Centre for Modeling and Simulation
Savitribai Phule Pune University

Master of Technology (M.Tech.)
Programme in Modeling and Simulation

Project Report

Classification of Promoters by Chromatin Structures

Sarvesh Nikumbh
CMS1005

Academic Year 2011-12

Centre for Modeling and Simulation
Savitribai Phule Pune University

Certificate

This is certify that this report, titled

Classification of Promoters by Chromatin Structures,

authored by

Sarvesh Nikumbh (CMS1005),

describes the project work carried out by the author under our supervision during the period from January 2012 to June 2012. This work represents the project component of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Center for Modeling and Simulation, Savitribai Phule Pune University.

Leelavati Narlikar, Scientist
Chemical Engineering Division
National Chemical Laboratory
Pune 411008 India

Sukratu Barve, Faculty
Centre for Modeling and Simulation
Savitribai Phule Pune University
Pune 411007 India

Anjali Kshirsagar, Director
Centre for Modeling and Simulation
Savitribai Phule Pune University
Pune 411007 India

Centre for Modeling and Simulation
Savitribai Phule Pune University

Author's Declaration

This document, titled

Classification of Promoters by Chromatin Structures,

authored by me, is an authentic report of the project work carried out by me as part of the Master of Technology (M.Tech.) Programme in Modeling and Simulation at the Center for Modeling and Simulation, Savitribai Phule Pune University. In writing this report, I have taken reasonable and adequate care to ensure that material borrowed from sources such as books, research papers, internet, etc., is acknowledged as per accepted academic norms and practices in this regard. I have read and understood the University's policy on plagiarism (`http://unipune.ac.in/administration_files/pdf/Plagiarism_Policy_University_14-5-12.pdf`).

Sarvesh Nikumbh
CMS1005

Abstract

The human genome with approximately 3×10^9 base-pairs stores a large amount of information crucial for the proper develeopment and functions of an organism. But there also exist mechanisms, external to the underlying DNA sequence of an organism that take part in regulating gene expressions in different cell types. These external mechanisms are typically called the epigenetic factors. Their examples include modifications in the chromatin structure through histone modifications, where histones are the chief protein components of the chromatin.

Study of epigenetic factors has gained interest and impetus in the last couple of years due to their roles in diseases. Their precise role in gene regulation is still poorly understood. Genome-wide maps of various histone modifications are now available across various cell-types. In this work we focus on high-resolution maps of 21 histone methylations from Barski et al. in CD4+ T cells.

Using support vector machines (SVMs), we show that histone modifications achieve near perfect accuracy in predicting the gene expression levels in CD4+ T cells. The results by SVM are in congruence with known facts in biology. Since, it is impractical to get such genome-wide maps of all tissues and with the remarkable characterizability of the histone modifications, we correlate the histone modifications in one cell type to the gene expression levels in other tissues. When applied specifically to Heart, the prediction accuracy has still been good.

On the other hand when attempted to use the relevant histone modifications data in identifying tissue specific signals in promoters across genome, their performance relatively decreased. Hence, for the second part, we propose a model that, along with the histone modifications, also includes an additional set of features deduced from performing a Markov analysis of the promoter sequences to help us identify tissue specific signals in promoters.

Acknowledgments

It gives me immense pleasure to have this opportunity to thank and acknowledge all those who made a difference in this endeavor. First and foremost, my sincere thanks to Leelavati Narlikar for being a wonderful guide. She made it possible for me to dive into computational biology, given that my last encounter with anything biology was way back in 2002. Leelavati has also been an excellent mentor. She was ever so patient and encouraging in all our discussions. It has been a wonderful experience throughout – thank you so much!

Prof. Mihir Arjunwadkar has been a tremendous source of inspiration and guidance all through the 2 years I have known him. It is just great to have someone like him to guide and mentor you. Every discussion with him, either academic or otherwise, has been a constant learning experience. My heartfelt gratitude to you.

I consider myself really fortunate to have interacted with Prof. Dilip Kanhere, the founding Director of our Centre. It was he who set the wheel rolling for me in the form of a summer internship at INRIA. All encounters with him have been extremely enriching. Thanks a ton, Sir!

My time at NCL for my masters project, will be one of the memorable ones, thanks to the Gadgil lab members, especially Avinash Ghanate, Sucheta Gokhale, Shraddha Puntambekar, Indhupriya, Priyanka Saxena, Dimple Nyaynit, Charudatta Navare, Mridula Prasad and Rossi D'Souza. Thanks for making this an enjoyable and productive experience! Also thanks to Dr. Tambe for allowing us to use his lab space when required.

Also, my sincere thanks to Prof. Anjali Kshirsagar, our Director at the Centre. She has been encouraging and very helpful in all my endeavors at the Centre. Prof. Sukratu Barve is another person at the Centre who encouraged us in higher mathematics with his wonderful teaching. I will always remember how mesmerized and engrossed it felt during his lectures and how we would even hate to take a break. Prof. Barve is also my internal guide; thanks! Dr. Abhijat Vichare also deserves a special mention. He has been an amazing personality whose lectures always left me thinking keenly on many essential aspects of any subject he taught. Discussions with him have always been interesting, demanding at most attention and dealing with very minute details. Thanks! Also, my special thanks to Dr. Jayaraman whom I have worked with apart from my course work and have also learnt a lot on many aspects of research from his insights and vast experience. His enthusiasm and appetite for research at this age always leaves me amazed.

Special thanks to Dinesh Mali, Suraj Meghwani and Rossi D'Souza for being such great friends. We have shared some of the best times of our lives together. Deepak Bankar, our system administrator at the Centre, has been another very nice friend helping us with various things at the Centre. Many thanks to him! I must also thank Mrunalini Dharmadhikari, our administrative associate at the Centre and Satish for being so helpful with respect to all my administrative tasks especially the conference travel grant procedures.

I would also like to thank the initial designers of this wonderful and unique M.Tech. course in modeling and simulation - Mihir Arjunwadkar, Abhay Parvate, Sukratu Barve, P. M. Gade and Dilip Kanhere. It is their vision and implementation which inspired me to take up this

program.

Most importantly, I thank my family for it is through their love, support and encouragement I stand where I am today!

Look deep into nature, and then you will understand everything better.

–Albert Einstein

Contents

Chapter 1

Introduction

"There is a paradox in the growth of scientific knowledge. As information accumulates in ever more intimidating quantities, disconnected facts and impenetrable mysteries give way to rational explanations, and simplicity emerges from chaos." This statement encompasses many of the challenges faced by sciences today – some of these challenges faced since even earlier than the last century. For a field like molecular biology, which has mainly advanced in the last 5-6 decades, especially with the discovery of the structure of the DNA in 1953, this is absolutely fitting. Technologies have evolved and advanced. Things which were unclear earlier are better understood now, but also making us realize the vastness we are dealing with. Overall, the underlying simplicity in nature has supposedly emerged albeit our comprehension still remains disconnected.

