotheses will thus be examined in this research:

- The correlation between DNA methylation and gene expression plays an important role in an epigenetic disease
- A machine learning technique, supplemented by a method to infer negative examples will have a better performance than previous methods that infer regulatory networks
- Combining the analysis of correlation between DNA methylation and gene expression with a machine learning technique that is supplemented with a method to infer negative examples can be used to identify important genes in an epigenetic disease

1.3 Objectives

The objective of this study was to investigate computational approaches which may provide insight on gene expression, epigenetics, and their link with a disease in order to help in the identification of potential epigenetic drug targets. Specifically, this was done by using machine learning to infer a transcriptional regulatory network
from gene expression data of control and disease samples. Differential expression and methylation analysis was incorporated into the network, from which an important subnetwork was isolated. A successful network contains genes and regulations that have previously been found to be involved with the disease being modeled, and should narrow the scope for identifying potential drug targets. Model results should also provide insight on the underlying genetic-epigenetic mechanisms of the disease, and a successful model may help advance drug discovery and development. Differential methylation and expression results should also be in agreement with previous studies.

Cancer data, specifically, Kidney Renal Clear Cell Carcinoma (KIRC) was used in this research. KIRC data was selected in this research due to interest; KIRC is the most common form of kidney cancer and can progress and develop without symptoms, resulting in a low rate of early detection [64, 85].

A successful model may also be applied to other disease datasets.

1.4 Contributions

Contributions in this thesis are:

1. A genome-wide KIRC DNA methylation analysis correlated with gene expression using an interaction analysis strategy

2. A novel multi-label machine learning approach, supplemented by semi-supervised learning to infer regulatory networks

3. An inferred regulatory network for KIRC which incorporates differential DNA methylation and gene expression analysis. Analysis of this network reveals interesting information about KIRC which may lead to potential drug targets.
1.5 Outline of Thesis

The remainder of the chapters are organised as follows.

Epigenetic mechanisms and drugs, the drug discovery and development process, and an overview of classification modeling are discussed in the next chapter. Related work, including current computational approaches in epigenetic drug discovery, as well as biological networks and machine learning in target identification, is presented in Chapter 3. Chapter 4 presents the first contribution: a genome-wide KIRC DNA methylation analysis correlated with gene expression using an interaction analysis strategy. Chapter 5 presents a novel multi-label machine learning approach, supplemented by semi-supervised learning to infer regulatory networks, which is evaluated on a benchmark \textit{E. coli} dataset. The final contribution is presented in Chapter 6, and describes an inferred regulatory network for KIRC which incorporates the correlation between DNA methylation and gene expression. The analysis of this network reveals interesting information about KIRC which may lead to potential drug targets. A discussion of the results is given in Chapter 7, and provides a summary and outlines future work.
Chapter 2

Background

2.1 Epigenetic Mechanisms

DNA can be compacted into chromatin by wrapping around an octamer of histone proteins into a nucleosome (Figure 2.1 on the next page). The interaction of multiple nucleosomes is what makes up chromatin. Histone modifications, such as acetylation, which is the addition of an acetyl group, can change the conformational state of chromatin. When histones become acetylated, they lose their positive charge, which decreases their interaction with DNA. This leads to a more open, active state of chromatin, known as euchromatin, which allows DNA to be accessible to transcription factors (TF) leading to increased gene expression. Histone acetyltransferases (HATs) are among the many enzymes that catalyse the acetylation of histone proteins. Histone deacetylase (HDACs) on the other hand have the opposite effect and lead to heterochromatin, a closed, inactive chromatin structure in which DNA is inaccessible thus leading to decreased gene expression [112].

Methylation of DNA, which is the addition of methyl groups to nucleic acids, is
2.1. EPIGENETIC MECHANISMS

(a) DNA Methylation

Figure 2.1: (a): Methylation primarily occurs in cytosine/guanine (CpG) rich promoter regions of DNA. Carried out by DNMTs and supplemented by SAM (producing SAH as a product), methylation alters accessibility of DNA to transcription [42, 78]. Modified from [53]. (b): Histone modifications, such as acetylation, alter the accessibility of DNA to transcription. Acetylation is mediated by HATs and HDACs [112]. Modified from [69]. (c): The energy from ATP hydrolysis alters the accessibility of nucleosomal DNA through nucleosome ejection, restructuring, or mobilization [119]. Modified from [119]
carried out by DNA methyltransferases (DNMTs) and primarily occurs within 5'-cytosine-phosphate-guanine-3' (CpG) dinucleotides. CpG dinucleotides are located preferentially in the genome, and can be found in gene promoter regions as clusters known as CpG islands (CGI) [42]. Promoters are generally found near the transcription start site (TSS) of genes and in vertebrates approximately 70% of annotated gene promoters have been associated with a CGI. However, CGIs are also found remotely from TSSs and often show characteristics of a promoter [56]. For example, a map of DNA methylation was created from the human brain in a study on the role of intragenic DNA methylation in regulating alternative promoters. The map encompassed 24.7 million of the 28 million CpG sites in the humans genome and it was found that CGIs located in intragenic, intergenic, and promoter regions all overlapped with RNA markers of transcription initiation [76]. Additionally, intragenic CGI methylation has been correlated with gene silencing in a 2011 study of intragenic CGI methylation in the immune system [31].

DNMTs promote the addition of a methyl group from a methyl donor such as S-Adenosyl methionine (SAM) to cytosine (resulting in S-Adenosyl homocysteine (SAH) as a product) [78]. The methylation status of an organism’s genome or DNA of a particular cell or tissue is known as its methylome. DNA methylation may reduce gene expression via two mechanisms. First, methylated DNA may physically impede the ability of a transcription factors to bind to genes. Second, methyl-CpG-binding domain proteins (MBDs) may bind to methylated DNA. MBD proteins then recruit other proteins, such as HDACs that act to change the conformation of chromatin into the transcriptionally inaccessible state, heterochromatin [23]. Additionally, gene body DNA methylation has been shown to have a positive correlation with gene expression
although its function is not well understood. In human a human colon cancer cell line, gene body DNA methylation is thought to lead to gene-over expression, which can for example be induced during carcinogenesis, by preventing the initiation of intragenic promoters or by affecting repetitive DNA in transcription units [76, 121].

2.2 Epigenetic Drugs

Currently, there exist two primary classes of epigenetic drugs; DNA methylation inhibitors and histone deacetylase inhibitors. DNA methyltransferase inhibitors can work in two ways. First, the DNA methyltransferase inhibitors may be nucleoside analogues. When these nucleoside-like inhibitors are phosphorylated into nucleotides, they are incorporated into DNA. There they can prevent methylation by trapping any DNMTs that attempt to methylate them. Second, DNA methyltransferase inhibitors that are non-nucleoside analogues can inhibit methylation by reversibly binding to the active sites of DNMTs, preventing them from binding to DNA. HDAC inhibitors are much more numerous and diverse and can act through a variety of mechanisms to modify gene expression and other cellular processes. Blocking the activity of HDACs leads to the acetylation of both histone and non-histone proteins. This may alter transcription either directly or indirectly, affect DNA replication and repair, and can influence cell differentiation and programmed cell death [115].

The majority of approved epigenetic drugs, or drugs that are in clinical trials are either HDAC or DNMT inhibitors [42]. These drugs include 5-azacytidine (AZA), 5-aza-2’deoxycytidine (decitabine (DAC)), suberoylanilide hydroxamic acid (SAHA), valproic acid and entinostat [29]. Developing drugs that target DNMTs is of particular interest due to their known association with certain diseases and because of the
complex and less understood effects of HDAC inhibitors. Additionally, targeting specific DNMT isoforms may also reduce the off-target effects that existing drugs suffer from. Currently, there is a need to develop DNMT inhibitors that do not incorporate into DNA. DNA-incorporating inhibitors such as AZA and DAC have been found to have a lack of DNA incorporation at high concentrations, are limited by cytotoxicity, and have variable potency. The variable potency of these drugs may be due to the ability of diseased cells, specifically tumour cells, to limit their incorporation into DNA [42].

2.3 Drug Discovery and Development

Drug discovery and development is a long and costly process. In the United States, bringing a new therapeutic drug to market typically takes an average of 10 years, and costs an average of $2.6 billion (Figure 2.2) [95]. Key stages of the drug discovery and development process are briefly outlined in Figure 2.3 on page 13 [50].

The drug discovery and development process begins with research, which results...
in a hypothesis regarding a protein or pathway and its association with a disease state that can be used to select a target. The target identification and validation phases are crucial in drug discovery and development since the inhibition or activation of the target should ultimately result in a therapeutic effect. Targets may include proteins, genes, and RNA, and a good target is efficacious, safe and “druggable”, meaning that a potential drug can bind to it and induce a biological response. 'The druggable genome' refers to the subset of genes in the human genome that express proteins that drug-like molecules can bind to. Drug targets need to be able to bind compounds with certain physio-chemical properties. The number of hydrogen-bond donors, molecular mass, lipophilicity, and the number of nitrogen and oxygen atoms all contribute to 'druggability' [48]. Data mining has significantly contributed to the identification of drug targets. Data sources include publications and patent information, transgenic phenotyping, compound profiling data, and genome wide association data. Alternatively, phenotypic screening has been used, which involves identifying a target that is found to alter the phenotype of a cell or organism [50].

The selected target must then undergo a thorough validation process. This usually involves the use of \textit{in vitro} and \textit{in vivo} models such as cells and whole animals respectively. There are many tools that are used in target validation. For example, antisense technology, which involves the use of RNA-like molecules to prevent the synthesis of an encoded protein, is often used. It allows researchers to study the role of a target in a given disease by preventing its synthesis and observing the effects. Transgenic animals are also frequently used since they allow the observation of the phenotype of whole animals that have been genetically manipulated, which gives insight on the potential functions of the gene [50, 95].
Next is the hit discovery process that involves developing assays to identify a hit molecule. A hit may be defined as a compound that has a desired and confirmed activity resulting from a compound screen. There are many screening techniques that exist, among which is high throughput screening (HTS). HTS is often an automated process that involves screening a compound library against the drug target or a cell-based assay and looking for a target induced response. The intent of these screens is to identify compounds that interact with the target, improve the potency, selectivity, and physiochemical properties of the compound and to verify the initial hypothesis that interaction with the target will elicit a desired biological response [50, 95].

