Exploring gene expression and protein binding data for gene regulation

Submitted by

Mohsina Mahmuda Ferdous

For the degree of Doctor of Philosophy

of the

Department of Computer Science

Brunel University London

July 2016

Declaration

I, Mohsina Mahmuda Ferdous, hereby declare that this thesis and the work presented in it is entirely my own. Some of the works have been previously published; this has been mentioned in the thesis. Where I have consulted the work of others, this is always clearly stated.

I dedicate this thesis to parents,

Mohammad Abdul Mannan
and
Maksuda Begum

Abstract

Gene expression is a tightly controlled process that is regulated by the epigenetic modifications and a series of interactions between the genes and the proteins across the genome. High-throughput technologies such as microarray and chromatin immunoprecipitation technique followed by the next generation sequencing (ChIP-seq) have enabled researchers to investigate the gene expression profile of large of number of genes and the locations of protein bindings and different epigenetic events at the genome-wide scale. To understand the underlying complex mechanisms that regulate gene expression, the computational biology community has proposed many methodologies and tools over the years to integrate the protein binding data; obtained by ChIP-seq and the gene expression data; generated by microarray technology. However, the integrative analysis is still in its infancy. Effective models that capture the complex characteristics of ChIP-seq data and integrate dynamic interactions between gene expression and regulatory factors across different genomic features are still lacking.

This thesis aims to provide robust and reliable methodologies to enable investigation of the relationship between different regulatory mechanisms and gene expression that incorporate the advanced and improved results from the ChIP-seq data and the epigenetic phenomena that are closely related to gene regulation. Here, the Markov Random Field model has been adapted to analyse the binding regions of proteins and epigenetic markers using ChIP-Seq technology where the complex characteristics of the data such as spatial dependency, IP efficiency are taken into consideration while modelling the data and demonstrated how this model along with the pre-analysis steps can improve the binding results. Two models have been proposed where these results are then assimilated in the integrative analyses between ChIP-seq and the gene expression data. Several classification techniques are also included in one of the models to find the association between different epigenetic markers, proteins, genomic features and gene expression profile. The models have been applied to public datasets and the results have been validated. With the proposed models, it has been shown how the

dynamic interactions between the regulatory proteins and gene expression can be investigated by integrating sets of genes regulated at successive time-points and different biological or experimental conditions as well as protein binding profiles across the genome.

If either the gene expression or the protein binding data is missing as it is often the case, studying the relationship between regulatory factors and gene expression with these models will help the biologists estimate gene expression from the available epigenetics data or assume the underlying epigenetics from the available gene expression data. In short, this thesis brings together different biological tools, data processing techniques, advanced machine learning techniques to make a systematic approach to advancing the state of the art in this important epigenetic field.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgements

First of all, I thank God for His blessings throughout my life to become the person that I am today. It was His guidance that made it possible for me to complete this work.

I would like to express my heartiest gratitude to my supervisor Xiaohui Liu, for his affectionate supervision, assistance, motivation and encouragement throughout this work. His methodology and patience have left a deep impression upon me. I would also like to thank David Gilbert and Paul Wilson for their guidance and would like to gratefully mention Veronica Vinciotti and Yanchun Bao for their advices and support throughout this PhD.

I would like to thank my family and friends for their support through my years as a PhD student. Very special thanks go to my husband Dr. M Hasan Shaheed who always encouraged me to get a PhD degree and it was his support that got me through the difficult days when I thought I could not do it. I would also like to thank my parents for not only being the constant source of support but also for taking care of my son so that I could finish the study. I would like to thank my sisters, Mity and Mahdia, for being there for me whenever I needed them. I would like to thank my daughter, Nuaimah, for being the joyful distraction from work and my son, Umair, who has made the biggest sacrifice for this PhD, staying with my parents since his birth five thousand miles away.

I also thank all my colleagues for their support including Neda Trifonova, Valeria Bo, Fadra Hassan, Djbreel Kaba, Miqing Li, Liang Hu, Izaz Rahman, Chuang Wang, Ali Tahrini and Khalid Eltayef.

Finally, I would like to express my gratitude to the EPSRC and GlaxoSmithKline, who through their funding have made this research possible.

Supporting publication

1. Ferdous M.M., Vinciotti V., Liu X., Wilson P. (2015) Exploring the link between gene expression and protein binding by integrating mRNA microarray and ChIP-Seq data. Statistical Learning and Data Sciences, Lecture Notes in Computer Science Volume 9047, pp 214-222

2. Ferdous M.M. (2016) Modelling ChIP-Seq Data using Markov Random Field model. Brunel Doctoral Consortium. (Awarded as the best paper)

3. Ferdous M.M., Vinciotti V., Liu X., Wilson P. "Prediction of underlying gene expression variance using genome-wide protein binding profile". (To be submitted in BMC Bioinformatics)

Chapter 1

Introduction

1.1 Background

In 1953 when American biologist James Watson and English physicist Francis Crick declared in a Cambridge pub that they had 'found the secret of life', their claim wasn't far from the truth. They indeed had solved the mystery of science of how genetic instructions were stored in any organism and transferred from one generation to another by discovering the structure of DNA [Watson et al. 1953]. DNA or deoxyribonucleic acid is the chemical compound that contains four basic building blocks or bases namely: adenine (A), cytosine (C), guanine (G) and thymine (T). The orders or the sequences of these bases form the instructions for making all the essential proteins in our bodies needed for the development of all living organism. These proteins perform essential functions in our body as enzymes, hormones and receptors. An organism's complete set of DNA is called its genome. Figure 1.1 demonstrates the structure of DNA and its components.

However, later with the flourish of the new science of epigenetics, researchers have realised DNA sequence is not the only factor that controls our biological make-up and in addition to nature and nurture, what makes us who we are is also determined by some tightly regulated chemical reactions that can switch parts of the genome off and on at strategic times and locations. These parts of the genome are genes that contain instructions to synthesise the gene products, typically proteins. The process in which information of genes is used to synthesis of these gene products is called gene expression. The chemical reactions mentioned above and bindings of regulatory proteins or transcription factors (TFs) occur at specific sequences of DNA to control gene expression so that the exact amount of proteins is produced when they are needed. Epigenetics is the study of these reactions and the factors that influences gene activity but does not involve a change in underlying DNA sequence.

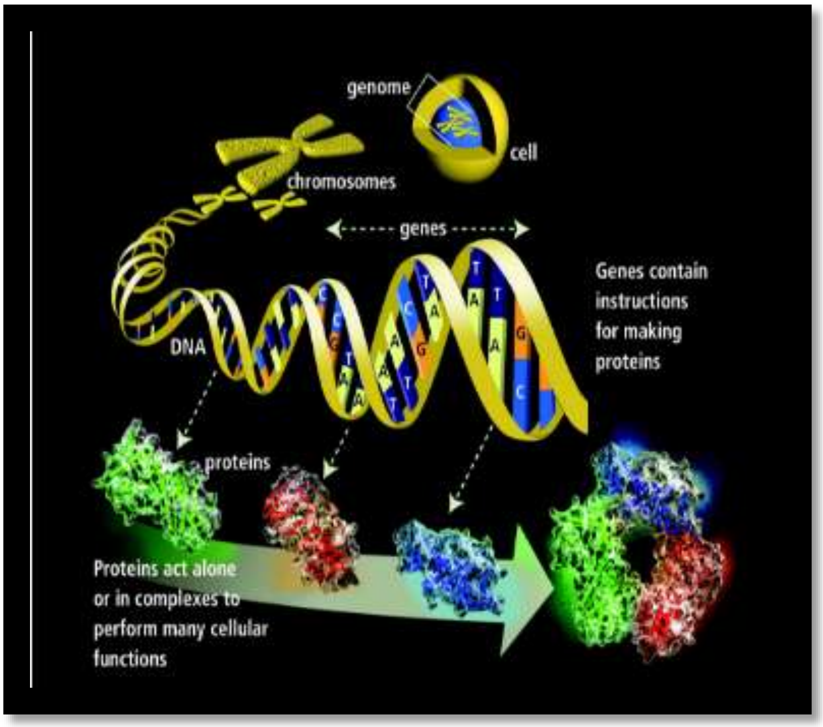

genome
cell
chromosomes
genes
Genes contain instructions for making proteins
DNA
proteins
Proteins act alone or in complexes to perform many cellular functions

Figure 1.1: The structure of DNA and its components. Gene is a unit of heredity which is composed of DNA occupying a fixed position on a chromosome that holds the instructions for creating proteins. Genome is defined as a group of all genes comprising of a set of chromosomes [Cheng 2006].

The transcription of the gene specifying a particular protein is a tightly controlled and complex process that intimately occurs in a context. To understand the process one must investigate what role the context plays in this process. The discovery of the complexity of the regulation mechanisms of gene expression has led the scientists to review their definition of gene and it is no longer viewed as a solo well-defined unit of DNA that contains specific information that is translated into proteins [Michel, 2010]. It is now recognised that all the developmental works in our body do not just rely on genes for protein production, rather the mechanism is much more complicated. A complex set of interactions between genes, RNA molecules, protein (including transcription factors) as well as the interactions of genes with their proximal and distal environments [Wright 1968] determine when and where specific genes are activated and the amount of protein or RNA products is produced.

However, for these interactions to happen, first the specific DNA sequence or gene needs to be opened up, which otherwise remains in an inactive state because it is tightly wound up in a structure called chromatin. Different chemical modifications also known as epigenetic mechanisms such as histone modification, DNA methylation and acetylation can alter the chromatin structure and make it accessible or inaccessible for transcription. These mechanisms occur at specific locations of the genome and these regions play important roles in gene regulation too.

Structural genes that code for proteins involve several different components such as introns and exons. Introns are the portions of the gene that do not code for amino acids and exons are the portions that do and also collectively determine the amino acid sequence of the protein product. There are also regulatory regions of the gene, such as, transcription start site, promoter, enhancer and silencer etc. These are the regions where different proteins bind and chemical modifications occur to interact with the genes to control transcriptional activity. Figure 1.2 shows the structure of eukaryotic gene with different regulatory element. Therefore different epigenetic mechanisms and other proteins or transcription factor binding patterns around these regions are of interest to the researchers to figure out which regions are important for the gene regulation.

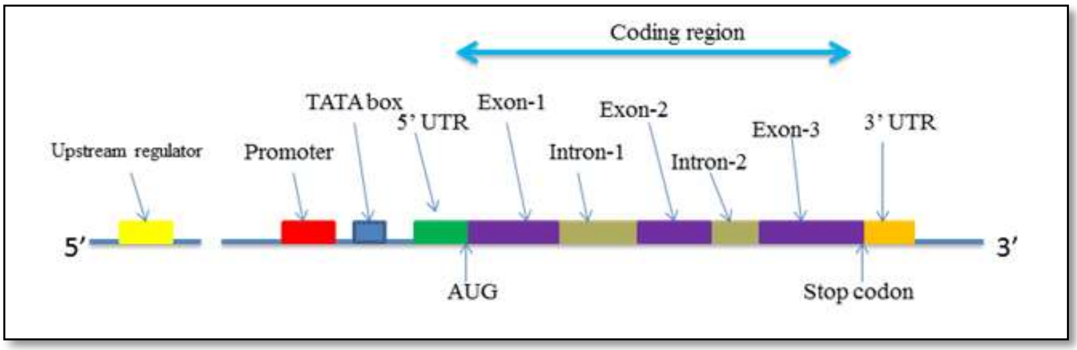


Figure 1.2: Eukaryotic Gene Structure with its component such as promoters, exons, introns etc [Eukaryotic gene structure].

Understanding and investigating all these epigenetic factors that regulate genes are critical in unravelling the complexity of various biological processes. Undue disruption of these processes can also lead to many diseases, some of which are life threatening. Therefore it is absolutely vital to study how genes are regulated and what controls gene expression. And that is what has made epigenetics a subject that is undergoing intense study among scientists. There are many technologies available today for studying different epigenetic mechanisms and gene expression. Two such high-throughput biological technologies are, microarray, which measures the expression level of large number of genes simultaneously, and chromatin immunoprecipitation technique followed by next generation sequencing (ChIP-Seq), which investigates the locations of proteins or transcriptions factors bindings and epigenetic modifications across genome.

In ChIP-Seq technology, a protein of interest is usually cross-linked with DNA site it binds to in an in vivo environment using formaldehyde. After the crosslinking is done, then the DNA is sheared by sonication or other mechanism. The next step is immunoprecipitation. From the resulting DNA strands and Protein of interest and DNA component, crosslinked DNAs are filtered out with antibody by the immunoprecipitation technique. Once the enrichment is convincing, the material is ready to be sequenced. The cross-linking of the protein and DNA is reversed and the DNA is purified and sequenced. These sequences are then further analysed to find the genomic locations that are bound by the protein under study.

The microarray experiments that analyse expression levels of selected gene involve the hybridization of an mRNA molecule to the DNA template from which it is originated. In this technique, an array is used where thousands of spotted samples known as probes are immobilized on a solid support, typically a microscope glass slide. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. Finally the data is collected and processed to generate a profile for gene expression. Both DNA microarray and ChIP-Seq have become indispensable tools in genome research as they both immensely help find out structural and functional characteristics of different genomes.

Next generation sequencing, no doubt has several advantages over microarray analysis, but microarray has its advantages too, which still makes it desirable for many studies. Microarray is an established tool with its mature analysis pipeline and it is a comparatively low cost experiment too. However, with microarray detailed underlying epigenetic landscape cannot be determined. On the contrary, ChIP-Seq offers detailed characterization of various types of chromatin marks on a genome-wide scale, but ChIP-Seq experiments are very costly and the analysis techniques are still evolving. Therefore, one might think that microarray will soon be replaced by these new sequencing technologies, experts rather think the cost-effectiveness and simplicity will play in its favour. Some also have suggested that microarray and ChIP-Seq should be integrated to study the gene regulation pattern and investigate whether microarray data alone can be used to predict underlying epigenetics. Experts have predicted that in the near future, these two technologies may complement each other and form a symbolic relationship [Hurd et al. 2009]. Integration of the result of these two technologies is biologically very significant as it enables the investigators to study how different epigenetic modifications and protein bindings are occurring across genome to control gene expression.

The integration techniques for both technologies are still at its infancy and researchers are working relentlessly to come up with different methodologies so that robust information can be achieved from such study. With the dawn of ChIP-Seq technology, researchers have begun to unravel how different epigenetic mechanisms and bindings of regulatory proteins work together to regulate genes. This has opened up possibility for not only getting new insights into the functional genomics of every living cell but also discovering drugs and treatments to diseases that are caused by disruption of normal regulation process. However, this exciting technological infancy comes at a price too.

ChIP-Seq data has very complex characteristics. To get robust information about protein binding locations, these characteristics need to be considered while modelling ChIP-Seq data. However, most of the integration methodologies of ChIP-Seq and gene expression data, to date, have used very basic analysis steps which may not capture all the information this next generation sequencing technology can offer.

With the advancements of different genome projects, more and more genomic locations are identified and annotated and rich datasets are produced. This progress enables researchers to investigate what roles different biological conditions such as treatment, non-treatment, time factors and also different genomic locations play role in gene regulation. There are still gaps in the literature where all these information are incorporated into the methodology to find the relationship between gene expression and epigenetic mechanisms. In most of the integrative study, protein binding at very common genomic locations such as promoters and transcription start sites are investigated, whereas experts have discovered that other genomic features underlie epigenetics too [Nott et al. 2003; Heyn et al. 2014].

Here in this thesis, the focus is on improved results from ChIP-Seq data and integrating the results of microarray and ChIP-Seq to find the relationship between gene regulation and epigenetic mechanisms. With this in mind, different methodologies have been proposed to study such relationship where advanced analysis techniques of ChIP-Seq data, proteins bindings at different genomic features, different biological conditions and time-factors relevant to underlying epigenetics are incorporated effectively.

1.2 The aim and objectives

The main aim of this project has been to search for effective ways of integrating protein binding and gene expression data to understand the underlying epigenetic mechanisms that regulate gene expression.

When this work began, the research community had already been excited about the ChIP-Seq technology and its potential to uncover underlying epigenetics. However as more datasets were made publicly available and genomic databases were updated, this field showed further potential for advancements to be made in the integrative analysis between ChIP-Seq and microarray data. The project started with the primary aim in mind that was to use advanced computational methodology to analyse comparable datasets from both technologies, microarray and ChIP-Seq obtained in the same biological settings and investigate how to effectively integrate both data to describe different epigenetic events that regulates gene expression.

The main objectives of the thesis are as follows:

1. Acquiring comparable datasets from both technologies obtained in the same biological settings.
2. Exploring the complex characteristics of ChIP-Seq data to find the most appropriate means for data pre-processing and effective modelling
3. Investigating techniques that integrate protein binding and gene expression data to uncover hidden relationships between them.
4. Understanding whether protein binding profile across genome can be predictive of gene expression changes thus finding associations between different epigenetic events and gene regulation.

1.3 Contribution to Knowledge

In this thesis, ways of modelling ChIP-Seq data have been explored where different characteristics of such data are taken into consideration. The Markov Random Field (MRF) model has been adapted for the analysis of ChIP-Seq data and comparative performance analysis has been carried out between the MRF model and other existing methods. One of the characteristics of ChIP-seq count data is that it is known to have spatial dependencies between regions. The reason is that a common pre-processing step to create count data is to divide the genome into fixed length windows and the count of sequences are summarised per window. As a result the bound regions can cross between two or more windows and that introduces spatial dependencies in the data. Another important characteristic of ChIP-Seq data is IP efficiency. The degree of enrichment found from the data depends on ChIP-efficiency or otherwise known as IP efficiency that means an efficient experiment will produce better signal to noise ratio than a less efficient one. The quality of antibodies plays an important role to determine the quality of the data. ChIP-efficiency also varies between data generated in different lab or experiment. Therefore when differential regions are sought between two experimental data, not considering efficiency may lead to over or under estimating the regions. These characteristics such as spatial dependency and IP efficiency have been incorporated in the ChIP-Seq analysis in this thesis which shown to improve the result. Pre-processing steps of the ChIP-Seq data before running statistical analysis to find the protein binding locations have been demonstrated to have great impact on the overall performances. It has been shown how the count correction step can further improve the results. The results have been validated using known biological information.

The next step in the study has been to find the effective way of integrating ChIP-Seq protein binding result with complementary microarray expression data. As most of the integrative analyses of these two ignore many important characteristics of ChIP-Seq data, I have proposed a method to integrate these characteristics of ChIP-Seq with the MRF model first. In this model the correlation between the differential binding probabilities for different proteins around transcription start sites (TSSs), estimated by the MRF model and microarray differential gene expression values associated with those TSSs, are investigated together. Using enrichment probabilities directly has the advantages of capturing many characteristics of ChIP-Seq data as opposed to using count data directly. Also the technique incorporates different biological conditions and time factors; therefore it can be applied to rich dataset that includes such variables. I have validated our results on the proteins investigated with known biological information in the field.

A novel approach is then proposed to investigate advanced machine learning techniques to find relationships between gene expression and protein binding profile across a genome where different genomic features such as exons, introns, distal intergenic along with promoters are integrated. It has been explored how predictive the binding profile of the proteins of interest at different features is of gene activity using several classification techniques such as neural network, decision tree and random forests. Other biological conditions such as treatment, non-treatment, time factors are also included in the model. Feature selection techniques by decision tree and random forests have identified important proteins, features and biological factors that mostly correlate with gene regulation among all the variables. Comparative analysis of different classifiers is also conducted to determine which one performs best at detecting potential relationships. It is anticipated that this study will provide a foundation for further opportunities for finding association between protein binding and gene expression where it can be investigated thoroughly how different proteins binding at different genomic features and the other factors such as time play role in the gene regulation mechanism.

Although the methods described above have been applied to a time-series ChIP-Seq data of six proteins and microarray data that are obtained with same biological conditions such as treatment and control, the methods can be applied on new datasets involving any number of proteins and biological conditions. It is believed with richer datasets the underlying epigenetic factors that are regulating gene expression are likely to be more apparent and also with this technique, genomic features other than promoters and TSS that are commonly used can be investigated for their roles in gene regulation. In future as genomic databases are updated and new annotations of more genomic feature are made available, this technique can help investigate their functionality in our biological process.

Major contributions of the thesis can be summarised as follows:

- Important insights have been obtained on how data pre-processing, particularly how to prepare the count data of ChIP-Seq experiments, can further improve the analysis results. (Chapter 3)

- The MRF model has been adapted for the analysis of real ChIP-Seq data and a comparative study has been conducted between this method and other existing algorithms to understand its strengths and weaknesses. (Chapter 3)

- A novel approach has been proposed where advanced analysis result of ChIP-Seq are incorporated in integrative analysis of protein binding and gene expression data to study the relationship between differential expressions and differential protein bindings around the transcription start site. In this approach it has been demonstrated how enrichment probabilities estimated by an advanced method that first incorporates all the characteristics can be generated around any genomic location to study its role in gene regulation. Different biological conditions and time factors are also included in the model. (Chapter 4)

- A methodology has been proposed to investigate how predictive the binding profile of different regulatory proteins at different genomic locations across genome is of gene activity. Experiments have been conducted using advanced machine learning technique to perform predictive analysis using protein binding profiles as a predictor and gene expression responses as a response to study which proteins at any binding location can best predict the gene status. (Chapter 5)

- It has been shown how dynamic interactions between regulatory proteins and gene expression may be explained by integrating sets of genes regulated at successive time-points and different biological or experimental conditions. This technique may help answer not only what proteins might be regulating genes but also where, when and at what condition they bind to do so. Comparative analysis between the classifiers has also been performed and the results are documented. (Chapter 5)

1.4 Roadmap to the thesis

The thesis is arranged as follows.

Chapter 2 provides general background knowledge relevant to each chapter of the thesis. It briefly introduces epigenetics, different epigenetic mechanisms and technologies that are currently aiding epigenetic studies. Mainly I focus on microarray and ChIP-Seq technology which are extensively used in this project to study biological phenomena such as gene expression changes and protein binding and also to investigate the relationship between them.

Chapter 3 gives some background information about the pipeline of ChIP-Seq analysis and how different characteristics of ChIP-Seq affect the peak results. Different techniques including those for data pre-processing are introduced in this chapter to obtain robust results from ChIP-Seq data. A MRF model is adapted for the analysis of the ChIP-Seq data used in this project. Experiments detailing the parameter sensitivity on the results obtained and those comparing this method with other existing methods are reported.

Chapter 4 presents a methodology that shows how ChIP-Seq data can be analysed using advanced methods that deal with different characteristics i.e. spatial dependency, overall distribution of the data and how this information in terms of enrichment probability can be incorporated in the integrative analysis of protein binding and gene expression data. It has also been shown how differential expression and differential bindings can be investigated around any genomic locations between different conditions such as time, treatment/non-treatment.

Chapter 5 reports a novel approach where the integrative analysis of protein binding and gene expression data incorporates binding locations at different genomic features such as exon, intron, promoters, distal intergenic region etc. and also other biological conditions such as treatment/non-treatment, time factors etc. The method and results show how dynamic interactions between regulatory proteins and gene expression can be explained by running predictive analysis on protein binding profiles across genome and complementary gene expression results. Several classification techniques, such as neural network, decision tree and random forest have been explored to find such associations.

Chapter 6 provides discussions of the work proposed in this thesis and highlights future research directions currently under investigation.

Chapter 2

Background

In this chapter many of the relevant terminologies, technologies and their respective associated analysis techniques will be briefly introduced. It is not in the scope of this chapter to present an overall review of the field, rather it is a concise introduction of the key biological concepts of the respective relevant subjects to enable an appreciation of the key concepts of the respective 'omics' technologies . However, references will be given throughout the chapter so that anyone who is interested can investigate further.

This project is focused on studying the relationship between epigenetic mechanisms and gene expression using different technologies such as microarray and ChIP-seq. This chapter will start with brief introduction of epigenetics, followed by description of those technologies and the available analysis techniques.

2.1 A brief Introduction to Epigenetics

In the 1950s and 60s when the genetic code and the structure of the genes were unravelled, scientists began to see genes as a collection of blueprints for proteins that are essential for the development and maintenance of any organism. Genes, can be conceptually thought of as a string of DNA that is capable of producing chains of amino acids that fold to from functional proteins, Some genes are constitutively active or 'on' regardless of organism's environment carrying out essential functions for our body, however, not all genes are always on or expressed to produce proteins and they only become active by some tightly regulated mechanisms when it is necessary for any specific biological process [Hoopes 2008]. Different chemical reactions and bindings of regulatory proteins at different genomic locations work together to turn the genes on or off at strategic times and locations and control the gene expression mechanism so that our body can have the right amount proteins when they are required. Epigenetics is the study of all mechanisms that control gene expression levels and the factors that influence them. Genetics and developmental biology were perceived as two separate research areas in the past. Developmental biologists or embryologists did not have much interest in genes and their roles; towards the middle of the twentieth century some leading biologists were waking up to the notion that these two fields were actually linked. Waddington being an expert on both fields defined epigenetics linking developmental biology and genetics together [Holliday, 2006]. In 1942, Conrad Waddington defined the term epigenetics as "the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being". The first concepts of epigenetics can be dated back as far as Aristotle (384-322 BC) and it continued as a conceptual theme through to the mid-19th century. However, slowly epigenetics has emerged to bridge the gap between nature and nurture. Today the most common definition describes epigenetics as 'the study of heritable changes in genome function that occur without a change in DNA sequence' [Riggs et al. 1996]. More recently however, Berger et al. [2009] has added a constraint to the definition that the initiation of the new epigenetic state should involve a transient mechanism separate from the one required to maintain it.

2.1.1 Epigenetics events

Deoxyribonucleic acid or DNA is the hereditary material, found in almost every organism. The structure of DNA is very complex and large and it is composed of several building blocks called nucleotides. In order to take less space in the cell, DNA is wrapped around histone proteins in repeating units of nucleosomes to form a structure known as chromatin [Campos et al. 2009; Fedorova et al. 2008]. This structure provides the first level of compaction of DNA into the nucleus. To achieve higher level of compaction, nucleosomes are sometimes spaced along the genome to form a nucleofilament that finally results in the highly condensed metaphase chromosome and chromatin is organized into functional territories within an interphase nucleus.

Change in the structure of chromatin plays a crucial role in whether transcription is allowed which is basically the first step of gene expression mechanism. In gene transcription process, a particular segment of DNA is copied into RNA (mRNA) by the enzyme RNA polymerase. Both RNA and DNA are nucleic acids, which use base pairs of nucleotides as a complementary language. Chromatin adopts different conformation in different contexts, for example; in different cell types. In simple words, chromatin can be in the open form that allows access of the machinery for transcription, and a closed form which does not allow transcription. Therefore as alteration of chromatin structure control gene expression, the events that are responsible for such alteration can be considered epigenetic events. Figure 2.1 shows how chromatin can be in open or closed form to control gene expression.

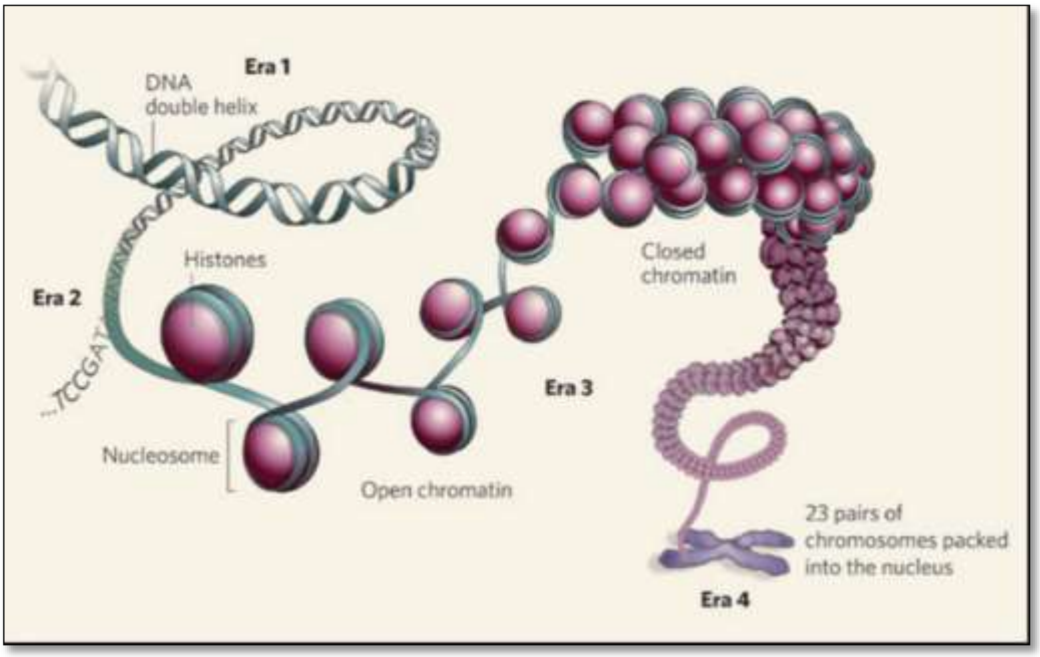


Figure 2.1 Chromatin remains in tight structure not to allow transcription and it opens up to initiate transcription [Carmona 2015]

When a specific sequence of DNA or gene is compactly bound with histone, that gene remain inactive or "off." However, the area where transcription should occur needs to be unbound or open before transcription process can start. This is a very multifaceted process that requires coordination of many mechanisms such as, histone modifications, transcription factor binding and other chromatin remodelling activities. Histone modifications is the chemical modification of the histones' NH_2-terminal tails and as a result of such change, DNA becomes unwound which allows its access to the transcriptional machinery [Karlic et al, 2010].

The family of the enzymes	**Type of epigenetic modification**	**Effect on gene expression**
DNMT1, DNMT3L, DNMT3A, DNMT3B (DNA methyltransferase)	Maintenance and de novo DNA methylation	Gene expression suppression
TET family (ten eleven translocation)	DNA demethylation	Induction of gene expression
IDH family (isocitrate dehydrogenase)	DNA demethylation	Induction of gene expression
HMTs (Histone methyltransferase)	Methylation of lysine in histone protein	H3k4me3 → transcription activation H3K9me or H3K27me → transcription repression
HDMs Histone demethylase)	Demethylation of lysine in histone protein	Transcription activation or repression based on the lysine residue
HATs (Histone acetyletransferase)	Histone acetylation	Transcription repression
HDACs classes I-IV	Histone deacetylation	Transcription repression

Table 2.1: Some important enzymes along with the types of modifications they cause and their effects on gene expression.

The chemical modification of the histone proteins can be caused by methylation and acetylation. The common types of such chemical modifications, enzymes involved in such modifications and their effect on gene regulation have been summarised in Table 2.1 [Abdolmaleky et al. 2013]. DNA methylation is a biochemical process that also forms the basis of chromatin structure, which enables a single cell to grow into different organs or different tissues. This DNA methylation process is important for the regulation of cellular differentiation and development, and can also serve as a biomarker for several human diseases. The important role of DNA methylation or demethylation in developmental biology was first proposed in 1969 [Grifith et al. 1969]. Scientists then suggested that DNA methylation can affect gene expression [Riggs et al. 1975; Holliday et al. 1975]. This mechanism usually appears to be coordinated with histone modifications, particularly those that lead to silencing of gene expression. However, when the tails of histone molecules are acetylated it removes positive charges, thereby reducing the affinity between histones and DNA thereby leaving it more open. In most case histone acetylation enhances transcription. Transcriptionally active, “open” chromatin generally has hyperacetylated and hypomethylated histones, whereas more inactive heterochromatin tends to be hypoacetylated and hypermethylated [Wild et al. 2010].

In addition to DNA methylation and histone modifications there are other mechanisms which also affect gene expression. For example, Eukaryotic genomes transcribe large numbers of RNAs that have no coding capacity. These noncoding RNAs include miRNA, piRNA etc. Chromosomal regions that are located far from each other can interact, in effect leading to the alternation of gene expression. This type of direct interaction can contribute to gene activation or repression by facilitating regulatory elements, influencing transcriptional state of associated genes [Grimaud et al. 2006; Lonvardas et al. 2006]. Therefore; the interactions between these chromosomal regions can be termed as epigenetic mechanisms. Figure 2.2 shows different epigenetic mechanisms. For example if DNA methylation Methyl marks added to certain DNA bases, it can repress gene activity, while histone modifications refer to covalent post-translational modification of N-terminal tails of four core histones (H3, H4, H2A and H2B).

Eukaryotic genomes also transcribe large numbers of RNAs that have no coding capacity.

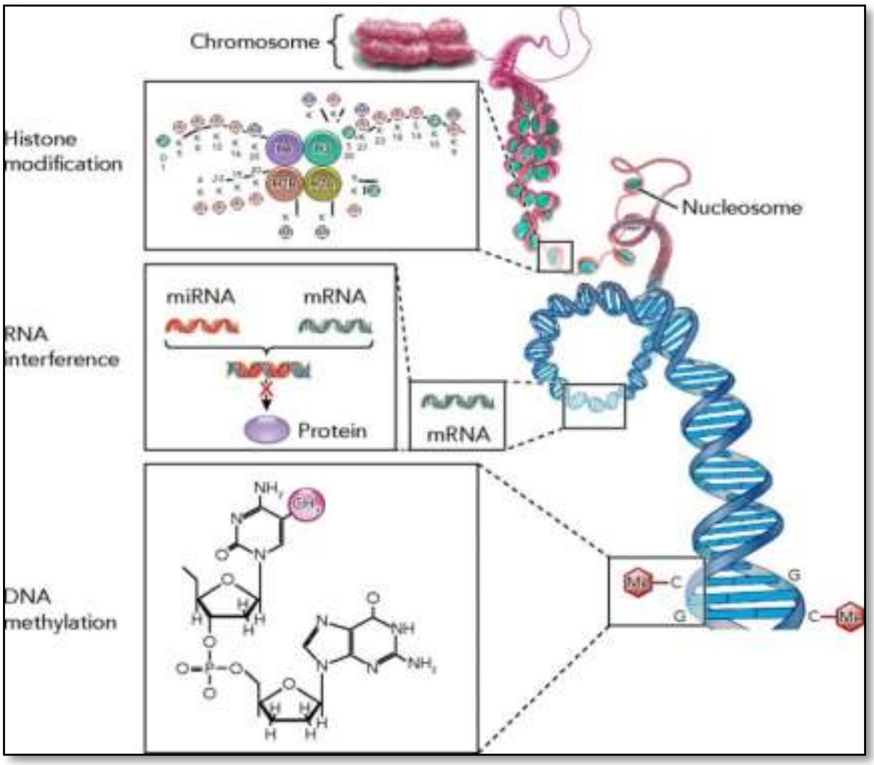


Figure 2.2 Schematic of different epigenetic mechanisms. Histone modification is the process in which several types of modification occur to amino terminals of the core histones to initiate transcription. DNA methylation is process by which methyl groups are added to DNA to regulate gene expression. RNA mediated gene silencing mechanisms also regulate genes [Hagood 2014].

Once the chromatin is in open form, specific DNA sequences are then accessible for specific proteins to bind. These proteins then act as activators or repressors for the genes and control gene expression. For a TF that is an activator, the effector region recruits RNA polymerase II which is the eukaryotic mRNA-producing polymerase that initiates transcription of any corresponding gene. These regulatory proteins bind at different locations of the genome, (i.e promoters, that resides just upstream of eukaryotic genes or enhancers, which can be oriented forward or backwards and are found upstream or downstream of transcription start sites) and activate gene expression. TFs have been observed to concurrently activate and repress multiple genes simultaneously.

Many genes are regulated together therefore studying gene expression across the whole genome via microarrays or massively parallel sequencing allows investigators to see which groups of genes are co-regulated given any particular biological state. Investigating the pattern of epigenetic mechanisms and regulatory proteins bindings across genome with next generation sequencing coupled with the gene expression study can tell us how exactly these genes are regulated.

2.1.2 Why Study Epigenetics

A eukaryotic cell requires different proteins in defined concentration at different times. That is why gene expression is one of the most tightly controlled processes in the body as any disruption to this protein making process can lead to serious consequences including disease conditions. Therefore it is absolutely vital to study how the genes are regulated and what controls gene expression which has made epigenetics such an interesting topic among scientists. Epigenetic changes are absolutely vital for our normal and healthy development; however they can also be the cause for many disease states. If normal epigenetic alterations of any of the systems that contribute to gene regulation is disrupted, that can be fatal and cause abnormal activation or silencing of genes. Such disruptions have been associated with many life-threatening diseases such as, cancer, syndromes involving chromosomal instabilities, and mental retardation [Portela et al. 2010]. By studying these epigenetic mechanisms one can understand how, why or where these changes are happening, what diseases they are causing, etc.

Cancer was the first human disease to be linked to epigenetics. Studies performed by [Feinberg et al. 1983], using primary human tumour tissues, found that genes of colorectal cancer cells were substantially hypomethylated compared with normal tissues. Another example can be given about prostate tumour where the enzymes that modify histones behave differently as tumours progress. Scientists can better understand potential disease conditions by looking at the way histone tails have been systematically modified in tumours from different patients. Apparently patterns of global histone modification can serve as an indicator for the future course of disease. Such epigenetic profiling of cancers, coupled with our knowledge of functional mutations, could pave the way for personalising cancer treatments in the near future.

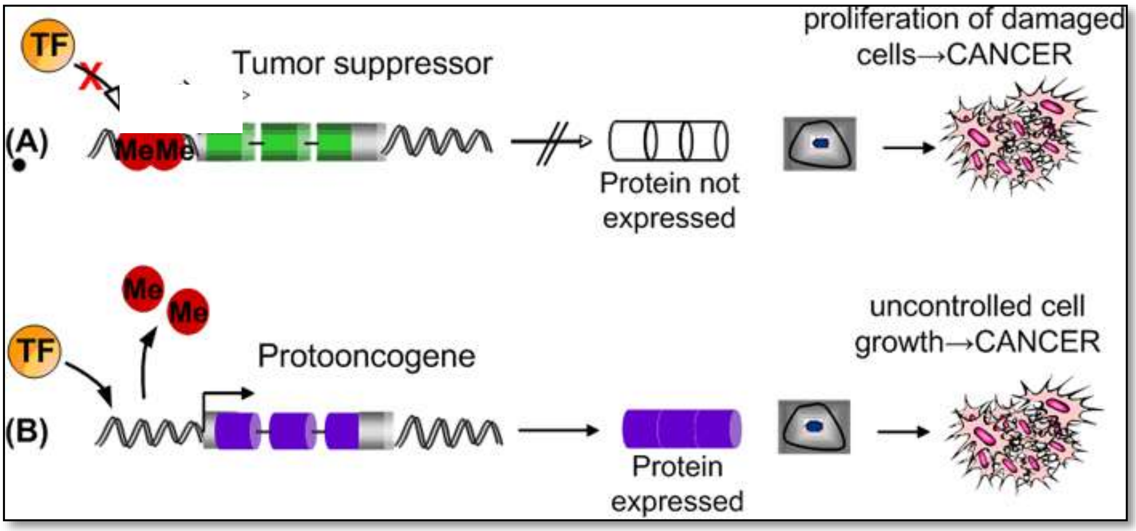


Figure 2.3 Possible mechanisms by which epigenetic modification can lead to cancer. (A) An undue methylation of a gene can cause disruption to transcription. As a result cells can be damaged and become cancerous. (B) A gene can also be demethylated when it is not required and the demethylation can initiate transcription and cause unnatural cell growth [Nelson 2008].