It is amazing how even after the discovery of the *cell* in 1665 by Hooke, it took about 174 years to understand and develop the cell theory in 1839. Every known living organism is composed of one or more of these cells – the basic structural and functional units. These cells amongst many other things store the necessary information required for regulating their individual functions, which is also passed on to the next generation of these cells. We call this information, the hereditary information, as – the DNA. It is a sequence of nucleotides composed of adenine (`A`), cytosine (`C`), guanine (`G`), and thymine (`T`). This complete DNA sits inside the nucleus of every single cell, divided into 23 pairs of linear molecules called chromosomes, 22 pairs of autosomes and a pair of sex chromosomes. These chromosomes are packaged by proteins into a structure called the chromatin which is a complex formed by the DNA and proteins. The DNA helps each cell determine its proper functioning. Every cell in an organism, though with the completely same DNA (called the *genome* of the organism) to work with, can understand its own distinct function. This genome is necessary to eventually form the proteins that make and operate an organism. And there are segments of this genome, called as 'genes', that correspond to a single protein. Every such cell, itself, 'knows' the function of each gene to produce the necessary gene products (the RNA and proteins) from them in *right quantities* and also knows '*when*' and '*where*' it is to be produced. In a nutshell this information flow happens as:

$$\text{DNA} \xrightarrow{transcription} \text{RNA} \xrightarrow{translation} \text{protein}$$

This production of other bio-molecules from the DNA is known as *gene expression*. No cell ever picks up information (forms products) from a gene that it doesn't require. So not all genes are always expressed and whenever they are, the expression level varies. This exercise of controlling the *gene expression* is called *gene regulation*. It can take place at any step in the above process from DNA to RNA to proteins and there are many factors effecting this regulation. The same factors can, at times, bring about regulation in multiple ways.

In this work we have focused on a certain kind regulatory factors, called *epigenetic factors*, that are responsible for chromatin structure remodeling, among other things, thus effecting changes and regulating gene expression levels. Further in this chapter, we will (a) delve little more into the cell and understand the packaging of the DNA in an eukaryotic cell; (b) get introduced to *histone* proteins that form the major portion of the chromatin and are the entities that undergo modifications, thus causing changes in the chromatin structure; and (c) talk about how the thesis has been laid out in the chapters ahead.

1.1 Cells and DNA

Cells are highly diverse. We have known several single and multi-cellular organisms. That there is variety in individual particulars and at the same time constancy in the fundamental mechanisms is astonishing. All cells on earth:

- store their hereditary information in the ***same*** linear chemical code (DNA)
- transcribe portions of hereditary information in the ***same*** intermediary form (RNA)
- translate RNA into protein the ***same*** way

These basic principles of biological information transfer are simple enough but we are very well aware that living cells are highly complex. Humans, the most complex of all known species, have about 210 different cell types that can be classified on the basis of the tissue of their origin. On the basis of how the cell, that holds the complete genome of the organism, structures itself, we classify living organisms into *prokaryotes* and *eukaryotes*. *Eukaryotes* are the ones that keep their DNA inside a nucleus, a membrane bounded intra-cellular compartment, while there is no distinct compartment to house the DNA in *prokaryotes*. A human body has enough DNA to span the complete solar system. Every single cell has the same DNA of approx. 2 metres in length when stretched out end-to-end (about 3.2 x 10^9 nucleotides long) and there are so many cells in the human body. This gives a compaction ratio of nearly 10,000-fold. This packaging of the DNA with such high compaction ratio is discussed ahead.

1.2 Packaging of DNA in Eukaryotic Cells

In *eukaryotes*, the DNA is stored inside the nucleus of a cell. This nucleus is nearly $6\mu m$ in diameter. The DNA is stored by dividing it into set of *chromosomes*. For instance, humans have their DNA distributed over 24 chromosomes, 22 pairs of autosomes and a pair of sex chromosomes. The Figure 1.1 below shows one such arbitrary chromosome. Each such chromosome consists of a long linear DNA molecule and the protein that binds and folds the DNA into a compact structure. This complex of DNA and proteins is called the *chromatin*. The protein that folds the DNA and causes the most basic level of compaction in the form of nucleosomes is a bead like structure called *histone* octamers. Exactly 146 base pairs of the DNA get wrapped 1.65 turns around a histone octamer. These histone octamer proteins have their Nitrogen (N)-terminals as tails hanging from them. The assembly of histone octamers in Figure 1.2 by two molecules each of histones H2A, H2B, H3 and H4 demonstrates these N-terminals. Thus, there are parts of the DNA that are inaccessible due to such compact winding and the genes in that portion of the DNA are inactive. While the part of the DNA that is accessible makes the corresponding gene(s) available for expression.

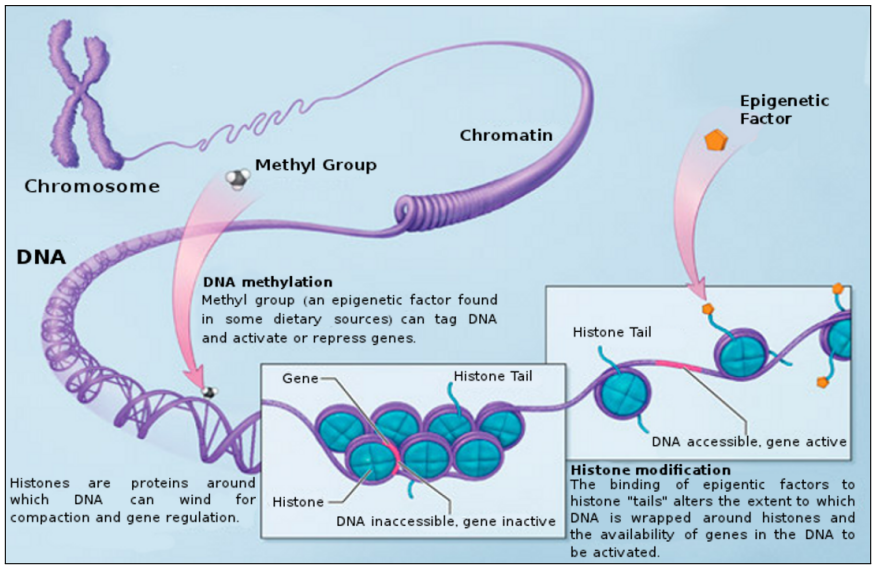

Figure 1.1: The chromatin structure. (Adapted from Epigenomics Scientific Background. ID: NBK45788, NCBI Bookshelf)

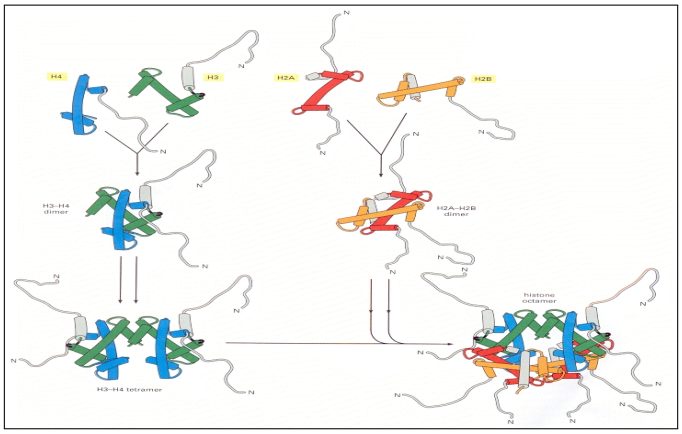

Figure 1.2: The assembly of a histone octamer. (Molecular Biology of the Cell. 4th edition. ID: NBK26887, NCBI Bookshelf)

1.3 Histones & their modifications

It is the tails of these histone cores that particularly undergo modifications [Kouzarides, 2007]. Since they are proteins, a particular amino acid of the 20 amino acids found in proteins undergo specific modifications like methylations, acetylations, phosphorylations etc. Depending upon the number of molecules of the modifiers taking part in modification it could be either a *mono-* , *di-* or *tri-* modification. These modifications can bring about changes like displacing the nucleosomes causing loosening/rearranging of the compactly bound structure that could make certain inaccessible regions accessible or vice-versa, in other words could cause remodeling of the chromatin structure and can regulate gene expression. These histone proteins are observed to be highly conserved across all *eukaryotes*. The Table 1.1 below enlists various types of identified modifications.

