6.047/6.878 Lecture 17: Epigenomics/Chromatin States

George Tucker, Victor Pontis 2012

November 6, 2012

Contents

List of Figures

1 Introduction

The human body contains approximately 210 different cell types, but each cell type shares the same genomic sequence. In spite of having the same genetic code, cells not only develop into distinct types from this same sequence, but also maintain the same cell type over time and across divisions. This information about the cell type and the state of the cell is called *epigenomic* information. The epigenome ("epi" means above in Greek, so epigenome means above genome) is the set of chemical modifications or marks that influence gene expression and are transferred between generations of cells [6] and between generations of people.

The study of the epigenome is particularly interesting because it is one of the first pieces of evidence that goes against Darwin's description of evolution. The human genome does not change much through the course of one's life and before epigenetics was studied, people thought that the DNA sequence was the sole factor in determining who and what a person is. Epigenetics bring up the interesting possibility that what you do while you are alive – how you live your life – changes your genetic makeup and the genetic makeup of your descendents. Does eating a lot of fatty, greasy foods make your children more susceptible to diabetes? Does habitual smoking change the expression of genes that are important in the respiratory system? These are the kinds of questions that can be posed and answered with epigenetics.

As shown in Figure 1, epigenomic information in a cell is encoded in diverse ways. For example, direct methylation of DNA (e.g. at CpG dinucleotides) can alter gene expression. Similarly, positioning of nucleosomes (unit of packing of DNA) determines which parts of DNA are accessible to TFs and other enzymes. Finally, a very powerful way to encoding epigenomic information is through chemical modifications (e.g. methylation, acetylation etc) of histone protein tails.

The 2012 Nobel Prize in Physiology and Medicine was awarded to John Gurdon and Shinya Yamanaka, for work that involved the reprogramming mature cells to become pluripotent. They used different techniques that modified the epigenomic information of the cell and modified the cell into an iPS (induced pluripotent stem cell). This research highlights the importance and applications of epigenetics. These stem cells could be used in the future to grow organs or be injected into people to repair tissue.

In this chapter we will go over histones which carry epigenetic information. We will then look at how we can gather information about histones and identify the location of histones on the genome using CHIP-seq and CHIP-chip. We then see how to analyze this information using the Burrows-Wheeler to allow for efficient search and mapping of our data. From this we then abstract a level and use a hidden Markov model (HMM) to look at chromatin states and group them by function. This will give us an idea of different chromatin states and their function on the human genome.

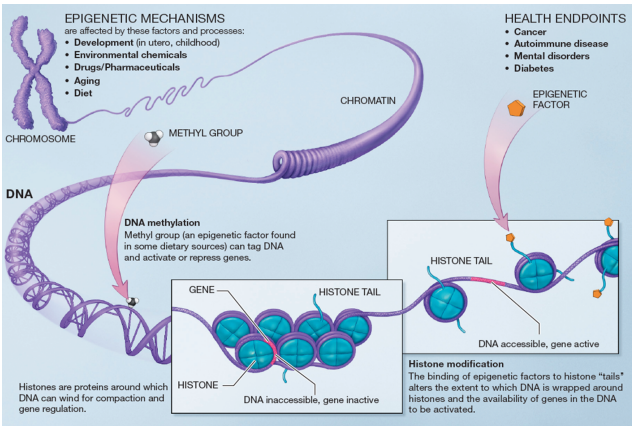

Figure 1: Types of epigenetic modifications

2 Epigenetic Information in Nucleosomes

A key method of encoding epigenetic data is right on the DNA itself. The DNA is annotated with histones that are further chemically modified. You can think of the histones as post-its on the genome. First we will see histones are and how the interact with DNA. Then we will look at how they can signal things and affect gene expression.

In order to fit two meters of DNA into a 5-20 μm diameter cell nucleus and arrange the DNA for easy access to transcriptional machinery, DNA is packaged into chromatin. Nucleosomes form the unit of this packaging. A nucleosome is composed of DNA approximately 150-200 bp long wrapped around a histone protein octamer consisting of two copies each of histone proteins H2A, H2B, H3, and H4 (and occasionally a linker histone H1 or H5). While the structure and importance of higher-level packaging of nucleosomes is less known, the lower-level arrangement and modification of nucleosomes is very important to transcriptional regulation and the development of different cell types. Histone proteins H3 and H4 are the most highly conserved proteins in the eukaryotic domain of life. If DNA contains the blueprints of life, nucleosomes contain the blueprints of life with multiple cell types.

Nucleosomes encode epigenetic information in two ways. First, their positions on the DNA are determine which parts of DNA are accessible. Nucleosomes are often bound to the promoters of inactive genes. To initiate transcription of a gene, transcription factors (TFs) and the General Factors have to bind to its promoter. Therefore, when a gene becomes active, the nucleosomes located on its promoter are often pushed aside or removed. The promoter will remain exposed until further modifications are made. Hence, nucleosome positioning on the DNA is stable, yet mutable. This property of stability and mutability is a prerequisite for any form of epigenetic information because cells need to maintain the identity of a particular cell type, yet still be able to change their epigenetic state to respond to environmental circumstances

Second, nucleosomes contain tails of amino acid residues protruding from the ends of their histones. These tails can undergo post-translational modification such as methylation, acetylation and phosphorylation which affect gene expression by recruiting initiation factors. There are over 100 distinct histone modifications that have been found experimentally. This has lead to the "histone code hypothesis" that combinations of chromatin modifications encode biological function. These modifications are so common that a shorthand has been developed to identify them. This shorthand consists of the name of the histone protein, the amino acid residue on its tail that has been modified and the type of modification made to this residue. To illustrate,

the fourth residue from the N-terminus of histone H3, lysine, is often methylated at the promoters of active genes. This modification is described as H3K4me3 (if methylated thrice). The first part of the shorthand corresponds to the histone protein, in this case H3; K4 corresponds to the 4th residue from the end, in this case a lysine, and me3 corresponds to the actual modification, the addition of 3 methyl groups in this case.

