POTTS MODEL CLUSTERING FOR DISCOVERING PATTERNS OF EPIGENETIC MARKS

BY JUNYI LI

A Dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

Written under the direction of

Professor Anirvan Sengupta

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2011

ABSTRACT OF THE DISSERTATION

Potts model clustering for discovering patterns of epigenetic marks

by Junyi Li

Dissertation Director: Professor Anirvan Sengupta

Study of epigenetics leads to understanding of the regulation of gene expression not caused by the changes in the underlying DNA sequences. This area of biological research has drawn much interest as large amounts of epigenetic data from numerous experiments were generated in recent past. In this thesis, we use the Potts model clustering method, which is based on statistical mechanics, to discover patterns in histone modification data. After a general overview of the epigenome and an introduction to common methods of clustering, we discuss why we need special cluster analysis methods for the data at hand. Then, we introduce our tool of choice, namely, the superparamagnetic Potts clustering method. We discuss the Potts model and the Swendsen-Wang method of Monte Carlo simulation, which avoids the usual slowing down experienced near phase transition. We apply Potts model clustering to histone modification data in highly conserved regions to discover the patterns of epigenetic marks and compare them with background reading. We also contrast the results from our method with that from usual K-means clustering approach. Finally, we discuss the biological significance of our computational results.

Acknowledgements

First and foremost, I would like to express deep and sincere gratitude to my Ph.D. advisor, Prof. Anirvan Sengupta, for his support, guidance, patience and great help. Anirvan taught me both a wide range of subjects and a large number of research skills. With his inspiration and enthusiasm, Anirvan also showed me his scientific and insightful thoughts during his classes and seminars. I truly feel that I have learned a lot from a brilliant professor and scientist.

I would like to thank Prof. Ronald Ransome for his kind help in my graduate study and life in Rutgers.

I would like to thank Prof. Gyan Bhanot and Prof. Vincenzo Pirrotta for offering kind help and being in my thesis committee. The same gratitude goes to my other committee members Prof. Robert Bartynski and Prof. Jerry Sellwood, for all the kind help and valuable suggestion since my first committee meeting.

I would also like to thank Adel Dayarian for collaborating with me on this project. It is very pleasant to work with him.

I would also like to thank many other members of the BioMaPS Institute: Alexandre Morozov, Viji Nagaraj, Deepangi Pandit, George Locke, Michael Manhart, Allan Haldane, Manjul Apratim, Ariella Sasson, Kevin Abbey and Katherine Lam.

I would like to thank Physics laboratory administrator Gabe Alba and Hsu-Chang Lu for their help when I worked as a teaching assistant in Physics Lab courses.

Thanks to my mother Shiping Zhao and my father Changfu Li, my grandparants, my aunts, my uncles and my cousins, for your great love and unconditional support. I would like to thank

my dear husband Xinjie Wang who has given me understanding, encouragement and love ever since we met in college. Thanks to my dear son Junwen for bringing me the priceless happiness of being a mother. I am deeply grateful to my family from the bottom of my heart.

Lastly, I would like to thank all my friends. I am honored to have your friendship.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to the Genome and the Epigenome

1.1 DNA and Gene Regulation

1.1.1 DNA and Gene Expression

DNA, or deoxyribonucleic acid, is a nucleic acid which provides a blueprint that directs all cellular activities and specifies the developmental plan of multicellular organisms [1]. Organisms have their genetic material in the form of one or more long DNA molecules. These molecules, often decorated with other proteins, are called the chromosomes.

In DNA, there are four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA bases can pair up specifically with each other, A with T and C with G. Each base is also attached to a sugar molecule and a phosphate molecule. One base, one five-carbon sugar, and phosphate together are called a nucleotide. DNA is a long polymer made of repeating nucleotides. The structure of DNA was first discovered James D. Watson and Francis Crick [2]. As shown in Fig. 1.1, DNA is a double helix formed with base pairs inside and sugar-phosphate backbone on the outside.

A gene is defined as a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions [3]. Genes are made up of DNA and act by determining the structure of proteins. Gene expression is the process of producing a biologically functional molecule of either RNA or protein. The collection of all genes in an organism were originally referred to

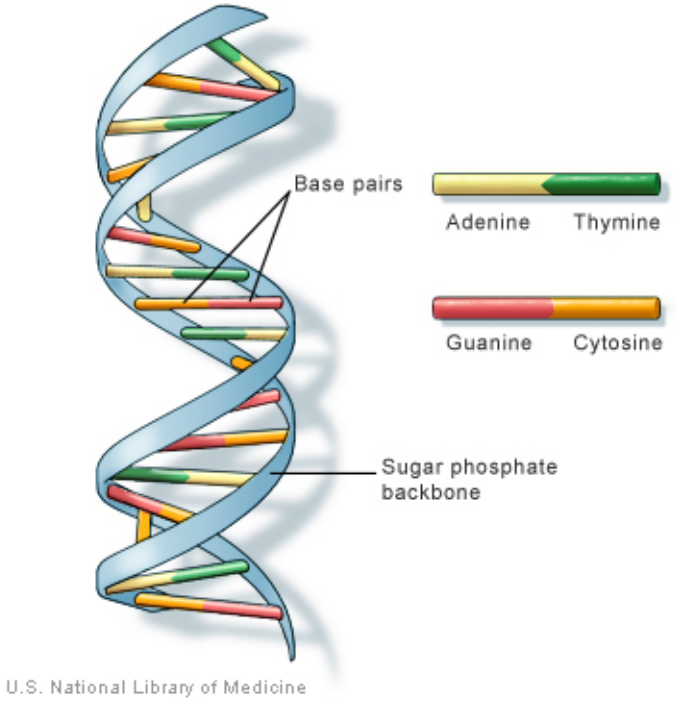

Figure 1.1: The structure of DNA. DNA is a double helix formed with base pairs inside and sugar-phosphate backbone on the outside. Source: U.S. National Library of Medicine.

as the genome, although today it often refers to all the DNA-related features on all the chromosomes, including features in the intergenic regions. The production of proteins directed by genes has two steps: genetic transcription and translation.

During the process of genetic transcription, the information stored in a genes DNA is transferred to RNA by enzymes called RNA polymerases. RNA usually is a single-stranded nucleic acid. The RNA that contains the information for making a protein is called messenger RNA (mRNA) because it carries the information from the DNA to the molecular machinery ready for translation.

Living organisms could be divided into two major groups: prokaryotes and eukaryotes. A eukaryotic cell has a nucleus which contains the genetics materials inside, while a prokaryotic

cell does not have such a well-defined structure. In prokaryotes, the process of transcription and translation may go on at the same place and time. In contrast, eukaryotic nascent RNAs need to be processed and transported out of the nucleus into a specific organelle for being translated.

During the process of genetic translation, the mRNA serves as a template and interacts with a specialized complex called a ribosome which synthesizes the proteins. Each sequence of three bases, called a codon, usually codes for one particular amino acid. Then a type of RNA called transfer RNA (tRNA), which is hybridized with the codon, plays a role bringing the appropriate amino acid for protein synthesis.

Fig. 1.2 shows the flow of genetic information from DNA to RNA to proteins. This is one of the fundamental principles of molecular biology and it is stated as the central dogma of molecular biology by Francis Crick [4].

1.1.2 Regulation of Gene Expression

Transcriptional Regulation

The first step of gene expression, the transcription of DNA into RNA, is the initial step where gene expression is regulated. In both prokaryotes and eukaryotes, transcriptional regulation often controls under what conditions transcription occurs and how much RNA is produced. The principal enzyme responsible for RNA synthesis is RNA polymerase (RNAP). RNAP initiates transcription at a specific DNA sequence known as a promoter. Particular subunits of RNAP bind to promoter elements and lead to the formation of initiation of complex. RNAP then starts to produce the RNA transcript which is complementary to the template DNA strand. The process of adding nucleotides to the RNA strand is known as elongation. RNAP will release its RNA transcript at specific DNA sequences encoded at the end of genes known as terminators. The whole process of transcription is shown in Fig. 1.3.

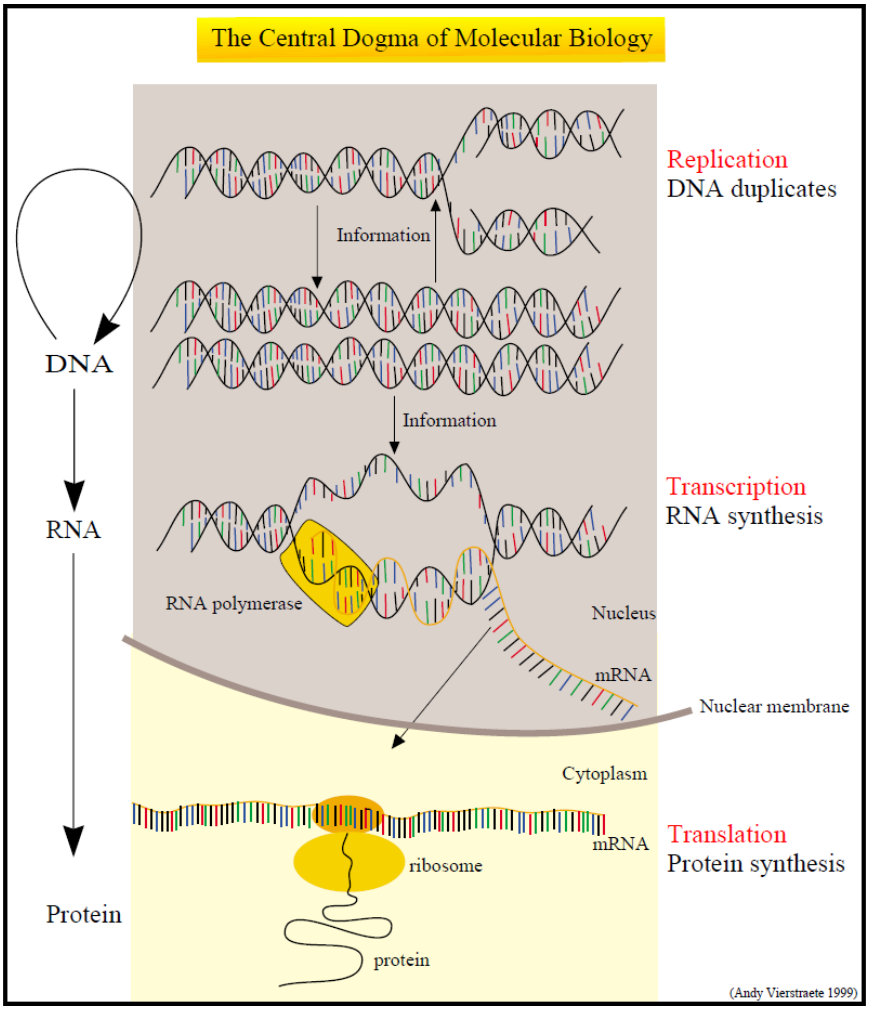

Figure 1.2: The central dogma of molecular biology. RNA molecules are synthesized DNA templates and proteins are synthesized from RNA. Source: http://users.ugent.be/~avierstr/principles/centraldogma.html.

An enhancer is a certain DNA sequence which can be bound with transcription factors to enhance transcription levels of genes. In higher eukaryotes, an enhancer can function even if it is located many hundred thousands of base pairs away from promoter and can be located upstream or downstream of the gene that it regulates. Under appropriate conditions, the enhancer is bound by activator proteins and activator proteins interact with the mediator complex, which recruits polymerase II and the general transcription factors. The DNA sequence corresponding to the enhancer could be reversed in orientation without affecting transcriptional enhancement.

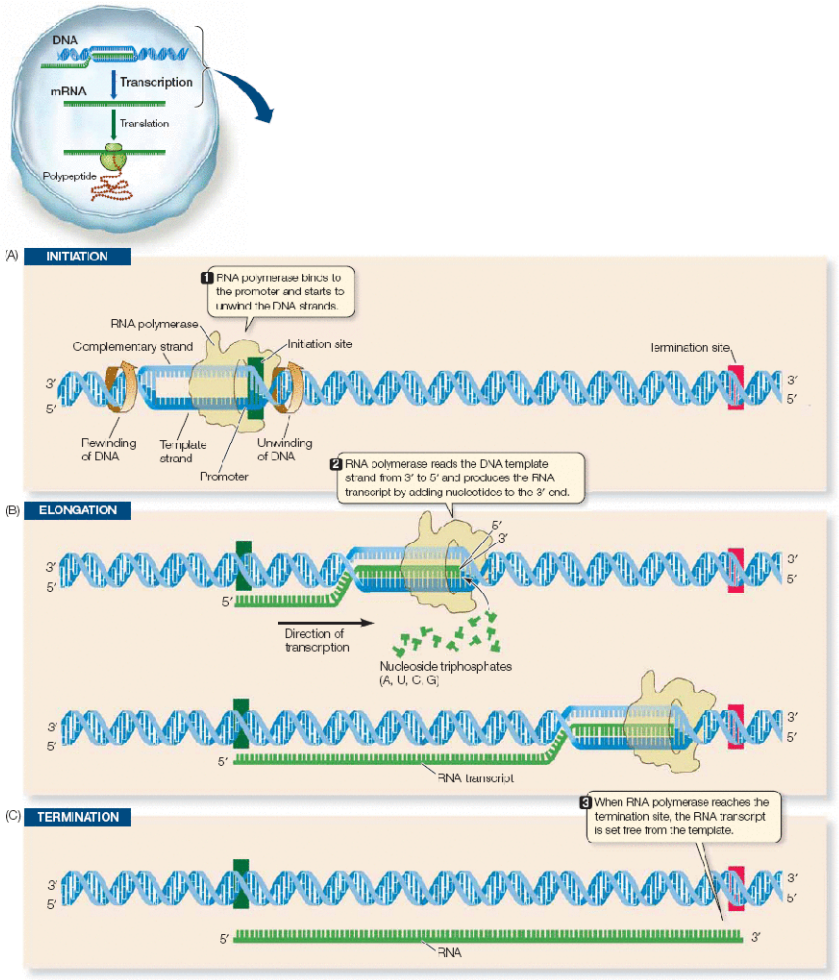

Figure 1.3: Transcription by RNA polymerase. During three distinct processes of transcription are initiation, elongation, and termination, RNAP initiates transcription at promoter, produce the RNA transcript and release RNA transcript at terminator. RNA polymerase is much larger in reality than indicated here, covering about 50 base pairs. Source: http://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393

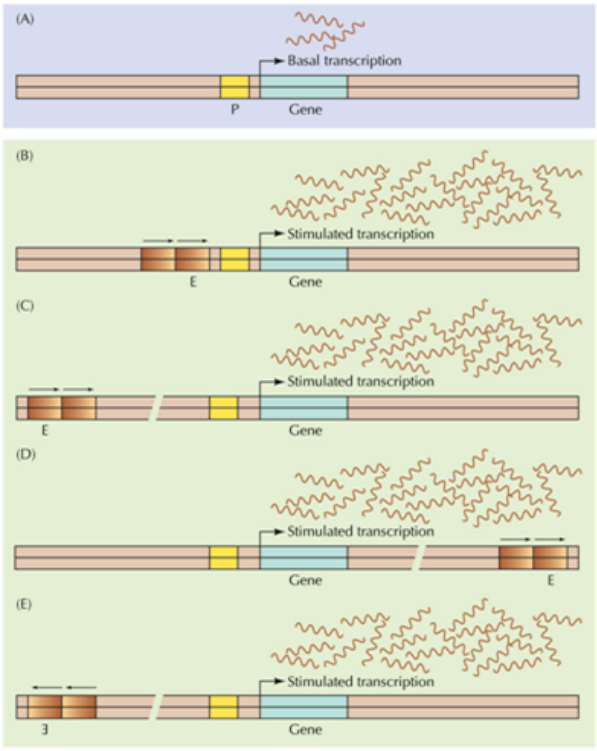

Figure 1.4: Actions of Enhancer. (a) Gene is transcribed at low base level without enhancer E. (b) Gene is transcribed at high level when enhancer E is added. (c) Gene is transcribed at high level when enhancer E is inserted several thousand base pairs upstream from promoter. (d) Gene is transcribed at high level when enhancer E is inserted several thousand base pairs downstream from promoter. (e) Gene is transcribed at high level when enhancer E is in reversed orientation. Source: Chapter 7 of The Cell: A Molecular Approach.

