Classification of Genes Using Clustering of Chromatin State Segmentations in Human Epigenomes

by

Nischay Kumar

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Masters of Engineering in Computer Science and Engineering at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

© Massachusetts Institute of Technology 2013. All rights reserved.

Author

Department of Electrical Engineering and Computer Science

May 24, 2013

Certified by

Manolis Kellis
Associate Professor
Thesis Supervisor

Accepted by

Dennis M. Freeman
Chairman, Masters of Engineering Thesis Committee
Classification of Genes Using Clustering of Chromatin State Segmentations in Human Epigenomes

by

Nischay Kumar

Submitted to the Department of Electrical Engineering and Computer Science on May 24, 2013, in partial fulfillment of the requirements for the degree of Masters of Engineering in Computer Science and Engineering

Abstract

Combinatorial patterns of chromatin marks have been shown to play a significant role in gene regulation activities by changing the landscape of the DNA through chemical means. Recent work has expanded on this observation using ChIP-seq signals of chromatin marks and supervised algorithms to build gene expression prediction models based on correlation analysis. However, no approach to date has attempted to use chromatin states to identify various classes of genes outside of the high-low expression classes. This research aims to fill this void by utilizing chromatin state segmentation and RNA-seq expression datasets from the NIH Roadmap Epigenomes project. A gene classification model was built using a k-fuzzy clustering approach of chromatin state features from a subset of training genes and then applied to a larger test set of genes. The models were found to be robust and show striking correspondence between training and test sets. 8 classes of genes that represent silent, repressed, and subsets of actively transcribed genes were identified and several metrics to validate the classes were computed. The systematic analysis outlined in this research is shown to be promising approach for gene classification and future de novo discovery of gene like regions.

Thesis Supervisor: Manolis Kellis
Title: Associate Professor
Acknowledgments

First and foremost I would like to acknowledge Professor Kellis for his support during the course of the MEng and his comments and questions during this project, which helped guide the research. I would also like to thank Anshul Kundaje, Wouter Meuleman, Matthew Eaton, and Jianrong Wang for their expertise in the field and insightful comments on how to improve the substance of the research. Lastly, I would like to thank the entire Computational Biology group for being a great group of people to interact with and learn from during my MEng.
Contents

1 Introduction 13
   1.1 Chromatin Marks .................................. 13
   1.2 Chromatin States .................................. 16

2 Related Works 21
   2.1 Gene Expression ................................... 21
   2.2 Focus of this Research: Gene Classification ........... 22

3 Experimental Methods 25
   3.1 NIH Roadmap Epigenomes Data ........................ 25
   3.2 Algorithmic Implementation .......................... 26
   3.3 Feature Vector Construction .......................... 27
   3.4 K-Fuzzy Cluster .................................... 28
      3.4.1 Training ........................................ 28
      3.4.2 Analysis ........................................ 30

4 Experimental Results and Discussion 31
   4.1 Final Model Selection ................................ 31
   4.2 k=6 Cluster Model ................................... 32
      4.2.1 Epigenome Specific Trends ....................... 32
      4.2.2 General Trends ................................ 35

5 Future Work 45
   5.1 Summary of Results ................................ 45
5.2 Future Work ............................................. 45

5.2.1 Expansion of Feature Vectors ...................... 46

5.2.2 Other Functional Enrichments ....................... 46

5.2.3 De Novo Discovery of Genes ......................... 46
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Packaging scheme for DNA inside Eukaryotic cells [1]</td>
<td>14</td>
</tr>
<tr>
<td>1-2</td>
<td>HMM scheme for discovering chromatin states in genome</td>
<td>17</td>
</tr>
<tr>
<td>1-3</td>
<td>chromHMM segmentation and raw chromatin signals for CAPZA2 gene</td>
<td>19</td>
</tr>
<tr>
<td>2-1</td>
<td>Caenorhabditis elegans gene expression prediction results</td>
<td>22</td>
</tr>
<tr>
<td>3-1</td>
<td>Pipeline of k-fuzzy cluster training and analysis</td>
<td>27</td>
</tr>
<tr>
<td>3-2</td>
<td>Heatmap of transition matrices from ChromHMM on left and E10 feature vectors on right.</td>
<td>29</td>
</tr>
<tr>
<td>4-1</td>
<td>PDIfF and WSS metrics for cluster models with training set across all epigenomes</td>
<td>37</td>
</tr>
<tr>
<td>4-2</td>
<td>PDIfF and WSS metrics for cluster models with test set across all epigenomes</td>
<td>38</td>
</tr>
<tr>
<td>4-3</td>
<td>Cluster centers for k=6 cluster model for epigenomes E09 and E19</td>
<td>39</td>
</tr>
<tr>
<td>4-4</td>
<td>Cluster specific chromatin state distributions in train and test set genes for epigenome E23</td>
<td>40</td>
</tr>
<tr>
<td>4-5</td>
<td>Cluster specific expression distributions in train and test set genes for epigenome E23</td>
<td>41</td>
</tr>
<tr>
<td>4-6</td>
<td>Cluster specific scatter plot of distance from cluster center for epigenome E23</td>
<td>42</td>
</tr>
<tr>
<td>4-7</td>
<td>Heatmap of class metrics and functional enrichments across epigenomes</td>
<td>43</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Chromatin marks and their functionalities [6] ........................................ 16