Figure 2.3 shows how epigenetic modification can lead to cancer, for example a previously unmethylated TS gene can be methylated and thus transcription factor(s) (TF) can no longer bind the promoter region, as a result the gene is not expressed, and damaged cells are allowed to proliferate and become cancerous. In other occasions, if a proto-oncogene can be demethylated, allowing TFs to initiate transcription and express the protein product, which can also lead to unnatural cell growth and cancer.

As many diseases are related to epigenetic changes, researchers are investigating if it is possible to counteract these modifications with epigenetic treatments. The most popular of these treatments aim to alter either DNA methylation or histone acetylation. Furthermore, epigenetic behaviours are understood to be reversible and therefore provide opportunities for novel therapeutic intervention in a number of chronic inflammatory diseases.

In Epigenetics studies, there are a number of issues that must be considered. Firstly, in genetic studies, scientists can collect DNA sample from any tissues and analyse them, however epigenetics studies are different in that respect. As epigenetic profiles may vary depending on the cell types, scientists need to collect samples from tissues and organs that are relevant to the phenotype of interest. For example, in order to study inflammatory bowel disease, samples must be collected from gut. Secondly the relationship between epigenetics and phenotype are not always straightforward, however, studying tissues of affected and unaffected subjects and keeping the study perspective may help identify the differences between causal associations and non-causal associations [Petronis 2010]. Currently there are many technologies are available for a close study of these relationships.

2.1.3 Technologies helping study Epigenetics

Epigenetics research continues apace in labs investigating a dazzling variety of topics. Many Bioinformatics tools have been proposed along with different experimental methodologies to analyse the epigenetic mechanisms [Bock et al. 2008; Lim et al. 2010; Laird 2010].One interesting direction is the application of high-throughput sequencing technologies to the characterization of hundreds of ‘epigenomes’ (epigenetic marks across the entire genome). Patterns of DNA methylation, six histone modifications, couple with gene activation from various normal and diseased cell types can serve as a baseline in many studies to identify changes associated with specific diseases.

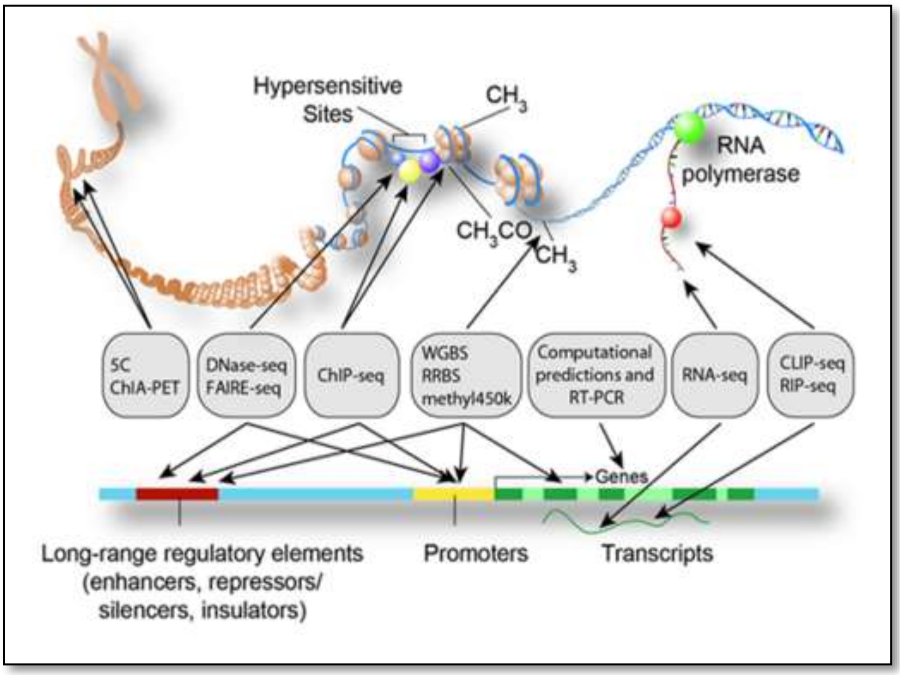


Figure 2.4: Some technologies available to investigate different epigenetic mechanisms [Technologies for studying epigenetics].

There are many technologies available for studying epigenetic modifications and gene expression. Figure 2.4 summarises several technologies that are used to investigate different epigenetic mechanisms. For example, methylation-specific PCR(MSP) provides the test for the methylation status of CpG dinucleotides in a CpG island making the technique applicable for high throughput analysis of clinical samples [Herman et al. 1996; Shanmuganathan et al. 2013; Wani et al. 2016], whole genome bisulfite sequencing enables differentially methylated sites to be detected on the genome at single nucleotide resolution [Frommer 1992], chromatin immunoprecipitation technique such as ChIP-on-chip is a microarray method that reveals the genome-wide location of DNA-bound proteins [Ren 2000] and MeDIP-seq [Jacinto et al. 2008; Down et al. 2008] is another technology available that can be used to detect or analyse DNA methylation. Microarray technology, which measures the expression level of large number of genes simultaneously, has been an established platform for studying epigenetic analysis for a long time now. ChIP-Seq, which is comparatively a new technology, produces DNA sequences that are bound by a protein of interest or other cellular markers. It offers high resolution mapping of TFs or epigenetic modifications' interaction sites to genomic locations [Furey 2012]. It is now an indispensable tool in medical and biological fields. As Microarray and ChIP-Seq are the two main technologies used in this thesis to analyse the relationship between gene expression and protein binding. A brief introduction of both technologies is given below.

2.2 A Brief Introduction to ChIP-Seq technology

Chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing (ChIP-Seq) is a relatively new technology to map genome-wide protein-DNA interaction. It has been extensively used for analysing how protein interacts with DNA and also the binding sites of DNA-associated proteins. In order to fully understand the biological processes and many disease states it is essential to understand how proteins interact with DNA to regulate gene expression. With ChIP-Seq technology, it is possible to determine how transcription factors and other chromatin associated proteins influence phenotype-affecting mechanisms.

Chromatin immunoprecipitation technique can isolate specific DNA binding sites that are in direct physical contact with transcriptional factor and other proteins. This produces a library of target DNA sites bound to protein under study in-vivo. [Gilmour et al. 1986] first developed the original ChIP technique, where they used UV irradiation to covalently cross-link proteins in contact with neighbouring DNA in intact living cells. Subsequently [Solomon et al. 1988] adapted formaldehyde cross-link replacing the UV cross-link technique.

This ChIP assay can then be combined with sequencing technology (ChIP-seq) to examine the interaction pattern of any Protein with DNA or the pattern of any epigenetic chromatin modifications. First genome-wide maps produced through ChIP-Seq were created in 2007 [Johnson et al. 2007]. Further studies suggested novel functions for histone modification and the importance of combinatorial patterns of modifications [Barski et al. 2007] and examines the correlations among histone modification patterns and their relationship to transcriptional activation [Wang et al. 2008].

2.2.1 How ChIP-Seq technology works

In this technique, a protein of interest is cross-linked with DNA site it binds to in an in vivo environment. Then the DNA is sheared by sonication or other mechanism. After fragmentation, the next step is immunoprecipitation. From the resulting DNA strands and Protein of interest and DNA component, crosslinked DNAs are filtered using antibody by the immunoprecipitation technique. Finally the cross-linking of the protein and DNA is reversed and the DNA is purified. These DNAs are then sequenced, which are known as 'reads'.

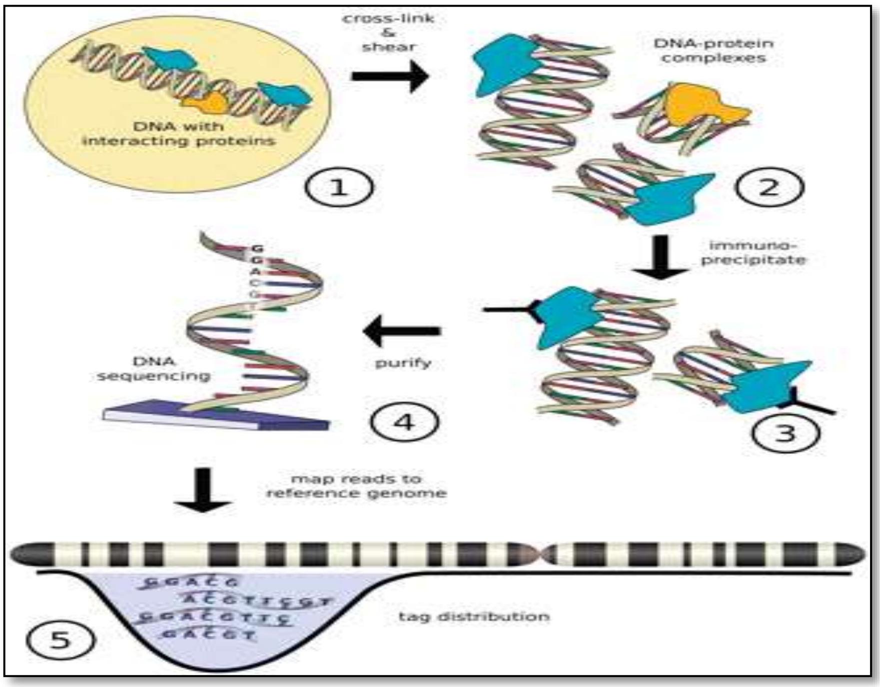


Figure 2:5 Schematic representation of ChIP-Seq technology. In Step 1, DNA and the protein are crosslinked and the DNA is sheared. In Step 2, DNA-protein complexes are obtained using immunoprecipitation technique. In Step 3 DNA and the protein is separated and DNA is purified. In Step 4, purified DNA is sequenced and finally in Step 5, the DNA sequences are mapped to the whole genome to analyse the location where the protein is bound [Szalkowski 2010].

Figure 2.5 shows step by step process of how ChIP-Seq data is produced. The success of a ChIP-Seq project depends crucially on strong enrichment of the chromatin specifically bound by the protein under study. Before any ChIP-Seq experiment, a number of antibodies, if available are evaluated and one is chosen that is with consistently high enrichment of DNA at a known binding site.

2.2.2 ChIP-Seq Analysis step

ChIP-Seq is a powerful technique that allows us to investigate the physical interaction with proteins or transcription factors. It also helps discover and understand the pattern of any epigenetic chromatin modification. Once the ChIP-Seq data is generated, the sequences are further analysed to determine the binding locations of protein under investigation. Figure 2.6 is the workflow diagram for steps involved in ChIP-Seq data analysis followed by the brief overview of some of those steps.

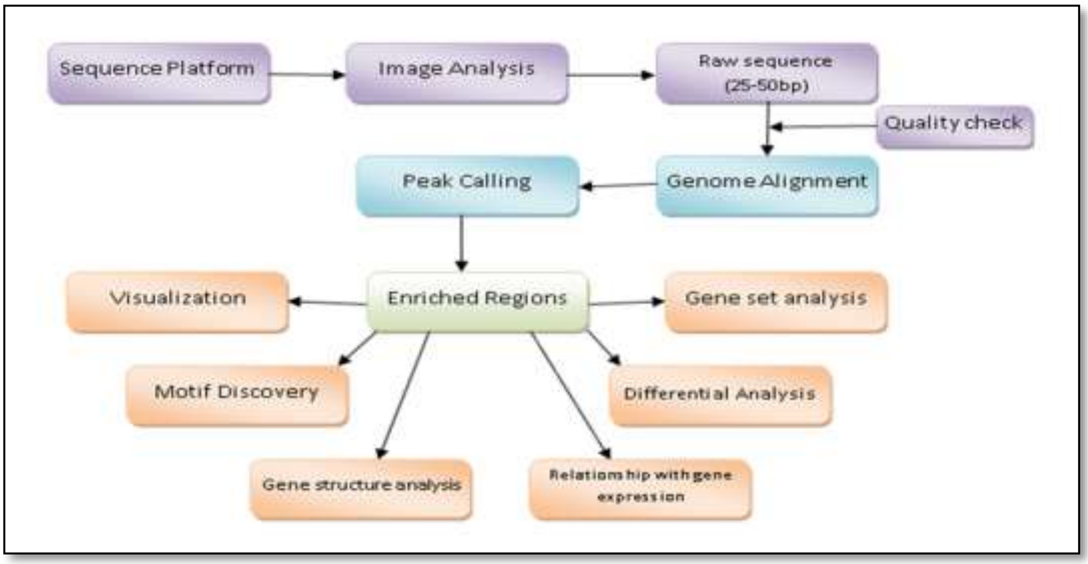


Figure 2:6 Schematics of analysis steps of ChIP-Seq data. The sequences are produced and their quality is checked, they are mapped to the whole genome and a peak-calling algorithm is applied to the aligned data to find the regions that are enriched by the protein. Further downstream analysis can be performed on the enriched results.

The raw data for chromatin immunoprecipitation followed by sequencing is generated by next generation platform and such platforms are Illumina (http://www.illumina.com/) and ABI SOLiD [Shah 2009]. The reads yielded by these platforms are short reads (typically around 25~30bp in length). However, recent platforms can result in longer reads (up to 50 ~ 100 bp) and extreme high throughput can result in up 700MB to 1GB per lane. Below each step that is involved in the workflow of the ChIP-Seq data analysis is described.

- **Quality Control ChIP-Seq Experiments**

After sequencing, before the sequences are mapped and analysed to find the protein bound locations, a number of quality controls can be used to determine if the data is worthwhile for any further investigation and validation. Packages such as FastQC [Andrews 2010] allow raw sequence quality to be assessed. There are several features that are used in assessing the quality of sequence data such as alignment independent features. Most sequencing hardware provides quality score for each base call in the read to report the confidence in assigning a specific nucleotide to each base.

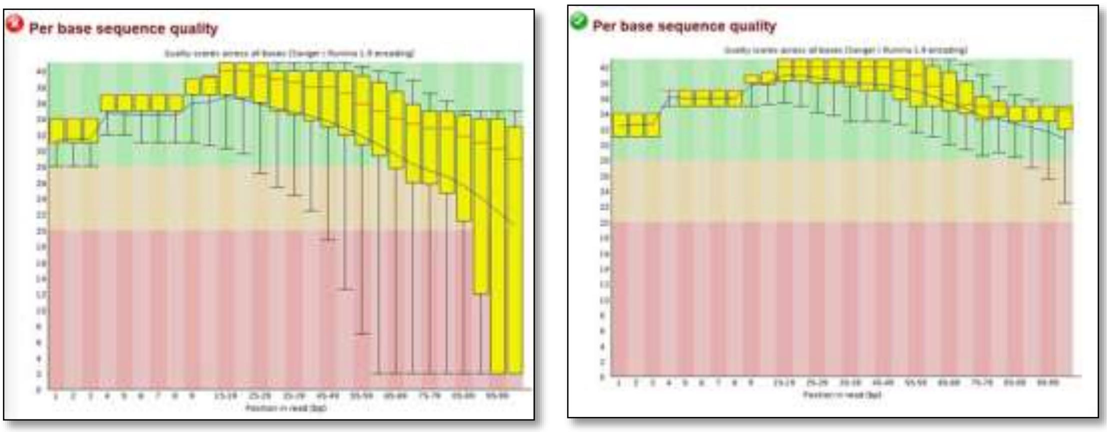


Figure 2.7: Per base sequence quality assessed by FastQC. (Left) shows sequence quality is unacceptable as good portions of the sequences scores very low in quality check and (right) shows good quality sequence data as most of the sequences scores high in quality check. In both plots, the X axis shows the position of the bases in read (1 – 99), and the Y axis shows the quality score (0 – 40).

The quality control software such as FastQC uses these scores to create plots and statistical reports about the overall quality of the data. Another feature is the number of bases that could not be called i.e the number of 'N's in the data also provides some insight to the quality of the data.

Figure 2.7 is an example of outputs by FastQC, which are the assessments of quality of per base sequence of two ChIP-Seq data. Read count enrichment can be calculated between ChIP and input samples and can help control for biases in the experimental methods. Visual inspection of the data allows for a simple but effective tool.

- **Genome Alignment**

ChIP-Seq analysis starts with mapping all the raw reads to the reference genome, the uniquely mapped reads from the ChIP experiment. In a typical ChIP-Seq experiment for a typical mammalian biological sample/biopsy, tens or even hundreds of millions of sequences must be aligned to gigabytes of a reference genome and for that reason; alignment is one of the most computationally challenging tasks in the ChIP-Seq data analysis process [Trapnell et al. 2009]. For alignment, Bowtie [Langmead et al. 2009], ELAND [Bentley et al. 2008], MACS [Zhang et al. 2008] are the most popular choices for the ChIP-Seq experiment.

There are several conditions or issues that need to be considered when choosing a mapping algorithm and its parameters. For example, one need to decide whether to keep only the reads that are found in unique position in the reference genome or whether to include reads that map to multiple locations. Accepting only unique reads, some true binding sites may not be found as they may be located in repeats or duplicated regions. On the other hand, multireads may improve signals but simultaneously may increase false positive rates. Therefore, a balance needs to be maintained between increased specificity and sensitivity while choosing the mapping algorithm [Pepke et al. 2009]. It also needs to be remembered that sequencing error can occur. Therefore alignment of reads should allow for a small number of mismatches (typically 2 ~ 3 mismatches).

- **Identification of enriched region**

After the sequenced reads are aligned to the genome, the next steps of the analysis are converting the mapped reads into a representative count number at each position in the genome and identification the regions or locations that are enriched significantly with reads or tags where significance is estimated from the distribution of the data along the genome or part of the genome that has been investigated. This step where enriched regions or peaks are identified is also known as 'peak calling'. There are several issues related to this step. The user needs to be careful while choosing a 'peak calling algorithm' as different peak callers may deal with different issues and each can be suitable for particular type of ChIP-Seq data.

A major challenge involved in detecting enriched region is that there are three types of such regions. Sharp peaks are usually found for protein-DNA binding or histone modifications at regulatory elements. Histone modifications marking domains for example transcribed or repressed regions usually have broad regions. The regions can be mixed as well. Figure 2.8 presents different types of peaks found in different data. Most of the available algorithms are designed for sharp peaks, while merging adjacent peaks for broad regions [Park et al 2009]. An effective method should take both types of regions into account and apply the relevant technique applicable for a given dataset. Peak detection algorithm is therefore a key to meaningful interpretation of ChIP-Seq data.

In peak calling, steps can be subdivided into several tasks such as, generating a signal profile for individual chromosome, defining the noise or background and true signal, identify peaks, assessing significance and finally removing artefacts [Pepke et al, 2008]. Different tools adapt different methods for these tasks.

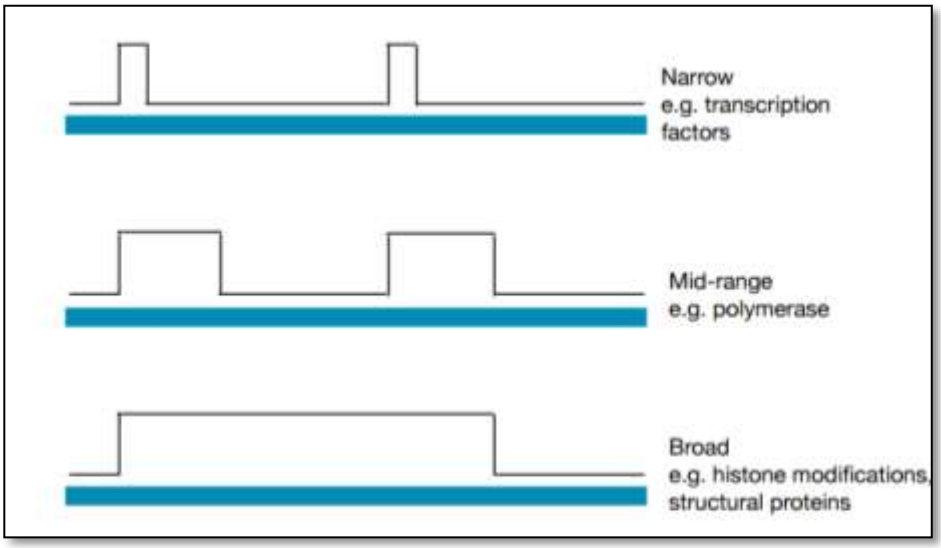

Narrow
e.g. transcription factors
Mid-range
e.g. polymerase
Broad
e.g. histone modifications, structural proteins

Figure 2.8: Different types of enriched regions depending on target proteins [Kotwaliwale 2013].

Building a signal profile is crucial in identifying enriched regions with confidence. Some tools slide a fixed length bin or window where each bin has the summation of the count at the centre. CisGenome [Ji et al. 2008] and SiSSRs [Jothi et al. 2008] both follow this method and also set criteria for consecutive windows to be merged. However, some peak calling algorithms take advantage of the direction of the reads. In this approach, the fragments are sequenced at the 5′ end and the positions of mapped reads form two separate distributions. One on the positive strand and the other on the negative strand and both is kept with a consistent distance between the peaks of the distributions. However, positive or negative strand peaks do not represent actual location of the enriched site.

To address these issues, some algorithms first construct a smoothed profile on each strand and then calculate the combined profile as showed in Figure 2.9. In order to achieve that, each distribution can be moved towards the centre or mapped location can be extended towards right fragments and fragments can be summed up.

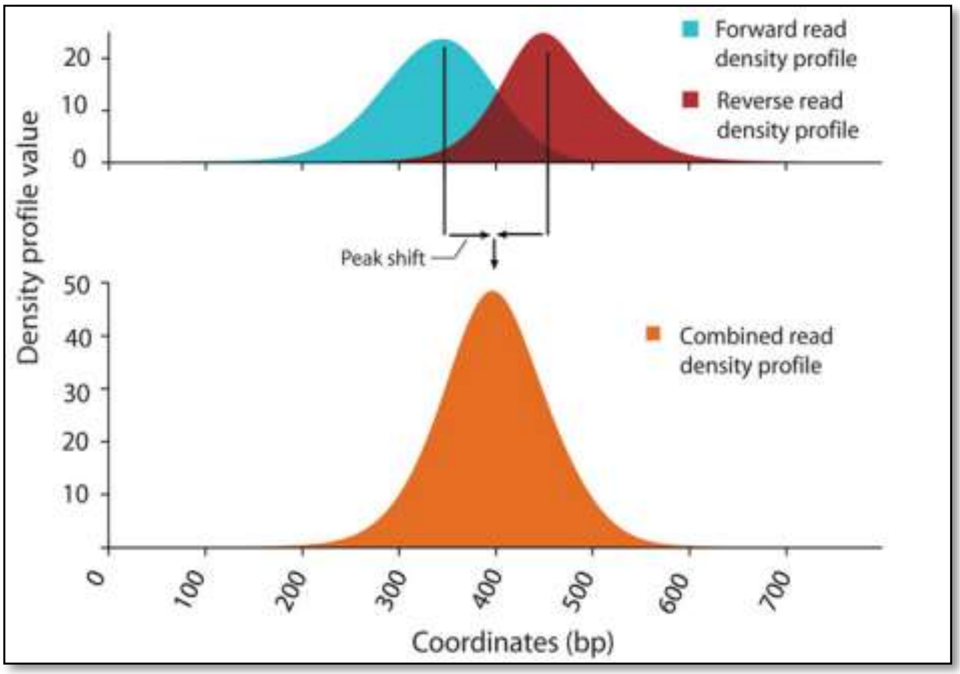


Figure 2.9: Forward and reverse (Blue and Red respectively) read density profile is used to make a combined density profile (orange) [Valouev et al. 2008].

MACS (Model-based Analysis of ChIP-Seq) [Zhang et al. 2008] shifts the read by $^{d}/_{2}$ where d is the fragment length, other methods such as FindPeaks [Fejes et al. 2008], PeakSeq [Rozowsky et al. 2009] etc. elongate the reads to a size of d where d is estimated from the actual data. This methodology should create better profile; however, there are some limitations of this approach. One needs a prior estimate of the fragment size and should assume that fragment size is uniform.

From the combined profile, peaks can be estimated. Random distribution of reads in a window of size w modelled using a theoretical distribution. Poisson model for tag distribution is a good approach as it takes into consideration both folds ratio and the absolute tag numbers. Poisson distribution has just one parameter, λ. If,

$\lambda =$ expected number of reads in window

$k =$ number of occurences of any read

Then the probability function takes the form,

$$P(X = k) = e^{-k}\frac{\lambda^k}{k!} \tag{2.1}$$

Binomial distribution is another good approach which has two parameters.

p = probability to start a read at particular position

n = window size

np = expected number of reads in a window

Then the probability function takes the form,

$$P(X = k) = C_n^k p^k (1-p)^{n-k} \tag{2.2}$$

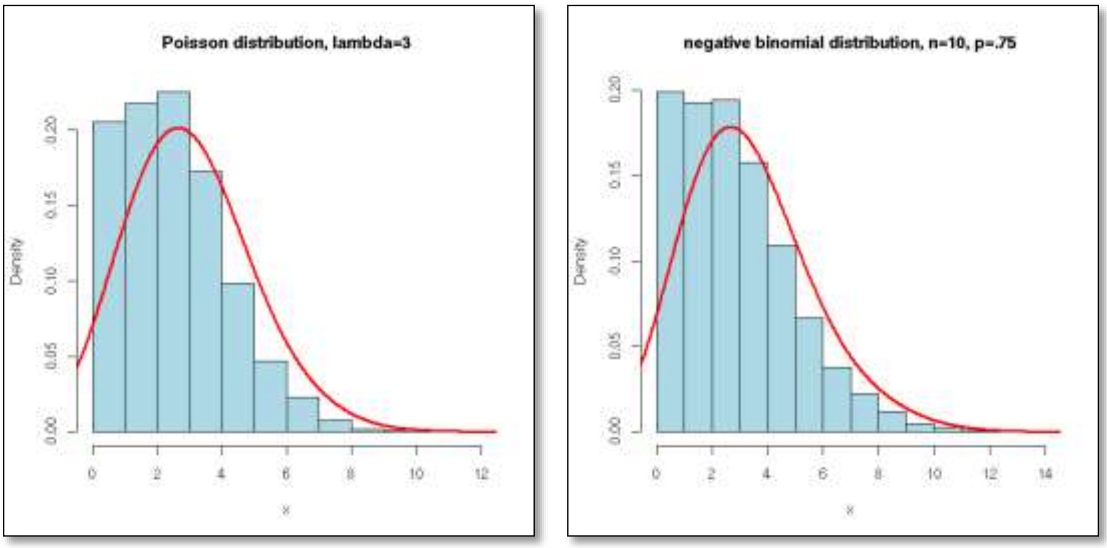


Figure 2.10: Poisson and Negative Binomial distribution.

However, the Poisson distribution has a single parameter, which is uniquely determined by its mean; its variance and all other properties follow from it; in particular, the variance is equal to the mean. However, it has been noted [Robinson et al. 2007] that the assumption of Poisson distribution is too restrictive as it predicts smaller variations than what is normally observed in the data to be investigated. Therefore, the resulting statistical test does not control type-I error (the probability of false discoveries) as required. To address this so-called over-dispersion problem, it has been proposed to model count data with negative binomial (NB) distributions [Whitaker, 1914].

Negative Binomial distribution has 2 parameters.

p = probability to start a read at particular position

r = number of sucsesses

And NB can have large variance.

$$\mathrm{Var}(X_{NB}) = \frac{\bar{X}}{1-p} \tag{2.3}$$

Depending on the underlying statistical model, a significance metric (e.g. p-value, q-value) is assigned to each putative peak.

In some experiments enriched regions are compared to a control sample, say where a non-specific antibody is used, in other cases differential binding of a protein between two or more biological conditions are also investigated.

There are several packages that are available to identify and analyse the enriched regions, all of which address different issues related to ChIP-Seq data analysis. PeakSeq [Rozowsky et al. 2009], Mosaics [Chung et al. 2014], MACS [Zhang et al. 2008], CisGenome [Ji et al. 2008], enRich [Bao et al. 2015] are among those tools to name a few. User needs to determine which one to choose in order to analyse their data depending on the type of the data in hand. Several reviews have been written summarising the methods used by different tools and their strengths and weaknesses [Ma et al. 2011; Shin et al. 2013; Steinhauser et al. 2016]. In table 2.2 profiling techniques of some of the tools along with their strengths and weaknesses are summarised.

Peak caller	Profiling of the count data	Peak Selection	Joint Modelling of two data together	Consideration of spatial dependency in adjacent windows
CisGenome	Strand specific window scan	Number of reads in window	No	No
MACS	Tag shifted then window scan	Number of reads in window	No	No
FindPeaks	Summation of overlapped tags	Height cut-off	No	No
PeakSeq	Extended tag aggression	Local region binomial p value	No	No
SICER	Sliding through windows and aggregating counts	Enrichment in relation to control	No	Yes
Mosaics	Window scan	Number of reads per window	No	Yes
enRich	Window scan	Number of reads per window	Yes	Yes

Table 2.2: Summary of some of the popular peak calling tools.

- **Downstream analysis**

After the peak is detected, there are two common downstream analysis tasks: gene annotation of the location of the enriched regions and the discovery of binding sequence motifs. Sequence motifs, the short recurring patterns in DNA play important role in regulation of gene expression. Different proteins and also RNA molecules bind to these motifs to initiate gene expression. There are several such programs available for motif discovery analysis from ChIP-seq data, for example MEME [Timothy et al. 2009], Weeder [Pavesi et al. 2004], TAMO [Gordon et al. 2005] etc. These algorithms return the details of potential motifs along with their statistical significance. Several tools for motif discovery analysis specifically designed for ChIP-seq data have been reviewed by Lihu et al. [2015].

The University of California Santa Cruz (UCSC) genome browser (genome.ucsc.edu) [Kent 2002] is a popular web-based application where alignment data can be visualized as signal overage. It also provides genomic annotations including genes (e.g. refseq, Ensembl), SNPs, evolutionary conservation, sequence properties, and patterns (e.g., CpG islands, repeats), as well as tracks for regulatory elements (e.g., transcription factor binding sites, methylation) from the ENCODE consortium [Encode], an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). An analyst can interpret the peaks in the context of functionally relevant genomic regions. There are other tools available that annotate peaks in relation to some known genomic features, for example, the transcriptional start site (TSS), exon/intron boundaries, and the 3′ ends of genes etc. ChIP-peak data can also be tested for biological pathways, Gene Ontology terms and other types of gene sets.

2.2.3 Advantages and limitations of ChIP-Seq technology

The progress in next-generation sequencing technology has been enormous. Owing to this advancement, ChIP-Seq offers higher resolution, less noise and greater coverage than its array-based predecessor ChIP–chip [Park et al. 2009]. ChIP-Seq has now become an essential technology for studying gene regulation and epigenetic mechanisms. Below are the some of the advantages this tool offers:

1) ChIP-Seq technology can help understand how transcription factor and other chromatin-associated proteins influence phenotype affecting mechanisms.
2) This technology can help determine how Proteins interact with DNA to regulate gene expression which is essential for fully understanding many biological processes and diseases.
3) Specific DNA sites in direct physical interaction with transcription factors and other proteins can be isolated by chromatin immunoprecipitation. ChIP

> produces a library of target DNA sites bound to a target in vivo. Massively parallel sequence analyses are used in conjunction with whole-genome sequence databases to analyse the interaction pattern of any protein with DNA [Johnson et al. 2007], or the pattern of any epigenetic chromatin modification.

ChIP-Seq technology has enabled many advantages and opportunities for the biomedical field; however the technology is not free from disadvantages. Sequencing error has one of the main artefacts of the technology, although the errors have been reduced significantly by process improvements. Another current disadvantage for ChIP-Seq is that the technology is still very expensive; especially for small-scale studies where fund is limited, it poses a significant problem. Its availability is thought to be another challenging issue.

The amount of data produced by a single high throughput sequencing run is huge as each experiment usually produces hundreds of millions of reads, which currently poses challenges for data management, storage and importantly, analysis. As the cost of the technology is reduced and availability increase, this problem will occur more frequently. The development of effective analysis tools are not advancing at the speed as the technology which is creating data bottleneck for the users. Another major problem is that the pipeline for ChIP-Seq data analysis is very complex, and again it is a major problem in small studies or in a study where thousands of samples are needed to be analysed. The complexity therefore makes using this technology a less cost-effective option, if not impossible.

2.3 A Brief introduction to Microarray technology

DNA microarray is an effective and rapid approach for analysing gene expression levels, at both cellular and organismic level. For last couple of decades, this technology has become an indispensable tool for biologists for analysing genome wide gene expression levels in organisms. A gene expression study involves analysing expression levels of thousands of genes simultaneously in an experiment and these large scale experiments have made gene discovery, disease sub-classification and understanding the gene regulatory network possible. Researchers have predicted that the result from DNA microarray technology with Next generation sequencing technology such as ChIP-Seq can help investigate regulatory mechanisms for gene expression and that can lead to many biological discoveries.

2.3.1 How does Microarray technology work?

A microarray most generally comprises of a glass slide, on which DNA molecules are spotted at specific locations that are called spots or features. A typical microarray can contain thousands of such features or spots and each location can have few million copies of identical DNA molecules that represent a section (generally the 3' UTR) of a gene. The DNA in a spot can either be genomic DNA or short stretch of oligo-nucleotide strands and is complementary to a gene nucleotide sequence. The spots are printed on to the microarray either by a robot or are synthesised by the technique of photolithography.

The most common application of DNA/oligonucleotide microarray is gene expression analysis. A typical experiment involves comparing expression level of a set of genes from a cell or tissue that are collected at a specific condition to the same set of genes from a reference cell or tissue that are collected at a normal condition [Lockhart 2000]. Microarray experiments can use either two-colour or one-colour techniques.

In two-colour microarrays, firstly, RNA is isolated from both samples as mentioned above. Those RNAs are then labelled with two different fluorochromes (generally the green cyanine 3 and the red cyanine 5 (Cy3, Cy5)). In the next step, they are hybridised to a microarray on which thousands of cDNAs/oligonucleotides are spotted in an orderly fashion.

After the hybridisation, the spots are excited by a laser and a scanner records the detection of green dye or red dye at suitable wavelength. The amount of fluorescence emitted after the excitation step is related to the amount of bound nucleic acid. If red and green dyes are being used in an experiment, a spot will be red or green depending whether the corresponding gene is expressed in any of the condition. If gene is expressed in both conditions, the spot will be yellow. If that gene is not expressed in any of the condition, the spot will be black. Therefore at the end of an experiment, image of the microarray is made and on that image fluorescence value of each location of the microarray that corresponds to a particular gene represents expression level of that particular gene.

Nowadays one-colour microarrays are very popular. One-colour microarrays give estimations of the absolute levels of gene expression. If genes from two separate conditions are needed to compared, two separate single-dye hybridizations are performed. These may be compared to other genes within a sample or to reference normalizing probes used to calibrate data across the entire array and across multiple arrays.

One of the advantages of one-colour microarrays over two-colour microarrays is that as each array chip is only used for only sample, anomalies from one data cannot affect the raw data derived from other samples. There are other advantages of one-colour microarrays such as it can reduce costs without compromising sensitivity and specificity [Schwarz et al. 2010]. Here, data collected at one experiment can be compared with data collected from several experiments. The absolute values of gene expression may be compared between studies conducted months or years apart. However, there are drawbacks with one-colour techniques too. Compared to the two-color system, twice as many microarrays are needed to compare samples within an experiment.

2.3.2 Analysis steps of microarray data

Multiple complicated steps or processes are involved in DNA microarray-based analysis. Various specific pieces of equipment are required to generate and analyse the data. Analysis requires not only the expertise in molecular biology but also in image analysis, computational methodologies and statistics. Figure 2.11 shows a typical microarray experiment. Below each step that is involved in the workflow of the Microarray data analysis is described.

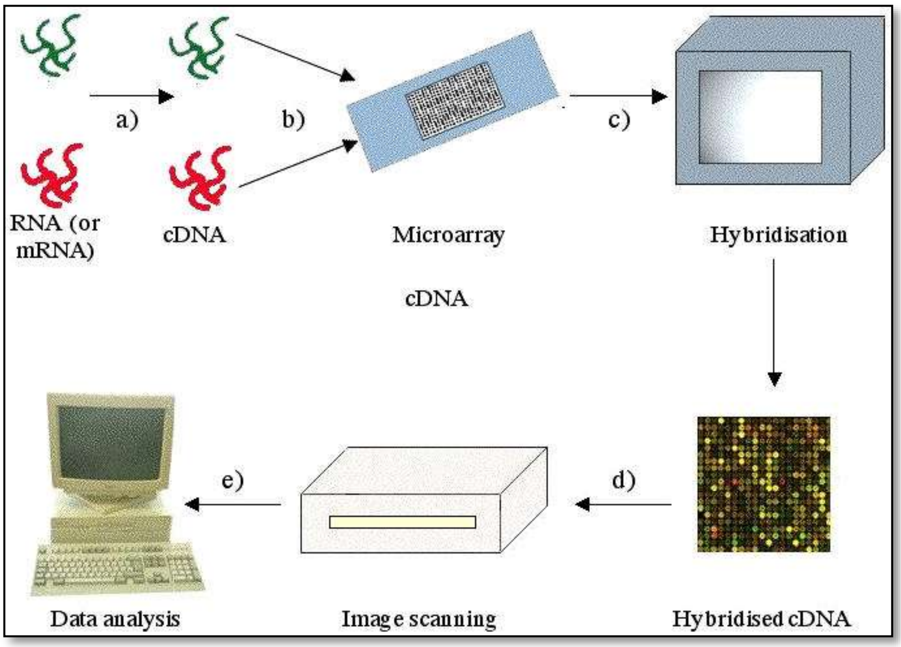


Figure 2.11: A typical microarray experiment a) mRNAs are isolated from different two samples and the mRNAs are color-coded using dye b) The DNA copy that is made (cDNA) is then spotted on microarray, c) The cDNA binds to complementary base pairs in each of the spots on the array in the hybridization process d) Based on how the DNA binds together, each spot will appear red, green, or yellow and the image of the array will be analysed to create gene expression profile e) Further computation analysis of the gene expression profile is performed to discover biologically meaningful results[Brown 2003].