In some cases, an enhancer also can be excised and inserted elsewhere in the chromosome, and still affect gene transcription. Fig. 1.4 illustrates many of these features of an enhancer.

A silencer is a certain DNA sequence located thousands of base pairs away from the gene it regulates. When certain protein factors bind to it, the silencer gives rise to particular modifications of the chromosomal materials around it. These modifications lead to the repression of one or more genes in that region.

The transcription step is followed by translation in gene expression. Gene expression is

still regulated at the stage of translation. Translation of mRNA can be regulated by the binding of repressor proteins, noncoding microRNAs, and controlled polyadenylation [1]. Interested readers are referred to general molecular biology textbooks for more information (e.g. [1]). Our focus would be on the transcriptional regulation rather than the translational regulation, since both epigenetic chromatin modifications and enhancers play an important role in the transcriptional regulation.

1.2 Epigenome

There has been remarkable progress in genetic research since the initial sequencing of human genome was completed in 2001 [5]. Increasingly, high-throughput technologies provide large-scale sequence data, expression maps and functional annotations of genes and their expression in many genomes. However, just DNA-level investigation is not able to explain the whole regulation of gene expression, especially in the context of development, where the same genome sequence can develop numbers of different cell types. Beyond the DNA sequence level, there is a layer of epigenetic information which is associated with gene expression in different developmental stages, cell types and disease states.

Epigenetics is defined as the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms [6], or of heritable traits that do not involve changes to the underlying DNA sequence [7]. The overall epigenetic state of a cell is often referred to as the epigenome, in analogy to the word genome. Epigenetic modifications normally fall into two main categories: DNA methylation and histone modification.

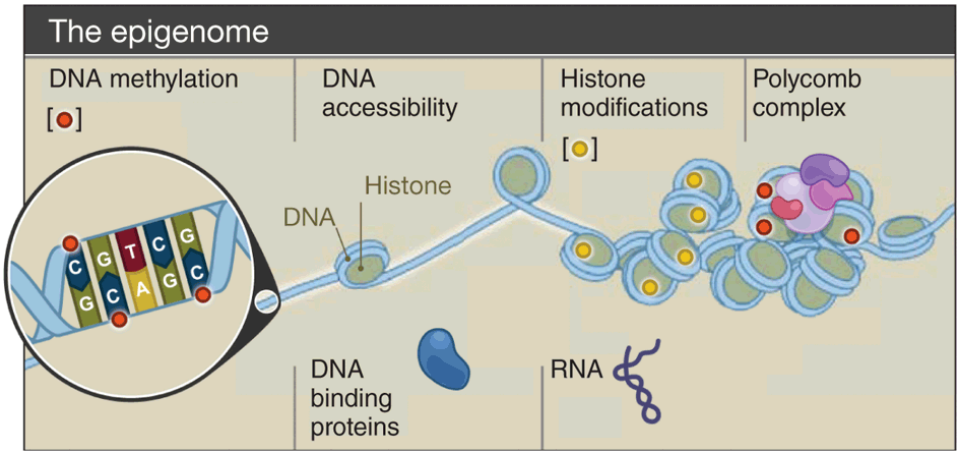

Figure 1.5: The epigenome. Epigenome refers to the overall epigenetic state of a cell. Epigenetic modifications fall into two main categories: DNA methylation and histone modifications. Source: The NIH Roadmap Epigenomics Mapping Consortium.

1.2.1 DNA Methylation

DNA methylation involves the addition of a methyl group to the position 5 of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring. This covalent modification plays an important role in gene regulation, chromosomal stability and parental imprinting. Research shows that CpG islands, which are genomic regions containing a high frequency of CpG sites (CpG indicates that the C and the G are linked through the phosphate group, as opposed to C and G from different strands which could be base paired), are resistant to methylation and are associated with most human genes [8]. Also, de novo methylation of promoters with CpG islands leads to gene inactivation.

Since we are mostly interested on histone modification data in the thesis, we will not discuss details of DNA methylation. Interested readers are referred to some reviews of DNA methylation research (e.g. [9]).

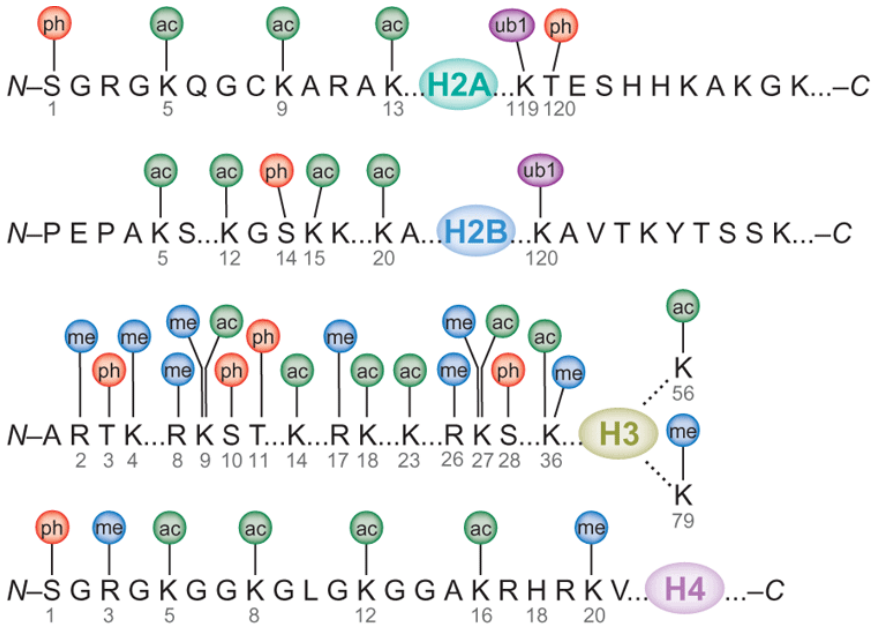

Figure 1.6: Modifications of histones. The modifications include acetylation(ac), methylation(me), phosphorylation(ph) and ubiquitination(ub1). Source: http://www.nature.com/nsmb/journal/v14/n11/full/nsmb1337.html

1.2.2 Histone Modification

Histone proteins are essential for packaging of DNA into chromosomes within the nucleus of a cell. They are highly conserved (meaning that these proteins are very similar across all eukaryotes) and can be grouped into five major classes: H1/H5, H2A, H2B, H3, and H4 [1]. There are a number of covalent modifications including methylation, acetylation, phosphorylation, ubiquitylation and ADP-ribosylation at the N-terminal tails of histones as shown in fig. 1.6. Histone modifications can alter DNA-histone interactions within and between nucleosomes and affect higher-order chromatin structures.

1.3 Genomic Profiles of Epigenetic Marks

1.3.1 Genome-wide Approaches to Study Histone Modifications

ChIP-chip

ChIP-chip is a technique that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip). ChIP-chip is used to investigate interactions between proteins and DNA. Whole-genome analysis can be performed by ChIP-chip to determine the locations of binding sites for almost any interesting proteins such as histones and histone modifications [10]. During the process of ChIP, chromatin fragments are isolated by antibodies that are specific to a feature of interest (like a specific chemical modification of a particular amino acid in one of the histone tails). After that the isolated fragments are amplified to generate micrograms of fluorescently labeled DNA which is hybridized to a DNA microarray (chip). Since ChIP-chip was first applied successfully to identify sites for histone modifications in 2002 [11, 12], it has become a powerful tool in determining genome-wide maps of histone modifications.

ChIP-Seq

ChIP-Seq is a recently developed technique that combines chromatin immunoprecipitation with high-throughput sequencing techniques. It is used to analyze protein interactions with DNA and can precisely map global binding sites for any protein of interest. During the process of ChIP, specific cross-linked DNA-protein complexes are enriched by using an antibody against a protein of interest. After that the ChIP DNA ends are repaired and ligated to a pair of adaptors, followed by limited PCR (Polymerase Chain Reaction) amplification. Then, all the resulting ChIP-DNA fragments are sequenced (hence the 'Seq' in the name) simultaneously using a genome sequencer. The number of sequenced reads, which is proportion to the modification

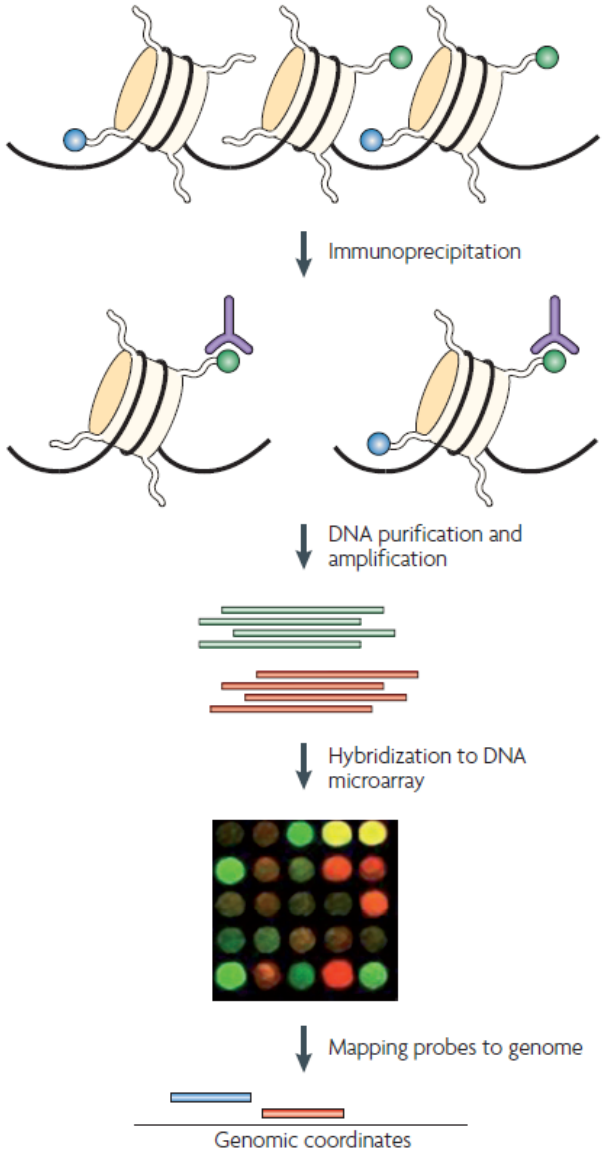

Figure 1.7: Chromatin immunoprecipitation combined with DNA microarrays (ChIP-chip). Modified chromatin is first purified by immunoprecipitating cross-linked chromatin using an antibody that is specific to a particular histone modification(shown in green). DNA is then amplified to obtain sufficient DNA. The colour labelled ChIP DNA, together with the control DNA prepared from input chromatin and labelled with a different colour, is hybridized to a DNA microarray. Source: http://www.nature.com/nrg/journal/v9/n3/full/nrg2270.html

level, is then mapped to the reference genome to obtain genomic coordinates. The first applications of ChIP-Seq to genome-wide profile of histone modifications were done in mouse embryonic stem (ES) cells [13, 14].

1.3.2 Genome-wide Profiles of Histone Modifications

Numerous genome-scale studies have provided data on the distribution of histone modifications [15]. As more histone modifications have been analyzed, it seems that regulatory regions such as enhancers and promoters have distinct histone modification patterns [14]. We will briefly introduce the types and patterns of histone modifications linked to regulatory elements, which are also demonstrated vividly by a dashboard of histone modifications (Fig. 1.9) in a recent review of histone modifications [16].

Studies in yeast show that histone H3 lysine 4 methylation (H3K4me) and histone acetylation correlate positively with transcription levels. They are highly enriched in promoter regions and extend significantly into the transcribed regions [17]. In addition, studies in higher eukaryotes also reveal that peaks of H3K4me3 associated with transcription start sites of many genes [18, 19]. Further research shows H3K4me3 broadly target to high CpG-content promoters [13]. Promoters with enriched H3K4me3 are also shown to be associated with other histone modifications such as histone acetylation [20].

Several of the histone marks are associated with gene silencing or repression. H3K27me3 is inversely correlated with gene activation and tends to spread over larger regions [21, 13]. H3K9me3 and H4K20me3 have been implicated transcriptional repression as well [22].

H3K4me1 enrichment at enhancers shows a chromatin signature of enhancer [23].

