

**Predicting enhancer regions and transcription
factor binding sites in *D. melanogaster***

by

Rachel Sealon

Submitted to the Department of Electrical Engineering and Computer
Science

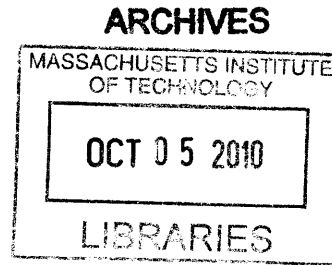
in partial fulfillment of the requirements for the degree of

Master of Science

at the


MASSACHUSETTS INSTITUTE OF TECHNOLOGY

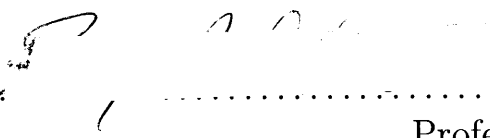
September 2010



© Massachusetts Institute of Technology 2010. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
September 3, 2010

Certified by .. 
Manolis Kellis
Associate Professor
Thesis Supervisor

Accepted by 
Professor Terry P. Orlando
Chairman, Department Committee on Graduate Theses

Predicting enhancer regions and transcription factor binding sites in *D. melanogaster*

by

Rachel Sealfon

Submitted to the Department of Electrical Engineering and Computer Science on 2010, in partial fulfillment of the requirements for the degree of Master of Science

Abstract

Identifying regions in the genome that have regulatory function is important to the fundamental biological problem of understanding the mechanisms through which a regulatory sequence drives specific spatial and temporal patterns of gene expression in early development. The modENCODE project aims to comprehensively identify functional elements in the *C. elegans* and *D. melanogaster* genomes. The genome-wide binding locations of all known transcription factors as well as of other DNA-binding proteins are currently being mapped within the context of this project [8]. The large quantity of new data that is becoming available through the modENCODE project and other experimental efforts offers the potential for gaining insight into the mechanisms of gene regulation. Developing improved approaches to identify functional regions and understand their architecture based on available experimental data represents a critical part of the modENCODE effort. Towards this goal, I use a machine learning approach to study the predictive power of experimental and sequence-based combinations of features for predicting enhancers and transcription factor binding sites.

Thesis Supervisor: Manolis Kellis
Title: Associate Professor

Acknowledgments

Thanks to my family for their love and support.

Thanks to Manolis for his guidance and encouragement. Thanks to Chris, Pouya, Jason, and the rest of the compbio group for their help and suggestions.

Contents

1	Introduction	9
2	Experimental Datasets and Biological Background	11
2.1	Data compendium	11
2.2	Drosophila life stages and cell lines	12
2.2.1	Drosophila embryogenesis	12
2.2.2	Cell lines	13
2.3	ChIP-chip and ChIP-seq assays	14
2.4	Histone marks	14
3	Cis-regulatory module prediction	19
3.1	Mechanism of action of CRMs	19
3.2	Enhancer Gold Standard	21
3.3	Previous approaches to CRM prediction	21
3.4	Enrichment of bound transcription factors and of histone marks in classes of CRM regions	24
3.5	CRM prediction by knowledge-based filtering	29

3.6	Integrating multiple data sources to predict novel blastoderm enhancers	30
3.7	Tissue-specific CRM prediction	34
3.7.1	Constructing tissue-specific training sets	34
3.7.2	Comparing the performance of classifier types	36
3.7.3	Optimizing predictor parameters	38
3.7.4	Predicting novel tissue-specific enhancer regions	44
3.8	Conclusions	45
4	Predicting Transcription Factor Binding	47
4.1	Background	47
4.2	Methods	48
4.3	Predictive power of sequence combinations	49
4.4	Conclusions	54
5	Conclusions	57
6	Appendix A	59
6.0.1	Table 1: Compendium of transcription factor binding experiments	59
6.0.2	Table 2: Chromatin mark timecourse (Kevin White group) . .	64
6.0.3	Table 3: Chromatin marks across cell lines (Gary Karpen group)	68

Chapter 1

Introduction

Development requires the spatial and temporal coordination of complex patterns of gene expression. Identifying the regions in the genome that have regulatory function is important to understanding the mechanisms through which a given regulatory sequence drives a specific pattern of gene expression in early development. The modENCODE project is a large scale multicenter collaboration directed towards comprehensively identifying functional elements in the *C. elegans* and *D. melanogaster* genomes. The genome-wide binding locations of all known transcription factors as well as of other DNA-binding proteins are currently being mapped within the context of this project [8]. Developing improved approaches to identify functional regions and to understand their architecture based on available experimental data represents a critical part of the modENCODE effort. Towards this goal, I use a machine learning approach to study the predictive power of experimental and sequence-based combinations of features for predicting enhancers and transcription factor binding sites.

I have pursued an integrative approach to enhancer prediction that leverages the

wealth of available experimental data on chromatin marks and transcription factor binding. Using a supervised learning framework, I have identified combinations of bound transcription factors, chromatin marks, chromatin-associated factors, motifs, and sequence conservation features that are characteristic of the enhancers in the CAD database, a compendium of experimentally validated enhancers [55]. I find that including multiple feature types improves the power of the classifier relative to using any individual class of features. The improvement in classifier performance using combinations of types of features relative to any individual feature type suggests that each class of functional elements plays distinct yet necessary roles in defining enhancer regions in the cell.

I have also applied supervised learning methods for predicting transcription factor binding locations based on combinations of regulatory motifs. For each experiment in a compendium of ChIP-chip studies, I constructed a classifier to distinguish between regions bound by the given factor and regions bound by any other factor. For each factor, I compared the performance of subsets of enriched and depleted motifs, and examined the improvement in classifier performance as individual motifs are added to the feature set. While the results differ across factors, I found that combinations of features typically outperformed individual motifs, and predictive power increased when depleted motifs were included as features. This result suggests that binding of an individual transcription factor at a given site may be highly dependent on the local combination of bound factors, which provide both synergistic and antagonistic influences.

Chapter 2

Experimental Datasets and Biological Background

In this thesis, I perform integrative analysis on multiple data types. This chapter describes the data sources included in the integrative analysis.

2.1 Data compendium

A compendium of ChIP-chip studies of transcription factors and other DNA-binding proteins was assembled (Appendix A). This compendium includes both new data produced as part of the modENCODE effort and previously published experimental results. There are a total of 196 experiments in the compendium, including data on the binding of 77 distinct proteins at a variety of time points throughout the fly developmental cycle.

Additionally, 185 chromatin mark timecourse experiments were available from

the White group as part of the modENCODE effort (Appendix A), and 52 chromatin mark timecourse experiments were performed by the Karpen group (Appendix A). The White chromatin timecourse includes data on six histone marks at twelve developmental stages, and also contains a timecourse of the binding of the CBP protein and of PolIII. The Karpen dataset includes information on 25 histone variants, as well as several other proteins, in BG3 and S2 cells.

2.2 Drosophila life stages and cell lines

The experimental compendium includes data gathered in multiple life stages and cell lines in the *Drosophila* embryo.

2.2.1 Drosophila embryogenesis

The early development of *D. melanogaster* has been subdivided into 17 stages (2.2.1) [6]. These stages are commonly used to refer to the parts of *Drosophila* development.

In stages 1 and 2 of *Drosophila* development, the nuclei divide and migrate to the periphery of the embryo. However, cells do not form at this stage. During stages 3 and 4, five nuclei move to the surface of the embryo's posterior pole, and are enclosed in cell membranes, forming the pole cells. The pole cells will generate the gametes of the adult fly. In the fifth stage, cellularization occurs. At this stage, all of the cells have the same appearance and shape.

At the sixth stage, gastrulation begins. During gastrulation, the mesoderm, endoderm, and ectoderm layers are segregated. The future mesoderm folds inward to

form the ventral furrow, while the endoderm forms pockets at each end of the ventral furrow and the pole cells move inward. The cephalic furrow also forms at this stage. Gastrulation completes during the seventh developmental stage. The cells remaining at the surface of the embryo are the ectoderm and the amnioserosa.

The eighth stage is marked by the formation of the germ band, a collection of cells which will form the embryo trunk. The germ band elongates during stage 9. Also during this stage, neuroblasts begin to differentiate from the ectoderm. During stage 10, the germ band continues to elongate, and the stomodeum, which will give rise to the foregut, invaginates. The eleventh stage is also marked by the formation of segmental boundaries in the embryo. In the twelfth stage, the germ band begins to retract, and germ band retraction is completed in stage 13. In the fourteenth and fifteenth stages, epidermal cells flatten and spread dorsally, the midgut closes dorsally, and the head continues to form. During the sixteenth stage, organs and somatic muscle tissue become visible; organogenesis continues in the seventeenth stage, which ends with the hatching of the embryo.

2.2.2 Cell lines

The experimental compendium includes studies performed in *Drosophila* cell lines derived from multiple cell types and developmental stages. Transcription factor binding has been mapped in multiple distinct cell lines, including cell lines derived from the nervous system, a blood cell line, and embryonic cell lines (2.2.2).

2.3 ChIP-chip and ChIP-seq assays

The experimental compendium includes both ChIP-chip and ChIP-seq assays. Both ChIP-chip and ChIP-seq are powerful techniques for identifying the genome-wide binding locations of a protein of interest. In ChIP-chip, proteins are first crosslinked to the DNA by treating nuclei with formaldehyde. The DNA is fragmented, and DNA segments bound by the protein of interest are isolated by immunoprecipitation with an antibody specific to the protein. The bound DNA is unlinked from the associated protein, isolated, and amplified using the polymerase chain reaction. The DNA is then hybridized to arrays tiled with genomic regions, so that the genomic location of the bound fragments can be identified [7]. In ChIP-seq, after immunoprecipitation of the protein of interest, the bound DNA is extracted and sequenced directly, and the sequence reads are mapped to the genome.

2.4 Histone marks

In the nucleus, DNA is wound around nucleosomes, which are composed of two copies of each of four distinct histone subunits. The tails of the histones project outward from the nucleosome surface. These histone tails are subject to a variety of post-translational modifications which regulate chromatin accessibility and transcriptional activity. Modifications include acetylation, methylation, phosphorylation, ubiquitinylation and sumoylation[7].

Distinct histone modifications are associated with distinct functional regions and states. Some marks, such as H3K9 acetylation and H3K4 methylation, are associ-

ated with active regions, while other marks, including H3K9 methylation and H3K27 methylation, are associated with repressed chromatin [3]. Heintzman et al. found that H3K4 monomethylation is associated with enhancer regions, while H3K4 trimethylation is associated with active promoters, but not with enhancers, in human HeLa cells [20].

Table 2.1: Stages of *Drosophila* Embryogenesis (from <http://flymove.uni-muenster.de>)

Stage	Time	Developmental events
1	0-0:25 h	Begins when the egg is laid; ends after 2 cleavage divisions complete
2	0:25-1:05 h	Cleavage divisions 3-8
3	1:05-1:20 h	Ninth nuclear division; polar bud formation
4	1:20-2:10 h	Syncytial blastoderm stage; blastoderm nuclei perform final 4 nuclear divisions; pole cells form.
5	2:10 - 2:50 h	Cellularization; blastoderm stage
6	2:50-3 h	Early gastrulation; formation of ventral furrow and cephalic furrow
7	3:00-3:10	Gastrulation completes
8	3:10-3:40 h	Germ band extension
9	3:40-4:20 h	Germ band elongation
10	4:20-5:20 h	Germ band continues elongating; formation of stomodeum
11	5:20-7:20 h	Segmentation
12	7:20-9:20 h	Shortening of germ band
13	9:20-10:20 h	Completion of germ band shortening
14	10:20 - 11:20 h	Head involution and dorsal closure
15	11:20-13:00	Dorsal closure; head involution continues; beginning of condensation of ventral nerve cord
16	13:00 - 16:00 h	Differentiation; somatic musculature, sensory organs, and heart become visible
17	16:00 - 22:00 h	Organogenesis completes; stage ends with hatching of embryo

Table 2.2: Drosophila cell lines

CL.8	larval imaginal wing disc
BG3	central nervous system of <i>D. melanogaster</i> 3rd instar larvae [45]
Kc	embryonic
Mbn2	hemocyte
S2	embryonic

Chapter 3

Cis-regulatory module prediction

3.1 Mechanism of action of CRMs

Gene expression is often modulated by genomic regions known as enhancers or cis-regulatory modules (CRMs). These may be located far from the genes that they regulate. CRMs are portions of DNA that interact with transcription factors to regulate a modular portion of the spatiotemporal expression pattern of a gene. CRMs have been defined experimentally by their ability to drive tissue-specific gene expression in transgenic studies, and computationally predicted by their increased sequence conservation in multiple related species, by their abundance of regulatory motif instances, and by specific signatures of chromatin marks and of bound proteins. A CRM may be located many thousands of base pairs away from the gene that it regulates, in an intron, in the coding sequence, or following the 3' end of the coding sequence [24]. Understanding which portions of the genome have regulatory function is both an important and a difficult step towards understanding the regulatory logic

that drives gene expression.

Several mechanisms of action have been suggested by which CRMs control the pattern of expression of genes that may be located many kilobases away. The most commonly accepted model suggests that enhancer regions physically interact with the target promoter by looping of the intervening sequence. For example, chromatin capture studies showed that the androgen receptor loops from the enhancer to the promoter of the prostate-specific antigen (PSA) gene when the gene is activated. ChIP-chip studies found similar bound proteins, including androgen receptor, PolIII and CBP, at the enhancer and promoter regions, perhaps because the proteins are crosslinked to both proximal sequences [47]. Another suggested mechanism, the DNA scanning model, involves tracking of transcription activators from enhancer to promoter regions. According to this model, a transcription factor complex assembles on a CRM, and then slides along the DNA sequence until it reaches the promoter of the target gene. This model could explain the activity of insulator regions, which would function to block the sliding DNA-protein complex. However, the scanning model has difficulty explaining the action of enhancers that skip over intervening promoters to regulate distant target genes, and could not explain the action of enhancers that regulate expression of a target gene on a different chromosome [24]. Intermediate mechanisms such as facilitated tracking, in which the enhancer loops part of the way to the promoter and tracks the rest of the way, and linking, in which a series of shorter loops form between the enhancer and the promoter to bring the enhancer region closer to the promoter, have also been suggested [52].