- **Image processing and data normalization**

In microarray experiments level of expression for each gene can be stored as an image. Therefore, image processing is the first step of analysis with microarray data. Microarray scanners come with their own software. There are following steps involved in processing of microarray image files.

1. Identification of the spots and distinguishing them from spurious signals: In this step, spots are identified in the image. As spots are usually systematically

arranged in microarray, identification of spots is simple. Users specified parameters also help software distinguish region as spot or not.

2. Identification of the spot area to be surveyed and local region to use for estimation of the background hybridization: After identification of spots, in this step, spot signal and background intensity is calculated.
3. Reporting summary statistics and assigning spot intensity after subtracting for background intensity: After the spot and background signals are estimated, a statistical report for each spot in each channel (red and green) is produced.

- **Expression ratios and transformation of expression ratios**

Expression ratio is the commonly used metric that represents the level of gene expression. So Expression ratio represents the amount of green or red light that is emitted after excitation for each gene. If expression ratio is represented by T_k, then it is,

$$T_k = \frac{R_K}{G_k} \tag{2.4}$$

For each gene k on the array, R_K is the spot intensity metric for the test sample and G_k is the spot intensity metric for the reference or control sample.

Expression ratio is an effective way to find the difference in expression levels of different genes. However, depending on whether a gene is up-regulated or down regulated, expression ratio could be mapped between 1 and infinity or 0 to 1. By performing inverse or logarithmic transformation this inconsistency can be eliminated.

- **Data mining techniques for Gene expression analysis**

After the normalization steps mentioned above, the data is represented in a form of numerical matrix, in which each row corresponds to a specific gene and each column represents either an experimental variable/condition or specific time points. Activity levels of the genes represented by the expression values for any given condition are described as the gene expression profile. The expression levels for all genes under one particular condition are called sample expression profile where expression data can be represented in many ways, such as absolute measurement, as expression ratio, discrete values and so on. One of the main objectives of carrying out microarray data analysis is to gain insight into underlying biology by monitoring expression level of genes at a genome scale. Using expression profiling it is possible to infer the active cellular signalling events under any particular biological or experimental condition. Over the years many statistical and data mining algorithms have been developed to effectively classify, or cluster, genes or biological samples into distinct groups based on gene expression.

There are different kinds of data mining methods including two main categories: one is supervised and the other is unsupervised technique. In supervised data mining techniques, each gene expression profile is labelled with a specific class. For example, the expression profile of each sample can be associated with the specific disease and the supervised methods make use of the class information in the learning process. While, unsupervised data mining methods have no prior knowledge about the label or the class information of the genes, they learn the pattern from the data. In the context of gene expression analysis, supervised data mining methods include class association rule mining and classification, while unsupervised data mining methods mainly refer to the various clustering methods.

(1) Clustering

In clustering techniques data are organised into clusters based on similarity. Patterns within the same cluster are more similar to each other than they are to a pattern belonging to a different cluster. In the context of gene expression data analysis, clustering methods have been used to find clusters of co-expressed/co-regulated genes which can be used to distinguish between diseases that a standard pathology might find it difficult to tell apart [Alizadeh 2000]. Clustering methods can also be grouped two categories: 1. hierarchical and 2. non-hierarchical clustering [Jain 1999]. A hierarchical clustering method builds a hierarchy of clusters or tree-like structure, which is basically a nested sequence of partitions. Non-hierarchical produces a particular number of clusters at a single step. K-means algorithm, graph-theoretic approaches via the use of minimum spanning trees, evolutionary clustering algorithms, self-organising maps are some of the commonly used hierarchical clustering methods just to name a few.

In clustering technique proximity to a cluster for each gene is usually measured by a distance/dissimilarity matric or a similarity function defined on pairs of patterns. The Euclidean distance, Manhattan distance, Minkowski distance are some of the popular distance measures techniques and Pearson's correlation, Spearman's Rank Correlation are some of the similarity measures in the context of gene expression profiling.

(a) Hierarchical clustering method

In this technique two types of algorithms are used, one is an agglomerative algorithm and divisive algorithm. Say there are n number of gene expression values across a set of arrays into an individual cluster. The agglomerative algorithm keeps merging the two most similar groups to form a new cluster and reducing the number of clusters by one until all the data fall within a single cluster. A divisive algorithm, on the hand, begins with a single group and then keeps dividing groups until there are n groups, each of a single individual. Agglomerative clustering algorithms are popular choices due to its computational efficiency. Hierarchical agglomerative clustering algorithms use different distance or similarity matrices for measuring the distance between two clusters where a cluster may consist of only a single object at a time. The most commonly used inter-cluster measures are described in Equations [2.5 - 2.7].

$$\mathrm{d_{AB}} = \mathrm{min}_{\substack{\mathrm{i \in A} \\ \mathrm{j \in B}}}(\mathrm{d_{ij}}) \tag{2.5}$$

$$\mathrm{d_{AB}} = \mathrm{max}_{\substack{\mathrm{i \in A} \\ \mathrm{j \in B}}}(\mathrm{d_{ij}}) \tag{2.6}$$

$$\mathrm{d_{AB}} = \frac{1}{\mathrm{n_A n_B}} \sum_{\mathrm{i \in A}} \sum_{\mathrm{j \in B}} \mathrm{d_{ij}} \tag{2.7}$$

Where $\mathrm{d_{AB}}$ is the dissimilarity between two clusters A and B and $\mathrm{d_{ij}}$ is the dissimilarity between two individual patterns i andj. $\mathrm{n_A}$ and $\mathrm{n_B}$ are the number of individuals in clusters A and B respectively.

(b) Non-hierarchical clustering method

In non-hierarchical clustering method a single partition of the data is created which is computationally less costly than hierarchical methods. The square error is the most commonly used criterion which is optimized to obtain the cluster. K-means is one of the popular choices which uses this criterion where the square error for each cluster j, $j = [1,2,3 \ldots \ldots K]$ is the sum of the squared Euclidean distances between each pattern $x_i^{(j)}$ in the cluster j and its centre, or mean vector, of the cluster, $m^{(j)}$.

$$E_j = \sum_{i=1}^{n_j} d^2_{x_i m^{(j)}} \tag{2.8}$$

Where,

$$m^{(j)} = \frac{\sum_{i=l}^{n_j} x_i^{(j)}}{n_j} \tag{2.9}$$

Where E_j is referred to as the within-cluster variation or sum of squares for cluster j and n_j is the number of patterns within cluster j, and is the Euclidean distance from pattern x_i, to the centre of the cluster to which it is assigned. Therefore the total square error for the entire clustering with K clusters is the sum of the within-cluster sum of squares:

(2) Classification technique

Classification is an important supervised data mining approach for gene expression analysis. Over the years, many classification techniques such as decision tree [Quinlan 1993], KNN (K-nearest neighbour) [Pan 2004], SVM (Support Vector Machine) [Cristianini 2000], neural network [Good 2001], have been applied to analyse gene expression data. A classifier is first built on training samples, and then its performance is tested on test samples. If the performance is at an acceptable level, the classifier should be able to classify samples of unknown class label. SVM and Neural network have gained popularity for gene classification and associative classification [Liu et al. 1998, Li et al. 2001] has been proposed which makes the decision with the most significant class association rules.

Many different approaches for clustering and classification of gene expression data are available these and among them one might be more efficient than the others. However, one approach cannot be uniformly called superior to others. While choosing an algorithm, goals of the analysis, the background knowledge and the specific experimental constraints are needed to be taken into consideration [Garrett-Mayerand 2004].

- **Differential expression analysis**

Many microarray studies are designed to detect genes that act differently in different biological conditions such as diseased and healthy cells. In some experiments, genetic mechanisms are also perturbed with various treatments to understand the effects of those treatments. Measurement of gene expressions on different genes are usually independent, however due to the high dimension of gene expression space and our lack of understanding of all biological mechanisms, most of the techniques take on gene by gene approach. One of the approaches could be to select genes using a fold-change criterion. However due to the presence of biological and experimental variation, which may differ from gene to gene, it is important to use statistical tests to assess differential expression. To scale the data, sometimes the logarithmic scale is used in order to make the distribution of replicated measurements per gene roughly symmetric and close to normal. A variance stabilizing transformation derived from an error model for microarray measurements are also applied to make the variance of the measured intensities independent of their expected value [Huber et al. 2002].

In the most popular R package `limma` [Smyth 2004], an Empirical Bayes approach is implemented that employs a global variance estimator s_0^2 computed on the basis of all genes' variances. The resulting test statistic is a moderated t-statistic, where instead of the single-gene estimated variances s_g^2, a weighted average of s_g^2 and s_0^2 is used. Under certain distributional assumptions, this test statistic can be shown to follow a t-distribution under the null hypothesis with the degrees of freedom depending on the data [Smyth, 2004].

2.3.3 Advantages and limitations of Microarray

Microarray permits parallel analysis of thousands of genes which has opened new opportunity for genomic studies and for epigenetic research as well. Although gene expression profiling is the main application of this technology, microarray-based methods have been adapted to reveal localization patterns of DNA-binding proteins and DNA methylation. It can also help predict protein interactions and protein functions. It has been shown that genes with similar expression profiles are more likely to produce proteins that interact with each other [Ge et al. 2001; Jansen et al. 2001]. The cost of running a microarray investigation has reduced significantly over the years, making it a very affordable method for many investigators.

Microarray technology has several disadvantages too. One of the main drawbacks is that before running a microarray experiment, some prior knowledge about the genome in question is required. Biological question regarding selected genes can only be answered by this technology. Another problem with this technology is the background hybridization which occurs due to repetitive DNA sequences. This cross-hybridization makes it difficult to identify differentially expressed genes, especially lower-abundant messages.

2.4 Integration of Microarray and ChIP-Seq data and challenges

Both DNA microarray and ChIP-Seq have played a crucial role in genome research. A variety of phenotypic changes important in normal development and in diseases are temporally and spatially controlled by chromatin-coordinated gene expression [Nowak et al. 2005]. Protein binding data collected by ChIP-Seq and gene expression data generated by microarray experiments can be combined to study the relationship between epigenetic mechanisms and transcriptional activation.

Studying epigenetic modifications that occur together with a change in gene transcription can also lead to understanding the underlying mechanism how genes are regulated and identification of additional functional genomic elements that impact gene expression. Different regulatory proteins bind at specific loci of the genome at different time to regulate gene expression. Exploring different protein binding patterns or the chromatin structures at different biological conditions can also help us understand the overall mechanism of gene regulation and it can also open the door for identifying previously unknown functional genomic elements that impact gene expression.

There are other advantages of integrating these two data types. Next generation sequencing is a new advanced technology for studying epigenetics mechanisms in genomic level and it has several advantages over microarray analysis. Many have predicted it will take over the job of microarray. In fact, the mature and established analysis pipeline and cost effectiveness still make microarray a viable option for small studies or studies where thousands of samples are involved. However, the underlying epigenetic landscape cannot be fully realised given such data as it is with next generation sequencing data such as ChIP-Seq. The general view of the scientific community is that these technologies should form a relationship [Hurd et al. 2009] and from such integrative analysis it can also be investigated if comparatively cheaper option microarray alone can be used to study epigenetic landscape.

The computational biology community has made several attempts to combine protein binding and mRNA expression data over the years. Several methodologies have been proposed to study the relationship between bindings of protein at different genomic locations and gene expression changes [Markowetz et al. 2010; Qin et al. 2011; Hoang et al. 2011; Guan et al. 2014]. In general, it is essential to measure the level of epigenetic modifications and probability of enrichment for proteins in any location accurately in order to find possible relationships between ChIP-Seq and gene expression data.

ChIP-Seq data analysis steps are very complicated and there are several characteristics of ChIP-Seq data that are needed to be considered while modelling such data. For example, in immunoprecipitation technique, along with protein bound sequences, some random sequences are also picked up which creates noise in the data. Therefore, if two datasets need to be compared i.e. differential protein binding location need to be investigated, IP efficiencies between the datasets need to be considered. Another issue is that once the protein bound sequences or reads are mapped back to the whole genome, most of the statistical techniques divide the whole genome into fixed length windows and generate number of reads per window data. Then an overall distribution of count is considered to find a cut-off to call a location bound by the proteins. However as binding may cross windows it introduces spatial dependency. The ChIP-Seq count data is also usually overdispersed.

The tools those are available today to call peaks deals with different aspects of ChIP-seq data. As discussed earlier in this chapter, some tools uses simple Poisson model that fails to recognise overdispersion that is observed in the count data of ChIP-seq, some utilises a negative binomial distribution which takes into account overdispersion or deals with spatial dependency but fails to take into IP efficiency into consideration while modelling the data. While investigating association between protein bindings and gene regulation, it is very important that the model used to analyse the data deals with all the characteristics of the data as much as possible to get robust information about protein binding locations, especially if different proteins and biological conditions are involved in the study.

Most integrative methods to date find the relationship between protein binding and gene expression at one biological condition [Qin et al. 2011; Guan et al. 2014]. The studies that investigate how differential bindings of proteins may correlate with differential gene expression at different biological conditions have used very primitive analysis of ChIP-seq data thus ignore lots of characteristics such as overall distribution of counts, spatial dependencies of counts for neighbouring regions of the genome and the different efficiencies of individual ChIP-Seq experiments as mentioned above. In some studies per-gene ChIP-seq enrichment has been estimated to find relationship between gene expression and protein binding where simply tags or sequences are counted associated with a given gene or promoter region of each gene [Yu et al. 2008; Karlic et al. 2010]. Some studies have used only control samples to deduct the noises and determine enrichment around transcription start sites (TSS) [Hoang et al. 2011; Markowetz et al. 2010; Nicodeme et al. 2010]. These methods have several limitations, firstly tag counting methods within a fixed region do not consider spatial component of the data. Secondly, basic enrichment estimation methods will not consider the distribution of the data throughout the genome thus may over or underestimate the significance of enrichment.

Advancement of ChIP-Seq and genome databases available today have allowed biologists to investigate how protein binds at different genomic features and it has been concluded in several literatures how these genomic features underlie epigenetics

[Bernstein et al. 2012]. The computational biology community that has proposed methodologies and tools to integrate protein binding and gene expression data to find relationship between the two mostly focus on protein binding at some common genomic features such as promoters or transcription start sites [Markowetz et al. 2010; Qin et al. 2011; Guan et al. 2014]. However, epigenetic mechanisms that regulate gene expression do not just occur in any particular area such as promoters and transcription start sites, therefore how other features may play into regulation of gene mechanism needs to be investigated and should be accommodated in the integrative analysis.

2.5 Summary

Integrating protein binding data obtained by ChIP-Seq and gene expression data obtained by microarray experiments to study the relationship between transcriptional activation and its regulation mechanisms has much significance. However, as an emerging field, it retains many problems and challenges too as discussed in the previous section. With these problems in mind, the next chapters explore possible ways to analyse ChIP-Seq data that improves protein binding results and also propose methodologies where these improved results can be incorporated in the integrative analysis of protein binding and gene expression data. The relationship between protein binding profile at different genomic locations and biological conditions and transcriptional activity is further investigated using advanced machine learning techniques. The corresponding experimental results are validated in various ways. It is envisaged that from these, future directions into the processing of ChIP-Seq and microarray data and their integration can be embarked upon.

Chapter 3

Adapting Markov Random Field for ChIP-seq data modelling

3.1 Introduction

Chromatin immunoprecipitation (ChIP) is an important experimental technique for studying interactions between specific proteins and DNA in the cell and determining their position on a specific genomic locus [Ren et al. 2000; Lieb et al. 2001; Iyer et al. 2001; Weinmann et al. 2002]. In recent years, the combination of ChIP with the next generation DNA-sequencing technology (ChIP-seq) has expanded the scope of these studies to identify binding locations of many transcription factors, histone modifications and other chromatin-associated proteins across genome with high resolution. In ChIP-seq technology, a protein of interest is usually cross-linked with DNA site it binds to in an in vivo environment using formaldehyde. After the crosslinking is done, the DNA is sheared by sonication or other mechanism. From the resulting DNA strands, DNA sequences crosslinked with the protein are filtered out with antibody by the immunoprecipitation technique. Then the cross-linking of protein and DNA is reversed and the DNA is purified. The DNA is then sequenced, which are known as 'reads'. Finally, the short sequenced fragments (known as reads or tags) are computationally mapped by an alignment algorithm to a reference genome and regions of enriched tag counts are identified in the step known as peak-calling.

ChIP-seq experiments produces enormous amount of data and the analysis of the data is very complex, which involves several steps. Figure 3.1 demonstrates the ChIP-seq data analysis steps. To get the most out of these experiments, it is absolutely vital to choose most appropriate computational analysis methods and tools which take different aspects of ChIP-seq data into account.

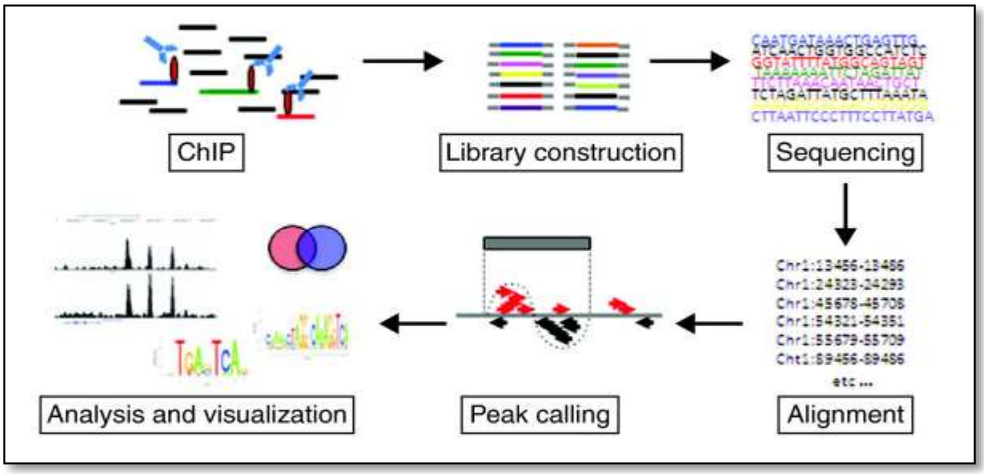


Figure 3.1: Flow scheme of the main steps in the ChIP-seq procedure. Using immunoprecipitation technique protein bound DNA sequences obtained and library is constructed. The DNA sequences are aligned to the whole genome and using appropriate peak calling method, the binding locations of the protein are found. After that, the bound regions can be visualised or further downstream analysis can be performed on the location information [Liu 2010].

In a ChIP-seq experiment, while generating the data, some random DNA sequences are also collected with the bound sequences which are usually scattered across the genome and considered as background noise. Also the background to signal ratio or ChIP-efficiency varies experiment to experiment. The antibodies that are used for specific transcription factors or protein in the immunoprecipitation technique have their own specificity for generating different signal to background ratios as well. Even with the same antibody, the technical or the biological replicates can have different specificity. When there are two or more experiments involved and the data from different conditions need to be compared, these issues may lead to over or under estimation of the overlapped or differential bound regions.

Once the sequences bound by the protein are mapped back to the whole genome, the next step is to find the enriched regions. That means, a peak calling algorithm considers the distribution of the tags or reads across the genome and finds out regions that are truly bound by the protein of interest. One of the most common methods is to divide the genome in question into fixed sized windows/bins and then summarise the counts per window. After the count data is summarised and prepared, a statistical model is used to filter out the windows with significant amount of counts that can be considered as enriched regions. However, an enrichment profile can cross between the neighbouring windows and thus spatial dependencies are introduced in the data.

Another characteristic of the data is that, as most of the regions in the genome do not have binding of the protein, there are excess numbers of zero counts in the background. Also, ChIP-seq data usually has over-dispersed per-bin read count distributions. It does not match with the existing computational method assumptions such as that read counts are generated according to a Poisson distribution with a local mean, so using a Poisson model for such data may result in incorrect assumption of statistical significance and eventually erroneous result [Hashimoto et al. 2014]. In this chapter, Markov Random Field model has been adapted to analyse ChIP-seq data to address some of these issues, such as spatial dependencies, excess zeros, overdispersion in the count data and joint modelling of replicates.

This chapter is organised as follows. In Section 3.2 some existing analytical methods for ChIP-seq data and their limitations are discussed. In Section 3.3, the ChIP-seq data analysis steps that have been adapted in this study are presented and Markov Random Field model is introduced. Section 3.4 is devoted to experimental studies and finally, the work is summarised in Section 3.5.

3.2 Background

There are several tools which deal with different aspects of ChIP-seq data. MACS [Zhang et al. 2008] is probably the most popular one to analyse ChIP-seq data. This model uses peak shifting mechanism to shift the forward and the reverse strands to create a combined profile and calls the peaks on the combined tags using a Poisson model through sliding windows. As mentioned earlier, because of the constraint of mean and variance equality in the Poisson distribution, this distribution fails to model peak data if the variability of tag counts far exceeds the mean.

CisGenome [Ji et al. 2008], another popular method, utilises a negative binomial distribution rather than a Poisson distribution to call peaks. MACS and CisGenome do not account for IP efficiencies or spatial dependencies in the data.

SCICER [Zang et al. 2009] uses a method that does not utilise fixed sized windows. Rather it scans the genome and identifies clusters of spatial signals that are unlikely to appear by chance and thus takes into account the issues with spatial dependencies and repeat regions that are not mappable by uniquely mapped reads.

MOSAiCS [Chung et al. 2011] deals with the reads that are randomly picked up by the IP technique and biases that are introduced in the data such as GC content [Dohm 2008] and mappability [Rozowsky 2008]. Later this model has been extended in MOSAiCS-HMM [Chung et al. 2014] to account for spatial dependency in ChIP. However, these models do not consider joint modelling of the data.

To overcome variability in ChIP-efficiency when two or more ChIP-seq data are investigated, Bao et al. [2013] has proposed a mixture model where multiple experiments can be modelled together where efficiency of each experiment is taken into consideration which leads to more accurate detection of enriched and differentially bound regions. The problems related to spatial dependency and excess zeroes are addressed in the approach proposed by Bao et al. [2015]. In this proposed method, Markov random Field (MRF) model has been implemented that accounts for the spatial dependencies in the ChIP-seq data and the large portion of zeroes are modelled using zero-inflated mixture distributions. The model also allows joint modelling of multiple experiments which deal with different ChIP efficiencies.

Here, this Markov Random Field (MRF) model has been adapted for the analysis of the ChIP-seq data to demonstrate how incorporating the characteristics such as spatial dependency, IP efficiency and excess zeroes can improve the result. The comparative performance analysis has been carried out between MRF and other existing methods. It has also been demonstrated how the pre-processing steps of the ChIP-seq data before running statistical analysis, such as the count correction step, can have great impact on the overall performances.

3.3 Method

ChIP-seq data analysis has several steps. The steps taken in the pre-processing of the data have significant effects on the overall binding result. Therefore those steps are also as important as modelling the data. After generating the ChIP-seq data for a particular protein, the analysis begins by aligning the raw data to the whole genome to obtain location information of each sequence in Step 1. The aligned data are usually converted into suitable formats such as, Sequence Alignment Map (SAM) or Binary Alignment Map (BAM). The chromosomes are usually modelled separately; therefore in Step 2, the aligned data is divided into different chromosomes. Using the length of each chromosome, an index is created where each entry is a fixed-size window with its co-ordinate information. Once the index is created, counts of sequences are generated per window from the aligned data. In Step 3, if there is any bias detected in the count data, it is corrected and the count data is updated before the modelling. After the correction, in Step 4, the data is modelled using MRF model and binding regions are identified using the chosen cut-off. The main steps involved in the analysis of one ChIP-seq data are illustrated in Algorithm 3.1.

Algorithm 3.1: Pseudocode for analysis of ChIP-seq data

```
Inputs

ChIP-seqData:  a ChIP-seq dataset

CutOff: an integer

Output

BindingRegions: A list of bound regions by the protein of
interest with their location information.

Function:- AnalyseChIP-seq(ChIP-seqData): BindingRegions
```

```
1. Align the ChIP-seqData to the genome.
2. Separate the chromosomes and generate count data per
   chromosome.
3. Perform count correction if needed.
4. Apply MRF model on each chromosome and using CutOFF
   identify binding regions and save them to BindingRegions
```

Step 1: Aligning the data

The sequences obtained in the ChIP-seq experiment are aligned in this step using an aligner that is capable of handling such data. The outputs are chronologically listed all the sequence reads with their genomic locations in terms of co-ordinates. A sequence can be mapped to different locations. Here, only uniquely mapped reads are retained to remove ambiguity. Two mismatches are allowed also in the algorithm for alignment. The aligned data is usually in SAM or BAM format.

Step 2: Separate the chromosomes and generate the count data

In this step, the aligned data is processed and count data is generated. The chromosomes are modelled separately, so from the aligned data, separate chromosome files are created. The lengths of the chromosomes are obtained from a genomic database and the size of the window is selected. Then, an index file is created per chromosome using the length information. The length of the chromosome is divided by the size of the window to get the number of windows per chromosome. Each entry of the index represents a window by its start and end co-ordinate. Finally, the counts of the sequences are generated per window using the co-ordinate information of each window from the aligned data. The steps are illustrated in Algorithm 3.2.

Algorithm 3.2: Pseudocode for generating count data

```
Inputs
ChIP-seqData:  an aligned ChIP-seq data
```

n: The number of chromosomes

l: An array containing the length of N chromosomes.

windowSize: An integer

Output

CountData: n number of m*3 matrix where m represents the number of fixed-length windows. The 3 columns represent the start and end co-ordinates of each window and the number of sequences found in them.

Function:- GenerateCountData(AlignedChIP-seqData): CountData

```
1. Separate the genome into n number of chromosomes and save
   them in ChromosomeList
2. For each chromosome c in ChromosomeList
          Obtain length from l of chromosome c
          Divide the length by windowSize to get w
          Generate an index with w entries with start and end
          co-ordinate and name of the chromosome
          Save the index in IndexList
   End For
3. For each index in IndexList
          For each window in index
                 Generate the number of sequences found in that
                 window using the co-ordinates
          End For
   Save it to CountData
   End For
```

Step 3: Count correction

In regions of elevated GC content, the numbers of reads are sometimes increased [Dohm 2008] and there are biases that come from the mapping algorithm as well. Some sequences can be mapped to multiple locations. It has also been observed that in some of the genomic regions abnormally a high number of sequences are mapped. These can be caused by a few factors, such as uneven chromatin structures or biased PCR amplification [Shen 2013]. These high counts may distort the background estimation.

Therefore, in this count correction step, looking at the distribution of each count data, if some unusual high counts (i.e outliers or variance is too big) are found in any concentrated areas, it is assumed that they have come from the biases. Therefore, those counts are removed and replaced by 0 in the method.

Step 4: Modelling the data with the MRF model

After the count data is prepared, the proposed methodology by Bao et al. [2015] for the analysis of ChIP-seq data has been followed for finding the enriched regions of the protein of interest. Given the count data, MRF model considers the distribution of the counts across the genome in question and associates a probability to each window of it being enriched or not. Additional information such as enrichment information of neighbouring regions is also considered while calculating this probability to incorporate spatial dependency. A brief overview of the model is given below.

In the MRF model, let M be the number of total bins in the regions of interest in a particular chromosome. Let, Y_{mc} be the counts in the mth bin, ($m = 1, 2, 3, \dots., M$) and c, the condition (time points or control/sample). The counts can be from either background (non-enriched region) or from the signal (enriched regions). So, given the data, the interest is in inferring the state of the latent variable X_{mc}.

That is $X_{mc} = 1$ if enriched, $X_{mc} = 0$ if not enriched, so the joint mixture model for Y_{mc} is as follows:

$$Y_{mc} \sim p_c f(y, \theta_{cr}^S) + (1 - p_c) f(y, \theta_{cr}^B) \qquad (3.1)$$

Where $p_c = P(X_{sc} = 1)$ is the mixture portion of the signal component and $f(y, \theta_{cr}^S)$and $f(y, \theta_{cr}^B)$ are the signal and background densities, respectively. Using this model, the enriched regions can be detected by controlling false discovery rate (FDR).

One of the characteristics that make this model attractive is that the probability p_c of a region being enriched does not depend on ChIP efficiencies. However, the parameters signal and background distributions θ_{cr}^{S} and θ_{cr}^{B} depend on ChIP efficiencies of replicates r. As the signal and background densities can take any form, the signal can be modelled using Poisson or Negative Binomial and their zero-inflated extensions to account for the excess number of zeros typical of this type of data.

So for the mixture components $f(y, \theta^{S})$ and $f(y, \theta^{B})$,

$$Y_{mc}|X_{mc} = 0 \sim ZIP(\pi_c, \lambda_{0c}) \ or \ ZINB(\pi_c, \mu_{0c}, \phi_{0c}), \tag{3.2}$$

$$Y_{mc}|X_{mc} = 1 \sim Poisson(\lambda_{1c}) \ or \ NB(, \mu_{1c}, \phi_{1c}) \tag{3.3}$$

The latent variable X_{mc}, which represents the binding profile is assumed to satisfy 1D Markov properties. Given the adjacent bins states, $X_{m-1,c} = i, X_{m+1,c} = j$ with $i, j \in \{1,0\}$

$$Y_{mc}|X_{m-1,c} = i, X_{m+1,c} = j \sim p_{c,ij} f(y, \theta_c^{S}) + \left(1 - p_{c,ij}\right) f(y, \theta_c^{B}) \tag{3.4}$$

Thus, the enrichment of a region depends on the state of the two adjacent regions. All the parameters in this model are estimated using Bayesian approach.

Finally to decide whether a region is enriched or not, a user can set a threshold on these probabilities. Different criteria can be used to set this cut-off, whereby each region is assigned to the state with the highest posterior probability.

If D is the set of declared enriched regions corresponding to a particular cut-off on the posterior probabilities, then the estimated false discovery rate for this cut-off is given by:

$$\widehat{FDR} = \frac{\sum_{m \epsilon D} \hat{P}(X_{mc}=0|Y)}{|D|} \tag{3.5}$$

3.4 Results

3.4.1 Data

In this experiment, time-series ChIP-seq datasets for Histone protein H3, RNA Polymerase II (RNA PolII) and Cyclin-dependent kinase 9 (CDK9) [Nicodeme et al. 2010] have been used. The data was collected from bone-marrow derived macrophages (BMDMs) stimulated with lipopolysaccharide (LPS). The data was also collected from LPS stimulated samples that are treated with a synthetic compound (I-BET) that, by 'mimicking' acetylated histones, disrupts chromatin complexes responsible for the expression of key inflammatory genes in activated macrophages. The LPS stimulated data are described as ‘control’ in this chapter and the data that are LPS stimulated and also treated with IBET compound are described as ‘drug’. The ChIP-seq data were collected at three time points: 0, 1 and 4 hours (which are described as 0H, 1H and 4H respectively in this chapter).

H3 is one of the main histone proteins. The reason behind using this data is that it has two technical replicates per biological condition, which has given the opportunity to assess the strength of joint modelling of the data. As RNA polymerase (RNA PolII) has known to be bound at the specific locations such as the promoter sequences of the genome to initiate gene regulation, the ChIP-seq datasets for RNA PolII have been used to validate the results biologically.

3.4.2 Pre-processing of the data

Alignment

The ChIP-seq datasets are in the FASTAQ format that contains the raw sequences. The alignment tool, `Bowtie`, has been used for aligning the ChIP-seq data used in the experiment. `Bowtie` is an ultrafast, memory-efficient short read aligner that can align very large sets of short DNA sequences or tags to large genomes in a very short period of time. As a reference genome, `Bowtie` requires indexed DNA sequences. The `bowtie-build` function can build an index in the form of six output files. These six files together are called the index and they are needed to align the reads to the reference genome. The algorithm that is used to build the index is based on the blockwise algorithm of Karkkainen.

Indexed mouse genome (mm9) has been used as the reference genome for the ChIP-seq data to be aligned. The reference genome obtained from UCSC Genome Browser [https://genome.ucsc.edu/]. This version was released in July 2007. Nicodeme et al. (2010) has used this version of the genome in their analysis, so this pre-processing step was kept similar so that the validation was easier if required. The aligned data produced by `Bowtie` is in standard SAM format. Due to some unusual findings in chromosome 2, where a large number of sequences are found to be aligned in a concentrated small region, some datasets have also been aligned using `BWA` to perform a comparative analysis. It has been concluded that both tools produce the similar results, as the unusual reads have also been reported by `BWA`. After the alignment with Bowtie, it produces some information about the whole alignment run including the total number of reads that were processed and the percentage of total reads for which Bowtie found at least one alignment.

Condition	**Proteins**						
	Brd4	**H3**	**H3K4me3**	**H4ac**	**RNA PolII**	**CDK9**	**RNA PolII S2**
0H minus	74.43	78.00	78.53	73.55	83.26	78.67	76.09
0H plus	74.52	64.16	82.80	74.05	82.17	75.80	82.98
1H minus	70.24	82.57	73.35	78.61	81.50	70.35	83.40
1H plus	78.03	64.05	73.07	77.74	82.69	80.89	80.03
4H minus	75.90	65.33	80.18	76.88	82.48	80.18	79.08
4H plus	60.86	62.24	79.05	71.35	81.18	69.79	76.89

Table 3.1: Percentage of sequences that are aligned per ChIP-seq dataset using Bowtie

Table 3.1 summarises the percentage of sequences that are aligned to the reference genome mm9 for all proteins/markers (used in different experiments throughout the thesis) at each condition by `Bowtie`. There is no consensus about the acceptability of the alignment percentage and it also depends on the number mismatches that are allowed for alignment. If allowed too little, the percentage goes down and lots of information can be thrown away, however if allowed too big, the result could be misleading. As large datasets have been compared for several proteins in this thesis, default number of mismatches that Bowtie allows has been used, which is 2. Allowing two mismatches only, it has been observed that the alignment scores are at the higher end for most of the datasets that have been used in this thesis.

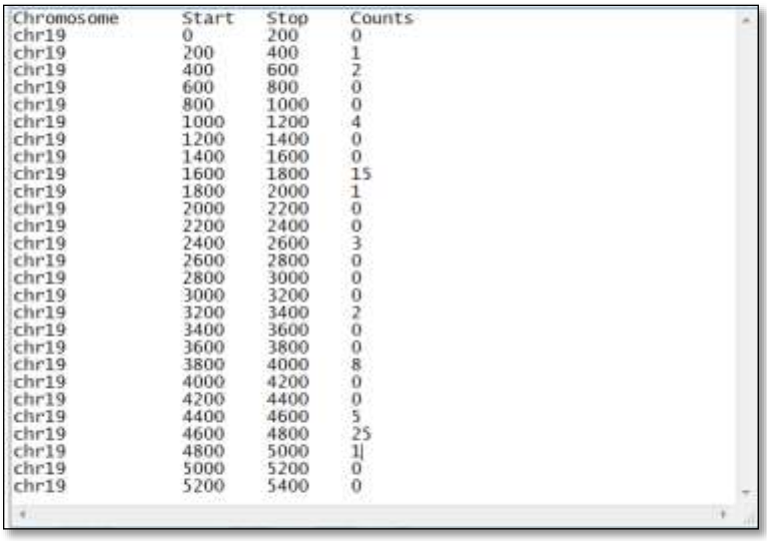
Generating count data

After the alignment step, the genome in question is needed to be divided into fixed size windows or bins to generate the count profile, so that the distribution of the data can be analysed to find the true enriched regions. To do so, a couple of information is required about the genome, such as, how many chromosomes there are in the genome, the length of each chromosome etc. The species in question, the mouse has 21 chromosomes including the X and the Y chromosome. The X and the Y chromosomes haven't been used in any of the experiments, therefore, 19 chromosomes are divided into separate files except X and Y. The lengths of all 19 chromosomes have been downloaded from UCSC.

A function written in python has been used for dividing each chromosome length into fixed size windows and creating an index per chromosome. The function takes a text file containing the length information of each chromosome and user-chosen size of the window or bin as input. The program then outputs text files, containing windows (start position and end position of each window) for each chromosome. In this experiment, the 200 base pair (bp) window size has been chosen (the reason for this decision is described at the end of this chapter).

In the next step, the number of counts is summarised per window and the count data is prepared. A function written in python has been used for this task. The function has been automated such a way so that several ChIP-seq data can be run together. For each dataset, it loops through all 19 chromosomes and produces a count data for each chromosome. The count data is a text file that contains the information of the numbers of tags per window, with the location information of the window in terms of the co-ordinate and the name of the chromosome that window comes from. Both of these python program are the implementation of Algorithm 3.2 for generating count data.

Figure 3.2 shows a tabular representation of the count data that is resulted in this post-alignment process for chromosome 19 generated from ChIP-seq data of H3. There are four columns in the file. The column, 'Chromosome' contains the name of the chromosome where the reads have come from. The columns, 'Start' and 'Stop' contains the location information of the regions in terms of co-ordinates and finally, the column 'Counts' contains the number of sequences found in each region. This count data is then provided to a model to find out the regions that are significantly enriched by the protein of interest.



Chromosome	Start	Stop	Counts
chr19	0	200	0
chr19	200	400	1
chr19	400	600	2
chr19	600	800	0
chr19	800	1000	0
chr19	1000	1200	4
chr19	1200	1400	0
chr19	1400	1600	0
chr19	1600	1800	15
chr19	1800	2000	1
chr19	2000	2200	0
chr19	2200	2400	0
chr19	2400	2600	3
chr19	2600	2800	0
chr19	2800	3000	0
chr19	3000	3200	0
chr19	3200	3400	2
chr19	3400	3600	0
chr19	3600	3800	0
chr19	3800	4000	8
chr19	4000	4200	0
chr19	4200	4400	0
chr19	4400	4600	5
chr19	4600	4800	25
chr19	4800	5000	1
chr19	5000	5200	0
chr19	5200	5400	0

Figure 3.2: A tabular representation of the count data that is given as the input to the statistical model to analyse the enriched regions

Applying MRF model

After the pre-processing step, the resulting count data can be used to find the locations of the genome that is likely to be bound by a protein in question. The MRF model for analysing the peaks of the ChIP-seq data is implemented in R. MRF model along with mixture model previously developed by Bao et al. [2013] can be found in R package called `enRich`. This tool also provides joint statistical modelling of ChIP-seq data, accounting for technical/biological replicates, multiple conditions and different ChIP efficiencies of the individual experiments.

To apply the MRF model on the count data, the mrf function provided by the R package `enRich` has been used.