3.1 Epigenomes studied in this research and their corresponding labels .......... 26

4.1 Annotated gene classes and their chromatin state associations .............. 33

4.2 Gene class fold enrichments and percentages over background (complete gene set) averaged over all epigenomes .......................... 36
Chapter 1

Introduction

The genomic regulatory network is a complex system of dynamic machinery and structures that use unique and efficient methods in order to regulate gene transcription and translation activity. The network relies on proximal and distal relationships between functional elements, such as those between gene promoters and enhancers, as well as temporal relationships, such as the control process to initiate gene translation, in order to coordinate such a vast system. More than the actual proteins produced or the chaotic organization of this network, one can’t help but find the hidden language of this network to be fascinating. The physical and chemical alterations of the moving components and landscape serve as a code to other members of the network and determine the production of the necessary proteins.

1.1 Chromatin Marks

In order to overcome space constraints, the DNA in Eukaryotic cells undergoes a complex and efficient packaging formation to reduce its footprint. The process behind this packaging scheme is shown in Figure 1-1 below. The DNA is first wrapped at 147 base pair(bp) intervals around 4 core histone proteins, H2A, H2B, H3, and H4, which form an octet structure known as a nucleosome. The nucleosomes are then compacted further into chromatin fibers, which form the basis of chromosomes [2]. As seen in the schematic, this packaging scheme leads to two different types of chro-
matin: heterochromatin or "closed" chromatin and euchromatin or "open" chromatin [3]. The "openess" or physical accessibility of a DNA regions will affect whether or not transcriptional regulators like activators and repressors will be able to transcribe or replicate the gene. In addition, the amino-terminal tails of the histone complexes are open to hundreds of post-translation modifications, such as methylation, acetylation, phosphorylation, ubiquitination, by various epigenetic factors. For example, H3K4me3 is a histone modification in which the fourth lysine in the H3 histone is trimethylated or H3K27me1 is a histone modification in which the twenty-seventh lysine in the H3 histone is methylated. Through chemical means, these histone modifications or chromatin marks affect the "openess" of a DNA region by changing the conformations of the local chromatin landscape and providing binding surfaces for activators and repressors [4].

![Figure 1-1: Packaging scheme for DNA inside Eukaryotic cells](image)

Strahl and Allis proposed that a histone language encoded through a combinatorial pattern of chromatin marks was read by other proteins in the genomic network.
However, it was unclear whether multiple chromatin marks appeared on the same tails and how these interactions were established. Strahl and Allis postulated that the modification activity of one enzyme would influence the activity of others that followed [2]. They thought that methylation or phosphorylation at a given site influenced the next enzyme's ability to modify or lack the ability to modify the histone tail with another epigenetic factor. The fundamental view that these chromatin marks existed in pairs and the absence or presence of one modification dictated the absence or presence of others forms the basis modern epigenomics research.

Strahl and Allis also proposed that these chromatin marks played the part of recruiters in attracting other complexes to the region for important biological processes such as transcription or replication. Experiments conducted in yeast had shown proteins involved in the repression of gene transcription, such as the Sir3 and Sir4 proteins, were binding to the tails of the H3 and H4 histone proteins. Several large scale, coordinated research projects have validated this hypothesis. The modEncode Consortium performed a study in order to better understand and annotate the chromatin landscape of Drosophila melanogaster. Chromatin analysis requires using chromatin immunoprecipitation (ChIP), where antibodies specifically target histone modifications and and cross-react. The ChIP is followed by ChIP-seq where all regions that reacted with the antibodies are sequenced and aligned to the genome [8]. They found that sequences in the genome acted as poor indicators for functional elements, whereas histone modifications served as great markers. For example, transcription start site (TSS) regions were lower in nucleosome density and were enriched for H3K4me3 and the transcribed regions of genes were enriched for H3K36me3 [5]. Table 3.1 below shows some of the commonly studied chromatin marks and their functionality in cellular activities. As a result of this research, histones are no longer thought of as the spools which DNA wraps around, but rather as a support structure which directs genomic activity. Chromatin marks are now thought to play significant roles in regulation of genes and as a result affect phenotypes and the onset of various diseases [7].
Table 1.1: Chromatin marks and their functionalities [6]

<table>
<thead>
<tr>
<th>Chromatin Mark</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3K4me1</td>
<td>Associated with enhancers and other distal elements</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Associated with promoters and enhancers</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Associated with promoters and TSS</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Associated with promoters</td>
</tr>
<tr>
<td>H3K9me1</td>
<td>Associated with 5' end of genes</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>Repressive mark associated with heterochromatin</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Associated with active regulatory elements</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Repressive mark associated with polycomb complexes</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>Elongation mark associated with transcribed parts of genes</td>
</tr>
<tr>
<td>H3K79me2</td>
<td>Transcription mark associated with 5' end of genes</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>Associated with 5' end of genes</td>
</tr>
</tbody>
</table>

1.2 Chromatin States

As mentioned earlier, the histone code, which states that the combinatorial pattern of chromatin marks act as biological markers in gene regulation, has formed the basis of present day epigenomics research. Researchers are still trying to uncover the true potential behind the association of combinatorial patterns of chromatin marks and functional elements in the genome. One of the most pervasive techniques currently used is segmenting the genome into chromatin states with multivariate Hidden Markov Models (HMMs), which in an unsupervised manner learns the spatial domains of operation for the chromatin marks. The HMM represents the presence or absence of the selected chromatin marks in windowed regions of the genome as observations to the algorithm to learn a probabilistic graph model. The graph is a representation in which the genome transitions between hidden states that generate our observed chromatin mark pattern. An example of the HMM scheme is shown in Figure 1-2 below.