Research also shows that histone modifications at promoters are generally cell-type-invariant

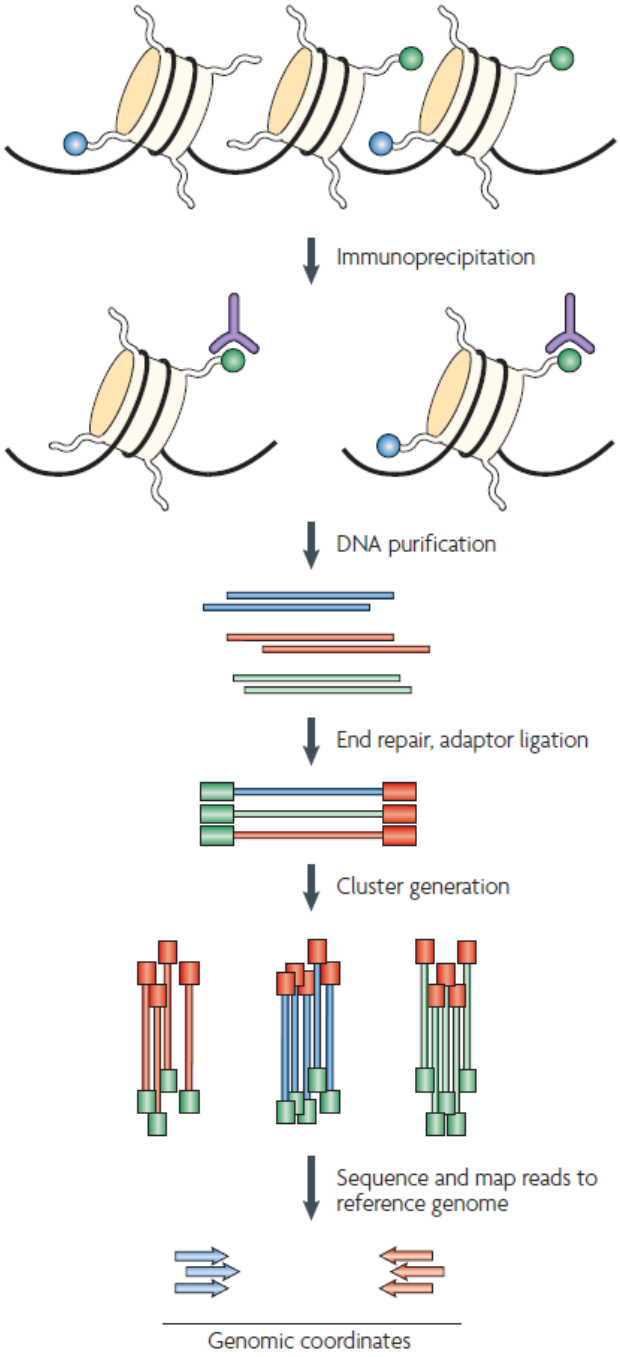

Figure 1.8: Chromatin immunoprecipitation combined with high-throughput sequencing techniques (ChIP-Seq). The first step is the purification of modified chromatin by immunoprecipitation using an antibody that is specific to a particular histone modification (shown in green). The ChIP DNA ends are repaired and ligated to a pair of adaptors, followed by limited PCR amplification. The DNA molecules are bound to the surface of a flow cell that contains covalently bound oligonucleotides that recognize the adaptor sequences. Clusters of individual DNA molecules are generated by solid-phase PCR and sequencing by synthesis is performed. The resulting sequence reads are mapped to a reference genome to obtain genomic coordinates that correspond to the immunoprecipitated fragments. Source: http://www.nature.com/nrg/journal/v9/n3/full/nrg2270.html

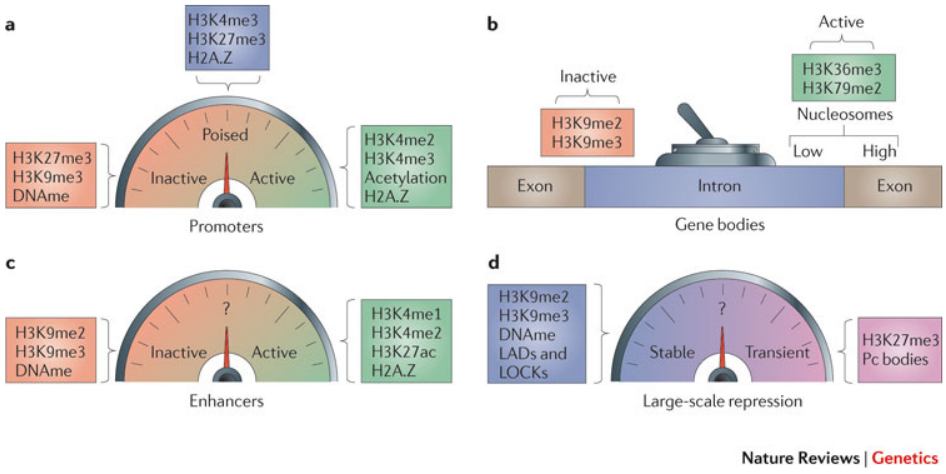

Figure 1.9: Dashboard of histone modifications for fine-tuning genomic elements. Source: http://www.nature.com/nrg/journal/vaop/ncurrent/full/nrg2905.html

whereas those at enhancers are cell-type-specific [24]. As shown in Fig. 1.10, K-means clustering was performed on the chromatin modifications from 414 promoters and similar patterns were observed across cell types. However, clustering on the enhancers revealed the cell-type-specificity of enhancers.

Histone modifications at enhancers are variable between cell types. Enhancers are probably of primary importance in driving cell-type-specific patterns of gene expression. Currently, our knowledge of what signals indicate active enhancers is far from completed. Some enhancers are known to have particular histone modification marks. The question of identification or prediction of histone modification patterns related to genome-wide and cell-type-specific enhancers is one of the main motivations of our research.

We will use Potts model clustering to analyze highly conserved regions that are candidates of enhancers and genome-wide histone modification data. We will like to discover the distinct histone modification patterns from the genome-wide data itself. Our special focus would be on the ultra conserved regions in genomes, rather than the genes (more about that later). However,

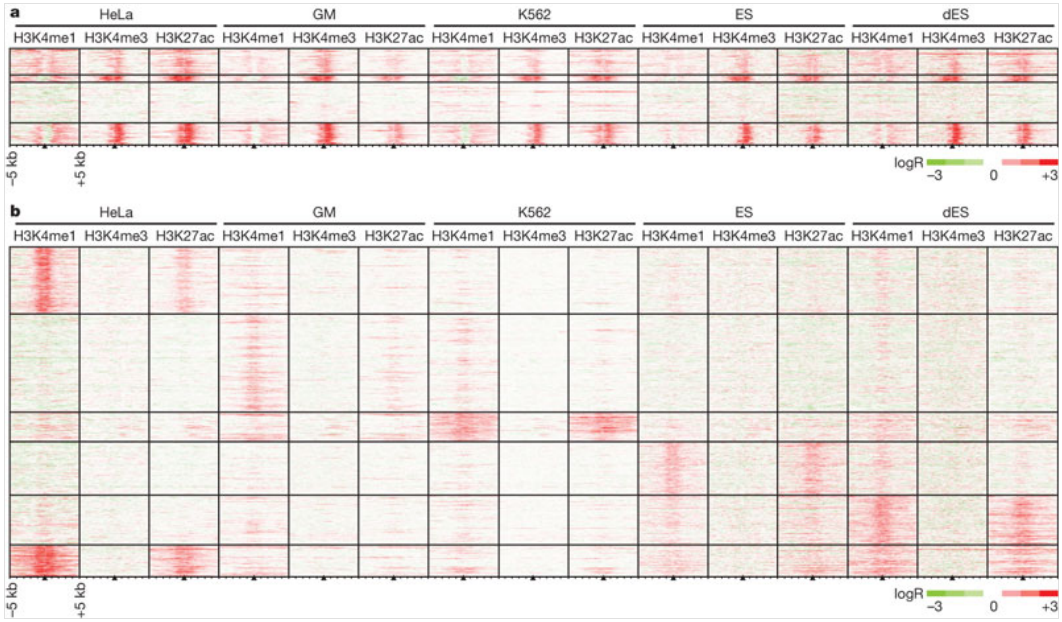

Figure 1.10: Histone modifications at promoters are cell-type-invariant while those at enhancers are cell-type-specific. Source: http://www.nature.com/nature/journal/v459/n7243/full/nature07829.html

for unsupervised discovery of potentially correlated histone mark signal, the natural tool is clustering. The next chapter introduces this important subject.

Chapter 2

Cluster Analysis

2.1 Overview of Cluster Analysis

Cluster analysis, or clustering, is the assignment of a set of observations into subsets which are called clusters. Cluster analysis groups data based only on the information found in the data that describes the data objects and their relationships. The final goal is to obtain clusters such that the observations in a cluster will be similar to each other but different from the objects in other clusters.

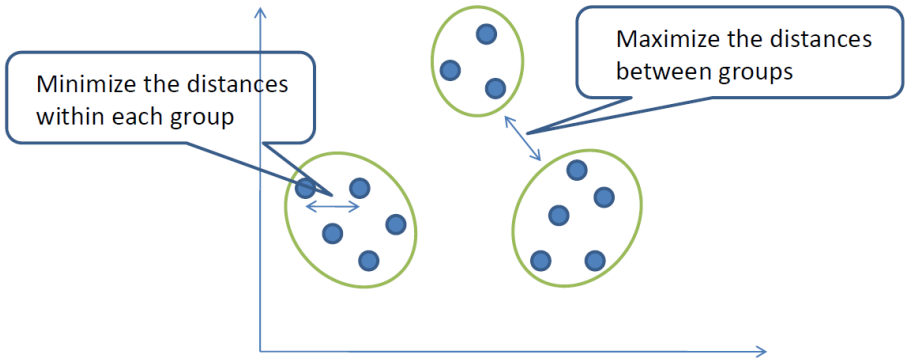

Figure 2.1: Cluster analysis. The greater the similarity within a group and the greater the difference between the groups the better the clustering.

2.2 Clustering in Computational Biology

Cluster analysis is widely applied in various areas in computational biology. Clustering by visualization has been used for many years for creating a taxonomy of living creatures: kingdom, phylum, class, order, family, genus and species. Originally, computational methods for constructing phylogeny trees were based on multiple morphological measurements. These tree construction methods are very similar to hierarchical clustering (see later). DNA sequence based phylogeny became popular as it became possible to sequence parts of an organism's genome [25].

Recently, with the rapid development of technology, a large amount of genome-wide expression data sets have been generated. Clustering has become one of the first steps in gene expression analysis, providing us with distinct classes of co-regulated genes, namely, a group of genes that behave similarly under a multitude of conditions the cell is subjected to [26]. In general, as data becomes very high dimensional, pattern discovery by visual inspection often becomes nearly impossible. Therefore, some unsupervised pattern discovery tools like clustering become essential for data exploration.

2.2.1 Measuring Similarity

The first step in clustering is to define the similarity. Similarity is often measured by distance between genes. Two most popular similarity measures as follows:

- Euclidean distance:

$$d_{xy} = \sqrt{\sum (x - y)^2} \ .$$

- Pearson correlation coefficient:

$$d_{xy} = 1 - r_{xy} \ ,$$

where

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} .$$

There are other similarity measures and variants. Interested readers are referred to some review papers on cluster analysis of gene expression (e.g. [27]). In our work, we will introduce a different measure of distance, based on differences in probability distributions.

2.2.2 Clustering Methods

There are two major types of clustering: Partitional clustering and Hierarchical clustering.

Partitional Clustering

Partitional clustering divides the data objects into non-overlapping clusters such that each data object is in exactly one subset. The K-means algorithm [28] is a widely used partitional clustering method. Its algorithm is very simple and direct. It attempts to minimize the sum of the squared distances of objects from the centroids of the clusters. K, the number of clusters, must be given in the very beginning. Each cluster is associated with a center point and each point is assigned to the cluster with the closest centroid.

The detailed algorithm is described as follows:

1. Select the number of clusters K and choose K points as the initial centroids.

2. Form K clusters by assigning points to its nearest centroid.

3. Calculate the new centroid of each cluster.

4. Repeat Step 2 and Step 3 until centroids remain the same.

K-means clustering is widely used but it has its disadvantages. The number of clusters, K, must be specified before clustering. However, it is hard to determine the right K value without knowing anything about the data structure. Apart from that, K-means has problems when clusters have different sizes, densities or shapes. In part (a) of Fig. 2.2, the data set contains four clusters of different sizes, shapes and numbers. However, in part (b) of Fig. 2.2 K-means (with K = 4) partitions the space into four subspaces, which are not as same as those clusters in (a) [27].

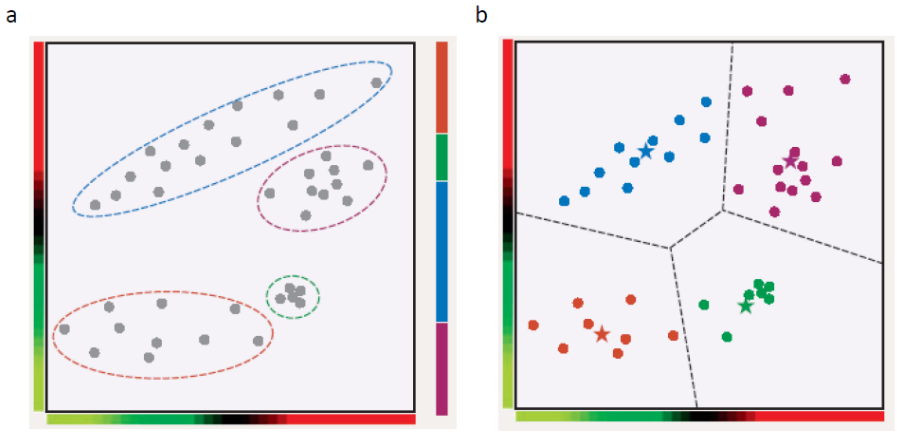

Figure 2.2: K-means clustering. (a) The data set contains four clusters of different sizes, shapes and numbers. (b) K-means (with K = 4) partitions the space into four subspaces, which are different from those in (a). Source: http://www.nature.com/nbt/journal/v23/n12/full/nbt1205-1499.html

Hierarchical Clustering

Hierarchical clustering creates a hierarchy of clusters. These are represented in a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations.