The nomenclature for these histone modifications is explained with the help of Figure 1.3.

On the basis of experiments performed and studies conducted, it is believed that these histone modifications can participate in gene regulation in 2 ways: directly affecting chromatin

Amino acid	Modification	Abbr.
(K) Lysine	mono-methylation	me1
	di-methylation	me2
	tri-methylation	me3
	acetylation	ac
	mono-ubiquitylation	ub
	poly-ubiquitylation	ubn
(R) Arginine	mono-methylation	me1
	di-methylation (symmetrical)	me2s
	di-methylation (asymmetrical)	me2a
(S) Serine	phosphorylation	ph
(T) Threonine	phosphorylation	ph
(E) Glutamate	ADP-ribosylation	ar

Table 1.1: Various types of identified modifications on some amino acids

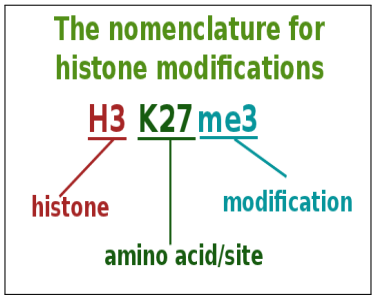

Figure 1.3: Nomen clature for histone modifications

structure by altering the charges of histone proteins and causing relaxation of chromatin structure; or indirectly by serving as recognition and binding sites for various classes of effector proteins.

These modifications are also termed as *epigenetic* factors because they are brought about by mechanisms external to the underlying DNA sequences. For instance the various methylations and acetylations can come from the food we eat. Also it is very well known that pregnant women are advised on more folic acid or folate as it aids rapid cell division and growth. These advices are made with the aim of epigenetically enhancing specific capabilities of an individual.

Some more information on the structure of the nucleosome core particle to know how the DNA is packaged can be looked at in Appendix A.

1.4 Outline

This chapter covered the necessary introduction to the domain – molecular biology. We will now move on and focus on our work in the following chapters. Chapter 2 describes the data – specifically, answers to the questions : 'how' and 'from where'; and the goals we set for ourselves in the 2 parts of this work. Chapter 3 discusses our classification model, methodology adopted and results for part I while chapter 4 looks into the corresponding model for part II. Concluding contributions of this thesis and proposed future directions that this work can take are presented in chapter 5.

Chapter 2

This Work

As mentioned earlier, we have attempted to focus on epigenetic factors influencing gene regulation. Thus, we set out to collect this histone modifications' data and the corresponding gene expression levels. In the sections that follow, we highlight the goals and discuss preprocessing of the collected data.

2.1 Goals

We decided to use the histone modifications at the promoter regions of genes in CD4+ T cells to help us characterize the average gene expression levels thereof. Though we started of with expression levels only in the T cells, while working along with the modifications, based on some initial results, we surmised that these modifications or specifically the methylations could also hold some characteristic information with respect to the tissue specificity of promoters.

Summarily, our goals are:

1. To predict expression level of a gene by histone modifications at its promoters.
2. To characterize tissue specificity of promoters using these histone modifications.

Quite evidently, the two parts of work we discussed at the end of the last chapter pertain to these two goals.

2.2 Collecting genome-wide histone modifications' data

Barski *et al.* in 2007 [Barski et al., 2007] have particularly generated high-resolution profile maps for genome-wide distribution of 20 histone lysine and arginine methylations as well as the distribution of histone variant H2A.Z, RNA polymerase II and the insulator binding protein CTCF across the human genome. Out of these we have particularly worked with only the distribution of histone methylations. They have identified typical patterns of these histone methylations exhibited at the promoters, insulators, enhancers and transcribed regions of the genome. They have successfully observed that monomethylation of H3K27, H3K9, H4K20, H3K79, and H2BK5 are all linked to gene activation, whereas trimethylation of H3K27, H3K9, H4K20, and H3K79 are linked to repression. The newly provided insights by their data into the function of histone methylations and chromatin organization in genomic functions also forms the basis for us to work with their data. With the genome wide histone modifications' data, we collected the corresponding gene expression levels in CD4+ T cells from Crawford *et al.* [Crawford et al., 2006].

The Figure 2.1 below is a snapshot of the UCSC Genome Browser, developed and maintained by the Genome Bioinformatics Group at University of California Santa Cruz (UCSC), USA. This browser contains the reference sequences and working draft assemblies for a large collection of genomes. It provides a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks of known genes, predicted genes, mRNAs, CpG islands etc. Most of these tracks are computed at UCSC from publicly available sequence data. The remaining tracks are provided by collaborators worldwide. Users can also add their own custom tracks to the browser for educational or research purposes.

The figure is a snapshot of the 84,000 bp long region on chromosome 12 from coordinate 6,742,459 to 6,826,779 of the human genome assembly 18. And the browser shows us all the necessary details like existence of certain histone methylation tag-counts, specifically H3K27me1 and H3K27me3, along with the genes present in this region. Similarly one could specify any other part of the genome to view many such details peculiar to those regions. The complete set of histone methylations from Barski *et al.* [Barski et al., 2007] can be browsed in this manner.

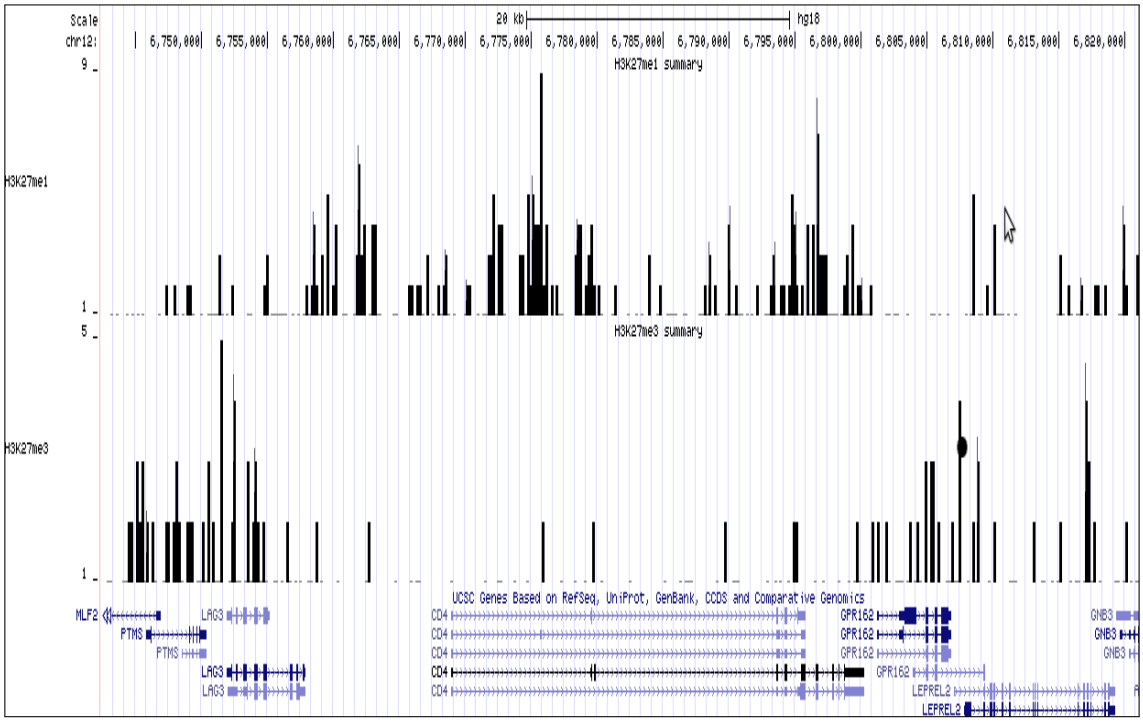

Figure 2.1: UCSC Genome browser snapshot: 84,000 bp long region on chromosome 12 from coordinate 6,742,459 to 6,820,000 of the human genome assembly 18

To determine if these modifications are associated with elevated levels of nearby gene expression, we determined the average expression value of genes that had nearby clusters of histone modifications. This modifications' data appears as tag counts lying in 200 bp long windows spread across the whole genome. We calculated the tag density (number of tags per base pair) located in the promoter regions of the genes in CD4+ T cells, typically 3 kb windows (2 kb upstream and 1 kb downstream) relative to the transcription start sites (TSS). These TSS for the genes in CD4+ T cells were learnt by mapping the expression probes from the GenBank database [#ref/appendix] to the UCSC known genes database.