Early chromatin state analysis was applied to model organisms such as Drosophila melanogaster and Caenorhabditis elegans, which have a great signal to noise ratio in their genomes since evolutionary pressures have kept their genomes compact. The modENCODE consortium’s study on Drosophila built a 9-state and 30-state HMM
model to identify broad regions of chromatin states as well as more detailed regions for functional element enrichment. Most researchers agree with Jason Lieb’s viewpoint: "There's not a magic number of states. The whole point of these is just to distill down the data into something that's interpretable" [3]. The smaller state model allowed them to observe that intergenic regions and silent genes were associated with a single state covering nearly 50% of the genome and lacking enrichment for many active marks. Regions enriched for active genes showed more complex biological functionality. TSS proximal regions were covered by states enriched in active promoter marks such as H3K4me3 and H3K9ac while other transcribed regions were covered
by states enriched for the elongation mark H3K36me3. With the 30 state model, the researchers were able to analyze functional elements such as regulatory motifs, which were found to be enriched in active marks and depleted in repressive marks. The researchers felt that this chromatin-centric view of the genome allowed for more robust predictions of functional elements such as transcription factors (TFs), compared to previously studied approaches. A study done by Kharchenko, et al verified the results found by modENCODE and built upon them too. They also found that chromatin state analysis separated repressive Polycomb domain genes into separate classes based on their chromatin state enrichment [8]. Some of the classes were enriched for the repressive mark H3K27me3, whereas others were enriched for the active marks H3K4me1/me2. They hypothesized that these gene classes represented various functional classes such as repressed, paused or transcribed genes or represented classes of regulatory genes that switched states due to environmental factors.

Ernst and Kellis built an innovative ChromHMM software package that binarizes the chromatin signals in windowed regions of the genome, leading to a simpler to interpret and more robust HMM. They applied chromHMM to human T cells and their results indicated a strong ability to discern noise from signal in the fairly large human genome. Figure 1-3 below shows their segmentation for the CAPZA2 gene located in chromosome 7 of the human T cell. Even though there is a lot of variation in the signal intensity and background noise in the signals, the binarization scheme is robust and creates continuous segments of chromatin annotated regions without overfitting. For example, at the beginning and end of the gene chromHMM picks up the H3K27me3 signal and labels the regions as repressed. The model also segments the functional elements such as the promoters upstream and downstream of the gene. Ernst and Kellis also found that using chromatin state segmentations resulted in better prediction performance of functional elements than using individual chromatin signals. In addition, the combination of active and inactive states allowed them to learn about validated and candidate functional elements across boundaries. For the reasons stated above and several others, chromatin state segmentations have become a fundamental tool in gene analysis studies.
Figure 1-3: chromHMM segmentation and raw chromatin signals for CAPZA2 gene [1]
Chapter 2

Related Works

Research studies had proven that chromatin state segmentations were a great tool for de novo annotation of the genome and validation and discovery of functional elements. They also showed it was possible to discern which of the functional elements of the genome were active or repressive based on the combinatorial patterns of the chromatin marks in the vicinity. Would it then be possible to predict if a gene was active or repressed—predict its expression values—based on the chromatin signals over the length of the gene body?

2.1 Gene Expression

Gerstein, et al. built a gene expression model for Caenorhabditis elegans by dividing the 4kb region flanking the transcription start site (TSS) and transcription termination site (TTS) into 100 bp bins and taking the average of the chromatin signal [9]. They used a support vector regression and found their predicted expression values to have a 0.75 correlation coefficient with the actual expression values. They also found that while proximity to the TSS and TTS led to higher correlation values between the expression value and average chromatin signal, the predictive capability of the chromatin signal extended as far 4kb flanking the anchors. The results for their research on gene expression prediction are shown in Figure 2-1 below.

Cheng, et al. also followed a similar binning scheme in analyzing modENCODE
Figure 2-1: Caenorhabditis elegans gene expression prediction results

datasets [4]. Their support vector regression model also generated a 0.75 correlation between predicted and actual expression values. They also found that they were able to account for 50% of the variation in gene expression. The study also built a simple linear regression model with singleton terms and a linear regression model with interaction terms. They found that the model with the interaction terms increased the prediction accuracy by 4%, once again supporting the histone code and the significance of the combinatorial pattern of chromatin marks. Lastly, the model was applied to cells from various developmental stages, which showed that models trained on a specific cell line had a 0.1 reduction in AUC when applied to other cell lines. Dong, et al. found that the deterioration in model performance was even worse when some of the cells were undifferentiated and others were committed cells [10]. This was due to the fact the genes in undifferentiated cells became paused— they had active promoters, but were repressed for the rest of the gene— before they became differentiated.