Prior to preclinical and clinical trials, the lead molecule selected from the hit discovery process enters the lead optimization phase. Here, the goal is to maintain and improve the desirable properties of the lead compound. The drug candidate is observed in various \textit{in vitro} and \textit{in vivo} models to ensure that it does not induce any genetic mutations or undesirable behavioural or physiological functions. Various pharmacological studies are also conducted to establish how the candidate is metabolised and to explore any stability issues and other chemical properties. This information is assembled along with control considerations to create a target candidate profile in order to be considered for preclinical and clinical trials [50].

### 2.4 Machine Learning using Support Vector Machines

Machine learning is a branch of artificial intelligence in computer science that uses statistical techniques or learning algorithms, such as support vector machines (SVM), to analyse data in order to make predictions. A dataset is composed of rows of observations (or samples), and columns of attributes (or features). The goal of a machine
2.4. MACHINE LEARNING USING SUPPORT VECTOR MACHINES

2.4.1 Preprocessing

Preprocessing is an important step in building machine learning models. Raw, unprocessed data is often noisy and contains missing entries. Additionally, there are
2.4. MACHINE LEARNING USING SUPPORT VECTOR MACHINES

several algorithms, such as SVM, that are sensitive to scaling and require the data to be standardised. This allows the often dynamic values of the features or attributes to be represented on the same scale [89].

2.4.2 Supervised Machine Learning

Classification and regression algorithms are both supervised learning algorithms, meaning that they estimate a relationship between features and a known categorical value. For example, email can be classified as “spam” or “not spam” based on different features, such as how frequently they are received, different keywords contained in the email, whether they have previously been marked as spam, and whether they are sent by a known contact. The classifications “spam” or not “spam” are referred to as the labels or class of the data. During training, the algorithm builds a model by estimating a relationship between the samples (the emails in this example) and labels (“spam” or “not spam”), and predicts a label for each sample. During testing this estimated relationship is evaluated by testing the model’s performance with emails whose labels are already known. Because in this example there are two labels (“spam” and “not spam”), it is referred to as a binary classification problem (labels are often represented as 1 - positive label, or 0 - negative label), however if there were several disjoint labels (“spam”, “not spam”, and “important” for example), then it would be referred to as a multi-class classification problem. However, it is also possible for a sample to belong to multiple categories (it may have more than one label, such as “important” and “personal” in the example of classifying emails), and this is called a multi-label classification problem.
Both classification and regression algorithms make predictions based on data, however classification algorithms predict discrete categorical values, whereas regression algorithms predict continuous values based on past data. Given the previous example, a regression model might be used to predict how much spam would be received given, for example, the amount of spam that was received in the last month, the time of year, and the amount of spam that has been received by other users [17].

Support Vector Machines

A support-vector machine is a supervised learning model that uses learning algorithms to analyse data used for classification. SVM uses a quadratic optimization problem and similarity measures based on dot products. SVM performs classification by generating a hyperplane (decision surface or decision boundary) that maximises a margin between binary data (data with two labels) (Figure 2.4). The support vectors, a subset of the training samples, are the data points closest to this decision surface, and fully specify the decision function (Equation 2.3 on page 17). SVM benefits from being versatile, allowing for the use of different kernel functions. Kernel functions are useful in classification when the relationship between the data and the labels is not necessarily linear; kernels allow for a non-linear mapping to feature space [44].

The input for SVM is the set of feature, label training pairs where $x_1, x_2, ..., x_n$ denote the input sample features where $n$ is the number of features, and the label (either +1 or -1) is denoted by $y$. The output of SVM is a vector of weights $w$ (or $w_i$) for each feature, whose combination determines the value of $y$. Support vectors are chosen by solving the optimization of maximizing a margin, which is done by minimizing the vector of weights. This can be written as:
2.4. MACHINE LEARNING USING SUPPORT VECTOR
MACHINES

Figure 2.4: Simplified representation of SVM. The algorithm finds a hyperplane that maximises a margin between the classes of data. Here two classes of data are depicted as red circles and green squares. Filled shapes represent the support vectors.

\[ \frac{1}{2} ||w||^2 \]  
\[ (2.1) \]

Because this equation is quadratic, solving it is a quadratic problem. Solving this equation minimises the number of nonzero weights that correspond to just the few features that are important in deciding the separating hyperplane. To avoid the problem of over-fitting, a penalty function (Equation 2.2) can be added to the quadratic problem directly:

\[ C \sum_{i=1}^{m} \xi_i^k \]  
\[ (2.2) \]

where the error term \( \xi_i \) is minimised, \( C \) is a regularization parameter, and \( k \) can be defined to specify sensitivity to outliers [34].

After training and finding \( w \), the prediction on a new, unknown point \( x \) can be made by looking at the sign of:
2.4. MACHINE LEARNING USING SUPPORT VECTOR MACHINES

\[ f(x) = \sum_{i=1}^{\ell} y_i \alpha_i K(x_i, x) \]  

(2.3)

where \( \ell \) is the number of training samples and \( y_i, \alpha_i, \) and \( x_i \) specify the label, weight, and features, respectively, of a support vector. A kernel function \( K(x, z) \) specifies the similarity measure used by SVM for different shapes of decision surfaces. Common kernel functions include linear, polynomial, radial basis function and sigmoid kernels.

Choosing optimal parameters is crucial to the performance of a SVM classifier. For example, the \( C \) parameter of the penalty function (Equation 2.2) affects the misclassification of training samples and the shape of the decision surface; a low value of \( C \) results in a smooth decision surface, while a large value of \( C \) allows for more support vectors to be chosen in order to correctly classify more training samples. In order to learn high dimensional data, it has been suggested that the importance of the regularization parameter \( C \) should be increased rather than fitting the data (with kernels for example). This means using a linear kernel and decreasing the \( C \) parameter (emphasizing the margin while ignoring outliers) [114].

When using a radial basis function (RBF) kernel, the width parameter \( \sigma^2 \) specifies the radius of influence of a training sample. Low values have a ‘far’ influence, while large values only have a ‘close’ influence [28, 89]. Parameters such as \( C \) and \( \sigma^2 \) can be optimised by performing a cross-validated grid-search over a parameter grid of all possible combinations of parameter values. The candidate parameter values are evaluated and the best combinations are retained [89].

Because SVM can have two complexities - at testing time and at training time, the computational complexity of certain implementations of SVM scale between:
respectively, where \( a \) is the number of features and \( b \) is the number of samples in either training or test sets.

Training (for kernel SVMs) involves selecting the support vectors and solving the quadratic problem. Solving the quadratic problem requires inverting the kernel matrix, which has the complexity of \( O(n^3) \), where \( n \) is the number of samples in the training set. Testing whether an optimal solution has been achieved involves computing \( O(n^2) \) dot products [14, 89].

### Multi-label Classification

Multi-label classification involves classifying samples that have more than one label. In a multi-label classification problem, each sample may belong to more than one class that are often correlated; a banana may be a fruit and yellow for example. This differs from a multi-class classification problem where the classes are disjoint; a fruit may be an apple or a banana but it cannot be both. There are two main techniques to solve multi-label classification problems: problem transformation and algorithm adaptation [111].

**Problem Transformation**  Problem transformation methods transform a multi-label problem into either several binary classification problems, or a single multi-class problem [111].

The binary relevance approach, also known as the one-vs.-rest (OvR) approach, is one of the main methods that transform multi-label problems into several binary
classification problems. This is done by separating the data by the labels, and building independent classification models for each of these subsets. However, because only one label is considered at a time, if a correlation among labels exists, it is not accounted for [109].

Unlike binary relevance, the label powerset method accounts for the correlation that may exist among labels. A multi-class dataset is created by making a single class from each unique combination of labels. The main shortcoming of this method occurs when some class values are only represented by a small number of samples, creating a multi-class dataset that is unbalanced [109].

Algorithm Adaptation Algorithm adaptation methods are classification algorithms that are extended to handle multi-label data. There are several classification algorithms that have been adapted by incorporating problem transformation methods, such as binary relevance, into the classification algorithm [109].

The Class Imbalance Problem Class imbalance is a common problem in multi-label classification. It may arise when using the label powerset method as described above or because of the nature of the data; some labels do not occur as frequently as others. This is a critical limitation in classification problems that may hinder a model's performance; a classifier may predict a label simply because it occurs more frequently. This problem may be addressed by over-sampling observations with labels that are under-represented for training, or by assigning weights to the labels so that under-represented classes are represented more equally.

However, these solutions are not always effective when many of the labels are unknown and when the dataset has high dimensionality in both labels and features.
Several methods have been proposed to improve multi-label classification performance in these scenarios. This may include dimensionality reduction techniques which reduce the number of features, and semi-supervised learning approaches among others [9, 87].

**Semi-supervised Learning to Address Class Imbalance**  
Label spreading is a semi-supervised learning algorithm. This method allows for data to be partially labeled; for example, labels can be 1, 0, (as in binary classification problems), or -1 indicating unknown, or unlabeled data, and performs well with only a small portion of labeled data, and a large portion of unlabeled data. Label spreading infers the labels of the unknown interactions using the assumption that data points close in proximity in feature space will be labeled similarly. Label spreading using label propagation, which assigns labels to previously unlabeled data points; for each unlabeled data point, the data point updates its label to the one that the maximum number of its neighbours belongs to [130].

**2.4.3 Unsupervised Machine Learning**

Clustering and association algorithms are unsupervised machine learning algorithms, meaning that they estimate relationships in data whose classification is unknown. Unsupervised learning includes algorithms that group or cluster data based on their features with the goal of modeling the underlying structure in the data in order to learn more about it. Following the example from 2.4.2 on page 14, an unsupervised clustering algorithm may be used to examine a collection of emails. It might be found that the data separates well with two or three clusters, from which further analysis might deduce the collection of emails contains mail marked as spam and others that
are not. An association algorithm can be used to discover rules that describe the data, for example, it might be found that emails with the term “winner” are sent more frequently than others [89].

