DISSERTATION

UNCOVERING THE ROLE OF EPIGENETICS IN ALTERNATIVE SPLICING

Submitted by

Fahad Ullah

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2020

Doctoral Committee:

Advisor: Asa Ben-Hur

Charles Anderson
Hamidreza Chitsaz
Anireddy SN Reddy

ABSTRACT

UNCOVERING THE ROLE OF EPIGENETICS IN ALTERNATIVE SPLICING

Alternative Splicing (AS) is a regulated phenomenon that enables a single gene to encode structurally and functionally different biomolecules (proteins, non-coding RNAs etc.), that play important roles in an organism's development and growth. Besides, it has been implicated in multiple diseases including cancer, thalassemia, and spinal muscular atrophy. Recent studies have shown that AS is widespread in both plants and animals. Moreover, it has been reported that splicing occurs co-transcriptionally and that chromatin state is important for understanding the regulation of AS. Most of the previous efforts made to elucidate the regulation of AS used sequence information alone. However, in this study our goal is to understand AS from an epigenetic perspective: how chromatin organization, accessibility, and modifications are involved in its regulation.

Intron Retention (IR) is the most frequent form of AS in plants, however, very little is known about its regulation, particularly regarding the role of chromatin state. Therefore, as a first step, we investigate the relationship between IR and chromatin accessibility in two plant species: arabidopsis and rice. We report a strong association between chromatin accessibility and IR. Our findings suggest that chromatin is more open and accessible in IR. Furthermore, we discover motifs associated with the regulation of alternative and constitutively spliced introns, many of which match those of known transcription factors and are conserved between arabidopsis and rice, a strong indication of their functional importance.

Recent studies have suggested that IR is highly prevalent in humans as well. Using the plethora of genomic data that is available in human, we design a deep learning model for predicting IR in regions of open chromatin. Our model exhibits good accuracy in terms of Area Under the ROC Curve (AUC), with median AUC = 0.80. Moreover, we identify motifs enriched in IR events with significant hits to known human transcription factors (TFs). The zinc finger family exhibits the

highest activity in IR events, a prediction that is validated using ChIP-Seq data. Experiments by our collaborators have validated our predictions in candidate IR events.

Finally, as an effort to capture the complete regulatory landscape of alternative splicing, we investigate the cooperativity and interactions between regulatory sequence features. To that end, we design a self-attention model that combines convolutional and recurrent layers with a self-attention layer that helps us capture a global view of the landscape of interactions between regulatory elements in a sequence. We evaluate our method on several datasets and compare it to existing methodology. In each experiment, our model identifies numerous statistically significant TF interactions, many of which have been previously reported. Finally, using this model with the chromatin accessibility in IR dataset, we identify many interactions primarily involving the zinc finger family of transcription factors. Our approach not only provides a global, biologically relevant set of interactions but, unlike existing methods, it does not require a computationally expensive postprocessing step.

In summary, this dissertation sheds light on the epigenetic regulation of alternative splicing by transcription factors, and also contributes methodologically by making the results of deep learning models more interpretable.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Dr. Asa Ben-Hur, for his invaluable support throughout my PhD career; without his insights and guidance, this would have not been possible. I am also indebted to Dr. ASN Reddy who has always been there to help and guide me in understanding the key biological concepts relevant to my research. Finally, I thank the rest of my doctoral committee members, Dr. Charles Anderson and Dr. Hamidreza Chitsaz, for their time, valuable suggestion, and feedback.

I would like to convey my gratitude to Mike Hamilton, a member of our research group, who helped me during the early years of my PhD. Besides, I would also thank the rest of past and current members of our group for their support.

In 2019, I was awarded the Wim Bohm and Partners PhD Support fellowship which helped expediting the completion of my PhD degree. I would like to thank and acknowledge Dr. Wim Bohm for taking that initiative to support PhD students.

Last but not least, I am grateful to my friends and family for their help and support throughout my PhD and my stay here at Fort Collins, Colorado.

DEDICATION

To my late grandmother.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
DEDICATION v
LIST OF TABLES ix
LIST OF FIGURES xii

Chapter 1 Introduction 1
1.1 Overview of chapters 3

Chapter 2 Biological and Bioinformatics Background 5
2.1 Biological background 5
2.2 Bioinformatics background 8
2.2.1 Transcriptome analysis 8
2.2.2 Chromatin profiling 10

Chapter 3 Deep Learning Background 14
3.1 Fully connected networks 14
3.2 Convolutional neural networks 15
3.3 Recurrent neural networks 16
3.4 Self-attention 20

Chapter 4 Related Work: Machine Learning in Genomics 23
4.1 Classical machine learning 23
4.1.1 Kernel based methods 24
4.1.2 Logistic regression 25
4.1.3 Random forests 25
4.2 Deep learning 26

Chapter 5 Intron Retention and Chromatin Accessibility in Plants 29
5.1 Introduction 29
5.2 Results 30
5.2.1 DHSs are enriched in IR events 30
5.2.2 IR events exhibit higher chromatin accessibility than IE events 31
5.2.3 Protein footprint analysis 33
5.3 Discussion 37
5.4 Conclusions 40
5.5 Materials and methods 41
5.5.1 Data collection 41
5.5.2 Alignment and processing 41
5.5.3 Extraction of IR/IE events and peak calling 42
5.5.4 Protein footprint analysis 42
5.5.5 Statistical tests 45