3 Technologies for measurement of epigenetic signals

Given the importance of epigenomic information in biology, great efforts have been made to study these signals on DNA. One common method for epigenomic mark measurement is chromatin immunoprecipitation (ChIP). The procedures of ChIP are described as follows and depicted in Figure 2:

1. Cells are exposed to a cross-linking agent such as formaldehyde, which causes covalent bonds to form between DNA and its bound proteins (e.g. histones with specific modifications);
2. Genomic DNA is isolated from the cell nucleus;
3. Isolated DNA is sheared by sonication or enzymes;
4. Antibodies against a specific epigenetic mark are then added to pull out its associated DNA. These antibodies are generated by exposing proteins of interest to mammals (e.g. goats or rats). The resulting immune response will cause the production of specific antibodies.
5. The cross-linking between the protein and DNA is reversed and the DNA fragments specific to the epigenetic marks are purified.

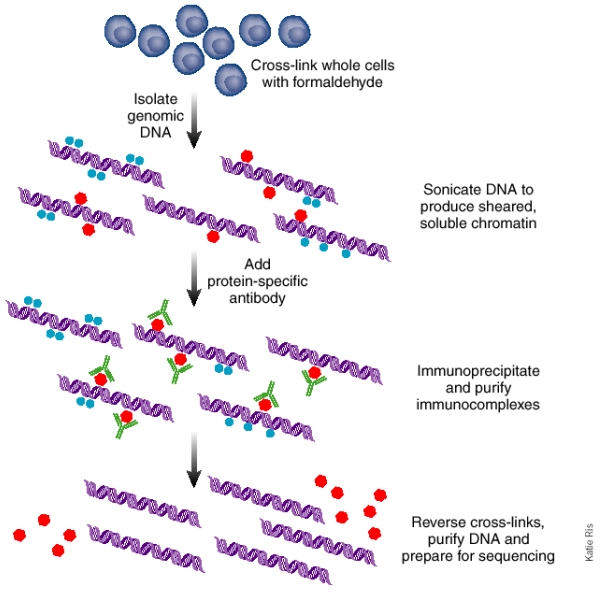

Figure 2: Chromatin immunoprecipitation [5]

So after this process we have different short sequences of DNA that correspond to spots where histones were bound to the DNA. To identify the location of these DNA fragments on the genome, one can hybridize them to known DNA segments on an array or gene chip and visualize them with fluorescent marks (called ChIP-chip). Alternatively, one can do massive parallel next-generation sequencing of these fragments (called ChIP-seq). Each sequence tag is 30 base pairs long. These tags are mapped to unique positions in the reference genome of 3 billion bases. The number of reads depending on sequencing depth, but typically there are on the order of 10 million mapped reads for each ChIP-seq experiment. ChIP-seq is preferred over

ChIP-chip nowadays because it has wider dynamic range of detection and can avoid problems such as cross-hybridization in ChIP-chip. Given this data from ChIP-chip or CHIP-seq, analysis of epigenetic data consists of various steps. First, the reads from ChIP must be mapped to the DNA (called read mapping). Next, we must determine which readings correspond to presence of a chromatin mark (called peak calling). After these preprocessing steps, we can build different supervised and unsupervised models to study chromatin states and their relation to biological function. We look at each of these steps in turn.

4 Read Mapping

The problem of read mapping seeks to assign a given read to the best matching location in the reference genome. Given the vast amount of reads and the size of human genome, one common requirement of all read mapping algorithms is the algorithmic time efficiency. We are going to have a huge number of reads and we want to place each of them on the genome which is billions of base pairs long, we need a very fast algorithm to allow us to do this in a reasonable time.

Based on previous lectures, we can image different ways to perform mapping of reads - sequence alignment (O(mn)), hash-based approaches such as MAQ, linear time string matching (O(m+n)) and suffix trees and arrays (O(m)). However, a problem with all these techniques is that they have a large memory requirement (e.g. O(mn)). Instead, state-of-the-art techniques based on the Burrows-Wheeler transformation [1] runs in O(m) time and require O(n) space.

The Burrows-Wheeler transform came from the need to compress information. It would take a long string and rearrange it in a way that there were often repeating letters. This string could be compressed because instead of writing 100 A's the computer could just indicate that there were now 100 A's in a row. The Burrows-Wheeler transform also has some other special properties that we will be exploiting to search in sublinear time.