2.4.4 Model Evaluation

Model performance can be evaluated through the use of confusion matrices, precision/recall curves, and area under the receiver operating characteristic curve (AUC-ROC) and are commonly used to summarise the performance of machine learning algorithms, including those used in bioinformatics [3, 18, 21, 22, 33, 62]. Confusion matrices show the number of correct and incorrect predictions made by a model in comparison to the actual, true labels of the data. AUC-ROC is a standard measure for evaluating prediction results of supervised learning techniques such as SVM. An AUC-ROC curve plots the percentage of true interactions as a function of the false positive rate [84]. An AUC-ROC value ranges between 0 and 1, where a value of 0.5 indicates a random classifier and 1 a perfect classifier. AUC-ROC is not affected by under-represented labels in a dataset and thus makes it a suitable metric for classification problems that suffer from class imbalance [33]. Although precision/recall curves are affected by false negatives, they can be used to compare a model’s performance with other models [101]. Precision/recall curves plot the proportion of true positives among all positive predictions made as a function of the proportion of true positives that were correctly identified [84]. A random classifier shows a straight horizontal line on a precision/recall curve, whereas a perfect classifier shows a combination of an horizontal line, followed by a vertical line.
Chapter 3

Related Work

Computational methods, such as the use of biological networks and machine learning techniques, have been shown to be powerful tools in drug discovery (the process of discovering novel candidate drugs) and development (the process of bringing a new drug to market once a candidate has been identified in drug discovery) [82]. The benefits of these approaches lie in their ability to represent real-time events in a fraction of the time, quickly analyse mass amounts of data, and find complex patterns. In epigenetic drug discovery, computational methods provide a means for gaining a better understanding of epigenetic mechanisms and their association with a disease, and identifying potential drugs, and drug targets.

3.1 Current Computational Approaches in Epigenetic Drug Discovery

Several computational approaches have been proposed to advance epigenetic drug discovery. Many of these approaches use molecular modeling techniques such as molecular dynamics simulations, molecular docking, homology modeling and pharmacophore modeling, or virtual screening, although several other methods have been proposed.
3.1.1 Virtual Screening

Also known as computational or *in silico* screening, virtual screening involves computationally searching databases such as small molecule libraries for specific structures of interest [124]. The criteria for these structures are often determined using information from X-ray crystallography or nuclear magnetic resonance and molecular modeling, with the intent of selecting a small number of compounds that are likely to be active. Virtual screening is an attractive method to guide hit identification and lead optimization, and in at least one research group, has been successfully used to identify potential epigenetic drugs [5, 77, 78, 124, 127].

In an early application, virtual screening was used to identify novel DNMT inhibitors. An initial set of 1990 compounds obtained from the Diversity Set available from the National Cancer Institute (NCI) was screened using molecular modeling and the top 2 ranking compounds were validated *in vitro* and *in vivo* [104]. Later, a larger subset of the NCI database consisting of 260,000 compounds was screened, out of which 65,000 compounds were selected. The application of several molecular modeling techniques created a set of 24 compounds out of which 13 continued for experimental testing. Seven of these compounds were found to have detectable DNMT inhibitory activity and at least 6 of these compounds were selective for a specific DNMT isoform [124].

In some cases, drugs have been found to treat diseases and disorders other than the ones they were originally approved for. For example, valproic acid, which is generally used for its anticonvulsant properties in the treatment of epilepsy, has also been found to exhibit HDAC inhibition [113]. This idea was explored in a study which aimed to re-purpose bioactive food compounds. In this study, a database of 4600
3.1. CURRENT COMPUTATIONAL APPROACHES IN EPIGENETIC DRUG DISCOVERY

bioactive food compounds was screened using 32 approved antidepressant drugs. On the basis of chemical similarity, the 10 compounds found to be the most similar to the antidepressant drugs were experimentally screened against an HDAC. Interestingly, these 10 compounds were most similar to valproic acid. Out of the 10 compounds, 2 showed HDAC inhibition equivalent to valproic acid [73].

3.1.2 Molecular Dynamics

Molecular dynamics simulations provide information on the dynamic behaviour of atoms and molecules. Although computationally expensive, molecular dynamic simulations offer many advantages, such as detailed structural data, the microscopic interactions between molecules, and time-dependent responses to perturbations, which complement traditional experiments [2]. Simulations may validate whether theoretical models predict empirical information and can give insight on details not available in experiments. For example, molecular dynamics simulations reveal information on protein dynamics at the atomic-level which may help improve experimental and predicted protein structures [67, 113]. Specifically, molecular dynamics simulations consist of algorithms which evaluate many mathematical physics equations, such as equations of motion [2]. Molecular dynamics provides more detail than other molecular modeling approaches and has applications in enhancing conformational sampling and calculating free-energy changes upon ligand binding [113].

In 1988, hydralazine, which is normally used as a potent arterial vasodilator, was found to exhibit DNA methylation inhibition, and in 2008 showed antitumour effects when combined with valproic acid during clinical trials [36, 124]. In order to gain a better understanding of the underlying molecular mechanisms of the methylation
inhibitory activity of hydralazine, molecular modeling techniques, which included molecular dynamics, were used to model the binding mode of a DNMT isoform [124]. These simulations revealed that hydralazine shares similar binding behaviour as nucleoside analogs, which are known to be important in DNA methylation mechanisms [77].

With the intent of gaining a better understanding of DNMTs, molecular dynamics simulations were used to model the catalytic domains of DNMTs upon binding to SAM. Crystal structures and other molecular modeling techniques were used to represent the different DNMT isoforms. However, on a nanosecond scale, no significant conformational changes were found upon binding of SAM. Nevertheless, the study provided insight on the the protein dynamics of DNMTs when binding to this cofactor [39].

3.1.3 Molecular Docking

Molecular docking often uses experimental data, such as the crystal structures of compounds, in parallel with molecular dynamics and other molecular modeling techniques, to predict how a molecule might fit into a specific binding site, such as the catalytic binding site of a DNMT [77, 126]. Each docking pose is scored in order to find the best position and orientation of the molecule within the binding site [124, 127]. Applications of this method benefit from the ability of scoring a large number of compounds in a small amount of time and generally produce meaningful results [78]. Docking has been used to study protein-ligand interactions of known DNMT inhibitors and has been shown to be important in the drug discovery process, as it can lead to a better understanding of the molecular interactions involved with
potential drugs and can also be used to improve existing epigenetic drugs [82, 127].

With the intent of gaining a better understanding of the different binding poses of DNMTs, 14 compounds with different structural classes, which included nucleoside and non-nucleoside inhibitors, were used for docking. Because the study was conducted prior to the availability of the crystallographic structure of the DNMT, a molecular model of the catalytic domain was used. A comparison between the docking score and experimental data was not possible, however docking revealed similar binding interactions among the different compounds with the binding site that are thought to be crucial in DNA methylation [127].

In a later study, the crystal structure of a DNMT bound to DNA containing unmethylated CpG sites was used to dock known DNMT inhibitors. First, molecular dynamics was used to model the catalytic binding site of the crystal structure, as it was in an inactive state, into an active conformation. The binding poses of the inhibitors were found to share common interactions with the catalytic domain of the DNMT that are involved with the proposed mechanisms of DNA methylation. To further the study, compounds that were recently identified through high-throughput screening were also docked in an attempt to understand their binding modes. These docking models were then used in virtual screening with the goal of finding other inhibitors in large databases. The compounds that were identified were found to have favourable docking scores and included approved drugs ideal for drug re-purposing [79].

Similar docking studies have been conducted to explore the binding of SGI-1027, a known DNMT inhibitor, and propose mechanisms for its inhibitory activity [78, 122].
3.1. CURRENT COMPUTATIONAL APPROACHES IN EPIGENETIC DRUG DISCOVERY

3.1.4 Homology Modeling

Homology modeling is among the top three three-dimensional (3D) structure prediction techniques and is often used as an alternative in the absence of experimental data or 3D structural information of a molecule [58, 113]. Homology modeling involves constructing a 3D model of a protein using its amino acid sequence and an homologous protein as a template. This approach is based on the observation that the amino acid sequence of a protein determines its structure and that related sequences will fold into similar structures [58]. The success of many homology models can be attributed to the well conserved catalytic domain of DNMTs [124]. Before the availability of crystal structures of DNMTs, many structure-based design studies, such as docking, relied on homology models that used the crystal structures of bacterial DNMTs [78, 127]. Homology modeling of this type was essential for the identification of novel DNMT inhibitors [78].

RG108, was the first non-nucleoside DNMT inhibitor identified using a homology model of a human DNMT in combination with virtual screening [78, 104]. Later, two more DNMT inhibitors were discovered using the same homology model and have been used in the optimization of novel DNMT inhibitors. Although the crystal structures of many human DNMTs have since increased, homology modeling still provides an excellent starting point in many molecular modeling studies.

Homology modeling was used in one of the first contributions of molecular modeling to the research of DNMT inhibitors. In 2003, the catalytic domain of a human DNMT isoform was modeled using the crystal structure of a related human DNMT isoform. This model was used to develop N4-fluoroacetyl-5-azacytidine, which was found to successfully inhibit DNA methylation in human tumour cell lines [105]. This
3.1. CURRENT COMPUTATIONAL APPROACHES IN EPIGENETIC DRUG DISCOVERY

homology model was also combined with docking and molecular dynamics to develop a binding mode of hydralazine [108]. Later, using the crystal structures of bacterial DNMTs, two homology models of the catalytic domain of a human DNMT were constructed. Although the two models were created using different homologous templates, both homology models shared common key interactions in the catalytic site. The models were later validated by superimposing them on their recently published crystal structure, and it was found that the homology model was in agreement with proposed mechanisms of DNA methylation [124].

3.1.5 Pharmacophore Modeling

The concept of a pharmacophore has existed for over a century [120]. Although the basic idea of a pharmacophore hasn’t changed, recently, the International Union of Pure and Applied Chemistry (IUPAC) has formally defined a pharmacophore as

an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response [117]. More simply, a pharmacophore model can be defined as a structure-based model that describes features necessary for molecular recognition [120]. It is often used in virtual screening, docking, lead optimization and is one of the major tools in drug discovery [120, 127].

In 2011, Yoo and Medina-Franco proposed a computational method that could be used in parallel with experimental data and molecular dynamics, to develop a pharmacophore. An active conformation of the catalytic binding site of a DNMT inhibitor was modeled using its crystal structure and molecular dynamics. Molecular
docking was then conducted by using known active compounds in the catalytic site. Comparison of the molecular docking results with previous research confirmed the model. They proposed that the information gained from this model could be used to determine binding behaviour of DNMTs, which may lead to the development of effective non-nucleoside inhibitors [125, 126]. This method has been shown to be important in the drug discovery process, as it can lead to a more thorough understanding of the molecular interactions involved with potential drugs and their possible structures. This technique may also be used in toxicity screening and to improve existing epigenetic drugs [82].