Two general paradigms for enhancer architecture have been proposed. In the

enhanceosome model, the precise order, number, and arrangement of bound proteins is crucial in determining the regulatory output of the module. The billboard model proposes that enhancers function as flexible regulatory units in which the output of the module depends on the identity and number of binding sites, but not on their exact arrangement. The human interferon- β enhancer is the canonical example of a CRM with enhanceosome architecture, in which the precise spacing and arrangement of binding sites is critical to its overall function [44]. Comparative genomics studies of enhancers, which have revealed enhancers in related species with conserved regulatory function but divergent number and arrangements of binding sites, provide support for the billboard model of enhancer function [5].

3.2 Enhancer Gold Standard

Applying machine learning approaches to enhancer prediction requires the use of a gold standard for training. I used the CRM Activity Database (CAD) [55] as a gold standard. CAD is a compendium of experimentally validated enhancers, assembled from a literature review, recent experimental results [55], and the REDFly database of enhancers [18]. CAD contains 525 non-redundant CRMs (Table 3.2).

3.3 Previous approaches to CRM prediction

The majority of previous approaches to predicting enhancer regions are unsupervised methods based on clustering of transcription factor binding sites. Several methods

Table 3.1: Most Common Tissue Annotations in CAD

blastoderm	92
dorsal mesothoracic disc	53
embryonic ventral nervous system	47
ectoderm	39
somatic muscle primordium	34
trunk mesoderm primordium	34
embryonic/larval somatic muscle	33
ventral thoracic disc	31
mesoderm anlage in statu nascendi	29
eye disc	27
visceral muscle primordium	26
ectoderm anlage	24
embryonic epidermis	24
ectoderm anlage in statu nascendi	22
embryonic/larval visceral muscle	22
amnioserosa	19
trunk mesoderm anlage	18
trunk mesoderm anlage in statu nascendi	17
ventral ectoderm anlage in statu nascendi	17
peripheral nervous system	15
ventral ectoderm primordium	15

(Ahab, Cister, Cis-analyst) require as input known transcription factor binding site sequences or position weight matrices in order to find regions enriched in clusters of binding sites [37, 12, 2]. Sinha et al.’s Stubb algorithm combines transcription factor binding site position weight matrix information with sequence conservation to identify clusters of conserved transcription factor binding sites [43]. Other approaches (Argos, CisModule) identify both likely transcription factor binding sites and cis-regulatory modules by searching for sequences with short, repeated words [37, 54]. PFR-searcher identifies clusters of conserved, repeated words, while CisPlusFinder predicts cis regulatory modules by locating clusters of perfectly conserved, ungapped subsequences in noncoding regions [17, 36]. Several supervised learning algorithms (HexDiff, LWF)

that distinguish between enhancers and non-enhancer sequences based on word frequencies have also been developed [9, 33]. While these methods have shown success in small-scale validation studies of predicted enhancers, more comprehensive validation of these approaches has been hampered by the time consuming and low-throughput nature of enhancer validation assays. Methods that draw upon the wealth of recently available biological data are likely to reveal additional enhancer regions that may not have been detected by previous enhancer prediction approaches.

The properties of enhancer regions differ from the properties of surrounding non-enhancer regions. Li et al. (2007) compared the properties of known enhancers in the REDFLY database and control non-enhancer regions. This study noted that enhancers have higher GC content, are more highly conserved, and are more likely to be transcribed than other noncoding regions. They also found that blastoderm enhancers, but not necessarily other classes of enhancers, are likely to contain clusters of transcription factor binding sites. Moreover, genome-wide studies of chromatin modification and transcription factor binding events have found that specific chromatin marks and transcription factor binding sites are associated with enhancer regions [20, 51]. Enhancer regions are associated with H3K4 monomethylation, and with an absence of H3K4 trimethylation. The acetyltransferase CBP has also been associated with CRMs by numerous studies [19, 47]. Visel et al. (2009) accurately predicted tissue-specific expression of enhancers in mice by identifying sites bound by the CBP homolog p300 in embryonic forebrain, midbrain, or limb [46].

I have developed supervised learning approaches for enhancer prediction that exploit the enrichment of transcription factor binding events and specific histone marks

in enhancer regions. This project differs from previous work in the use of heterogeneous experimental data sources (histone modifications, transcription factor binding events observed in ChIP-chip assays, motifs, and sequence conservation), rather than purely sequence and conservation-based information, as a feature set to train classifiers. The supervised learning framework also distinguishes this project from most previous methods for enhancer prediction, which generally employ unsupervised approaches.

3.4 Enrichment of bound transcription factors and of histone marks in classes of CRM regions

I performed enrichment analysis to identify the factors and proteins that are most enriched in each class of enhancer regions. I examined enrichment of binding of all transcription factors and chromatin marks in the set of enhancers with at least 15 annotated examples in the CAD database (3-1, 3-2, 3-3).

Enrichment was computed according to the following formula:

$$Enrichment = \frac{|B||E \cap F|}{|E \cap B||B \cap F|}, \quad (3.1)$$

where $|B|$ is the number of base pairs included in the array background, $|E \cap B|$ is the number of base pairs in the given set of enhancers that is included in the array background, $|E \cap F|$ is the number of enhancer base pairs bound by the factor, and $|B \cap F|$ is the total number of base pairs that are bound by the factor.

Enrichment of bound factors varied by enhancer class, Transcription factors whose binding was most enriched for many enhancer classes are known to be functionally involved in the development of the related tissue. For example, the factor that is most enriched in enhancers that drive expression in the dorsal mesothoracic disc is engrailed, with over tenfold enrichment above background. In engrailed mutants, the dorsal mesothoracic disc fails to develop normally; instead, it takes on the characteristics of the anterior region of the mesothoracic disc [14]. The experiments most enriched in known blastoderm enhancers probed factors known to be involved in blastoderm development (such as knirps, tailless, Schnurri, and bicoid) during the blastoderm stage (2-3h). Enhancers that drive expression in somatic muscle tissue are most enriched in muscle-associated transcription factors, including bagpipe, myocyte enhancer factor-2, biniou, and twist. Binding of tll, which controls genes that promote normal development of the head and posterior of the embryo, is among the transcription factors most enriched in ectoderm enhancers. The transcription factor whose binding is most enriched in peripheral nervous system enhancers is twist; twist knockouts have mutant nervous system phenotypes [21]. The transcription factor whose binding is most enriched in enhancers that drive expression in the trunk mesoderm primordium is tinman, which is functionally implicated in mesodermal patterning [22].

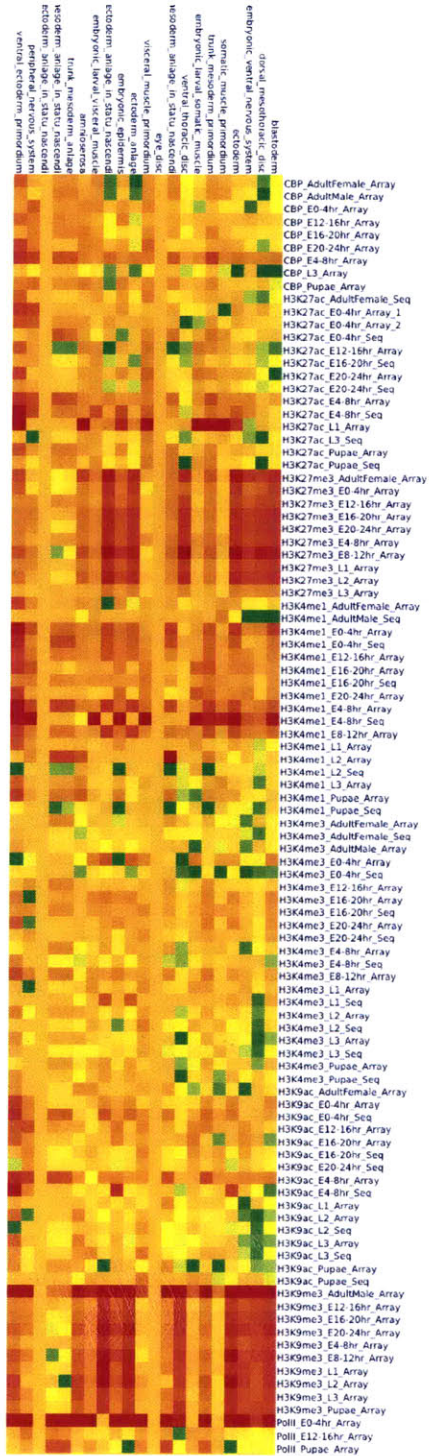


Figure 3-2: Enrichment of chromatin marks (White lab timecourse) in enhancer regions. Green indicates marks whose binding is depleted, while red indicates marks whose binding is enriched.

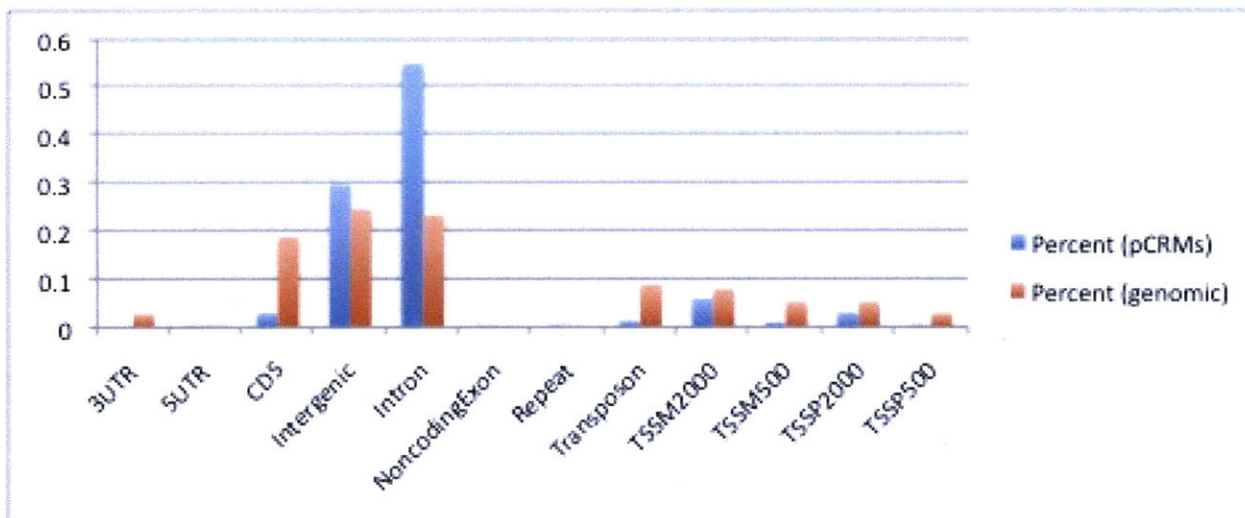


Figure 3-4: Distribution of region types of CRMs (predicted by unsupervised filtering) and all genomic regions.

3.5 CRM prediction by knowledge-based filtering

I applied a filtering method to identify potential enhancer regions based on the known characteristics of enhancers. I segmented the genome into 100-base-pair windows, and identified regions with the following characteristics typical of enhancer regions:

- absence of H3K4 trimethylation, a chromatin mark which is typical of promoters
- presence of H3K4 monomethylation, a chromatin mark which is typical of enhancers
- CBP binding
- presence of transcription

After merging adjacent regions, there are 545 regions of the genome meeting these criteria. These regions overlap ten of the known enhancers in the CAD database, and are enriched in intronic regions, depleted in coding and promoter sequences (3-4).

3.6 Integrating multiple data sources to predict novel blastoderm enhancers

Since enhancer regions are enriched in the binding of many distinct types of transcription factors and chromatin marks, combining multiple data types to create an integrated predictor for enhancer regions is likely to result in more accurate predictions than identifying enhancer regions based on any individual data source in isolation. In order to predict novel enhancer regions, I combine chromatin marks, transcription factor binding data, and conservation within a supervised learning framework.

Because of the availability of early embryonic-stage experimental data and of known blastoderm enhancers, I construct a predictor for blastoderm-stage enhancer regions. I segment the genome into 1203813 nonoverlapping 100-base pair windows, and construct a feature vector for each window based on the counts of transcription factor binding events, chromatin marks, and sequence conservation based on phast-Cons score, a sequence conservation metric [42]. As a positive gold standard, I use the union of the set of blastoderm enhancers in the CAD database with the set of blastoderm enhancers compiled by Papatsenko et al [35, 55]. These gold standards represent a compendium of experimentally validated enhancer regions, and contain 140 distinct enhancer regions spanning 2196 windows. All other windows were taken to be negative examples.

To explore the performance of classifiers constructed using various subsets of features, I create a classifier using all of the experimental features in the compendium, as well as various feature subsets. Feature subsets include the five transcription factor

binding experiments most enriched in known blastoderm enhancers and the five transcription factor binding experiments most enriched in known blastoderm enhancers as well as the five experiments most depleted in known blastoderm enhancers. I also examined the performance of feature subsets including only a single experiment for each transcription factor and chromatin mark. (Since many factors were examined in multiple experiments, including only one experiment per factor reduces the number of features). I selected the representative experiments and chromatin marks on the basis of experimental stage (selecting an experiment for each factor that was closest to the blastoderm stage, 71 features), highest information gain with known blastoderm enhancers (105 features), and greatest enrichment in known blastoderm enhancers (105 features).