The Burrows-Wheeler transform creates a unique transformed string that is shorter than the original string. It also can be reversed easily to generate your original string, so no information is lost. The transformed string is in sorted order which allows for easy searching. The details of Burrows-Wheeler transformation are described as follows and illustrated in Figure 3:

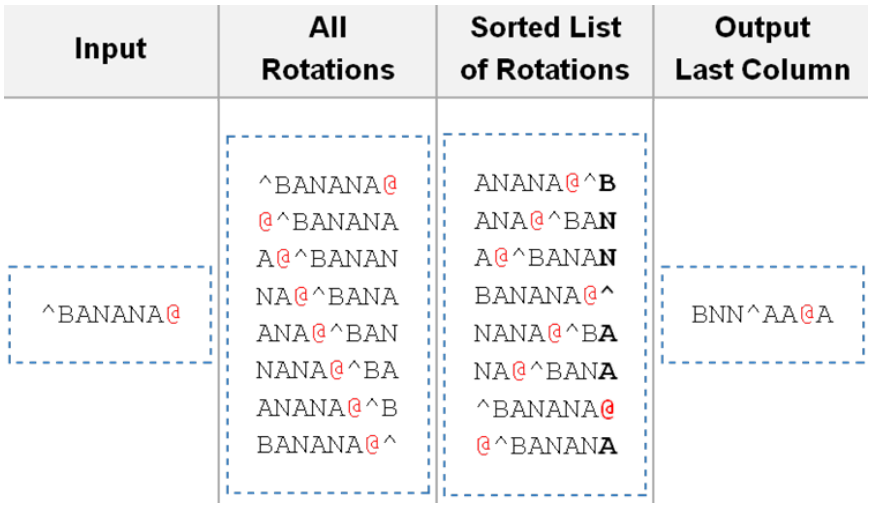

Figure 3: The Burrows-Wheeler forward transformation

First, producing a transform from an original string consists of the following steps:

1. For a given reference genome, add a special character at the beginning and end of the string (e.g. BANANA becomes `^BANANA@`). Then generate all the rotations of this string, e.g., `NANA@^BA`.
2. Sort the rotations lexicographically, i.e. in alphabetical order, with special characters sorted first.
3. The last column of the sorted list of rotations contains the transformed string

Add 1	Sort 1	Add 2	Sort 2	Add 3	Sort 3	Add 4	Sort 4	Add 5	Sort 5	Add 6	Sort 6	Add 7	Sort 7	Add 8	Sort 8
B	A	BA	AN	BAN	ANA	BANA	ANAN	BANAN	ANANA	BANANA	ANANA@	BANANA@	ANANA@^	BANANA@^	ANANA@^B
N	A	NA	AN	NAN	ANA	NANA	ANA@	NANA@	ANA@^	NANA@^	ANA@^B	NANA@^B	ANA@^BA	NANA@^BA	ANA@^BAN
N	A	NA	A@	NA@	A@^	NA@^	A@^B	NA@^B	A@^BA	NA@^BA	A@^BAN	NA@^BAN	A@^BANA	NA@^BANA	A@^BANAN
^	B	^B	BA	^BA	BAN	^BAN	BANA	^BANA	BANAN	^BANAN	BANANA	^BANANA	BANANA@	^BANANA@	BANANA@^
A	N	AN	NA	ANA	NAN	ANAN	NANA	ANANA	NANA@	ANANA@	NANA@^	ANANA@^	NANA@^B	ANANA@^B	NANA@^BA
A	N	AN	NA	ANA	NA@	ANA@	NA@^	ANA@^	NA@^B	ANA@^B	NA@^BA	ANA@^BA	NA@^BAN	ANA@^BAN	NA@^BANA
@	^	@^	^B	@^B	^BA	@^BA	^BAN	@^BAN	^BANA	@^BANA	^BANAN	@^BANAN	^BANANA	@^BANANA	^BANANA@
A	@	A@	@^	A@^	@^B	A@^B	@^BA	A@^BA	@^BAN	A@^BAN	@^BANA	A@^BANA	@^BANAN	A@^BANAN	@^BANANA

Figure 4: The Burrows-Wheeler reverse transformation

Once a Burrows-Wheeler transform has been computed, we can reverse the transform to compute the original string via the procedure in (Figure 4). Briefly, the reverse transformation works as follows: given the transformed string, sort the string characters in alphabetical order; this gives the first column in the transform. Combine these two columns to get pairs of characters. Sort the pairs and repeat.

From the Burrows-Wheeler transform we observe that all occurrences of the same suffix are effectively next to each other rather than scattered throughout the genome. Moreover, the i^{th} occurrence of a character in the first column corresponds to the i^{th} occurrence in the last column. Searching for substrings using the transform is also easy. Suppose we are looking for the substring ANA in the given string. Then the problem of search is reduced to searching for a prefix ANA among all possible sorted suffixes (generated by rotations). In the case of read mapping, the genome is transformed using the Burrows-Wheeler transform and stored in O(n) space. We can then efficiently search for substrings (reads) using the strategy described above.

5 Peak Calling

After reads are aligned, signal tracks as shown in Figure 5 can be computed. This data can be ordered into a long histogram spanning the length of the genome and indicating the number of reads (or degree of fluorescence in the case of ChIP-chip) found at each position in the genome. More reads (or fluorescence) means the epigenomic marker of interest is most often present at this particular location. For histone modifications, peak calling methods are based on univariate Hidden Markov Model, or scan statistics (count the reads within bins of certain size and apply statistical analysis). Problems tend to arise with broader domains due to the ambiguity of calling them one large peak or multiple smaller peaks. Instead, a simple but effective strategy consists of *Data binarization*. During data binarization, the only call to be made is whether a chromatin marker is observed in an interval or not. Typically, one can set a threshold on the amount of signal that must be present and then any interval with more than that quantity of signal gets a value of 1 and gets 0 otherwise. Data binarization makes the data easy to interpret, less prone to overfitting and amenable to simpler models.