Chapter 6 Predicting Intron Retention using Deep Learning . . . 46
6.1 Introduction . . . 46
6.2 Methods . . . 48
6.2.1 Data collection, processing, and representation . . . 48
6.2.2 Network architecture . . . 49
6.2.3 Network training and evaluation . . . 51
6.2.4 Gapped kmer SVM . . . 51
6.2.5 Motif extraction and analysis . . . 51
6.2.6 TF ChIP-Seq analysis . . . 52
6.3 Results . . . 52
6.3.1 Predicting DHSs associated with IR . . . 52
6.3.2 Embeddings lead to poor interpretability . . . 53
6.3.3 The zinc finger transcription factor family is enriched in IR events . . . 55
6.3.4 Evidence from Chip-Seq data . . . 55
6.3.5 Experimental validation . . . 57
6.4 Discussion . . . 58

Chapter 7 A Self-Attention Model for Inferring Regulatory Interactions . . . 61
7.1 Introduction . . . 61
7.2 Methods . . . 63
7.2.1 Model architecture . . . 63
7.2.2 Network training and evaluation . . . 66
7.2.3 Motif extraction . . . 66
7.2.4 Quantifying feature interactions . . . 67
7.2.5 Data collection and processing . . . 70
7.3 Results and Discussion . . . 70
7.3.1 Benchmark 1: embedded motif interactions in simulated sequences . . . 70
7.3.2 Benchmark 2: Inferring TAL-GATA motif interactions from ChIP-Seq data . . . 70
7.3.3 The TF interaction landscape across human promoters . . . 72
7.3.4 Genome-wide regulatory interactions in arabidopsis . . . 74
7.3.5 Comparison: SATORI and FIS-based interactions . . . 75
7.3.6 Regulatory interactions in IR events in human . . . 77
7.4 Conclusions and Future Work . . . 78

Chapter 8 Conclusions . . . 81
8.1 Open Problems . . . 81
8.1.1 Predict chromatin accessibility in AS in plants . . . 82
8.1.2 Investigate epigenetic regulation of other forms of AS . . . 82
8.1.3 Use evidence from other chromatin marks . . . 83
8.1.4 Towards a comprehensive epigenetic splicing code: tissue and condition specific splicing . . . 84

Bibliography . . . 86

Appendix A Chapter 5 Supplementary Material . . . 111

Appendix B Chapter 6 Supplementary Material . . . 121
B.1 Generating transcription factor family distributions . . . 121

Appendix C Chapter 7 Supplementary Material . . . 124
C.1 Data collection and processing . . . 124
C.1.1 Experiment 1: simulated dataset . . . 124
C.1.2 Experiment 2: TAL-GATA ChIP-peaks . . . 124
C.1.3 Experiment 3: human promoter DHSs . . . 125
C.1.4 Experiment 4: genome-wide arabidopsis regions of open chromatin . . . 126
C.2 Limitations of the TomTom motif comparison tool . . . 126
C.3 Additional tables . . . 127

Appendix D Chapter 8 Supplementary Material . . . 133
D.1 Methods: Differential AS and chromatin accessibility . . . 133
D.1.1 Data collection and processing . . . 133
D.1.2 Differential IR and chromatin accessibility analysis . . . 133

LIST OF TABLES

2.1 AS landscape for 5 different eukaryote species where the number of splicing events are provided for each type of AS. In general, exon skipping is prevalent in mammalian species whereas intron retention is the most common form of splicing in plants. The data was generated using Splicgraher [1] with the following gene annotations: UCSC hg19 (human), UCSC mm9 (mouse), TAIR10 (arabidopsis), MSU v7 (rice), and JGI Phytozome V12 (sorghum). 6

5.1 Enrichment of DHSs in IR and IE events. DHS content is the fraction of IR/IE events with an overlapping hypersensitive site. The significance of the difference is quantified by the Fisher exact test. 32

5.2 Enriched hexamers exhibiting a footprint. For each of the four datasets we provide the number of hexamers that exhibit a footprint and are also enriched in either IR or IE events. The number of enriched footprint-hexamers are shown in each of the three regions of an event: 5' exon, intron and 3' exon. An HMM score cutoff of $S = 0.30$ was used to generate the footprint hexamers. 34

5.3 Common enriched footprint-hexamers between arabidopsis and rice. The number of hexamers in common between the arabidiopsis and rice leaf samples, and the corresponding significance levels of the overlap are shown for all three regions of IR and IE events. The hexamers in each region were clustered, and motif consensus sequences are shown. When there is no clear consensus in a given position, that is denoted by an x. Leading or trailing positions without a clear consensus were omitted, so some consensus sequences are less than 6 nucleotides long. In the intronic region of IR events only 2 hexamers were detected so no clustering was performed. Here, an HMM score cutoff of $S = 0.20$ was used with manual verification of the footprints of the overlapping hexamers. 35

6.1 Enrichment of C2H2 ZF transcription factors binding in IR vs non-IR events quantified using ChIP-Seq peaks of the corresponding TF. 56

7.1 Summary of the datasets used in the four experiments we designed to test and analyze SATORI. The first two datasets have binary labels whereas the last two experiments deal with a multi-label, multi-calss problem. 65