Algorithms for hierarchical clustering are generally either agglomerative, in which one

starts at the leaves and successively merges clusters together; or divisive, in which one starts at the root and recursively splits the clusters. The red numbers in part (a) of Fig. 2.3 present the order of the merges. Meanwhile, part (b) of Fig. 2.3 demonstrates how the hierarchical tree is built.

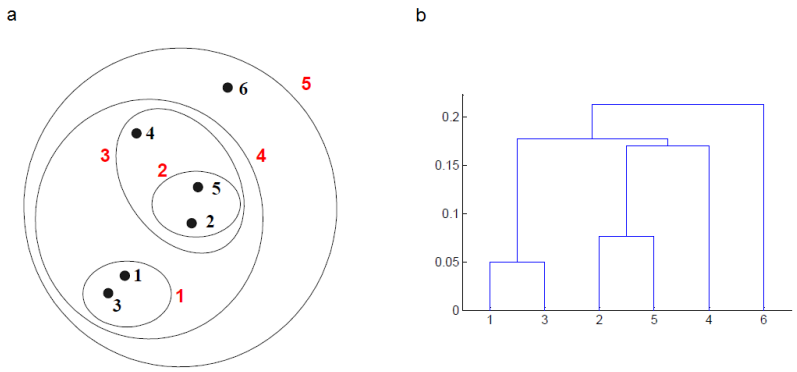

Figure 2.3: Hierarchical clustering. This method produces a set of nested clusters organized as a hierarchical tree. (a) The red numbers present the order of the merges. (b) Demonstration of how the hierarchical tree is built.

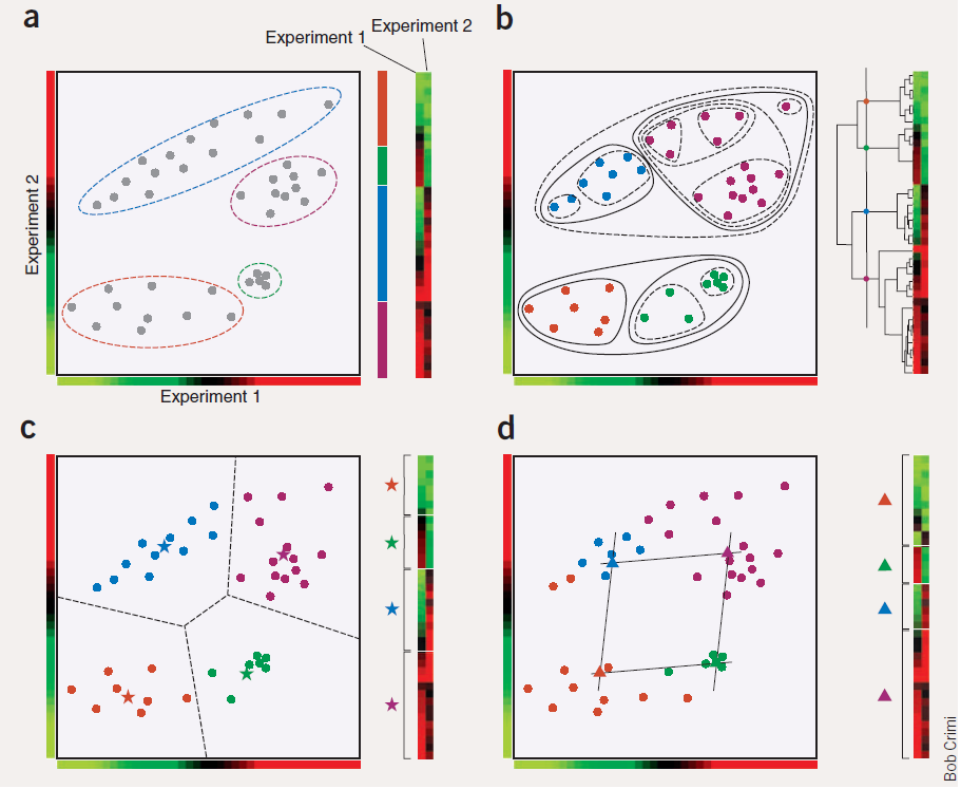

Figure 2.4: A simple clustering example with genes measured under two different conditions. (a) The data set contains four clusters of different sizes, shapes and numbers of genes. Left: each dot represents a gene, plotted against its expression value under the two experimental conditions. Euclidean distance, which corresponds to the straight-line distance between points in this graph, was used for clustering. Right: the standard red-green representation of the data and corresponding cluster identities. (b) Hierarchical clustering finds an entire hierarchy of clusters. The tree was cut at the level indicated to yield four clusters. Some of the superclusters and subclusters are illustrated on the left. (c) k-means (with k = 4) partitions the space into four subspaces, depending on which of the four cluster centroids (stars) is closest. (d) SOM finds clusters, which are organized into a grid structure. Source: http://www.nature.com/nbt/journal/v23/n12/full/nbt1205-1499.html

Many cluster analysis methods are applied to gene expression data. There are more than one hundred published clustering algorithms, dozens of which have been used in gene expression data [27]. Each algorithm has its own biases when it constructs the clusters. In Fig. 2.4, Patrik D'haeseleer gave us a very good example that clustering algorithms can lead to different results

on data of just 40 genes [27], which has much simpler structure comparing with the genome-wide data.

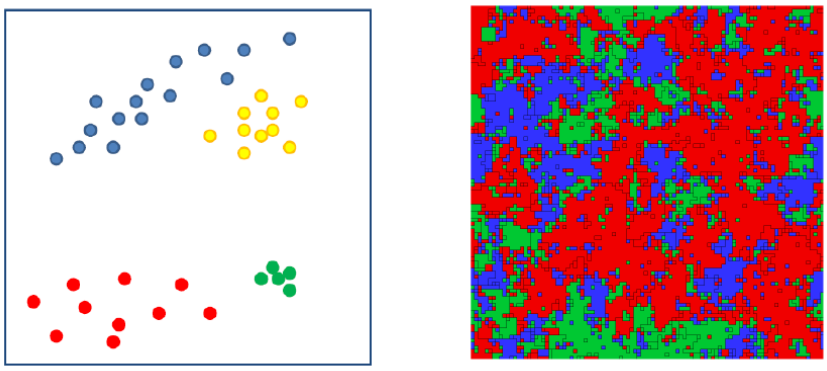

Figure 2.5: Similarity between clusters and Potts ferromagnet. The clustering result looks similar to an intermediate state of inhomogeneous Potts model at a certain temperature.

As we mentioned before, K-means clustering, the most common method of partitional clustering, has problems when the individual clusters have shapes that differ seriously from sphere. A preliminary examination of our data suggests there may be clustered shaped like long cigars in our data set. For dealing with such clusters, there is a plethora of alternative clustering methods. Some of these are Self-organized Map [29], graph partitioning [30] as well as Markov Random Field (alternatively called Gibbs Random Field) based methods [31]. Additionally, in the context of discovering pattern for histone marks, it may not always be meaningful to compute the distance between every pair. For example, for the mark patterns related two genes of very different lengths, defining distance would be a problem. However, we might want gene A and gene B in the same cluster if there is a chain of intermediate genes with defined separation connecting A and B indirectly. Clustering based on some of these alternative methods, including Markov Random Fields, can achieve this goal.

Markov Random Field (MRF) [32] was originally used for analyzing problems related to

statistical mechanics. It is defined as a graphical model in which a set of random variables have a Markov property described by an undirected graph, with the distribution of a variable living on a vertex is conditional only on the variables living on the neighboring vertices of the graph. The natural MRF for clustering problem is related to the Potts model. Fig. 2.5 demonstrates that the clustering result looks similar to an intermediate state of inhomogeneous Potts model at a certain temperature. Therefore, the data points of clustering problem can be looked as the sites of an inhomogeneous Potts ferromagnet [33]. In this approach, solving clustering problem is converted to calculating the thermal average of the spin-spin correlation between every pairs of spins at a certain temperature T of the system. We believe that several of the alternatives to K-means clustering, mentioned in the previous paragraph, produce similar answers, but as physicists, we find thinking of clustering in terms of Potts model a natural choice. Chapter 3 describes both Potts model and its application to clustering.

Chapter 3

Clustering Based on Statistical Physics: Potts Model Clustering

3.1 Potts Model

3.1.1 Introduction

The Potts model is named after Renfrey Potts [34], who described the model as his thesis topic, which was suggested by his advisor Cyril Domb. The Potts model is a model of interacting spins on a crystalline lattice. At each site of the lattice, the spin variable s takes one of the values $1, 2, 3, ..., q$. If $q \geq 2$, this is called the Potts model. The Hamiltonian of a Potts model with N sites and a spin configuration $(s_1, s_2, ..., s_N)$ is given by:

$$H(S) = \sum_{\langle i,j \rangle} J_{ij}(1 - \delta_{s_i,s_j}) \ , \tag{3.1}$$

where the sum is running over all neighboring sites denoted as $\langle i, j \rangle$, J_{ij} is the interaction between a pair of spins associated with the site i and j, and δ_{s_i,s_j} is 1 if $s_i = s_j$, otherwise 0 (This excludes the contribution from the site itself). The distance between the site i and j is denoted as d_{ij}. If the interaction J_{ij} is a monotonical decreasing function of d_{ij}, the spin s_i and s_j are inclined to have the same value as d_{ij} becomes smaller.

Once Hamiltonian and temperature T are given, the thermodynamic average of a generic physical quantity Q can be calculated as,

$$\langle Q \rangle = \sum_{S} Q(S)P(S) \ , \tag{3.2}$$

where $P(S)$ is the Boltzmann factor given by:

$$P(S) = \frac{1}{Z} \exp(-\frac{H(S)}{T}) \ . \tag{3.3}$$

Here $Z = \sum_S \exp(-\frac{H(S)}{T})$ is the partition function. According to statistical mechanics, $P(S)$ is the probability that the system in equilibrium is found in configuration S. Moreover, the logarithmic derivative of Z with respect to temperature is related to the free energy of the system. Singularities in the derivative of Z generally corresponds to phase transitions. The temperature at which there is a phase transition is called the critical temperature.

3.1.2 Spin-spin Correlation Function in Thermodynamic Phases

The spin-spin correlation function C_{ij} is the thermal average of δ_{s_i,s_j}:

$$C_{ij} = \left\langle \delta_{s_i,s_j} \right\rangle \ . \tag{3.4}$$

C_{ij} can be calculated by Eq. (3.2). It represents the probability that spin variables s_i and s_j have the same values.

There are two phases in a homogeneous system where J_{ij} is constant. At high temperatures, the system is in the paramagnetic phase and the spins are in disorder. $C_{ij} \approx \frac{1}{q}$ for sites i and j and q is the number of possible spin variables. At low temperatures, the system turns into the ferromagnetic phase and all the spins align to the same direction. $C_{ij} \approx 1$ for sites i and j.

If the system is inhomogeneous, the phases are not just ferromagnetic or paramagnetic. We will discuss inhomogeneous system along with magnetization and susceptibility.

3.1.3 Magnetization and Susceptibility in Thermodynamic Phases

The magnetization, $m\,(S)$, is the order parameter of the system. It is defined as

$$m\,(S) = \frac{qN_{max}(S) - N}{(q-1)\,N}$$

with

$$N_{max}(S) = max\left\{N_1(S), N_2(S), ..., N_q(S)\right\},$$

where $N_\mu(S)$ is the number of spins with the value μ.

The susceptibility χ, which is related to the variance of the magnetization, is defined as

$$\chi = \frac{N}{T}\left(\langle m^2\rangle - \langle m\rangle^2\right).$$

The susceptibility also reflects different thermodynamic phases of the system. In a inhomogeneous system, some spins form magnetic grains where there are strong couplings between spins in the same grain. At high temperatures, all the couplings are weak and the system is in the paramagnetic phase. As the temperature is decreased, the grains act by producing large fluctuations in the magnetization because of relatively strong couplings within each grains. This is an intermediate (pseudo-) phase called superparamagnetic phase. A peak of χ can be observed at superparamagnetic phase [33]. When the temperature is decreased further, the system turns into ferromagnetic and all the spins tend to align to the same direction. Since the magnetization is close to 1 at low temperature, χ is very small. We use the peak in the profile of susceptibility vs. T to identify the superparamagnetic phase.

3.2 Simulations of Potts Model

Simulation of a Boltzmann distribution is often done via a Markov chain, as a stochastic process happening in the space of configuration of the model. This Markov process is arranged in such a way that the stationary state happens to be the target Boltzmann distribution. More precisely,

$$P(S) = \sum_{S'} M(S, S')P\left(S'\right),$$

where $M(S, S')$ is the transition probability from configuration S' to configuration S and the sum is running over all possible configurations.

There are many ways to generate a Markov chain. Here we introduce the Metropolis algorithm and the Swendsen-Wang algorithm.

3.2.1 The Metropolis Algorithm

Detailed balance shows that configurations of equilibrium system have probability proportional to the Boltzmann factor. The Metropolis algorithm [35] based on detailed balance is used to sample the space of possible configurations in a thermal way. Consider two configurations of the system $S1$ and $S2$, they occur with the probabilities which are proportional to the Boltzmann factor:

$$\frac{P(S1)}{P(S2)} = \frac{\exp(-\frac{E_{S1}}{T})}{\exp(-\frac{E_{S2}}{T})} = \exp(-\frac{E_{S1} - E_{S2}}{T}) \ . \tag{3.5}$$

We can use Eq. (3.5) to generate a Markov chain where S1 is the initial configuration of the system. Simulation is described as follows:

1. Starting from a configuration $S1$ with energy E_{S1}, make a change in the configuration to obtain a new configuration $S2$.

2. The change of energy ΔE is $E_{S2} - E_{S1}$.

3. If $\Delta E < 0$, accept $S2$, since it has lower energy.

4. If $\Delta E > 0$, accept $S2$ with probability $p = \exp(-\frac{E_{S1}-E_{S2}}{T})$. Here a random number R is generated between 0 and 1. Only when $p = \exp(-\frac{\Delta E}{T}) > R$, $S2$ is accepted.

Usual implementation of the Metropolis algorithm is widely used. However, this single spin alteration process is not suitable for the Potts model especially in superparamagnetic phase. The local moves employed by Metropolis algorithm will take very long time to explore the configuration space. The typical spin value in a large correlated domain tend not to change for a long time. To do better, one would need to modify spins in a large domain together. One clever way to do this is the Swendsen-Wang method.