Selecting subsets of features results in better performance than including all features. I examine performance with a variety of feature sets (3-5). Including only the 5 transcription factor binding experiments that are most enriched in the set of known blastoderm enhancers as features, the performance of the classifier based on six-fold crossvalidation is almost as good as when the full set of features is included. Including the 5 transcription factors most depleted in the set of known enhancers further improves the power of the classifier. When only one experiment per unique transcription factor is included in the feature set, the classifier power improves still further. Unique transcription factor feature sets were constructed by including the most stage-appropriate experiment only; the experiment with the highest information gain for the set of known enhancer regions; and the experiment that is most enriched in known enhancer regions. Based on six-fold cross-validation, the performance of the

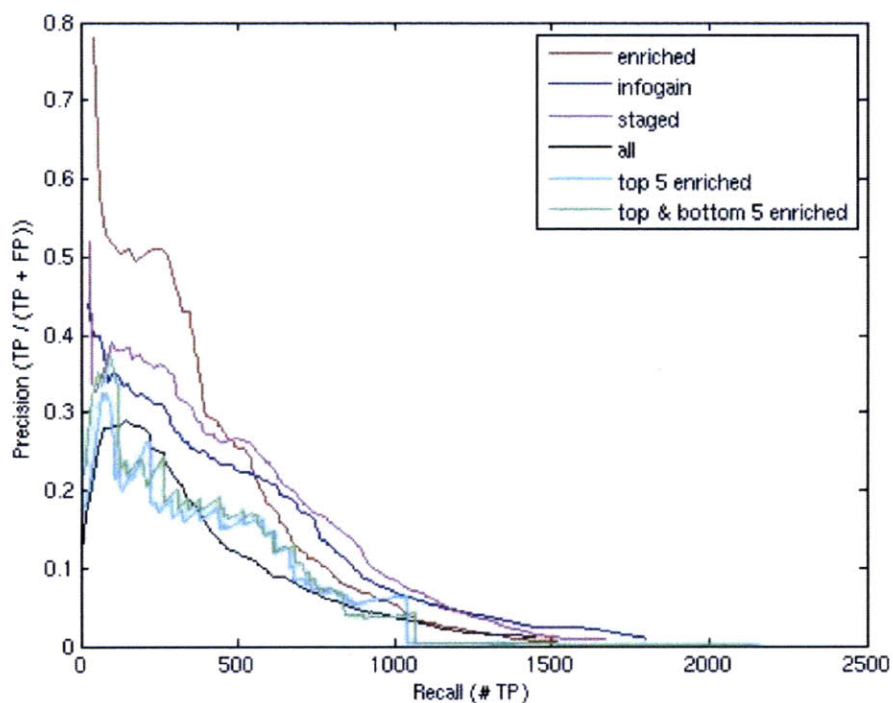


Figure 3-5: Performance of feature subsets for enhancer prediction.

classifier is highest using the classifier constructed from the unique experiments that are most enriched in enhancer regions, as well as chromatin mark and conservation data.

A number of intriguing observations emerged from this analysis. Firstly, although chromatin marks or chromatin remodeling factors have poor predictive power in isolation, combining chromatin mark and chromatin factor binding data with TF binding data substantially improved classifier performance. This result suggested that separate classes of functional elements play distinct and important roles in defining enhancer regions. I examined the set of true positive windows that are among the top 100 predictions of the classifier including chromatin marks as well as transcription factor binding data, but not among the top 500 predictions of the classifier using

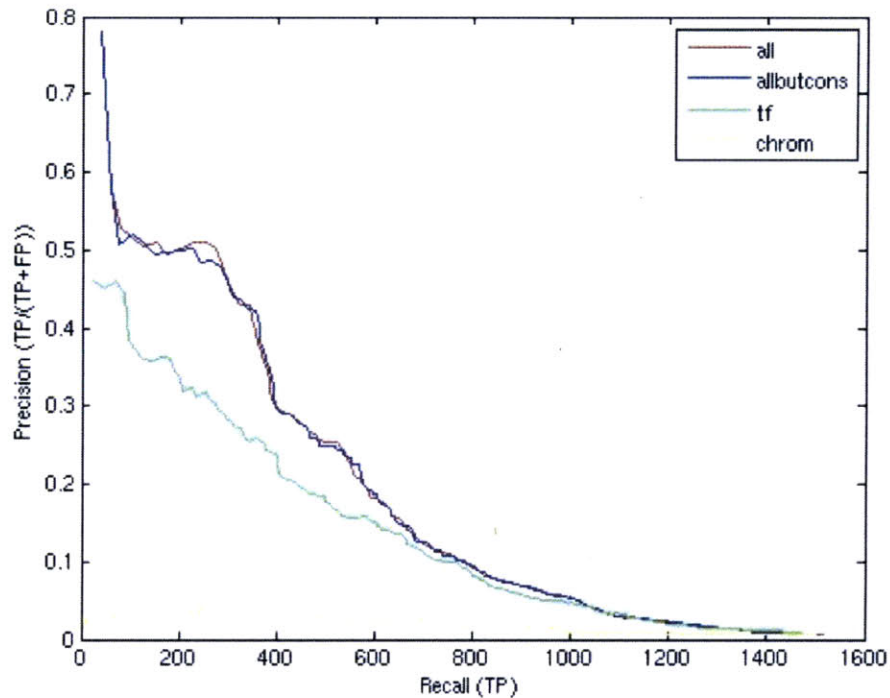


Figure 3-6: Performance of individual feature types for enhancer prediction.

transcription factor binding data only. There are 14 such windows from three distinct parts of the genome, compared with six true positives among the top 100 predictions of the classifier using only transcription factor binding data, but not among the top 500 predictions of the classifier using all data types. Windows correctly classified as positives by the classifier including chromatin mark features, but not by the classifier containing transcription factor binding features only, contained multiple chromatin marks. The number of false positive predictions that are present in the top 100 predictions of one classifier but absent from the top 500 predictions of the other are similar for the classifier using only transcription factor binding features and the classifier using all feature types (5 for the former, 6 for the latter).

Enhancer predictions were validated using cross-validation, as well as by exam-

ining top predictions for characteristics of known enhancer regions. To ensure that adjacent windows (which are likely to have similar feature sets, since experimentally determined transcription factor-bound regions and chromatin marks are generally longer than the 100-base pair window) are placed in the same crossvalidation fold, the classifier was trained on enhancer regions in five of the six chromosomes (chr2L, chr2R, chr3L, chr3R, chr4, chrX), and tested on the remaining chromosome. Six-fold cross-validation confirmed the ability of the classifier to recover known enhancer regions while excluding regions that are not likely to be enhancers. Top predictions overlapped recently validated enhancers more than a negative set of previously predicted enhancers for which experimental validation failed [55]. Moreover, predicted enhancers were enriched near genes patterned in the blastoderm, and top predictions included a higher percentage of blastoderm-patterned genes than regions bound by most individual transcription factors.

3.7 Tissue-specific CRM prediction

3.7.1 Constructing tissue-specific training sets

Since enhancers that drive gene expression in distinct tissues are enriched in distinct bound factors, I examined the ability of a supervised classification approach to predict tissue-specific expression. For each of the 21 IMAGO categories with at least fifteen annotated enhancers, I extract the central 1500 base pairs of each enhancer in the category. I choose size-matched random negative regions, construct a feature vector

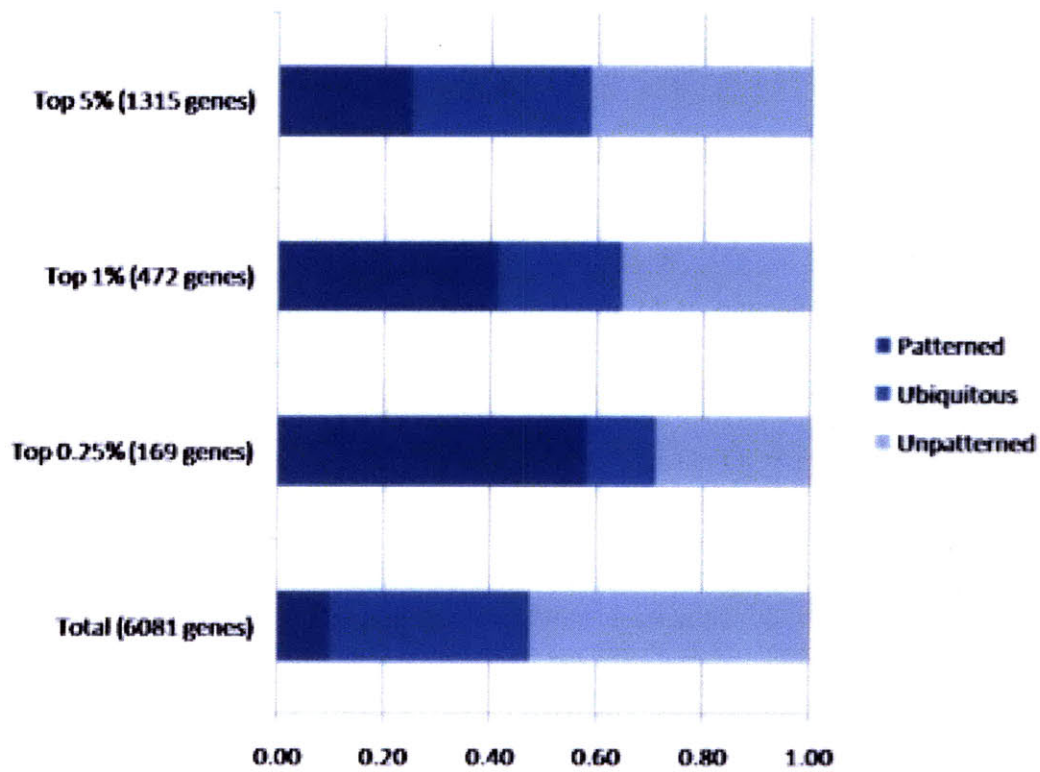


Figure 3-7: Predicted enhancers are enriched near genes that are patterned in the early blastoderm embryo

consisting of overlap with bound transcription factors, conserved motifs, and chromatin marks, and learn a classifier to distinguish between the negative and positive regions.

One concern with setting up the prediction in this way is overfitting due to adjacent enhancer regions that overlap shared features. To determine if overfitting occurred due to adjacent enhancer regions overlapping similar or identical sets of bound regions, I examined the effect of varying the crossvalidation scheme on classifier performance. I split the training examples into test and training sets using either random assignments, or in genomic order, with the first half of enhancers constituting the training set and the second half comprising the test set. I also examined the performance of fivefold and tenfold crossvalidation, with regions assigned randomly to folds. In ten of the 21 IMAGO categories, the genomic order split had the lowest predictive power, and in fourteen of the 21 categories, the genomic order split performs worse than 2-fold crossvalidation with random fold assignments. While the result does not achieve statistical significance ($p=0.09$), the result suggests that genomic proximity of enhancers may result in a small amount of overfitting using this method when crossvalidation fold assignments are random (3-8).

3.7.2 Comparing the performance of classifier types

I then tested the performance of a variety of classifier types and parameter settings (SVM, C4.5 decision tree, Logistic regression, and Naive Bayes classifiers). For the logistic regression classifier, I tested ridge parameter values 0.1, 1, 10, and 100. (Higher

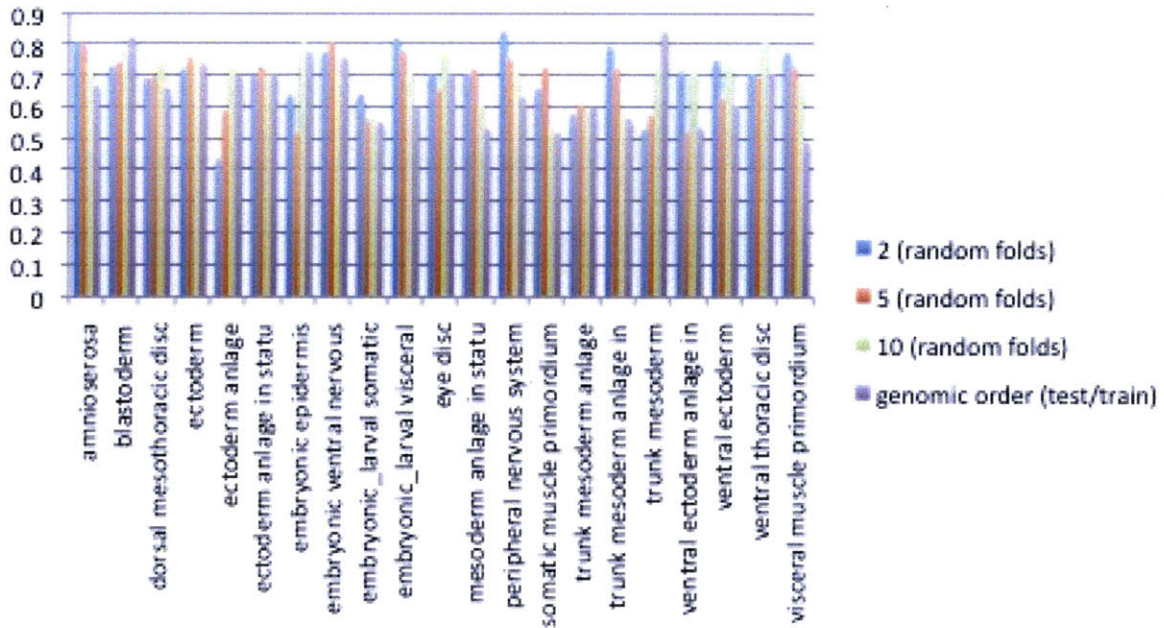


Figure 3-8: Effect of crossvalidation scheme on classifier performance

values for the ridge parameter encourage low feature coefficients and control overfitting). The classifier achieved the best performance with ridge parameter of 100 in fifteen of the 21 categories when the ridge parameter was set to 100 (using the genomic split), and in fourteen of 21 categories using ten-fold crossvalidation. For the SVM classifier, I examined performance using all combinations of kernel degree 1,2,3,4 and slack parameter C set to 0.1, 1, 10, and 100. (Increasing the slack parameter C increases the penalty for errors). In general, linear or second-degree kernel and slack penalty of 0.1 showed the best performance). For an RBF kernel, I tested gamma set to 0.0001, 0.001, 0.01, 0.1, 1, 10, and 100 with C set to 1. (Lower values of gamma cause a higher-width kernel, resulting in a smoother classifier). The best performance was achieved with an intermediate setting for gamma, gamma=0.01. High gamma values overfit the training data and assign all test data to the same class, while low

values of gamma fail to separate the training data. For the C4.5 decision tree, I examined a range of values for the parameters M (minimum number of examples in a leaf node) and C (pruning confidence, with lower values of C causing heavier pruning of the tree), and found that the highest predictive power was for M=8 in 10 of the 21 categories. The performance of the classifier was stable across the various tested values of C. Across the classifiers tested, the best performance was achieved using the logistic regression classifier in 6 of the 21 IMAGO categories, and the Naive Bayes classifier in 7 of the 21 IMAGO categories.