2.2 Focus of this Research: Gene Classification

In addition to creating expression models in the study, Cheng, et al performed two way hierarchical clustering on chromatin features and annotated genes [4]. They found two overall clusters that separated the genes into a low and high expression clustering. The high expression cluster was enriched for the active transcriptional elongation mark H3K36me3 whereas the low expression cluster was enriched for repressive mark
H3K9me3.

This research aims to build on the approach taken on by Cheng, et al and use clustering methods to create classes of genes across various epigenomes for the NIH Roadmap Epigenomes project. Annotated chromatin state segmentations were used in this research rather than the raw chromatin signals, which makes the final model more robust and increases its interpretability. To further the understanding of the biological significance of the gene classes, several metrics were analyzed including the cluster specific expression distribution, protein gene enrichment, and pseudogene enrichment. The research also annotates these gene classes across epigenomes and provides a higher level picture of the operation of genes. The methodology provided in this paper can be a great complement to existing gene classification methods and can lead to de novo discovery of possible gene candidates. It can also lead to a new method of classifying epigenomes based on the breakdown of gene classes within each epigenome.
Chapter 3

Experimental Methods

3.1 NIH Roadmap Epigenomes Data

As part of the Roadmap Epigenomes initiative, the NIH has taken on a large scale study to systematically analyze chromatin signals and expression data across over 90 epigenomes. ChIP-seq signals for a multitude of marks were available for each of the epigenomes, but only the core histone marks: H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3 were available across all of the epigenomes. Using these histone marks, a joint HMM with 25 states was trained across all of the epigenomes using the ChromHMM package. The transition matrix of chromatin states is representative of transitions across all of the epigenomes, providing a much more robust and higher level representation of state transitions in human cell lines. The state labels were annotated according to their biological enrichments to determine active, repressed, quiescent and other various states. RNA-seq expression data was also provided for over 51,000 genes, unfortunately this was only available for 26 specific epigenomes. In order to maximize the potential to understand the biological significance of the models in these study, the analysis was limited only to those 26 epigenomes that had RNA-seq expression data. The epigenomes used in this study are shown below in Table 3.1.
Table 3.1: Epigenomes studied in this research and their corresponding labels

<table>
<thead>
<tr>
<th>Epigenome Label</th>
<th>Cell Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>E01</td>
<td>Breast</td>
</tr>
<tr>
<td>E07</td>
<td>H1 BMP4 Derived Trophoblast Cultured Cells</td>
</tr>
<tr>
<td>E08</td>
<td>H1 Derived Mesenchymal Stem Cells</td>
</tr>
<tr>
<td>E09</td>
<td>H1 Derived Neuronal Progenitor Cultured Cells</td>
</tr>
<tr>
<td>E10</td>
<td>H1 Cell Line</td>
</tr>
<tr>
<td>E17</td>
<td>Mobilized CD34 Primary Cells</td>
</tr>
<tr>
<td>E19</td>
<td>Penis Foreskin Fibroblast Primary Cells - Donor 1</td>
</tr>
<tr>
<td>E20</td>
<td>Penis Foreskin Fibroblast Primary Cells - Donor 2</td>
</tr>
<tr>
<td>E21</td>
<td>Penis Foreskin Keratinocyte Primary Cells - Donor 2</td>
</tr>
<tr>
<td>E22</td>
<td>Penis Foreskin Melanocyte Primary Cells - Donor 1</td>
</tr>
<tr>
<td>E23</td>
<td>Adult Liver</td>
</tr>
<tr>
<td>E24</td>
<td>Brain Germinal Matrix</td>
</tr>
<tr>
<td>E25</td>
<td>Brain Hippocampus Middle</td>
</tr>
<tr>
<td>E26</td>
<td>Breast Myoepithelial Cells</td>
</tr>
<tr>
<td>E29</td>
<td>CD4 Memory Primary Cells</td>
</tr>
<tr>
<td>E30</td>
<td>CD4 Naive Primary Cells</td>
</tr>
<tr>
<td>E31</td>
<td>CD8 Naive Primary Cells</td>
</tr>
<tr>
<td>E36</td>
<td>Fetal Brain - Donor 1</td>
</tr>
<tr>
<td>E37</td>
<td>Fetal Brain - Donor 2</td>
</tr>
<tr>
<td>E38</td>
<td>hESC Derived CD184+ Endoderm Cultured Cells</td>
</tr>
<tr>
<td>E41</td>
<td>Neurosphere Cultured Cells Ganglionic Eminence Derived</td>
</tr>
<tr>
<td>E42</td>
<td>Penis Foreskin Keratinocyte Primary Cells - Donor 3</td>
</tr>
<tr>
<td>E43</td>
<td>Penis Foreskin Melanocyte Primary Cells - Donor 2</td>
</tr>
<tr>
<td>E44</td>
<td>Penis Foreskin Melanocyte Primary Cells - Donor 3</td>
</tr>
<tr>
<td>E50</td>
<td>H1 BMP4 Derived Mesendoderm Cultured Cells</td>
</tr>
<tr>
<td>E51</td>
<td>Neurosphere Cultured Cells Cortex Derived</td>
</tr>
</tbody>
</table>