Pharmacophore modeling is frequently performed alongside virtual screening and molecular docking, and can use either homology models or crystal structures [79, 125–127]. This ensemble of techniques allows researchers to explore ligand-binding interactions of DNMTs, giving insight into possible inhibitors [79]. For example, Yoo et al. developed pharmacophore models for 16 known DNMT inhibitors. Using the best scoring docking poses and a homology model of the catalytic binding site, researchers found that many of the inhibitors matched the pharmacophore features of the model [123]. Results led to the identification of aurintricarboxylic acid (ATA) as a novel DNMT inhibitor.

3.1.6 Additional Methods

Identifying epigenetic targets has recently become an area of particular interest for successful drug research. However, current research is limited by a lack of thorough
understanding of the underlying biology of many epigenetic targets [26]. In this regard, researchers developed a method to gain a better understanding of genome functionality through database analysis to determine epigenetic targets. A 2015 study utilised several databases such as The Encyclopedia of DNA Elements (ENCODE), the University of California Santa Cruz (UCSC) Genome Browser and The National Center for Biotechnology Information (NCBI) to collect information that may help drive future in vivo studies. The goal of the ENCODE Project is to reveal and characterise functional elements of the human genome. This data can then be passed to the UCSC Genome Browser to identify epigenetic modifications, which can be further superimposed with functional genomic studies from NCBI. Through this process, researchers may gain a better understanding of relevant systems and may use this information to guide in vivo epigenetic target studies [47].

Computational approaches are also used to gather more information about epigenetic systems. Epigenetic protein targets and drug molecules are difficult to identify because they are highly dynamic. In 2012, Baron and Vellore approached this problem by using high-performance computing technology to dynamically visualise the conformational changes of different molecules. They performed a molecular dynamics computer simulation on X-ray crystal structures of a protein involved in epigenetic changes. This involved taking several static images and creating a dynamic visualization of the mechanisms of the protein. Visualization of these dynamic changes helped explain the ability of the protein to affect a variety of molecules involved in epigenetic regulation, and gave researchers insight into designing more targeted epigenetic drugs [5].
3.2 Biological Networks and Machine Learning in Target Identification

In biology, complex systems can be analysed using networks; a collection of nodes that represent biological elements (biomolecules, diseases, phenotypes, etc.) and edges between them that represent a specific relationship (physical interactions, shared metabolic pathway, shared gene, shared trait, etc.) [20]. Protein-protein interaction networks (molecular networks), genetic regulatory networks (DNA-protein interaction networks), gene co-expression networks (transcript-transcript association networks), transcriptional regulatory networks, metabolic networks, and signalling networks are examples of the many biological networks that are used to model systems [70]. Networks have been found to be excellent models of biological systems and can provide insight on the behaviour and functional states of a system [106]. However, these networks are often large and have intricate structures that are difficult to decipher. Fortunately, recent developments in network theory have made advances in characterizing the relationships in biological networks [4].

Machine learning has had applications not only in inferring these biological networks, but also in directly assessing biological data. For example, machine learning can be used to recognise patterns in DNA sequences and identify TSS, splice sites, and promoter and enhancer regions. Additionally, when applied to gene expression data, machine learning techniques can be used to distinguish between disease phenotypes and identify disease biomarkers [65].

Machine learning can be used to infer biological networks. SIRENE (supervised inference of regulatory networks) is a method that uses the SVM algorithm to infer gene regulatory networks from gene expression data and known TF-gene associations. Unlike other techniques, SIRENE divides the problem of gene regulatory network
inference into several binary problems. For each problem SVM predicts whether or
not each gene in the dataset is regulated by a TF. This is then repeated for every TF,
resulting in a network of known and predicted TF-gene associations. In comparison
to other state-of-the-art inference methods, SIRENE predicted on the order of 6 times
more known regulations when tested on a benchmark experiment aimed at predicting
regulations in *E. coli*. Knowledge on transcriptional regulatory networks can help
understand underlying mechanisms of a disease and can be useful for identifying
novel therapeutic targets [84].

Metabolic, signalling, and regulatory networks such as those previously described,
can narrow the scope for identifying drug targets [85] and have also been used to assess
the relationship between gene expression and DNA methylation [85]. Furthermore,
biological networks can lead to the generation of experimentally testable hypotheses
[63]. In a 2009 study of allergic inflammation, a large number of gene expression
microarray experiments were used to construct a hypothetical module or subnetwork.
It was hypothesised that the module genes would show significant gene expression
changes in allergen-challenged CD4\(^+\) cells taken from patients with seasonal allergic
rhinitis (SAR). It was found that 23 of the 33 genes in the module had significant
expression changes in response to stimulation with grass pollen extract. It was also
hypothesised that treatment with glucocorticoids (known to be effective in the treat-
ment of allergic inflammation) would significantly reverse gene expression of these 23
genes - 16 out of the 23 genes satisfied this hypothesis. Additionally, experiments
were conducted to validate a novel disease gene in the module. By looking at expres-
sion levels and gene products, interleukin receptor 7 (IL7R) was found to have an
inhibitory role in allergen inflammation [83].
Chapter 4

Genome-wide Statistical Analyses Reveal
Correlation Between DNA Methylation and Gene
Expression in Kidney Renal Clear Cell Carcinoma

4.1 Introduction

Kidney renal clear cell carcinoma (KIRC) is among the three common forms of renal cell carcinoma (RCC), alongside papillary renal cell carcinoma, and chromophobe renal cell carcinomas. In 2013, RCC was the most common form of kidney cancer to affect adult populations in Western countries [64]. KIRC is the most common form of kidney cancer, accounting for approximately 92% of all cases of kidney cancer among Americans [85].

In KIRC, cancer cells are found in the lining of tubules in the kidney (very small tubes that filter waste from the blood and make urine). KIRC usually affects people over the age of 55 and is more predominant in men. Most cases of KIRC can be treated effectively through surgery, radiation therapy, or chemotherapy when detected early, however survival rates are low when the cancer has spread to other parts of the body.
Additionally, in the early stages of KIRC, the disease often progresses and develops without symptoms, resulting in a low rate of early detection [64].

The role of epigenetics has been gaining interest in cancer research. Specifically, the link between DNA methylation and cancer is being investigated. Although there are numerous cancer studies that explore which genes, and how genes are abnormally regulated by DNA methylation, the relationship between methylation changes and gene expression is still poorly understood [100]. This study presents a series of genome-wide statistical analyses of KIRC data that identifies 37 genes with differentially methylated CpGs and matched differentially expressed transcripts. Among these 37 genes, a correlation was found between 5 differentially methylated CpGs and their matched differentially expressed transcripts, that have previously been identified to have a role in kidney cancer.

4.2 Methods

All statistical analyses were performed in Matlab R2014b.

4.2.1 Data Collection

Level 3 whole-transcriptome RNA-sequencing expression data (Reads Per Kilobase of transcript per Million mapped reads (RPKM)) and whole-genome bisulfite sequencing methylation data (β-values mapped to the genome) for cohorts of healthy controls and participants with KIRC were collected from The Cancer Genome Atlas (TCGA) in March 2016 [52]. The gene expression dataset included 20,532 genes (68 healthy control samples, 475 samples from participants with KIRC). The DNA methylation dataset contained 485,577 probes for 160 healthy control samples, and 325 samples
from participants with KIRC. Healthy control tissue samples were taken from the matched anatomical site of the tumour from the same participant.

**Differential Expression Analysis**

In order to perform valid statistical analyses between control samples and KIRC samples, only the expression data for participants with matched healthy control samples and cancer samples were kept for differential expression analysis. This reduced the data to 65 participants that had both healthy control samples and KIRC samples.

Prior to analysis, missing values were imputed using the median of the column, where a column contained the expression values for a gene.

The dataset reported gene expression values in RPKM. This posed some initial challenges; often RNA-Seq experiments report many low-intensity genes which may be identified as differentially expressed because their fold-change (ratio of average gene expression between two conditions) might be high, when their actual expression values differ very little. This contributes to the false-discovery-rate (FDR), which is the proportion of false positives of genes initially identified as differentially expressed. The accumulation of false positives that result from conducting many statistical tests (the multiple testing problem) can be controlled by adjusting the FDR [30].

Intensity specific thresholds have been proposed as a solution to address the challenges of fold-change [30, 116]. Such a method was proposed by Warden, Yuan, and Wu, who calculated differential expression using analysis of variance (ANOVA) [116]. ANOVA assumes a normal distribution which can be approximated by the log2 of the RPKM expression values. A cutoff value (which was found to be optimal between 0.01-1) can be added to address fold-change, giving:
The $p$-value, or probability value, is the probability of obtaining the observed result when the null hypothesis is true, where the null hypothesis is a hypothesis of “no difference” (in this case the null hypothesis would state that there is no difference between control samples and KIRC samples). The null hypothesis is rejected when the $p$-value is less than the chosen significance level (0.05 in this case), and it can be accepted that there is reasonable evidence to support that there is a difference between the two compared groups.

The Benjamini and Hochberg method calculates the FDR by first ordering the $p$-values of a statistical test in ascending order and assigning them ranks (for examples the smallest value is ranked 1). Next, a Benjamini-Hochberg critical value is calculated for each $p$-value using:

$$\frac{i}{m}Q$$  \hspace{1cm} (4.2)

where $i$ is the $p$-value’s rank, $m$ is the total number of tests, and $Q$ is the false discovery rate (a chosen percentage). The largest $p$-value that is smaller than is critical value is then found, and all $p$-value smaller than it are retained.

Genes were defined as differentially expressed if their fold-change was greater than 1.5, and the FDR was less than 0.05. FDR was calculated using the method of Benjamini and Hochberg [8], and the fold-change for each gene was calculated on a linear scale using the mean. Following this method, genes with a $p$-value less than 0.05 were considered differentially expressed, resulting in 868 differentially expressed
4.2. METHODS

Differential Methylation Analysis

As above, in order to perform valid statistical analyses between control samples and KIRC samples, only the methylation data for patients with matched healthy control samples and cancer samples were kept for differential expression analysis. This reduced the data to 160 patients that had both healthy control samples and KIRC samples. Prior to differential methylation analysis, probes that contained missing data, that contained SNPs or that mapped to sex chromosomes were removed, as they may lead to incorrect methylation analysis [6].