3.7.3 Optimizing predictor parameters

To predict novel enhancer regions, I constructed logistic regression classifiers trained on known enhancers. As a training set, I used enhancers in the CAD database [55] as positive examples, and randomly selected size-matched regions as negative examples. I initially selected the thirteen controlled vocabulary terms with the greatest number of annotated known enhancers in the CAD database for analysis. I also defined several aggregate categories by combining related terms.

The 592 features provided as input to the classifier included experimental (Chip-ChIp and Chip-seq) and computational (conserved motifs) features, comprising 189 chromatin mark datasets, 85 chromatin remodeling factor datasets, 171 conserved motifs, and 147 transcription factor interaction site datasets.

I performed logistic regression using the weka machine learning library (version 3.6.1) [49] after optimization of regression parameters. To determine the logistic

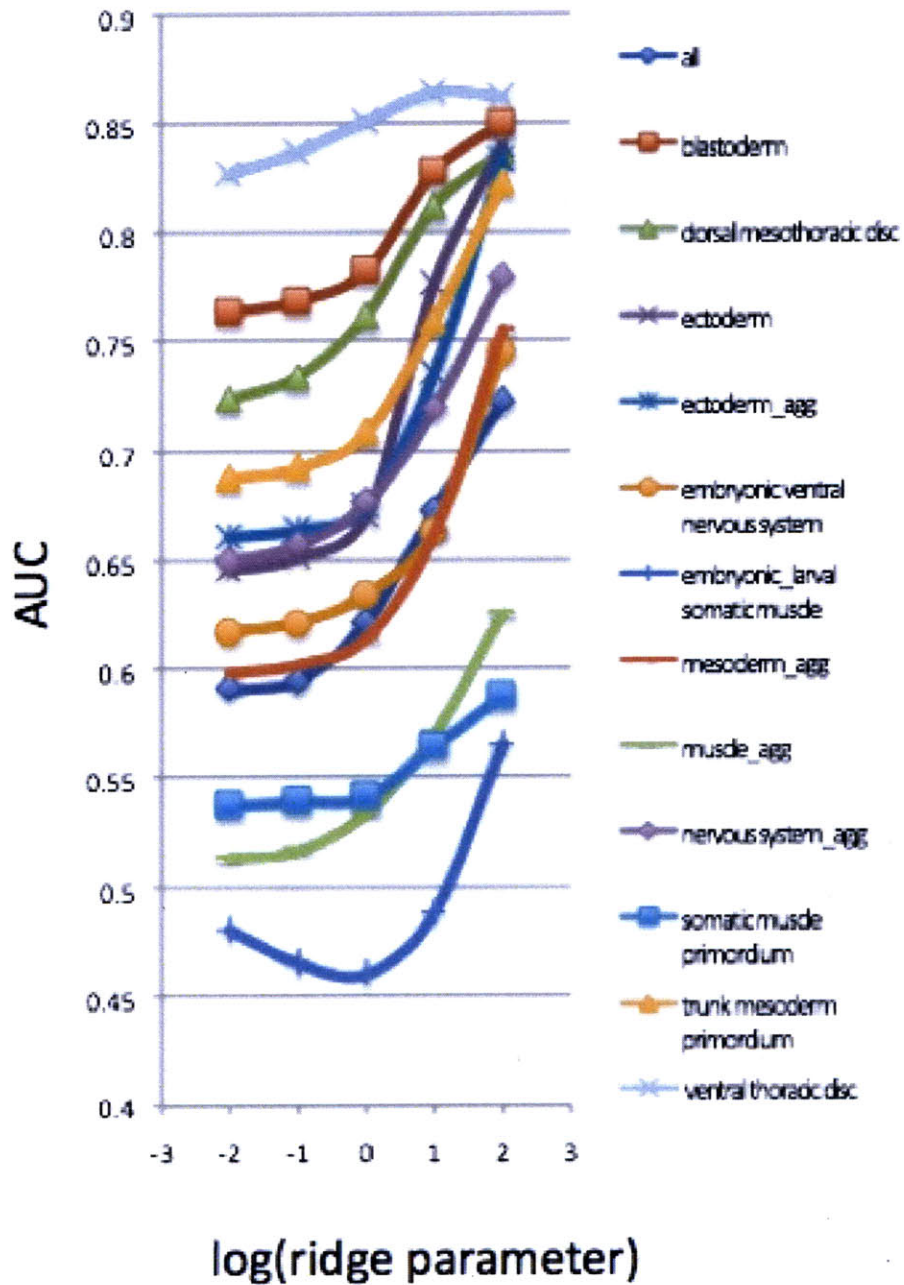


Figure 3-9: Optimizing ridge parameter. Ridge parameter values between 10^{-2} and 10^2 were selected for analysis. The classifier with ridge parameter = 100 performed best across tissue types.

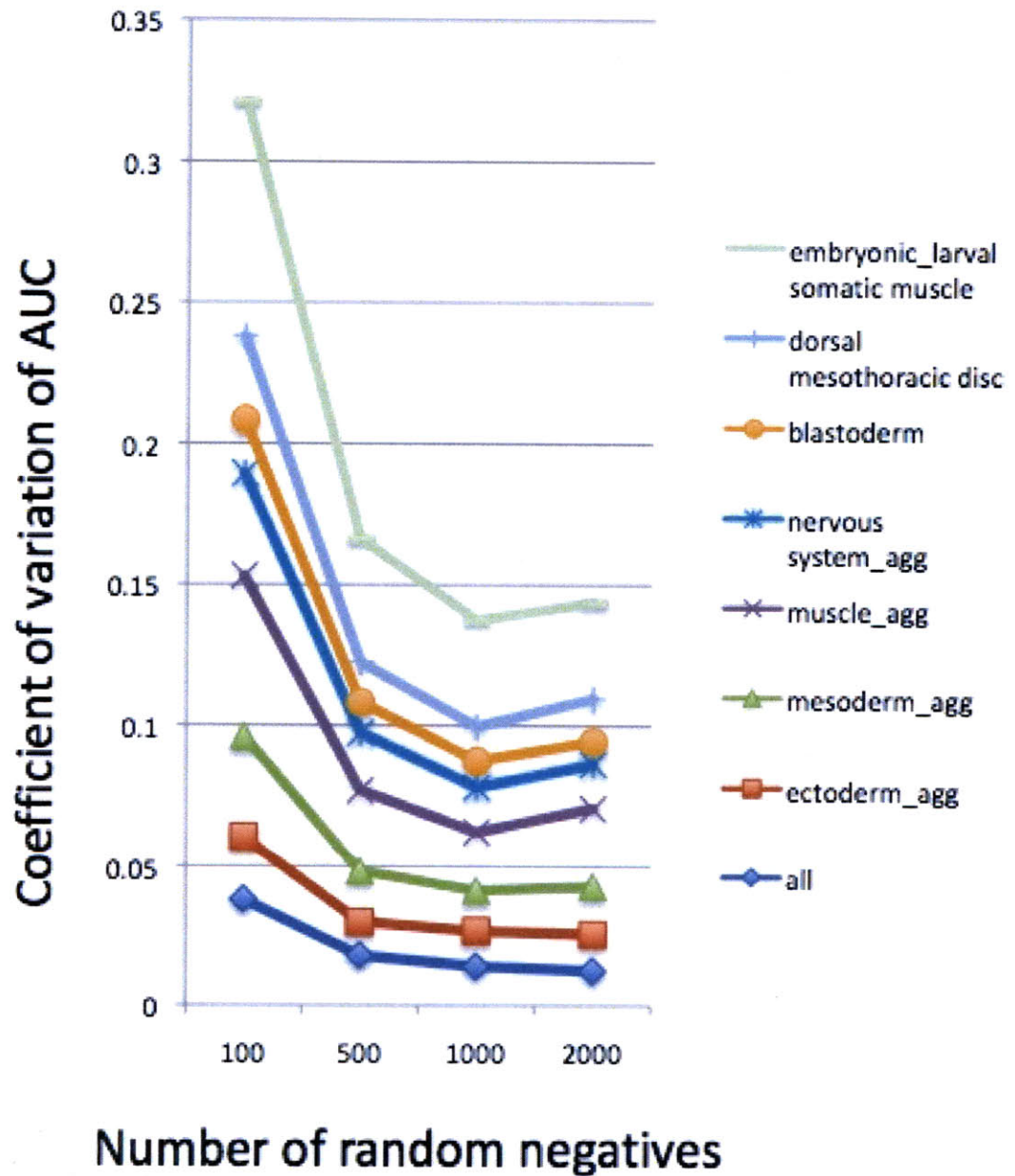


Figure 3-10: Determining adequate number of random negative training examples to reduce variation in AUC. The coefficient of variation of the AUC was examined when between 100 and 2000 random regions were selected. 500 trials were repeated per test.

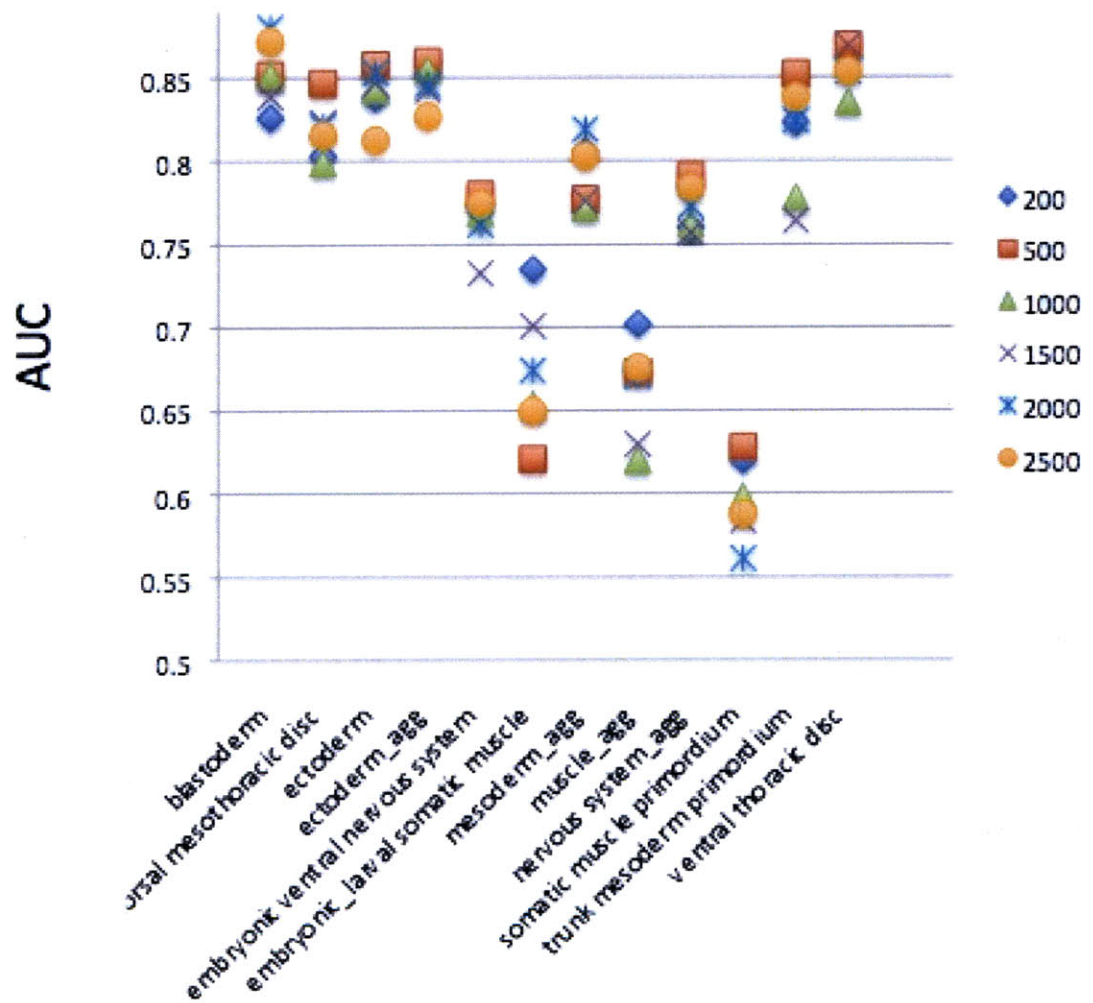


Figure 3-11: Determining optimal window size. Windows between 200 and 2500 base pairs were tested centered on enhancer regions. For the categories with the highest AUC, a 500-base-pair window performed best.

regression classifier ridge parameter, I compared the area under the curve (AUC) with values ranging from 0.01, 0.1, 1, 10, 100 under 10-fold crossvalidation on the training set. A ridge parameter setting of 100, which showed the highest AUC across tissue types, was selected for further analysis (3-9). The adequate number of randomly selected negative genomic regions to provide consistent results was determined by comparing the coefficient of variation of the AUC for 500 trials using between 100 and 2000 random negative regions, with random regions selected separately for each trial (3-10). Based on this analysis, negative training sets of 1000 bp were used for further analysis. In order to determine the optimum-sized window for training, I compared the AUC obtained for each category with windows sizes ranging from 200 to 2500, selecting 500 bp windows as having the best performance for most tissues (3-11).