The Figure 2.2 shows an example set of histone methylation patterns at active and inactive genes in the genome. Arbitrarily, at chromosome 2 in the window 191600000 to about 192000000, it shows an active and inactive region and the corresponding genes in those regions namely MYO1B (inactive), STAT1 and STAT4 (active), not necessarily the CD4+ T cells'

genes. Consider the red-marked histone methylation H3K27me3. Very easily, even visible to the naked eye, this modification follows a structure. It appears more dense in the inactive region and comparatively almost nil or very less density in the active region.

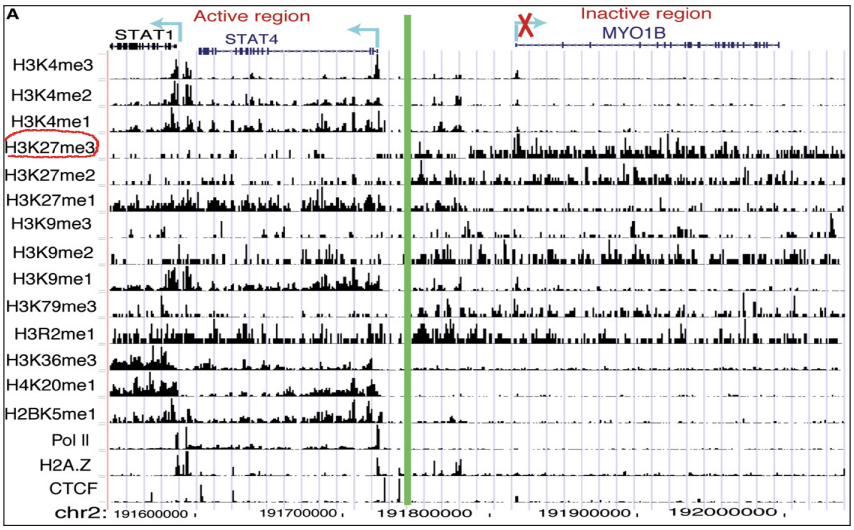

Figure 2.2: A typical example of histone methylation patterns at active and inactive genes (Barski *et al.*, Cell 129:823-837, 2007.)

When we compared the gene expression levels in CD4+ T cells with the density of this particular methylation, H3K27me3, we did observe a correlation value of $\rho = -0.4633$. It is shown in Figure 2.3 below.

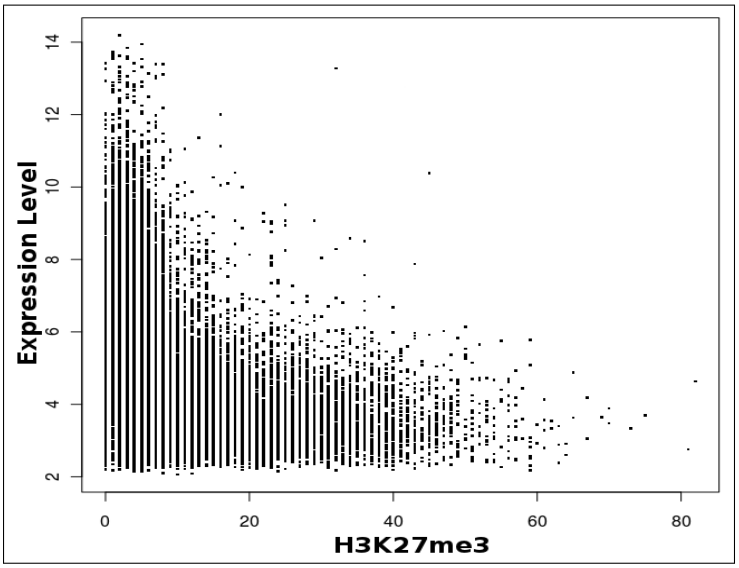

Figure 2.3: Scatter analysis of H3K27me3. This modification shows a negative correlation with gene expression levels in CD4+ T cells; $\rho = -0.4633$

With so much preprocessing to get the data into the required format and satisfactorily attempting to re-establish relations between the modifications and gene expression levels in CD4+ T cells, we move on to discuss our model, methodology adopted and results in the next chapter.

Chapter 3

Part I – Model, methodology and results

By our first goal, we intend to build a model that on the basis of the histone modifications' data can predict for us the expression level of a gene in CD4+ T cells. More technically, we want this system to act as a *classifier* that is capable of predicting the gene expression levels in the concerned cells based only on these histone modifications as features. Details of this classification model are discussed ahead.

3.1 Classification Model

Intuitively, in brief, *classification* is the problem of identifying the class or the category of a new observation from an already known set of classes on the basis of data/observations that the classifier is trained on. This *'training'* data is a collection of observations whose class memberships are known in advance. For observations of a particular class, the classifier tries to note the values their characteristic features take or the pattern they follow, if any, and accordingly it later attempts to classify any new observations it comes across. This is very typically known as *learning from data*. A more mathematical explanation of classification is deferred to a later point in this thesis.

For our work, from the histone modifications' data discussed, we have exactly 21 histone methylations' information as our features of each observation input to the classifier. The corresponding gene expression level is the response variable for each input. But what we have overlooked about the gene expressions until now is that the gene expression levels are basically intensity values represented by their logarithms taken to base 2. In fact these intensities are the number of mRNA molecules that are produced from each gene whose logarithms are then computed. Hence these are inherently *'continuous'* values. With an aim to predict the expression level we have only talked about classes, which are always to be discrete. The next section describes in short how we obtained discrete classes from the continuous response variable.

3.2 Methodology

We collected about 26,000 instances, each with 21 features, and their corresponding expression levels. These features were the 21 methylations' counts falling in the promoter regions. The gene expression levels of these instances ranged approximately from 2.0 to just over 14.0 on the logarithmic scale. You may recollect from Figure 2.2 earlier that the negative correlation between a methylation and active genes was clearly visible. So we simply decided upon 2 optimal threshold values, one that marks the instances where the genes are active or highly expressed

and the other marking repressed or inactive genes. The Figure 3.1 depicts this via a frequency histogram of gene expression levels of all the 26,000 instances. It is worth to note here that a repressed gene does not necessarily mean an inactive gene.