3.2 Algorithmic Implementation

In order to train the cluster models and evaluate their biological significance, the pipeline shown in Figure 3-1 below was implemented. The pipeline begins with the feature construction from the chromHMM segmentation data. The feature instances were split into a training and test set for clustering. K-fuzzy clustering was then performed on the training set and then test instances were fit to the cluster centers. Lastly, the biological significance of the clusters was evaluated using expression data and other enrichments.
3.3 Feature Vector Construction

Using the Gencode Version 10 annotations, the TSS and TTS of all of the genes in human epigenomes were extracted. The chromatin state distribution along the body of every gene needed to be extracted in order to build the feature vectors. The kentutils bigWigSummary function, recommended by Dong et al. [10], is a utility which allows you to very quickly index a start and end point in a bigWig file and extract the relevant labels. A Python script, written as a wrapper for the bigWigSummary function, extracted the chromatin state labels for every gene in order to compute the feature vector instance. Two different feature vector constructions were experimented with before the final feature vector set was chosen. The first feature vector set is described by Equation (3.1). Feature vector $X_{i,g}$ for a given gene $g$ was represented by a normalized counted of the occurrences of state $i$. $P(state = i)$ was represented by the prior or background distribution over state $i$ over the entire genome. This feature vector construction limited $X_{i,g}$ to values between 0 and 1 and not be biased by the length of a given gene.

$$X_{i,g} = \frac{P(state = i) \cdot \text{Counts}(state = i|gene = g)}{\sum_{j=1}^{25} P(state = j) \cdot \text{Counts}(state = j|gene = g)} \quad (3.1)$$

However, this approach was found to be lacking one of the more significant aspects of a chromatin state model—the transitions between $state_i \rightarrow state_j$. Consequently, the feature vector set was reformulated to that described by Equation (3.2). This
formulation creates $25 \times 25$ features vectors, which were normalized using the prior or background distribution over all $state_i \rightarrow state_k$ transitions. The prior distribution was obtained using the state transition matrix learned from ChromHMM. Once again, this feature vector construction limited $X_{i,j,g}$ to values between 0 and 1 and not be biased by the length of a given gene.

$$X_{i,j,g} = \frac{P(state = i \rightarrow state = j) \cdot Counts(state = i \rightarrow state = j|gene = g)}{\sum_{k=1}^{25} P(state = i \rightarrow state = k) \cdot Counts(state = i \rightarrow state = k|gene = g)}$$

With this prior formulation however, the diagonal elements of the transition matrix were found to be dominating the feature vector set. The HMM model is fairly robust and creates long continuous segments of a given $state_i$, so many self loop or $state_i \rightarrow state_i$ transitions occur. In order to limit the dominance of this effect, the diagonal terms or self loops were removed from the feature vector set and the normalized state counts previously described in Equation (3.1) were added in. The difference in the feature vector relative weights is shown in Figure 3-2. The heatmap on the left represents the background state transition distribution learned from ChromHMM on genomes from the 90 epigenomes. The heatmap on the right represents the state transition distribution learned from the feature vectors of all genes in epigenome E10. The state transition counts over all genes: $\sum_{g} Counts(state = i \rightarrow state = j|gene = g)$ were added up and normalized according to Equation (3.2). As you can see the diagonal terms are 0 in the right matrix, allowing us to pick up the more subtle transitions such as the E11.EnhWk1 transition to E24.Quies3.

### 3.4 K-Fuzzy Cluster

#### 3.4.1 Training

In order to learn how representative a clustering based on sub sample of genes would be of the entire gene set, a training set of 10000 randomly selects genes were used
Figure 3-2: Heatmap of transition matrices from ChromHMM on left and E10 feature vectors on right.

to build the model and then the remaining test set of genes were fit to that model. Also, this methodology would allow de novo learning of different classes of gene like regions in the genome by fitting the regions to the trained cluster model. This would be a very promising and exciting avenue that needs to be explored as well.

K-fuzzy clustering was performed for all values of k from 3 to 10 for every epigenome with 20 different model runs. The final selected for analysis was based on a few metrics. The sum of the within cluster sum of squares (WSS) for each clustering, defined by Equation (3.3) was computed as one confidence metric. It was observed that for some of the lower value k models such as k= 4, the k-fuzzy probabilities of cluster assignment for some genes were nearly uniform. In order to pick a model where the k-fuzzy probabilities and the confidence in the clustering were maximized, the probability differential metric shown in Equation (3.4) was utilized.
\[ WSS = \sum_{c=1}^{NumClusters} \sum_{g \in label_c} |X_g - \mu_c|^2 \quad (3.3) \]

\[ PDIFF = \frac{\sum_{g \in Genes} [\max(P(label_c|g, \mu_c)) - 2^{nd} \max(P(label_d|g, \mu_d))]^2}{|Genes|} \quad (3.4) \]

### 3.4.2 Analysis

After selecting the best clustering model, \( P(label_c|g, \mu_c) \) was computed for all \( genes \in Test.Genes \). The argmax of the conditional label probabilities was chosen as the label for a given gene. The WSS and and PDIFF of the test set were computed to compare to the values obtained from the training set and to determine how well the training model represents the entire gene set.