To examine the performance of feature subsets, I compared the AUC values obtained using the following feature groups: chromatin timecourse features only, chromatin cell line features only, ChIP-features only, motif features only. Enhancer regions in the training data that overlapped within a 2500-base-pair window were excluded. I observed that the performance of chromatin timecourse features only and of chromatin cell line features only was better than that of any individual group of features in isolation.

I also investigated the effects of feature selection on classifier performance to determine whether exclusion of low information gain features would improve classification. The information gain of each feature was computed on half the training set (selected from half the genome), and performance was evaluated using the top 5 to 100 features

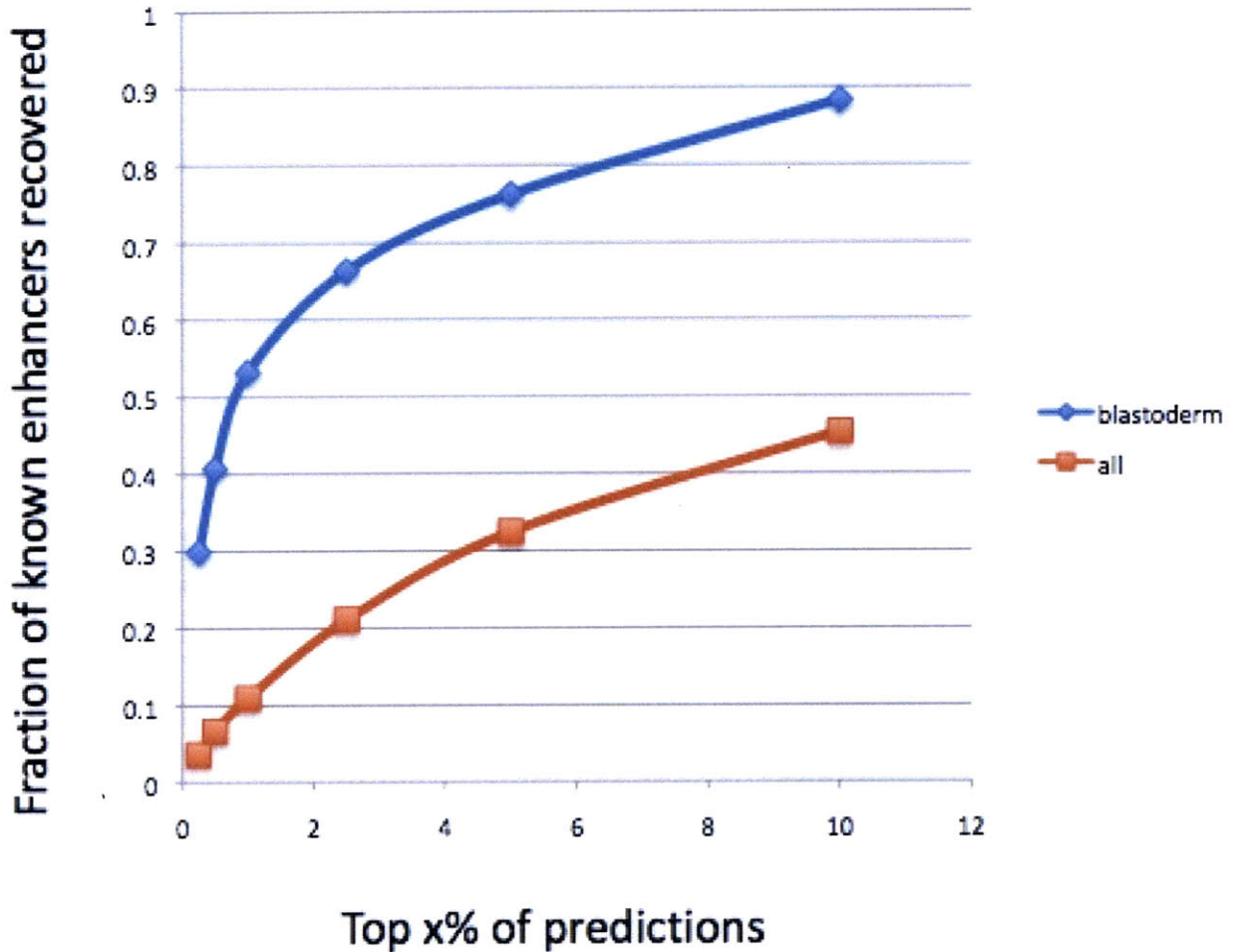


Figure 3-12: Predicted enhancer enrichment in REDfly database. The predicted enhancers were recover enhancers in the REDfly database that were not included in the CAD database training set.

on the remaining half of the test dataset. I also examined the effect of merging features, so that the value of the i th entry in the feature vector for a given transcription factor is the sum of the number of time points and cell lines for which the transcription factor binds the region of interest. This analysis showed that the full feature set displays a higher AUC than any of the feature subsets examined.

3.7.4 Predicting novel tissue-specific enhancer regions

To predict novel enhancers, a logistic regression classifier was trained on the known enhancers of each tissue category using the optimized parameter values. The genome was tiled with 500-base pair windows, with each window offset by 100 base pairs, and the classifier was applied to all windows. To assign a threshold for enhancer predictions, I estimated the false discovery rate (FDR) for the predictions. The total number of enhancers in the genome was assumed to be comparable to the number of genes in the genome. The number of true enhancers in each tissue category is likely to differ; to estimate this number, I scaled by the fraction of enhancers in CAD in a given tissue category. I computed the FDR for the total number of enhancers as follows: I compute the FDR on the training set, which has ratio r of true positives to false positives, and scale the FDR based on r' , the estimated ratio of true positives to false positives in the genome. Thus, if the estimated FDR based on the training set is $FP/(TP+FP)$, the scaled estimate of the FDR is: $\frac{FP(r/r')}{TP+FP(r/r')}$. For each tissue category, we select as a cutoff threshold the enhancers that have FDR threshold less than 0.5. Three tissue categories (all, blastoderm, and nervous system aggregate) met this threshold, and the number of enhancers meeting the cutoff is 26807 (all enhancers), 6263 (blastoderm), and 40 (nervous system aggregate).

Downstream analysis supported the biological significance of the enhancer predictions. The genomic distribution of the predicted enhancer regions was examined, with most predicted enhancers located in introns and in intergenic regions. For the predicted blastoderm enhancers, the fraction of predicted enhancers located near genes

known to be patterned in the blastoderm embryo was examined. The most confident predictions were enriched in genes that are patterned in the blastoderm. Finally, the ability of the predictions to recover enhancers included in the REDFly database (downloaded from redfly.ccr.buffalo.edu/ on 1/19/10) but not in CAD was examined. Top predictions were also enriched in known enhancers not included in the CAD database (3-12).

3.8 Conclusions

This analysis suggests that multiple data types contribute to predicting enhancer locations, with conserved motifs, chromatin marks, transcription factor binding sites, and chromatin remodeling factors all improving classifier power. Enhancers are predicted by a combination of genetic and epigenetic elements, indicating that multiple types of features combinatorially define enhancer regions. This complexity may have arisen to provide the precision needed to control tissue-specific, stage-specific, and stimulus-specific expression of individual transcripts.

Enhancers in different tissue classes are enriched in the binding of distinct transcription factors, and I constructed tissue-specific enhancer predictions by training on known enhancers in distinct tissue categories. I predict 26807 general enhancers, 6263 blastoderm enhancers, and 40 nervous system enhancers at a false discovery rate of 0.5. Downstream analysis supported the biological significance of predicted enhancers, indicating that predicted blastoderm enhancers were enriched near blastoderm-patterned genes and in known enhancers not included in the CAD

database.

Chapter 4

Predicting Transcription Factor

Binding

4.1 Background

Understanding the relationship between genome sequence and the observed genome-wide binding locations of transcription factors is an important problem in computational biology. Additionally, understanding the range of sequence motifs that are positively or negatively correlated with binding of specific transcription factors may provide insight into co-operative and competitive interactions among transcription factors.

Previous studies have successfully applied sequence and motif-based features within a supervised framework to predict transcription factor binding. For example, Zhou and Liu (2008) examined a variety of learning methods for predicting the binding of transcription factors to DNA [53]. Ernst (2010) combined multiple evidence sources

using logistic regression to predict ChIP-chip binding sites across the human genome [11]. Using varying combinations and subsets of features, I have predicted binding of transcription factors within a supervised learning framework. The goals are to elucidate the feature sets that are most predictive of binding, providing insight into the combinations of sequence features that are associated with binding of individual transcription factors and the synergistic and antagonistic interactions among factors.

Transcription factor motifs were obtained from Transfac, Jaspar, and FlyReg databases, as described in Kheradpour et al (2007) [25]. Motif conservation was assessed based on conservation of a motif instance in the phylogenetic tree of twelve *Drosophila* species [25]. Thus, a motif at 0.5 conservation is conserved through half of the phylogenetic tree. 304 distinct motifs are included in the full compendium.

4.2 Methods

I constructed a classifier to predict the binding locations of individual transcription factors based on sequence features. I segmented the genome into 2000-base pair windows, constructed a feature vector consisting of counts of 7mer motifs in each window, and predicted transcription factor binding using logistic regression. As a positive training set, all windows that are bound in a given experiment and contain at least one motif were used. All windows that are not bound in any experiment by the factor, but are bound by some factor in at least one experiment and contain at least one motif, were used as a negative training set. Thus, the classifier learns to distinguish between regions bound by a given factor at a given timepoint, and regions

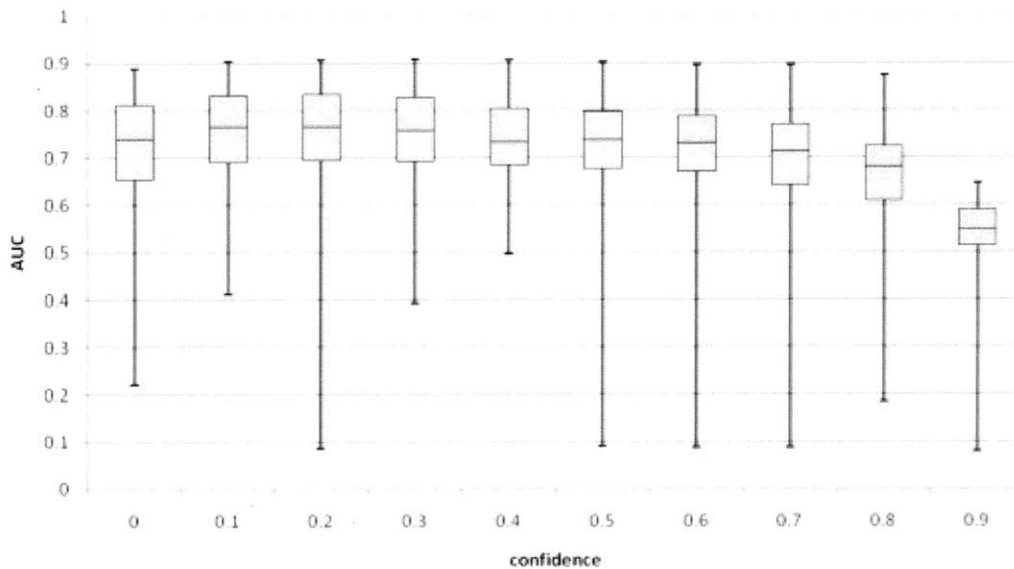


Figure 4-1: Comparison of performance of classifier using motifs conserved at various confidence levels. The highest median AUC was obtained with motifs conserved at the 0.2 confidence level

that are never bound by the factor of interest, but are bound by some factor at some timepoint.

4.3 Predictive power of sequence combinations

I investigated the combinations of features that were most predictive of transcription factor binding (4-1). Firstly, binding in nearly all experiments was better predicted by the smaller set of conserved motifs than by the larger set of motifs without conservation information included. This result suggested that including motif conservation information improves predictive power. The exceptions were experiments on factors binding low-complexity motifs, like ultrabithorax, and also several experiments that bound a very small number of regions. Additionally, I compared performance using

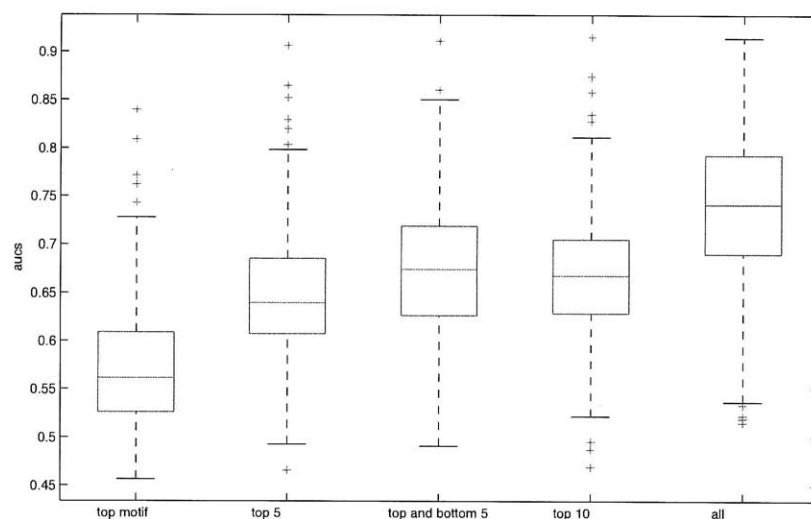


Figure 4-2: Performance of feature subsets. The performance of the classifier using the most enriched motif, the five most enriched motifs, the five most enriched and most depleted motifs, and the ten most enriched motifs, were examined. The median AUC when the feature set included both enriched and depleted motifs was higher than the median AUC when the feature set included enriched motifs only.

various subsets of features, including only the most enriched motif reported by motif discovery algorithms, the top 5 most enriched motifs, the top 5 most enriched as well as the 5 least enriched motifs, and the ten most enriched motifs (4-2). I found that including depleted as well as enriched motifs improved the power of the classifier, suggesting that this approach may provide insight into antagonistic as well as synergistic relationships among transcription factors. The result is consistent with the hypothesis that additional motifs provide information on synergistic and antagonistic binding, improving predictive power.