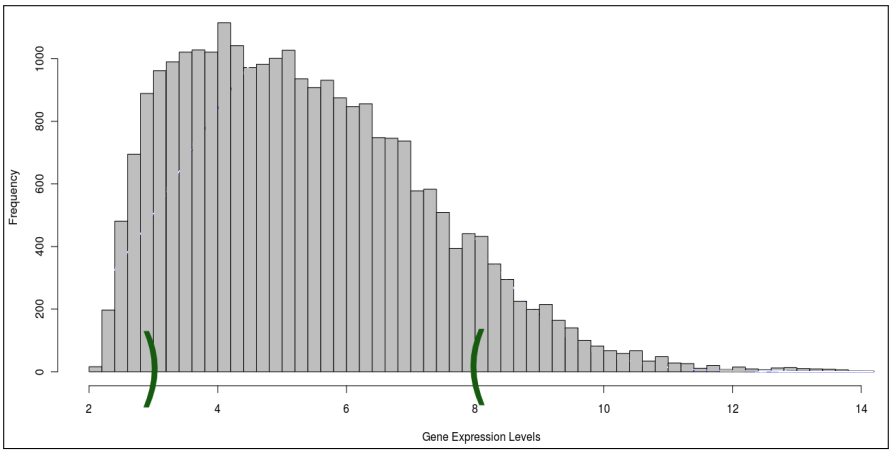

Figure 3.1: Frequency histogram showing distribution of gene expression levels in CD4+ T cells

One heuristic to decide the lower threshold as 3.0 and upper threshold as 8.0 is that we also looked to have just enough number of instances for the classifier to learn the classes well. We thus formed a binary classification model – class label '+1' if the expression level is beyond 8.0 and class label '−1' if under 2.0.

3.3 Results

Here we discuss and analyze the model performance and what measures were taken in terms of tuning the model parameters to maximize its performance.

Since we have built a binary classification model here, what we mean by maximizing the model performance is that we would like the model to learn the intricacies of the problem at hand and be able to accurately discriminate between the two classes. Ideally, the classifier could correctly predict the class labels for each and every new observation it makes, giving an accuracy of 100%. We employed a support vector classifier, also called a support vector machine (SVM) [Hastie et al., 2009, Boser et al., 1992]. The details of the support vector classifier with the parameter values we used are given next.

3.3.1 Support Vector Classification

We used 'LIBSVM - A Library for Support Vector Machines' created by Chih-Chung Chang and Chih-Jen Lin [Chang and Lin, 2011] for implementing our classification model. A simple support vector classifier is a binary linear classifier which can perform a non-linear classification with what is popularly known as the kernel trick [Boser et al., 1992]. We worked with linear and polynomial kernels to start with. Though the dimension of our feature space was only 21, which isn't so high considering the typical dimensionality issues SVM is capable of handling otherwise, we still had a large number of example instances to train SVM on, about $26,000$ instances. This made it difficult for SVM employing linear and polynomial kernel to converge to optimal maximum margin hyperplanes within 100,000 iterations. We thus opted to work with a radial basis kernel. More on the kernel trick in appendix.

Thus, the tuned set of values for all parameters for the classifier are given below:

Parameter	Values
Cost	10
Gamma for radial basis kernel	1 x 10^{-5}
Folds for cross-validation	10

Table 3.1: Tuned parameter values for SVM

3.3.2 Receiver Operating Characteristics Curve

Simply put, this curve measures the operating characteristics of a signal receiver. In other words it tells us the receiver's capability to accurately distinguish signal from noise. Receiver Operating Characteristics of a classifier shows its performance as a trade off between *selectivity* and *sensitivity*. It is a plot of *'true positives'* vs. the *'true negatives'*. In place of *'true negatives'*, one could also use *'false positives'* which are essentially {1 - *'true negatives'*}. This is plotted on a scale of 0-1 for both the axes. This curve always goes through (0,0) and (1,1).

A classifier's performance can be evaluated by knowing its confusion matrix (appendix). We use that matrix to plot a ROC curve for a classifier and the area under the curve (AUC) of this plot pictorially tells us the classifier's accuracy – the discriminative power of the classifier. The ROC curve of an ideal classifier (100% accuracy) has an AUC of 1, with 0.0 *'false positives'* and 1.0 *'true positives'* (all new observations correctly classified). On the contrary, the ROC curve of what we call a 'random guess classifier', when the classifier is completely confused and cannot at all distinguish between the two classes, has an AUC of 0.5, the *'$x = y$'* line in the plot.

Figure 3.2 is such a ROC curve of binary classification with only the 21 histone modifications predicting gene expression levels. In this and all the following ROC curves we have also denoted the responses for (a) an ideal classifier in blue and (b) the random guess classifier in **turquoise blue**.

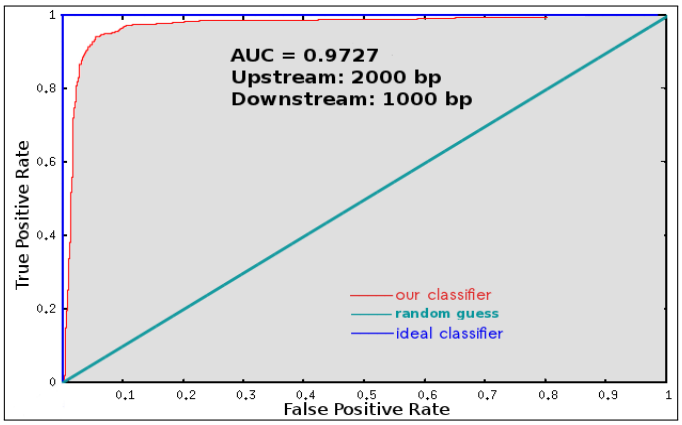

Figure 3.2: ROC curve of binary classification with only histone modifications predicting gene expression levels

We used a window size of 3000 bp, relative to the TSS, as the promoter region, 2000 bp upstream and 1000 bp downstream. But this window size can really be anything, because we only know that the promoter region for a gene is a few base pairs long and lies in vicinity to that gene. Hence, we varied this window size with combinations of different upstream downstream sizes relative to the TSS. The bar-plot in Figure 3.3 depicts the AUCs of the ROC curves obtained with these different combinations of upstream downstream window sizes – upstream abbreviated as U and downstream as D.

We observe that our classification model features have demonstrated a high discriminative

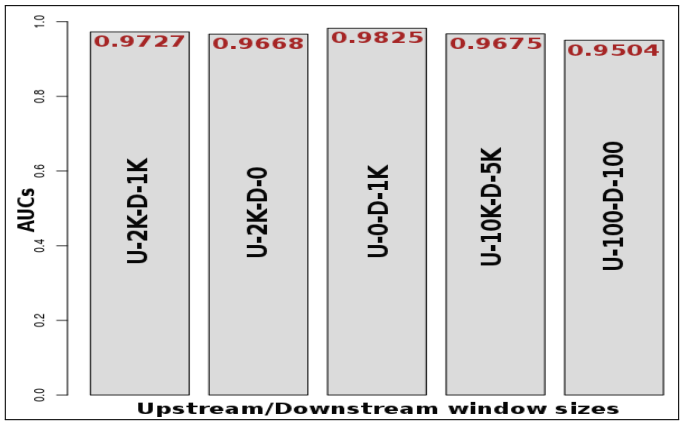

Figure 3.3: Barplot of the AUCs for combinations of upstream/downstream window sizes

ability with all the various window sizes. But why, we thought, should this be the case? Is our model so convincingly capable of discriminating the classes? To satisfy ourselves of our classifier, that it itself isn't doing something outright wrong insulated from the domain perspective, we attempted to confuse our classifier. For all the training instances fed to the classifier, we permuted their class labels with uniform randomness. So the feature set for the training data now may or may not exhibit the earlier structure/pattern since all instances with the same true class label have now got randomly dispersed in either of the classes. In such a scenario we expect the classifier to thus behave as a 'random guess classifier'. Our ROC curve for this training data with randomized class labels is Figure 3.4. We have retained the upstream/downstream window sizes to 2000 and 1000 bp respectively.