The values in original dataset, known as \( \beta \)-values, range from 0 to 1 and measure the percentage of methylation. Methylation can also be represented as a \( \log_2 \) ratio, and values in this form are known as M-values. The \( \beta \)-values were normalised to have a common distribution of intensities using quantile normalization, and the M-values were calculated from the \( \beta \)-values. M-values were used in the analysis as opposed to the \( \beta \)-values as they have been found to be more statistically valid in differential methylation analyses [35]. Probes were aggregated by CpG and an ANOVA as described above was performed. Differentially methylated CpG probes were linked to their transcripts using the definitions in the chip manifests obtained from TCGA. 870 genes with differentially methylated CpGs were found following this method.
4.3. RESULTS AND DISCUSSION

Spearman Rank Correlation

Among the 870 genes with differentially methylated CpGs and 868 differentially expressed genes, 37 of these genes were found to be both differentially methylated and expressed. 20 patients had matched expression and methylation data which were used in the following analysis.

A Spearman rank correlation coefficient ($\rho$) was found between each differentially methylated CpG and its matched differentially expressed transcript using the beta-values of the CpG and the expression values of the transcripts across the 20 patients. Differentially methylated CpGs and differentially expressed transcripts were considered to be correlated if a p-value of less than 0.05 was obtained from the Spearman rank correlation. A correlation was found between 5 differentially methylated CpGs and their matched differentially expressed transcripts. The sign of $\rho$ was used to identify CpGs and their transcripts as being positively or negatively correlated: $\rho > 0$ may either indicated hypermethylation and overexpression, or hypomethylation and underexpression. Likewise, $\rho < 0$ may either indicate hypermethylation and underexpression, or hypomethylation and overexpression [37].

4.3 Results and Discussion

The 37 genes that were found to be both differentially methylated and expressed were analysed using the Protein Analysis Through Evolutionary Relationships (PANTHER) Classification System [81]. 36 of the 37 gene IDs were successfully mapped and results can be seen in Table 4.1 on the next page as well as their methylation status (hyper- or hypo- methylated) and expression status (over or under expressed).
4.3. RESULTS AND DISCUSSION

Table 4.1: List of genes, and their PANTHER classifications, that were found to be both differentially expressed and methylated. The first column includes the mapped gene ID as well as methylation/expression status: (+) denotes hypermethylation or overexpression, (-) denotes hypomethylation or underexpression.

<table>
<thead>
<tr>
<th>Mapped Gene ID with methylation/expression status</th>
<th>Entrez Gene Definition</th>
<th>PANTHER Protein Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANKDD1A+/+</td>
<td>Ankyrin repeat and death domain-containing protein 1A</td>
<td></td>
</tr>
<tr>
<td>B3GNTL1+/-</td>
<td>UDP-GlcNAc:betaGal beta-1,3-N-acetylg glucosaminyltransferase-like protein 1</td>
<td>glycosyltransferase</td>
</tr>
<tr>
<td>UNKL+/+</td>
<td>Putative E3 ubiquitin-protein ligase UNKL</td>
<td></td>
</tr>
<tr>
<td>EP400NL+/+</td>
<td>EP400 N-terminal-like protein</td>
<td></td>
</tr>
<tr>
<td>RIN1-/+</td>
<td>Ras and Rab interactor 1</td>
<td>guanyl-nucleotide exchange factor; membrane trafficking regulatory protein</td>
</tr>
<tr>
<td>PDE4C+/+</td>
<td>cAMP-specific 3',5'-cyclic phosphodiesterase 4C</td>
<td></td>
</tr>
<tr>
<td>SLC6A9+/+</td>
<td>Sodium- and chloride-dependent glycine transporter 1</td>
<td>cation transporter</td>
</tr>
</tbody>
</table>
# 4.3. RESULTS AND DISCUSSION

Continuation of Table 4.1

<table>
<thead>
<tr>
<th>Mapped Gene ID</th>
<th>Entrez Gene Definition</th>
<th>PANTHER Protein Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHPN1+/+</td>
<td>Rhophilin-1</td>
<td>G-protein modulator</td>
</tr>
<tr>
<td>GRK4-/+</td>
<td>G protein-coupled receptor kinase 4</td>
<td>non-receptor serine/threonine protein kinase</td>
</tr>
<tr>
<td>PTGFR-/+</td>
<td>Prostaglandin F2-alpha receptor</td>
<td>G-protein coupled receptor</td>
</tr>
<tr>
<td>ZC3H12A-/-</td>
<td>Endoribonuclease ZC3H12A</td>
<td>nucleic acid binding</td>
</tr>
<tr>
<td>C2orf60-/+</td>
<td>tRNA wybutosine-synthesizing protein 5</td>
<td></td>
</tr>
<tr>
<td>ZNF767-/+</td>
<td>Protein ZNF767</td>
<td></td>
</tr>
<tr>
<td>EPHX4+/-</td>
<td>Epoxide hydrolase 4</td>
<td>serine protease</td>
</tr>
<tr>
<td>WDR90+/+</td>
<td>WD repeat-containing protein 90</td>
<td></td>
</tr>
<tr>
<td>LRRRC27+/+</td>
<td>Leucine-rich repeat-containing protein 27</td>
<td></td>
</tr>
<tr>
<td>HIC1-/+</td>
<td>Hypermethylated in cancer 1 protein</td>
<td>KRAB box transcription factor</td>
</tr>
<tr>
<td>ZNF549+/+</td>
<td>Zinc finger protein 549</td>
<td></td>
</tr>
</tbody>
</table>
## 4.3. RESULTS AND DISCUSSION

Continuation of Table 4.1

<table>
<thead>
<tr>
<th>Mapped Gene ID</th>
<th>Entrez Gene Definition</th>
<th>PANTHER Protein Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMC8-/+</td>
<td>Transmembrane channel-like protein 8</td>
<td></td>
</tr>
<tr>
<td>ZNF589-/+</td>
<td>Zinc finger protein 589</td>
<td>KRAB box transcription factor</td>
</tr>
<tr>
<td>HOXB3+/+</td>
<td>Homeobox protein Hox-B3</td>
<td></td>
</tr>
<tr>
<td>HOXD3-/+</td>
<td>Homeobox protein Hox-D3</td>
<td></td>
</tr>
<tr>
<td>HCN3-/+</td>
<td>Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 3</td>
<td></td>
</tr>
<tr>
<td>DMPK+/+</td>
<td>Myotatin-protein kinase</td>
<td>annexin; calmodulin; non-receptor serine/threonine protein kinase; transfer/carrier protein</td>
</tr>
<tr>
<td>CNTNAP1+/+</td>
<td>Contactin-associated protein 1</td>
<td></td>
</tr>
<tr>
<td>C19orf71-/+</td>
<td>Uncharacterised protein C19orf71</td>
<td></td>
</tr>
<tr>
<td>RNPC3-/+</td>
<td>RNA-binding protein 40</td>
<td></td>
</tr>
<tr>
<td>WDR27+/-</td>
<td>WD repeat-containing protein 27</td>
<td></td>
</tr>
</tbody>
</table>
A correlation was found in 5 differentially methylated CpGs and their matched differentially expressed transcripts. These genes, as well as their Spearman rank correlation coefficients ($\rho$) have been listed in Table 4.2 on the next page. These genes have previously been identified to be involved in kidney cancer. Their potential roles in KIRC are summarised in the following paragraphs.
4.3. RESULTS AND DISCUSSION

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>$\rho$</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>B3GNTL1</td>
<td>-0.5835</td>
<td>hypermethylation, underexpression</td>
</tr>
<tr>
<td>HCN3</td>
<td>-0.5536</td>
<td>hypomethylation, overexpression</td>
</tr>
<tr>
<td>HOXB3</td>
<td>0.5639</td>
<td>hypermethylation, overexpression</td>
</tr>
<tr>
<td>RHPN1</td>
<td>0.4541</td>
<td>hypermethylation, overexpression</td>
</tr>
<tr>
<td>ZC3H12A</td>
<td>0.4887</td>
<td>hypomethylation, underexpression</td>
</tr>
</tbody>
</table>

Table 4.2: List of genes where a correlation was found between their differential expression and methylation values. The associated Spearman rank correlation coefficient $\rho$ indicates a negative or positive correlation: $\rho > 0$ may either indicated hypermethylation and overexpression, or hypomethylation and underexpression. Likewise, $\rho < 0$ may either indicate hypermethylation and underexpression, or hypomethylation and overexpression.

4.3.1 B3GNTL1 is among Metabolic Genes Identified as Differentially Expressed in KIRC

Because KIRC develops and progresses asymptptomatically in the early stages of the disease, the discovery of diagnostic makers and novel therapeutic targets is crucial for patient prognosis. A recent study that was interested in cancer metabolism for the understanding of the molecular mechanism of carcinogenesis (the formation of cancer), analysed the expression of metabolism-associated genes in an attempt to identify metabolic changes between control samples and KIRC samples at different disease stages. KIRC gene expression data was obtained from the TCGA, and the Recon2 model was used to obtain data on metabolism-associated genes. B3GNTL1 was among 89 metabolic genes that were identified as differentially expressed in late stages of KIRC. B3GNTL1 has been found to be involved in the metabolism of proteins and O-linked glycosylation, and gene ontology (GO) annotations include transferase activity, and transferring glycosyl groups. Glucosyltransferase is a member of a glycosyltransferase family [16], and inhibition of glucosyltransferase enzymes has been shown to reverse the disease phenotype and has been proposed as a potential
treatment strategy for kidney disease [64]. In this regard, it may be beneficial to explore inhibitors of glycosyltransferase enzymes that may be related to B3GNTL1 activity.

4.3.2 HCN3 was found to be Significantly Correlated to Lower Survival Rates in Kidney Cancer

A recent study examined the expression of hyperpolarization-activated cyclic nucleotide-gated genes (HCN1-4) in multiple types of cancer. Hyperpolarization-activated cyclic nucleotide-gated (HCN) channels are one of many intra-membrane ion channels that control the flow of ions through a cell. Ion channels are a crucial component of the nervous system, as well as muscle contractions, the transport of nutrients and ions in epithelial cells (found on the surface of skin, blood vessels, urinary tract, and other organs), T-cell (a subtype of white blood cell involved in immune response) activation, and insulin release. However, the role of HCN channels in cancer is unknown. This study investigated the correlation between overexpression of HCN genes and survival rate of cancer patients using expression data collected from public microarray research databases Oncomine and NextBio. Results suggested that HCN genes are good potential candidates for cancer diagnosis and prognosis. HCN3 specifically was found to be over-expressed and significantly correlated with low survival rates in kidney cancer. Inhibition of HCN channels has previously been found to decrease cell proliferation, and has been suggested as a potential target for tumour suppression [90].
4.3. RESULTS AND DISCUSSION

4.3.3 Tumour Suppressor Gene Epigenetically Silenced through HOXB3 in a Variety of Cancers

The RASSF1A gene is a tumour suppressor gene that has been found to be epigenetically silenced in a variety of cancers. The inactivation of this gene has been linked to over 40 types of cancers, including renal cell carcinoma [88, 110]. A genome-wide human RNA analysis found that the epigenetic silencing of RASSF1A requires the homeobox (a gene that encodes a protein that binds to DNA) HOXB3 protein. It was found that the HOXB3 protein binds to a DNMT and increases its expression. This leads to the hypermethylation of the RASSF1A promoter, and the silencing of its expression. Over-expression of HOXB3 was strongly correlated with RASSF1A silencing and has also been found in renal carcinomas [24, 88].