7.2 The most frequent interacting families of human transcription factors in the TAL-GATA ChIP-peaks in human K562 cell-line. All interactions are significant with adjusted p-value < 0.05. 71

8.1 Preliminary results in terms of AUC scores for the three types of alternative splicing (ES, A3, and A5) data used with our deep CNN model. The table also shows the number of positive examples for each dataset. Note that the number of negative examples were roughly twice the size of the positive set. 83

A.1 Alignment statistics for different Arabidopsis thaliana (AT) and rice samples. Note that the aligned reads went through preprocessing and then aligned using tophat2 for RNAseq and bowtie/STAR for DNase I-seq (see Methods in the main text). The reads in both cases (DNase I-seq and RNA-seq) were filtered for multiple alignments and filtered for spurious junctions for the RNAseq. Also, in all samples, biological and technical replicates were pooled. As mentioned in the main text, we used pre-aligned DNase I-seq and RNA-seq from [2]. 116
A.2 Hotspot was used to call DHS peaks in all DNase I-seq samples. Rice samples, on average, had more DHS peaks identified. Since hotspot can't handle replicates, we pooled DNase I-seq libraries. 116
A.3 Number of IR and IE events extracted at different coverage levels are listed below. Evidence from known gene models and RNA-seq data was used to extract the events as described in the Methods section in the main paper. 117
A.4 DHS overlap statistics are shown for the four samples in both IR and IE events at the four levels of read coverage. For both IR and IE events, the number of DHS (peaks) overlapping the events is shown at both individual parts (5' exon, Intron, and 3' Exon) and the whole event (shown in the column titled "All"). Finally, the fisher exact test p-value is shown for each case, indicating that the overlap is significant in IR events in contrast to IE events. 118
A.5 The HMM's transition probabilities for all 13 states. The probabilities were derived from the training data (8 hexamers that were manually detected to have a footprint). Some of the probabilities were manually tweaked to adjust for the noise in our data. The highlighted probabilities (as described in figure 2) are relatively higher than the other transition from the same state. This is to force our HMM to prioritize detection of the primary footprint. 119
A.6 Emissions for the all HMM's 13 states are listed. These emissions are modeled by Gaussian distributions with the corresponding mean and standard deviation (std) shown. Note that these values are derived after standardization of raw hexamer profile coverage to the background score calculated from the training data. The BG_1 and BG_2 (intermediary/secondary backgrounds) were calculated (and tweaked) based on the measured BG_0 and BG_3 values (somewhere in between the two). 120
A.7 The overlap stats between all significantly enriched arabidopsis IR/IE hexamers and transcription factor motifs from Plant Cistrome Database are summarized below. The actual overlaps are provided in the Additional file 3. 120

B.1 List of neural network hyperparameters. 123

C.1 List of neural network hyperparameters. 130
C.2 Summary of all significant interactions in the simulated/toy dataset. Our model is able to recover multiple interactions involving SIX5 and ELF1 TF motifs. We also provide the actual CNN filter interactions in the first column, named based on the total number of filters in the convolutional layer. 131

C.3 All significant interactions between TAL1 and GATA transcription factors. Note that in this case, a custom TF database was used containing motifs for TAL1, GATA1, and GATA2. In case of TAL1, other TFs (LYL1, NHLH2, and TAL2) also shared the same binding site motif and hence are mentioned here. 131
C.4 A list of known TF interactions identified by our model in the human promoter regions. TRRUSTv2 [3] database was used as a reference of all known interactions. The level of significance (adjusted p-value) assigned by our model to each interaction is provided in the last column. 132
C.5 Summary of the number of unique statistically significant TF interactions reported by SATORI and FIS for the three real-world datasets. 132
C.6 A list of known TF interactions in the IR events. TRRUSTv2 [3] database was used as a reference of all known interactions. The level of significance (adjusted p-value) assigned by SATORI to each interaction is provided in the last column. 132

D.1 The overlap between differential IR events with the differentially occuring DHSs in *K562* vs. eight other human cell-lines. The significance of overlap is shown in the last column in terms of p-value (Fisher test). 135

LIST OF FIGURES

2.1 Different types of AS are shown in (a) where introns are represented by black lines connecting the exons. The structure and organization of chromatin within a cell nucleus is shown in (b) along with the accessible regions (DNase I hypersensitive sites) and some of the key factors that regulate gene expression. Adapted with permission from [4]. Copyright 2012, Springer Nature. 7
2.2 The steps of RNA-Seq workflow are shown in (a): RNA extraction, fragmentation, reverse transcription, sequencing, mapping to the genome, and finally quantification of expression (Adapted with permission from [5]. Copyright 2009, Springer Nature). An IR event is shown in (b) with evidence from RNA-Seq data across two biological replicates for human *K562* leukemia cell-line. The coverage plots in green are calculated using the number of reads aligned to that region of the genome. The high RNA-Seq coverage across the intron indicates its retention in both K562 replicates. This figure is generated using the Integrated Genome Viewer [6]. 10
2.3 High throughput massively parallel sequencing experiments for profiling chromatin accessibility and modifications. Adapted with permission from [7]. Copyright 2014, Springer Nature. 11
2.4 An example of chromatin profiling for a hypothetical gene. The gene model annotations are shown in the top row where the boxes represent exons/coding sequences. The corresponding chromatin state is depicted in the second row. Finally, the quantification of chromatin accessibility, histone modifications, and DNA methylation is shown in third, fourth, and fifth row, respectively. 12