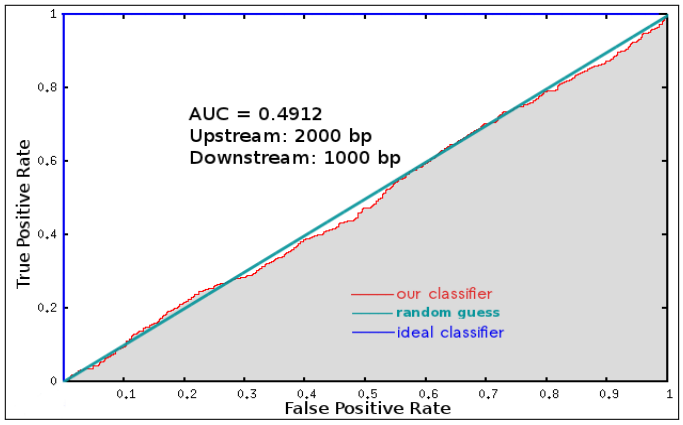

Figure 3.4: ROC curve with randomized class labels

As the curve depicts, our classifier does get confused with the AUC falling to 0.4912. This corroborates our methodology. What remains to be checked is whether we can garner a similar support for these accuracies from the domain perspective.

We performed similar classification procedures in 2 more, different ways. Since we had collected these feature values from the promoter regions of the genes to predict the gene expression levels, as a first we moved out of the promoter region of these genes and performed the same procedure of collecting the modifications' values in 3000 bp long windows. More specifically, we moved away from the TSS by 100,000 bp and considered similar windows there to predict the expression levels of genes lying hereof. The ROC curve for these specifications are given in Figure 3.5. The AUC has reduced comparatively; AUC = 0.7622 << 0.9727.

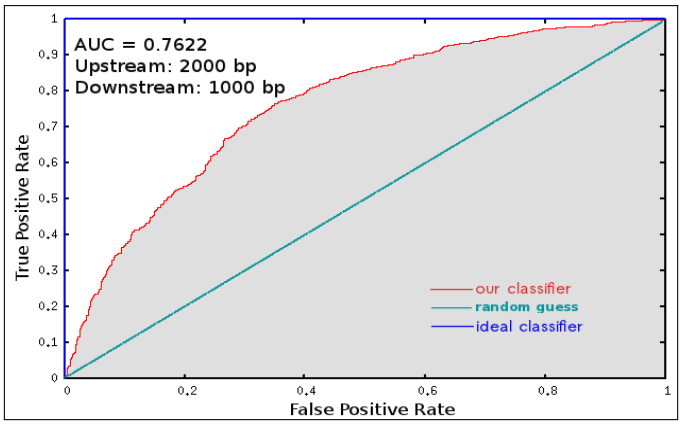

Figure 3.5: ROC curve after moving away by 100,000 bp from the promoter region

As a second, we attempted to play with the way we have generated the two classes from our response variable. The earlier lower threshold was changed from 2.0 to a range of $4.0 - 5.0$. Thus, effectively, we diminished the distance between the two classes expecting to have made it at least little more difficult for the classifier to discriminate between the two class members now than the initial setting. The distance here being their separation on the x-axis, denoting the expression levels. Figure 3.6 and 3.7 show the change in class boundaries and the corresponding ROC curves respectively.

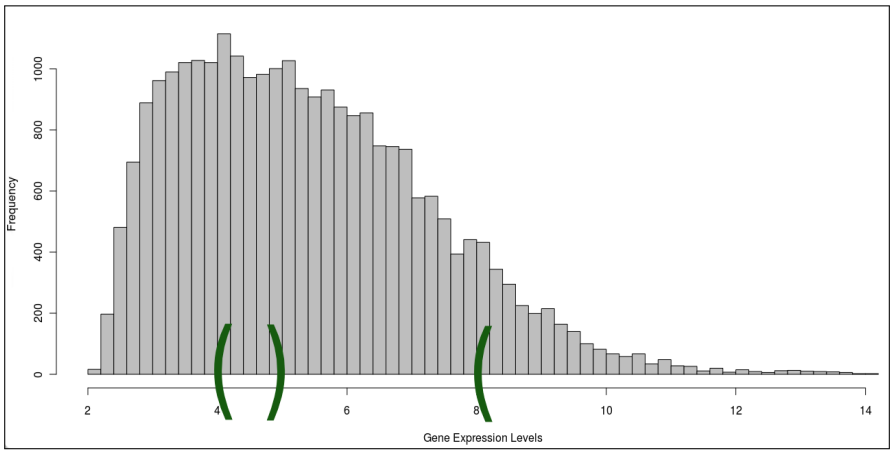

Figure 3.6: Distribution of gene expression levels in CD4+ T cells with a varied class boundary I

We moved the classes more closer to each other changing the lower class boundaries to $5.0 - 6.0$. Each time we made sure that the classifier also has enough examples to train on. The AUC now reduced to 0.8154 (Figure 3.8 and 3.9).

Thus we satisfied ourselves on both methodology and domain fronts. The histone modifications do seem to have a considerably high discriminative capability in terms of predicting whether the gene is highly expressed or not.

3.4 Goals revisited

Of our goals:

1. to predict expression level of a gene by histone modifications at its promoters; and

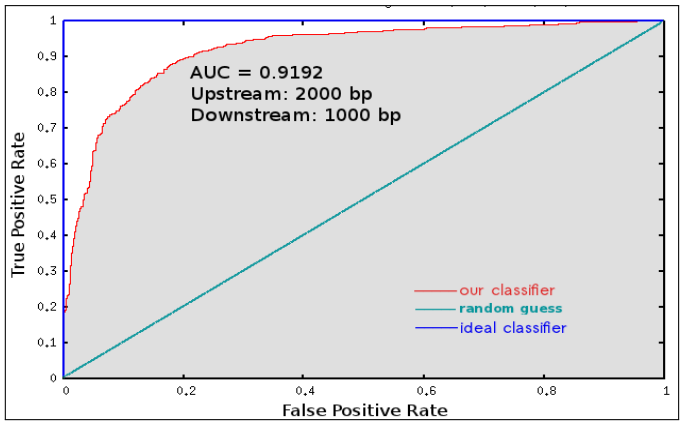

Figure 3.7: ROC curve for varied class boundary I

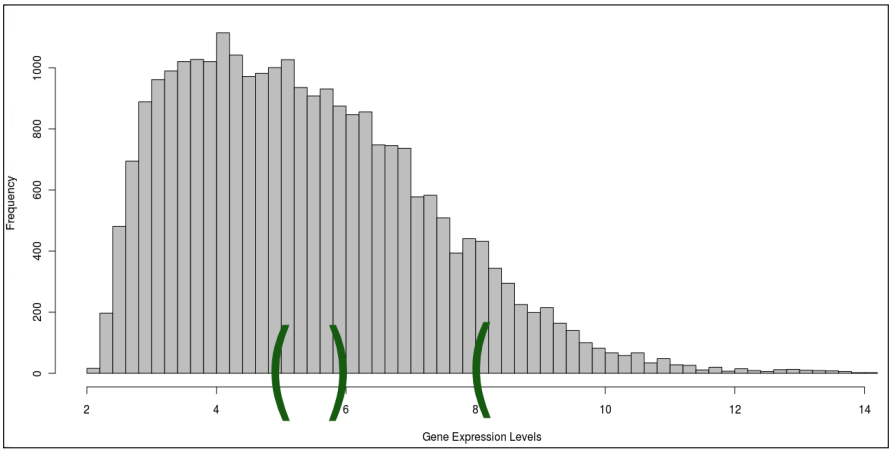

Figure 3.8: Distribution of gene expression levels in CD4+ T cells with varied class boundary II