4.3.4 RHPN1 found to be a Key Regulator of the Podocyte - A Promising Target for Kidney Disease

The biological function of RHPN1, a Rho GTPase-interacting protein is relatively unknown, however, Rho GTPases regulate cell cytoskeleton remodeling and cell migration [15, 60]. In an in vivo mouse study, RHPN1 was identified as a novel podocyte-specific protein of the kidney glomerulus (the kidney filter). Podocytes are cells in the glomerulus that are essential for its integrity [60]. It was found that RHPN1 knock-out mice developed abnormal podocytes and a thickening of glomerular membranes, and the study results suggest that RHPN1 is essential for healthy podocyte formation [15]. The evaluation of RHPN1 and podocytes may therefore be beneficial in the early detection of KIRC.
4.3.5 ZC3H12A Encodes Protein which Contributes to KIRC Development

ZC3H12A encodes MCPIP1, a protein that was recently found to contribute to KIRC development. MCPIP1 regulates inflammatory processes through multiple mechanisms. KIRC samples and adjacent normal tissue samples from KIRC patients were analysed to estimate the level of transcripts coding for MCPIP1, which were found to be downregulated in KIRC. The role of MCPIP1 in the development in KIRC was further examined using cell lines [66]. The differential methylation and expression of ZC3H12A may therefore contribute to KIRC development by way of MCPIP1.

4.4 Conclusion

The statistical analyses presented in this chapter reveal 5 genes that may have an important role in kidney cancer - an epigenetic disease. Further investigation of these 5 genes in vitro, as well as analysis of the 37 genes with differentially methylated CpGs and matched differentially expressed transcripts, may be beneficial to gaining a better understanding of the link between methylation changes and gene expression in KIRC. Additionally, such research may help identify biomarkers for the early detection of KIRC, and new therapeutic targets. An example of this research is presented in Chapter 6, where the results that were presented in the current chapter were used to supplement a computational model to infer a regulatory network of KIRC.
Chapter 5

Multi-label Support Vector Machines to Infer Regulatory Networks

5.1 Introduction

Interest in elucidating gene regulatory networks (GRN) - networks inferred from gene expression data, has increased in recent years due to the increased availability of multiple ‘omics’ data. GRNs may provide information on regulatory interactions such as gene-gene interactions, protein-protein interactions, and gene-protein interactions. This has led to the development of several GRN models which have been used to better understand diseases, and have aided in the discovery of novel therapeutics [38, 51].

Cell activity is affected by genes through proteins known as transcription factors (TF). TFs control the rate at which genetic information is transcribed from DNA to messenger RNA (mRNA), from which ribosomes can synthesise proteins. These proteins may include more TFs that also control the expression of one or more genes [63]. This complicated process can be represented by GRNs, where TFs and genes are represented by nodes, and edges are interactions between them.
Many computational approaches exist to infer regulatory networks, including unsupervised, semi-supervised, and supervised techniques, each offering their own benefits. The majority of these methods are unsupervised, meaning that they do not require prior knowledge of known interactions. The simplest of these models are based on Boolean logic, which assumes that genes exist in discrete states: on (active or expressed) or off (inactive or unexpressed). Because of their simplicity, Boolean models are not computationally expensive, and have provided insights on the design and properties of GRNs [51, 63]. Boolean network models are often easy to implement; however they are not always able to offer the same level of information that continuous models are able to represent.

Discrete Bayesian network models are also a popular choice for building GRNs. A Bayesian network is an directed acyclic graph where the edges between nodes describe a regulatory relationship, as well as conditional dependencies based on probability distributions of a set of variables. Bayesian models may either be discrete or continuous when derived from time-series data [131]. Although they are often found to be successful, modeling Bayesian networks involves many statistical and probabilistic calculations and can be computationally taxing [63].

The statistical analysis of dependencies between expression patterns has also been used to successfully infer large-scale GRNs. For example, coexpression networks use correlation coefficients that are derived from the expression patterns of all pairs of genes. Mutual information (MI) measures have been proposed to capture more complex, non-linear relationships between expression patterns and are used in Relevance Networks; models where the MI is calculated for all pairs of genes and a regulatory interaction, represented by an edge, is inferred when the MI falls above a threshold.
However, Relevance Networks methods often suffer from false positives that arise from one or more indirect relationships between genes that are highly co-regulated, which the model cannot reduce. To address this issue, several variations of Relevance Networks have been proposed to distinguish between direct and indirect relationships in the network. The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) algorithm [72] uses the concept of the Data Processing Inequality to filter out indirect relationships from all triplets of genes. The Minimum Redundancy Networks (MRNET) [80] approach uses the maximum relevance/minimum redundancy technique to select variables with the highest MI with a target to infer relationships. Finally, the Context of Likelihood of Relatedness (CLR) algorithm [41], which has been considered as state-of-the-art among methods, modifies the MI score based on the distribution of all MI scores [51, 63].

Prior to the development of the CLR algorithm, no benchmark of biologically validated regulatory relationships had existed to assess GRN model performance. In order to evaluate CLR, the study compiled an *E. coli* benchmark and compared several approaches which included Bayesian networks and ARACNe, and found that CLR had the best model performance [41, 63]. Soon after, the Supervised Inference of Regulatory Networks (SIRENE) algorithm [84] was applied to this benchmark. The SIRENE algorithm uses several binary classification models to predict TF-target gene interactions using the known regulatory relationships of the *E. coli* benchmark. The SIRENE method’s performance is reported to outperform the CLR and ARACNe algorithms, however further experiments in the current study have revealed that the SIRENE model does not perform as well as other methods.

In addition to the above unsupervised techniques and SIRENE, a small number
of supervised machine learning techniques have also been used to infer regulatory networks [11, 84]. Machine learning uses learning algorithms which analyse data in order to make predictions. For example, GRNs can be inferred from gene expression data and known interactions by learning a pattern that associates interactions with gene expression profiles [101]. Supervised learning techniques have been found to perform better than unsupervised methods which infer GRNs, however only a small number of such methods have been explored [71].

Although supervised learning algorithms have been successful, they are limited by existing information, for example, the availability of known TF-target gene regulations [101]. This may negatively affect model performance; when considering all possible TF-target gene interactions, a very small ratio of interactions are actually known. In classification modeling this is referred to as class-imbalance, a problem which leads to a bias in predicting non-interacting pairs [40]. Additionally, supervised learning techniques require knowledge of both positive (interacting) and negative (non-interacting) examples, however often true non-interacting pairs are not available for biological networks. Most techniques simply treat the unlabeled data (pairs with no known interactions), or a subset of it, as negative examples, thereby risking the presence of false negatives affecting the performance of the classifier. The model will also be biased towards predicting low scores for these examples [101].

Here is presented a multi-label classification technique to infer regulatory networks (MInR) from gene expression and TF data. Unlike other methods, this technique was supplemented by a semi-supervised learning step to address the problem of class-imbalance; rather than assume that all non-interacting TF-target gene pairs are true negatives simply because the information is missing, MInR allows for the assumption
5.2. METHODS

that some of these pairs are ‘unknown’. The model was tested on a benchmark experiment [41] which includes expression and regulation data for *E. coli*, and is commonly used to evaluate GRNs. Model performance was found to be comparable to other methods, and improved when semi-supervised learning was applied to address class imbalance.

5.2 Methods

5.2.1 Data and Data Preprocessing

Publicly available regulation and expression data for *E. coli* of a benchmark experiment [41] was used in the following model. The expression data included 445 expression profiles for 4327 genes. The expression data for each gene was standardised using a Gaussian distribution to have zero mean and unit variance.

Regulation data was retrieved from the RegulonDB database [98] and included 4357 regulations (where 3277 have been confirmed with strong evidence) between 189 TFs and 1807 genes. This information was used to create a labels matrix; for each of the 4327 genes of the benchmark data and 189 TFs from RegulonDB, a 1 between a gene and TF indicates a regulation and 0 otherwise.

A list of 899 known operons in *E. coli* were also retrieved from RegulonDB. An operon is a functional unit comprised of a group of genes that are transcribed into a single mRNA. Gene expression data was sorted according to operon groups so as not to split genes within an operon between training and testing during cross-validation; generally, a TF regulates all genes within an operon which tend to have similar expression profiles. If genes within the same operon were split between testing and training, a classifier would correctly predict a regulation for the test gene simply
because it was within the same operon as the training gene, rather than predicting regulations for new operons [84].

### 5.2.2 Semi-supervised Learning to Infer Labels

Label spreading, a semi-supervised learning algorithm was applied prior to classification modeling. Assuming that 0s in the labels matrix include unknown, as well as non-interacting TF-target gene pairs, a third of the 0s were set to be unknown (-1) and label spreading was executed three times to ensure that all zero labels were included in inferencing.

### 5.2.3 Classification Modeling

SciKit Learn’s [89] implementation of multi-label, one-vs.-rest linear SVM was used for classification modeling. SVM has been found to perform well in computational biology [11], and SVM models that use linear kernels have been found to perform better on biological data such as gene expression [114]. Parameter selection was done using a grid-search with cross validation (2/3 of the data was used for training and 1/3 for testing). Starting from -2, a logarithmic grid with a basis of 10 was used to generate 50 C values. Through this method it was found that the ideal value for the C parameter was 0.01. Additionally, the class weight parameter was set to ‘balanced’ so that both interacting (1) and non-interacting (0) labels were represented equally to avoid prediction bias.

The data was then split for a 3-fold cross-validation (2/3 was used for training and 1/3 for testing).
5.3. RESULTS

5.2.4 Model Evaluation

Classification performance was averaged over the 3-folds and evaluated using the AUC-ROC, and precision/recall curves. The lack of knowledge of true non-interacting pairs (a 0 in the labels may indicate either a non-interacting pair, or an interaction that is not known) poses a problem in model evaluation; evaluation measures commonly used to evaluate models rely on the availability of both interacting and non-interacting pairs. Fortunately, it was found that the presence of false negatives among unlabeled data, that are used as true negatives, tends to have a negligible effect in AUC-ROC curves. Although precision/recall curves are affected by false negatives, they were used in order to compare performance with other models [101].