A few functional enrichments were determined for the genes in each of the clusters. An RNA-seq expression density curve was determined for each of the cluster models by intersecting the gene clusters with their respective RNA-seq expression values. A similar enrichment was run for gene types to determine if there was an enrichment for protein-coding genes or pseudogenes in a specific gene class. The lengths of all genes were normalized and an average chromatin state distribution along the length of the normalized genes was determined for each cluster. The cluster labels output by the k-fuzzy algorithm in R have no biological significance and are not uniform across epigenomes. For example, a cluster labeled 1 in epigenome E10 may represent a repressed group of genes whereas in epigenome E23 it may represent a highly expressed group of genes. Consequently, the previous functional enrichments were used to manually annotate all clusters across epigenomes.
Chapter 4

Experimental Results and Discussion

4.1 Final Model Selection

As discussed in Chapter 3, the metrics WSS (3.3) and PDIFF (3.4) to gauge the confidence in a given k-clustering model. Figure 4-1 below shows a plot of these metrics for the various cluster models run across all epigenomes for the training set of genes. The PDIFF metric converges fairly quickly for a given epigenome as the value of k is increased, which intuitively makes sense since as the number of clusters increases the model begins to overfit to the data and create very specialized clusters. The WSS metric doesn’t converge nearly as quickly to a final value for a given epigenome as k is increased. The same metrics were plotted for the various cluster models run across all epigenomes for the test set of genes, which is composed of 41761 genes or a little more than four times the size of the train set. The plots are shown in Figure 4-2 below. The trends observed in the train set are exhibited strongly in the test set as well. The PDIFF values are all roughly 0.1 higher in the test set exhibiting a 16% increase averaged across all epigenomes and cluster models. The WSS values are higher but that metric is a summation rather than an average so it should be expected to be roughly four times larger due to the larger test set of genes.
Another thing to point out is that there are a few outlier epigenomes for each metric in both the train and test sets. For example, epigenome E09 has a PDIFF metric significantly greater than the average, whereas epigenome E19 has a PDIFF metric significantly less the average. The same outlier observations can be made about the WSS of the two epigenomes with respect to the average WSS observed across epigenomes. In order, to further understand this various plots of the two epigenomes were analyzed and some interesting observations were noted. As seen in Figure 4-3, the training genes for E09 on the top show very little variability in their cluster centers compared to E19 on the bottom. This result is due to the lack of variability in chromatin states in the E09 epigenome. The "Gene Feats" subplot was calculated by taking the mean of the features for the entire training set, representing a clustering with $k=1$. It is evident that epigenome E09 is significantly enriched for the E25.K9acLow state whereas E19 shows variable enrichment with some Quies, ReprPCWk and Tx states.

Individual epigenomes exhibited variation in the cluster metrics analyzed in this research due to chromatin state bias, therefore the optimal $k$ was chosen based on the average of metrics across all epigenomes. For the training set, the average PDIFF exhibited a 27% increase from $k=3$ to $k=6$, but less than a 1% increase from $k=6$ to $k=10$. For the test a similar observation was noticed, the average PDIFF showed a 19% increase from $k=3$ to $k=6$, but less than 0.5% from $k=6$ to $k=10$. The WSS also reduced by 50% in both sets from $k=3$ to $k=6$. Based on this change in metrics, especially the PDIFF metric derived from the k-fuzzy cluster assignment probabilities, and the functional enrichments the final model chosen was the $k=6$ model.

4.2 k=6 Cluster Model

4.2.1 Epigenome Specific Trends

Having selected the final cluster model, the task of annotating the cluster labels was undertaken. Although there were only 6 clusters in each model, the final annotated
classes of genes contained 8 unique classes. The classes of genes and their functional enrichments are shown below in Table 4.1 The colors of the gene shown in the table are part of the annotations and will be the colors used in any future plots or table referenced to those states.

Table 4.1: Annotated gene classes and their chromatin state associations

<table>
<thead>
<tr>
<th>Label</th>
<th>Gene Class</th>
<th>Associated Chromatin States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sil1</td>
<td>Silent</td>
<td>Quies1</td>
</tr>
<tr>
<td>Sil2</td>
<td>Silent 2</td>
<td>Quies2</td>
</tr>
<tr>
<td>Sil3</td>
<td>Silent 3</td>
<td>K9acLow</td>
</tr>
<tr>
<td>Rep</td>
<td>Repressed</td>
<td>ReprPCWk</td>
</tr>
<tr>
<td>Het</td>
<td>Pseudogenes</td>
<td>Het</td>
</tr>
<tr>
<td>Weak</td>
<td>Weakly Expressed/Paused</td>
<td>Quies3, TxWk</td>
</tr>
<tr>
<td>Mod</td>
<td>Moderately Expressed</td>
<td>TxWk, Tx, Quies3</td>
</tr>
<tr>
<td>High</td>
<td>Highly Expressed</td>
<td>TxWk, Tx</td>
</tr>
</tbody>
</table>