I also compared the predictive power of motifs and kmers as features. While for some factors, motifs were more predictive than kmers, kmers were more predictive of binding for other factors. In general, when motifs were far more predictive, the

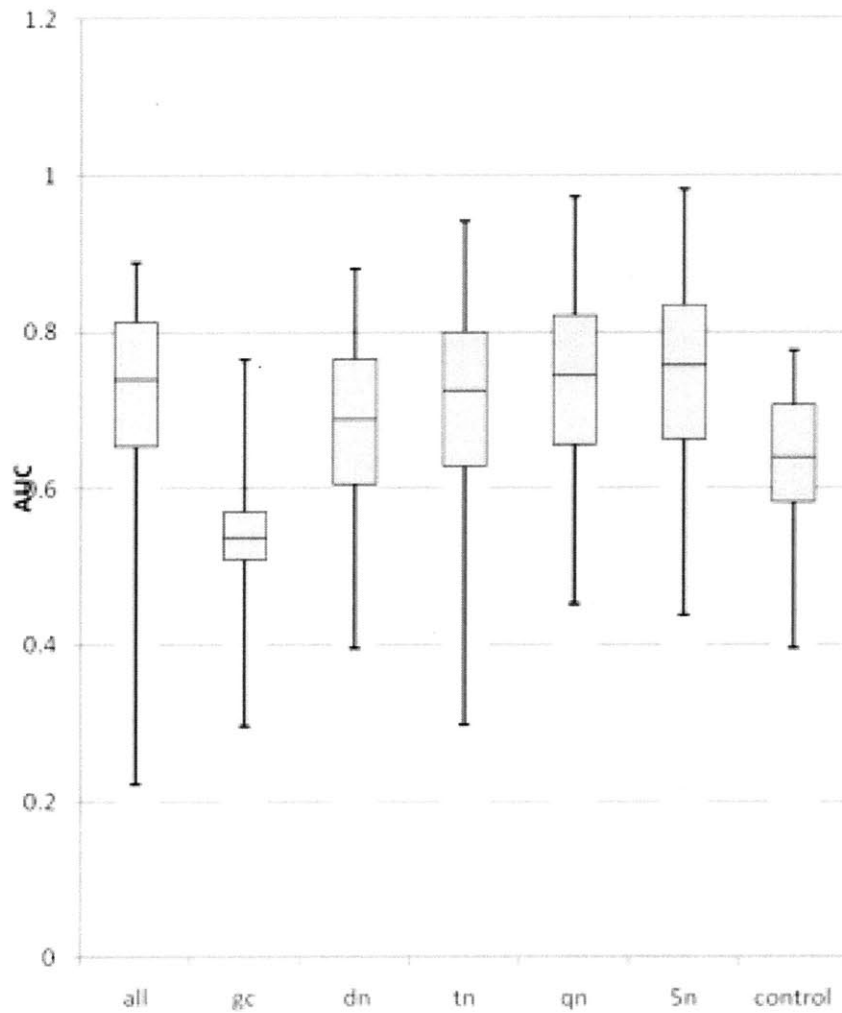


Figure 4-3: Performance of shuffled control motifs and kmers as features. The boxplot displays the AUC for predicting transcription factor binding using a feature set consisting of all motifs, GC-content, dinucleotide composition, 3-mers, 4-mers, 5-mers, and shuffled control motifs. The set of all real motifs were more predictive of transcription factor binding than the set shuffled control motifs, in 117 of 131 experiments. For some factors, motifs were more predictive than kmers, while kmers were more predictive of binding for other factors.

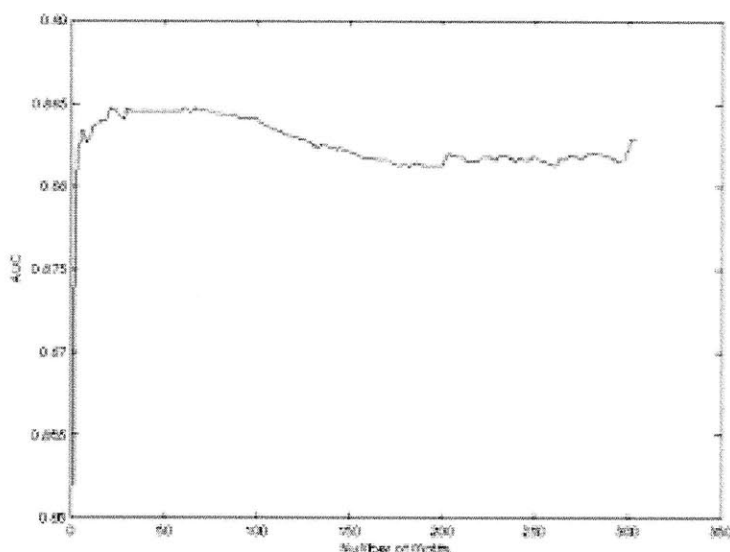


Figure 4-4: Features were ranked based on information gain, and classifiers were constructed by incrementally adding additional features. The performance of the classifier to distinguish between regions bound by Su(Hw) and by all other factors reaches nearly its maximum after the first few features are added; for this factor, performance does not improve when additional features are added.

motif for the factor was high-information content, while when the k-mers were more predictive, the associated motif was highly repetitive or has low information-content. Also, real motifs were more predictive of transcription factor binding than shuffled control motifs, in 117 of 131 experiments (4-3).

I also clustered the motifs using the method described in Xie (2005), grouping together motifs with high Pearson correlation [50]. Clustering the motifs reduces the number of features by merging features that represent motifs that are similar. I clustered motifs whose similarity based on Pearson correlation ranged from 0.4 to 0.9, at 0.1 increments. However, in 180 of 195 experiments, performance was better using the full set of motifs than using any clustered subset

Finally, motifs were ranked based on their information gain for each experiment,

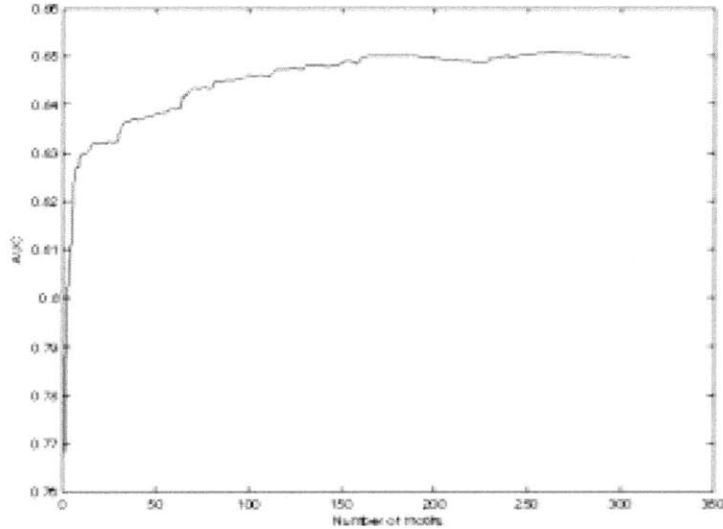


Figure 4-5: Features were ranked based on information gain, and classifiers were constructed by incrementally adding additional features. The performance of the classifier to distinguish between regions bound by Trl continues to increase as additional motifs are added as features.

and the predictive power of the top k motifs was evaluated for each experiment, for k ranging from 1 to the total number of motifs. The number of features at which optimal predictive power was achieved varied by experiment and by factor (4-4, 4-5). Some factors were well predicted by just a few motifs, and little additional predictive power was gained as more features are added. For some factors, predictive power began to degrade due to the addition of extra motifs. Su(Hw) is an example of a factor for which across all experiments studying the factor, the binding was well predicted with just a few features, while Trl is an example of a factor for which the binding required many features to reach the optimal predictive power across multiple experiments. Intriguingly, consistent with the computational finding that more features contribute information towards Trl binding than towards Su(Hw) binding, in the *Drosophila* interaction database DroID, Trl has twice the number of physical

interactions with transcription factors as Su(Hw) (8 interactions vs. 4 interactions). This result suggests that the larger number of motifs that have predictive power for Trl relative to Su(Hw) may capture physical interactions between factors that stabilize or destabilize binding.

4.4 Conclusions

The variation in the optimal subset size to predict different factors suggests that there is a wide range of mechanisms underlying the combined affects of motifs in predicting factor binding. Binding that is well predicted by just a few motifs may indicate a factor with highly specific sequence recognition that is well represented by existing motifs, while binding that requires many motif features to reach optimal predictive power suggests cooperative or competitive interactions, recognition of multiple distinct motifs, less specific binding, or sequence specificities that are not well captured by the existing motif compendium.

Previous studies based on sequence analysis and gene expression have suggested that expression is controlled via combinations of transcription factor sites [1, 29]. Studies of a mouse strain carrying human chromosome 21 support the hypothesis that gene expression is determined not by the cellular environment but by differences in the regulatory sequence [48, 10]. A thermodynamic model in conjunction with studies of synthetic promoters in yeast suggests that interaction between transcription factors contributes to binding stabilization and cooperativity [16]. Previously, testing whether in fact the binding of specific proteins is determined by cooperative motif ef-

fects has not been possible. The present study shows using genome-wide experiments and analysis that positive and negative cooperativity of sequence motifs contributes directly to determining the probability of transcription factor binding in a eukaryotic species. These results are a step towards correlating sequence with regulation of gene expression and help set the stage towards a better understanding of the combinatorial sequence code underlying regulation.

Sequence analysis and expression-based computational studies, predominantly in yeast, suggest that relative position, orientation, location relative to transcription initiation sites, and presence of multiple binding sites may all contribute to the control of gene regulation. Future studies of the modENCODE data are likely to contribute to a greater understanding of the detailed mechanism by which DNA sequences determine transcription factor binding and gene expression.

Chapter 5

Conclusions

In this thesis, I identify novel enhancer regions in *D. melanogaster* by integrating multiple data types, including transcription factor binding site, conservation, and chromatin mark data. I find that the combination of all data types performs better in recovering known enhancers than any individual type of data. I also generate tissue-specific enhancer predictions by training classifiers on sets of known enhancers in specific tissues. Finally, I predict transcription factor binding sites from combinations of motifs, and show that for distinct factors, the number of motifs required to achieve optimal predictive power differ. Also, including both enriched and depleted motifs contributes to classifier performance, and the optimal number of motifs to reach peak predictive power differs by factor.

This analysis is relevant in the context of the modENCODE project, which seeks to generate a comprehensive catalog of all sequence-based genomic elements and has examined the genome-wide binding of a diverse compendium of proteins as well as the genome-wide occurrence of multiple histone marks. Previous work to examine the

relationship between combinatorial transcription factor binding and gene regulation has relied primarily on sequence analysis[1, 29, 41]. Analysis based on observed binding of proteins permits a direct examination of the relationships between sequence elements and observed binding as well as between transcription factor binding and gene expression. Recent large-scale experiments, which map the genome-wide binding of proteins and occurrence of histone marks, hold the promise of allowing a better understanding of the combinatorial code underlying gene regulation.

This work has examined the relationship between combinations of sequence motifs and transcription factor binding within defined windows of the genome, and generated a predictor based on diverse data types for enhancer regions. Future directions include developing methods to link enhancer regions to specific genes, and incorporating gene expression data to examine the relationship between sequence, observed binding and gene expression. Examination of finer relationships between sequence and observed binding, such as the location, spacing, and orientation of sequence motifs, would facilitate an understanding of the relationship between sequence and expression. Highly specific, precise, and reproducible gene expression, both in space and time, is necessary for critical biological processes such as development. Thus, it is likely that there are subtle relationships between sequence and binding sites relevant to the control of gene expression that could be elucidated from protein binding data.