3.1 A simple fully connected network, part of a deep learning model [8]. Every unit in layer A is connected to all other units in layer B. Note that a fully connected network is usually followed by a read-out layer (not shown here). 15
3.2 Summary of convolution operation followed by a non-linear ReLU function and max-pooling. In convolution, a filter of fixed size is scanned across the entire input matrix. Next, the ReLU operation gets rid of negative values. Finally, the output is reduced via max-pooling which takes a maximum value across a window of fixed size. 17
3.3 A simple recurrent neural network with sequential processing is shown in (a). In (b), the detailed structure of a long short-term memory unit is depicted. These figures were taken from [9], originally adapted from [10] and [11]. 18
3.4 (a) The process of self-attention is summarized. The attention matrix A helps amplify the input signals that are relevant for the task at hand. (b) An example output, taken from a model that uses self-attention in a machine translation problem. In regards to self-attention, the top row shows the queries whereas the bottom row represents the keys. The results are highlighted for a specific query: the verb “making”. Finally, the attention values are represented by the color boxes in the bottom row, each corresponding to one of the 8 single heads. This figure has been taken from [12], with permission from the author. 22

4.1 Summary of Basset architecture: to predict DHS occupancy across 164 human cell lines, Basset used three convolutional layers followed by multiple fully connected layers. This figure has been taken from [13]. 27

5.1 DHS content profiles in IR and IE. For each sequence bin within an IR/IE event we show the frequency with which that bin overlaps a DHS. Profiles are computed for arabidopsis leaf samples (a), arabidopsis flower samples (b), rice leaf samples (c), and rice callus samples (d). In all samples, we see overall higher DHS occupancy across IR events compared to IE events, suggesting a more open chromatin in IR. Moreover, the DHS content is much higher in the 3' exons of IR events. 31
5.2 Methylation profiles in IR and IE Methylation levels are shown across introns and their flanking exons in IR and IE events in arabidopsis (a) and rice (b). 32
5.3 HMM footprint detection. The hexamer `CCGCCG` was detected to have a footprint at the location of the hexamer (red bar) in the standardized DNase I-seq data profile in both arabidopsis (a) and rice (b). The number of occurrences of the hexamer in IR/IE events is shown next to the k-mer in the title of each sub-figure. The profile extends 100bp upstream and downstream of the hexamer location, and is used by our HMM to score the k-mer for a potential footprint. In both cases, we see a clear dip in coverage indicating a possible footprint at the hexamer location. 33
5.4 Hexamer positional preference. Average positional preference is shown for `AT`-rich footprint-exhibiting hexamers in the 3' exon region of IR events (a), and for comparison, the same hexamers in the 3' exon region of IE events. Similarly, (c) and (d) show average positional preference for `GC`-rich hexamers in the 3' exon region of IR and IE events, respectively. To demonstrate the positional preference of footprint-exhibiting hexamers that are associated with IE events we show the average positional profile of those hexamers in IE events (f) and IR events(e). 36
5.5 Significance of overlap of enriched hexamers. The significance of overlap among enriched hexamers exhibiting a footprint is shown in the 3' exon region of IR events (a), and in the intronic region of IE events (b) for two or more samples. The overlap is shown in circular layout for all possible combinations of two or more of the four samples. The four inner sections of each slice represent the four samples and a sample is labeled in green if it is included in a particular combination. The right most slide provides the labeling of the samples. The size of the fifth section in each slice represents the number of hexamers in an intersection of the corresponding samples. The actual number of overlapping hexamers is also shown. Finally, the color of the fifth section indicates the significance of overlap (p-value). The intersections are sorted based on p-value, starting at the labelled segment in an anti-clockwise fashion. 38
5.6 HMM Architecture The core continuous HMM states used to discover footprints are shown. The five states represent different regions of the DNase I-seq coverage profile: leading background (BG_1), down (DN), footprint (FP), Up (UP), and trailing background (BG_2). The footprint state is shown in the center, within the "dip" in the DNase I-seq coverage. 43

6.1 The distribution of different transcription factor families in the promoter, intragenic, and intergenic regions of the human genome. These statistics were obtained by training the Basset-like network [13] and analyzing the motifs learned by the network (see supplementary methods in Appendix B for more details). 47
6.2 Summary of the different model variants explored when predicting DHS occupancy in IR events. Every architecture is represented by the corresponding colored arrows connecting different network components. The output represents a binary class prediction: IR vs. non-IR DHSs. 49
6.3 ROC and Precision-Recall curves are shown for the different deep learning architectures as well as the gkm-SVM in (a) and (b) respectively. The median AUC and AUPRC values are also provided in the legends. These results were generated using a 10 fold cross validation strategy. 53
6.4 In (a), the mean information content is summarized for different cases: whether we use word2vec embeddings and exponential activations in the first convolutional layer. The distribution of TF families enriched in IR vs non-IR events are summarized in (b). Finally, the top 3 matches (based on the adjusted p-value) for the IR and non-IR convolutional layer filters against the CISBP database are shown in (c). In each match, the target transcription factor motif in the database is shown in the top row whereas the bottom row shows the actual CNN filter/motif. 54
6.5 TF occupancy profiles across IR and non-IR events are shown for two transcription factors, (a) EGR1 and (b) MAZ. To generate the profile, PWM of the corresponding transcription factor was used to score the actual ChIP-Seq peaks (their DNA sequences). The regions with the highest score were then used to determine the TF occupancy within the events. 56
6.6 Evidence of MAZ, a C2H2 ZF transcription factor, regulating intron retention in the human K562 cell line. 58