3.2.2 The Swendsen-Wang Algorithm

The Swendsen-Wang Algorithm

The other way to generate the Markov chain is the Swendsen-Wang (SW) algorithm [36] , which works very well in the superparamagnetic phase. It overturns an aligned cluster instead of one single spin in one Monte Carlo step. One way to approach the Swendsen-Wang algorithm is to rewrite the partition function in terms of spin variables along with an additional bond occupation variable [37]. The resultant procedure alternates between generating new spin configurations and then generating new bond occupation configurations.

The detailed Swendsen-Wang algorithm is described as follows:

1. Generate initial configuration of system $S1 = (s_1, s_2, ..., s_N)$ randomly.

2. Generate the next configuration of system $S2$ based on $S1$:

   (a) Visit all pairs of spins $\langle i, j\rangle$ which have interaction $J_{ij} > 0$. Spin i and spin j are frozen together with probability

   $$p^f_{i,j} = 1 - \exp(-\frac{J_{ij}}{T}\delta_{s_i,s_j}) \ . \tag{3.6}$$

   If i and j have the same spin variable values, $s_i = s_j$, the probability that they are frozen is

   $$p^f = 1 - \exp(-\frac{J_{ij}}{T}).$$

   If $s_i \neq s_j$, there is no chance that i and j can be frozen together. Calculate all pairs of spins and put a frozen bond between any frozen spin pairs.

   (b) We define SW cluster as the cluster contains all spins that have a path of frozen bonds connecting all of them. Since spins are frozen only if they have the same spin

variables, we just need to identify the Swendsen-Wang clusters from the same-color (same-spin variable) sites. We use the Hoshen-Kopelman algorithm to connect frozen spin pairs into a path of frozen bonds. The Hoshen-Kopelman algorithm is described below.

(c) For each SW cluster, we draw random number from $1, 2, ..., q$ and assign this number to the values of all spins of this cluster. After going through all SW clusters, the new configuration $S2$ is generated.

3. Iterate Step 2 for M times. At each Monte Carlo step the physical quantity is known. And Eq. (3.2) can be easily calculated by:

$$\langle Q \rangle \approx \frac{1}{M} \sum_{i}^{M} Q(S_i) \ . \tag{3.7}$$

The Swendsen-Wang algorithm is widely used in lots of applications and it has been tested in various Potts models [38].

The Hoshen-Kopelman Algorithm

The Hoshen-Kopelman (HK) algorithm [39] is an algorithm to label clusters on a grid, where the cells may be either occupied or unoccupied. The HK algorithm is used to identify clusters of contiguous cells.

The general idea is to scan the whole grid and mark all occupied cells. To each occupied cell we assign a label corresponding to the cluster to which the cell belongs. If the cell has zero occupied neighbors, then we assign to it a new cluster label. If the cell has one occupied neighbor, then we assign to the current cell the same label as the occupied neighbor. If the cell has more than one occupied neighbors, we choose the lowest-numbered cluster label of the

occupied neighbors to use as the label for the current cell and change all these neighbors' labels to be the lowest-numbered label.

We use the Union-Find algorithm to implement the HK algorithm. The function Union (i, j) makes i and j members of the same cluster. Because equivalence relations are transitive, all items equivalent to i are equivalent to all items equivalent to j. The function Find (i) finds representative members of the equivalence class to which i belongs.

The HK algorithm is described in terms of union and find operations as follows:

1. Initialize the labels of N points, written as a vector L with integer components:

$$L = (1, 2, ..., N).$$

2. Check the signs of left and above points and decide whether current label is a connected to its left or above points.

   - If there is no connection, give the point new label.
   - If there is a connection, use Find function to find the labels of the above and left points and use the Union function to unite all of them.

3. Correct labels so that every point is labeled by the representative label of its same cluster.

Since HK algorithm employed in our method is in non-lattice environment, we use the extension of Hoshen-Kopelman (EHK) algorithm to non-lattice environment [40]. Please refer to [40] where implementation of EHK for non-lattice is described in detail.

3.3 Clustering Data Based on Potts Model

Marcelo Blatt, Shai Wiseman and Eytan Domany introduced a new clustering approach based on Potts model [33, 41]. The general idea of this clustering method is that the data points of

clustering problem are looked as the sites of an inhomogeneous Potts ferromagnet. Solving clustering problem involves calculating the thermal average of the spin-spin correlation between every pairs of spins at a certain temperature T of the system. There are three main steps as follows:

1. Suppose there are N data points to be clustered and each data point is x_i with L dimensions. These data points are used to build analogous physical Potts model. The Hamiltonian and interaction J_{ij} are calculated.

2. Decrease the temperature of the system slowly. At each temperature, measure the spin-spin correlation C_{ij}.

3. Observe the change of the clusters and choose temperature we prefer. At the chosen temperature, use C_{ij} to get clusters.

In this section, we describe these three steps in details.

3.3.1 Building Potts Model Related to Data

The Hamiltonian

To go ahead, we need the Hamiltonian, which is the most important quantity in the system. According to Eq. (3.1), we have to define J_{ij}. So that the Hamiltonian can be evaluated for every S. There are so many ways to define J_{ij}. Here, we need strong interactions between spins that correspond to data from a high-density region and weak interactions between neighbors that are in low-density regions. So the interaction function J_{ij} is defined as:

$$J_{ij} = \begin{cases} \frac{1}{K}\exp(-\frac{d_{ij}^2}{2a^2}), & \text{if } i \text{ and } j \text{ are neighbors} \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

where K is average number of numbers and a is local length scale which is the average distance of K nearest neighbors. We will discuss K and a in the next part. Please remember that this K, is not the number of clusters we finally get (in K-means approach, K refers to the target number of clusters).

Neighboring Points and Related Parameters

It is very important to define the right neighbors because only when i and j are neighbors the interaction between i and j has the nonzero value. Here is how we define the neighboring points: when i is one of the K-nearest neighbors if j and j is also one of the K-nearest neighbors, i and j are neighbor and K is their mutual neighborhood value. K is chosen to be the value that makes all data points to be one connected community.

3.3.2 Calculation of Physical Quantity

At each fixed temperature T, we always perform the following steps to calculate the spin-spin correlation for each pair of spins:

1. Choose the iteration value M.

2. Generate initial configuration of system $S1 = (s_1, s_2, ..., s_N)$ randomly.

3. Figure out all neighboring points and test whether they are connected by a frozen bond by p_{ij}^f in Eq. (3.6).

4. Use Extension of Hoshen-Kopelman algorithm [40] to calculate the SW clusters.

5. For each SW clusters in $S1$, assign a new spin variable to all spins of the SW cluster. New configuration $S2$ is then generated.

6. Repeat iteration M times, record all configurations $S_1, S_2, ..., S_M$.

7. Use Eq. (3.7) to calculate the thermal average of physical quantity such as spin-spin correlation at each T.

The spin-spin correlation function C_{ij} is defined in Eq. (3.4). A two-point connectedness function R_{ij} to used to calculate the C_{ij} [42]. R_{ij} is defined as the probability that i and j are in the same SW cluster:

$$R_{ij} = \frac{1}{M} \sum_{S} r_{ij} \tag{3.9}$$

where

$$r_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are in the same SW cluster} \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

R_{ij} can be calculated from every configurations of system and C_{ij} is:

$$C_{ij} = \frac{(q-1)R_{ij} + 1}{q} \tag{3.11}$$

3.3.3 Clustering the Data

We can do the Potts model clustering by using C_{ij} values at temperature T. If $C_{ij} > 0.5$, i and j are connected. Also link i to its neighbor j which has the largest C_{ij} value among all i's neighbors. Then the Hoshen-Kopelman algorithm is used to connect the indirectly connected points. All the connected points belong to the same cluster.

Chapter 4

Potts Model Clustering of Histone Modification Data

About a decade ago, Jenuwein and Allis [43] proposed that the combinatorial nature of histone amino-terminal modifications reveals a "histone code" that extends the information potential of the genetic code, by controlling access to DNA for various biological processes. They seem to suggest that each these 'codes' involve several different marks, with the pattern of combinations distinguishing between different codes. Discovering such 'codes' should be an obvious application of clustering.

Recent research indicate that enhancers are more variable than other classes of transcriptional regulatory element between cell types [24]. Enhancer is also considered probably the primary important factor for driving cell-type-specific patterns of gene expression on a global scale. We use Potts model clustering to analyze genome-wide data of histone modifications and study if some of the patterns/codes discovered strongly correlate to enhancer activity.

4.1 Genome-wide Histone Modification Data

Our histone modification data is genome-wide ChIP-Seq data from the Broad Institute(http://www.broadinstitute.org/scientific-community/science/programs/epigenomics/chip-seq-data).

The primary datasets for published Broad Institute ChIP-Seq experiments were deposited in the NCBI GEO database. The data sets that we use are from three GSE accessions: GSE12241 [13], GSE11074 [44] and GSE11172 [45]. We choose different histone modifications in pluripotent and lineage-committed cell. They are H3K4me1, H3K4me3, H3K9me3, H3K27me3,

H3K36me3, H4K20me3 in murine embryonic stem cells, neural progenitor cells and embryonic fibroblasts.

4.2 Enhancer and VISTA Enhancer Browser

The VISTA Enhancer Browser is our other data source(http://enhancer.lbl.gov/). It is a central resource for experimental validated human non-coding sequences with enhancer activity as tested in transgenic mice. In the following, we briefly describe how the enhancers listed in VISTA Enhancer Browser were tested.

1. Preparation of enhancer candidates. The candidates of enhancer sequences are prepared by comparative analysis for conserved regions between the human genome and a wide range of available species (mouse, rat, chicken, frog, fugu, tetraodon and zebrafish). Because of the goal to identify gene enhancer sequences, these comparative alignments were filtered for overlap with exons of known genes and mRNAs. In the end, suitable conserved noncoding sequences were the candidates of enhancer.

2. Test of enhancer activity. Fig. 4.1 show the flow-chart to couple comparative genomic conservation to a mouse transgenic enhancer screen.

   Candidate human gene enhancers, selected based on ultraconservation, were amplified and cloned into a reporter vector containing the lacZ gene. This reporter vector was microinjected into fertilized mouse eggs. As the embryo developed, if the enhancer becomes activated, the reporter would produce the lacZ gene which could later be assayed with a particular stain. Embryos were harvested at embryonic day 11.5, and lacZ stained.

   Each positive developmental enhancer was tested based on the observed pattern of expression as shown in last step of Fig. 4.1in day 18. These annotations were done by

multiple curators in a group setting. If an element was observed reproducible expression in the same structure in at least three independent transgenic embryos, it was defined as a positive enhancer. For each structure the reproducibility of the observed pattern was provided in the database. If elements have been obtained at least five transgenic embryos, but there is no reproducible expression in any structure was observed in at least three different embryos, these elements were defined as negative.

4.3 Computational Details

4.3.1 Read Density Calculation

We use the aligned sequence data where reads (length 25 bp) from each IP experiment have been aligned to the mouse reference genome (mm8). We calculate the read density by counting the numbers of mapped reads overlapping with a 300 bp window centered at that position. We sample the read density at every 25 bp.

4.3.2 The Data Matrix

We obtain the coordinates of candidate enhancer region from the VISTA enhancer database and compare the coordinates with the read density data. Although the read density data are genome-wide, there are still some masked regions. As a result, we do not have the histone information from these regions. So, if the coordinates of any conserved regions overlap with these masked regions, we do not include these conserved regions. After this filtering step, we have totally 860 conserved sequences.

For each conserved sequences, we calculate the density reading every 100 bp from the 25bp resolution data from 25 features. The detailed description of the features is listed in Table. 4.1. Our data is composed of 860 elements ($S_1, S_2, ..., S_{860}$). Each element S_i has 25 features and

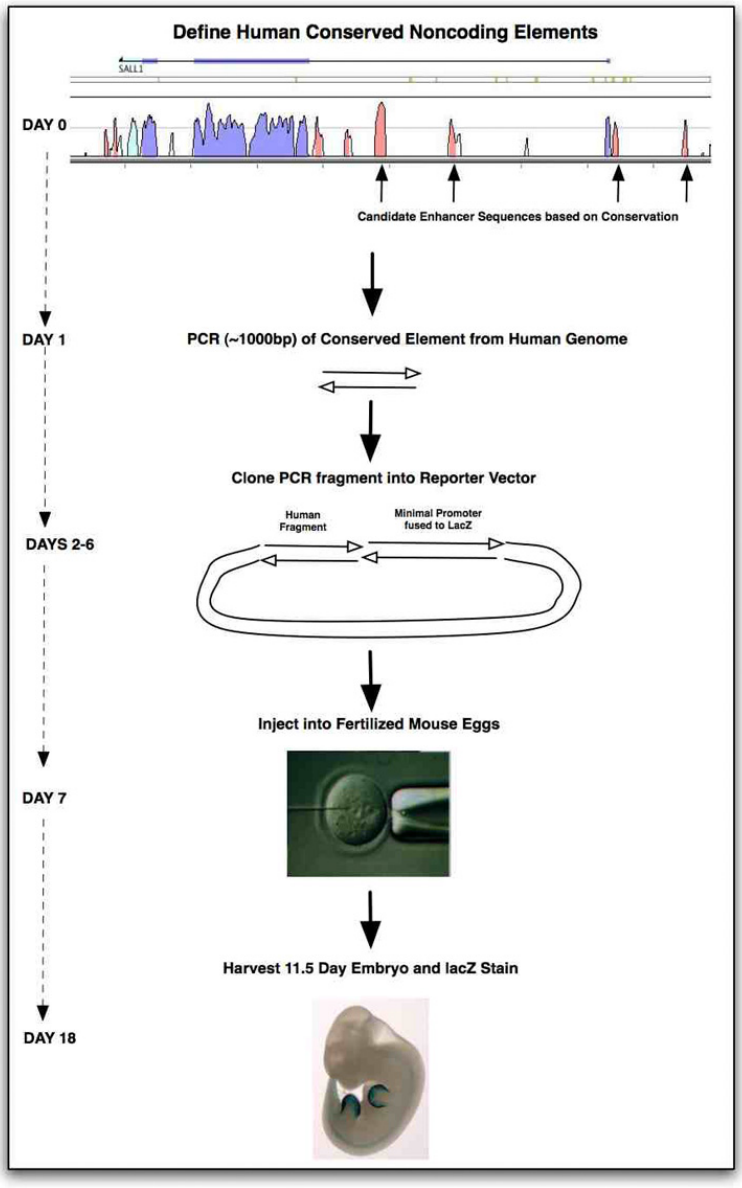

Figure 4.1: Test of the enhancer activity. The conserved noncoding elements can be PCR amplified, cloned, microinjected, and assayed for enhancer activity. Source: http://enhancer.lbl.gov

each feature is a vector with the length which is the length of the elements in the unit of 100 bp.