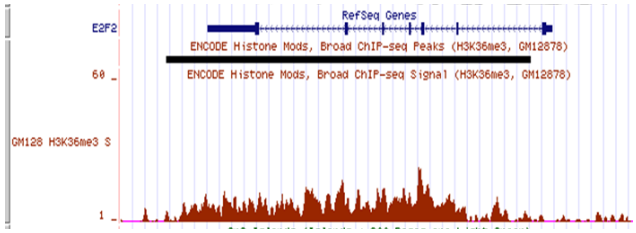

Figure 5: Sample signal tracks

We now move on to techniques for interpreting chromatin marks. There are many ways to analyze epigenomic marks such as aggregating chromatin signals (such as H3K4me3) on known feature types (e.g. promoters of genes with high or low expression levels), performing supervised or unsupervised machine learning methods to derive epigenomic features that are predictive of different types of genomics elements such as promoters, enhancers or large intergenic non-coding RNAs. In particular, in this lecture, we examine in detail the analysis of chromatin marks as done in [3].

6 Annotating the Genome Using Chromatin Signatures

The histone code hypothesis suggests that chromatin-DNA interactions are guided by **combinatorial histone modifications**. These combinatorial modifications, when taken together, can in part determine how a region of DNA is interpreted by the cell (i.e. as a transcription factor binding domain, a splice site, an enhancer region, an actively expressed gene, a repressed gene, or a non functional region). We are interested in interpreting this "code" (i.e. determining from histone marks at a region whether the region is a transcription start site, enhancer, promoter, etc.). With an understanding of the combinatorial histone marks, we can annotate the genome into functional regions and predict novel enhancers, promoters, genes, etc. The challenge is that there are dozens of marks and they exhibit complex combinatorial effects.

In this section, we explore a technique for interpreting the "code" and its application to a specific dataset [3], which measured 41 chromatin marks across the human genome.

6.1 Data

Data for this analysis consisted of 41 chromatin marks including acetylations, methylations, H2AZ, CTCF and PolII in CD4 T cells. First, the genome was divided into 200 bp non-overlapping bins in which the binary absence or presence of each of the 41 chromatin marks was determined. This data was processed using **data binarization** as described in the previous section. Specifically, let C_{ij} be the number of reads detected by ChIP-seq for mark i, mapping to the 200bp bin j. Let λ_i be the average number of reads mapping to a bin for mark i. The mark i is determined to be present in bin j if $P(X > C_{ij}) < 10^{-4}$ **TODO: expand** *@scribe: How is the threshold chosen?* where X is a Poisson random variable with mean λ_i and absent otherwise. In order words, the read enrichment for a specific bin has to be significantly greater than a random process of putting reads into bins. An example for chromatin states around the CAPZA2 gene on chromosome 7 is shown in Figure 6. So in this way, for each mark i, we can label each bin j with a 1 if the mark is present and a 0 if it isn't. Looking at the data as a whole, we can think of it as large binary matrix, where each row corresponds to a mark and each column corresponds to a bin (which is simply a 200bp region of the genome).

Additional data used for analysis included gene ontology data, SNP data, expression data, and others.

6.2 HMMs for Chromatin State Annotation

Our goal is to identify biologically meaningful and spatially coherent combinations of chromatin marks. Remember that we broke the genome up into 200bp blocks, so by spatially coherent we mean that if we have a genomic element that is longer than 200bps, we expect the combination of chromatin marks to be consistent on each 200bp bin in the region. We'll call these biologically meaningful and spatially coherent combinations of chromatin marks **chromatin states**. In previous lectures, we've seen HMMs applied to genome annotation for genes and CpG islands. We would like to apply the same ideas to this situation, but in this case, we don't know the hidden states a priori (e.g. CpG island region or not), we'd like to learn them *de novo*. This model can capture both the functional ordering of different states (e.g from promoter to transcribed regions) and the spreading of certain chromatin domains across the genomes. To summarize, we want to learn an HMM where the hidden states of the HMM are chromatin states.

As we learned previously, even if we don't know the emission probabilities and transition probabilities of an HMM, we can use the Baum-Welch training algorithm to learn the maximum likelihood values for those parameters. In our case, we have an added difficulty, we don't even know how many chromatin states exist! In the following subsections, we'll expand on how the data is modeled and how we can choose the number of states for the HMM.

6.2.1 Emission of a Vector

In HMMs from previous lectures, each state emitted either a single nucleotide or a single string of nucleotides at a time. In the HMM for this problem, each state emits a combination of epigenetic marks. Each combination can be represented as an n-dimensional vector where n is the number of chromatin marks being analyzed ($n = 41$ for our data). For example, assuming you have four possible epigenetic modifications: H3K4me3, H2BK5ac, Methyl-C, and Methyl-A, a sequence containing H3K4me3 and Methyl-C could be