7.1 Model architecture variants. We use a convolutional layer followed by a multi-head self-attention layer (a); optionally, we add a recurrent layer between the two (b). The input in both cases is a one-hot encoding of the DNA sequence. The output of the model is either be a binary or multi-label prediction. 63
7.2 Summary of the process of inferring interactions from self-attention layer values. For a given example, we collapse the attention heads into a single matrix. Next, at each pair of positions, the corresponding active CNN filters are identified and the attention value is assigned to the interacting pair. This is repeated for all examples to generate interaction profiles for all filter-pairs. Finally, we use a background set to test the significance of filter-filter interactions. 67
7.3 The most frequent TF interactions in human promoters (a). The distribution of TF-TF interaction distances (b). 73
7.4 The regulatory interaction landscape in accessible chromatin in the arabipdosis genome. The most frequently interacting families of plant transcription factors (a). The distribution of distances between inferred TF-TF interactions (b). 74

7.5 Common interactions in the top predictions of SATORI and FIS. Interactions predicted by FIS are sorted by frequency. Those predicted by both methods are shown in blue, and ones predicted only by FIS are shown in red. Top predictions are shown for the TAL-GATA dataset (a) the human promoter dataset (b), and the genomewide arabidopsis dataset (c). For each experiment, the 10 most frequent TF family interactions are shown in (d), (e), and (f) respectively. 76
7.6 Run time in minutes for SATORI and FIS-based interaction estimation for the four datasets. 77
7.7 The most frequent transcription factor interactions in intron retention events are depicted in (a). A majority of these interactions involve C2H2 ZF family. In (b), the distribution of distances is shown for all the statistically significant interactions. 78
7.8 Common interactions in the top predictions of SATORI and FIS for the DHS occupancy in IR dataset are shown in (a). The 10 most frequent TF family interactions are shown in (b). 79

8.1 Differential occurrence of a DHS in a DIR event when comparing (a) the cell-lines *K562* and *H1-hESC*. The DIR event is evident from the RNA-seq coverage plots in two biological replicates of the corresponding cell-lines. The DHS is overlapping the up-regulated IR event (*K562*) while entirely absent in the down-regulated event (*H1-hESC*). That is, the differential DHS and IR event are in the same direction. The opposite direction DHS and IR event are shown in (b) for *K562* and *HCT-116* cell-lines. This plot is generated using Integrated Genome Viewer [6]. 85

A.1 Average DNase I-seq coverage profile is shown across IR and IE events in the four samples: Arabidopsis leaf (a) and flower (b); rice leaf (c) and callus (d). The profile is centered at the 5' and 3' splice sites (indicated by "0" on x-axis in the split figure), and goes 50bp into the intron and 100bp into the flanking exons. Note that we chose all three parts of an IR/IE event to be at least 100bp. These profiles do not include events that come from the first intron of a gene. Moreover, to avoid bias, for each IR event, we selected IE events with similar relative positions within the gene. 111
A.2 Average DNase I-seq coverage profile for the four samples: Arabidopsis leaf (a) and flower (b); rice leaf (c) and callus (d). In each case, a pool of genes up to 5000bp in length are used (roughly 95% of total genes). The profile encompasses the gene body and 1000bp upstream of the transcription start site represented by '0' on the x-axis. Each figure shows the profile for three sub-categories: genes with first intron retained (purple), genes with intron(s) retained anywhere else but the first one (red), and genes without any retained intron (green). 112

A.3 The complete state diagram for the continuous HMM used to predict hexamers with potential footprints. The diagram shows all 13 states. The HMM consists of three modules, to enable us to model leading/trailing footprints in addition to the primary footprint. Each module has copies of the five core states. The size of the arrow (transition) as in $BG_0 \longrightarrow DN$ and $UP \longrightarrow BG_3$ represents higher probabilities than the other transitions from the same state. These probabilities are highlighted in the supplementary table 4. This is used to emphasize the primary footprint detection by our model in all cases. The figure also summarizes the HMM states in the rectangular box to the right. 113
A.4 Positional preference is shown for AT-rich hexamers (top), GC-rich hexamers (middle) in 3' exon region of IR events, and all hexamers in intron region of IE events (bottom). All hexamers mentioned in the figure exhibit a footprint. 114
A.5 Motifs generated after clustering the IR and IE enriched hexamers exhibiting a footprint across in leaf samples in both species. Motif logos were generated using the weblogo tool. In the table, these motifs are grouped based on the type of event (IR and IE) they are enriched in and part of the event from which their respective hexamers were found (5' exon, intron, and 3' exon). 115

B.1 The distribution of different transcription factor ChIP-Seq peaks in the promoter, intragenic, and intergenic regions of the human genome. The ChIP-Seq peaks for the corresponding TFs were downloaded from the ENCODE database [14]. 122
B.2 AUC box and whiskers plot is shown for the different network architectures and gapped kmer SVM, using the leave-one-chromosome-out strategy. For each model, the green line in the box represents median AUC across the 22 chromosomes whereas average AUC value is represented by the red marker. All deep learning methods use embedded representation of the input except Basset. 122