Chapter 6

Appendix A

6.0.1 Table 1: Compendium of transcription factor binding experiments

Table 6.1: Compendium of transcription factor binding experiments

	factor	timepoint/cell line	lab	reference
1	BEAF-32	E0-12hr	kw	[8]
2	BEAF-32	Kc	vc	[4]
3	BEAF-32	Mbn2	vc	[4]
4	BEAF-32	S2	gk	[8]
5	CBP	E0-4hr	kw	[8]
6	CTCF	BG3	gk	[8]
7	CTCF	BG3	gk	[8]
8	CTCF	BG3	gk	[8]
9	CTCF	BG3	gk	[8]
10	CTCF	E0-12hr	kw	[8]
11	CTCF	E0-12hr	kw	[8]
12	CTCF	Kc	kw	[8]
13	CTCF	Kc	vc	[4]
14	CTCF	Mbn2	vc	[4]
15	CTCF	S2	gk	[8]
16	CTCF	S2	gk	[8]
17	CTCF	S2	gk	[8]
18	CTCF	S2	kw	[8]
19	CTCF	S2	kw	[8]

20	CTCF	S2	kw	[8]
21	Chro	BG3	gk	[8]
22	Chro	CL8	gk	[8]
23	Chro	Kc	gk	[8]
24	Chro	S2	gk	[8]
25	Chro	S2	gk	[8]
26	Cp190	E0-12hr	kw	[8]
27	Cp190	Kc	vc	[4]
28	Cp190	Mbn2	vc	[4]
29	Cp190	S2	gk	[8]
30	CtBp	E0-12hr	kw	[8]
31	D	E0-12hr	kw	[8]
32	D	E0-8hr	kw	[8]
33	D	E2-3hr	b	[31]
34	Dll	E0-12hr	kw	[8]
35	Dsp1	E4-12hr	cv	[40]
36	E(z)	E8-16hr	kw	[8]
37	E(z)	S2	gk	[8]
38	E2f2	Kc	ma	[15]
39	EcR	L3	kw	[8]
40	EcR	WPP10-11hr	kw	[8]
41	EcR	WPP30-33hr	kw	[8]
42	GATAe	E0-8hr	kw	[8]
43	HP1	S2	gk	[8]
44	Kr	E0-8hr	kw	[8]
45	Kr	E2-3hr	b	[28]
46	Kr	E2-3hr	b	[28]
47	Kr	Kc	kw	[8]
48	Mad	E2-4hr	z	[51]
49	Mad	E2.5-3.5hr	b	[31]
50	Med	E2.5-3.5hr	b	[31]
51	Mef2	E10-12hr	f	[39]
52	Mef2	E10-12hr	f	[39]
53	Mef2	E2-4hr	f	[39]
54	Mef2	E2-4hr	f	[39]
55	Mef2	E4-6hr	f	[39]
56	Mef2	E4-6hr	f	[39]
57	Mef2	E6-8hr	f	[39]
58	Mef2	E6-8hr	f	[39]
59	Mef2	E8-10hr	f	[39]
60	Mef2	E8-10hr	f	[39]
61	Myb	Kc	ma	[15]
62	NELF-B	S2	b	[27]
63	Nelf-E	S2	b	[27]

64	O-GlcNAc	L3:disc	jm	[13]
65	Pc	BG3	gk	[8]
66	Pc	E0-16hr	rw	[26]
67	Pc	E4-12hr	cv	[40]
68	Pc	L3:haltere	rw	[26]
69	Pc	S2	gk	[8]
70	Sfmbt	L3:disc	jm	[34]
71	Snr1	pupae	kw	[8]
72	Stat92E	E0-12hr	kw	[8]
73	Trl	BG3	gk	[8]
74	Trl	BG3	gk	[8]
75	Trl	E0-12hr	kw	[8]
76	Trl	E4-12hr	cv	[40]
77	Trl	Kc	kw	[8]
78	Trl	S2	b	[31]
79	Trl	S2	gk	[8]
80	Trl	S2	gk	[8]
81	Trl	S2	gk	[8]
82	Ubx	E0-12hr	kw	[8]
83	Ubx	E3-8hr	kw	[8]
84	Ubx	E3-8hr	kw	[8]
85	Ubx	E3-8hr	kw	[8]
86	bab1	E0-12h	kw	[8]
87	bab1	pupae	kw	[8]
88	bap	E6-8hr	f	[23]
89	bap	E6-8hr	f	[23]
90	bcd	E2-3hr	b	[28]
91	bcd	E2-3hr	b	[28]
92	bin	E10-12hr	f	[23]
93	bin	E10-12hr	f	[23]
94	bin	E12-14hr	f	[23]
95	bin	E6-8hr	f	[23]
96	bin	E6-8hr	f	[23]
97	bin	E8-10hr	f	[23]
98	bin	E8-10hr	f	[23]
99	brm	pupae	kw	[8]
100	cad	AdultFemale	kw	[8]
101	cad	E0-4hr	kw	[8]
102	cad	E0-4hr	kw	[8]
103	cad	E0-4hr	kw	[8]
104	cad	E2-3hr	b	[28]
105	cad	E4-8hr	kw	[8]
106	cad	E4-8hr	kw	[8]
107	chinmo	E0-12hr	kw	[8]

108	cnc	E0-12hr	kw	[8]
109	da	E2-3hr	b	[31]
110	disco	E0-8hr	kw	[8]
111	dl	E2-3hr	b	[31]
112	dl	E2-4hr:Toll10b	z	[51]
113	en	E0-12hr	kw	[8]
114	en	E7-24hr	kw	[8]
115	eve	E1-6hr	kw	[8]
116	ftz-fl	E0-12hr	kw	[8]
117	ftz	E2.5-3.5hr	b	[31]
118	gro	E0-12hr	kw	[8]
119	gro	E0-12hr	kw	[8]
120	gro	E0-12hr	kw	[8]
121	gsb-n	E0-12hr	kw	[8]
122	gsb-n	E7-24hr	kw	[8]
123	gt	E2-3hr	b	[28]
124	h	E0-8hr	kw	[8]
125	h	E2.5-3.5hr	b	[31]
126	h	E2.5-3.5hr	b	[31]
127	hb	E2-3hr	b	[28]
128	hb	E2-3hr	b	[28]
129	hkb	E0-8hr	kw	[8]
130	hkb	E2-3hr	b	[31]
131	hkb	E2-3hr	b	[31]
132	hkb	E2-3hr	b	[31]
133	insv	E0-12hr	kw	[8]
134	insv	E2-6hr	kw	[8]
135	insv	E6-10hr	kw	[8]
136	inv	E0-12hr	kw	[8]
137	inv	E0-12hr	kw	[8]
138	jumu	E0-8hr	kw	[8]
139	kn	E0-12hr	kw	[8]
140	kni	E2-3hr	b	[28]
141	kni	E2-3hr	b	[28]
142	lin-52	Kc	ma	[15]
143	mip120	Kc	ma	[15]
144	mip130	Kc	ma	[15]
145	mod(mdg4)	BG3	gk	[8]
146	mod(mdg4)	E0-12hr	kw	[8]
147	mod(mdg4)	E8-16hr	kw	[8]
148	ph-p	E4-12hr	cv	[40]
149	ph-p	L3:disc	jm	[13]
150	pho	E0-16hr	rw	[26]
151	pho	E4-12hr	cv	[40]

152	pho	E6-12hr	jm	[34]
153	pho	L3:disc	jm	[34]
154	pho	L3:haltere	rw	[26]
155	phol	E4-12hr	cv	[40]
156	prd	E2.5-3.5hr	b	[31]
157	prd	E2.5-3.5hr	b	[31]
158	run	E0-12hr	kw	[8]
159	run	E2.5-3.5hr	b	[31]
160	run	E2.5-3.5hr	b	[31]
161	sbb	E0-12hr	kw	[8]
162	sbb	E0-4hr	kw	[8]
163	sens	E4-8hr	kw	[8]
164	sens	E4-8hr	kw	[8]
165	sens	E4-8hr	kw	[8]
166	shn	E2.5-3.5hr	b	[31]
167	shn	E2.5-3.5hr	b	[31]
168	slp1	E2.5-3.5hr	b	[31]
169	sna	E2-3hr	b	[31]
170	sna	E2-3hr	b	[31]
171	sna	E2-4hr:Toll10b	z	[51]
172	su(Hw)	E0-12hr	kw	[8]
173	su(Hw)	E0-12hr	kw	[8]
174	su(Hw)	Kc	vc	[4]
175	su(Hw)	Mbn2	vc	[4]
176	su(Hw)	S2	gk	[8]
177	su(Hw)	S2	gk	[8]
178	tin	E2-4hr	f	[30]
179	tin	E4-6hr	f	[30]
180	tin	E6-8hr	f	[30]
181	tll	E2-3hr	b	[31]
182	trx	E4-12hr	cv	[40]
183	trx	E4-12hr	cv	[40]
184	trx	S2	gk	[8]
185	ttk	E0-12hr	kw	[8]
186	twi	E2-3hr	b	[28]
187	twi	E2-3hr	b	[28]
188	twi	E2-4hr:Toll10b	z	[51]
189	twi	E2-4hr	f	[38]
190	twi	E2-4hr	f	[38]
191	twi	E4-6hr	f	[38]
192	twi	E4-6hr	f	[38]
193	twi	E6-8hr	f	[38]
194	z	E7.5-9.5hr	b	[32]
195	zfh1	E0-12hr	kw	[8]

(Lab key: kw = Kevin White; gk = Gary Karpen; b = Mark Biggin; f = Eileen Furlong; z = Julia Zeitlinger; vc = Victor Corces; rw = Robert White; ma = David Macalpine; cv = Giacomo Cavalli; jm = Jurg Muller)

6.0.2 Table 2: Chromatin mark timecourse (Kevin White group)

Table 6.2: Compendium of chromatin marks (Kevin White group)[8]

	factor	timepoint/cell line	platform
1	CBP	AdultFemale	Array
2	CBP	AdultFemale	Seq
3	CBP	AdultMale	Array
4	CBP	AdultMale	Seq
5	CBP	E0-4hr	Array
6	CBP	E0-4hr	Seq
7	CBP	E12-16hr	Array
8	CBP	E12-16hr	Seq
9	CBP	E16-20hr	Array
10	CBP	E16-20hr	Seq
11	CBP	E20-24hr	Array
12	CBP	E20-24hr	Seq
13	CBP	E4-8hr	Array
14	CBP	E4-8hr	Seq
15	CBP	E8-12hr	Array
16	CBP	E8-12hr	Array
17	CBP	L1	Array
18	CBP	L1	Seq
19	CBP	L3	Array
20	CBP	L3	Seq
21	CBP	Pupae	Array
22	CBP	Pupae	Seq
23	H3K27ac	AdultFemale	Seq
24	H3K27ac	AdultMale	Array
25	H3K27ac	AdultMale	Array
26	H3K27ac	AdultMale	Seq
27	H3K27ac	E0-4hr	Array

28	H3K27ac	E0-4hr	Array
29	H3K27ac	E0-4hr	Seq
30	H3K27ac	E12-16hr	Array
31	H3K27ac	E12-16hr	Seq
32	H3K27ac	E16-20hr	Array
33	H3K27ac	E16-20hr	Seq
34	H3K27ac	E20-24hr	Array
35	H3K27ac	E20-24hr	Seq
36	H3K27ac	E4-8hr	Array
37	H3K27ac	E4-8hr	Seq
38	H3K27ac	E8-12hr	Array
39	H3K27ac	E8-12hr	Seq
40	H3K27ac	L1	Array
41	H3K27ac	L1	Seq
42	H3K27ac	L2	Array
43	H3K27ac	L2	Seq
44	H3K27ac	L2	Seq
45	H3K27ac	L3	Array
46	H3K27ac	L3	Seq
47	H3K27ac	Pupae	Array
48	H3K27ac	Pupae	Seq
49	H3K27me3	AdultFemale	Array
50	H3K27me3	AdultMale	Array
51	H3K27me3	AdultMale	Array
52	H3K27me3	AdultMale	Seq
53	H3K27me3	E0-4hr	Array
54	H3K27me3	E0-4hr	Seq
55	H3K27me3	E12-16hr	Array
56	H3K27me3	E12-16hr	Seq
57	H3K27me3	E16-20hr	Array
58	H3K27me3	E16-20hr	Seq
59	H3K27me3	E20-24hr	Array
60	H3K27me3	E20-24hr	Seq
61	H3K27me3	E4-8hr	Array
62	H3K27me3	E4-8hr	Seq
63	H3K27me3	E8-12hr	Array
64	H3K27me3	E8-12hr	Seq
65	H3K27me3	L1	Array
66	H3K27me3	L1	Seq
67	H3K27me3	L2	Array
68	H3K27me3	L2	Seq
69	H3K27me3	L3	Array
70	H3K27me3	L3	Seq
71	H3K27me3	Pupae	Seq

72	H3K4me1	AdultFemale	Array
73	H3K4me1	AdultFemale	Seq
74	H3K4me1	AdultMale	Array
75	H3K4me1	AdultMale	Seq
76	H3K4me1	E0-4hr	Array
77	H3K4me1	E0-4hr	Seq
78	H3K4me1	E12-16hr	Array
79	H3K4me1	E12-16hr	Seq
80	H3K4me1	E16-20hr	Array
81	H3K4me1	E16-20hr	Seq
82	H3K4me1	E20-24hr	Array
83	H3K4me1	E20-24hr	Seq
84	H3K4me1	E4-8hr	Array
85	H3K4me1	E4-8hr	Seq
86	H3K4me1	E8-12hr	Array
87	H3K4me1	E8-12hr	Seq
88	H3K4me1	L1	Array
89	H3K4me1	L1	Seq
90	H3K4me1	L2	Array
91	H3K4me1	L2	Seq
92	H3K4me1	L3	Array
93	H3K4me1	L3	Seq
94	H3K4me1	Pupae	Array
95	H3K4me1	Pupae	Seq
96	H3K4me3	AdultFemale	Array
97	H3K4me3	AdultFemale	Seq
98	H3K4me3	AdultMale	Array
99	H3K4me3	AdultMale	Seq
100	H3K4me3	E0-4hr	Array
101	H3K4me3	E0-4hr	Seq
102	H3K4me3	E12-16hr	Array
103	H3K4me3	E12-16hr	Seq
104	H3K4me3	E16-20hr	Array
105	H3K4me3	E16-20hr	Seq
106	H3K4me3	E20-24hr	Array
107	H3K4me3	E20-24hr	Seq
108	H3K4me3	E4-8hr	Array
109	H3K4me3	E4-8hr	Seq
110	H3K4me3	E8-12hr	Array
111	H3K4me3	E8-12hr	Seq
112	H3K4me3	L1	Array
113	H3K4me3	L1	Seq
114	H3K4me3	L2	Array
115	H3K4me3	L2	Seq