Table 4.1: Description of histone modification features

Feature	Type of modification	Histone	Type of cell
ES.H3K4me3	tri-methylation	H3 Lysine 4	embryonic stem cells(V6.5)
ES.H3K4me1	mono-methylation	H3 Lysine 4	embryonic stem cells(V6.5)
ES.H3K4me2	di-methylation	H3 Lysine 4	embryonic stem cells (V6.5)
ES.H3K27me3	tri-methylation	H3 Lysine 27	embryonic stem cells (V6.5)
ES.H3K9me3	tri-methylation	H3 Lysine 9	embryonic stem cells (V6.5)
ES.H4K20me3	tri-methylation	H3 Lysine 20	embryonic stem cells (V6.5)
ES.H3K36me3	tri-methylation	H3 Lysine 36	embryonic stem cells (V6.5)
ES.RPol2		RNA polymerase II	embryonic stem cells (V6.5)
ES.H3		H3	embryonic stem cells (V6.5)
ES.WCE		Whole cell extract	embryonic stem cells (V6.5)
ESHyb.H3K4me3	tri-methylation	H3 Lysine 4	embryonic stem cells (hybrid)
ESHyb.H3K36me3	tri-methylation	H3 Lysine 36	embryonic stem cells (hybrid)
ESHyb.H3K9me3	tri-methylation	H3 Lysine 9	embryonic stem cells (hybrid)
MEF.H3K4me3	tri-methylation	H3 Lysine 4	Embryonic fibroblasts
MEF.H3K27me3	tri-methylation	H3 Lysine 27	Embryonic fibroblasts
MEF.H3K36me3	tri-methylation	H3 Lysine 36	Embryonic fibroblasts
MEF.H3K9me3	tri-methylation	H3 Lysine 9	Embryonic fibroblasts
MEF.WCE		Whole cell extract	Embryonic fibroblasts
NP.H3K4me3	tri-methylation	H3 Lysine 4	ES-derived neural precursor cells
NP.H3K4me1	mono-methylation	H3 Lysine 4	ES-derived neural precursor cells
NP.H3K4me2	di-methylation	H3 Lysine 4	ES-derived neural precursor cells
NP.H3K27me3	tri-methylation	H3 Lysine 27	ES-derived neural precursor cells
NP.H3K36me3	tri-methylation	H3 Lysine 36	ES-derived neural precursor cells
NP.H3K9me3	tri-methylation	H3 Lysine 9	ES-derived neural precursor cells
NP.WCE		Whole cell extract	ES-derived neural precursor cells

4.3.3 Defining the Similarity

Suppose two conserved sequences are A and B. We define the distance between A and B to be:

$$D = \sum_{i=1}^{n} d_i \ , \tag{4.1}$$

where d_i is the distance between each features and here $n = 25$.

For calculation of d_i, we suggest two definitions which are described as follows:

1. The first definition that we use is:

$$d_i = \begin{cases} \ln \frac{\left(\frac{a_i+b_i}{2}\right)!\left(\frac{a_i+b_i}{2}\right)!}{a_i!b_i!}, & \text{if } a_i + b_i \text{ is even number} \\ \ln \frac{\left(\frac{a_i+b_i+1}{2}\right)!\left(\frac{a_i+b_i-1}{2}\right)!}{a_i!b_i!}, & \text{if } a_i + b_i \text{ is odd number} \end{cases} \tag{4.2}$$

   where a_i and b_i are the means of the i th feature values of A and B. This definition is very straightforward. In the next section, we present the results using this definition.

2. An updated definition is considered as:

$$d_i = \frac{P_{max}}{P_{a_i}} \tag{4.3}$$

   where the function $P(a)$ is:

$$P(a) = \left(\frac{a_i + b_i}{a}\right)\left(\frac{L_1}{L_1 + L_2}\right)^a \left(1 - \frac{L_1}{L_1 + L_2}\right)^{a_i+b_i-a} \tag{4.4}$$

   a_i and b_i are the total readings of the i th feature of A and B. L_1 and L_2 are the lengths of the i th feature of A and B in unit of 100bp. And P_{max} is the maximum value of Eq. (4.4).

4.3.4 Applying Potts Clustering

The number of Potts model states q does not directly imply any assumption about the number of clusters present in the data. Although the value of q determines the sharpness of the transitions and also affects the iteration of simulations to the equilibrium state. Our simulation shows the influence of q on the clustering result is very weak. We set the value of q to be 10 in our

calculation. For each domain, q affects the entropy logarithmically, thereby we expect to see major differences only if q changes by an order of magnitude.

The average mutual neighborhood value K was chosen to be 25 since it allowed all the vertices to be connected. The local length scale a is 8, in Eq. (3.8), which is the average of all distances d_{ij} between neighboring i and j.

4.4 Results

4.4.1 Choice of Temperature and the Number of Clusters

By applying our procedure, we obtain the size of the clusters as a function of the temperature as presented in Fig. 4.2. This allows us to observe the change of the clusters as the temperature T varies. We notice that the system keeps its largest cluster, which size is around 830, at low temperature. The largest cluster starts breaking into smaller clusters at $T = 0.09$. After $T > 0.12$, all the clusters of the system are very small and their sizes are smaller than 50.

An obvious peak is also observed at $T = 0.1$ in the profile of susceptibility vs. temperature in Fig. 4.2. We decide to choose $T = 0.1$ to do more analysis. At $T = 0.1$, we obtain 5 major clusters. Table. 4.2 lists the size of each cluster. The centers of clusters in each feature are calculated and listed in Table. 4.3.

Table 4.2: Size of the clusters at T = 0.1

Cluster	Clustersize
c1	528
c2	147
c3	63
c4	21
c5	20

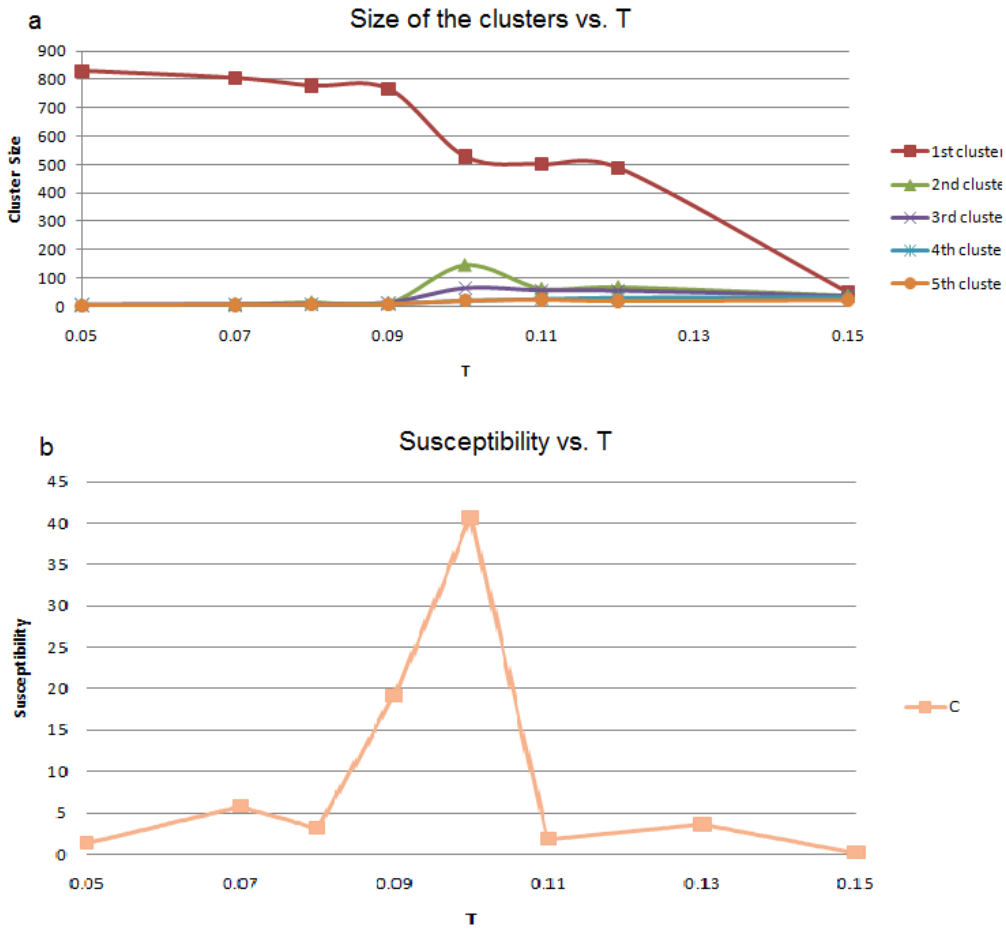

Figure 4.2: Size of clusters vs. temperature and Susceptibility vs. temperature. (a) Size of the five biggest clusters obtained at each temperature. (b) Susceptibility vs. temperature. An obvious peak is observed when temperature is 0.1.

Table 4.3: Centers of the clusters

Cluster	ES.H3K4me3	ES.H3K4me1	ES.H3K4me2	ES.H3K27me3	ES.H3K9me3
c1	2.216	1.907	1.660	1.454	0.804
c2	2.081	1.640	5.000	1.454	0.942
c3	1.338	2.121	2.553	1.825	0.844
c4	6.332	3.354	1.648	2.223	1.010
c5	2.776	4.838	0.917	1.406	0.915
	ES.H4K20me3	ES.H3K36me3	ES.WCE	ES.RPol2	ES.H3
c1	0.213	1.114	3.701	3.221	1.326
c2	0.250	1.740	2.281	3.043	1.249
c3	0.222	10.944	3.606	2.211	1.413
c4	0.370	3.138	4.496	5.957	1.448
c5	0.288	1.834	14.690	3.136	1.008
	MEF.H3K27me3	MEF.H3K36me3	MEF.H3K9me3	MEF.H3K4me3	MEF.WCE
c1	2.456	86.649	8.615	5.075	6.552
c2	3.080	2.418	8.263	57.201	81.472
c3	2.730	167.384	0.354	2.722	5.300
c4	3.305	6.135	57.638	3.989	5.770
c5	3.664	2.221	3.790	3.636	3.392
	ESHyb.H3K4me3	ESHyb.H3K36me3	ESHyb.H3K9me3	NP.H3K4me2	NP.H3K4me1
c1	1.140	1.760	2.278	1.028	1.944
c2	1.588	21.123	2.269	1.244	1.756
c3	1.251	2.429	2.150	0.953	2.488
c4	2.328	93.457	3.496	1.437	1.739
c5	1.732	15.949	4.280	1.188	1.151
	NP.H3K4me3	NP.H3K27me3	NP.H3K36me3	NP.H3K9me3	NP.WCE
c1	148.582	5.894	2.543	9.564	1.931
c2	9.975	100.923	3.036	4.744	1.974
c3	10.484	8.354	3.472	9.342	2.218
c4	13.502	6.938	2.356	9.883	1.561
c5	5.610	2.505	1.603	21.875	1.473

4.4.2 Tests of Significance

Significance tests are performed between the clusters and the background readings for each features. The background is composed of around 6000 randomly generated regions from the whole genome. The methods for testing variances and means are as follows:

- Test for equality of two variances

  A two-tailed F-test is used to test the same variance with the significance level $\alpha = 0.05$, which is the criterion used for rejecting the null hypothesis. The test statistic

  $$f = \frac{S_x^2}{S_y^2},$$

  has an F-distribution with $n - 1$ and $m - 1$ degrees of freedom if the null hypotheses of equality of variances is true. Here, S_x^2 and S_y^2 are sample variances and n and m are the degrees of freedom of x and y. If the value of statistic f is between $F(0.975, n-1, m-1)$ and $F(0.025, n-1, m-1)$, the two variances are the same. If the value of f is smaller than $F(0.975, n-1, m-1)$ or greater than $F(0.025, n-1, m-1)$, the two variances are different.

  Table. 4.4 is an example of the tests and shows the test for variances in the histone modification ES.H3K4me3. The result indicate that there are no significant difference between the variances of five clusters and the background in the histone modification of H3K4me3 in embryonic stem cells.

- Test for equality of two means

  Two different tests with the significance level α are used according to the equal or unequal variances from previous variance test result.

  1. Test of mean equality with equal variance

Table 4.4: Part of the significance tests

ES.H3K4me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	2.355	81.049				
c1	528	2.216	83.744	1.033	1.130	0.879	-0.037
c2	147	2.081	4.785	0.059	1.246	0.782	-0.262
c3	63	1.338	0.664	0.008	1.384	0.678	-0.994
c4	21	6.332	13.392	0.165	1.711	0.479	3.377
c5	20	2.776	0.774	0.010	1.731	0.468	0.411

The test statistic t has a degree of freedom $n + m - 2$ and tests whether the means are different. t is calculated as follows:

$$t = \frac{\bar{x} - \bar{y}}{s_{xy}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$s_{xy} = \sqrt{\frac{(n-1)\, s_x^2 + (m-1)\, s_y^2}{n + m - 2}}$$

If $|t| > t(1 - \frac{\alpha}{2}, n + m - 2)$, there is significant difference between two means. If $|t| \leq t(1 - \frac{\alpha}{2}, n + m - 2)$, there is no significant difference between two means.

Table. 4.4 indicates the test of mean as well. it shows that only the means between cluster 4 and background are significantly different.

2. Test of mean equality with unequal variance

   The test statistic t for unequal variance is calculated by:

$$t = \frac{\bar{x} - \bar{y}}{s_{xy}}$$

   where

$$s_{xy} = \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$$

   The degrees of freedom of the test statistic is call the Welch-Satterthwaite equation:

$$d.f. = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{(n-1)} + \frac{\left(\frac{s_y^2}{m}\right)^2}{(m-1)}}$$

Actually, the values of the degrees of freedom of all tests are approximate to 1000. The t distributions are therefore approximately the same as a normal distribution. Therefore, we set a threshold of t value for all the tests as 3.0. This is because the critical absolute value of standard normal deviate is 3.09 for a two-tailed test for a normal distribution with the significance level $\alpha = 0.002$. Thus, if the absolute value of the statistic t is greater than 3, the cluster mean is considered to be significantly deviant from the background mean. The whole results of the tests are listed in the table in the Appendix. And we will discuss the biological significance from these results in the next section.