C.1 Distribution of the attention weights for the main test and the background sets. The actual frequencies (y-axis) are normalized by total sizes of the test and background sets. This figure helps in selecting the appropriate attention cutoff, one of the parameters of SATORI. We use a default value of 0.10. 127
C.2 Similarities between motifs of GATA variants (a). Similarly, TAL1 and TCF15, both belonging to the bHLH family, have very similar motifs (CAGCTG consensus) (b). . . 128
C.3 AUC scores for DHSs in human promoters across 164 cell types, achieved by the two model variants. Each circle represents performance on detecting DHSs in that cell line. 128
C.4 Limitations of the TomTom motif comparison tool. Matches shown here are statistically significant (q-value < 0.01) for both (a) HOXA2 and (b) ZNF263. The top row depicts the gold standard motif in the CISBP database and the bottom row shows the CNN filter/motif. 129
C.5 The most frequent interacting transcription factor families in human promoter regions. 129
C.6 The most frequent interacting transcription factor families in the intron retention events. 130

D.1 Genome-wide differential AS in four lines (cultivars) of sorghum is shown in (a) for IR events and (b) for ES events. Each slice represents one of the 10 sorghum chromosomes. From the center, the first four concentric circles represent sorghum lines 1, 2, 3, and 7 respectively. The outer most circle shows the genomic coordinates with a step size of 10 million bp. The gene expression levels are shown by purple and blue coverage plots for the treated and control samples, respectively. Finally, across the coverage plots, the up-regulate and down-regulated differential IR events are marked by green and red lines, respectively. This figure is generated using CIRCOS [15]. . . . 134

Chapter 1

Introduction

Alternative Splicing (AS) is a regulated phenomenon that enables a single gene to encode structurally and functionally different biomolecules (proteins, non-coding RNAs, etc.) that play important roles in an organism's development and growth [16, 17]. Recent studies involving high throughput RNA-seq data show that AS is widespread in both plants and animals. There are different forms of AS but the notable and prevalent ones are Exon Skipping (ES), Intron Retention (IR), and alternative $3'$ (A3) and $5'$ (A5) splicing. Interestingly, these forms of AS have different frequencies of occurrence in plants and animals. For instance, exon skipping is the most prevalent form of AS in animals whereas intron retention is the most frequent one in plants [18]. This difference in frequencies might be because of the number of differences in the architecture of plant and animal genes. For instance, introns in plants are much shorter than in animals. This compositional bias is important for identification of splice sites by the splicing regulating proteins and for efficient splicing of introns [19, 20]. Note that the biological background necessary for understanding alternative splicing is discussed in a greater detail in the next chapter.

Alternative splicing is highly prevalent in both plants and animals. This was first found when the genomes of human and several other animal species were sequenced. It was observed that humans do not have a significantly higher number of genes than other organisms such as mice, fruit flies, and worms, yet have a higher behavioral and morphological complexity. Later, in several studies, the correlation between the extent of AS and organismal complexity was reported [21–24]. Besides its prevalence, AS has been shown to be important in several diseases and cancer-related studies. It is known that disease mutations can affect splicing by altering either the splice sites or the corresponding sequence motifs in exons and introns. This has been reported in thalassemia [25, 26] and spinal muscular atrophy related mutations [27, 28]. Even mutations deep within the introns—that are non-coding—have been shown to control splicing and RNA-processing. This indicates that these apparently unimportant base changes need to be accounted

for in the future studies on the role of AS in normal development and in disease [29]. Such links have been found between DNA damage and repair factors and mutations in the non-coding regions, which have a direct role in cancer pathways [30–32]. Moreover, in the Cancer Genome Project, a similar association between genetic mutations and splicing regulating factors was reported in several studies [33–39].

To understand the regulation of splicing, we require an extensive knowledge of AS regulating sequence elements that determine splice site choice. These elements essentially form the so-called *splicing code*; a comprehensive set of features that governs the regulation of AS. Until recently, sequence based elements have been the primary features listed in the splicing code [40]. However, in the past few years, numerous studies have reported that splicing occurs co-transcriptionally [41–44] and is influenced by chromatin state, in both plant and animal species [45–47]. More specifically, multiple aspects of chromatin state have been implicated in the regulation of AS, such as chromatin accessibility [48], DNA methylation [49, 50], and histone modifications [45, 47]. It follows that the splicing code needs to include sequence elements that determine chromatin state as part of a holistic model of AS. Therefore, in this work, we investigate the role of chromatin state in the regulation of alternative splicing.

As mentioned earlier, IR is the most frequent form of splicing in plants. Nevertheless, very little is known about the regulation of IR, particularly from an epigenetic perspective. To that end, we explore the role of chromatin accessibility in the regulation of IR in two plant species: rice and arabidopsis. We report a significant association between open chromatin and intron retention and present a mechanistic hypothesis of its regulation from the perspective of co-transcriptional nature of splicing. Moreover, we identify numerous conserved sequence elements for DNA-binding proteins that affect splicing in the two plant species.