116	H3K4me3	L3	Array
117	H3K4me3	L3	Seq
118	H3K4me3	Pupae	Array
119	H3K4me3	Pupae	Seq
120	H3K9ac	AdultFemale	Array
121	H3K9ac	AdultFemale	Seq
122	H3K9ac	AdultMale	Array
123	H3K9ac	AdultMale	Array
124	H3K9ac	AdultMale	Seq
125	H3K9ac	E0-4hr	Array
126	H3K9ac	E0-4hr	Seq
127	H3K9ac	E12-16hr	Array
128	H3K9ac	E12-16hr	Seq
129	H3K9ac	E16-20hr	Array
130	H3K9ac	E16-20hr	Seq
131	H3K9ac	E20-24hr	Array
132	H3K9ac	E20-24hr	Seq
133	H3K9ac	E4-8hr	Array
134	H3K9ac	E4-8hr	Seq
135	H3K9ac	E8-12hr	Array
136	H3K9ac	E8-12hr	Seq
137	H3K9ac	L1	Array
138	H3K9ac	L1	Seq
139	H3K9ac	L2	Array
140	H3K9ac	L2	Seq
141	H3K9ac	L3	Array
142	H3K9ac	L3	Seq
143	H3K9ac	Pupae	Array
144	H3K9ac	Pupae	Seq
145	H3K9me3	AdultFemale	Array
146	H3K9me3	AdultMale	Array
147	H3K9me3	E0-4hr	Array
148	H3K9me3	E0-4hr	Array
149	H3K9me3	E0-4hr	Seq
150	H3K9me3	E12-16hr	Array
151	H3K9me3	E12-16hr	Seq
152	H3K9me3	E16-20hr	Array
153	H3K9me3	E16-20hr	Seq
154	H3K9me3	E20-24hr	Array
155	H3K9me3	E20-24hr	Seq
156	H3K9me3	E4-8hr	Array
157	H3K9me3	E4-8hr	Seq
158	H3K9me3	E8-12hr	Array
159	H3K9me3	E8-12hr	Seq

160	H3K9me3	L1	Array
161	H3K9me3	L1	Seq
162	H3K9me3	L2	Array
163	H3K9me3	L2	Seq
164	H3K9me3	L3	Array
165	H3K9me3	Pupae	Array
166	H3K9me3	Pupae	Seq
167	PolII	AdultFemale	Array
168	PolII	E0-4hr	Array
169	PolII	E12-16hr	Array
170	PolII	E12-16hr	Seq
171	PolII	E16-20hr	Array
172	PolII	E16-20hr	Seq
173	PolII	E20-24hr	Seq
174	PolII	E4-8hr	Array
175	PolII	E4-8hr	Seq
176	PolII	E8-12hr	Seq
177	PolII	L1	Seq
178	PolII	L2	Array
179	PolII	L2	Array
180	PolII	L2	Seq
181	PolII	L3	Array
182	PolII	L3	Array
183	PolII	L3	Seq
184	PolII	Pupae	Array
185	PolII	Pupae	Seq

6.0.3 Table 3: Chromatin marks across cell lines (Gary Karpen group)

Table 6.3: Compendium of chromatin marks (Gary Karpen group)[8]

	factor	timepoint/cell line
1	Chro	BG3
2	Chro	S2
3	H1	BG3
4	H1	S2
5	H2BK5ac	S2
6	H2Bubiq	BG3

7	H2Bubiq	S2
8	H3K18ac	BG3
9	H3K18ac	S2
10	H3K23ac	BG3
11	H3K23ac	S2
12	H3K27ac	BG3
13	H3K27ac	S2
14	H3K27me3	BG3
15	H3K27me3	S2
16	H3K36me1	BG3
17	H3K36me1	S2
18	H3K36me3	BG3
19	H3K36me3	S2
20	H3K4me1	BG3
21	H3K4me1	S2
22	H3K4me2	BG3
23	H3K4me2	S2
24	H3K4me3	BG3
25	H3K4me3	S2
26	H3K79me1	BG3
27	H3K79me1	S2
28	H3K79me2	BG3
29	H3K79me2	S2
30	H3K9ac	S2
31	H3K9me2	BG3
32	H3K9me2	S2
33	H3K9me3	BG3
34	H3K9me3	S2
35	H4acTetra	S2
36	H4	BG3
37	H4K12ac	S2
38	H4K16ac	BG3
39	H4K16ac	S2
40	H4K5ac	S2
41	H4K8ac	S2
42	H4	S2
43	HP1	BG3
44	HP1c	BG3
45	HP1c	S2
46	HP1	S2
47	Pc	BG3
48	Pc	S2
49	RpII	BG3
50	RpII	S2

51	Su(var)3-9	BG3
52	Su(var)3-9	S2

Bibliography

- [1] M.A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, 2004.
- [2] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, and M.B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757, 2002.
- [3] B.E. Bernstein, T.S. Mikkelsen, X. Xie, M. Kamal, D.J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.
- [4] A.M. Bushey, E. Ramos, and V.G. Corces. Three subclasses of a *Drosophila* insulator show distinct and cell type-specific genomic distributions. *Genes & development*, 23(11):1338, 2009.
- [5] R.A. Cameron and E.H. Davidson. Flexibility of transcription factor target site position in conserved cis-regulatory modules. *Developmental biology*, 336(1):122–135, 2009.
- [6] J.A. Campos-Ortega and V. Hartenstein. *The embryonic development of Drosophila melanogaster*. Springer Berlin, 1997.
- [7] M. Carey, Craig L. Peterson, and S. T. Smale. *Transcriptional regulation in eukaryotes: concepts, strategies, and techniques*. Cold Spring Harbor Laboratory Pr, 2009.
- [8] S. Celniker, L. Dillon, M. Gerstein, K. Gunsalus, S. Henikoff, G. Karpen, M. Kellis, E. Lai, J. Lieb, D. MacAlpine, et al. modENCODE Consortium: Unlocking the secrets of the genome. *Nature*, 459:927–930, 2009.
- [9] B.Y. Chan and D. Kibler. Using hexamers to predict cis-regulatory motifs in *Drosophila*. *BMC bioinformatics*, 6(1):262, 2005.
- [10] H.A. Collier and L. Kruglyak. Its the Sequence, Stupid! *Science*, 322:380–381, 2008.

- [11] J. Ernst, H.L. Plasterer, I. Simon, and Z. Bar-Joseph. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research*, 2010.
- [12] M.C. Frith, U. Hansen, and Z. Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878, 2001.
- [13] M.C. Gambetta, K. Oktaba, and J. Muller. Essential role of the glycosyltransferase Sxc/Ogt in Polycomb repression. *Science*, 325(5936):93, 2009.
- [14] A. Garcia-Bellido and P. Santamaria. Developmental analysis of the wing disc in the mutant engrailed of *Drosophila melanogaster*. *Genetics*, 72(1):87, 1972.
- [15] D. Georlette, S. Ahn, D.M. MacAlpine, E. Cheung, P.W. Lewis, E.L. Beall, S.P. Bell, T. Speed, J.R. Manak, and M.R. Botchan. Genomic profiling and expression studies reveal both positive and negative activities for the *Drosophila* Myb–MuvB/dREAM complex in proliferating cells. *Genes & development*, 21(22):2880, 2007.
- [16] J. Gertz, E.D. Siggia, and B.A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218, 2008.
- [17] Y.H. Grad, F.P. Roth, M.S. Halfon, and G.M. Church. Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*. *Bioinformatics*, 20(16):2738, 2004.
- [18] Marc S. Halfon, Steven M. Gallo, and Casey M. Bergman. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*. *Nucleic Acids Research*, 2007.
- [19] P. Hatzis and I. Talianidis. Dynamics of enhancer-promoter communication during differentiation-induced gene activation. *Molecular cell*, 10(6):1467–1477, 2002.
- [20] N.D. Heintzman, R.K. Stuart, G. Hon, Y. Fu, C.W. Ching, R.D. Hawkins, L.O. Barrera, S. Van Calcar, C. Qu, K.A. Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–318, 2007.
- [21] A.I. Ivanov, A.C. Rovescalli, P. Pozzi, S. Yoo, B. Mozer, H.P. Li, S.H. Yu, H. Higashida, V. Guo, M. Spencer, et al. Genes required for *Drosophila* nervous system development identified by RNA interference. *Proceedings of the National Academy of Sciences*, 101(46):16216, 2004.
- [22] K. Jagla, M. Bellard, and M. Frasch. A cluster of *Drosophila* homeobox genes involved in mesoderm differentiation programs. *Bioessays*, 23(2):125–133, 2001.

- [23] J.S. Jakobsen, M. Braun, J. Astorga, E.H. Gustafson, T. Sandmann, M. Karzynski, P. Carlsson, and E.E.M. Furlong. Temporal CHIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes & development*, 21(19):2448, 2007.
- [24] D.M. Jeziorska, K.W. Jordan, and K.W. Vance. A systems biology approach to understanding cis-regulatory module function. In *Seminars in Cell & Developmental Biology*, volume 20, pages 856–862. Elsevier, 2009.
- [25] P. Kheradpour, A. Stark, S. Roy, and M. Kellis. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome research*, 17(12):1919, 2007.
- [26] C. Kwong, B. Adryan, I. Bell, L. Meadows, S. Russell, J.R. Manak, and R. White. Stability and dynamics of polycomb target sites in Drosophila development. *PLoS Genetics*, 4(9), 2008.
- [27] C. Lee, X. Li, A. Hechmer, M. Eisen, M.D. Biggin, B.J. Venters, C. Jiang, J. Li, B.F. Pugh, and D.S. Gilmour. NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Molecular and cellular biology*, 28(10):3290, 2008.
- [28] X. Li, S. MacArthur, R. Bourgon, D. Nix, D.A. Pollard, V.N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C.L. Luengo Hendriks, et al. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol*, 6(2):e27, 2008.
- [29] A. Lindlof, M. Brautigam, A. Chawade, O. Olsson, and B. Olsson. In silico analysis of promoter regions from cold-induced genes in rice (*Oryza sativa* L.) and *Arabidopsis thaliana* reveals the importance of combinatorial control. *Bioinformatics*, 25(11):1345, 2009.
- [30] Y.H. Liu, J.S. Jakobsen, G. Valentin, I. Amarantos, D.T. Gilmour, and E.E.M. Furlong. A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. *Developmental cell*, 16(2):280–291, 2009.
- [31] S. MacArthur, X.Y. Li, J. Li, J. Brown, H.C. Chu, L. Zeng, B. Grondona, A. Hechmer, L. Simirenko, S. Keranen, et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology*, 10(7):R80, 2009.
- [32] A.M. Moses, D.A. Pollard, D.A. Nix, V.N. Iyer, X.Y. Li, M.D. Biggin, and M.B. Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10):e130, 2006.
- [33] A.G. Nazina and D.A. Papatsenko. Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC bioinformatics*, 4(1):65, 2003.

- [34] K. Oktaba, L. Gutiérrez, J. Gagneur, C. Girardot, A.K. Sengupta, E.E.M. Furlong, and J. Muller. Dynamic regulation by polycomb group protein complexes controls pattern formation and the cell cycle in *Drosophila*. *Developmental cell*, 15(6):877–889, 2008.
- [35] D. Papatsenko, Y. Goltsev, and M. Levine. Organization of developmental enhancers in the *Drosophila* embryo. *Nucleic Acids Research*, 37(17):5665, 2009.
- [36] N. Pierstorff, C.M. Bergman, and T. Wiehe. Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, 22(23):2858, 2006.
- [37] N. Rajewsky, M. Vergassola, U. Gaul, and E.D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC bioinformatics*, 3(1):30, 2002.
- [38] T. Sandmann, C. Girardot, M. Brehme, W. Tongprasit, V. Stolc, and E.E.M. Furlong. A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. *Genes & development*, 21(4):436, 2007.
- [39] T. Sandmann, L.J. Jensen, J.S. Jakobsen, M.M. Karzynski, M.P. Eichenlaub, P. Bork, and E.E.M. Furlong. A temporal map of transcription factor activity: *mef2* directly regulates target genes at all stages of muscle development. *Developmental cell*, 10(6):797–807, 2006.
- [40] B. Schuettengruber, M. Ganapathi, B. Leblanc, M. Portoso, R. Jaschek, B. Tolhuis, M. Van Lohuizen, A. Tanay, and G. Cavalli. Functional anatomy of polycomb and trithorax chromatin landscapes in *Drosophila* embryos. *PLoS Biol*, 7:e13, 2009.
- [41] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451(7178):535–540, 2008.
- [42] A. Siepel, G. Bejerano, J.S. Pedersen, A.S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L.D.W. Hillier, S. Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034, 2005.
- [43] S. Sinha, Y. Liang, and E. Siggia. Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic acids research*, 34(Web Server issue):W555, 2006.
- [44] D. Thanos and T. Maniatis. Virus induction of human IFN [beta] gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–1100, 1995.
- [45] Kumiko Ui, Shoko Nishihara, M. Sakuma, S. Togashi, R. Ueda, Y. Miyata, and T. Miyake. Newly established cell lines from *drosophila* larval CNS express neural specific characteristics. *In Vitro Cell Dev. Biol.*, 1994.

- [46] A. Visel, M.J. Blow, Z. Li, T. Zhang, J.A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, 2009.
- [47] S.X. Wang, P.K. Elder, Y. Zheng, A.R. Strauch, and R.J. Kelm. Cell cycle-mediated regulation of smooth muscle α -actin gene transcription in fibroblasts and vascular smooth muscle cells involves multiple adenovirus E1A-interacting cofactors. *Journal of Biological Chemistry*, 280(7):6204, 2005.
- [48] M.D. Wilson, N.L. Barbosa-Morais, D. Schmidt, C.M. Conboy, L. Vanes, V.L.J. Tybulewicz, E. Fisher, S. Tavare, and D.T. Odom. Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900):434, 2008.
- [49] I.H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S.J. Cunningham. Weka: Practical machine learning tools and techniques with Java implementations. In *ICONIP/ANZIIS/ANNES*, volume 99, pages 192–196. Citeseer, 1999.
- [50] X. Xie, J. Lu, EJ Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.
- [51] J. Zeitlinger, R.P. Zinzen, A. Stark, M. Kellis, H. Zhang, R.A. Young, and M. Levine. Whole-genome ChIP–chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the *Drosophila* embryo. *Genes & development*, 21(4):385, 2007.
- [52] H. Zhao and A. Dean. Organizing the genome: enhancers and insulators. *Biochemistry and Cell Biology*, 83(4):516–524, 2005.
- [53] Q. Zhou and J.S. Liu. Extracting sequence features to predict protein-DNA interactions: a comparative study. *Nucleic acids research*, 2008.
- [54] Q. Zhou and W.H. Wong. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33):12114, 2004.
- [55] Robert P. Zinzen, Charles Giradot, Julien Gagneur, Martina Braun, and Eileen E. M. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 2009.