The function uses an MCMC algorithm to fit a one-dimensional Markov random field model for the latent binding profile from ChIP-seq data. The emission distribution of the enriched state (signal) can be either Poisson or Negative Binomial (NB), while the emission distribution of the non-enriched state (background) can be either a Zero-inflated Poisson (ZIP) or a Zero-inflated Negative Binomial (ZINB) as described in the method section. As input, the count data is provided. Negative Binomial has been used as the method in this experiment that refers to the densities of the mixture distribution. It means a ZINB distribution has been used for the background and a Negative Binomial for the signal. 2000 for MCMC iteration steps and 1000 for burn-in steps have been used for modelling the data. The function outputs the estimates of the parameters. It also produces a sample matrix drawing from the posterior distributions of the parameters. The samples are collected one from every ten steps right after burn-in step. The posterior probability list for each window of being enriched is also produced. For the joint modelling of two or more ChIP-seq datasets that are either biological/ technical replicates or that needs to be analysed for differential enrichment are modelled using mrf.joint function. In joint modelling the model allows the splitting of the background and signal components of the data that gives different efficiency ratio for individual dataset. When data are modelled jointly these ratios are taken into account to detect enriched regions found jointly in multiple datasets or differentially enriched regions [Bao et al. 2013].

After the data is analysed, the locations that are enriched or bound by the protein of interest are extracted. For this purpose enrich.mrf function has been used that detects the enriched and differentially bound regions for fitting results of mrf and mrf.joint by controlling a given FDR level. enrich.mrf also calculates the IP efficiencies for each experiment. The input for this function is the output of mrf or mrf.joint. The user can also choose the FDR for identifying enriched regions. As the cut-off value, 0.5 has been used for identifying the enriched regions in this experiment.

For comparative analysis, the datasets are also analysed using a mixture model with Negative Binomial distribution. In that case, the function mix has been used on the count data that adopts an EM algorithm to fit the data by a latent mixture model with two components. One component is the signal density and the other is the background density. Function mix can deal with more than one experiment at the same time. In this case, it fits individual models to each experiment. For joint modelling function mix.joint has been used. After modelling the data, enrich.mix has been used to identify enriched regions at the chosen FDR. Below are the examples of some commands that are run to call these functions on the ChIP-seq count data.

Say, there is ChIP-seq data for protein, P and the count file generated for chromosome 1 from this dataset is called P_chr1_countData. The count file has 4 columns, where first three columns represent the name of the chromosome and the regions and the fourth column has the count data as seen in Figure 3.2. To find the enriched regions by modelling this count data with the MRF model at 5% FDR (which is a cutoff typically used in most of the statistical analysis), the following commands are called in R console.

```
1. library(enrich)
2. P_chr1=list()
3. P_chr1$region= P_chr1_countData [,1:3]
4. P_chr1$count= P_chr1_countData [,4]
5. P_chr1_mrf= mrf(P_chr1, method="NB", exp="P chr1",
   Niteration=2000, Nburnin=1000, cr=0.05)
6. P_chr1_enrich=enrich.mrf(P_chr1_mrf, analysis="separate")
```

If there are two technical replicates for protein P and the count files for the replicates of chromosome 1 of P are called P_rep1_chr1_countData and P_rep2_chr1_countData respectively, joint modelling can be on the data and get the enriched regions by running the following commands in R.

```
1. P_replicates_chr1=list()
2. P_replicates_chr1$region= P_rep1_chr1_countData [,1:3]
3. P_replicates_chr1$count=cbind(P_rep1_chr1_countData [,4],
   rep2= P_rep1_chr2_countData [,4])
4. P_replicates_chr1_mrf =mrf.joint(P_replicates_chr1,
   method="NB", exp=c("rep1", "rep2"), rep.vec=c(1,1),
   p.vec=c(1,1), Niteration=2000, Nburnin=1000, cr=0.05)
5. P_replicates_chr1_enrich=enrich.mrf(P_replicates_chr1_mrf
   )
```

To model the data with the negative binomial distribution, mrf can be replaced with mix in the above commands.

3.4.3 Comparative analysis of the MRF model and the Negative binomial distribution model

The ChIP-seq datasets for histone protein H3 have been analysed with both the latent mixture model and the MRF model in order to evaluate the performances of both models. Each biological condition has two technical replicates; therefore, it has been also possible to investigate the strength of the joint modelling of the data. The comparative performance has been conducted and the results have been published in [Ferdous et al 2015]. The result is also given below. For each condition, Table 3.2 shows the number of regions bound by H3 at 5% FDR by both MRF model and the mixture model with negative binomial (NB) distribution. At each biological condition, two replicates have been modelled jointly to get the bound regions.

Conditions	**MRF**	**Mixture model with NB**
	Number of enriched regions (200 bp) at 5% FDR	
0H control	3604	3113
1H control	3412	3006
4H control	4886	2962
0H drug	4448	2793
1H drug	3783	3181
4H drug	3921	2978

Table 3.2: The number of 200bp enriched regions found by the MRF and the mixture models with NB at 5% FDR

From Table 3.2, it can be observed that in each condition, MRF produces more regions than the other method. Observing differentially bound regions between two models, it has been found that MRF assigns a high probability to a region that has low tag counts but has neighbouring regions with some significant number of counts as it incorporates spatial dependency in the model. On the other hand, the mixture model assigns a very low enrichment probability to those regions, thus, may discard lots of useful information [Zang et al. 2009]. Table 3.3 shows the number of regions uniquely bound by each method, that is, these regions are bound by one method but ignored by the other.

Conditions	**MRF**	**Mixture model with NB**
	Unique regions (200 bp) at 5% FDR	
0H minus	1262	771
1H minus	1160	754
4H minus	2630	706
0H plus	2347	692
1H plus	1549	947
4H plus	1756	813

Table 3.3: Number of unique regions found by the MRF and the Mixture model with mixture with NB models

Some of the bound regions uniquely found by just one model have been observed in the integrated genome browse viewer (IGV) to investigate why these regions may be considered as enriched by one model while ignored by the other. Figure 3.3 (Left) shows an island of counts shown in IGV which has been reported as enriched by the MRF model but not by the mixture model. Due to the spatial dependency, all the small counts in that island are given high probabilities by the MRF model. However, as the Negative Binomial (NB) distribution does not account for spatial dependency, these small counts are given low probabilities individually and as a result are not considered enriched. In Figure 3.3 (Right) it shows a window that has a high count of sequences and has been considered bound by the NB model. Due to the lack of counts in the neighbouring regions, this window has been ignored as non-bound regions by the MRF.

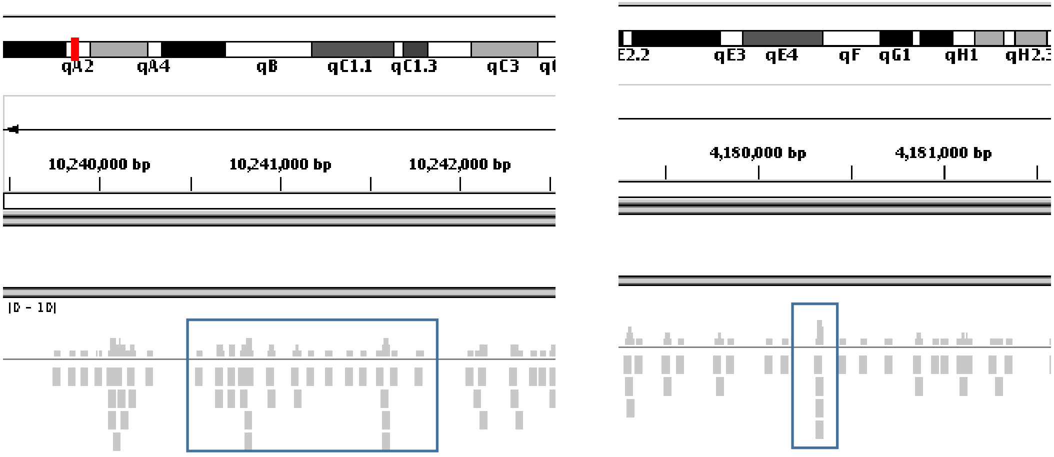


Figure 3.3: Integrated genome browser view of count data in some selected regions (Left) A region that is reported as significant by the MRF model due to spatial dependency but ignored by the NB model. (Right) A stand-alone high count that is reported as significant by the NB method but ignored by the MRF model due to lack of counts in the adjacent bins.

3.4.4 Biological Validation

In order to verify the result biologically, the ChIP-seq dataset for RNA PolII has been also analysed using both methods. Table 3.4 summarises the number of regions found by the MRF and the latent mixture model using NB for two chromosomes. In this case, it has been noted that the difference between the numbers of enriched regions found by two different methods is very large. The mixture model with negative Binomial picks up only sharp peaks, whereas the MRF model considers the spatial dependency, therefore broad peaks are identified and more regions are reported than the other model.

Chromosome	**MRF**	**Mixture model with NB**
	Number of enriched regions (200 bp) at 5% FDR	
Chr1	45,560	8950
Chr19	17027	3805

Table 3.4: The number of enriched regions found for RNA PolII reported by the MRF and the mixture model with NB models

Chromosome	**Mixture model with NB**	**MRF**
	Percentage of Promoters	
Chr1	31.24%	21.27%
Chr19	41.59%	31.75%

Table 3.5: Percentage of promoters found in enriched regions of RNA PolII reported by the mixture model with NB and the MRF models

RNA polymerase II is known to bind to the promoter sequences and initiate transcription. The R package `ChIPseeker` has been used to summarise the enriched regions obtained by two different methods in terms of the percentages of different genomic features they fall into (Table 3.5 and Figure 3.4). It has been investigated which model yields more bound regions in the promoter area. The percentage of promoters is higher in the enriched regions generated by the mixture model (NB), however as MRF produces more enriched regions, it yields more promoters in the enriched regions than the mixture model.

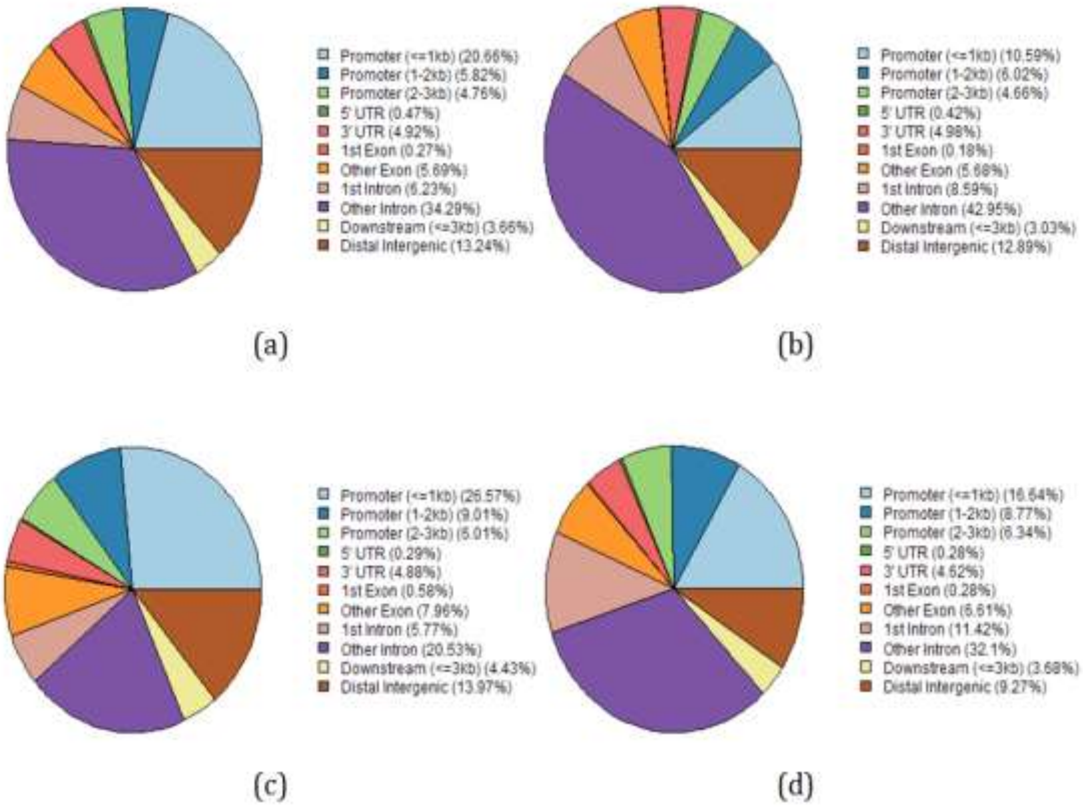
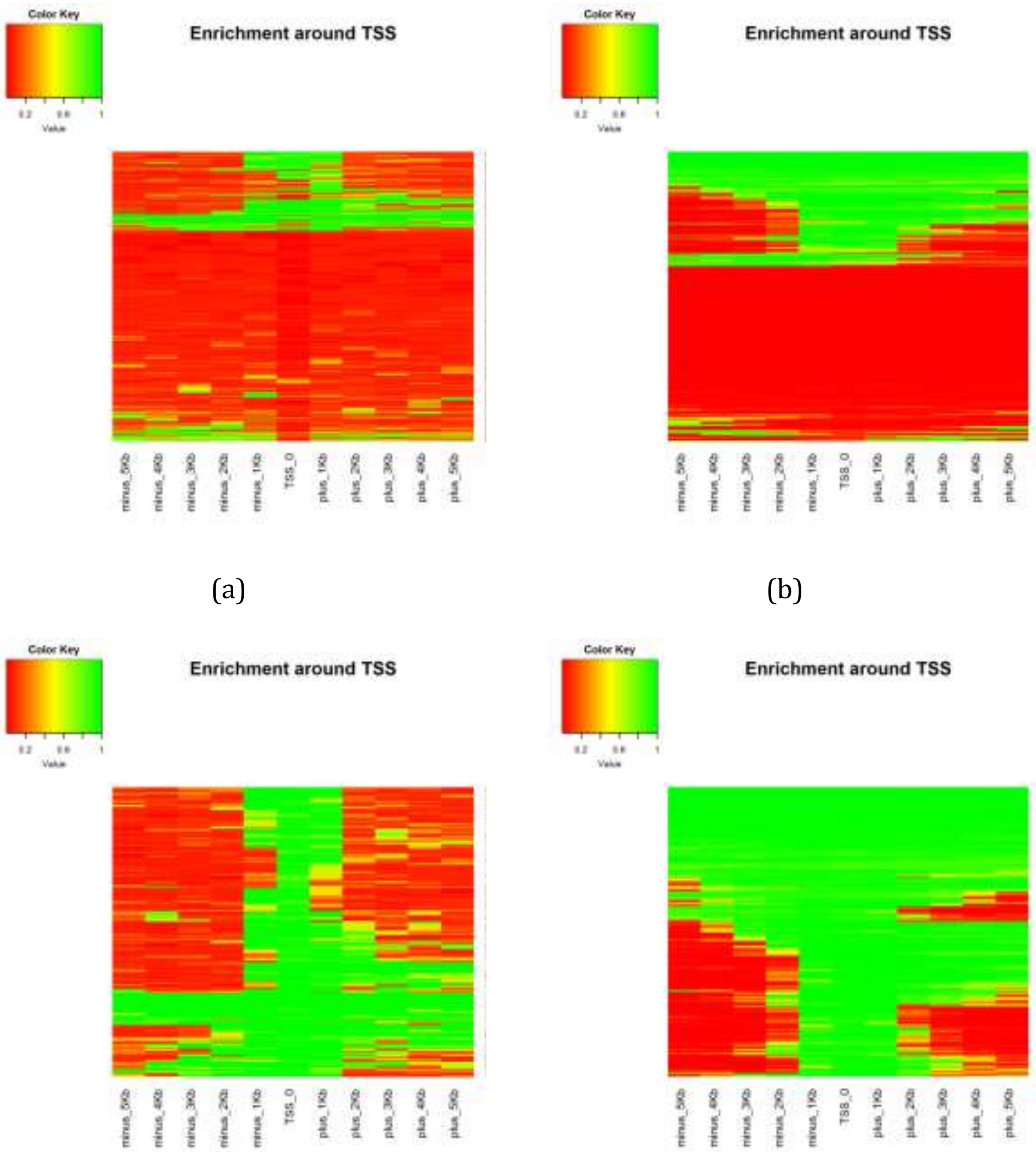


Figure 3.4: (Top) Distribution of the enriched regions of chromosome 1 of RNA PolII (a) reported by the NB model (b) reported by the MRF model (Bottom) Distribution of the enriched regions of chromosome 19 of RNA PolII (c) reported by the NB model (d) reported by the MRF model



(a) (b)

(c) (d)

Figure 3.5: (Top) The distribution of binding probabilities of RNA PolII around TSSs in chromosome 1 (a) by the NB model (b) by the MRF model. (Bottom) The distribution of binding probabilities of RNA PolII around selected TSSs that are significantly enriched (c) by the NB model (d) by the MRF model

Promoter sequences are typically located directly upstream or at the 5' end of the transcription start sites (TSS). As RNA PolII is known to bind at promoters, which resides around the transcription start sites, enrichment probabilities have been generated (reported by both model) at the start sites and 5kb upstream and downstream (upstream and downstream represented as minus and plus in the heatmaps) from the probability results yielded by both models. The heatmaps in Figure 3.5 show the distribution of binding probability of RNA PolII around the TSSs in chromosome 1. Figure 3.5(a) and Figure 3.5(b) show binding probabilities at all TSS sites in chromosome 1 reported by NB and MRF respectively. There are some TSSs which are not bound by RNA PolII where genes are inactive (represented by red). In Figure 3.5(c) and Figure 3.5(d) the bound TSS are only selected to observe the distribution. The heat maps in Figure 3.5 show that the portions of transcription start sites that are bound by RNA PolII, where the bindings in the surroundings of those TSSs are very smooth and obvious with the MRF result. However, with the mixture modelling, it is not so apparent.

- **Joint model Vs individual modelling (H3)**

To check the merit of the joint modelling technique, the technical replicates have been analysed separately to investigate how the results of binding regions differ from the experiments where the replicates are modelled jointly. For this experiment, the ChIP-seq data for H3 protein has been used for three time points (0H, 1H and 4H) for two experimental conditions (drug and control).

For separate modelling, two replicates at each condition have been modelled individually and overlapped regions from both replicates found after modelling at 5% FDR have been recorded. For joint modelling, both replicates have been modelled jointly, while IP efficiency of both replicates are taken into consideration and one peak list per condition has been generated by the model. For both experiments, Markov random Field model has been used with the negative binomial method and others parameters have been kept the same for both joint and separate modelling.

The result shows that the joint modelling produces more enriched regions at each experiment at 5% FDR than if the replicates are modelled separately and overlapped regions are taken. Joint modelling has produced 622, 336, 2260, 1912, 929 and 1230 more regions respectively in six experiments than separate modelling. The results of the number of regions produced in both experiments and time taken to model the data have been summarised in Table 3.6. The experiments have also been timed. On a computer with 8 GB RAM and duel processors, timewise, the models do not show any significant differences.

Conditions	**Number of enriched regions**		**Time taken to model the data**	
	Joint modelling	**Separate modelling**	**Joint modelling**	**Separate modelling**
0H control	3604	2982	2.54	2.62
1H control	3412	3076	2.34	2.69
4H control	4886	2626	2.68	2.28
0H drug	4448	2536	2.40	2.51
1H drug	3783	2854	2.35	2.65
4H drug	3921	2691	2.45	2.55

Table 3.6: The comparative analysis results between joint and separate modelling techniques of the ChIP-seq data.

- **Count correction**

In all of the ChIP-seq datasets, an abnormally high number of sequences is mapped at some highly concentrated genomic regions in some of the chromosomes. As an example, in Figure 3.6, the histograms shows the read distribution of count data in the chromosome 2 in ChIP-seq data for CDK9, obtained at 4 hour time point from drug data. On the right, the distribution includes the regions that have more than 20 but less than 200 counts. On the left, the histogram shows an overall read distribution in the regions with more than 200 counts. And it has been observed that a very small number of regions has abnormally high counts. Figure 3.7 shows a tabular representation of the count data at those genomic regions. From genomic co-ordinate 98502200 to 98507400, 1124474 sequences have been mapped.

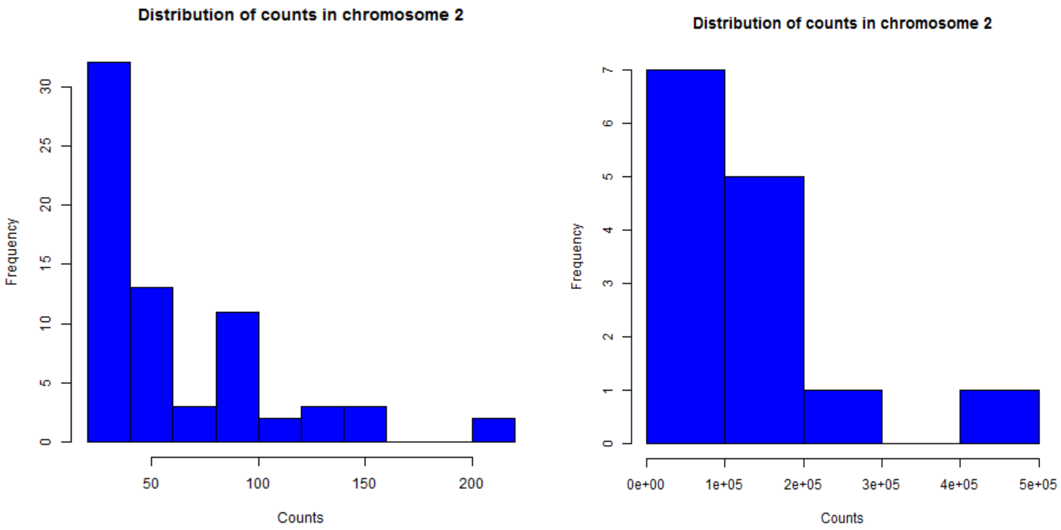
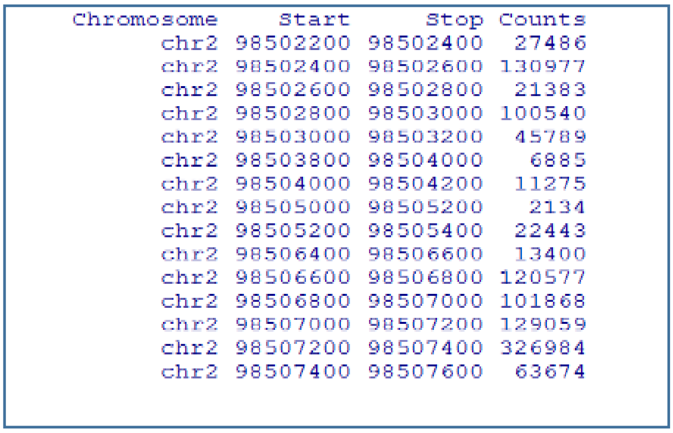


Figure 3.6 In chromosome 2 (Left) Frequency of counts, in the range of 0 and 200 per 200 bp (Right) Frequency of counts, larger than 1000 per 200bp



```
 Chromosome     Start      Stop Counts
       chr2  98502200  98502400  27486
       chr2  98502400  98502600 130977
       chr2  98502600  98502800  21383
       chr2  98502800  98503000 100540
       chr2  98503000  98503200  45789
       chr2  98503800  98504000   6885
       chr2  98504000  98504200  11275
       chr2  98505000  98505200   2134
       chr2  98505200  98505400  22443
       chr2  98506400  98506600  13400
       chr2  98506600  98506800 120577
       chr2  98506800  98507000 101868
       chr2  98507000  98507200 129059
       chr2  98507200  98507400 326984
       chr2  98507400  98507600  63674
```

Figure 3.7: A tabular representation of the tag counts at the genomic position from co-ordinate 98502200 bp to 98507400 bp at chromosome 2 of ChIP-seq data for protein CDK9.

The datasets have been modelled with these high counts removed and it has been found that at each experiment, more enriched regions have been found by the model than if the data has been modelled with those unusually mapped tags still in place. Table 3.7 summarises the numbers of enriched regions found by modelling chromosome 2, 9 and 12. Abnormal high counts are found in all of those chromosomes. The count data has been modelled with and without those high counts and it has been observed how that affects the number of enriched regions reported by the MRF model.

Chromosome	**Count corrected data**	**Original count data**	**Count corrected data**	**Original count data**
	No. of 200 bp enriched regions		**Time needed to model the data**	
Chromosome 2	12072	6788	1.363841	1.327059
Chromosome 9	9178	8027	1.013104	53.37
Chromosome 12	10585	7903	57.7018	55.63262

Table 3.7: The enriched regions found by the MRF model at 5% FDR in the ChIP-seq data with and without the unusual counts in the concentrated regions.

The experiments have been timed and it does not show any significant difference in terms of length of time to model the data with or without the high counts.

3.5 Summary

In this chapter, the Markov Random Field model that has been adapted to analyse the binding regions from ChIP-seq data has been described. The model incorporates spatial dependency between adjacent regions and also accounts for excess zeroes that are common in ChIP-seq data. It has been investigated here how the final list of bound regions can be improved by incorporating spatial dependency in the algorithm to model ChIP-seq data and how this can identify broad peak regions even if the adjacent windows have small counts which otherwise would have been ignored.

The MRF model has been compared with the mixture model using the negative binomial method. The performances of the methods have been compared in terms number of peaks generated and the quality of the peaks. Observing the bound regions produced by the two models, it has been found that the MRF model has produced more regions than the other method in all of the six experiments that have been conducted using different datasets. In six experiments, on average it has produced 34% more regions than the negative binomial method in the experiments conducted on histone protein H3. With RNA PolII, the MRF model produces around 80% more regions than the NB method. It is due to the fact that RNA PolII is known to have broad regions therefore, it is apparent without incorporating spatial dependency a model will ignore a vast number of bound regions for the proteins or markers like RNA PolII. Also, when enrichment probabilities have been generated around TSSs where RNA PolII is known to have bound to regulate gene expression, the enrichment profile estimated by the MRF looks smoother in those regions than the NB model.

It has been observed that the MRF assigns a higher probability to a region that has low tag counts but has neighbouring regions with some significant number of counts, but the mixture model assigns a very low enrichment probability to those regions, thus discards lots of regions that should have been identified as bound regions.

Bao et al (2015) compared the MRF model and NB distribution in their paper where they proposed MRF model, however they made the comparison on data they simulated and also on two transcription factors that are known to have broad enriched regions, whereas in this thesis the comparison has been applied on real datasets of different proteins (variable shaped peaks), generated in different biological and experimental conditions. The result of comparison has also been validated using known biological information.

It has also been showed how the joint modelling of the replicates where the IP efficiency of both replicates are taken into account can produce more bound regions together than if the replicates are modelled separately and the overlapped regions are considered. It has been demonstrated how removing the high (over-dispersed) counts found in a concentrated region caused by some kind of bias from the data, can significantly affect and improve the peak lists that are reported by the chosen model.

The parameters such as, size of the windows has been constant throughout this experiment. Counts summarised per fixed sized windows help analyse the distribution of the data and find out the true enriched regions. However, there is no universal rule for choosing the window size. It controls the compromise between count size and spatial resolution. For example, large window size can yield higher read counts but spatial features can no longer be distinguished [Lun et al. 2015]. Humberg et al. [2011] recommended 150 bp window size for histone markers, however as different types of epigenetic markers or proteins have been used in this thesis, 200 bp window size has been used for consistency. For spatial dependency only adjacent windows are considered for simplicity. However, in future this method can be extended for higher order spatial dependency.

Chapter 4

Relationship between gene expression and protein binding

4.1 Introduction

Epigenetic mechanisms such as histone modifications, DNA methylation coupled with other transcriptional regulatory events such as transcription factors bindings at different genomic locations control gene activity in different cell types [Jones et al. 2007; Jaenisch et al. 2008]. Transcription factors, proteins that initiate and regulate the transcription of genes, have DNA-binding domains that give them the ability to bind to specific sequences of DNA called enhancer or promoter sequences near the transcription start site (TSS) as depicted in Figure 4.1.

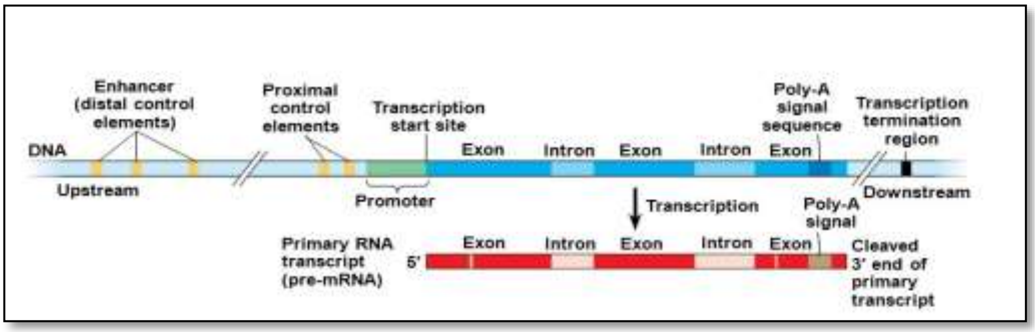

Enhancer (distal control elements)
Proximal control elements
Transcription start site
Poly-A signal sequence
Transcription termination region
DNA
Exon
Intron
Exon
Intron
Exon
Upstream
Promoter
Transcription
Poly-A signal
Downstream
Primary RNA transcript (pre-mRNA)
5'
Exon
Intron
Exon
Intron
Exon
Cleaved 3' end of primary transcript

Figure 4.1: A general structure of a eukaryotic gene with all its elements including transcription start and stop sites [Pearson Education, Inc. 2014].

Other regulatory sequences also reside within thousands of base pairs upstream or downstream from TSS [Maston et al. 2006]. That makes TSS a very vital feature to investigate in many biological, disease and developmental studies that explore the relationship between binding of proteins and gene regulation. Next generation sequencing technology such as ChIP-seq providing information about localization of binding of proteins and microarray experiment exhibiting gene expression information are both used to study such relationship.

Several attempts have been made to identify correlation between these two data platforms. Some methods concentrate on specific genomic features such as TSS, promoter, enhancer etc [Hoang et al. 2008; Markowetz et al. 2009; Nicodeme et al. 2010], while others take binding regions and gene expression data from one biological condition only [Qin et al. 2011; Guan et al. 2014]. To date the comparative analyses to determine how differential bindings are correlated with differential expression between two conditions have used very basic techniques to analyse ChIP-seq data, where absolute tag counts are used for enrichment estimation, or peak signals are collected by subtracting background using control data and thus have omitted important characteristics of ChIP-seq data. In this chapter a novel approach has been proposed to integrate gene expression and protein binding data that addresses these issues.

This chapter is organised as follows. In Section 4.2 some existing methods for integrative analysis between protein binding and gene expression data and their limitations are discussed. In Section 4.3, a novel approach has been proposed where advanced analysis result of ChIP-Seq are incorporated in the integrative analysis of protein binding and gene expression data to study the relationship between differential expressions and differential protein bindings around the transcription start sites. Section 4.4 is devoted to experimental studies where the proposed model is applied to a set of ChIP-seq and microarray data obtained at different biological and experimental conditions and the results are discussed. And finally, the work is summarised in Section 4.5.

4.2 Background

The wider research community has made several attempts to integrate protein binding and gene expression data to identify the mechanisms that control or regulate gene expression [Markowetz et al. 2010; Qin et al. 2011; Geevan et al. 2012; Guan et al. 2014]. Transcription start site has been studied extensively by biologists to investigate how different epigenetic mechanisms around this genomic region regulate genes [Roh et al. 2006; Bernstein et al. 2005; Heintzman et al. 2007].

Correlation of histone acetylation around TSS with gene expression data has been studied by Markowetz et al. [2010], however the acetylation level has been measured using the ChIP-ChIP technology and it has been concluded in the study that ChIP-seq will give greater resolution of histone acetylation profiles.

A web-server based solution called ChIParray performs an integrative analysis of ChIP technology and microarray data, to detect direct and indirect target genes regulated by a TF, using the details of the bound regions of a targeted TF under a given biological condition as input [Qin et al. 2011]. The model is limited in that it does not include differential binding information between different biological conditions.

Another web-based server, PTHGRN, analyses interaction between transcription factors and their effect on gene expression, using information on bound and non-bound regions for one biological condition [Guan et al. 2014].

A method called GEMULA, proposed by Geevan et al. [2012] uses linear models to predict TF-gene expression association or TF-TF interaction. However, they implemented binding affinity values as the predictor for classification of genes. Other attempts have also been made to infer relationship between gene expression and histone modification where absolute tag counts around a feature, such as promoter, are considered.

Several methods have estimated ChIP-seq enrichment levels for particular genes, where the Multivariate Adaptive Regression Splines (MARS) algorithm [Friedman 1991] have been applied for each estimation method using the estimation of enrichment levels as predictors and gene expression levels as responses. However, to determine enrichment levels these methods used model based methods including absolute tag counts. Each of these methods concluded that instead of tag counting method, incorporating model based approach that includes spatial distribution of enrichment improves the result of integrative studies [Hoang et al. 2011].

Most integrative methods investigate protein binding data and gene expression data in one biological condition but fail to consider how differential bindings of proteins between biological conditions may correlate with respective differential expression values. Although a number of studies have investigated such correlations, they have implemented imprecise analysis of ChIP-seq data by ignoring several important characteristics, such as overall distribution of counts, spatial dependencies of counts for neighbouring regions of the genome and the different efficiencies of individual ChIP-seq experiments. These, if not accounted for, can lead to false results, especially where differential bindings of proteins are considered between different conditions [Bao et al. 2013; Bao et al. 2015]. Several studies have suggested that it is absolutely vital to measure the ChIP-seq enrichment accurately to optimise integrative analysis, therefore a model to analyse ChIP-seq data must account for these issues.

In this study, these issues have been addressed by analysing the ChIP-seq data with an adaption of the Markov Random Field method proposed by Bao et al. [2015] that incorporates spatial dependency and ChIP-efficiency, while modelling the ChIP-seq data. After modelling the ChIP-seq data, the model estimates enrichment probabilities per fixed-length windows across genome of those factors mentioned above. After retrieving the TSS location information of each gene from the UCSC genome browser, enrichment probabilities around per transcription start site is generated and its correlations with associated gene expression values investigated. This approach demonstrates how results from advanced analysis of ChIP-seq data better define the relationship between protein binding and gene expression incorporating information from different conditions. This approach produces robust information that describes the binding profile in more detail than previously observed, and may elucidate aspects of cellular signalling mechanisms more effectively.

4.3 Method

In the proposed methodology, microarray technology is used to define the differential expression status of a set of genes between two biological or experimental conditions. Differential expression analysis is conducted on the microarray datasets and genes are identified with significant variance of expression level between two conditions. For these genes the differential expression changes are calculated. If there are technical or biological replicates available, the mean values of expression of the replicates are considered for each gene. To obtain the values denoting differential expression, the log2 normalised expression values of each gene under one condition is deducted from the respective value reported for the second condition. Therefore, given two conditions, a gene can be defined as upregulated, downregulated or unchanged.

Once the differentially expressed genes are defined, the differential bindings of a set of proteins in proximity (5000 bp up and downsteam) to the TSSs of these genes are investigated to determine significant associations with the differential expressions, between the two conditions. To extract the differential binding information, the ChIP-seq datasets for the proteins are analysed using Markov Random Field (MRF) model. The model yields posterior probability of binding per fixed-length region of the genome. Note that the size of the window is chosen by the investigator. Transcription start site information was obtained along with the precise location co-ordinates provided by the Ensemble genome database. From this list, TSSs of those selected genes are identified. Next a binding probability profile of the proteins close to the TSSs is created. For each TSS, the enrichment probability is determined from the probability result, mentioned above. Using the genomic co-ordinates, the window that includes the start coordinate is selected. The probability of that window is assigned to the start point of the TSS. For 5 Kilo base (KB) upstream and downstream regions from the start site, the mean probability of those regions is derived from the probability results. Once the profile has been estimated, the differential probability is then obtained by deducting the probability observed under one condition from the second condition, for each region. The differential expression and differential probability results are then integrated to further evaluate the correlations. Figure 4.2 gives a schematic representation of the proposed model demonstrating the correlation between differential expression of genes and differential bindings of proteins around TSS.

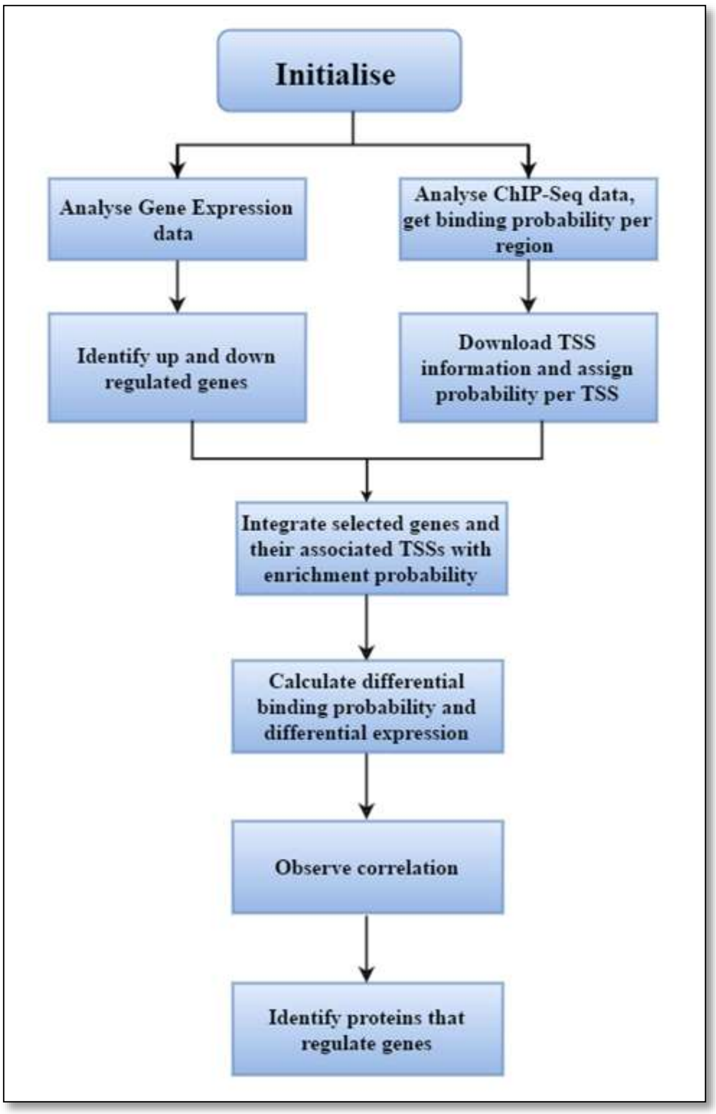

Initialise
Analyse Gene Expression data
Analyse ChIP-Seq data, get binding probability per region
Identify up and down regulated genes
Download TSS information and assign probability per TSS
Integrate selected genes and their associated TSSs with enrichment probability
Calculate differential binding probability and differential expression
Observe correlation
Identify proteins that regulate genes

Figure 4.2: Proposed model to find correlation between differential expression and differential binding.