Recent studies have shown that IR is also highly prevalent in humans [51]. Moreover, unlike in the case of plants, the ENCODE project [14] provides a wealth of human genomic and epigenomic datasets that are well-suited to be used with complex computational models. Therefore, in the next step, we design a deep learning model that predicts chromatin accessibility in IR events across the

entire human genome. The basis for this study lies in the fact that the proteins that regulate AS should bind in the vicinity of the splicing events, in the regions of open and accessible chromatin. Using the model, we identify motifs enriched in IR events with significant hits to known human Transcription Factors (TFs). The zinc finger family exhibits the highest activity in IR events, a prediction that is validated using ChIP-Seq data. Experiments by our collaborators have validated our predictions in candidate IR events.

We note that to render an accurate regulatory landscape of alternative splicing, the splicing code should encompass cooperativity and interactions between the sequence features. Therefore, we develop a a self-attention based model that captures the regulatory interactions between motifs identified within the genomic sequences. We first evaluate our method on simulated data and three complex datasets. In each of the three experiments, our model identifies numerous statistically significant TF interactions, many of which have been previously reported. Finally, using this model with the chromatin accessibility in IR dataset, we identify multiple interactions primarily involving the zinc finger family of transcription factors. Our approach not only provides a global, biologically relevant set of interactions but, unlike existing methods, it does not require a computationally expensive postprocessing step.

1.1 Overview of chapters

The next set of chapters describe the relevant background and related work. In Chapter 2, we provide a detailed overview of the relevant biological and bioinformatics background. Chapter 3 provides a primer on different concepts in deep learning that have been used in this work. Finally, in Chapter 4, we review the related work in genomics, relevant to the topic of this dissertation.

Chapter 5 investigates the association between chromatin accessibility and IR in two plant species; arabidopsis and rice [52]. We analyze the chromatin accessibility in IR and non-IR events using DNase I-Seq and Bisulfite-Seq data. Moreover, a Hidden Markov Model (HMM) is used with the DNase I-Seq data to identify binding sites for the IR-regulating proteins.

Chapter 6 uses a deep learning model to predict chromatin accessibility in IR events in human. Because of the limited chromatin profiling data in plants, for our deep learning model, we use data from ENCODE Consortium [14] which provide a plethora of human and mouse datasets. In this study, we also report the IR regulating motifs and known transcription factor binding sites that are identified by our model.

In Chapter 7, we present a self-attention based model to infer regulatory interaction between DNA-binding proteins. We test our model on both simulated and real-world datasets (including the IR dataset) and report numerous statistically significant TF-TF interactions.

Finally, in Chapter 8, we provide a summary of our contributions and discuss the potential future work in this area, with evidence from our preliminary experimentation and results.

Chapter 2

Biological and Bioinformatics Background

2.1 Biological background

The central dogma of molecular biology was presented by Francis Crick back in 1958 [53]. In its simplest form, it states that the DNA in our cells is transcribed into RNA which in turn is translated into proteins. Our focus is on eukaryotic cells which have a distinct membrane-bound nucleus. Almost all of the DNA in eukaryotic cells is found in their nuclei, organized into structures called chromosomes. The DNA is in highly compact form held together by protein complexes known as histones. This packaged DNA coiled around the histone complexes is measured in the units called nucleosomes. This whole arrangement of DNA is called chromatin (see figure 2.1(b)), and its organization plays a central role in the regulation of vital biological processes such as gene expression and alternative splicing, as described next.

Genes in eukaryotes are sequences of DNA that encode functional biomolecules: proteins, non-coding RNAs, etc. The coding parts of a gene that eventually form the functional biomolecule are called exons, while the non-coding segments in between the exons, that are spliced out, are called introns. Gene expression starts with the process of transcription, where the DNA is transcribed into pre-mRNA by an enzyme known as RNA Polymerase II. In the next step, it goes through a process known as splicing where exons, the sub-parts of the coding sequence, are joined together and the introns are spliced out by complex molecular machinery called the spliceosome. The mature mRNA with $5'$ cap and $3'$ poly-A tail is then transported to the cytoplasm where it is translated into protein in the ribosome. As simple as it seems, this whole process has several complex aspects. For instance, genes do not encode proteins in a one-to-one correspondence: the exons can be combined in different ways, the $3'$ and $5'$ ends of the exons can be alternatively selected, and finally the introns can be retained and find their way out of the cell, as part of the coding sequence. This phenomenon is called alternative splicing and is highly prevalent in both plants and animals. For instance,