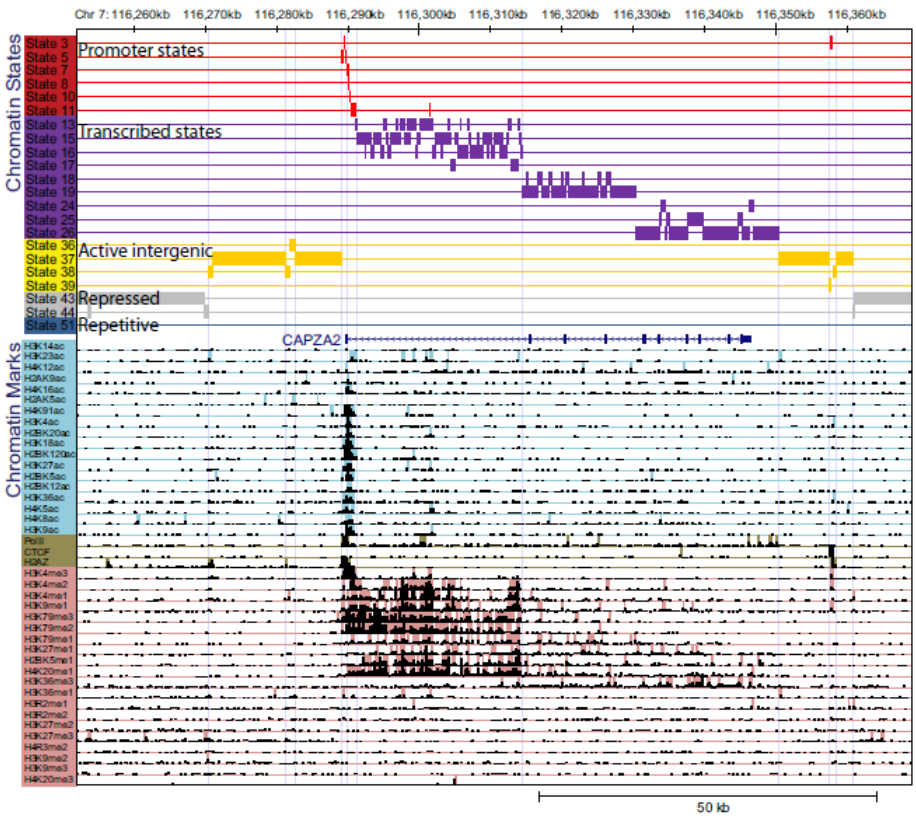

Figure 6: Example of the data and the annotation from the HMM model. The bottom section shows the raw number of reads mapped to the genome. The top section shows the annotation from the HMM model.

presented as the vector (1, 0, 1, 0). One could imagine many different probability distributions on binary n-vectors and for simplicity, we assume that the marks are independent and modeled as Bernoulli random variables. So we are assuming the marks are independent given the hidden state of the HMM (note that this is not the same as assuming the marks are independent).

If there are n input marks, each state k has a vector $(p_{k1}, .., p_{kn})$ of probabilities of observing marks 1 to n. Since the probability is modeled as a set of independent Bernoulli random variables, the probability of observing a set of marks given that we are in the hidden state k equals the product of the probabilities of observing individual marks. For example if $n = 4$, the observed marks at bin j were $(1, 0, 1, 0)$ and we were in state k, then the likelihood of that data is $p_{k1}(1 - p_{k2})p_{k3}(1 - p_{k4})$.

The learned emission probabilities for the data are shown in Figure 7.

6.2.2 Transition Probabilities

Recall that the transition probabilities represent the frequency of transitioning from one hidden state to another hidden state. In this case, our hidden states are chromatin states. The transition matrix for our data is shown in Figure 8. As seen from the figure, the matrix is sparse, indicating that only a few of the possible transitions actually occur. The transition matrix reveals the spatial relationships between neighboring states. Blocks of states in the matrix reveal sub-groups of states and from these higher level blocks, we can see transitions between these meta-states.

6.3 Choosing the Number of states to model

As with most machine learning algorithms, increasing the complexity of the model (e.g. the number of hidden states) will allow it to better fit training data. However, the training data is only a limited sample of

Figure 7: Emission probabilities for the final model with 51 states. The cell corresponding to mark i and state k represents the probability that mark i is observed in state k.

the true population. As we add more complexity, at some point we are fitting patterns in the training data that only exist due to limited sampling, so that the model will not generalize to the true population. This is called **over-fitting** training data; we should stop adding complexity to the model before it fits the noise in the training data.

Bayesian Information Criterion (BIC) is a common technique for optimizing the complexity of a model that balances increased fit to the data with complexity of the model. Using BIC, we can visualize the increasing power of the HMM as a function of the number of states. Generally, one will choose a value for k (the number of states) such that the addition of more states has relatively little benefit in terms of predictive power gain. However, there is a tradeoff between model complexity and model interpretability that BIC cannot help with. The optimal model according to BIC is likely to have more states than an ideal model because we are willing to trade some predictive power for a model with fewer states that can be interpreted biologically. The human genome is so big and the chromatin marks so complex that statistically significant differences are easy to find, yet many of these differences are not biologically significant.

To solve this problem, we start with a model with more hidden states than we believe are necessary and prune hidden states as long as all states of interest in the larger model are adequately captured. The Baum-Welch algorithm (and EM in general) is sensitive to the initial conditions, so we try several random initializations in our learning. For each number of hidden states from 2 - 80, we generate three random initializations of the parameters and train the model using Baum-Welch. The best model according to BIC had 79 states and states were then iteratively removed from this set of 79 states.

As we mentioned earlier, Baum-Welch is sensitive to the initial parameters, so when we pruned states, we used a nested initialization rather than a randomized initialized for the pruned model. Specifically, states were greedily removed from the BIC-optimal 79 state model. The state to be removed was the state that such that all states from the 237 randomly initialized models were well captured. When removing a state, the emission probabilities would be removed and any state transitioning to the removed state would have that