Predicted TF-target gene interactions were ranked according to decision values; only interactions whose decision values were greater than or equal to 1 were kept, as these typically correspond to confident predictions [11]. These interactions were then visualised as a network using Cytoscape [97] where nodes represent TFs or genes, and edges represent interactions between them.

5.3 Results

5.3.1 Model Performance

MInR’s performance was compared to the CLR, ARACNE, MRNET, and SIRENE algorithms on the E. coli benchmark [41] by comparing precision/recall curves (Figure 5.1 on the next page), and evaluated by computing AUC-ROC curves (Figure 5.2 on the next page). Overall, model accuracy was found to be 65% with an AUC-ROC score of 0.86. MInR was able to correctly predict 82% of known TF-target gene pairs, and 65% of non-interacting TF-target gene pairs.
5.3. RESULTS

(a) MInR’s precision/recall curves
(b) Other models’ precision/recall curves

Figure 5.1: Comparison of MInR’s precision/recall curves with the curves of various algorithms including SIRENE, CLR, MRNET, and ARACNE on *E. coli* benchmark. Note scale. Shows MInR’s performance with and without semi-supervised label spreading. Modified from [51].

Figure 5.2: ROC curves for MInR with and without label spreading on *E. coli* benchmark.
5.3. RESULTS

Semi-supervised label spreading was found to positively affect model performance; the AUC-ROC score increased by 37%, and the model was able to predict 22% more known TF-target gene pairs (Figure 5.2 on the previous page).

Model comparison with SIRENE found that SIRENE model performance was not consistent with the results that were previously reported [84] (Figure 5.3 on the next page).

5.3.2 Regulatory Network

MInR predicted 1811 TF-target gene interactions with decision values greater than 1 between 153 TFs and 711 genes. 847 of these interaction have been confirmed in RegulonDB, leaving 947 new TF-target gene interactions that were predicted by MInR. Removing non-TF genes in order to better illustrate some of the predicted interactions resulted in 35 TF-target gene interactions between 18 genes and 22 TFs (Figures ?? on page ??).
5.3. RESULTS

(a) Obtained SIRENE precision/recall curve

(b) Reported SIRENE precision/recall curve

(c) Obtained SIRENE ROC curve

(d) Reported SIRENE ROC curve

Figure 5.3: **Fig. 5.3a:** Obtained precision/recall curve for SIRENE. **Fig. 5.3b:** Reported precision/recall curve for SIRENE on *E. coli* benchmark. Modified from [84]. **Fig. 5.3c:** Obtained AUC-ROC curve for SIRENE on *E. coli* benchmark. **Fig. 5.3d:** Reported AUC-ROC curve for SIRENE on *E. coli* benchmark. Modified from [84].
Figure 5.4: Inferred *E. coli* regulatory network produced by MInR. Part 1/2. Nodes are either genes or transcription factors, edges indicate interactions between them. Predicted new interactions are green, confirmed interactions are blue.
5.3. RESULTS

Figure 5.5: Inferred E. coli regulatory network produced by MInR. Part 2/2. Nodes are either genes or transcription factors, edges indicate interactions between them. Predicted new interactions are green, confirmed interactions are blue.
5.4 Discussion

A multi-label classification model, supplemented by a semi-supervised learning, MInR, was presented to infer regulatory networks. This new method has the benefit of being relatively computationally inexpensive and easy to implement. Additionally, unlike previous methods, MInR addresses the problem of class-imbalance and the selection of negative examples, factors that limit many supervised learning models that rely on pre-existing information, by implementing label-spreading. This additional semi-supervised learning step was found to positively affect model performance (Figure 5.2 on page 54) and may have applications in other biological models that suffer from class-imbalance.

When evaluated on the *E. coli* benchmark [41], MInR’s new supervised machine learning approach was found to have a better performance than previous unsupervised methods [41, 72, 80, 84] that “reverse engineer” regulatory networks from gene expression data (Figure 5.1 on page 54). MInR was also found to perform better than a comparable supervised machine learning method, SIRENE [84]. However, it was found that the obtained performance of SIRENE was not consistent with what was previously reported by the developers of SIRENE, and is not likely a reliable model to compare to. When running the freely available Matlab code for the SIRENE model (available at: http://projects.cbio.mines-paristech.fr/sirene/), the precision/recall and ROC curves that were outputted were not consistent with those that were previously published (Figure 5.3 on page 56). An attempt to contact the authors of the SIRENE paper to clarify their results was unsuccessful.

*PosOnly*, an interesting supervised learning approach that infers regulatory networks from only positive labels has been found to outperform unsupervised methods
as well as SIRENE [19]. This method avoids the problem of selecting negative examples by separating positive examples from unlabeled data (as opposed to separating positive examples from negative examples) by learning the probability that a given example will be positive given its gene expression profile. Although this method may outperform existing approaches, it still depends on the availability on the few interactions that are known and may benefit from the semi-supervised approach proposed here. However, more evaluation is needed in order to compare PosOnly to MInR.

The inferred regulatory network was reduced to include only novel regulations of TFs by other TFs (Figure ?? on page ??); TF-TF interactions have been shown to be important determinants in cell functions, are frequently altered in disease, and might function as biomarkers [94, 118]. A quick analysis using ClueGO, a Cytoscape plug-in that integrates gene ontology and pathway annotation [10], revealed that the cluster around the cspA node (the gene encoding the cold-shock protein), that contained most of the newly predicted interactions, was comprised of regulations related to positive regulation of transcription, DNA-templated, and glyoxalase III activity. A quick literature search confirms the association of cold shock protein expression with glyoxalase activity [99].

An important limitation of MInR is the inability to predict new TF-target gene interactions of TFs with no previous known interactions - only TFs with known regulations were used in modeling. Incorporating TF binding site information from databases such as JASPAR [74] and TRANSFAC [75], as well as chromatin immunoprecipitation data from ENCODE [27] may provide additional insight on previously unknown TF-target gene interactions.
5.5 Conclusion

It was found that machine learning has potential in inferring regulatory networks. A multi-label classification model supplemented by semi-supervised learning, MInR, was found to outperform previous techniques that infer regulatory networks. MInR can likely be applied towards more complex regulatory networks and advance the discovery of novel therapeutic targets. In the next chapter, MInR is used alongside the differential methylation and differential expression results obtained in the previous chapter, to infer a regulatory network for KIRC. This was done in order to gain a better understanding of KIRC, and to identify potential therapeutic targets and biomarkers for the disease.
Chapter 6

Inference of Kidney Renal Clear Cell Carcinoma Regulatory Network using MInR Supplemented with Gene Expression and Methylation Analysis Reveals Therapeutic Opportunities

6.1 Introduction

The development of drugs, including epigenetic drugs is a long and expensive process. The discovery and development of epigenetic drugs is particularly difficult due to limited knowledge of epigenetic mechanisms, the relationship between epigenetics and gene expression, and their link with disease. Computational approaches, primarily molecular modeling techniques such as those outlined in Chapter 3, have been explored in order to help expedite the discovery and development of epigenetic drugs. Although these approaches have been successful, they provide little information on the relationship between gene expression and epigenetics, and their roles in an epigenetic disease such as KIRC. Methods that integrate gene expression and epigenetic information, such as DNA methylation data, are of particular interest because they
6.2. METHODS

may reveal novel epigenetic drug targets [54]. Additionally, networks have been found to be excellent models of biological systems and have been used in target identification [83, 106].

Here, KIRC is used to demonstrate a workflow for a model that integrates DNA methylation and gene expression with network analyses to identify potential drug targets. A regulatory network was generated from gene expression data using the MInR technique presented in Chapter 5. MInR was found to be an appropriate technique for the inference of regulatory networks as the data contained a high ratio of unknown regulations. Next, the genome-wide statistical analysis presented in Chapter 4, which revealed a correlation between DNA methylation and gene expression in KIRC, was added to the network. Network analysis highlighted HOXD3 as a potential therapeutic target.

6.2 Methods

6.2.1 Data and Data Preprocessing

Whole-genome expression data and the matching DNA methylation data for cohorts of healthy controls and patients with KIRC were collected from the TCGA in March 2016 [52]. The gene expression dataset included 20,532 genes (68 healthy control samples, 475 samples from patients with KIRC). In order to perform valid analyses, only the expression data for patients with matched healthy control samples and cancer samples were kept for modeling. This reduced the data to 65 patients that had both healthy control samples and KIRC samples.

Regulation data was retrieved from the Transcriptional Regulatory Relationships
Unraveled by Sentence-based Text mining (TRRUST) database [46] as well as A Public Database of Transcription Factor and Regulatory Sequence Annotation (PAZAR) [91] and combined. This resulted in a TF dataset that included 12,739 regulations between 765 TFs and 5125 genes. This information was used to create a labels matrix; for each of the 20,532 genes in the gene expression dataset and 765 TFs from the TF dataset, a 1 between a gene and TF indicates a regulation and 0 otherwise.

Prior to network creation using MInR, genes with no known TF regulations were removed. This resulted in 4573 genes out of the original 20,532. This reduction was both desired and necessary for further modeling - applying MInR’s multi-label SVM on such a large dataset is computationally expensive. Although MInR performs relatively well with data that contains a high ratio of negative to positive labels, it would be difficult to validate the high number of resulting novel regulations that are additionally evaluated as false positives. MInR also required that more than 1 regulation existed for each TF, reducing the original number of 765 to 441.

6.2.2 Model Evaluation

MInR’s performance was averaged over 3-folds and evaluated using AUC-ROC curves on the healthy control dataset.

6.2.3 Network Creation

A grid-search was performed to find the optimal C value used by MInR. Using a logarithmic grid with a basis of 10 to generate 50 C values, it was found that the ideal value for the C parameter was 0.01. Using 2/3 of the data for training and 1/3 for testing, MInR was used to create two networks; one from the healthy control
samples, and a second was created using the samples from patients with KIRC, where nodes represent TFs or genes, and edges represent interactions between them. As per the MInR technique, predicted TF-target gene interactions were ranked according to decision values; only interactions whose decision values were greater than or equal to 1 were kept. In order to illustrate the disease state, a single network was created by focusing on the KIRC network: First, all interactions that appeared in both the healthy control network were removed from then KIRC network. Second, a subnetwork was isolated by only retaining the nodes that were found to be both differentially expressed and differentially methylated that were previously found in Chapter 4, and their first neighbours. These interactions were then visualised using Cytoscape [97]. Gene ontology and pathway annotation was then performed using ClueGO [10].