4.5 Discussion about Comparison Between the Results from Potts Clustering and K-means Clustering

In this section, we compare the result from Potts model clustering and K-means clustering. Since K-means clustering requires the pre-determined value of K, we set K to be the number of clusters from Potts model clustering at $T = 0.1$.

We choose first five biggest clusters from K-means clustering and compare the members of each cluster with those from Potts clustering. Table 4.5 shows the result of comparison.

The clusters $k1, k2, ...k5$ are the five largest ones from K-means and their sizes are $595, 89, 57, 28, 26$. Notice first that there is significant overlap between the results of the two clustering methods, giving us some confidence that these cluster identities are associated with real clustering in data. Looking in detail, we remark that, broadly speaking, k1 includes c1, k2 and k4 are inside of c2, k3 is inside of c3, k5 and c5 are mostly like. To contrast the results better, we create the

Table 4.5: Overlap of K-means and Potts clustering results

Overlap	Cluster size	k1	k2	k3	k4	k5
Cluster size	837	595	89	57	28	26
c1	528	513	2	4	0	1
c2	147	28	85	1	28	0
c3	63	10	1	43	0	0
c4	21	8	1	2	0	0
c5	20	4	0	0	0	16

3D plot of the clusters in three principal components.

As shown in Fig. 4.3, the clustering results are not the same and K-means clustering divides the second Potts cluster, represented by red points, into two clusters. In part (b) of Fig. 4.3, these two K-means clusters are actually part of a continuous structure, as far as we can see. The cluster c2 has significantly higher level of ES.H3K36me3 modification than other groups, which is the distinguishing feature of this cluster. Fig. 4.4 is the histogram of ES.H3K36me3 value in this cluster. There does not seem to be any need to break the cluster c2 up with the distribution, shown in Fig. 4.4, indicating no sign of bimodality. However, K-means clustering has the tendency to divide elongated shapes into multiple clusters, since the underlying data probability model for k-means is a mixture of isotropic Gaussians [46]. As a result, we have k2 and k4, capturing the medium high and the very high levels of the mark, respectively.

The advantage of Potts model clustering is well demonstrated here: no assumption is made regarding the underlying distribution of the data. The main advantage of this method is it generic applicability. It is natural to apply Potts model clustering to any underlying distribution of data. When there is no previous knowledge of the data distribution, the performance of normal methods such as K-means could be problematic. Therefore, it may be better to use methods like Potts model clustering to explore the distribution of the data points.

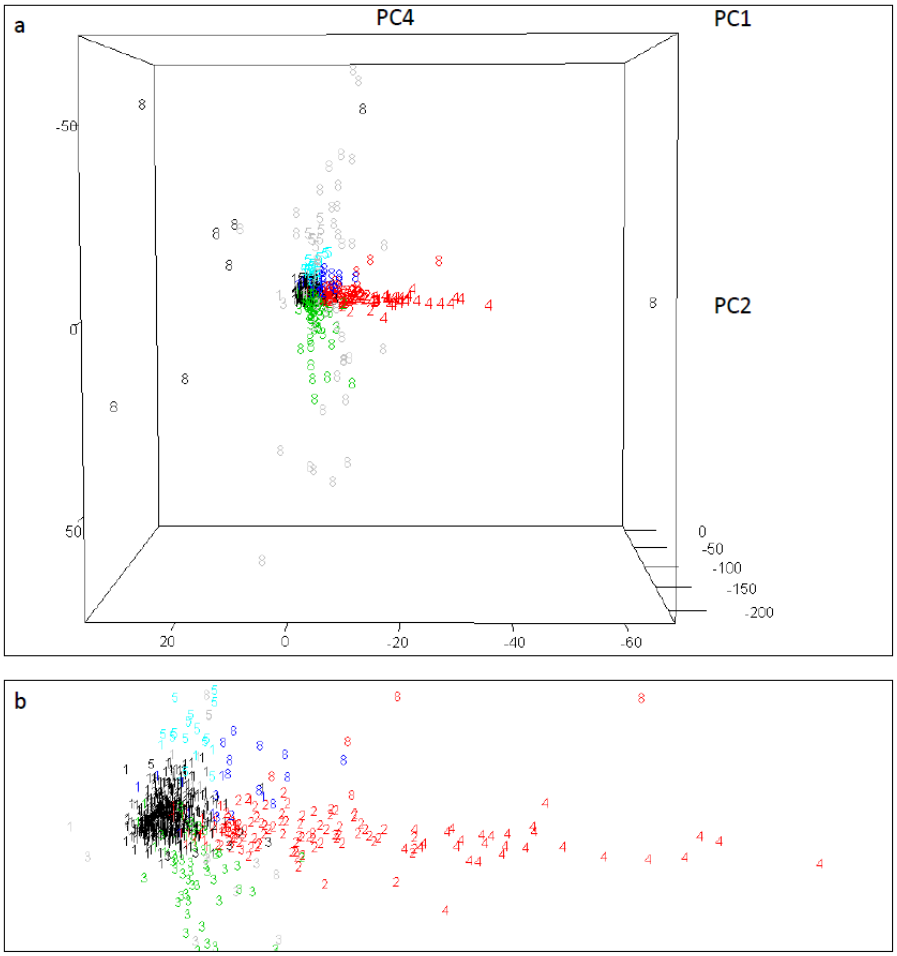

Figure 4.3: Clusters from K-means and Potts clustering. The different colors indicate the clusters from Potts clustering. The different numbers indicate the clusters from K-means clustering. (a) The whole data is projected along three principal components. (b) K-means clustering divides one cluster from Potts clustering into two clusters.

	Cluster size	ES.H3K36me3
k1	595	1.2288067
k2	89	4.3350562
k3	57	1.4168421
k4	28	13.3232143
k5	26	1.5446154
c1	528	1.2675
c2	147	5.6581
c3	63	1.38937
c4	21	1.69286
c5	20	1.5155

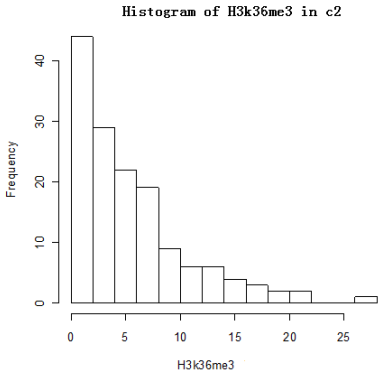

Figure 4.4: Histogram of H3k36me3 in cluster 2 of Potts clustering.

4.6 Biological Significance

4.6.1 Summary of Distinguished Epigenetic Marks

We summarize the distinguished epigenetic marks of the clusters in Table 4.6. These could be identified by inspecting the profiles of the clusters as well as by the significance tests for equality of cluster means for individual features.

Table 4.6: Summary of distinguished epigenetic marks

Cluster	Features	Possible functions
C1	No special features	
C2	High H3K36me3	Marks actively transcribed regions.
C3	High H3K4me1, H3K4me2 in neural precursor cells	Mark the transcription start site region of active gene. Related to activity of enhancers.
C4	High H3K4me1, H3K4me2, H3K4me3 in embryonic stem cells	Mark the transcription start site region of active gene. Related to activity of enhancers.
C5	High H3K27me3	Widespread across silent genes

We note that, primarily, one or two epigenetic marks distinguish most clusters. Cluster 3 has high H3K4me1 and H3K4me2 readings in neural precursor cell. Cluster 4 has high H3K4me1, H3K4me2 and H3K4me3 readings in embryonic stem cells. Cluster 5 has high H3K27me3 readings. Our results show that, for the data set at hand, there is not much combinatorial

histone coding [43, 11], which is suggested by the histone code hypothesis that multiple histone modifications act in a combinatorial way to specify distinct chromatin states.

4.6.2 Correlation between Epigenetic Marks and Enhancer Activity

The experiments described in last section show that some of the candidate sequences have enhancer activity. Table 4.7 is the proportion of elements with enhancer activity within each cluster. Notably in the table, the proportions of cluster 3 and cluster 4 have significantly higher values than that of the whole data set.

Table 4.7: Proportion of enhancer within each cluster

Cluster	Size	Number of tested elements with enhancer activity	Proportion of enhancer within each cluster
C1	528	237	0.45
C2	147	74	0.5
C3	63	40	0.63
C4	21	12	0.57
C5	20	9	0.45
Whole data set	860	412	0.48

There is significant enrichment of functional enhancers in cluster 3 ($p = 0.008$). It is possible that cluster 4 would have shown statistically significant enrichment as well, had there been more members in this cluster. For rest of the clusters, our results do not show much correlation. The correlation between H3K4 methylation and enhancer activity has been noted before [24]. In the meanwhile, high levels of the epigenetic marks related to H3K4 methylation in cluster 3 and cluster 4 come from different cell types. This shows that some enhancers are marked with cell-type-specific histone modification patterns and functionally active in different cell types at least in the early development stage. Also, c5, showing marks of polycomb silencing, seem to include regions which act as enhancers under some other circumstances, where the H3K27me3

marks are removed and other marks are put in. Our study suggests that it is important to associate the enhancer study with the cell types and the timing of the gene expression in the developmental program.

Our approach in this study has been to cluster first to discover different patterns and then explore what biological functions we can discover that are associated with these patterns. This exploration sets up more biologically interesting questions rather than answer them. Had we set ourselves the goal of making a model predictive of enhancer activity, we would have proceeded differently. Although, we may have had slightly higher predictive success in that regard, we are not sure it would have shed any more light, especially on the issue of cell-type specific marks.

Chapter 5

Epilogue

In the epilogue, we conclude the thesis by summarizing our contributions and propose some future work.

5.1 Summary of Contributions

We apply Potts model clustering to histone modification data in conserved regions and discover the significant patterns of marks in clusters. These patterns have some predictive values for cell-type-specific activity of enhancers, but not too much. Our study raises the question whether it is possible to make these predictions without having histone modification data for each specific cell type. In many studies related to gene expression, it has been shown that model trained on part of gene expression data is incapable of predicting something about the part that was withheld [47]. In our context, how much cell-type-specific data is necessary remains an open question. Our natural direction for our study, based on data from cell types early in development, is to see if we successfully predict enhancers active early in development as opposed to those which come into action later.

5.2 Future Directions

There are several open problems for our further research.

- We are including more conserved regions which are not listed in the VISTA enhancer data base to discover the epigenetic marks.

- We are interested in exploring the epigenetic marks from both enhancers and the their associated genes in different stage of gene expression.

In the post genomic era, we are still far from understanding how epigenome works in detail. It is my hope that the study in this thesis, which employs physics, statistics and biology, can help to reveal a small piece of the jigsaw puzzle of the epigenetics.

Appendix

The t tests for equality of two means in each epigenetic marks, the values of the degrees of freedom are approximate to 1000. The t distributions are therefore approximately the same as a normal distribution. Therefore, we set a threshold of t value for all the tests as 3.0. This is because the critical absolute value of standard normal deviate is 3.09 for a two-tailed test for a normal distribution with the significance level $\alpha = 0.002$. Thus, if the absolute value of the statistic t is greater than 3, the cluster mean is considered to be significantly deviant from the background mean. The whole results of the tests are listed in the table in the Appendix. And we will discuss the biological significance from these results in the next section.

Table 1: Tests of significance in ES.H3K4me3

ES.H3K4me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	2.355	81.049				
c1	528	2.216	83.744	1.033	1.130	0.879	-0.037
c2	147	2.081	4.785	0.059	1.246	0.782	-0.262
c3	63	1.338	0.664	0.008	1.384	0.678	-0.994
c4	21	6.332	13.392	0.165	1.711	0.479	3.377
c5	20	2.776	0.774	0.010	1.731	0.468	0.411

Table 2: Tests of significance in ES.H3K4me1

ES.H3K4me1							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.461	8.322				
c1	528	0.807	1.107	0.133	1.130	0.879	-5.655
c2	147	2.886	12.737	1.531	1.246	0.782	2.527
c3	63	1.120	1.153	0.139	1.384	0.678	-2.921
c4	21	8.359	20.256	2.434	1.711	0.479	7.771
c5	20	2.853	7.402	0.889	1.731	0.468	4.109

Table 3: Tests of significance in ES.H3K4me2

ES.H3K4me2							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.054	11.252				
c1	528	0.655	8.185	0.727	1.130	0.879	-1.041
c2	147	1.405	6.276	0.558	1.246	0.782	1.139
c3	63	0.423	0.172	0.015	1.384	0.678	-4.437
c4	21	6.154	17.966	1.597	1.711	0.479	6.417
c5	20	1.468	1.527	0.136	1.731	0.468	2.639

Table 4: Tests of significance in ES.H3K27me3

ES.H3K27me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.786	9.347				
c1	528	1.907	4.108	0.439	1.130	0.879	0.563
c2	147	1.640	2.188	0.234	1.246	0.782	-0.963
c3	63	2.121	1.811	0.194	1.384	0.678	2.363
c4	21	3.354	3.601	0.385	1.711	0.479	7.993
c5	20	4.838	5.436	0.582	1.731	0.468	11.543

Table 5: Tests of significance in ES.H3K9me3

ES.H3K9me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.209	2.285				
c1	528	1.241	0.734	0.321	1.130	0.879	0.746
c2	147	1.306	0.552	0.242	1.246	0.782	2.578
c3	63	1.193	0.704	0.308	1.384	0.678	-0.382
c4	21	1.589	0.851	0.372	1.711	0.479	8.095
c5	20	1.657	1.397	0.611	1.731	0.468	6.658

Table 6: Tests of significance in ES.H4K20me3

ES.H4K20me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.429	2.253				
c1	528	1.454	0.702	0.312	1.130	0.879	0.614
c2	147	1.454	0.588	0.261	1.246	0.782	0.660
c3	63	1.825	1.158	0.514	1.384	0.678	6.848
c4	21	2.223	1.433	0.636	1.711	0.479	11.587
c5	20	1.406	0.858	0.381	1.731	0.468	-0.494