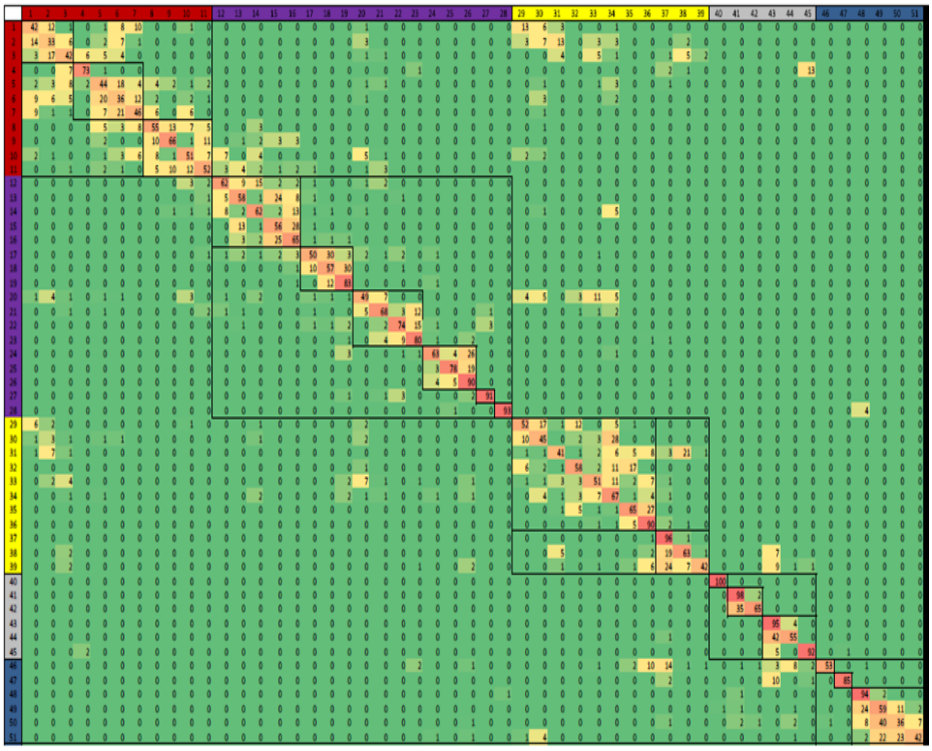

Figure 8: Transition probabilities for the final model with 51 states. The transition probability increases from green to red. Spatial relationships between neighboring chromatin states and distinct sub-groups of states are revealed by clustering the transition matrix. Notably, the matrix is sparse, so indicating that most are not possible.

transition probability uniformly redistributed to the remaining states. This was used as the initialization to the Baum-Welch training. The number of states for a model to analyze can then be selected by choosing the model trained from such nested initialization with the smallest number of states that sufficiently captures all states offering distinct biological interpretations. The resulting final model had 51 states.

We can also check model fit by looking at how the data violates model assumptions. Given the hidden state, the HMM assumes that each mark is independent. We can test how well the data conforms to this assumption by plotting the dependence between marks. This can reveal states that fit well and those that do not. In particular, repetitive states reveal a case where the model does not fit well. As we add more states, the model is better able to fit the data and hence fit the dependencies. By monitoring the fit on individual states that we are interested in, we can control the complexity of the model.

6.4 Results

This multivariate HMM model resulted in a set of 51 biologically relevant chromatin states. However, there were no one-to-one relationship between each state and known classes of genomic elements (e.g. introns, exons, promoters, enhancers, etc) Instead, multiple chromatin states were often associated with one genomic element. Each chromatin state encoded specific biological relevant information about its associated genomic element. For instance, three different chromatin states were associated with transcription start site (TSS), but one was associated with TSS of highly expressed genes, while the other two were associated with TSS of medium and lowly expressed genes respectively. Such use of epigenetic markers greatly improved genome annotation, particularly when combined with evolutionary signals discussed in previous lectures. The 51 chromatin states can be divided in five large groups. The properties of these groups are described as follows and further illustrated in 9:

1. **Promoter-Associated States (1-11):**

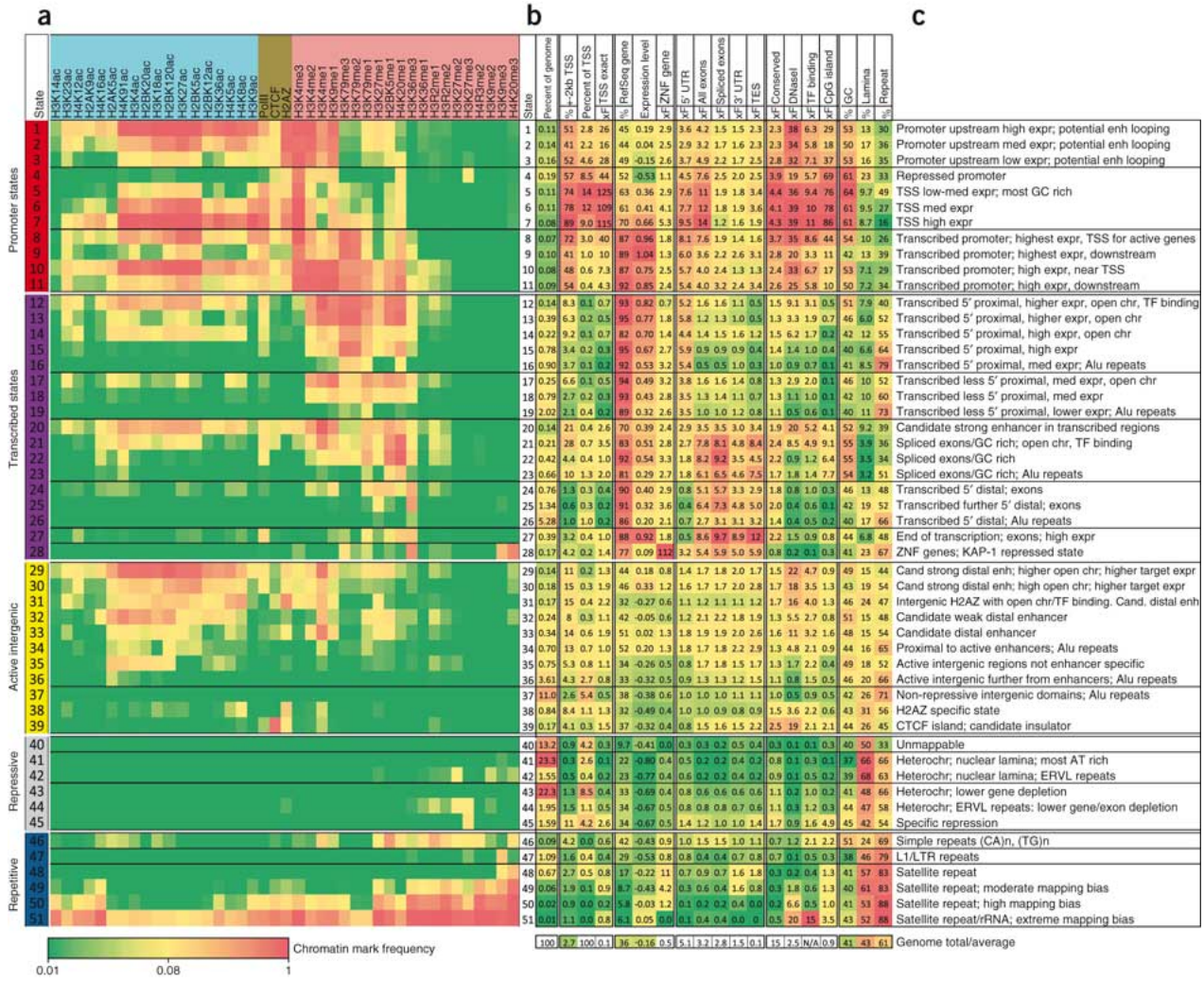

Figure 9: Chromatin state definition and functional interpretation. [3] a. Chromatin mark combinations associated with each state. Each row shows the specific combination of marks associated with each chromatin state and the frequencies between 0 and 1 with which they occur in color scale. These correspond to the emission probability parameters of the HMM learned across the genome during model training. b. Genomic and functional enrichments of chromatin states, including fold enrichment in different part of the genome (e.g. transcribed regions, TSS, RefSeq 5 end or 3end of the gene etc), in addition to fold enrichment for evolutionarily conserved elements, DNaseI hypersensitive sites, CpG islands, etc. All enrichments are based on the posterior probability assignments. c. Brief description of biological state function and interpretation (chr, chromatin; enh, enhancer).

These chromatin states all had high enrichment for promoter regions. 40-89% of each state was within 2 kb of a RefSeq TSS. compared to 2.7% genome-wide. These states all had a high frequency of H3K4me3, significant enrichments for DNase I hypersensitive sites, CpG islands, evolutionarily conserved motifs and bound transcription factors. However, these states differed in the levels of associated marks such as H3K79me2/3, H4K20me1, acetylations etc. These states also differed in their functional enrichment based on Gene Ontology (GO). For instance, genes associated with T cell activation were enriched in state 8 while genes associated with embryonic development were enriched in state 4. Additionally, among these promoter states there were distinct positional enrichments. States 1-3 peaked both upstream and downstream of TSS; states 4-7 were concentrated right over TSS whereas states 8-11 peaked between 400 bp and 1200 bp downstream of TSS. This suggests that chromatin marks

can recruit initiation factors and that the act of transcript can reinforce these marks. The distinct functional enrichment also suggests that the marks encode a history of activation.

2. **Transcription-Associated States (12-28):**

   This was the second largest group of chromatin states and included 17 transcription-associated states. There are 70-95% contained in annotated transcribed regions compared to 36% for rest of genome. These states were not predominantly associated with a single mark but rather they were defined by a combination of seven marks - H3K79me3, H3K79me2, H3K79me1, H3K27me1, H2BK5me1, H4K20me1 and H3K36me3. These states have subgroups associated with 5'-proximal or 5'-distal locations. Some of these states were associated with spliced exons, transcription start sites or end sites. Of interest, state 28, which was characterized by high frequency for H3K9me3, H4K20me3, and H3K36me3, showed a high enrichment in zinc-finger genes. This specific combination of marks was previously reported as marking regions of KAP1 binding, a zinc-finger specific co-repressor.

3. **Active Intergenic States (29-39):**

   These states were associated with several classes of candidate enhancer regions and insulator regions and were associated with higher frequencies for H3K4me1, H2AZ, several acetylation marks but lower frequencies of methylation marks. Moreover, the chromatin marks could be used to distinguish active from less active enhancers. These regions were usually away from promoters and were outside of transcribed genes. Interestingly, several active intergenic states showed a significant enrichment for disease SNPs, or single nucleotide polymorphism in genome-wide association study (GWAS). For instance, a SNP (rs12619285) associated with plasma eosinophil count levels in inflammatory diseases was found to be located in the chromatin state 33, which was enriched for GWAS hits. In contrast, the surrounding region of this SNP was assigned to other chromatin states with no significant GWAS association. This can shed light on the possible functional significance of disease SNPs based on its distinct chromatin states.

4. **Large-Scale Repressed States (40-45):**

   These states marked large-scale repressed and heterochromatic regions, representing 64% of the genome. H3K27me3 and H3K9me3 were two most frequently detected marks in this group.

5. **Repetitive States (46-51):**

   These states showed strong and distinct enrichments for specific repetitive elements. For instance, state 46 had a strong sequence signature of low-complexity repeats such as (CA)n, (TG)n, and (CATG)n. States 48-51 showed seemingly high frequencies for many modification but also enrichment in reads from non-specific antibody control. The model was thus able to also capture artifacts resulting from lack of coverage for additional copies of repeat elements.