The major steps of the model are illustrated in Algorithm 4.1. In step 1, differential expression analysis is carried out between two microarray datasets to identify a set of upregulated or downregulated genes. In step 2, the ChIP-seq datasets of any protein obtained at the same biological conditions are analysed using the MRF model, and binding probabilities of fixed-length windows are estimated. In the following step, TSS information of the identified genes is obtained and probability of binding is assigned to each of them. Differential expression and differential probability between two conditions for each gene are calculated and correlation between them investigated.

Algorithm 4.1: Pseudocode for implementation of the major steps of the model

```
Inputs

Microarray data: Two Microarray datasets obtained at two
conditions

ChIP-seq data: Two ChIP-seq datasets of a protein obtained at
same biological conditions as microarray data

TSS: Transcription start site information of the genome of
interest.

Output

Correlation: Correlation between differential expression of a
list of genes and differential binding probability of a
protein around TSS.


Function:- ObserveCorrelation(Microarray data, ChIP-seq data,
TSS): Correlation

      1. Run differential expression analysis on microarray
         datasets to determine a list of up or down regulated
         genes and save them in GeneList
      2. Analyse ChIP-seq datasets using MRF model and obtain
         binding probability per FixedLengthWindow across genome
         and save them in BindingResult with their co-ordinates.
      3. Select TSS for each gene in GeneList and assign binding
         probabilities to start point, 5KB upstream and 5 KB
         downstream regions from BindingResult using co-
         ordinates.
      4. Calculate differential expression value per gene and
         differential binding probability per TSS between
         conditions.
```

```
    5. Integrate differential expression and differential
       probability result and observe correlation.
```

Step 1: Analysis of Microarray data

Microarray data can be obtained at two different biological or experimental conditions and each experiment may have any number of biological/technical replicates. Differential expression analysis is carried out between the two conditions to identify up and down regulated genes. The pseudocode in algorithm 4.2 illustrates the implementation of differential expression analysis between two microarray datasets.

Algorithm 4.2: Pseudocode for analysis of the microarray data

```
Inputs

MicroarrayData: Two Microarray datasets

Method: A method to run differential expression analysis

Threshold: an integer to choose the cutoff for significantly
differentially expressed genes. (i.e fold2 change or pvalue)

Output

GeneList: A list of genes whose expressions have changed
significantly (decided by the threshold) between two
conditions

Function:- AnalyseMicroarrayData(MicroarrayData, Method,
Threshold): GeneList

  1. For each gene in microarray data
          Run differential expression analysis between
          datasets using Method
  2. End For
```

```
3. Select genes whose expressions have changed at least by
   the threshold.
4. Save genes in Genelist
```

Step 2: Analysis of the ChIP-seq data

ChIP-seq data for all the proteins of interest are analysed in this step. The proposed model incorporates any number of proteins. However, the biological conditions must be the same as microarray data. A MRF model (described in Chapter 3) has been used to analyse each ChIP-seq data. The output of this step is the binding probability per fixed-length region across genome. Algorithm 4.3 presents the pseudo implementation of one ChIP-seq data analysis steps. R code that implements this algorithm is attached in Appendix 1.

Algorithm 4.3: Pseudocode for analysis of ChIP-seq data

```
Inputs

ChIP-seqData:  A ChIP-seq dataset
WindowSize: An integer to specify the size the window.

Output

BindingProbability: A m*3 matrix where m represents the number
of WindowSize length windows. The 3 columns represent the
start and end co-ordinates of each window and the associated
binding probability yielded by MRF.

Function:- AnalyseChIP-seq(ChIP-seqData): BindingProbability

    1. Align ChIP-seqData
    2. Separate the chromosomes
    3. For each chromosome
           Divide the chromosome length into WindowSize windows
```

```
      Generate count of sequences per window and save in
    countData
      Model the countData with MRF model
      Obtain binding probability per window
      Save them to BindingProbability
    End For
 4. Return BindingProbability
```

Step 3: Generating the enrichment probability around TSS

In this step, TSS information is obtained from the latest release of the UCSC genome database [https://genome.ucsc.edu/] along with their precise associated co-ordinates. The TSSs are selected for the each of the differentially expressed genes identified by the investigation in the microarray analysis step. If a gene has multiple TSSs with the same start point, only one is kept and all others are discarded. For each TSS, the window is searched from binding probability result where transcription start coordinate lies and probability of that window is assigned to the start point. For 5Kb up and downstream regions from the start point, the mean probability of those regions is assigned. Algorithm 4.4 presents the pseudo implementation of generating the enrichment probability around TSS steps. R code that implements this algorithm is attached in Appendix 1.

Algorithm 4.4: Pseudocode for generating the enrichment probability around TSS

```
Inputs
GeneList: A list of genes
BindingProbability: A matrix containing binding probability
per WindowSize length window across genome
```

Outputs

BindingProfileTSS: A m x 4 matrix where m is the number of TSSs and 4 columns represents gene name, binding probability at start co-ordinate, 5KB upstream and 5 KB downstream regions

Function:- CreateBindingProbabilityTSS(GeneList, BindingProbability): BindingProbabilityTSS

```
1. Download TSS information from a genome database
2. Select TSSs associated with genes in GeneList and save
   them in TSSList
3. For each gene in GeneList
        If(number of TSS with same starting point > 1)
             Keep 1 and discard the rest
             update TSSList
   End For
4. For each TSS in TSSList
        Take start co-ordinate
        Search BindingProbability for the window where the
        co-ordinate lies
        Assign bindingProababability of the window to
        startPoint
        Take mean probability of 5KB windows upstream and
        assign it to 5KBupstreamPobability
        Take mean probability of 5KB windows downstream and
        assign it to 5KBdownstreamPobability
   End For
5. Return BindingProbabilityTSS
```

Step 4: Calculating the differential binding probability and the differential expression

The binding probability yielded by the MRF model incorporates the overall distribution of the data and spatial dependency. In order to generate differential binding probability between two conditions at any genomic location (i.e transcription start point) the difference is determined by subtracting one binding probability from another.

$$D_{prob} = Prob_{pgx} - Prob_{pgy} \tag{4.1}$$

Where p is the protein of interest, g is the genomic location, and x and y represent two biological/experimental conditions.

The expression values that are used in this experiment are log2 normalized. For each condition, if there are replicates, the mathematical mean of expression values for each gene is used. To determine differential expression values, for each gene, the log2 normalized expression value is subtracted from one condition to another.

$$D_{expr} = Expr_{gx} - Expr_{gy} \tag{4.2}$$

Where g is the gene and x and y represent two biological conditions.

Step 5: Integrating the differential expression and probability result and observing correlations

Once the differential expression result and differential binding probability result around TSS are ready, the two data are integrated using the official gene names. Each gene has a differential expression value associated with it. However as any gene can have more than one TSS, it may also have more than one differential binding probability result associated with it. From this result, correlation is observed using plots and other methods.

4.4 Results

4.4.1 Data

In this experiment, the described methodology has been applied to time-series ChIP-seq and microarray data provided by Nicodeme et al. [2010]. Both data have been collected from bone-marrow derived macrophages (BMDMs) stimulated with lipopolysaccharide (LPS) and also from samples that are stimulated with LPS but pre-treated with a synthetic compound (I-BET) that, by 'mimicking' acetylated histones, disrupts chromatin complexes responsible for the expression of key inflammatory genes in activated macrophages. The LPS stimulated datasets have been described as ‘control’ and IBET treated datasets as ‘drug’ throughout this chapter.

The time-series ChIP-seq datasets have been obtained for 6 proteins/markers. They are Bromodomain-containing protein 4 (Brd4), Acetylated Histone H4 (H4ac), Histone H3 lysine 4 tri-methylation (H3K4me3), RNA Polymerase II (RNA PolII), subunit of RNA polymerase II (RNA PolII S2) and Cyclin-dependent kinase 9 (CDK9). The ChIP-seq data have been collected at three hourly time points (0H, 1H and 4H) and microarray data at four time points (0H, 1H, 2H and 4H). For each marker/protein, ChIP-seq data has only one replicate per condition; however, microarray data has 3 replicates per condition.

The main reason for choosing these datasets has been that the study has been very extensive that includes lots of different experimental conditions such as time points and drug and control. The study also includes 6 epigenetic markers for ChIp-seq experiment and complimentary microarray data which is very suitable for the experiments in this thesis.

4.4.2 Data Pre-processing

Microarray Data analysis

Illumina beadarray gene expression datasets have been pre-processed using `beadarray` R package [Dunning et al., 2007]. After the pre-processing steps, the differential expression analysis has been performed on the expression profile using the R package `limma` [Smyth 2005].

First the data is prepared for analysis using a normalisation function that is applied to the data so that any systematic trends which arise from the microarray technology, rather than from differences between the probes or between the target RNA samples hybridized to the arrays, can be removed from the expression values. In the next step, `limma` fits a linear model to the expression data for each probe/gene. The resulting coefficients of the fitted model describe the differences between the RNA sources hybridized to the arrays.

Finally the empirical Bayes method [Smyth 2005] has been applied to compute moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes shrinkage of the standard errors towards a common value. These functions are used to rank genes in order of evidence for differential expression. The empirical Bayes method is used to shrink the probe-wise sample variances towards a common value and to augment the degrees of freedom for the individual variances.

In this analysis, the differential expression value for each gene between drug and control datasets at each time point has been investigated. The differential expression analysis has also been performed to check how LPS changes the expression values of the genes between 0 and other (1, 2 and 4 hour) time points. The design of the analysis is as follows.

To find out the effect of LPS, the LPS induced samples taken at 1, 2 and 4 hour have been compared with sample taken at 0 hour. To observe the effect of IBET treatment on LPS induced genes, LPS+IBET treated samples have been compared with control samples (LPS only) at each time point.

From 25,000 genes/probes, the genes that have changed at least 2-fold between two conditions have been selected as significantly differentially expressed genes. From the analysis 836 genes were defined as up regulated by LPS at 4 hour time point. Among these, 183 genes were identified as downregulated by IBET treatment. 4 hour time point has been selected as the number of genes affected by the LPS and IBET are greater at this time point than the previous time point. These subsets of genes are the genes that are further investigated in this analysis. Figure 4.3 shows the number of differentially expressed genes found in different experiments at 1 and 4 hour time points.

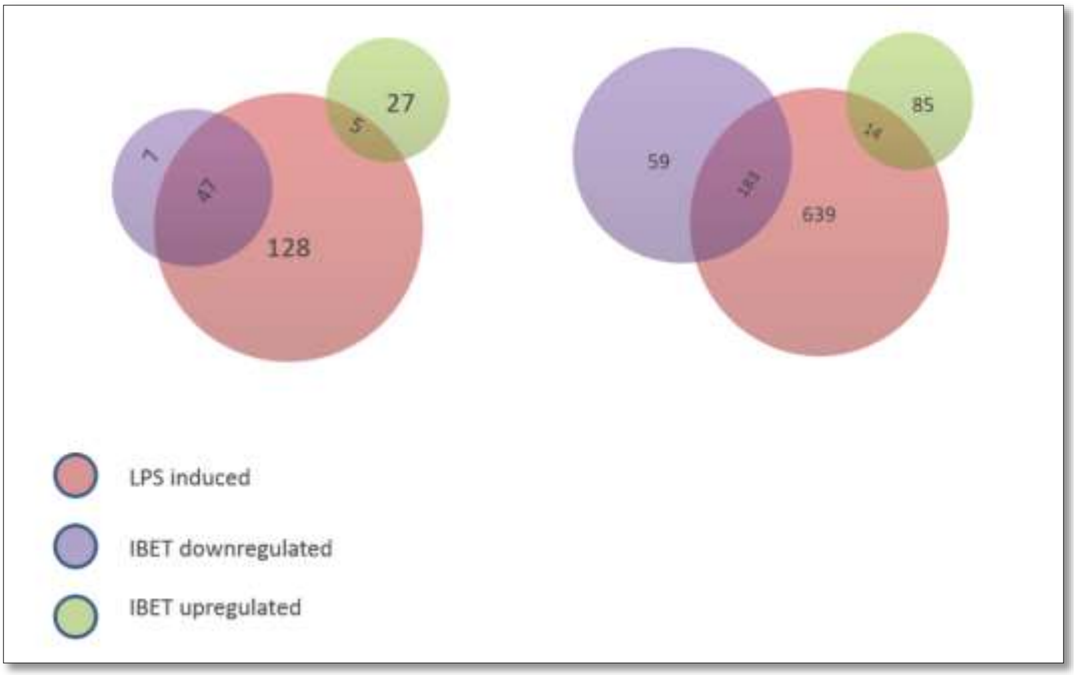


Figure 4.3: Upregulated and downregulated genes by LPS and IBET at 1 hour (left) and 4 hour (right) time points.

ChIP-seq data analysis

The ChIP-seq datasets were each aligned against mouse genome (version mm9) using `bowtie` [Langmead et al. 2009] and only the uniquely mapped reads have been retained for further analysis. The reference genome was obtained from UCSC Genome Browser. Note that all chromosomes have been analysed separately, as the count of sequence distribution varies across genome. For each dataset, all 19 chromosomes have been separated and each chromosome has been divided into 200bp long windows. The counts of sequences have been then generated per window. This count data has been supplied as the input for MRF model and analysed as described in Chapter 3. The model estimates posterior probability of enrichment per 200bp region. The enrichment/binding probability results of all 19 chromosomes have been assembled together for each dataset.

Generating enrichment probability around TSS

The transcription start site (TSS) information of chromosome 1 to 19 of mouse genome using NCBI mm9 assembly has been downloaded from UCSC database. Each TSS has its location information in terms of the txStart (transcription start coordinate) and the txEnd (transcription end coordinate) and also the associated gene information attached to them. Many genes have several TSSs with the same txStart co-ordinate. In that case, only one TSS is retained. For 19 chromosomes, information about 55419 TSSs has been downloaded from UCSC. After the selection process, 37351 TSSs have been used for the subsequent experiments. From this list, the TSSs have been selected that are associated with the genes that are identified in the microarray data analysis step. The enrichment probability has been then assigned to each start point of the TSS and also to the regions, 5 KB upstream and downstream from the start point as described in Algorithm 4.4. Each TSS is associated with a gene and the official gene symbols are used to integrate this data with microarray result.

Integration of the data

From Microarray data the log2 normalized expression values of 836 genes that are up regulated by LPS at 4 hour time point and 183 of those genes that are downregulated by IBET treatment have been selected for investigation for this study. Each microarray experiment has 3 replicates and averaged expression value of three replicates per probe/gene has been considered. To determine the differential expression values, equation 4.2 has been used for each gene. From the TSS result, the differential probability has been also calculated per TSS using equation 4.1. These two data are then integrated using the official gene symbols.

4.4.3 The results of observing correlations

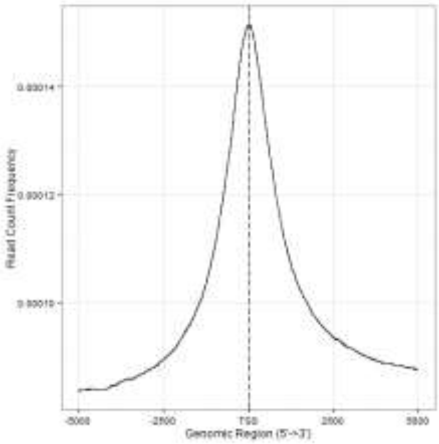
In order to understand how all the proteins bind around transcription start sites of the genes, binding regions of each protein around transcription start and upstream and downstream of TSS have been plotted. R package `ChIPseeker` [Yu et al. 2015] has been used to create these plots [Figure 4.4].

From the microarray data, it has been observed that LPS has upregulated the most number of genes at 4 hour time point. That is why ChIP-seq data for all markers collected in the same biological condition are selected for these plots to visualise how these proteins bind around TSS at that particular condition. The peaks have been obtained by modelling the ChIP-seq data by MRF model as described above. Table 4.1 shows the number of 200 bp long peaks that are found at 5% FDR for each protein.

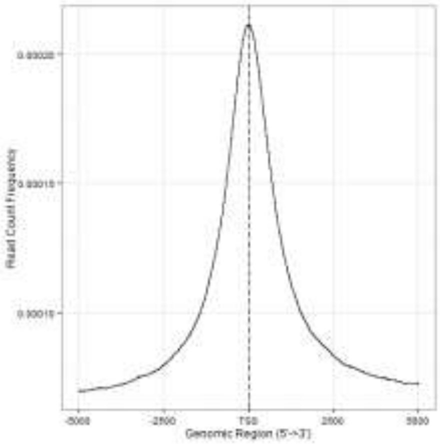
Proteins	The number of enriched regions
RNA Polymerese II	705177
RNA Polymerese II S2	1282471
H3k4me	327854
H4ac	218960
Brd4	135101
CDK9	105905

Table 4.1: The number of the enriched regions found at 5% FDR at LPS stimulated data at 4 hour time point.

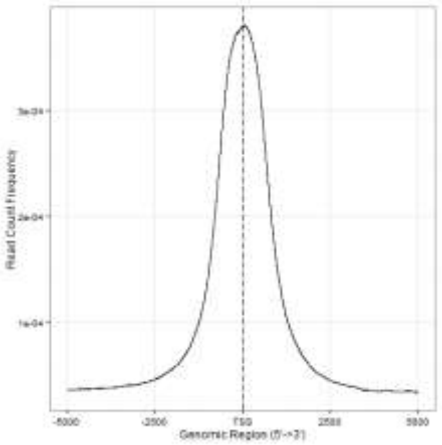
To calculate the profile of ChIP peaks binding to the TSS regions, the data has been processed in the following ways. In `ChIPseeker`, the tagMatrix (frequency of counts in any given genomic location) generated for such plots is not restricted to just TSS regions, and the user can define the upstream and downstream areas they want to include in the plots. For these plots, 5kb upstream and 5Kb downstream regions from the transcription start have been selected to be included in the plots. `ChIPseeker` plotting function aligns the peaks that are mapped to these regions and generate the tagMatrix.



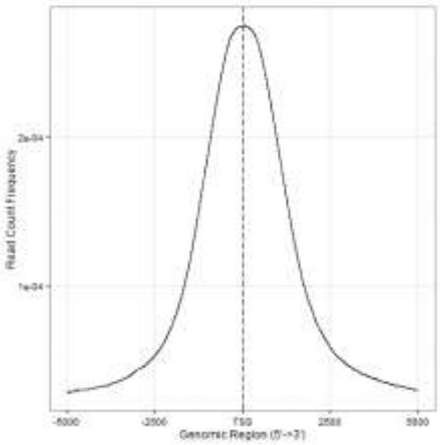
PolII S2



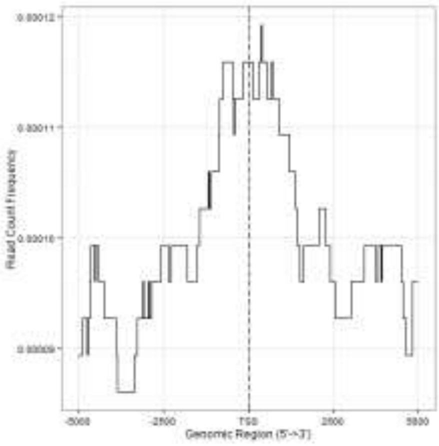
PolII



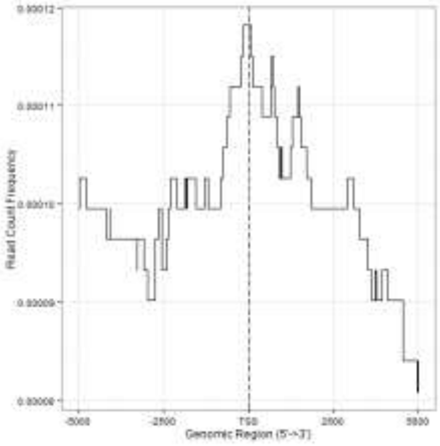
H4ac



H3K4me3



Brd4



CDK9

Figure 4.4: Average profile of the ChIP peaks binding of 6 proteins to TSS region at 4 hour time point with LPS.

In Figure 4.4, it has been observed that the peaks of RNA PolII, RNA PolII S2, H3K4me and H4ac show smooth profiles around the TSSs. Brd4 and CDK9 show significant percentage of peaks binding around TSS, but the numbers of peaks observed with these two markers are generally less than that observed for the other four markers. Also the peaks around TSS for CDK9 and Brd4 are not as smooth as other 4 markers.

After creating the binding profile for all proteins in question around TSS as described above, how these bindings correlate with gene expression data has been investigated. The primary objective is to determine whether differential expression i.e upregulation or downregulation of genes between different biological conditions significantly correlates with upregulation or downregulation of bindings around TSS for these proteins.

First the 183 LPS induced genes downregulated by IBET treatment were investigated. As some genes have more than 1 TSS, 227 transcription start sites were associated with the 183 genes. These genes are downregulated from control data at 4H to drug data at 4H by IBET, therefore, the differential binding probabilities at transcription start points, 5Kb downstream and 5 Kb upstream regions, have been calculated by subtracting binding probabilities found at control data from the probabilities found at drug data at 4 hour time point. For each region, if the difference is positive, the binding probability has gone up from drug to control data, but if the difference is negative then the overall binding has been downregulated from control to drug data.

Proteins	**5Kb upstream**	**Transcription start**	**5Kb downstream**
RNA PolII	179	183	195
RNA PolII S2	199	199	201
H3K4me3	149	117	174
H4ac	114	129	127
Brd4	212	211	213

CDK9	129	107	128

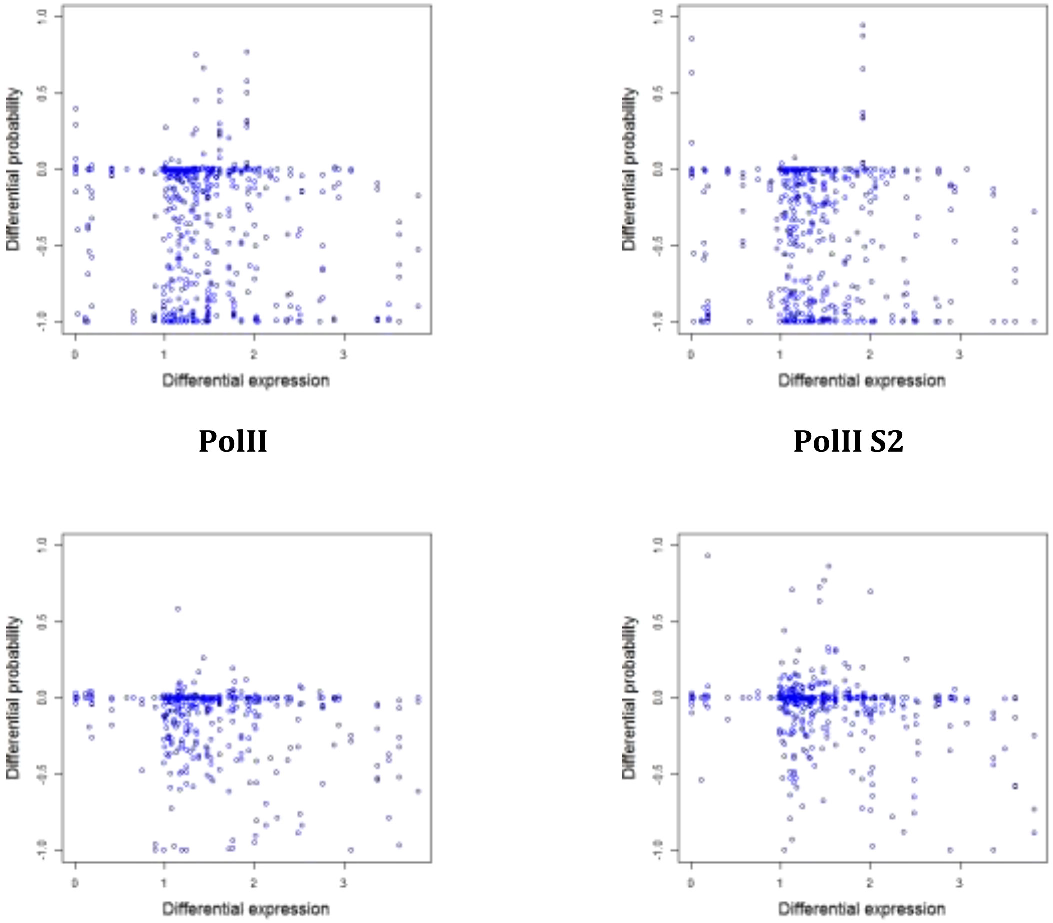
Table 4.2: The number of sites around TSSs associated with downregulated genes that show downregulation of bindings of proteins.

Table 4.2 summarises the result of the number of regions that show downregulation of probabilities for each protein. From this result, it has been observed that, the overall count has decreased for Brd4 in most of the 227 TSS regions for the downregulated genes, followed by RNA PolII S2 and RNA PolII. Which suggest positive correlation between protein binding around TSS and gene expression for these markers.

Secondly the LPS induced upregulated genes have been investigated. After the samples were stimulated with LPS, 836 genes were reported to be upregulated from 0 hour to 4 hour time-point. 989 TSSs associated with these genes were identified. Again it has been investigated how many sites correlate in terms of upregulation of gene expression. The differential binding at transcription start points, 5Kb downstream and 5Kb upstream regions are calculated by subtracting binding probabilities of control data from 4 hour to 0 hour time point. The number of sites is searched that show positive enrichment changes around the TSSs associated with the upregulated genes. From Table 4.3, it can be observed that the overall count has increased for RNA PolII S2 around most of the TSSs. For other 5 markers, significant number of TSSs associated with upregulated genes show upregulation of enrichment around them, which suggests positive correlation between protein binding around TSS and gene expression.

Proteins	5Kb upstream	Transcription start	5Kb downstream
RNA PolII	666	670	711
RNA PolII S2	852	885	891
H3K4me3	684	509	801
H4ac	541	591	582
Brd4	598	553	626
CDK9	645	531	667

Table 4.3: The number of sites around TSS associated with upregulated genes that show upregulation of bindings of proteins.

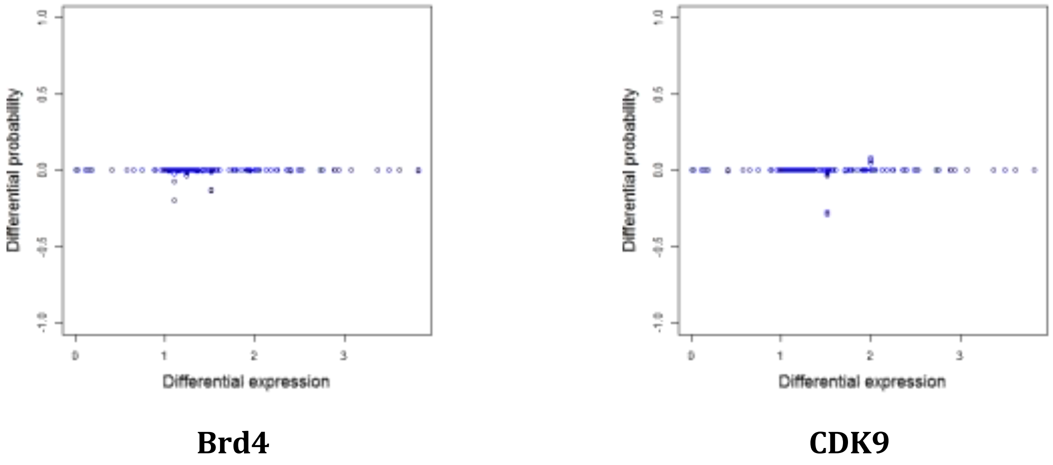


PolII

PolII S2

H3K4me

H4ac



Brd4 **CDK9**

Figure 4.5: Plots to show the correlation between downregulation of genes with downregulations of bindings.

The number of sites that shows positive association between up or downregulation of enrichment of regulatory proteins and up or downregulation of gene expression gives us partial information about the relationship between gene expression and protein binding. The genes have been selected with criteria of at least 2-fold change between two conditions. To quantify how significantly protein bindings change around TSS, relative to the change in expression values, differential enrichment and differential expression have been plotted. To visualise overall change in probability, the start site, 5KB upstream and downstream regions are plotted together.

In the plots in Figure 4.5, the x axis represents the differential expression values. Though the genes that are downregulated are selected between drug and control data; absolute values are used illustrate the degree of changes. In the y axis, the difference of probability of bindings range from 1 to -1 to show changes in both directions (up or down).

Proteins	5Kb upstream	Transcription start	5Kb downstream
RNA PolII	0.023	0.152	0.175
RNA PolII S2	0.036	0.126	0.185

H3K4me3	0.107	0.220	0.342
H4ac	0.216	0.269	0.232
Brd4	-0.039	-0.005	-0.023
CDK9	-0.052	0.001	-0.041

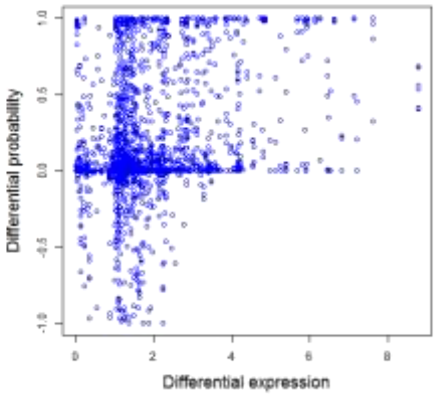
Table 4:4: The correlation results for differential expression and differential probabilities around TSSs for downregulated genes at 4 hour time point.

It is evident from the plots in Figure 4.5 that RNA PolII and RNA PolII S2 show significant downregulation of enrichment at transcription start and 5KB upstream and downstream regions that are associated with the genes that are downregulated by 2-fold due to IBET treatment at 4 hour time point. The enrichment probability of H3K4me and H4ac also show significant downregulation for some of those genes. However, though the enrichment probability around TSS for Brd4 and CDK9 change in the same direction for most of the sites as it has been seen in previous results, the change is insignificant and it can be concluded that the bindings of these two proteins do not show a significant correlation between differential expression and differential enrichment around TSS.

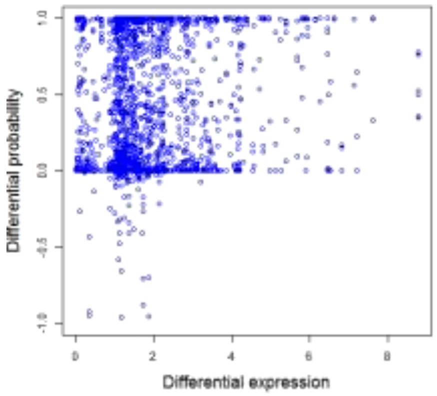
To confirm this result, Pearson correlation coefficients have been calculated for differential expression values for these downregulated genes and differential probabilities of each protein around associated TSSs. The results are summarised in Table 4.4 and the reported values suggest that, differential probabilities of H3K4me3 in the downstream region are most correlated with the gene expression variation between the conditions, while Brd4 shows negative correlation.

In the plots in Figure 4.6, with expression changes for 836 upregulated genes between 0 and 4 hour time point are in the x axis and differential bindings are in the y axis. Again it appears that both the RNA PolII and RNA PolII S2 values represent a significant upregulation of enrichment probability at transcription start and 5KB upstream and downstream regions. Enrichment probability of H3K4me and H4ac also show upregulation for some of those genes. As with the previous analysis, no significant changes can be observed in the bindings of Brd4 and CDK9 in and around TSS.

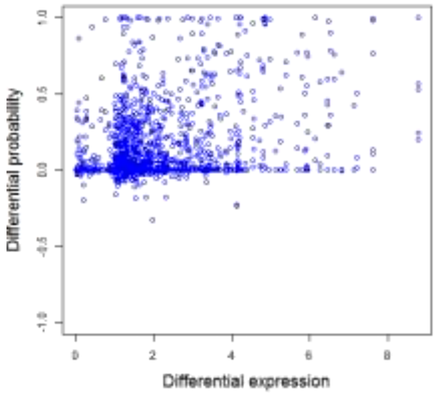
Again Pearson correlation coefficients have been calculated for differential expression values for these downregulated genes and differential probabilities of each protein around the TSSs. The result is summarised in Table 4.5 where it is evident that, differential probabilities of H3K4me3 in the downstream region are most correlated with the gene expression variation between the conditions, while Brd4 shows the least correlation.



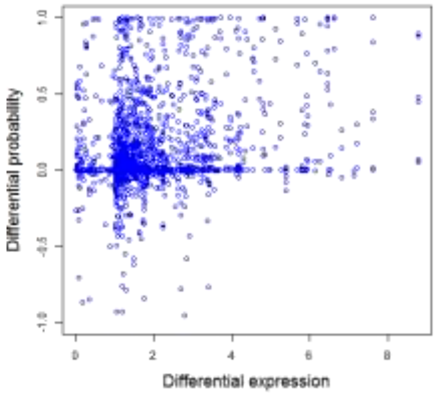
PolII



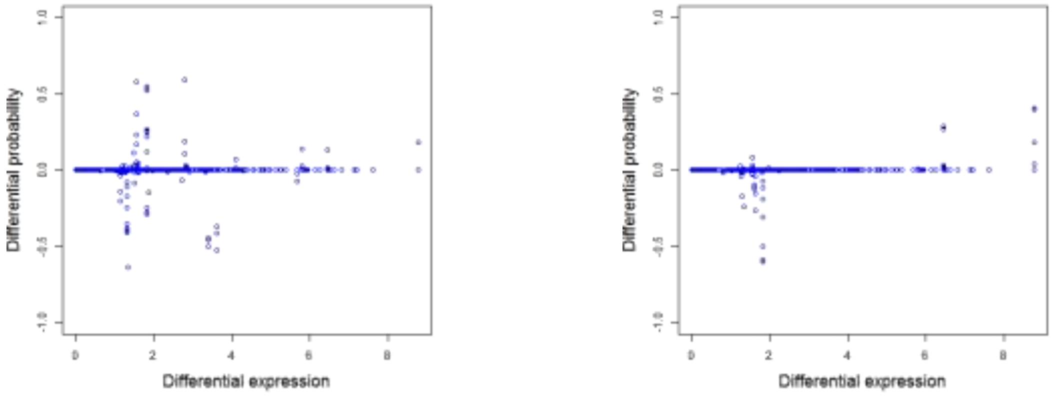
PolII S2



H3K4me



H4ac



Brd4 **CDK9**

Figure 4.6: Plots to show the correlation between upregulation of genes with upregulations of bindings for 6 proteins.

Proteins	5Kb upstream	Transcription start	5Kb downstream
RNA PolII	0.304	0.233	0.327
RNA PolII S2	0.223	0.173	0.206
H3K4me3	0.304	0.273	0.439
H4ac	0.223	0.202	0.331
Brd4	-0.020	0.0362	0.0360
CDK9	0.011	0.140	0.174

Table 4:5: The correlation values between differential expression and differential probabilities around TSSs for upregulated genes at 4 hour time point.

Comparing the plots and correlations coefficient results, a simple focus on the direction of change in protein binding probabilities may lead us to wrong conclusions concerning those proteins/markers most significantly associated with gene expression changes. For example, the RNA PolII and RNA PolII S2 plots highlight a greater change in probabilities, while correlation estimates indicate that the binding probability changes for H3K4me3 in downstream region most strongly correlate with the gene expression changes.

Proteins	LPS upregulated genes	IBET downregulated genes
RNA PolII	0.144	-0.279
RNA PolII S2	0.313	-0.394
H3K4me3	0.056	-0.058
H4ac	0.089	-0.046
Brd4	-0.001	-0.002
CDK9	-0.001	-0.001

Table 4.6: The average changes in protein binding probabilities around TSSs of upregulated and downregulated genes at 4 hour time point.

Therefore, in order to look at the overall changes in protein binding around TSS associated with the genes that have been investigated, average change of probability around TSS for each protein has been generated. Table 4.6 summarises the result. The mean changes of the bindings of the proteins support the same conclusion that that have been summarized from the plots. It can be concluded that, average probability changes of RNA PolII S2 around TSSs are greater than any other 5 markers for both upregulated and downregulated genes followed by RNA PolII. Brd4 shows the least change in probabilities around TSSs for both upregulated and downregulated genes.

4.5 Summary

The methodology presented in this chapter clearly illustrates how the analysis result of ChIP-seq data obtained by using an advanced method that incorporates important characteristics of such data i.e spatial dependency, overall distribution etc. can be incorporated in the integrative analysis of protein binding and gene expression data. The method demonstrates how the information retrieved from the ChIP-seq in terms of enrichment probability can be used to investigate the correlation between differential expression and differential bindings. The enrichment probability has been generated around genomic feature, TSS in the study. However, given location information, the proposed model can be used for any number of genomic features to study the relationship between gene regulation and protein binding around them. Different biological conditions such as time, treatment/non-treatment are also incorporated in the study.

The method has been applied to a rich set of ChIP-seq data of six proteins and microarray data that includes a range of different biological and experimental conditions. From the results it can be concluded that the bindings of RNA PolII and RNA PolII S2 around TSS show the most correlation with both upregulation and down regulation of gene expression. Histone methylation and histone acetylation, represented by bindings of H3K4me3 and H4ac, also show positive correlation. Pearson correlation estimates indicate that differential probabilities of H3K4me3 around 5Kb downstream show the most correlation with gene expression variation for both upregulated and downregulated genes. However, bindings of Brd4 and CDK9 around TSS do not show significant correlation with either up or down regulation of gene expression.

These findings are in agreement with several literature reports that describe how that RNA PolII and RNA PolII S2 both are observed to be correlated with active genes. High enrichment of RNA PolII and RNA PolII S2 are also observed around TSS of active genes [Sun et al. 2011]. Histone acetylation (H4ac) and histone methylation (H3K4me3) are considered reliable epigenetic regulator of transcriptional activation. It has also been reported that H3K4me3 binds heavily around the transcriptional start sites (TSSs) of genes, while the enrichment of H4ac is slightly lower around the area. Enrichment of both H3K4me3 and H4ac around TSSs are positively related to the extent of gene activity. Furthermore, enrichment of H3K4me3 occurs just downstream from the TSS, with lower levels of enrichment of H4ac occurring farther downstream [Koch et al. 2007].