The cluster specific distributions for every epigenome were computed by normalizing the lengths of all genes from TSS to TTS to 1000 bins and then using the kentutils bigWigSummary function to extract out the chromatin states for each bin for every gene. The value for \( state_i \) in \( bin_j \) is given by the formula in Equation 4.1. If the bin does not cross any chromatin state boundaries for a given gene, then a single \( state_i \) will have a value of 1 whereas if multiple states exist then the sum of their values will add up to 1. This resulted in a chromatin state distribution feature for each cluster of genes, which was then averaged across all the genes in that cluster according to the formula in Equation 4.2. Figure 4-4 below shows the cluster specific chromatin state distributions for the train and test set of epigenome E23. For clarity of visualization 9 of the more interesting states were selected for the plot. The plot illustrates all of the chromatin state-class associations shown in Table 4.1. For example, the Sil1 class shows heavy enrichment of the Quies1 state and the Rep class shows heavy enrichment of the ReprPCWk state. A few more of the subtle associations are also visible, for example the Weak, Mod, and High classes all have enrichment of the TssA state, presumably associated with the promoters at the beginning of these expressed genes.
The Weak class also shows heavy enrichment of Quies3 which may be indicative of paused genes. Pol2 binding site data would be required to verify this fact, but the further evidence to support this claim is provided below. In addition, as discussed by Cheng et al, the negative or silencing states tend to have more uniform distributions over the length of the gene whereas the positive or transcription states tend to show variability [4].

\[ B_{i,j,g} = \frac{\text{Counts}(\text{state}_i|\text{bin}_j, \text{gene}_g)}{|\text{bin}_{j, \text{in} \_ \text{gene}_g}|} \]  

\[ B_{i,j,c} = \frac{\sum_{\text{gene}_g \in \text{label}_c} B_{i,j,g}}{|\text{gene}_g \in \text{label}_c|} \]  

Figure 4-5 below shows the cluster specific expression distributions for the train and test genes of E23. The expression distributions correspond well with the annotate gene classes. The Sil1, Sil2, and Rep classes of genes all have greater than 0.7 probability mass at 0 expression. Consequently, they have very low means to their gene expression distribution and fairly low variances as well. The Weak, Mod, and High classes all have less than 0.4 probability mass at 0 and exhibit fairly high means. One point to note is the high variance in the Mod class expression distribution, which corresponds with the variable chromatin state distribution shown in Figure 4-4 and multiple centers of low weight in Figure 4-3. To inspect this difference in variability even further, a scatter plot of the distance of each gene from the cluster center versus a labeled gene number was created. The scatter plot, shown in Figure 4-6 does in fact show that genes in the Mod class exhibit the highest cluster variability, whereas the Sil1 and Sil2 classes exhibit very low variance. What is even more interesting is that the Weak and High class exhibit very low variance relative to the Mod class. This is perhaps indicating some of the mechanisms behind gene expression. Dong, et al. discussed in their research that the chromatin features involved in determining if genes are expressed are different from those that determine the level of expression. Moderately expressed genes have multiple chromatin marks in their vicinity which may act in a sort of checks and balances methodology to keep the genes expressed but
prevent them from being on the low or high end of the expression spectrum. These genes could also be dynamic in their expression, sometimes being weakly expressed and other times highly expressed due to the abundances of chromatin marks that can change the physical and chemical openness of the DNA.

The other noticeable fact is the relative sizes of each class, the Sil1 class appears to be the largest class of genes in this epigenome and the Mod class appears to be the next largest class. This could be due to the fact that the E23 epigenome, adult human liver, has become differentiated so a smaller subset of specialized genes are now only being expressed. Further gene ontology enrichments for specific functionality in epigenomes would be required to verify this fact. Another remarkable fact is the correspondence between the train set and the test set. Using only 1/5 of the genes available, a robust clustering was created that translates extremely well to the test set visually and in terms of the metrics defined. This is a very promising approach in terms of categorizing genes and even epigenome types based on a very a limited subset of chromatin data available.

4.2.2 General Trends

To further understand the differentiation between the various classes, the above functional enrichments were observed across all epigenomes for each class type. Figure 4-7 shows a heatmap of several of the differentiating metrics between classes discussed in 4.2.1. The expression, protein-coding gene and pseudogene fold enrichment over the entire epigenome and percentage gene set size and WSS with respect to the total sum were calculated for each epigenome. The heatmap color scale is normalized by column to allow illustration of enrichment of a certain feature in a class relative to the other classes. The trend in class sizes that we noticed in the E23 epigenome is observed across all epigenomes: the Sil1 class is nearly always the largest class in the epigenome. The only exception is the E09 epigenome, which discussed earlier, has a remarkable 72% of genes in the Sil3 class enriched for the K9acLow chromatin state. As expected, the gradient in expression strength is observed from Weak to High class expression. The silent classes show nearly no expression in any epigenomes. The
Mod class as discussed before exhibit the largest variability in WSS. The annotation of the classes also matches the types of genes that would be expected in the relevant classes. The results discussed above are averaged over all epigenomes and summarized in Table 4.2 below.