Table 7: Tests of significance in ES.H3K36me3

ES.H3K36me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.777	8.920				
c1	528	1.268	3.466	0.389	1.130	0.879	-2.708
c2	147	5.658	24.969	2.799	1.246	0.782	3.552
c3	63	1.389	0.761	0.085	1.384	0.678	-3.304
c4	21	1.693	0.987	0.111	1.711	0.479	-0.700
c5	20	1.516	1.224	0.137	1.731	0.468	-2.101

Table 8: Tests of significance in ES.RPol2

ES.RPol2							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	0.770	0.948				
c1	528	0.804	0.350	0.369	1.130	0.879	1.773
c2	147	0.942	0.353	0.372	1.246	0.782	8.868
c3	63	0.844	0.334	0.352	1.384	0.678	3.935
c4	21	1.010	0.515	0.543	1.711	0.479	9.446
c5	20	0.915	0.675	0.712	1.731	0.468	4.570

Table 9: Tests of significance in ES.H3

ES.H3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.259	1.468				
c1	528	1.367	0.583	0.007	1.130	0.879	-0.965
c2	147	1.544	0.680	0.008	1.246	0.782	-0.793
c3	63	1.480	0.505	0.006	1.384	0.678	-0.855
c4	21	1.571	0.718	0.009	1.711	0.479	-0.766
c5	20	1.635	0.857	0.011	1.731	0.468	-0.703

Table 10: Tests of significance in ES.WCE

ES.WCE							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	0.209	0.228				
c1	528	0.213	0.082	0.360	1.130	0.879	0.741
c2	147	0.250	0.095	0.418	1.246	0.782	8.025
c3	63	0.222	0.076	0.331	1.384	0.678	2.860
c4	21	0.370	0.117	0.511	1.711	0.479	27.614
c5	20	0.288	0.163	0.717	1.731	0.468	10.176

Table 11: Tests of significance in NP.H3K4me2

NP.H3K4me2							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	2.446	102.941				
c1	528	1.940	82.879	1.023	1.130	0.879	-0.111
c2	147	1.121	1.151	0.014	1.246	0.782	-1.204
c3	63	8.770	89.253	1.101	1.384	0.678	1.597
c4	21	2.111	1.465	0.018	1.711	0.479	-0.238
c5	20	1.060	1.041	0.013	1.731	0.468	-1.265

Table 12: Tests of significance in NP.H3K4me1

NP.H3K4me1							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.760	19.259				
c1	528	1.114	4.166	0.216	1.130	0.879	-2.129
c2	147	1.740	4.260	0.221	1.246	0.782	-0.065
c3	63	10.944	66.673	3.462	1.384	0.678	3.154
c4	21	3.138	6.925	0.360	1.711	0.479	3.559
c5	20	1.834	2.728	0.142	1.731	0.468	0.272

Table 13: Tests of significance in ESHyb.H3K4me3

ESHyb.H3K4me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	3.304	98.180				
c1	528	3.221	105.188	1.071	1.130	0.879	-0.017
c2	147	3.043	9.406	0.096	1.246	0.782	-0.200
c3	63	2.211	0.847	0.009	1.384	0.678	-0.881
c4	21	5.957	11.009	0.112	1.711	0.479	1.996
c5	20	3.136	1.162	0.012	1.731	0.468	-0.136

Table 14: Tests of significance in ESHyb.H3K36me3

ESHyb.H3K36me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	2.246	11.462				
c1	528	1.650	3.826	0.334	1.130	0.879	-2.704
c2	147	6.896	33.596	2.931	1.246	0.782	3.165
c3	63	1.989	1.282	0.112	1.384	0.678	-1.659
c4	21	1.949	1.793	0.156	1.711	0.479	-1.810
c5	20	1.652	0.528	0.046	1.731	0.468	-4.058

Table 15: Tests of significance in ESHyb.H3K9me3

ESHyb.H3K9me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.339	1.902				
c1	528	1.326	0.726	0.382	1.130	0.879	-0.333
c2	147	1.249	0.647	0.340	1.246	0.782	-2.425
c3	63	1.413	1.601	0.842	1.384	0.678	1.007
c4	21	1.448	0.534	0.281	1.711	0.479	3.263
c5	20	1.008	0.476	0.250	1.731	0.468	-10.456

Table 16: Tests of significance in MEF.H3K4me3

MEF.H3K4me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	3.104	103.991				
c1	528	3.447	136.678	1.314	1.130	0.879	0.056
c2	147	2.821	1.962	0.019	1.246	0.782	-0.215
c3	63	2.725	2.938	0.028	1.384	0.678	-0.287
c4	21	3.110	3.372	0.032	1.711	0.479	0.004
c5	20	2.011	1.365	0.013	1.731	0.468	-0.832

Table 17: Tests of significance in MEF.H3K27me3

MEF.H3K27me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	3.320	21.289				
c1	528	3.701	11.060	0.136	1.130	0.879	1.190
c2	147	2.281	1.415	0.017	1.246	0.782	-0.072
c3	63	3.606	5.794	0.071	1.384	0.678	1.187
c4	21	4.496	3.723	0.046	1.711	0.479	2.066
c5	20	14.690	21.413	0.264	1.731	0.468	8.912

Table 18: Tests of significance in MEF.H3K36me3

MEF.H3K36me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	2.275	16.302				
c1	528	1.467	3.738	0.229	1.130	0.879	-3.082
c2	147	8.464	28.792	1.766	1.246	0.782	4.874
c3	63	2.063	4.032	0.247	1.384	0.678	-0.787
c4	21	2.002	2.738	0.168	1.711	0.479	-1.148
c5	20	1.288	0.763	0.047	1.731	0.468	-4.737

Table 19: Tests of significance in MEF.H3K9me3

MEF.H3K9me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.031	1.413				
c1	528	1.028	0.544	0.385	1.130	0.879	-0.097
c2	147	1.244	0.717	0.507	1.246	0.782	5.930
c3	63	0.953	0.454	0.321	1.384	0.678	-2.909
c4	21	1.437	0.690	0.488	1.711	0.479	11.641
c5	20	1.188	0.481	0.340	1.731	0.468	5.726

Table 20: Tests of significance in MEF.WCE

MEF.WCE							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.376	1.689				
c1	528	1.402	0.573	0.339	1.130	0.879	0.791
c2	147	1.579	0.526	0.312	1.246	0.782	6.469
c3	63	1.646	0.718	0.425	1.384	0.678	7.139
c4	21	1.338	0.481	0.285	1.711	0.479	-1.293
c5	20	1.034	0.482	0.285	1.731	0.468	-11.462

Table 21: Tests of significance in NP.H3K4me3

NP.H3K4me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.817	14.923				
c1	528	1.944	14.681	0.984	1.130	0.879	0.190
c2	147	1.756	0.708	0.047	1.246	0.782	-0.319
c3	63	2.488	1.425	0.095	1.384	0.678	3.383
c4	21	1.739	0.559	0.037	1.711	0.479	-0.411
c5	20	1.151	0.395	0.026	1.731	0.468	-3.522

Table 22: Tests of significance in NP.H3K27me3

NP.H3K27me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	2.339	4.430				
c1	528	2.595	2.818	0.636	1.130	0.879	1.895
c2	147	1.916	1.048	0.236	1.246	0.782	-5.864
c3	63	2.703	1.920	0.433	1.384	0.678	3.616
c4	21	2.795	1.700	0.384	1.711	0.479	4.911
c5	20	4.075	4.443	1.003	1.731	0.468	8.621

Table 23: Tests of significance in NP.H3K36me3

NP.H3K36me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.962	5.632				
c1	528	1.660	1.822	0.022	1.130	0.879	-0.677
c2	147	5.000	10.722	0.132	1.246	0.782	2.352
c3	63	2.553	3.570	0.044	1.384	0.678	0.191
c4	21	1.648	1.524	0.019	1.711	0.479	-0.690
c5	20	0.917	0.330	0.004	1.731	0.468	-1.406

Table 24: Tests of significance in NP.H3K9me3

NP.H3K9me3							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.150	1.346				
c1	528	1.199	0.490	0.364	1.130	0.879	1.790
c2	147	1.201	0.522	0.388	1.246	0.782	1.782
c3	63	1.287	0.535	0.398	1.384	0.678	4.742
c4	21	0.988	0.539	0.401	1.711	0.479	-5.613
c5	20	0.977	0.471	0.350	1.731	0.468	-6.509

Table 25: Tests of significance in NP.WCE

NP.WCE							
	Cluster Size	Mean	Variance	f	F0.025	F0.975	t
Background	6275	1.107	1.828				
c1	528	1.140	0.955	0.522	1.130	0.879	0.705
c2	147	1.588	1.670	0.913	1.246	0.782	6.308
c3	63	1.251	0.687	0.376	1.384	0.678	3.820
c4	21	2.328	1.856	1.015	1.711	0.479	14.530
c5	20	1.732	0.452	0.247	1.731	0.468	20.588

References

[1] G. M. Cooper and R. E. Hausman. The cell: a molecular approach. *Sinauer Associates, Inc.*, 2006.

[2] J. D. Watson and F. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737, 1953.

[3] H. Pearson. Genetics: what is a gene? *Nature*, 441:398, 2006.

[4] F. Crick. On protein synthesis. *Symp. Soc. Exp. Biol.*, 7:139, 1958.

[5] E.S. Lander, L.M.Linton, and B. Birren et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860, 2001.

[6] R. Holliday. Mechanisms for the control of gene activity during development. *Biol. Rev.*, 65:431, 1990.

[7] V. E. A. Russo, R. A. Martienssen, and A. D. Riggs. Epigenetic mechanisms of gene regulation. *Cold Spring Harbor Laboratory Press, Plainview, NY.*, 1996.

[8] H. H. Ng and A. Bird. DNA methylation and chromatin modification. *Curr. Opin. Genet. Dev*, 9:158, 99.

[9] B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128:669, 2007.

[10] O. Aparicio, J. V. Geisberg, and K. Struhl. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Cell Biol.*, 17:17, 2004.

[11] S. L. Schreiber and B. E. Bernstein. Signaling network model of chromatin. *Cell*, 128:707, 2002.

[12] D. Robyr et al. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell*, 109:437, 2002.

[13] T. S. Mikkelsen et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448:553, 2007.

[14] A. Barski et al. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823, 2007.

[15] D. E. Schones and K. Zhao. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9:179, 2008.

[16] V. W. Zhou, A. Goren, and B. E. Bernstein. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 2010.

[17] D. K. Pokholok et al. Map of nucleosome acetylation and methylation in yeast. *Cell*, 122:517, 2005.

[18] B. E. Bernstein et al. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120:169, 2005.

[19] T. H. Kim et al. A high-resolution map of active promoters in the human genome. *Nature*, 436:876, 2005.

[20] Z. Wang et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40:897, 2008.

[21] L. A. Boyer et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441:349, 2006.

[22] J. H. Martens et al. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.*, 24:800, 2005.

[23] N. D. Heintzman et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39:311, 2007.

[24] B. Ren et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459:108, 2009.

[25] J. Felsenstein. Inferring phylogenies. 2004.

[26] S. Draghici. Data analysis tools for DNA microarrays. 2003.

[27] P. Dhaeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23:1499, 2005.

[28] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical*, 1967.

[29] T. Kohonen. The Self-Organizing Map. *Proceedings of the IEEE*, 78:1464, 1990.

[30] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1101, 1993.

[31] H. H. Nguyen and P. Cohen. Gibbs random fields, fuzzy clustering, and the unsupervised segmentation of textured images. *CVGIP: Graph. Models Image Process*, 55:1, 1993.

[32] R. Kindermann and J. L. Snell. Markov random fields and their applications. *AMS*, 1980.

[33] M. Blatt, S. Wiseman, and E. Domany. Super-paramagnetic clustering of data. *Physical Review Letters*, 76:3251, 1996.

[34] R. B. Potts. Some generalised order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, 48:106, 1952.

[35] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087, 1953.

[36] R. H. Swendsen and J-S. Wang. Non-universal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86, 1987.

[37] R. G. Edwards and A. D. Sokal. Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Phys. Rev. D*, 38:2009, 1988.

[38] A. Billoire, R. Lacaze, A. Morel, S. Gupta, A. Irback, and B. Petersson. Dynamics near a first-order phase transition with the Metropolis and Swendsen-Wang algorithms. *Nuclear Physics*, 358:231, 1991.

[39] J. Hoshen and R. Kopelman. Percolation and cluster distribution. i. cluster multiple labeling technique and critical concentration algorithm. *Physical ReviewB*, 14:3438, 1976.

[40] A. Al-Futais and T. W. Patzek. Extension of Hoshen-Kopelman algorithm to non-lattice environments. *Physica A*, 321:665, 2003.

[41] M. Blatt, S. Wiseman, and E. Domany. Data clustering using a model granular magnet. *Neural Computation*, 9:1805, 1997.

[42] F. Niedermayer. Improving the improved estimators in O(n) spin models. *Physical Letters*, 237:473, 1990.

[43] T. Jenuwein and C. Allis. Translating the histone code. *Science*, 293:1074, 2001.

[44] T. S. Mikkelsen et al. Dissecting direct reprogramming through integrative genomic analysis.. *Nature*, 454:49, 2008.

[45] T. S. Mikkelsen et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454:766, 2008.

[46] C. M. Bishop. Pattern recognition and machine learning. *Springer*, 2006.

[47] M. Middendorf et al. Predicting genetic regulatory response using classification. *Bioinformatics*, 20:232, 2004.

[48] B. E. Bernstein, et al., E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28:1045, 2010.

[49] L. A. Pennacchio et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444:499, 2006.

[50] F. Poulin, et al., and L. A. Pennacchio. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics*, 85:774, 2005.

Curriculum Vitae

Junyi Li

Education

2005-2011 Ph.D. in Physics, Rutgers University, Piscataway, NJ, USA

2007-2009 MS in Statistics, Rutgers University, Piscataway, NJ, USA

1998-2002 BS in Physics, Peking University, Beijing, China

Experiences

2008-2010 Graduate Assistant, Department of Physics and Astronomy, Rutgers University, NJ

2005-2008 Teaching Assistant, Department of Physics and Astronomy, Rutgers University, NJ

Research Projects

Potts model cluster analysis of large histone modification data sets in different DNA regions such as highly conserved region. (Programming with R, Perl and Python)

Application of discriminant analysis on the dendritic branching of Hippocampal Neurons. (Programming with R)

Honors and Awards

2001-2002 Fellowship of Jun Zhen scholar in Peking University.