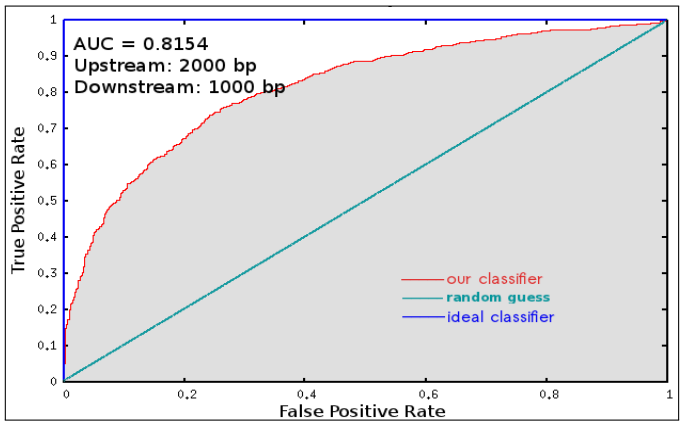

Figure 3.9: ROC curve for varied class boundary II

2. to characterize tissue specificity of promoters using these histone modifications,

the tissue specificity of promoters is still to be dealt with. The next chapter deals with this and explains what prompted us to believe in histone modifications, the epigenetic factors, to characterize such tissue specificity.

Chapter 4

Part II – Model and methodology

In part I, we looked at histone modifications in CD4+ T cells located at the promoters and the corresponding results were discussed. These epigenetic factors did seem to have a considerable discriminative ability in predicting the gene expression levels in T cells. In part II, with confidence from part I results, we used the same features of 21 histone methylations and replaced the gene expression levels in T cells by the expressions from heart tissue. In the sections that follow, we discuss the source of these tissue specific promoter regions, initial results with heart tissue which then proved to be our motivation and premise for the part II.

4.1 Preliminary results for part II

Schug *et al.* in 2005 [Schug et al., 2005] performed a genome-wide analysis of promoters in the context of gene expression patterns with tissue specificity. They studied 25 tissues in humans and mouse. We have used the heart tissue data with few other human tissues listed in Table 4.1.

Adrenaline gland	Amygdala	Cerebellum	Cortex
Kidney	Liver	Lung	Ovary
Pancreas	Pituitary gland	Placenta	Prostate
Salivary gland	Spinal cord	Spleen	Testis
Thalamus	Thymus	Thyriod	Trachea
Uterus	Caudate nucleus	Corpus callosum	Drg

Table 4.1: 25 tissues in humans studied by Schug *et al.*

We arbitrarily selected heart tissue and replaced our earlier model's response variable by the expression levels of genes in heart tissue but retained the same feature values which were collected at the promoter regions of genes in CD4+ T cells. The remaining model with all the tuned parameter values was completely retained and used as it is. The ROC curve for this initial setting is given in Figure 4.1.

Schug *et al.* have measured two kinds of tissues specificity 'overall' tissue specificity and 'categorical' tissue specificity. Overall tissue specificity ranks a gene according to the degree to which its expression pattern differs from ubiquitous uniform expression and categorical tissue specificity places special emphasis on a particular tissue of interest and ranks a gene according to the degree to which its expression pattern is skewed toward expression in only that particular tissue. This categorical tissue specificity is denoted by Q. It is near its minimum of zero when a gene is relatively highly expressed in a small number of tissues including the tissue of interest, and becomes higher as either the number of tissues expressing the gene become higher, or as the relative contribution of the tissue to the gene's overall pattern becomes smaller. We thus

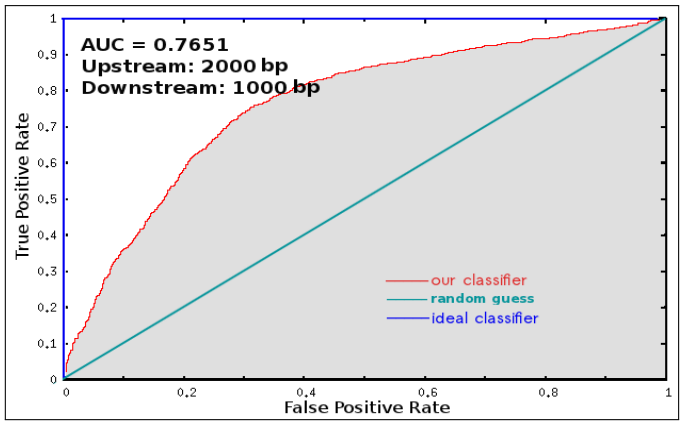

Figure 4.1: ROC curve with heart data (*AUC = Area Under the Curve)

have used this Q value as our response variable with a certain threshold deciding whether the Q value is low enough and the corresponding class label is 'more tissue specific' and vice versa. The feature values too now came from promoter regions of the genes in heart tissue. The ROC curve with this new response variable for our model is plotted in Figure 4.2.

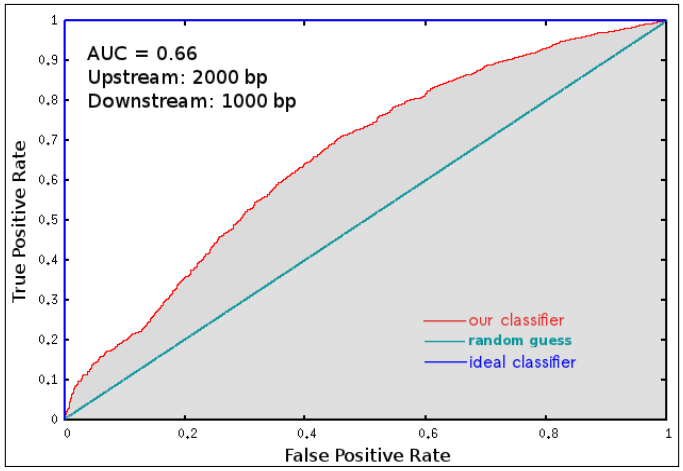

Figure 4.2: ROC curve using Q of heart data

We observe that when classes made out of the gene expression levels in heart tissue were used in our model, the classifier performed well to achieve an accuracy of 0.7651 in spite of having the modifications' values pertaining to promoter regions of genes in CD4+ T cells. On the contrary, when we used the Q values from Schug *et al.* denoting tissue specificity, the performance went down but the classifier did not loose its discriminative ability completely, having an AUC of 0.66. The histone modifications have served well as features to under-estimate them so easily. Hence, we decided to supplement these histone modifications with another set of features in characterizing the tissue specificity of promoters.

4.2 Additional set of features

We performed a Markov analysis of the genomic sequence data composed of adenine (`A`), cytosine (`C`), guanine (`G`), and thymine (`T`. We collected this sequence information located at the genes for each tissue provided by Schug *et al.* (25 tissues in all). We selected a gene on the basis of the Q value assigned to it. The ones with a low Q value were selected. 3000 bp long sequences

relative to each TSS were considered in any tissue. 5500 such sequences were collected in this manner.

The sequences were modeled as Markov chains of orders 1 to 9. We observed that modeling a promoter sequence as a Markov chain of order 4 facilitates good predictions of whether a given sequence drives the expression of a gene specific to a given tissue. Thus using a Markov chain of order 4, we computed the log-probabilities that a sequence belongs to each tissue. Since we dealt with 25 tissues, every sequence from the collection of 5500 sequences had corresponding 25 log-probabilities. These 25 log-probabilities are used as an additional set of features along with the 21 histone modifications' data for the same region as the sequences. Thus in our classification model for part II:

$$\#\text{features} = \underset{\text{(histone modifications)}}{21} + \underset{\text{(tissue specific log-likelihoods)}}{25}$$

$$\#\text{Class labels} = \underset{\text{(tissues)}}{25}$$

Since we have moved on to identify tissue specific signals in promoters in part II, our class labels are now no longer binary as in part I. Every sequence in the training data will now have the particular tissue that it belongs to as its class label. Thus we have 25 classes and the problem is now a multi-class classification problem.