In summary, these combined literature observations support the results obtained in this experiment. 88% of TSSs associated with downregulated genes and 90% of TSSs associated with upregulated genes have shown downregulation and upregulation of RNA PolII S2 bindings respectively. 82% of TSSs associated with downregulated genes and 69% of TSSs associated with upregulated genes have shown downregulation and upregulation of RNA PolII bindings respectively. Both H3K4me3 and H4ac markers also exhibit positive correlation with gene expression around TSS and the enrichment of H3K4me3 and H4ac change more in the downstream regions from TSS with the upregulation of gene expression as suggested in the literature.

Chapter 5

Prediction of gene activity using protein binding profile

5.1 Introduction

Structural genes, that code for amino acid sequences, can be described in terms of several integral components such as introns, exons, transcription start sites, promoters, enhancers and silencers. The pattern of epigenetic modification distributions across the genome and within different mammalian cells indicates that these genomic features combine with biochemical modifications of the DNA molecule define epigenetic mechanisms. To understand the gene regulation mechanisms and biological significance of epigenetic marks, it is necessary to identify the distribution of epigenetic modifications and bindings of different regulatory proteins. This means that where they occur (globally or regionally at which genomic features) among different tissue or cell types and when they occur (different biological conditions such as normal development or disease processes) need to be investigated. Studying the binding profiles of different proteins and gene expression together can tell us how different genomic features and also other factors such as biological conditions and time factors play the part in the regulation of genes. Figure 5.1 shows how regulatory proteins and transcription factors bind to different components of the genome such as promoters and enhancers to initiate gene expression.

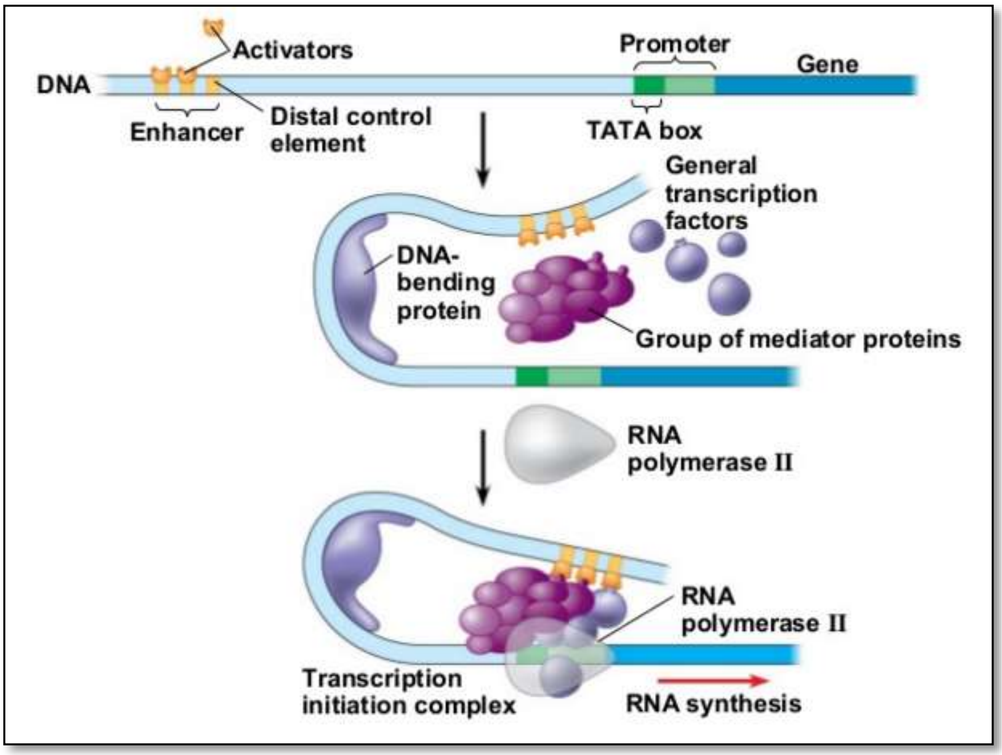


Figure 5.1 Schematic representation of how regulatory proteins bind at different genomic locations to initiate the transcription [Initiation of the transcription].

The field of gene regulation has made significant advancement recently as different large-scale genome projects have made progress in annotating genome wide protein coding regions of different species and have made these annotation databases available to the research community [Douglas 2009]. With the advancement of next generation sequencing technology, biologists can now look closely at how regulatory proteins binding at different features such as promoters, exons, introns, enhancers may impact on gene regulation. The computational biology community has also proposed methodologies and tools to integrate protein binding and gene expression data to identify causal relationships between the two, however those methods primarily focus on protein binding located at common genomic features such as promoters or transcription start sites. Furthermore, analysis of next generation sequencing data have been very basic in those studies that do not fully exploit most of the information the NGS can offer [Li et al. 2015]. Some studies [Markowetz et al. 2010; Geevan et al 2012] have shown how classification techniques may be used to find relationship between epigenetic mechanism and gene expression, but again focusing on only one single feature.

This chapter is organised as follows. In Section 5.2, methodologies and tools for integrative analysis between protein binding and gene expression data as well as the gap in the field are discussed. In Section 5.3, a methodology has been proposed to investigate how predictive the binding profile of different regulatory proteins at different genomic locations across genome is of gene activity that integrates different advanced machine learning techniques. Section 5.4 is devoted to experimental studies where the proposed model is applied to a set of ChIP-seq and microarray data obtained at different biological and experimental conditions and the results are discussed. Finally, the work is summarised in Section 5.5.

5.2 Background

Several groups have reported methods to integrate protein binding and gene expression data to identify the mechanisms that control or regulate gene expression. For example, the correlations of histone acetylation around TSS with gene expression data have been explored by Markowetz et al. [2010]. Though differential bindings and differential expressions between different biological conditions are considered in the study, the acetylation level has been measured using the ChIP-ChIP technology and it has been concluded ChIP-seq will give greater resolution about the histone acetylation profile.

Some web-based server solutions are also available today that look for interactions between transcription factors and their effect on gene expression, by using information on bound and non-bound regions. However these approaches have only considered the binding at promoter and transcription sites [Qin et al 2011; Guan et al 2014].

A linear model called GEMULA has been proposed by Geevan et al. [2012] and the model predicts TF-gene expression association or TF-TF interactions from the experimental data. It has been suggested in the literature that high-throughput techniques such as ChIP-ChIP or ChIP-Seq can be very useful to study transcription factors and their interaction with target genes, however, these techniques are very expensive and there is a practical need for methods that can predict TF–TF interactions from gene expression or DNA sequence data alone. Using TF binding affinity at gene promoter, the model describes underlying gene expression variation, hence finds out association between how TFs interacts to regulate gene. However the approach too, only considers the genomic location promoter.

Along with transcriptions start sites, promoters and enhancers, biologists have suggested that other genomic regions such as introns and exons are very important for gene regulation. Importance of the number and length of exons and introns as regulatory players has been described in several studies [Nott et al. 2003; Heyn et al. 2014] Current opinion is that after the transcription is initiated, elongation of RNA PolII can be influenced by density of exons which is due to the fact that RNA PolII pauses over exons during gene regulation [Heyn et al. 2014]. It has also been found that the first exons are shown to have more defined peaks of activating histone marks closer to the transcription start sites (TSS) and enhance the transcription accuracy. It is evident that to fully understand the mechanism of how the genes are regulated by the bindings of proteins, binding locations at exons, introns, promoters along with other genomic features should be included in subsequent molecular models of transcription.

Here, a technique has been proposed where integrative analysis of protein binding and gene expression data includes binding locations at different genomic features such as exon, intron, promoter, distal intergenic region etc. It has also been shown how dynamic interactions between regulatory proteins and gene expression can be explained by integrating sets of genes regulated at successive time-points and different biological or experimental conditions which makes it possible to answer not only what proteins might be regulating genes but also where and when they bind to do so. Several classification techniques have been used to find out the associations between protein binding profiles across genome and underlying gene expression variations.

5.3 Method

In the proposed methodology microarray technology is used to identify expression status of a set of genes at a biological or experimental condition. A gene can be ‘active’ or ‘inactive’. However, in comparison to another biological condition, it can be ‘upregulated’ or ‘downregulated’. Then ChIP-Seq data is used to create binding profiles of a set of proteins for those genes at the same condition. The binding profiles indicate whether the proteins bind at those genes and if they do, which the genomic features they bind to. The binding profile and gene status data are integrated and modelled using different classification techniques. How well the protein binding profiles can predict the gene status is then investigated. Observing the performance of different models, proteins and genomic features performing well in predicting gene status among all the variables are identified. Figure 5.2 gives a schematic representation of the proposed model demonstrating the relationship between gene expression and protein binding.

The model is illustrated as pseudo code in Algorithm 5.1. In Step 1, microarray data is analysed to identify a set of genes with their activity status. In Step 2, the ChIP-Seq datasets for a list of proteins obtained under the same biological conditions are analysed to identify their binding regions. In Step 3, protein binding profiles at different genomic locations for the selected genes are created from the annotated binding regions or peaks. In Step 4, these two sets of data are modelled with different classification techniques and finally in Step 5, the corresponding prediction performances are observed and evaluated where the proteins and genomic features closely related to gene expression are identified.

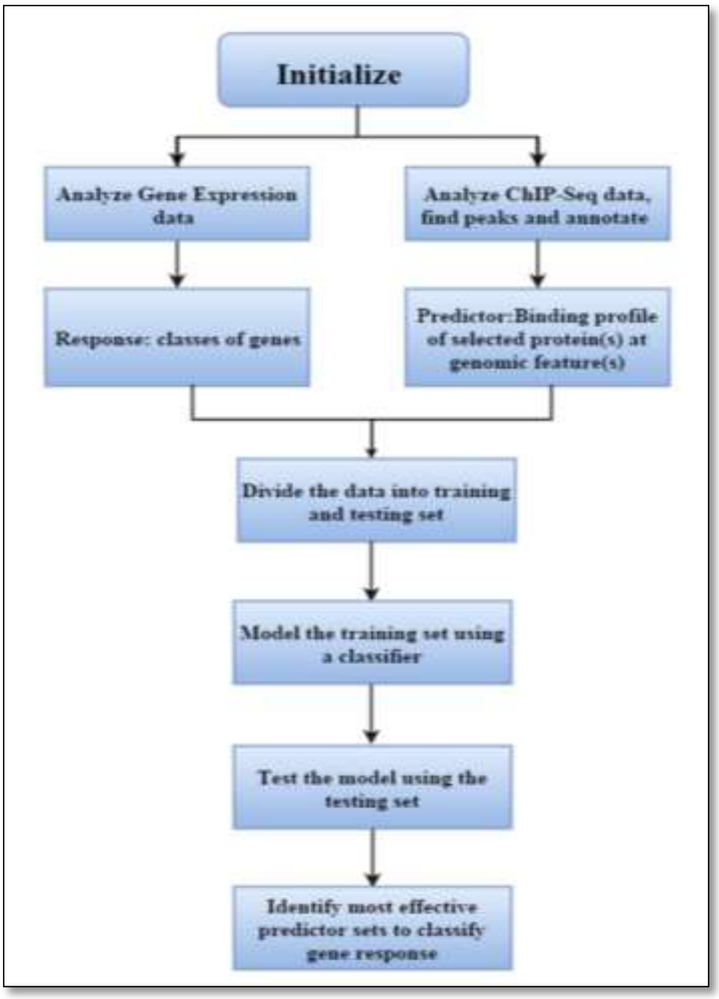

Initialize
Analyze Gene Expression data
Analyze ChIP-Seq data, find peaks and annotate
Response: classes of genes
Predictor:Binding profile of selected protein(s) at genomic feature(s)
Divide the data into training and testing set
Model the training set using a classifier
Test the model using the testing set
Identify most effective predictor sets to classify gene response

Figure 5.2: Proposed model to predict gene response from binding profile of proteins at different genomic features.

Algorithm 5.1: Pseudocode for implementation of the major steps of the model

```
Input

Microarray data: A list of Microarray datasets
ChIP-Seq data: A list of ChIP-Seq datasets

Output

ImportantFeatures: A list features (proteins and genomic
features) that can predict gene response.
```

```
Function:- IndentifyRelationship(Microarray data, ChIP-Seq
data): ImportantFeatures
      1. ANALYSE microarray data to get status of a set of genes
         and save them in GeneList
      2. ANALYSE ChIP-SEQ data of all proteins in ProteinList
         to get annotated peaks
      3. CREATE BindingProfile for each protein in ProteinList
         for all Genes in GeneList
      4. RUN classification using BindingProfile as preditor and
         status of genes in GeneList as response
      5. Identify features that can predict gene status well
         observing classification performance and save them in
         ImportantFeatures
```

Step 1: Analysis of Microarray data

Microarray data can be obtained at different biological conditions and each experiment can have any number of biological/technical replicates. Differential expression analysis is carried out between two biological conditions to identify up and down regulated genes. Based on a single data set, genes can also be classified as active or inactive. For each gene in the list a class, 0 or 1 is assigned to represent its status (i.e. upregulated/downregulated or active/inactive). The status of these genes is used as response variable in the classification step. The pseudocode in Algorithm 5.2 illustrates the implementation of differential expression analysis between two microarray datasets.

Algorithm 5.2: Pseudocode for analysis of microarray data

```
Inputs

MicroarrayData: Two microarray datasets
```

```
Method: A method to run differential expression analysis
Threshold: an integer

Output

GeneMatrix: n x 2 matrix with n number of genes with their
status.

Function:- AnalyseMicroarrayData(MicroarrayData, Method,
Threshold): GeneMatrix
   1. For each gene in microarray data
           Run differential expression analysis between
           datasets
      End For
   2. Use Threshold to select genes that is significantly
      changed between conditions i.e foldchange or pvalue
   3. Assign status to those genes
   4. Save the genes along with their status to GeneMatrix
   5. Return GeneMatrix
```

Step 2: Analysis of ChIP-Seq data

ChIP-Seq data for all the proteins of interest are analysed in this step. The proposed model can incorporate any number of proteins. However, the biological conditions need to be the same as the microarray data. A peak calling method of choice is used to locate the genomic regions that are bound by the protein in each ChIP-Seq data. The peak calling method should consider all the characteristics of ChIP-Seq data such as spatial dependency, IP efficiency, excess zeroes while modelling the data. Once the binding regions of all proteins are identified by the model, they are annotated to their proximal genes and genomic features using a genomic database. Algorithm 5.3 presents the implementation of ChIP-Seq data analysis steps.

Algorithm 5.3: Pseudocode for analysis of ChIP-Seq data

```
Inputs

ChIP-SeqData:  A list of aligned ChIP-Seq datasets
Method: A method to model the data
GenomicDatabase: A database containing loci information of
genes and genomic features.

Output
AnnotatedPeakList: A list of annotated peaks of all ChIP-Seq
data in ChIP-SeqData

Function:- AnalyseChIP-Seq(ChIP-SeqData, Method,
GenomicDatabase): AnnotatedPeaks

  1. For each dataset in ChIP-SeqData
           Model the data with Method
           Select peaks
           Annotate the peaks with nearest genomic features
           (i.e promoter, exon, intron, TSS etc.) and genes
           using GenomicDatabase
           Save the peaks in AnnotatedPeakList
     End For
  2. Return AnnotatedPeakList
```

Step 3: Creating Binding profile of proteins

Once the protein binding regions are identified and annotated, the binding profile of the proteins for data integration is generated. The method for creating the binding profile is as follows. Assume a set of m genes to have been selected in the microarray data analysis step in a biological condition c. Say, annotated binding regions of p proteins are identified in the ChIP-Seq analysis step. Each binding region is annotated to the nearest gene, represented by gene symbol and genomic feature. Again let f, the number of genomic features, be included in the study. The binding profiles of p proteins for m genes and f genomic features are created that take the form $X_1, \ldots \ldots, X_p$ where, X_{ijkc} represents binding status (1 or 0) of protein j to the feature k of gene i at an biological condition c. The implementation of the step of creating the binding profile is illustrated in Algorithm 5.4.

Algorithm 5.4: Pseudocode for creating binding profile of proteins

```
Inputs
AnnotatedPeakList: A list of annotated peaks of proteins to be
investigated
GeneList: A list of genes for which binding profile of
proteins need to be created
GenomicFeatureList: A list of genomic features

Outputs
BindingProfile: A m x p*f matrix where m is the number of
genes, p is the number of proteins and f is the number of
features included in the experiment.

Function:- CreateBindingProfile (AnnotatedPeakList, GeneList,
GenomicFeatureList): BindingProfile


   1. Populate BindingProfileMatrix
   2. For each Genomicfeature in GenomicFeatureList
            For each Annotatedpeaks in AnnotatedPeakList
                 For each gene in GeneList
                 Populate BindingProfile
                     If a peak in Annotatedpeaks found in
                 near gene and GenomicFeature
                             put 1
                      Else
                             put 0
            End For
       End For
  End For
3. Return BindingProfile
```

Step 4: Run classification

For classification purpose, the binding profile of the p proteins for m genes and f genomic features is used as predictor and status of m genes is considered as response variable. Note that this model could be extended to include more than one biological condition. In such a scenario, assume that there are c number of biological conditions and from each there is a set of genes with their activity status. For each set s_l, where l represents the experimental condition, the binding profile of the proteins need to be created from the same biological condition l. The binding can then be integrated with their complementary gene status for classification. Algorithm 5.5 demonstrates the classification process, in accordance with the proposed model, using the integrated data.

Algorithm 5.5: Pseudocode for running classification techniques

```
Inputs
Predictor: it is an m x n matrix where each column represents
a protein, binding at a genomic feature in a biological
condition and each row represents a gene.
Response: This is a matrix with same index (genes) Predictor,
but with 1 column. The column represents status of the gene
Classifier: A classifier of choice
```

Outputs

Performance: How well Predictor can predict Response in 10-fold cross validation.

ImportantFeatures: The important features in Predictor among all features that can predict the gene status well.

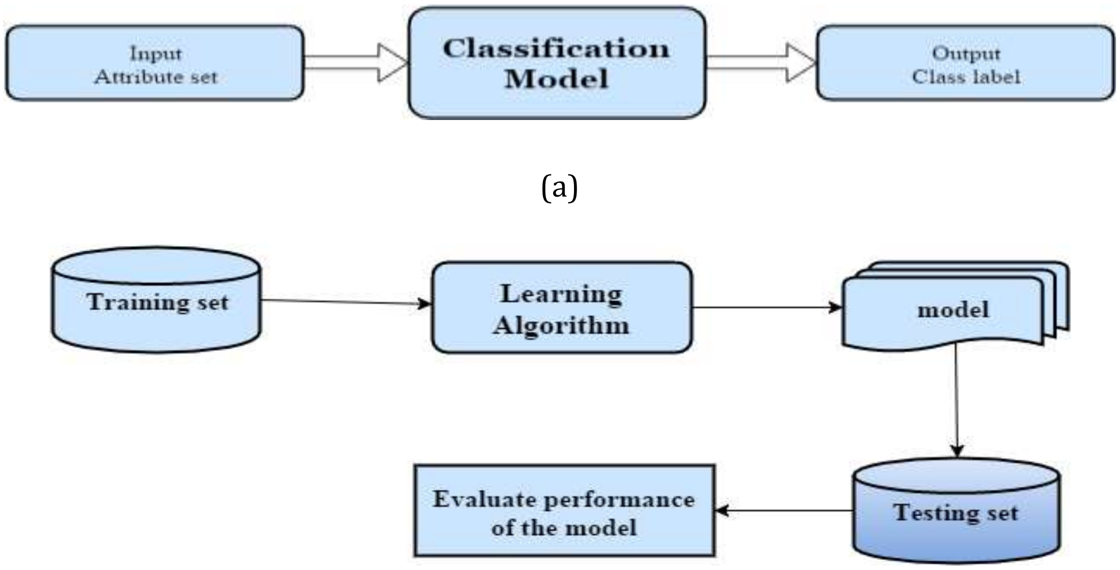
Function:- RunClassification(BindingProfile, GeneStatus, Classifier): (Accuray, ImportantFeature)

```
1. Divide the dataset in 10 Subsets.
2. For each set in Subsets
      Run classifier with binding profile as the predictor
   and gene status as the response with remaining 9 sets
      Test the model with set and observe performance
      Save performance in a list
   End For
3. Performance = mean(performances in the list)
4. Select Feature/Features that can predict the response
   well with a threshold accuracy and save them to
   ImportantFeature
5. Return Performance, ImportantFeature
```

General working principles of classification techniques are described below. The proposed model is attractive in that it could be implemented with most classification techniques. In this thesis, three popular classifiers namely neural network, decision tree and random forest have been used to demonstrate the model. Neural network is very good at finding pattern in the data and decision tree and random forest do feature selection which is very useful when many variables are available and manual selection is not possible. Also, with decision tree and random forest, one can visualise the association between variables of the data. Description of the three classifiers used in this study, is given in this section. Approaches to observe the performance of the classifiers are also included below.

Classification techniques

Classification is a data mining or machine learning technique to discover pattern and relationship in large data set. It is also a systematic approach that is used to predict class labels for data instances. Examples include decision tree classifiers, rule-based classifiers, neural networks, support vector machines, and Naïve Bayes classifiers. Briefly, each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data.



(a)

(b)

Figure 5.3: (a) Workflow of building a classification model. (b) A general Approach of how to evaluate performance of a classification model.

To have any utility the model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records of an independent test dataset. Therefore a key objective of the learning algorithm is to build models with a strong generalisation capability, i.e models that accurately predict the class labels of previously unknown data. Figure 5.3 shows a general approach for solving classification problems.

Generally, for classification, a training set consisting of records whose class labels are known must be provided. The training set is used to build a classification model, which is subsequently applied to the test set, which consists of records with unknown class labels. Below key features of the classification techniques used in this analysis have been described.

Neural Network

An artificial neural network (ANN) or neural network (NN) [McCulloch et al. 1943] is a non-linear statistical data modelling tool that is inspired by the way brain processes information. The novel structure of its information processing system is capable of learning relationships among variables and finds patterns in the data. A neural network consists of an interconnected group of units and like biological systems; learning involves adjustments to the connections between these units. Figure 5.4 shows a connected multilayer neural network with one input layer, four hidden layers and one output layers.

In a typical neural network, the input units receive various forms of information and as it passes through the inner layers, the network learns, recognises and processes this information and then produces a signal to the output units. Most neural networks are fully connected, which means each input unit and each output unit is connected to every unit in the layers either side. The connections between units are represented by a number called a weight.

This common design of neural network is called a feedforward network. Each unit receives inputs from the units to its left, and the inputs are multiplied by the weights of the connections they move along. Every unit adds up all the inputs it receives in this way and reports an output. The neural network learns by a method called backpropagation [Werbos 1975]. In backpropagation algorithm the output generated by the model is compared with desired output. Weights are updated from the outer layer to inner layer according to the difference between actual and desired output. This is an iterative process and in time backpropagation causes the network to learn, as the difference between produced and desires output is reduced.

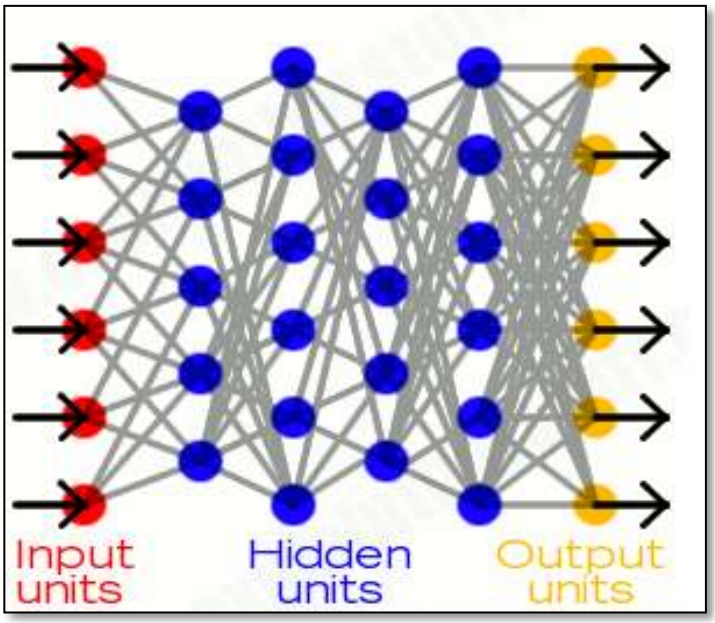


Figure 5.4: A fully connected neural network is made up of input units (red), hidden units (blue), and output units (yellow).

Let, j and i are two units in the output and input layers respectively. Firstly, the unit in the output layer computes the total weighted input $x_j$using the formula,

$$x_j = \sum_i y_i w_{ij} \tag{5.1}$$

Where y_i is the activity is level of the ith unit in the previous layer and w_{ji} is the weight of the connection between the ith and jth unit.

Secondly, the unit calculates the activity y_j using some function, usually sigmoid function, of the total weighted input.

$$Y_j = \frac{1}{1+e^{-x_j}} \tag{5.2}$$

Once the result of all the output units have been calculated, the error E is computed using the following equation,

$$E = \frac{1}{2}\sum_j (y_j - d_j)^2 \tag{5.3}$$

Where y_j is the activity level of the jth unit in the top layer and d_j is the desired output of the jth unit. The calculated error is backpropagated to adjust the weight for next iteration. Once the network has been sufficiently trained usually determined the error reaching to a threshold given by the user, it is tested with a new set of unseen examples and the performance is evaluated.

Decision Tree

Decision tree is a recursive partitioning method that helps explore the structure of a set of data and visualise decision rules for prediction of the outcome. The tree has three types of nodes: A root node that has no incoming edges and zero or more outgoing edge. Each internal node has exactly one incoming edge and two or more outgoing edges and each leaf or terminal node has exactly one incoming edge and no outgoing edges. In decision tree, each leaf node is assigned a class label. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics.

The main algorithms used for learning decision tress are ID3 algorithm (Quinlan 1986) and C4.5 (Quinlan 1993). The main steps of the algorithms are as follows: Let, there be a set of training samples T:

Step 1: If all examples in T are positive, then create YES node and stop.

If all instances in T are negative, create a NO node and stop.

Otherwise select a feature, F with values $v_1 \ldots\ldots\ldots\ldots . v_n$ and create a decision node.

Step 2: Divide the training examples in T into subsets $T_1, T_2 \ldots \ldots \ldots . T_n$ according to the values of V.

Step 3: apply the algorithm recursively to each of the sets T_i.

The algorithm incorporates feature selection methods. It searches and identifies the attribute that best partitions the data. If any attribute perfectly classifies the training sets then the algorithms halts, if not it carries on until it identifies the best set of attributes that partitions the data among the classes. The feature selection method is called information gain. This method calculates how well each attribute partitions the data. Attribute with the highest information gain is selected. Amount of information in each attribute is measure by entropy.

Given a collection S of c outcomes,

$$Entropy(S) = \sum -p(I)log2p(I) \qquad (5.4)$$

Where $p(I)$ is the proportion of S belonging to class I.

$Gain(S, A)$ is information gain of example set S on attribute A is defined as,

$$Gain(S, A) = Entropy(S) - \sum(\left(\frac{|S_v|}{|S|}\right) * Entropy(S_v)) \qquad (5.5)$$

Where,

S_v = subset of S for which attribute A has value v

$|S_v|$= number of elements in S_v

$|S|$= number of elements in S

Random Forests

Random Forests [Breiman et al. 2001] is an ensemble learning method used for classification and regression where methods generate many classifiers and aggregate their results. It is an extension of tree methods, where a number of independent trees are generated with the subset of input variables. Generally in standard tree methods, each node is split using the best split among all variables. However, in random forests method, each node is split using the best among a subset of predictors randomly chosen at that node, chosen using the least squared error.

The method of growing each tree is as follows:

Say there are N numbers of cases in the training set, and this set will be used to grow the tree.

If the number of input variables is M, then a number m which is less than M $(m < M)$, is specified and at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. During the forest growing process, the value of m is held constant.

Each tree is grown to the largest extent possible. There is no pruning.

Random forest has been a very useful method for detecting variable interactions. With this method importance of variables can be observed for classification techniques, it proves very efficient when large databases and large set of variables are involved [Ziegler et al. 2014]. The Random forest is a particularly suitable method as it can readily incorporate thousands of input variables without variable deletion. Protein-protein interactions, gene expression analysis and other image processing analysis routinely use this method [Bosch et al. 2007; Kruppa et al. 2013; Qi 2012].

In this technique, the importance of attributes is measured by two ways, one is 'mean accuracy decrease', which tests how worse the model becomes without each variable, and therefore a high decrease in accuracy would be expected for variables that are important for prediction. The 'mean decrease Gini' measures how pure the nodes are at the end of the tree so again it tests to determine the result if each variable is taken out and a high value indicates the variable is important.

Step 5: Evaluation of the performance of the classifiers

Evaluation of the performance of a classification model is based on the counts of test set correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix. Each entry f_{ij} in such table denotes the number of records from class i predicted to be of class j. For instance, f_{01} is the number of records from class 0 incorrectly predicted as class 1. Based on the entries in the confusion matrix the total number of incorrect predictions made by the model is (f_{11} + f_{00}) and the number of incorrect predictions is (f_{10} + f_{01}).

Summarizing this information with a single number would make it more convenient to compare the performance of different models. This can be done using performance metric such as accuracy, which is defined as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}} \qquad (5.6)$$

Equivalently, the performance of a model can expressed in terms of its error rate, which is given by the following equation:

$$Error\ rate = \frac{Number\ of\ wrong\ predictions}{Total\ number\ of\ predictions} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}} \qquad (5.7)$$

Most classification algorithms seek models that attain the highest accuracy, or equivalently, the lowest error rate when applied to the test set.

Cross-validation

Cross validation is the procedure where the experiment is repeated a specific number of times, say *n*. The original datasets are partitioned *n* number of times randomly, and each time different samples are used as training set and testing set. At the end the *n* results are again averaged (or otherwise combined) to produce a single estimation. In this experiment 10-fold cross validation has been used. The accuracies and errors estimated from each run are then averaged at the end to evaluate the performance of the model.

5.4 Results

5.4.1 Datasets

In this experiment, the proposed model has been applied on the time-series ChIP-seq data for Bromodomain-containing protein 4 (Brd4), Acetylated Histone H4 (H4ac), Histone H3 lysine 4 tri-methylation (H3K4me3), RNA Polymerase II (RNA PolII), subunit of RNA polymerase II (RNA PolII S2) and Cyclin-dependent kinase 9 (CDK9) proteins/markers and gene expression data provided by Nicodeme et al. [2010] as described in Chapters 3 and 4. In this study LPS stimulated ChIP-Seq datasets (control data) obtained at 0 hour and 4 hour time points and IBET treated datasets (drug data) obtained at 4 hour time point have been used. The complementary microarray data provided by Nicodeme et al. [2010] obtained at the same biological conditions have also been used to investigate at the gene expression profile.

5.4.2 Data Pre-processing

Microarray Data analysis

The microarray data analysis method is described in chapter 4. Gene expression data produced by Illumina beadarray technology have been pre-processed using `beadarray`, an R package [Dunning et al., 2007] and then analysed using linear models to define differentially expressed gene transcripts using the R package `limma` [Smyth 2005]. The design of the analysis is as follows. To define the effect of LPS, LPS stimulated expression profiles obtained at 1, 2 and 4 hour time points are compared with the expression profile at 0 hour. To define the effect of IBET treatment on LPS induced genes, LPS+IBET treated samples have been compared with control samples (LPS only) at each time point. Out of the 25,000 gene transcripts estimated in the expression profile (for simplicity probes are refered as gene transcripts throughout this chapter) the genes that have changed at least 2-folds at 4 hour time point have been selected. The analysis reported that 836 genes were up regulated by LPS between 0 and 4 hour time points. Among them 183 genes are downregulated by IBET treatment. For integrative analysis the probes have been annotated with the official gene symbols and these gene symbols have been used to integrate the data with ChIP-Seq.

ChIP-seq data analysis

The ChIP-seq reads have been aligned against mouse genome (version mm9) using Bowtie [Langmead et al. 2009] and only uniquely mapped reads have been retained for further analysis. As distribution of the counts of sequences varies from chromosome to chromosome, all chromosomes have been modelled separately. After the alignment process, 19 chromosomes have been separated and after obtaining the lengths of the chromosomes from UCSC, the counts of sequences have been generated per 200 bp region for each chromosome. The count data of each chromosome has been analysed using the MRF model as described in the chapter 3. The regions that are found enriched at 5% FDR have been selected. Table 5.1 reports the number of binding regions (200bp) found at 5% FDR for the proteins of interest at the three biological conditions used in this experiment.

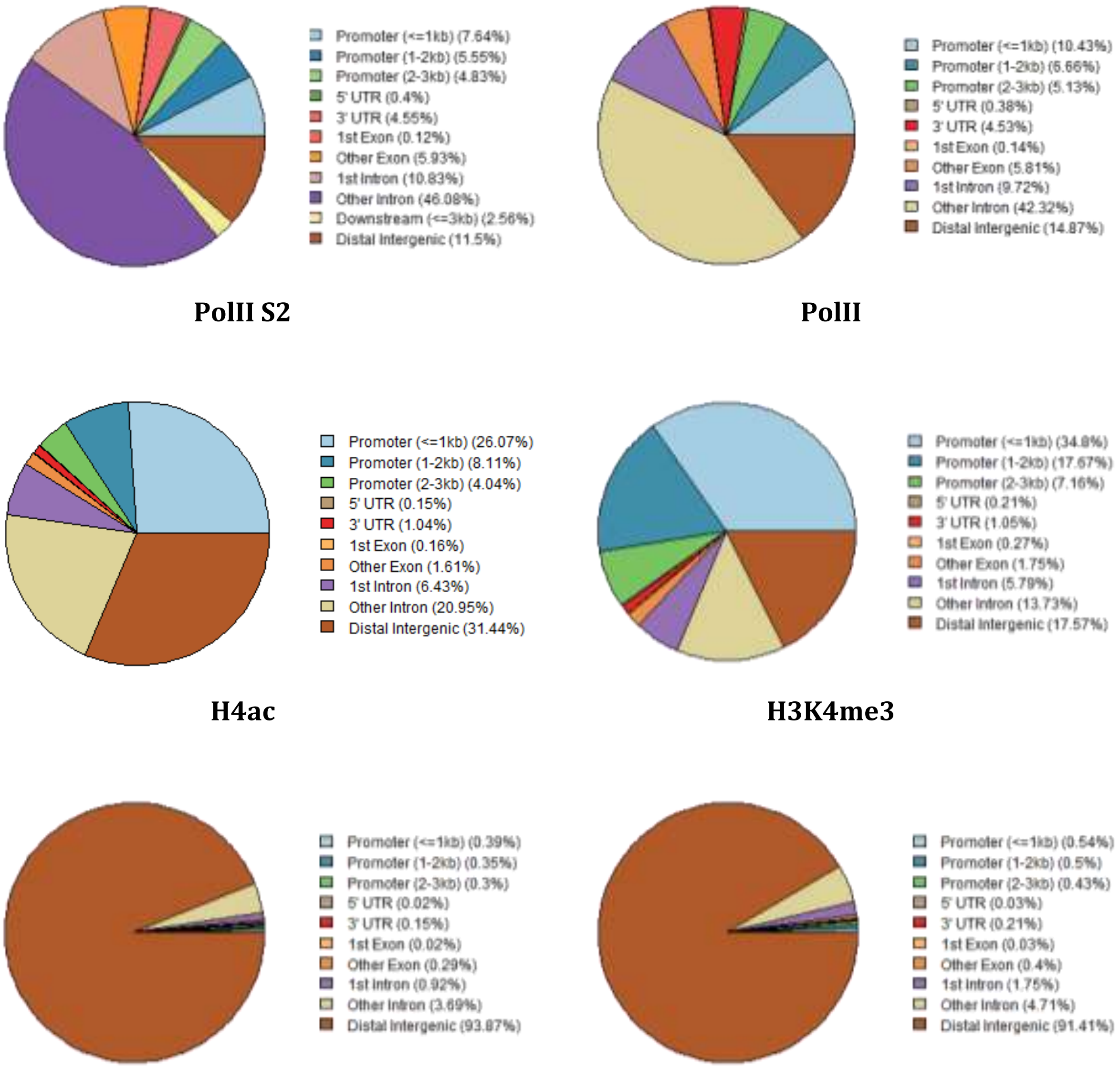
Proteins	LPS stimulated at 0H	LPS stimulated at 4H	IBET treated at 4H
	Number of 200 bp enriched regions at 5% FDR		
RNA Polymerese II	1132284	705177	625282
RNA Polymerese II S2	1020916	1282471	666159
H3K4me3	293266	327854	318679
H4ac	170087	218960	166806
Brd4	151048	135101	38831
CDK9	166600	105905	122004

Table 5.1: The number of enriched Regions found at 5% FDR from LPS stimulated data at 0 and 4 hour and IBET treated data at 4 hour time-point.

Annotation of the Peaks

All ChIP-seq datasets have been analysed using MRF model, and the results of bound genomic regions (200 bp long) obtained. These regions are then annotated with the nearest gene names and genomic features. For this purpose R package `ChIPseeker` [Yu et al. 2015] has been used. The input for the package is binding regions of the ChIP-Seq data in BED format [https://genome.ucsc.edu/FAQ/FAQformat.html#format1].

Using the annotation database for the species in question, the peaks are annotated with the gene symbol, gene name and genomic feature. For instance, if a peak is located in 5'UTR of a gene, it will be annotated with 5'UTR and the name or symbol of the gene. Annotated genomic features are promoter, exon, 5' UTR, 3' UTR, intron, and distal intergenic. The `TxDb` object containing the transcript-related features of a particular genome is used to generate the annotations which can be prepared using information from UCSC and BioMart data resources. R package `TxDb.Mmusculus.UCSC.mm9.knownGene` [Carlson et al.] has been used in the study that contains `TxDb` object of mouse species, genome version mm9.



PolII S2 **PolII**

H4ac **H3K4me3**

BRD4 **CDK9**

Figure 5.5: Feature distribution of the binding regions of the proteins.

The feature distributions have been plotted to compare summary of the peaks for each protein. Figure 5.5 shows the distribution of different features in the peaks that have been found for different proteins in IBET treated samples at 4 hour time points. The legends, attached with each pie chart summarise the percentage of features bound by the specific protein. From the plots, it is apparent that H3K4me3 and H4ac are mostly bound at promoters, CDK9 and Brd4 are mostly bound at distal intergenic. And most of the bound regions by RNA PolII and RNA PolII S2 fall in the intron regions.

Generation of protein binding profile and integration of both datasets

After annotating the peaks, the binding profile of each protein has been generated at four genomic features at different biological conditions using Algorithm 5.4. Firstly, 652 unique genes have been selected that are upregulated by LPS at 4 hour time points and classified as expressed. Their expression values are at the upper end in the profile (>9.52). They have been assigned to class 1. Another 609 genes whose expression values are at lower end in the profile (<5.72) at 4 hour time points have also been selected as lowly/non expressed genes and they have been assigned to class 0.