Since many of the chromatin states were described by multiple marks, the contribution of each mark to a state was quantified. Varying subsets of chromatin marks were tested to evaluate their potential for distinguishing between chromatin states. In general, increasing subsets of marks were found to converge to an accurate chromatin state when marks were chosen greedily.

The predictive power of chromatin states for discovery of functional elements consistently outperformed predictions based on individual marks. Such unsupervised model using epigenomic mark combination and spatial genomic information performed as well as many supervised models in genome annotation. It was shown that this HMM model based on chromatin states was able to reveal previously unannotated promoters and transcribed regions that were supported by independent experimental evidence. When chromatin marks were analyzed across the whole genome, some of the properties observed were satellite enriched states (47-51) enriched in centromere, the zinc-finger enriched state (state 28) enriched on chromosome 19 etc. Thus, such genome-wide annotation based on chromatin states can help better interpret biological data and potentially discover new classes of functional elements in the genome.

6.5 Multiple Cell Types

All of the above work was done in a single cell type (CD4+ T cells). Since epigenomic markers vary over time, across cell types, and environmental circumstances, it is important to consider the dynamics of the chromatin states across different cell types and experimental conditions. The ENCODE project [2] in the Brad Bernstein Chromatin Group has measured 9 different chromatin marks in nine human cell lines. In this case, we want to a learn a single set of chromatin marks for all of the data. There are two approaches to this problem: concatenation and stacking. For concatenation, we could combine all of the 9 cell lines as if they were a single cell line. By concatenating the different cell lines, we ensure that a common set of state definitions are learned. We can do this here because the profiled marks were the same in each experiment. However, if we profiled different marks for different cell lines, we need to use another approach. Alternatively, we can align the 9 cell lines and treat all of the marks as a super-vector. This allows us to learn cell line specific activity states, for example there might be a state for ES-specific enhancers (in that state there would be enhancer marks in ES, but no marks in other cell types). Unfortunately, this greatly increases the dimension of the vectors emitted by the HMM, which translates to an increase in the model complexity needed to adequately fit the data.

Suppose we had multiple cell types where we profiled different marks and we wanted to concatenate them. One approach is to learn independent models and then combine them. We could find corresponding states by matching emission vectors that are similar or by matching states that appear at the same places in the genome. A second approach is to treat the missing marks as missing data. The EM framework allows for unspecified data points, so as long as pairwise relationships are observed between marks in some cell type, we can use EM. Lastly, we can predict the missing chromatin marks based on the observed marks using maximum-likelihood as in the Viterbi algorithm. This is a less powerful approach if the ultimate goal is chromatin state learning because we are only looking at the most likely state instead of averaging over all possibilities as in the second approach.

In the case with 9 marks in 9 hyman cell lines, the cell lines were concatenated and a model with 15 states was learned [4]. Each cell type was analyzed for class enrichment. It was shown that some chromatin states, such as those encoding active promoters were highly stable across all cell types. Other states, such as those encoding strong enhancers, were highly enriched in a cell-type specific manner, suggesting their roles in tissue specific gene expression. Finally, it was shown that there was significant correlation between the epigenetic marks on enhancers and the epigenetic marks on the genes they regulate, even though these can be thousands of base pairs away. Such chromatin state model has proven useful in matching enhancers to their respective genes, a problem that has been largely unsolved in modern biology. Thus, chromatin states provide a means to study the dynamic nature of chromatin across many cell types. In particular, we can see the activity of a particular region of the genome based on the chromatin annotation. It also allows us to summarize important information contained in 2.4 billion reads in just 15 chromatin states.

7 Current Research Directions

Several large-scale data production efforts such as ENCODE, modENCODE and Epigenome Radmap projects are currently in progress and therefore there are several opportunities to computationally analyze this new data.

Epigenomic data is also being used to study how behavior can alter your genome. There are studies being done that look at diet and exercise and their effects on disease susceptibility.

8 Further Reading

There are several interesting papers that are looking at chromatin states and epigenetics in general. I have listed several urls below to begin your exploration:

1. http://www.nature.com/nmeth/journal/v8/n9/full/nmeth.1673.html
2. http://www.nature.com/nature/journal/v473/n7345/full/nature09906.html
3. http://www.nature.com/nbt/journal/v28/n8/abs/nbt.1662.html

4. http://www.nytimes.com/2012/09/09/opinion/sunday/why-fathers-really-matter.html?_r=1

9 Tools and Techniques

ChromHMM is the HMM described in the text. It is available free for download with instructions and examples at: http://compbio.mit.edu/ChromHMM/.

Segway is another method for analyzing multiple tracks of functional genomics data. It uses a dynamic Bayesian network (HMMs are a particular type of dynamic Bayesian network) which enables it to analyze the entire genome at 1-bp resolution. The downside is that it is much slower than ChromHMM. It is available free for download here: http://noble.gs.washington.edu/proj/segway/.

10 What Have We Learned?

In this lecture, we learnt how chromatin marks can be used to infer biologically relevant states. The analysis in [3] presents a sophisticated method to apply previously learnt techniques such as HMMs to a complex problem. The lecture also introduced the powerful Burrows-Wheeler transform that has enabled efficient read mapping.

References

[1] Langmead B, C Trapnelli, M Pop, and S Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 10(3), 2009.

[2] Encyclopedia of dna elements.

[3] J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28:817–825, 2010.

[4] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh, L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[5] Elaine R Mardis. Chip-seq: welcome to the new frontier. *Nat Meth*, 4(8):614–614, 2007.

[6] Ncbi nih epigenomics.