We expect that our classifier with these additional features, from Markov analysis of genomic sequences, supplemented to the epigenetic factors could perform comparatively better than the epigenetic factors alone as features. Overall, the combination of both of these feature sets could posses a benefit of higher characterizability of tissue specific signals in promoters.

We convey some concluding remarks and future work in the next chapter.

Chapter 5

Conclusions and Future work

In this chapter we summarize our work. We further analyze some results produced earlier and reason about why certain things worked the way they did. We also discuss some future research directions.

We demonstrated that the epigenetic factors, more specifically the histone methylations are capable of discriminating between genes that are highly expressed and otherwise, or in other words the histone modifications can be successfully deployed to predict the expression levels of genes in CD4+ T cells, which can be extended to others. This will certainly work at least when the classes are far away from each other that the classifier can easily distinguish between them.

We worked with a radial-basis kernel function in order to transform the feature space. For the linear and polynomial kernels the *max-iterations* of 100,000 weren't enough for the optimizers to converge to a set of optimal hyperplane equations for the support vector classifier. Hence, we wouldn't claim that the linear and polynomial kernels are 'incapable' of finding the optimal hyperplane equations, but are computationally intensive. The dimension of our feature space is not very high, it is 21 which is very much workable without any hassles for the optimizers. On the contrary, what happened to be more important for this problem than its feature space dimension is the number of training examples. Having a large number of those can make the problem computationally intensive.

We also saw that when we replaced the gene expression levels in CD4+ T cells by those in heart tissue but kept the corresponding histone modification values from the promoter regions of genes in CD4+ T cells as features, the classifier still managed an accuracy of $\tilde{0}.76$. We believe the reason that this was so is that T cells are immune cells, they assist other white blood cells in immunologic processes. Since every part of our body, even the tissues would require some immune cells amongst them to protect from or fight infections and their causes, the genes in CD4+ T cells may be expressed to some extent is all these tissues. And hence that accuracy.

Also, when we moved 100,000 base pairs away from the promoter region, we expected a further dip in accuracy than we actually observed. Probably, one reason why expected dip did not really happen could be because our DNA usually has long modules (parts of the DNA) in which the adjacent modules may be insulated from each other but the way the DNA folds itself there still could be chances of interaction between 2 very distant regions which can't be ruled out completely.

So this convinces us that the histone modifications can exhibit a good characterizability when it comes to predicting gene expression levels. But this discriminative quality of theirs declined to some extent when we attempted to extract tissue specific signals in promoters. The experiment that we performed with heart could be re-performed with some other tissues like the skin tissues, lung tissues or the cerebellum tissue. There may be some pattern that the histone modifications may work well for selected tissues like skin and lung but not for cerebellum. Thinking more of it, these are epigenetic factors which may be governed by external/environment factors that do

not govern all the tissues in a similar fashion.

More over, we know that the genomic sequences composed of adenine (`A`), cytosine (`C`), guanine (`G`), and thymine (`T` are the most basic entities participating in every gene expression or regulation taking place in the body. Hence we wouldn't be completely wrong in believing that even when the histone modifications saw some drop their discriminative power, the inclusion of additional set of features from the Markov analysis of the genomic sequences can boost the classifier's ability again. And thus came our classification model of part II, from binary classification to multi-class classification to deal with 25 target tissues.

In closing, there are some possible future directions. We could perform renewed Markov analysis by varying the upstream/downstream window sizes and effectively vary the sequence lengths. Also, depending upon the kind of problem at hand, we could let the model completely switch from histone modifications as features to the sequence analysis information as features or just have a complete set of features with weighted selection. Currently, we only have genome-wide information for histone methylations available. May be with advances in technologies all the other modifications like the genome-wide acetylations, phosphorylations etc. might become available and including them could throw up some more intriguing results.

Bibliography

[Barski et al., 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhaocorrespondenceemail, K. (2007). High-resolution profiling of histone modifications in the human genome. *Cell*, 129:823–837.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA. ACM.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

[Crawford et al., 2006] Crawford, G. E., Davis, S., Scacheri, P. C., Renaud, G., Halawi, M. J., Erdos, M. R., Green, R., Meltzer, P. S., Wolfsberg, T. G., and Collins, F. S. (2006). Dnase-chip: A high-resolution method to identify dnase i hypersensitive sites using tiled microarrays. *Nature Methods*, 3:503–509.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, USA, second edition.

[Kouzarides, 2007] Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128:693–705.

[Schug et al., 2005] Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., and Stoeckert, C. J. (2005). Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biology*, 6:R33.

Appendix A

Molecular Biology Excerpts

A.0.1 Structure of a nucleosome

Figure alongside depicts the structure of a nucleosome. It contains a protein core made of eight histone molecules. The nucleosome core particle is released from chromatin by digestion of the linker DNA with a nuclease, an enzyme that breaks down DNA.

After dissociation of the isolated nucleosome into its protein core and DNA, the length of the DNA that was wound around the core can be determined. It is precisely 146 base pairs long. This length of 146 nucleotide pairs is sufficient to wrap 1.65 times around the histone core.

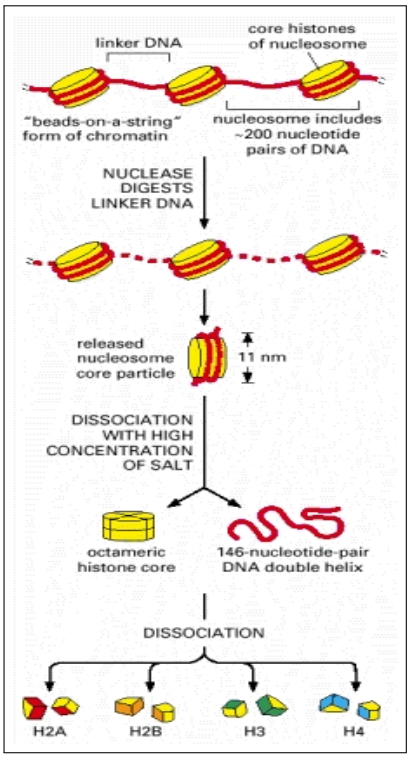

Figure A.1: Nucleosome structure. (Molecular Biology of the Cell. 4th edi- Bookshelf)

A.0.2 Promoters

Promoters

- are a special sequence of nucleotides proximal to the starting point for RNA synthesis, in other words, a transcriptional start site of a gene.
- typically lie within few base-pairs upstream and downstream of the gene they regulate.

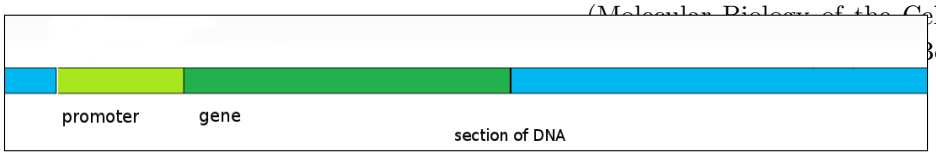

Figure A.2: Pictorial depiction of a section of DNA showing a gene and the corresponding promoter region

Appendix B

Acronyms

DNA	Deoxyribo Nucleic Acid
RNA	Ribo Nucleic Acid
mRNA	messenger RNA
TSS	Transcription Start Site
bp	base pairs
DRG	Dorsal Root Ganglia (tissue)
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
SVM	Support Vector Machines