Proteins	Promoter	distal intergenic	Exon	Intron
RNA Pol II	0.525	0.429	0.525	0.524
RNA PolII S2	0.491	0.384	0.483	0.465
H3k4me	0.274	0.184	0.198	0.275
H4ac	0.384	0.273	0.222	0.290
Brd4	-0.002	-0.003	0.039	0.029
CDK9	0.029	-0.003	0.047	-0.013

Table 5.2: Correlation values of binding profile of different proteins at different genomic features with state of the genes.

The binding profile for these 1261 genes created using the annotated peak file for genomic features, promoter, exon, intron and distal intergenic region. The integrated data containing the binding profile of proteins along with the status of genes have been modelled using different classifiers. Prior to that, the Pearson correlation coefficients were calculated between each input variable and output. This way it can be checked how binding and non-binding of a protein/marker at a certain genomic feature is correlated to gene regulation.

Table 5.2 shows the correlation values for all proteins. From this it can be concluded that the bindings of RNA Polymerese II at promoter, intron, exon and distal intergenic show most correlations with gene regulation and Brd4 shows the least correlation.

5.4.3 Results of running classification on the data

Result from Neural Network

The integrated data from both ChIP-Seq and Microarray has been modelled with neural network to check whether the binding profile of the protein can identify the class of the genes. The `nnet` is an R package that implements feed-forward neural networks with a single hidden layer. In this study, R package `e1071` and its wrapper function for `nnet` package has been used to model the data with neural network and run 10-fold cross validation. There are options to choose the different parameter in nnet package, however as there are many variables in the data, the default option has been used to keep the model simpler. No initial weights were provided, and they were chosen randomly. The range of hidden layers have been 5, at the end model keeps the structure that gives the least error. Different combinations of proteins have been selected as predictors of the gene status for modelling with neural network. Some combinations are presented here that show the highest accuracy. The results are summed up at Table 5.3.

From Table 5.3, it can be observed that the binding profile of RNA PolII, RNA PolII S2 and H4ac bound at promoter classify the data most accurately. For all combinations, the feature promoter does better than the rest of the features. The binding profile at distal intergenic regions does the prediction least accurately among the features.

Combination of variables	Genomics features			
	Promoter	Exon	Intron	Distal Intergenic
PolII+PolII_S2+H4ac	83.16	81.90	82.18	80.35
PolII+PolII_S2+H3K4me+H4ac	82.36	82.35	82.30	80.37
PolII+H4ac	82.48	82.16	81.96	80.07
PolII+PolII_S2	82.67	81.76	82.03	80.40
PolII+PolII_S2+H3K4me	81.82	81.33	80.20	79.87
PolII_S2 + H4ac	81.97	80.19	80.88	79.44
H3K4me+H4ac	79.99	76.76	78.39	77.18

Table 5.3: Performance of neural network in terms of accuracy (%) after 10-fold cross validation.

The data containing protein binding profile at promoter, exon, intron and distal intergenic for the genes that are upregulated at 4H (LPS only) and then downregulated by IBET at 4H (LPS + IBET) has been also modelled using the neural network. There are 366 (183 +183) data points here. The protein binding profile for upregulated 182 genes has been generated from LPS induced ChIP-seq results obtained from LPS induced samples and class/status of the genes has been assumed as 1 (upregulated). For the same genes, the protein binding profile has also been generated from IBET treated ChIP-seq peaks and these genes have been classed as 0 (downregulated).

Combination of variables	Genomic Features			
	Promoter	Exon	Intron	Distal Intergenic
PolII+PolII_S2+H4ac	77.20	78.34	78.98	76.36
PolII+PolII_S2+H3K4me+H4ac	77.39	78.56	77.02	75.46
PolII+H4ac	75.34	77.89	76.91	76.07

PolII+PolII_S2	77.40	78.90	78.04	76.80
PolII+PolII_S2+H3K4me	77.09	78.68	78.45	76.73
PolII_S2 + H4ac	78.90	78.23	78.67	76.05
H3K4me+H4ac	75.74	75.67	75.08	75.86

Table 5.4: Performance of neural network in terms of accuracy (%) after 10-fold cross validation.

The combined data has been modelled with neural network to check whether binding profile of the protein can identify the classes of the genes. Same combinations of proteins have been used here as above. In this case, most of the experiments show similar accuracies (around 75-78%). However, as different combinations of predictors at different features produce the similar results, it is difficult to identify which proteins or features predict better than the rest of the variables. The results are summarised up in Table 5.4.

Results from Decision tree

To model the data with the decision tree, the dataset with 1261 genes with classes has been used. The R package, `rpart` has been used to fit the data that implements recursive partitioning and regression Trees. Here too, the class of the genes has been used as the response variable and the binding profile of the six proteins at promoter for those genes as the predictors. The Package `rpart` creates the tree with only important variables that can classify the response well. There are options to specify different parameters before running rpart. The default parameter values has been used for this experiment. No initial weights, maximum size of the tree or method have been provided. Firstly the binding profile of all proteins at promoter has been used as the input. The tree is depicted in Figure 5.6.

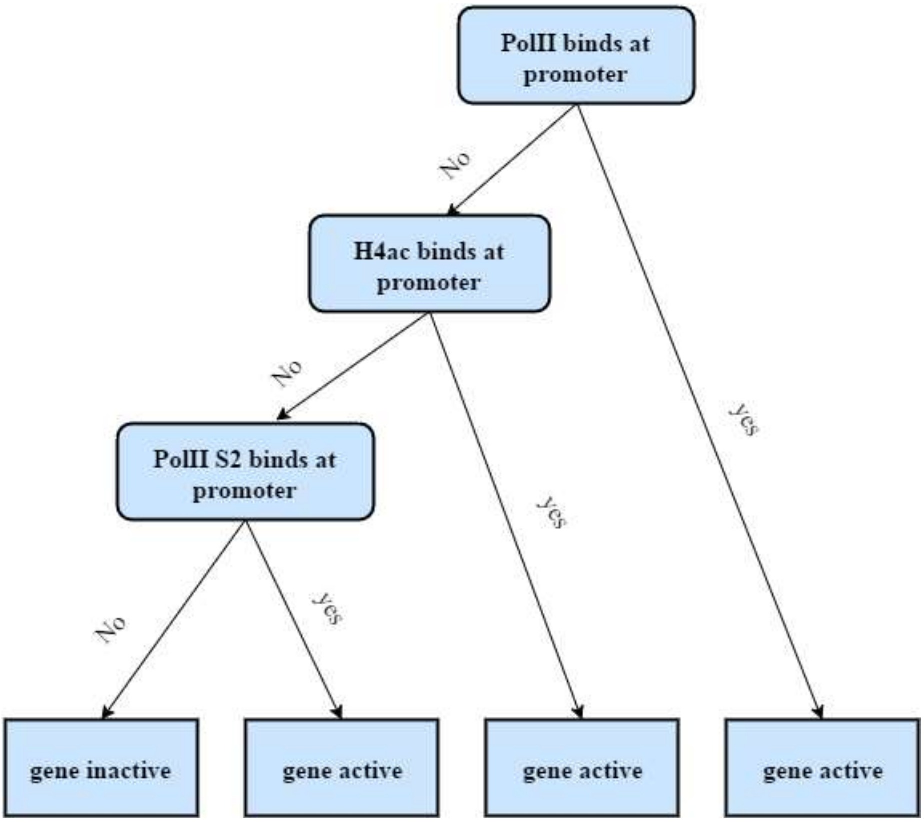


Figure 5.6: Resulted tree where leaf nodes represent class of the genes and the root node and internal nodes represent binding of protein at promoter region.

The variables that have been used to construct the tress are the bindings of RNA PolII, H4ac and RNA PolII S2 at promoter. The tree indicates that if RNA PolII binds at a gene promoter, the gene will be active, however if it does not but H4ac binds at promoter, that gene will be active, else if PolII S2 binds at the promoter the gene will be active. The gene is classified as inactive for other status of the protein. The accuracy for 10-fold cross validation is 83.94%.

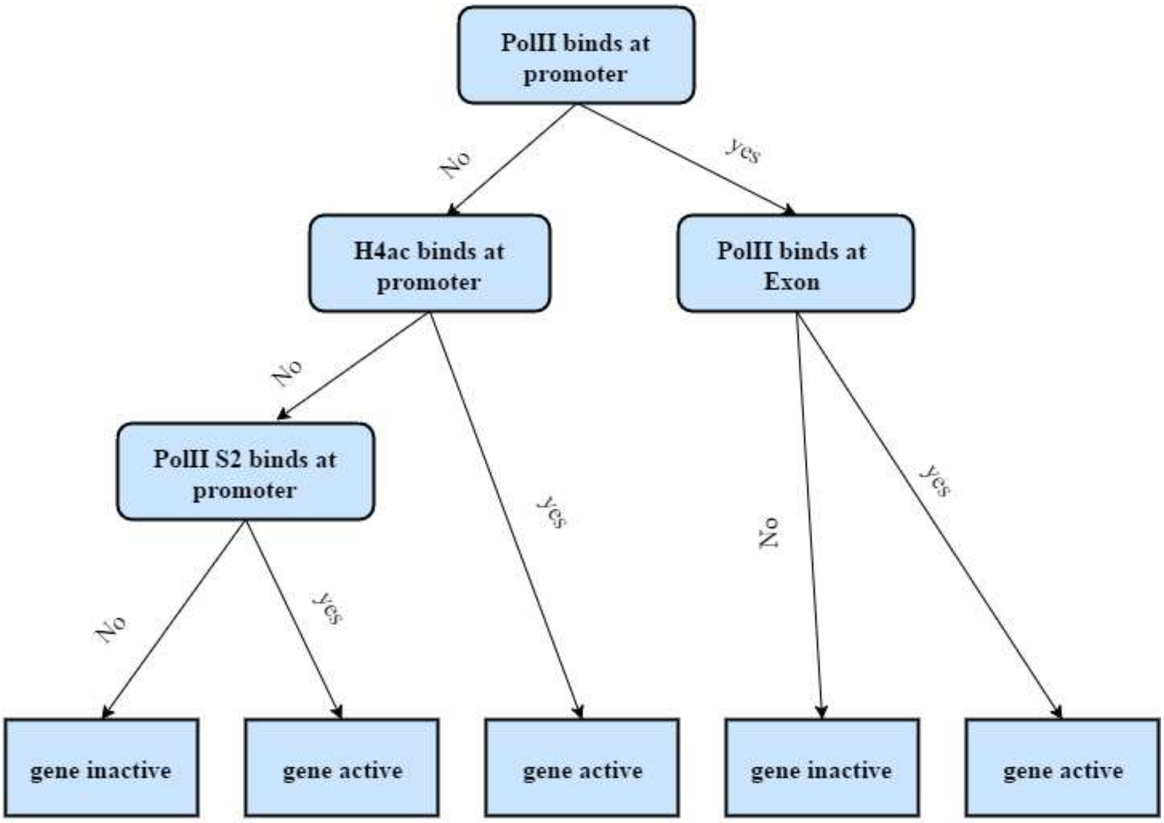


Figure 5.7: Resulted tree where leaf nodes represent class of the genes and the root nodes and internal nodes represent binding of protein at different gnomic region (promoter, exon etc).

Next, the profile of all the proteins bound at different genomic features, such as promoter, exon, intron, distal intergenic for all 1261 genes have been combined and the data has been modelled with the decision tree. In this case, the tree has been constructed with the features, the bindings of RNA PolII, H4ac, RNA PolII S2 at promoter regions and RNA PolII at exon. Here the tree concludes that when RNA PolII is not bound at promoter for a gene, the tree appears similar to that above Figure 5.6, but the right side of the tree suggests that RNA PolII bindings at promoter and exon would classify a gene as active. The tree is depicted in the Figure 5.7.

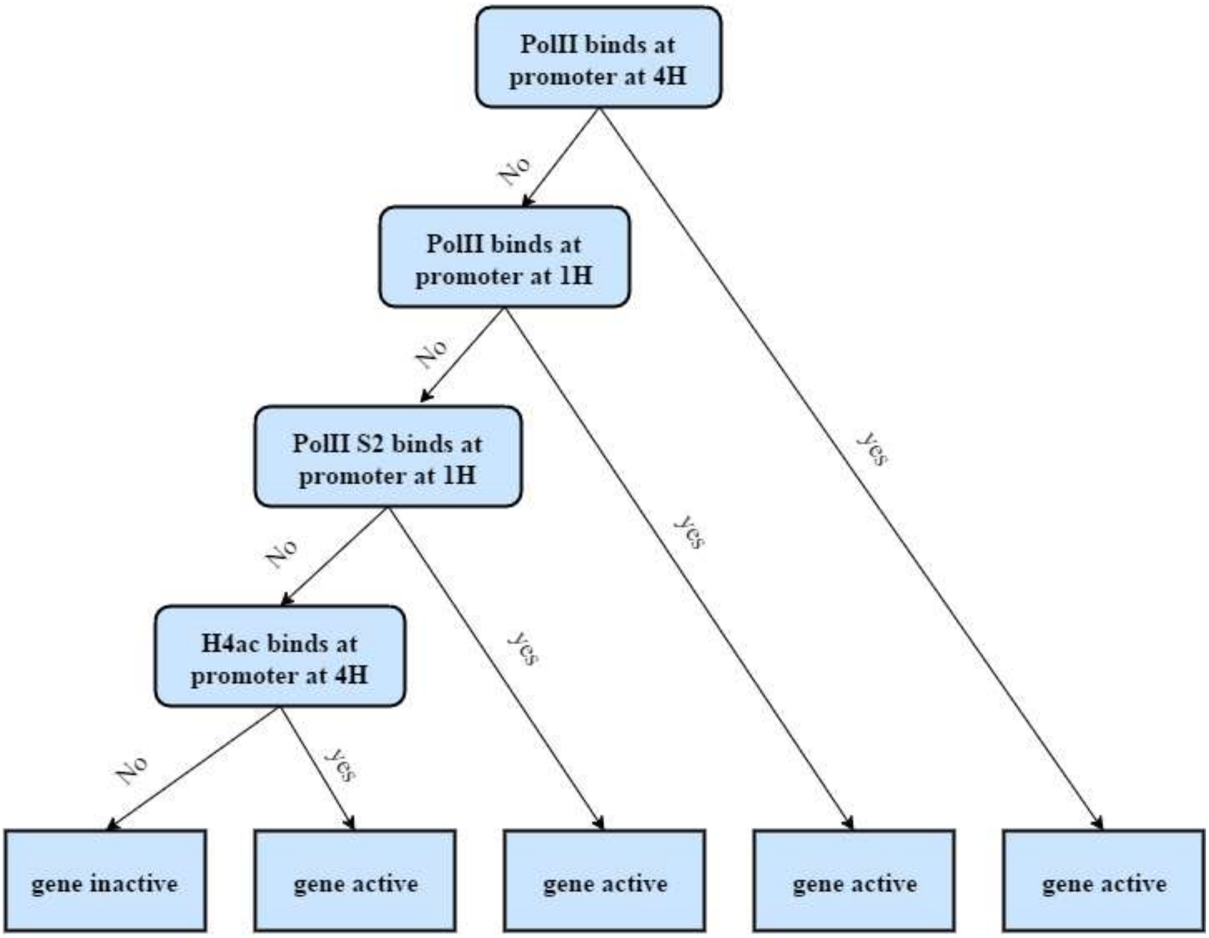

PolII binds at promoter at 4H
No
PolII binds at promoter at 1H
No
PolII S2 binds at promoter at 1H
No
H4ac binds at promoter at 4H
No
yes
yes
yes
yes
gene inactive
gene active
gene active
gene active
gene active

Figure 5.8: Resulted tree where leaf nodes represent the class of the genes and the root node and internal nodes represent binding of protein at promoter at different time points.

Next how protein binding across times points affects gene expression is investigated. Epigenetic mechanism and gene expression might not happen at the same time for the former to regulate the latter. For this experiment, the feature promoter has been selected and the profiles of all proteins bound at promoter at different time points have been combined, such as 0, 1 and 4 hour time points (0H, 1H and 4H respectively) for all 1261 chosen genes. The tree is depicted in Figure 5.8.

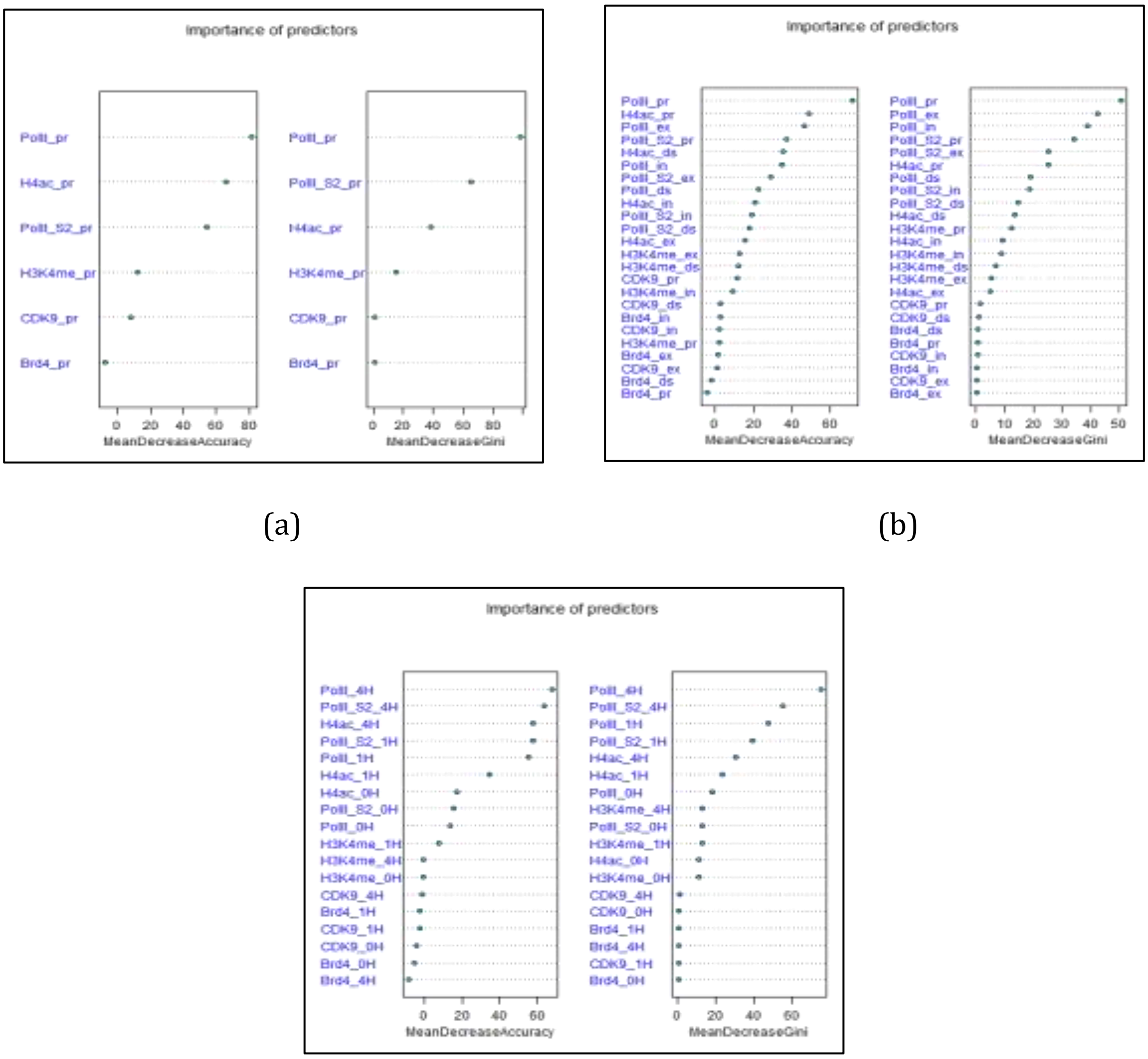
In this case, the tree has been constructed with the features/variables such as, the bindings of RNA PolII, H4ac, RNA PolII S2 at promoter at 1H and 4H time points. The tree suggests that if RNA Polll binds at promoter at 4H or 1H hour, or PolII S2 binds at promoter at 1H or H4ac binds at promoter at 4H time point the gene will be active, else the gene will be inactive at 4H time point. In the tree 0, 1 or 4 hour time-points are denoted as 0H, 1H and 4H respectively.

Result from Random Forests

To apply the Random Forests method for the classification of the data, R package `randomForest` [Liaw et al. 2002] has been used.

As input the same dataset mentioned above is used. Here too, the class of the genes is used as the response variable and the binding profile of the six proteins at different genomic locations as the predictors. In Figure 5.9, the importance of different variables obtained by the random forests method has been presented.

When only binding at promoter is considered, in terms of mean accuracy and mean Gini, RNA PolII, H4ac and RNA PolII S2 are at the top of the table with more than 40% importance. However, when the binding profile of all the variables are aggregated, RNA PolII, H4ac and RNA PolII S2 bindings at promoter and PolII binding at exon are selected as the most important features with more than 40% importance from the mean decrease accuracy results. This results match with the features that are selected by the decision tree in the previous section.



(a)

(b)

(c)

Figure 5.9: Importance of variables by random forests. (a) result from protein binding profile at promoter. (b) result from protein binding profile at different features, exon (ex), intron (in), promoter (pr), distal intergenic (ds). (c) result from protein binding profile at promoter at different time points (0H, 1H and 4H).

5.4.4 Comparative performance between three classifiers

After looking at the performance of individual classifier, comparative analysis among all the classifiers has been performed. Different combinations of variables have been selected for this experiment (e.g the variables that are used to draw the decision tree or reported as important by random forest etc). The combinations of variables used for this experiment are as follows:

1. RNA PolII, RNA PolII S2 and H4ac at promoter (pr)
2. RNA PolII, RNA PolII S2 and H4ac at promoter and RNA PolII at exon (ex)
3. RNA PolII and H4ac at promoter at 4 hour time point (4H) and RNA PolII and RNA PolII S2 at promoter at 1 hour time point (1H)

Predictors	**Neural Network**	**Decision Tree**	**Random Forest**
	Accuracy (%)		
PolII_pr + H4ac_pr + PolII_S2_pr	80.02	84.08	78.83
PolII_pr + PolII_ex H4ac_pr + PolII_S2_pr	82.93	84.75	78.43
PolII_1H + PolII_4H+ H4ac_4H + PolII_S2_1H	83.48	84.70	80.41

Table 5.5: The performances of different classifiers in terms of accuracy after the 10-fold cross validation.

From Table 5.5 it is observed that for all combinations of variable, decision tree performs the best among three classifiers and neural network is the second best in terms of accuracy. Decision trees have several advantages over neural networks and random forests. Firstly, decision trees are very interpretable as it provides visual representation of the data. For example, here from the tree itself it is very apparent which variables are responsible for gene activation. Neural networks can predict the response from the unknown data well, but it remains unknown which states of the predictors are responsible for each response class.

A decision tree based classification model also automatically selects features that are important for the prediction and discards those input features that are not useful to the prediction. However, a neural network based model does not report feature selection automatically; therefore it uses all the features that are provided. Unless a user manually implements feature selection as a pre-processing step, with large set of variables, neural networks often provide a poor prediction if there are features in the dataset that are not useful. Random forests method also provides the feature selection steps.

For this reason, when the binding profiles of all proteins at all the genomic features and the protein binding profiles at promoters at different time-points have been used for prediction, only decision tree and random forests technique have been used as these datasets have 26 variables, and neural network has been excluded from the following comparative study.

A comparative performance analysis between decision tree and random forests has been performed on the datasets that include:

1. The binding profile of all proteins at promoters
2. The protein binding profiles of all protein at all the genomic features
3. The binding profiles of all proteins at promoters at different time-points

The accuracy results of these two classifiers for the selected datasets are summarised in Table 5.6. Again decision tree performs better than random forests in each instance.

Predictors	**Decision Tree**	**Random Forest**
	Accuracy (%)	
Promoter only	83.94	79.38
All genomic features combined	83.80	79.70
Promoter at different time-points	85.01	80.73

Table 5.6: The performances of decision tree and random forests in terms of accuracy after the 10-fold cross validation.

5.5 Summary

In this study, the central question has been whether protein binding at different genomic features can be predictive for gene expression changes. The proposed methodology described in this chapter illustrates how different genomic locations, biological and experimental conditions can be incorporated in the predictive study to identify associations between protein bindings at different locations and gene regulation events. To demonstrate the model, the protein binding profiles at four genomics features have been generated for six proteins and these have been integrated with gene expression results. Different classification approaches have been used to understand how these profiles describe the underlying gene expression variations. Different time-points and biological conditions have also been included in the investigation.

The results show that the binding profiles of different proteins at different genomic features can describe variations in the underlying gene expression. Among the six markers/proteins, RNA PolII, RNA PolII S2, H3K4me3 and H4ac have been found to be the most predictive of the gene expression profile. Among the features, classification with neural network proposes that the promoter feature performs best as the binding location for proteins for predicting gene expression. However, for some combination of proteins other genomics features also provide acceptable prediction of 80% accuracy, which suggests protein binding profiles at genomic features correlate strongly with gene regulation. Protein binding profiles at all four genomic features have been combined and the data has been modelled using the decision tree and random forests methods to identify the most important features predictive of the response value. Both classifiers have identified the same set of variables as important for predictions as neural network, however as combined profiles of proteins at different genomic locations have been used here, both classifiers have identified RNA PolII binding at exon as an important feature to describe the underlying gene expression profile.

It is known that epigenetic events may not occur simultaneously with gene expression changes, so how proteins bindings at promoter at different time-points correlate with gene expression activity has also been investigated. The decision tree suggests that, the gene activity at 4 hour time-point is correlated with RNA PolII and H4ac binding at 4 hour time point and RNA PolII and RNA PolII S2 binding at 1 hour time point. The random forests method also identifies these variables as mostly correlated with the gene activity.

The comparative analyses of the performances of the classifiers have been performed on the variables that are selected by decision tree and random forests as important features. In each case, decision tree has classified the data with most accuracy, followed by neural network. The combined protein binding profiles at all genomic features and at three time points at promoters have been used to observe the performance of decision tree and random forests. Again, decision tree performs better than the random forests in terms of accuracy.

The findings confirm existing knowledge on how genes are regulated by different regulatory proteins binding at different features. The proposed approach has given new insights how other regulation factors can be integrated such as different genomic locations, time points and biological conditions to find out dynamic regulation of gene expression.

Next generation data is expensive. It is not always possible to generate epigenetic data in clinical trials where thousands of samples are involved; therefore it is not possible for the biologist to know the underlying epigenetic profile that is causing the gene expression changes. On the other hand, there are times when epigenetic data is available but the gene expression data is missing when it is difficult to know how the epigenetic changes between conditions are affecting gene expression. These machine learning based models could feel that gap, by finding the association between gene expression and regulatory mechanisms. Once the association between up and down regulation of genes with epigenetics changes is created, it will be possible to predict one from the other.

Chapter 6

Conclusion and Future Direction

Gene expression that produces the essential proteins that maintain and support normal cellular development and functioning of the eukaryotic cell is a tightly controlled biological process. Disruption to this process may present as a variety of disease conditions and therefore, investigating all the mechanisms such as epigenetics events and transcription factor regulation that control gene expression; have generated considerable interest among scientists.

Recently, several pivotal gene regulation studies have made significant progress and this field is currently expanding very rapidly. High-throughput technologies such as chromatin immunoprecipitation technique followed by the next generation sequencing (ChIP-seq) and microarray expression studies enable researchers to investigate the relationships between different epigenetic mechanisms and gene regulation on a genome-wide scale. Several attempts at integrative analyses have identified a number of direct relationships between the two processes; however, a comprehensive understanding of the regulatory events remains elusive.

It is anticipated that high-throughput sequencing technologies such as ChIP-seq will prove to be of immense value to biomedical research, though they are currently hampered by a scarcity of robust analytical methods. For example, current integration methodologies have implemented basic analysis steps of ChIP-Seq data which do not fully leverage all the information this sequencing technology can offer. There are also gaps in the literature regarding how differential expression and differential binding change together between biological or experimental conditions and how time factors and distant genomic locations coalesce to direct gene regulation events.

The primary objective of this thesis has been to acquire complementary ChIP-Seq and microarray datasets and provide robust and reliable methodologies that will contribute to the investigation of the relationship between epigenetic mechanisms and gene regulation. It has also been the goal of this thesis to explore whether protein binding profile across genome can be predictive of gene expression changes thus finding associations between different epigenetic events and gene regulation. However, before finding the association another important goal has been to study the complex characteristics of ChIP-Seq data to find the most appropriate means for data pre-processing and effective modelling.

In this thesis these goals have been achieved by incorporating the advanced and improved results of ChIP-seq data, the protein binding profiles at different genomic features, different biological conditions and time-factors relevant to underlying epigenetics effectively within integrative analyses to detail putative causal relationships. Thus, the Markov Random Field model has been adapted to analyse the binding regions of ChIP-Seq data where the complex characteristics of the data are taken into consideration while modelling the data. Two methodologies have been proposed where the advanced analysis methods can be used in the integrative analyses. Various classification methods are also included in the model to determine the relationship between different epigenetic markers, proteins, genomic features and gene expression profile.

Often in biological study either the gene expression or the protein binding data is unavailable. I believe that the studying the relationship between regulatory factors and gene expression with these models will help the biologists estimate gene expression from the available epigenetics data or assume the underlying epigenetics from the available gene expression data.

The models have been applied to the time-series ChIP-seq datasets for 6 proteins/markers and the complementary microarray datasets. The ChIP-Seq datasets are for Bromodomain-containing protein 4 (Brd4), Acetylated Histone H4 (H4ac), Histone H3 lysine 4 tri-methylation (H3K4me3), RNA Polymerase II (RNA PolII), a subunit of RNA polymerase II (RNA PolII S2) and Cyclin-dependent kinase 9 (CDK9).

Both ChIP-Seq and microarray data have been obtained at various biological and experimental conditions, for example, the data have been collected at consecutive time points after stimulating the biological samples with different compounds.

6.1 ChIP-Seq data analysis

In this thesis, the Markov Random Field (MRF) model has been adapted for the analysis of ChIP-Seq data where complex characteristics of the data have been considered while modelling the data to accurately locate binding loci of a protein (Chapter 3). Comparative performance analysis has been carried out between the MRF and other existing methods and it has been shown that incorporating the characteristics such as spatial dependency, IP efficiency and excess zeroes in the model can improve the final list of bound regions. It has also been demonstrated how steps taken in the pre-processing of the ChIP-Seq data such as count correction before running the statistical analysis to find the protein binding locations can affect the performance of the model.

The MRF model has been compared with the mixture model using the negative binomial method on six ChIP-Seq datasets for histone protein, H3. The performances of the methods have been compared in terms of the number of peaks generated and also the quality of the peaks. In all six experiments, it has been found that the MRF model has produced on average 34% more regions than the negative binomial method. Considering spatial dependency provides the MRF model the strength to identify regions with low tag counts that have neighbouring regions with significant numbers of counts. It has been demonstrated that a model that does not incorporate spatial dependency will overlook those counts as insignificant, and focus only on the overall distribution. RNA Polymerase II is known to bind at promoter regions around the transcription start sites (TSS) to control gene expression. It has been demonstrated that the enrichment probabilities produced by the MRF model show a smooth profile of bindings around TSSs, whereas with the negative binomial method, the binding profiles in proximity to the TSSs are not so apparent.

The count correction step taken while pre-processing the data can also significantly improve the results. If unusually large counts are found in a concentrated region, it is assumed to have come from an undisclosed bias. If these counts are removed the results improve significantly. Three chromosomes (2, 9 and 12) have been found to have those high counts of sequences in a very concentrated region. With those counts removed, the model identifies 5284, 1151 and 2682 more bound regions respectively than if the counts were retained.

6.2 Integrative analysis between ChIP-Seq and microarray

In this thesis, two models (Chapters 4 and 5) have been proposed for integrative analysis of the protein binding and the gene expression data.

Most of the integrative analyses to date have used very basic analysis steps to obtain information from ChIP-Seq data while ignoring many important characteristics of the data. A Method has been proposed in this thesis to find the correlation between the differential binding probabilities for different proteins around transcription start sites and differential gene expression values associated with those TSSs. The binding probabilities are estimated by the MRF model and thus capture all the complex characteristics of the data. The probabilities are then generated around TSS to integrate the data with the gene expression results from the microarray data. This model also incorporates different biological conditions such as time factors, treatment etc., therefore, it can be applied to rich datasets that include such variables. The results have been validated on the proteins investigated using known biological characteristics.

The methodology exemplifies how extended information about the ChIP-Seq data such as spatial dependency, overall distribution can be incorporated in terms of enrichment probability in the integrative analysis of protein binding and gene expression data and illustrates how such analysis can investigate the correlation between differential expression and differential bindings of different proteins.

The method has been applied to the datasets described above that include ChIP-Seq data of six markers/proteins and microarray data obtained at different biological and experimental conditions. The result concludes that the bindings of RNA PolII and RNA PolII S2 around TSS are significantly correlated with both upregulation and down regulation of gene expression. Histone methylation and histone acetylation, represented by bindings of H3K4me3 and H4ac also show a positive correlation. However, bindings of Brd4 and CDK9 around TSS are weakly correlated with either up or down regulation of gene expression, while Brd4 is negatively correlated. From the experimental results it has also been demonstrated how binding events upstream or downstream of TSS also correlate with gene expression changes.

In this study the enrichment probability is generated around TSS. However, with the proposed model, any number of genomic features could be investigated their location. The applied model also allows us to incorporate any number of biological or experimental conditions.

A novel approach has also been proposed where the integrative analysis of the protein binding and the gene expression data includes binding locations of proteins and epigenetic markers at different genomic features such as exon, intron, promoter, distal intergenic region etc. It has been shown how the dynamic interactions between the regulatory proteins and gene expression can be investigated by integrating sets of genes regulated at successive time-points and different biological or experimental conditions and protein binding profile across the genome. This method also makes it possible to not only identify those proteins or markers that might be regulating genes but also where and when they bind to do so. Several classification techniques have been used to define the association between the protein binding profiles across the genome and the underlying gene expression variations. To demonstrate the utility of the model, the protein binding profiles at four genomics features have been generated for 6 proteins and this data has been integrated with gene expression results. Different time-points and biological conditions are also included in the investigation. A comparative performance analysis between the classification techniques has also been performed to determine which classification technique better predicts gene expression status from the protein binding profile.

Our results show, that among the proteins investigated, RNA PolII, RNA PolII S2, H3K4me3 and H4ac correlate with gene expression. Results from the neural network modelling indicate that the promoter performs best as the binding location for proteins to predict gene expression, but the bindings at other features can also predict the gene expression status with 70-75% accuracy.

The combined protein binding profile at all four genomic features are modelled using both decision tree and random forests to identify the most important features for predicting the status of the gene. The results confirm that the same variables that perform well with the neural network are again identified as important for predictions by these two methods. From the combined profile, RNA PolII binding at exon is identified as an important feature along with bindings of RNA PolII, RNA PolII S2 and H4ac at promoter to describe the underlying gene expression profile.

Of course epigenetic events may not occur at the same time as the gene expression changes, so how protein bindings at promoters at the different time-points correlate with gene expression activities has also been investigated. The decision tree reports that gene activity at the 4 hour time point is correlated with RNA PolII and H4ac binding at the 4 hour time point, and that RNA PolII and RNA PolII S2 binding at the 1 hour time point. The random forests method also identifies these variables as mostly correlated with the gene activity. The comparative analyses on the performances of all three classifiers have shown that the decision tree has classified the data with most accuracy, followed by the neural network model.

Our findings confirm existing knowledge on how genes are regulated by different regulatory proteins binding at different features. Furthermore, the proposed approach has given us new insights on how other regulation factors can be integrated such as different genomic locations, time points and biological conditions to better describe the dynamic regulation of gene expression. These models will help the scientist explore deeper into epigenetic regulations. It will also help discover new relationships between proteins, gene regulation and different genomic features.

6.3 Future work

In this thesis, effective ways of modelling ChIP-Seq data to improve results for the integrative analysis of protein binding and gene expression data has been investigated. Although considering all the complex characteristics gives us the opportunity to evaluate most of the available information that this type of next generation sequencing technology can offer, it remains challenging to assess its full potential with one method. Even after 10 years of its adaptation, artefacts are still discovered today and also as sequencing technologies are making advancement continuously, it is difficult to standardise the analysis techniques of ChIP-seq profile [Park et al. 2013; Teytelman et al. 2013; Cusco et al. 2016]. The shapes of the ChIP-seq peaks vary protein to protein. The peaks can be either sharp or broad and some peaks are in mixed mode. Most of the peak callers have been designed to deal with just one specific kind of peak. If an experiment involved ChIP-seq data of different proteins and markers, different peak-calling strategies are required for different shapes. Prior knowledge about the shape of the proteins or markers will help chose the peak callers. However, one can also adapt the multi-tool approach using several peak callers to generate consensus peak lists. This way it will be possible to use the strengths of different peak callers together and thus, the result of the peaks would be more robust.

Integrative analyses between ChIP-seq data and other types of genomic assays, such as, gene expression require both types of data to be obtained from the same samples in the same biological and experimental conditions. ChIP-seq and other next generation sequencing data are very expensive to generate and that is why it is not always possible to find a single experiment that generates ChIP-seq data along with complementary gene expression data. In this thesis the data provided by Nicodeme et al. [2010] which includes ChIP-seq data for seven markers obtained at different conditions and also microarray data obtained at the same conditions has been used. The models proposed in this thesis has applied to analyse these datasets; however it will be interesting to apply the learned classifiers to similar kind of datasets to check the robustness of the model that includes ChIP-seq data for different combinations proteins, epigenetic markers and transcription factors. Projects like ENCODE and other genome projects are producing rich datasets nowadays which will open up the opportunity to do so in the future. The ChIP-seq data obtained for the same proteins or markers used in this thesis but in different conditions can also be investigated along with complementary gene expression data to verify our conclusion. It will also be very interesting to compare our models with existing methods to investigate if the same biological findings can be achieved. It will not only validate our models but also will provide confidence for the biologists about the result found by the model.

Four genomic features, namely exon, intron, promoter and distal intergenic have been explored in our model, but there are several other important features that have close association with the gene regulation. For example, databases now also have annotation information on the positions of the exon and intron, such as first exon, first intron etc. It has been mentioned in the literature, first exon or first introns have a close relationship with gene expression. Therefore, the relationship of those features with gene regulation also needs to be investigated. Furthermore, there are other classification techniques that can be investigated to check whether the performances of those classifiers are better than the classifiers investigated in this thesis.