Table 4.2: Gene class fold enrichments and percentages over background (complete gene set) averaged over all epigenomes

<table>
<thead>
<tr>
<th>Label</th>
<th>Size of Gene Set (%)</th>
<th>Expression (FE)</th>
<th>WSS (%)</th>
<th>Protein Genes (FE)</th>
<th>Pseudo Genes (FE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sil1</td>
<td>0.406</td>
<td>0.412</td>
<td>0.277</td>
<td>0.796</td>
<td>1.173</td>
</tr>
<tr>
<td>Sil2</td>
<td>0.057</td>
<td>0.269</td>
<td>0.043</td>
<td>0.465</td>
<td>1.282</td>
</tr>
<tr>
<td>Sil3</td>
<td>0.238</td>
<td>0.829</td>
<td>0.174</td>
<td>0.894</td>
<td>1.066</td>
</tr>
<tr>
<td>Rep</td>
<td>0.085</td>
<td>0.250</td>
<td>0.068</td>
<td>0.918</td>
<td>1.067</td>
</tr>
<tr>
<td>Het</td>
<td>0.055</td>
<td>0.171</td>
<td>0.061</td>
<td>0.505</td>
<td>1.143</td>
</tr>
<tr>
<td>Weak</td>
<td>0.157</td>
<td>1.421</td>
<td>0.108</td>
<td>1.217</td>
<td>0.687</td>
</tr>
<tr>
<td>Mod</td>
<td>0.175</td>
<td>1.787</td>
<td>0.423</td>
<td>1.224</td>
<td>0.745</td>
</tr>
<tr>
<td>High</td>
<td>0.124</td>
<td>2.471</td>
<td>0.155</td>
<td>1.562</td>
<td>0.481</td>
</tr>
</tbody>
</table>
Figure 4-1: PDIFF and WSS metrics for cluster models with training set across all epigenomes
Figure 4-2: PDIFF and WSS metrics for cluster models with test set across all epigenomes.
Figure 4-3: Cluster centers for \( k = 6 \) cluster model for epigenomes E09 and E19
Figure 4-4: Cluster specific chromatin state distributions in train and test set genes for epigenome E23
Figure 4-5: Cluster specific expression distributions in train and test set genes for epigenome E23
Figure 4-6: Cluster specific scatter plot of distance from cluster center for epigenome E23
Figure 4-7: Heatmap of class metrics and functional enrichments across epigenomes
Chapter 5

Future Work

5.1 Summary of Results

In this research, a robust methodology for classifying genes using chromatin state segmentations has been presented. The approach used 10000 randomly selected genes from a set of 51671 genes to train the k-fuzzy cluster models for every epigenome. The remaining test set of genes were then fit to the clustering model. This methodology resulted in good correspondence of chromatin state features and other enrichments such as RNA-seq expression between the train set and test set as shown in Figures 4-4 and 4-5. It also identified 8 different gene classes which represent silent, repressed, pseudogenes, and actively expressed genes. As the modENCODE consortium illustrated, the chromatin state signals were very good discriminators between subsets of active genes [5]. Even more the defined PDIFF and WSS metrics provided confidence that the clustering was in fact a reduction into proper subclasses.

5.2 Future Work

As stated, this method shows a lot of promise in gene classification without prior knowledge of the functionality of the genes. There are several areas that can be explored to expand upon this approach and result in an classifications that have greater biological significance.
5.2.1 Expansion of Feature Vectors

Currently, there are 625 feature vectors used that measure the posterior probabilities of the given states and state transitions for each gene. These features are computed over the entire gene body, but this approach can be made more granular by creating feature vectors for specific landmarks such as the first and second exon, the first intron, regions flanking the TSS and TTS, etc. This will lead to the discovery of subclasses within the gene classes discovered in this research. Consequently, the number of clusters with this approach will probably be a few more than the k=6 number used in this paper. In addition, other signals for gene expression can be added to the feature vectors to provide further differentiation between classes. For example, Pol2 binding signals indicating initiation and transcription can be added to increase confidence of clustering and differentiate the weakly expressed genes from the paused ones.

5.2.2 Other Functional Enrichments

There are several other functional enrichments that could validate our clusterings and provide further biological significance to them. Besides RNA-seq expression, CAGE data would also be useful since CAGE captures transcription initiation whereas RNA-seq captures elongation [10]. In addition, gene ontology analysis of the relevant gene classes would also be useful in determining what types of genes are silent or repressed in certain epigenomes and active in other ones. This would further support the earlier claims that differentiated epigenomes have larger classes of silent genes due to specialization of protein production.

5.2.3 De Novo Discovery of Genes

The algorithm presented here learns the chromatin state signatures that differentiate various classes of genes. Instead of using actual gene regions for the test set, virtual genes can be created by identifying promoter-annotated regions in the chromatin state segmentations. These virtual genes can then be fitted to the epigenome-specific cluster models that were trained to identify new classes of gene like regions. The probability
of a label given the region being analyzed would provide further confidence in the
discovery of the gene like regions. This provides an alternative approach to gene
discovery which has yet to be explored.
Bibliography