6.3 Results

6.3.1 Model Performance

MInR’s performance was evaluated on the healthy control dataset. Overall, model accuracy was found to be 59%, with and AUC-ROC score of 0.60 (Figure 6.1 on the next page). MInR was able to correctly predict 52% of known TF-target gene interactions and 59% of non-interacting pairs.

6.3.2 Regulatory Network

MInR predicted 13,372 TF-target gene interactions with decision values greater than 1 between 379 TFs and 588 genes in the control dataset, and 14,494 TF-target gene interactions with decision values greater than 1 between 362 TFs and 637 genes in the KIRC dataset. Removing interactions that appeared in both KIRC and control
6.4. DISCUSSION

Figure 6.1: ROC curve of MInR on healthy control dataset.

networks from the KIRC network removed 4,499 interactions leaving 9,996 interactions in the KIRC network. The final isolated subnetwork contained 12 of the 37 differentially expressed and methylated genes and 94 interactions between 35 TFs and 28 genes (Figure 6.2 on the next page).

6.4 Discussion

MInR was used for the inference of a KIRC regulatory network, from which a subnetwork of interest could be isolated when supplemented with differential expression and differential methylation analysis. Further analysis with ClueGO revealed that the subnetwork was comprised of regulations related to fat cell regulation, various cancers, and cell adhesion mediated by integrins.
Figure 6.2: Inferred KIRC regulatory network. Nodes are either transcription factors or genes, edges represent interactions between them. Predicted new interacting nodes in KIRC are red, differentially expressed and methylated genes are highlighted in yellow.
6.4. DISCUSSION

6.4.1 Model Performance

MiNIR’s performance was found to be low (AUC-ROC value of 0.60, an accuracy of 60%, and correctly predicts 52% of known interactions) in comparison with its performance on the benchmark *E. coli* dataset in Chapter 4 (AUC-ROC value of 0.86, an accuracy of 65%, and correctly predicts 82% of known interactions). This is likely due to the transcription regulation differences between *E. coli* and humans. In humans (and other eukaryotes) as opposed to prokaryotes like *E. coli*, transcription involves a large class of transcription factors. For example, a transcription factor might only activate a set of genes needed in certain cell types [25]. SVM may therefore be able to capture the regulatory network complexity in *E. coli*, but may not perform as well in a more complex organism such as humans. Further refinement of the model may be needed, or a model that is better suited to capturing complex systems, such as neural networks, may be applied.

Other contributing factors may include the availability of known human TF-target gene interactions. The TRRUST database claims (to the best of their knowledge) to be the largest publicly available database of literature-curated human TF-target interactions, and a reliable benchmark for the reconstruction of human transcriptional regulatory networks [46]. The TRRUST database is likely reliable, however it was constructed using a text-mining approach which extracted TF-target gene interactions from abstracts involving human biology, which encompasses many tissue and cell types. Many TFs are common to several cell types, however others are cell-specific [1]. This suggests that not all the TF-target gene interactions used in training MiNIR were indeed present in the cell that the control samples were taken from, thereby negatively affecting performance (MiNIR was trying to learn an interaction from a
pattern in gene expression that was not present). Future work may focus on creating a labels matrix from TF-target gene pairs that are expressed in the given sample.

6.4.2 Network Analysis

Recent studies have found that disorders in fat metabolism play an important role in the initiation of cancer formation, and lipid-lowering drugs and anti-lipid treatments have shown potential in comparison to other cancer therapies that are highly toxic [68].

Further analysis of the presence of integrins in the KIRC network and their role in the disease is of particular interest. First, attention is brought to the HOXD3 gene which was found to be both differentially expressed and differentially methylated, and regulates 4 integrins (ITGB3, ITGAV, ITGB1, and ITGA5) in the KIRC network. Second, integrins are known to be important to the initiation, progression, and metastasis of tumours [32]. Furthermore, a 2006 study found that the overexpression of HOXD3 resulted in the increased expression of a type of integrin, leading to the enhanced motility and dissociation of human lung cancer cells [86]. More recently, it was found that HOXD3 was overexpressed in breast cancer tissue and had an important role in the ability of breast cancer cells to self-renew and differentiate by way of integrin mediated signaling [128]. In renal cell carcinomas, changes in integrin expression have been correlated with their degree of malignancy [57], and drugs specifically targeting integrins for renal cancers have been developed [93].

Because HOXD3 was found to be overexpressed and hypomethylated in the current study, it may be beneficial to investigate HOXD3 as an epigenetic therapeutic target for the improvement of KIRC therapies. Although hypermethylation is commonly
6.5. CONCLUSION

associated with cancers, hypomethylation may also play a role in cancer formation by inducing chromosomal instability and irregular gene expression [45]. Therefore, increasing the methylation of HOXD3 through DNMTs may be beneficial in treating KIRC. Increased methylation of DNA methyltransferase Dnmt3b targets for example, was found to impair cancer development in mice [102] and a similar approach may be advantageous in KIRC therapies.

6.5 Conclusion

In Chapter 4, a series of statistical analyses were performed in order to identify correlations between differential methylation and differential expression. 37 genes were identified that were found to be both differentially methylated and differentially expressed, and among them 5 genes were identified where differentially methylation and differential expression were correlated. In Chapter 5, a model, MInR, was developed to infer regulatory networks that addressed some of the limitations of existing models. MInR was found to outperform previous models when evaluated on the same benchmark *E. coli* dataset. In the current chapter, MInR was used to infer a regulatory network for KIRC, and the results of Chapter 4 were used to isolate a subnetwork. Further analysis of the subnetwork revealed genes that have previously been found to play a role in KIRC and may lead to potential drug targets. Although more research is needed to investigate these results, the use of MInR alongside statistical analyses shows promise as a tool in drug development.
Chapter 7

Conclusions

7.1 Summary

Interest in epigenetic drug research has grown in recent years due to the potential of epigenetic drugs to treat the misregulation of epigenetic mechanisms that have been linked to several diseases. The role of DNA methylation in cancer has been of particular interest. This has led to the development of a small number of epigenetic drugs. Unfortunately many of these drugs are limited by variable potency, instability \textit{in vivo}, and lack target specificity. This is largely due to a lack of thorough understanding of epigenetic mechanisms at the molecular level, especially pertaining to DNMTs which most epigenetic drugs are aimed to target. Because of this, many of the current computational methods in epigenetic drug discovery complement phases of drug discovery and development by exploring molecular modeling techniques to gain a better understanding of DNMTs as potential targets, screen small molecule libraries for candidate drugs, and predict potential molecular structures from existing data for further analysis.

Although these molecular modeling techniques have been found to be successful,
more research is needed on the integrative analysis of genetics, epigenetics, and their link with disease in order to elucidate novel epigenetic drug targets. The increased interest in epigenetic research has led to an increase in available genetic and epigenetic data, however few tools exist for their analysis. In this regard, a workflow was developed which included differential gene expression and differential methylation analysis to identify genes of interest in KIRC, a novel method, MInR, to infer GRNs, and finally the incorporation of differential gene expression and methylation information into a KIRC network created with MInR.

7.2 Discussion

This thesis described a workflow for the integrative analysis of gene expression and DNA methylation data, and presented the following contributions:

1. A genome-wide KIRC DNA methylation analysis correlated with gene expression using an interaction analysis strategy

2. A novel multi-label machine learning approach, supplemented by semi-supervised learning to infer regulatory networks

3. An inferred regulatory network for KIRC which incorporates differential DNA methylation and gene expression analysis. Analysis of this network reveals interesting information about KIRC which may lead to potential drug targets.

The correlation analyses between DNA methylation and gene expression in KIRC that was presented in Chapter 4 revealed 5 genes that have previously been identified to have a role in kidney cancer - an epigenetic disease. The goal of these analyses was to gain a better understanding of the link between methylation changes and gene
expression in KIRC, which is still poorly understood [100], and to help focus future research, such as the study presented in Chapter 6.

In Chapter 5 it was found that a machine learning technique, MInR, which is supplemented by semi-supervised label spreading to infer negative examples, had a better performance than previous methods that infer regulatory networks. Knowledge on transcriptional regulatory networks can help understand the underlying mechanisms of a disease and can be useful for identifying novel therapeutic targets [84]. MInR was evaluated using a benchmark *E. coli* dataset, and further investigation of MInR with more complex datasets, such as KIRC in Chapter 6, may be used in parallel with drug discovery and development to expedite the development of novel epigenetic drugs.

In Chapter 6, combining the analysis of correlation between DNA methylation and gene expression with a machine learning technique that is supplemented with a method to infer negative examples, was used to identify important genes in KIRC. MInR was used to infer a regulatory network for KIRC, and the results of the statistical analyses presented in Chapter 4 were used to isolate a subnetwork. Analysis of the subnetwork revealed genes that have previously been found to play a role in KIRC and, with further research, may lead to the identification of potential drug targets. Although more investigation is needed *in vitro* and *in vivo*, the use of MInR alongside statistical analyses shows promise as a tool in drug development.

Overall, the objectives of this study were met. A computational approach which provided insight on gene expression, epigenetics, and their link with a disease was explored. This computational approach was used to analyse gene expression and DNA methylation in KIRC, and a subnetwork that revealed systems that may play
a role in KIRC was identified. This network contained genes that have previously been found to be associated with cancer. The study therefore presented a successful workflow and offers a new tool in epigenetic drug discovery.

7.3 Future Work

This workflow offers a new tool that explores the association of gene expression and DNA methylation with a disease, and can be used to help identify potential drug targets in epigenetic drug research. This work may be expanded by incorporating molecular modeling techniques to gain a better understanding of the potential drug targets that were identified, and to explore potential drugs. These drugs and drug targets can then be validated in vitro, and if found successful, the workflow can be evaluated on other epigenetic diseases.

7.4 Conclusion

Computational approaches, may help provide insight on, and help gain a better understanding of biological systems, thereby narrowing the scope for drug discovery and development. Further investigation of computational approaches that complement phases of epigenetic drug discovery, such as those described in this thesis, show promise in the improvement of current epigenetic drugs and the development of novel therapeutics.
Bibliography


[34] N. Drakos and R. Moore. Support Vector Machines SVM, 1993-1999. [Online; Translation was initiated on 2016-08-09].