References

Abdolmaleky H.M., Shafa R., Ming T. et al. (2013) Psychiatric Epigenetics: A Key to the Molecular Basis of and Therapy for Psychiatric Disorders. http://www.neurologytimes.com/special-reports/psychiatric-epigenetics-key-molecular-basis-and-therapy-psychiatric-disorders

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C. et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503-511.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. (URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc)

Archer T.K. Chromatin, Epigenetics & Transcription (URL: http://www.niehs.nih.gov/research/atniehs/labs/escbl/pi/cge/index.cfm). Last reviewed 2016.

Bailey T.L., Boden M., Buske F.A., Frith M. et al. (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, 37:W202-W208.

Bao Y., Vinciotti V. et al (2015) Joint modeling of ChIP-seq data via a Markov random field model. Biostat (15 (2): 296-310.

Bao Y., Vinciotti V. et al. (2013) Accounting for immunoprecipitation efficiencies in the statistical analysis of ChIP-seq data. BMC Bioinformatics 2013, 14:169

Barski A., Cuddapah S., Cui K., Roh T.Y. et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129, 823–837.

Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P. et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Berger S.L., Kouzarides T., Shiekhattar R., Shilatifard A. (2009) An operational definition of epigenetics. Genes Dev 23: 781–783.

Bernstein B.E., Kamal M., Lindblad-Toh K., Bekiranov S. et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell. 120(2):169-81.

Bock C., Lengauer T. (2008) Computational epigenetics. Bioinformatics 24(1):1-10.

Bosch A, Zisserman A, Muoz X (2007). "Image Classification Using Random Forests and Ferns." In ICCV 2007. IEEE 11th International Conference on Computer Vision, pp. 1–8. IEEE Computer Society, Washington, DC, USA.

Breiman L. (2001) Random forests. Machine Learning, 45(1): 5–32.

Brown I, Heys S.D., Schofield A.C. (2003) From peas to "chips" – the new millennium of molecular biology: a primer for the surgeon. World Journal of Surgical Oncology. < http://wjso.biomedcentral.com/articles/10.1186/1477-7819-1-21>

Campos E.I., Reinberg D. (2009) Histones: annotating chromatin. Annu Rev Genet. 43:559–599

Carlson M. (Maintainer BP.) TxDb.Mmusculus.UCSC.mm9.knownGene: Annotation package for TxDb object(s). R package version 3.2.2.

Carmona F.J. (2015) Dog's DNA methylome uncovers hints on human cancer metastasis. [Image] Available at: http://mappingignorance.org/2015/01/16/dogs-dna-methylome-uncovers-hints-human-cancer-metastasis/. [Accessed at 12 August, 2014]

Cheng J. (2006) Introduction of Bioinformatics. [Image] Available at: http://calla.rnet.missouri.edu/cheng_courses/CAP5937.htm. [Accessed at: 2 August 2014]

Chung D., Kuan P.F., Pan G., Thomson J.A. et al. (2011) A Statistical Framework for the Analysis of ChIP-seq data. Journal of American Statistical Association 106(459), 891–903

Chung D., Zhang Q., Keles S. (2014) MOSAiCS-HMM: A Model-Based Approach for Detecting Regions of Histone Modifications from ChIP-Seq Data. Statistical Analysis of Next Generation Sequencing Data Part of the series Frontiers in Probability and the Statistical Sciences pp 277-295

Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines, Cambridge University Press.

Cusco P., Filion G. (2016) Zerone: a ChIP-seq discretizer for multiple replicates with built-in quality control

Dohm J., Lottaz C., Borodina T., Himmelbauer H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36(16), e105

Douglas JE (2009) Cis-regulatory mutations in human disease. Brief Funct Genomic Proteomic. 8(4): 310–316.

Down T.A., Rakyan V.K., Turner D.J., Flicek P., et al. (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26: 779–785. doi: 10.1038/nbt1414

Dunning M.J., Smith M.L., Ritchie M.E., Tavaré S. (2007) beadarray: R classes and methods for Illumina bead-based data. Bioinformatics. 2007 Aug 15;23(16):2183-4.

ENCODE Project Consortium, Bernstein B.E., Birney E. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414), 57–74

Eukaryotic gene structure. [Image] Available at: http://nptel.ac.in/courses/102103013/module1/lec1/8.html. [Accessed at: 1 of August, 2014]

Fedorova E., Zink D. (2008) Nuclear architecture and gene regulation. Biochim Biophys Acta. 1783:2174–2184.

Feinberg A.P., Vogelstein B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 301, 89-92.

Fejes A.P., Robertson G., Bilenky M., Varhol R. et al. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics; 24:1729-1730.

Ferdous M.M., Vinciotti V., Liu X., Wilson P. (2015) Exploring the link between gene expression and protein binding by integrating mRNA microarray and ChIP-Seq data. Statistical Learning and Data Sciences, Lecture Notes in Computer Science Volume 9047, pp 214-222

Friedman J.H. (1991) Multivariate Adaptive Regression Splines. Ann. Statist. Volume 19, Number 1, 1-67.

Frommer M., McDonald L.E., Millar D.S., Collis C.M. et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89: 1827–1831.

Furey T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet., 13, 840–852.

Garrett-Mayerand E, Parmigiani G (2004) "Clustering and classification methods for gene expression data analysis." Johns Hopkins Univ., Dept. of Biostatist. Working Papers. Working Paper 70. [Online]. Available: http://biostats.bepress. com/jhubiostat/paper70

Ge H., Liu Z., Church G.M., Vidal M. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. Nat. Genet. 29, 482-486.

Geeven G., van Kesteren R.E., Smit A.B., de Gunst M.C. (2012) Identification of context-specific gene regulatory networks with GEMULA--gene expression modeling using LAsso. Bioinformatics. 15;28(2):214-21.

Gilmour D.S., Lis J.T. (1984) Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. Proc Natl Acad Sci; 81:4275-9.

Gilmour D.S., Lis J.T. (1985) In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. Mol Cell Biol; 5:2009-18; PMID:3018544

Gilmour D.S., Lis J.T. (1986) RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol Cell Biol; 6:3984-9;

Good B, Peay J, et al. (2001) Class prediction based on gene expression: Applying neural networks via a genetic algorithm wrapper. Genetic and Evolutionary Computation Conference Late Breaking Papers, pages 122–130, July 2001.

Gordon D.B., Nekludova L., McCallum S., Fraenkel E. (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics. 21:3164–3165.

Grimaud C., Bantignies F., Pal-Bhadra M., Ghana P. et al. (2006) RNAi components are required for nuclear clustering of Polycomb group response elements. Cell124: 957–971.

Guan D., Shao J., Zhao Z., Wang P. et al. (2014) PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data. Nucl. Acids Res.

Hagood J. (2014) Beyond the Genome: Epigenetic Mechanisms in Lung Remodeling. Physiology May 2014, 29 (3) 177-185

Hashimoto T., Edwards M.D. (2014) Universal Count Correction for High-Throughput Sequencing. PLoS Comput Biol 10(3): e1003494.

Heintzman N.D., Stuart R.K., Hon G., Fu Y., Ching C.W. et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genet.; 39:311–18.

Herman J.G., Graff J.R., Myöhänen S., Nelkin B.D. et al. (1996) Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proceedings of the National Academy of Sciences of the United States of America. 93(18):9821–9826.

Heyn P., Kalinka A.T., Tomancak P., Neugebauer K. M. (2014): Introns and gene expression: Cellular constraints, transcriptional regulation, and evolutionary consequences. BioEssays, Volume 37, Issue 2, pages 148–154.

Hoang S.A., Xu X., Bekiranov S. (2011) Quantification of histone modification ChIP-seq enrichment for data mining and machine learning applications. BMC Res. Notes 4, 288.

Hoen P.A., Ariyurek Y., Thygesen H.H., Vreugdenhil E. et al. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucl. Acids Res. (2008) 36 (21): e141.

Holiday R. (2006) Epigenetics – A Historical overview. Epigenetics. 1(2):76-80

Holliday R. et al. (1975) DNA modification mechanisms and gene activity during development. Science. 1975; 187:226-32.

Hoopes L. (2008) Introduction to the gene expression and regulation topic room. Nature Education 1(1):160

Huber W., Sultmann H. et al. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 18 Suppl 1:S96–104.

Humburg P., Helliwell C.A., Bulger D., Stone G. (2011) ChIPseqR: analysis of ChIP-seq experiments. BMC Bioinformatics, 12:39.

Hurd P.J., Nelson C.J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomics Proteomics 8: 174-183

Initiation of the transcription. [Image] Available at: https://cellularphysiology.wikispaces.com/Genetic+Control+via+Activators+and+Enhancer+Elements. [Accessed at 2 February 2015]

Iyer V.R., Horak C.E., Scafe C.S., Botstein D., et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409: 533–538

Jacinto F.V., Ballestar E., Esteller M. (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. Biotechniques 44: 35, 37, 39 passim.

Jaenisch R., Young R. (2008) Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell; 132: 567–82.

Jain A.K., Dubes D.C. (1999) Data clustering, ACM Computing Survey, 31:264-323.

Jansen R., Grestein M. (2000). Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. Nucl. Acids Res. 28, 1481-1488.

Ji H., Jiang H., Ma W., Johnson D.S. et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-Seq data. Nature Biotechnology, 26: 1293-1300.

Johnson D.S., Mortazavi A., Myers R.M., Wold B. (2007) Genome-wide mapping of in vivo protein–DNA interactions. Science 316: 1497–1502

Jones P.A., Baylin S.B. (2007) The epigenomics of cancer. Cell.;128:683–92.

Jothi R., Cuddapah S., Barski A., Cui K. et al. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. Nucleic Acids Res. 36, 5221–5231

Karlic R., Chung H.R., Lasserre J., Vlahovicek K. et al. (2010) Histone modification levels are predictive for gene expression. Proc Natl Acad Sci USA 107: 2926–2931.

Kent W.J., Sugnet C.W. Furey T.S., Roskin K.M. et al. (2002) The human genome browser at UCSC. Genome Res 2002; 12:996-1006

Koch C.M., Andrews R.M., Flicek P., Dillon S.C. et al. (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res. 17(6):691-707.

Kruppa J, Schwarz A, Arminger G, Ziegler A (2013). "Consumer Credit Risk: Individual Probability Estimates Using Machine Learning." Expert Systems with Applications, 40(13), 5125–5131.

Laird P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. Nature reviews Genetics. 11(3):191-203.

Langmead B., Trapnell C., Pop M., Salzberg S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Lewin et al. (2008) Introduction to Genetics. Molecular biology of the cell, 5th ed.

Li J., Ching T., Huang S., Garmire L.X. (2015) Using epigenomics data to predict gene expression in lung cancer. BMC Bioinformatics, 16(Suppl 5):S10

Li W, Han J, et al. (2001) Cmar: Accurate and efficient classification based on multiple class-association rules. Proc in IEEE Int. Conf. on Data Mining, pages 369–376, 2001.

Liaw A., Wiener M. (2002) Classification and Regression by randomForest. R News. 2(3), 18-22.

Lieb J.D., Liu X., Botstein D., Brown P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28: 327–334

Lihu A., Holban S. (2015) A review of ensemble methods for de novo motif discovery in ChIP-seq data. Brief Bioinform. 16(6):964-73.

Lim S.J., Tan T.W., Tong J.C. (2010) Computational Epigenetics: the new scientific paradigm. Bioinformation. 4(7):331-337.

Liu B., Hsu W., Ma Y. (1998) Integrating classification and association rule mining. In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining.

Liu E.T., Pott S. et al. (2010) Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biology 8:56Lockhart D.J, Winzeler E.A. (2000) Genomics, gene expression and DNA arrays. Nature 405, 827-836Lomvardas S., Barnea G., Pisapia D.J., Mendelsohn M. et al. (2006) Interchromosomal interactions and olfactory receptor choice. Cell126: 403–413.

Lou S., Lee H., Qin H., Li J. et al. (2014) Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. Genome Biology. 15:408

Lun A.T.L, Smyth G.K (2015) csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucl. Acids Res

Ma W., Wong WH. (2011) The analysis of ChIP-seq data. Methods Enzymol, 497:51-73.

Markowetz F., Mulder K.W., Airoldi E.M., Lemischka I.R. et al. (2010) Mapping Dynamic Histone Acetylation Patterns to Gene Expression in Nanog-Depleted Murine Embryonic Stem Cells. PLoS Comput Biol 6(12): e1001034

Maston G.A., Evans S.K. Green M.R. (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet. 7:29–59.

McCulloch W., Pitts W. (1943) A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics 5 (4): 115–133.

Michel G.F. (2010) The roles of environment, experience, and learning in behavioral development. In: Hood K, Halpern C, Greenberg G, Lerner R, editors. Handbook of Developmental Science, Behavior and Genetics. Wiley; Malden, MA: pp. 123–165.

Mo Q. (2012). A fully Bayesian hidden Ising model for ChIP-Seq data analysis. Biostatistics 13(1), 113-128

Nelson S. (2008) Comparative methylation hybridization. Nature Education 1(1):55

Nicodeme E., Jeffery K.L., Schaefer U., Beinke S. et al (2010) Suppression of inflammation by a synthetic histone mimic Nature. 23;468(7327):1119-23.

Nott A., Meislin S.H., Moore M.J. (2003) A quantitative analysis of intron effects on mammalian gene expression. RNA. 9:607–617.

Nowak DE, Tian B, et al. (2005) Two-step cross-linking method for identification of NF-kB gene network by chromatin immunoprecipitation. Biotechniques; 39:715-25; PMID:16315372

Pan F., Wang B. et al. (2004) Comprehensive vertical sample-based knn/lsvm classification for gene expression analysis. In J Biomed Inform.

Park, D., Lee, Y., Bhupindersingh, G., and Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. PLoS ONE, 8(12), e83506.

Pavesi G., Mereghetti P., Mauri G., Pesole G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 32, W199–W203.

Pearson Education, Inc. 2014 [Image] Available at: http://preuniversity.grkraj.org/html/10_MOLECULAR_BIOLOGY.htm [Accessed at: 3 March, 2015]

Pepke S., Wold B. Mortazavi A. (2009) Computation for ChIP-seq and RNA-seq studies. Nature Methods **6**, S22 - S32

Petronis A. (2010) Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature 465, 721-727.

Portela A. Esteller M. (2010) Epigenetic modifications and human disease. Nature biotechnology. 28(10):1057-1068.

Qi Y. (2012). “Random Forest for Bioinformatics.” In C Zhang, Y Ma (eds.), Ensemble Machine Learning, pp. 307–323. Springer-Verlag, New York. ISBN 978-1-4419-9325-0.

Qin J., Li M.J., Wang P., Zhang M.Q. et al. (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. Nucl. Acids Res

Quinlan J.R. (1986) Induction of decision trees. Machine Learning, l(1). 81-106

Quinlan J.R. (1993) C4.5: Programs for machine learning. In Morgan Kaufmann, San Mateo, CA.

Quinlan J.R. (1993) C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

Ren B., Robert F., Wyrick J.J., Aparicio O. et al. (2000) Genome-wide location and function of DNA binding proteins. Science. 22;290(5500):2306-9.

Riggs A.D., Martienssen RA., Russo V.E.A. (1996) Introduction. In Epigenetic mechanisms of gene regulation (ed. Russo VEA, et al.), pp. 1–4. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Riggs AD. (1975) X inactivation, differentiation and DNA methylation. Cytogenet.Cell Genet. 14:9-25

Robinson M.D., McCarthy D.J., Smyth G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26:139-140.

Robinson M.D., Smyth G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, 23(21):2881-2887.

Roh T.Y., Cuddapah S., Cui K., Zhao K. (2006) The genomic landscape of histone modifications in human T cells. Proc Natl Acad Sci U S A. 103(43):15782-7.

Rozowsky J., Euskirchen G., Auerbach R.K., Zhang Z.D. et al. （2009）PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. Nat Biotechnol 27:66-75.

Rozowsky J., Zhang, Z.D., Snyder M. Chang J. et al. (2008). Modeling ChIP sequencing in silico with applications. PLoS Computational Biology 4(8), e1000,158

Schwarz R., Joseph B. Gerlach G. et al. (2010) Evaluation of One- and Two-Color Gene Expression Arrays for Microbial Comparative Genome Hybridization Analyses in Routine Applications. J Clin Microbiol. 48(9): 3105–3110.

Sha, K. and Boyer, L. A. (2009), The chromatin signature of pluripotent cells StemBook, ed. The Stem Cell Research Community, StemBook, doi/10.3824/stembook.1.45.1.

Shah A. (2009) Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system. Nature Methods 6

Shanmuganathan R., Basheer N.B., Amirthalingam L., Muthukumar H. et al. (2013) Conventional and nanotechniques for DNA methylation profiling. Journal of Molecular Diagnostics. 15(1):17–26.

Shen L, Shao N., Liu X, Maze I. et al. (2013) diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. PLoS One. 2013; 8(6): e65598

Shin H., Liu T., Duan X., Zhang Y. et al. (2013) Computational methodology for ChIP-seq analysis. Quantitative Biology, 1(1): 54–70

Smyth GK (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology Volume 3, Issue 1, Article 3

Smyth GK (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Springer, New York, pages 397–420

Solomon M.J., Larsen P.L., Varshavsky A. (1988) Mapping protein-DNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. Cell 1988; 53:937-47.

Steinhauser S., Kurzawa N., Elis R., Hermann C. (2016) A comprehensive comparison of tools for differential ChIP-seq analysis. Brief Bioinform, 17 (3) Oxford University Press

Sun H., Wu J., Wickramasinghe P., Pal S. et al. (2011) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. Nucleic Acids Res. 39(1): 190–201.

Szalkowski, A.M., Schmid, C.D. (2011) Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. Brief Bioinform; 12(6):626-33

Technologies for studying epigenetics [Image] Available at: www.encodeproject.org [Accessed at: 4 January 2016]

Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. U.S.A., 110(46), 18602–7.

Trapnell C., Salzberg S.L. (2009) How to map billions of short reads onto genomes. Nat. Biotechnol. 27, 455–457.

Triantaphyllopoulos A.K., Ikonomopoulos L. and Bannister A.J (2016) Epigenetics and inheritance of phenotype variation in livestock. Epigenetics & Chromatin 9:31

UCSC genome browser. http://genome.ucsc.edu/

Wang Z., Zang C., Rosenfeld J.A., Schones D.E. et al., (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. Nature Genet. 40, 897–903

Wani K., Aldape K. (2016) PCR techniques in characterizing DNA methylation. Methods in Molecular Biology. 1392:177–186.

Watson J. D., Crick F. H. C. (1953) A structure for deoxyribose nucleic acid. Nature 171, 737–738

Weinmann A.S., Yan P.S., Oberley M.J., Huang T.H., et al. (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 16: 235–244

Werbos P.J. (1975). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.

Whitaker L. (1914) On the Poisson law of small numbers. Biometrika 1914, 10:36-71.

Wild L., Flanagan J.M. (2010) Genome-wide hypomethylation in cancer may be a passive consequence of transformation. Biochimica et biophysica acta. 1806(1):50-57.

Wright S. (1968) Evolution and the Genetics of Populations, Volume 1: Genetic and Biometric Foundations. University of Chicago Press.

Xiao H., Teng X. (2009) Perspectives of DNA microarray and next-generation DNA sequencing technologies. Science in China Series C: Life Sciences, Volume 52, Issue 1, pp 7-16

Yu G., Wang L.G., He Q.Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics, 31(14):2382-2383.

Yu H., Zhu S., Zhou B., Xue H. et al. (2008) Inferring causal relationships among different histone modifications and gene expression. Genome Res. 18(8):1314–1324.

Zang C., Schones D.E., Zeng C., Cui K. et al. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25 (15): 1952-1958).

Zhang Y., Liu T., Meyer C.A., Eeckhoute J. et al. (2008). Model Based analysis of ChIP-Seq (MACS). Genome Biol.; 9(9): R137

Ziegler A., K¨onig I.R. (2014). “Mining Data with Random Forests: Current Options for Realworld Applications.” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(1), 55–63.

Glossary

ANN	Artificial neural network
BED	Browser Extensible Data
BRD4	Bromodomain-containing protein 4
BWA	Burrows-Wheeler aligner
CDK9	Cyclin-dependent kinase 9
ChIP	Chromatin immunoprecipitation
ChIP-Seq	ChIP sequencing
DNA	Deoxyribonucleic acid
ENCODE	Encyclopaedia of DNA elements
FDR	False discovery rate
H3K4me3	Histone H3 lysine 4 tri-methylation
H4ac	Acetylated Histone H4
BET	Bromodomain and extra-Terminal motif
LPS	Lipopolysaccharide
MRF	Markov Random Field Model
mRNA	Messenger RNA
NN	Neural Network

RNA	Ribonucleic acid
RNA PolII	RNA Polymerase II
RNA PolII S2	subunit of RNA polymerase II
SAM	Sequence alignment map
TF	Transcription factor
TSS	Transcription start site
UCSC	University of California Santa Cruz

Appendix 1

This section provides the functions that implements Algorithm 4.3 and 4.4 mentioned in Chapter 4 in the method section.

Code:

****

Given a ChIP-seq count data of any chromosome for a protein, this function models the data with the MRF model using the R package enRich and returns the enrichment probability result per window estimated by the model

Input/s

1. The ChIP-seq count data of a protein

Output

1. The enrichment probability result per window for given chromosome

****

```
Find_ProbabilityResult_oneChr <- function (countData)

{

      library(enrich)

      countData <- read.table(countData, header = T)
```

"Preapare the count data"

```
      countDataList =list()

      countDataList$region = countData [,1:3]

      countDataList$count = countData [,4]
```

"Model the data with mrf function with 2000 iteration and 1000 burnin value"

```
result_mrf=        mrf(countDataList,        method="NB",
Niteration=2000, Nburnin=1000, cr=0.05)
```

“Generate enrichment probability result per window that are estimated by the model”

```
result_mrf_allPP   <-   data.frame(start   =   result_mrf
$data$region$Start,  end  =  result_mrf$data$region$Stop,
count = result_mrf$data$count, PP = result_mrf$PP)
```

****

Given the path to the directory where all ChIP-seq count data of all chromosomes for a protein are stored, this function models each count data with the MRF model and returns the enrichment probability per window for all chromosomes together.

Input/s

1. Path to the directory where all count data are saved.

Output

1. The enrichment probability result per window for all chromosomes together

****

```
Find_ProbabilityResult <- function(filePath)

{
```

"Set the path to the directory where count data are saved"

```
setwd(filePath)
```

"Read the name of the files in the folder and assign them to a variable"

```
files = (Sys.glob("*.txt"))
```

"Create an empty vector to hold the enriched regions"

```
vec = vector()
```

"For each file in the variable files do:"

```
      for (i in 1:length(files))

            {
```

"Call the function `Find_ProbabilityResult_oneChr` on each count data"

```
            enrichmentProb =
            Find_ProbabilityResult_oneChr(files[i])
```

"Save the enrichment probability result in the vector"

```
            vec = rbind(vec, enrich)

            }
```

"Return the enrichment probability result for all cromosomes"

```
vec

}
```

****

Given the path to the directory of where TSS information are saved, prefix of the file names (up to chromosome number) and the number of chromosomes, this function reads all the TSS information into the R workspace and assign them to appropriate variables.

Input/s

1. The path to the directory where TSS data are saved
2. The prefix of the TSS file names
3. The number of the chromosomes

Output

1. Reads the TSS information into the workspace

****

```
GetTSSinfo <- function (filePath, fileExt, chrN)

{

      for (j in 1:chrN)
```

```
          {

          tss_file   =   paste(filePath,   fileExt,   j,".txt",
     sep="")
```

“Read TSS file into the workspace”

```
          tss <- read.table(tss_file, header=T, sep="\t")
```

“Remove the TSSs that have same startsite”

```
          tss <- tss[!duplicated(tss$mm9.knownGene.txStart),]
```

“Create the name of the variable to hold the TSS information”

```
          tss_chr = paste("tss_chr",j, sep="")
```

“Assign the TSS to a variable created above”

```
          assign(tss_chr, tss)

          }

}
```

****

Given loci of TSSs of any chromosome and the enrichment probability result per fixed length window for that chromosome obtained by the MRF model, this function returns a matrix with enrichment probability profile around TSS

Input/s

1. TSS information of any chromosome
2. Enrichment probability results for that chromosome, obtained by the MRF model

Output

1. Enrichment probability profile around TSS

****

```
FindRegionAroundTSS <- function(TSS, enrichProb)

     {
```

“Number of TSS regions”

```
     len1 = dim(TSS)[1]
```

"Creating empty vectors to hold the probability results around start site"

```
     TSS_0 = numeric()

     plus1Kb = numeric()

     minus1Kb = numeric()

     plus2Kb = numeric()

     minus2Kb = numeric()

     plus3Kb = numeric()

     minus3Kb = numeric()

     plus4Kb = numeric()

     minus4Kb = numeric()

     plus5Kb = numeric()

     minus5Kb = numeric()

     vec = numeric()
```

"For each TSS do:"

```
for(i in 1:len1)

     {

     tss_start = TSS$mm9.knownGene.txStart[i]
```

"Calculate index/row number from proabability result that matches index of transcription start"

```
      j = (tss_start%/%200) + 1
```

"Assign the probability result for the regions 5KB up and downstream around TSS"

```
     TSS_0[i] = enrichProb$PP[j]

     plus1Kb[i] = mean(c(enrichProb$PP[j+1],
     enrichProb$PP[j+2],
     enrichProb$PP[j+3],enrichProb$PP[j+4],
     enrichProb$PP[j+5]))
```

```
minus1Kb[i] = mean(c(enrichProb$PP[j-1], enrichProb$PP[j-2], enrichProb$PP[j-3],enrichProb$PP[j-4], enrichProb$PP[j-5]))

plus2Kb[i] = mean(c(enrichProb$PP[j+6], enrichProb$PP[j+7], enrichProb$PP[j+8],enrichProb$PP[j+9], enrichProb$PP[j+10]))

minus2Kb[i] = mean(c(enrichProb$PP[j-6], enrichProb$PP[j-7], enrichProb$PP[j-8],enrichProb$PP[j-9], enrichProb$PP[j-10]))

plus3Kb[i] = mean(c(enrichProb$PP[j+11], enrichProb$PP[j+12], enrichProb$PP[j+13],enrichProb$PP[j+14], enrichProb$PP[j+15]))

minus3Kb[i] = mean(c(enrichProb$PP[j-11], enrichProb$PP[j-12], enrichProb$PP[j-13],enrichProb$PP[j-14], enrichProb$PP[j-15]))

plus4Kb[i] = mean(c(enrichProb$PP[j+16], enrichProb$PP[j+17], enrichProb$PP[j+18],enrichProb$PP[j+19], enrichProb$PP[j+20]))

minus4Kb[i] = mean(c(enrichProb$PP[j-16], enrichProb$PP[j-17], enrichProb$PP[j-18],enrichProb$PP[j-19], enrichProb$PP[j-20]))

plus5Kb[i] = mean(c(enrichProb$PP[j+21], enrichProb$PP[j+22], enrichProb$PP[j+23],enrichProb$PP[j+24], enrichProb$PP[j+25]))

minus5Kb[i] = mean(c(enrichProb$PP[j-21], enrichProb$PP[j-22], enrichProb$PP[j-23],enrichProb$PP[j-24], enrichProb$PP[j-25]))

}
```

“Put the enrichment profile together in a matrix”

```
vec = cbind(TSS$mm9.knownGene.txStart,minus5Kb, minus4Kb, minus3Kb, minus2Kb, minus1Kb, TSS_0, plus1Kb, plus2Kb, plus3Kb, plus4Kb, plus5Kb)
```

“Return matrix with enrichment probability profile around TSS”

```
     vec

}
```

****

Given the path to the directory where enrichment probabilities (obtained by modelling the ChIP-seq data with MRF model) of a number of chromosomes is stored, prefix of the name of the file (before the number of the chromosome and the number of chromosomes, this function returns a vector that holds enrichment probability profiles around TSS for all chromosomes of a protein.

Input/s

1. The path to the directory of enrichment probability results
2. The prefix of the names of files of probability results (before the number of the chromosome)
3. The number of chromosomes

Output

1. Enrichment probability profiles around TSS for all chromosomes of a protein

****

```
GetProbabilityProfile <- fucntion(filePath, fileExt, chrN)

     {
```

“Create an empty vector to hold the enrichment probability profile”

```
     TSSProfileofProtein = c()
```

“For each chromosome do:”

```
          for (j in 1:chrN)

          {
```

“Read the enrichment probability file for chromosome j into the workspace”

```
          PP_file = paste(filePath, fileExt, j, sep="")

          PP_chr =  read.table(PP_file, header = T)

          chr = paste("tss_chr", j, sep="")
```

“Call the function `FindRegionAroundTSS` on chromosome j”

```
        region_tss = findRegionAroundTSS(get(chr), PP_chr )
```

“Annotate the enrichment profile with gene symbol”

```
        region_tss_annotated <- data.frame(region_tss, Genes
        = get(chr)$mm9.kgXref.geneSymbol, chr =
        paste("chr",j,sep=""))

        TSSProfileofProtein = rbind(TSSProfileofProtein,
        region_tss_annotated)

        }
```

“Return the enrichment probability profile around TSS”

```
    TSSProfileofProtein

}
```

****

Given enrichment probability profile of all chromosomes of a ChIP-seq data and a list of genes, this function returns a matrix that holds enrichment profile around TSS associated with the genes provided.

Input/s

1. Enrichment profile of a protein around all TSSs
2. A list of genes

Output

1. Enrichment profile around TSS associated with the genes provided

****

```
Integrate_Marray_ChIP-seq <- function(TSSProfile, listOfgenes)
    {

    TSS_gene <- data.frame(Genes= TSSProfile$Genes,
    as.is=TRUE)
```

“Create an empty matrix to hold the result”

```
    TSSProfileSelected <-matrix(,nrow=0, ncol=1)

    for (j in 1:dim(TSS_gene)[1])

        {
          found <- match(TSS_gene[j,1], listOfgenes[,1])

          if(!is.na(found))

               {

               TSSProfileSelected <-rbind(TSSProfileSelected,
               TSS[j,])

               }

          }

TSSProfileSelected

}
```

****

If there is a list of genes and the enrichment probability profiles of a protein at two biological conditions around TSSs associated to those genes are needed to be investigated, following codes can be run to create enrichment probability profile around selected TSSs for both datasets. Once the profiles are created, it can be studied with the gene expression result of those genes.

****

“Read TSS file into the workspace”

```
GetTSSinfo(filePath1, fileExt1, chrN)
```

“Create enrichment probability profile for the protein in two conditions”

```
Protein_con1 <- getProbabilityProfile(filePath2, fileExt2,
chrN)

Protein_con2 <- getProbabilityProfile(filePath3, fileExt3,
chrN)
```

“Create enrichment probability profile for the protein in two conditions for selected genes”

```
Protein_con1_selected <- Integrate_Marray_ChIP-
seq(Protein_con1, listOfgenes)

Protein_con2_selected <- Integrate_Marray_ChIP-
seq(Protein_con2, listOfgenes)
```

Appendix 2

This section provides the functions that implements Algorithm 5.3 and 5.4 mentioned in Chapter 5 in the method section.

Code:

****

Given a ChIP-seq count data of any chromosome for a protein this function models the data with the MRF model using the R package enRich and returns the enriched regions found by the model.

Input/s

1. The ChIP-seq count data of one chromosome

Output

1. The list of the enriched regions

****

```
Find_enrichedRegions_oneChr <- function (countData)

{
      library(enrich)

      countData <- read.table(countData, header = T)
```

"Format the count data to provide it as input to the model"

```
      countDataList =list()

      countDataList$region = countData [,1:3]

      countDataList$count = countData [,4]
```

"Model the data with mrf function from enRich with 2000 iteration and 1000 burnin"

```
     result_mrf= mrf(countDataList, method="NB",
     Niteration=2000, Nburnin=1000, cr=0.05)
```

"Find the enriched regions at 5% FDR"

```
     enrich_regions = enrich.mrf(result_mrf,
     analysis="separate")
```

"Return the enriched regions found by the model"

```
     enrich_regions$enrich[[1]]

}
```

****

Given the path to the directory where ChIP-seq count data for a set of chromosomes of a protein are saved, this function models each count data with the MRF model and returns the enriched regions of all chromosomes together in a vector.

Input/s

1. Path to the directory where the count data of a number of chromosomes are saved.

Output

1. The list of the enriched regions found in the given chromosomes

****

```
Find_enrichedRegions <- function(filePath)

{
```

"Set the path to the directory where count data are saved"

```
     setwd(filePath)
```

"Read the name of the files in the folder and assign them to a variable"

```
     files = (Sys.glob("*.txt"))
```

"Create an empty vector to hold the enriched regions"

```
    vec = vector()
```

"For each file in files do:"

```
    for (i in 1:length(files))

        {
```

"Call the function `Find_boundRegions_oneChr` on each count data"

```
        enrich = Find_boundRegions_oneChr(files[i])
```

"Save the enriched regions in the vector"

```
        vec = rbind(vec, enrich)

        }
```

"Return the enriched regions found in all the chromosomes together"

```
vec

}
```

****

Given a list of regions enriched by a protein, a TxDb object where annotation information are stored and an annotation package for a particular organism, this function annotates the regions with gene symbols and genomic features using the R package ChIPseeker.

Input/s

1. A list of enriched Regions
2. TxDb object
3. An annotation package
4.

Output

1. Enriched regions annotated with gene symbols and genomic feature.

****

```
Annotate_enrichedRegions <- function(enrichedRegions, txDb,
annoDb)

{

     library(ChIPseeker)
```

"Annotate the enriched regions"

```
     peakAnno <- annotatePeak(enRichedRegions, TxDb=txdb,
     annoDb=annoDb)
```

"Convert the annotation result into appropriate format"

```
     peakAnno <- as.GRanges(peakAnno)

     peakAnno <- as.data.frame(peakAnno)
```

"Return annotated enriched regions"

```
     peakAnno

}
```

****

Given annotated enriched regions of a protein and a genomic feature, this function finds the bound regions annotated with gene symbols that fall in the given genomic feature and returns the list of those genes.

Input/s

1. Annotated bound regions of a protein from ChIP-seq data
2. A genomic feature

Output

1. List of regions enriched by a protein in the given genomic feature

****

```
Enrichment_near_feature <- function(Annotation_file, feature)
{
```

"Read the annotated enriched regions into the workspace"

```
     anno_file <- read.table(Annotation_file, header=T,
     sep="\t", fill=TRUE)
```

"Select the regions that are bound by the given feature"

```
anno_file <- anno_file[grep(feature, anno_file$annotation),]
```

"Return the name of the genes that are bound by the protein in the feature"

```
genes <- data.frame(SYMBOL = anno_file$SYMBOL)

genes

}
```

****

Given path to the directory that contains annotated enriched regions of number of proteins this function writes names of the genes bound by each protein in given genomic feature in separate files in a directory.

Input/s

1. Path to the directory where all the annotated files are

2. A genomic feature

Output

1. Create files containing the name of the genes that are bound by all proteins in the given feature

****

```
Enrichment_nearFeature_all <- function (filePath, feature) {
```

"Set the path to the given directory"

```
      setwd(filePath)
```

"Read the name of the files in the folder and assign them to a variable”

```
      files = (Sys.glob("*.txt"))
```

"Create a new directory to write the results"

```
      dir.create(feature)
```

"For each file in files do:"

```
      for (i in 1:length(files))

            {
```

"Call function `Enrichment_near_feature` on each file"

```
             enrich <- Enrichment_near_feature(files[i],
             feature)
```

"Create file name for the result to be written into. Write the gene names on the file and save it to the directory created for this feature"

```
            name <- strsplit(files[i], "[.]")

            fname <- name[[1]][1]

            fname = paste(feature, "\\", fname, "_", feature,
            ".txt", sep="")

            write.table(unique(enrich$SYMBOL), fname,
            col.names=TRUE, row.names=FALSE, quote=FALSE,
            sep="\t")

                 }

}
```

****

Given path to the directory that contains annotated enriched regions of number of proteins, a genomic feature and a list of genes, this function returns a biding profile of each protein near the feature for those genes

Input/s

1. Path to the directory where all the annotated files are
2. A genomic feature
3. A list of genes

Output

1. The binding profile of each protein near the feature for given genes

****

```
Feature_profile <- function (filePath, feature, listOfGenes)

{
```

"Call the function `Enrichment_nearFeature_all` to create separate files for each protein that will contain the name of the genes if the protein is bound to it in the given feature"

```
      Enrichment_nearFeature_all(filePath, feature)
```

"Set the directory to a folder created for feature"

```
      setwd(feature)
```

"Read the name of the files into a variable"

```
      files = (Sys.glob("*.txt"))
```

"Create an empty matrix to hold the binding profile"

```
      profile <- matrix(ncol = length(files), nrow =
      length(listOfGenes$Genes))
```

"Set the name of the genes as index for the matrix"

```
     row.names(profile)= listOfGenes$Genes
```

"Create an empty vector"

```
     vec = vector()
```

"Set the names of the columns of the matrix"

"For each file in the folder do:"

```
     for (i in 1:length(files))

          {

          name <- strsplit(files[i], "[_]")

          cname <- name[[1]][1]

          vec = rbind(vec, cname)

          }

     colnames(profile)= vec
```

"Create the binding profile for each protein saved in the given directory"

```
     for (i in 1:length(files))

          {
```

"Read the file into the workspace"

```
          genes <- read.table(files[i], header = T)
```

"For each gene in the listOfgenes do: if the gene is bound by the protein in the feature put 1 in the corresponding cell else put 0"

```
          for (j in 1:length(listOfGenes$Genes))

                    {

                    found <- match(listOfGenes$Genes[j],
                    genes[,1])

                    if (is.na(found))
```

```
                                        {
                                        profile[j, i] = 0
                                        }
                                  else profile[j, i] = 1
                                  }
                            }
```

"Return the binding profile"

```
      bindingProfile <- as.data.frame(profile, col.names=TRUE,
      row.names=FALSE, quote=FALSE)

      bindingProfile <- data.frame(Genes = listOfGenes$Genes,
      bindingProfile, status = listOfGenes$status)

      bindingProfile

}
```

****

Say there are annotated binding regions created for n proteins using the function `Find_enrichedRegions` and `Annotate_enrichedRegions`. Say, the annotation files are in a folder whose full path is assigned to a variable called, filePath. If binding profile needs to be created for these proteins for a feature, say "feature1" using a set of genes saved in a vector called listOfGenes. The following command can be run to create the binding profile.

****

```
profile_promoter <- Feature_profile(filePath, "feature1",
listofGenes)
```