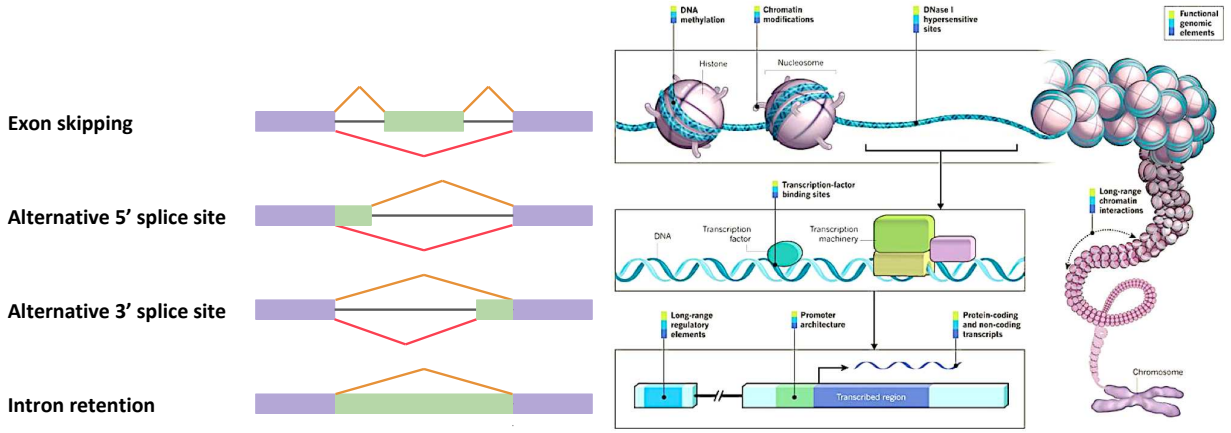
nearly 95% of human genes have been shown to undergo AS [23, 24]. It is worth mentioning that transcription is not a two-step process; splicing has been shown to occur co-transcriptionally i.e. while the polymerase enzyme is transcribing DNA into RNA, the spliceosomal machinery can act upon the nascent RNA simultaneously, stripping off the introns.

Figure 2.1(a) depicts the different forms of AS. The most common forms are Exon Skipping (ES), Intron Retention (IR), Alternative 5' (donor) site splicing (A5), and Alternative 3' (acceptor) site splicing (A3). Table 2.1 summarizes the distribution of these types in different animal and plant species. In general, exon skipping is the most prevalent form of splicing in animal species. On the other hand, in plants, intron retention is the most common type of AS.

Table 2.1: AS landscape for 5 different eukaryote species where the number of splicing events are provided for each type of AS. In general, exon skipping is prevalent in mammalian species whereas intron retention is the most common form of splicing in plants. The data was generated using Splicgraher [1] with the following gene annotations: UCSC hg19 (human), UCSC mm9 (mouse), TAIR10 (arabidopsis), MSU v7 (rice), and JGI Phytozome V12 (sorghum).

Kindom	**Species**	**AS Type**			
		ES	***IR***	***A3***	***A5***
Animal	Homo sapiens (human)	19985	1583	5763	5798
	Mus musculus (mouse)	4470	434	1775	1637
Plant	Arabidopsis thaliana (thale cress)	1138	3334	2540	2378
	Oryza sativa (rice)	1207	4321	2959	1937
	Sorghum bicolor (sorghum)	2531	3670	3446	2380

As mentioned earlier, the spliceosome primarily executes the process of splicing in eukaryotic cells. This protein complex is formed by five nuclear ribonucleoprotein particles (snRNPs) along with numerous auxiliary proteins. These sub-components of the spliceosomal machinery recognize core splicing signals through a series of biochemical reactions. The splicing signals include the splice sites: donor ($5'$ splice sites) and acceptor ($3'$ splice sites), the polypyrimidine tract, and the branch point sequence. However, these features alone are insufficient for AS: Splicing Regulatory Elements (SREs), which are other sequence elements in pre-mRNA, act as the binding sites for splicing regulating proteins. These SREs, exonic and intronic splicing enhancers/silencers, play an important role in the regulation of alternative and constitutive splicing [16, 54, 55]. Usually



(a) Types of AS (b) Chromatin organization

Figure 2.1: Different types of AS are shown in (a) where introns are represented by black lines connecting the exons. The structure and organization of chromatin within a cell nucleus is shown in (b) along with the accessible regions (DNase I hypersensitive sites) and some of the key factors that regulate gene expression. Adapted with permission from [4]. Copyright 2012, Springer Nature.

$6 - 10$ nucleotides long [56], these SREs work by accommodating splicing factors that activate or suppress splice site recognition or spliceosome assembly [54, 57]. These sequence elements, generally spaced in regular intervals or in clusters, impact the splice site selection through specific binding of splicing regulatory proteins [58].

DNA/RNA sequence characteristics have also been implicated as splicing determinants. Some of these features are the length of exons and introns [59], GC-content, and divisibility by 3 etc. In addition, the secondary structure of RNA in the vicinity of AS events has been shown to affect splicing outcomes. For instance, secondary structures have been computationally identified that aid in the prediction of splice sites [60, 61]. Moreover, genome-wide analyses of conserved RNA secondary structure overlapping splice sites have been shown to affect AS [62]. These secondary structures can shorten the distance between splice sites and aid in their recognition [63, 64].

Another aspect of the regulation of AS is tissue- and/or condition-specific binding of splicing factors [65, 66] that affect gene expression in the respective stages and/or cell types. A number of such elements have been identified, however, their binding sites are not well characterized. For instance, many of these factors/binding sites involve loosely defined motifs such as “CA-rich” for

hnRNP-L and "CU-rich" for $PTB1$ and $PTB2$ [67, 68]. These motifs alone are not sufficient to accurately predict tissue- and condition-specific alternating splicing [57].

Finally, we discuss chromatin, a cellular structure that has been linked in several ways with splicing [46, 69–71]. As shown in figure 2.1(b), it is a compact structure of hereditary material (DNA and the histone complexes) within the nucleus. Moreover, the figure provides details on how regulatory proteins—for instance, transcription factors—can bind the DNA in regions of open chromatin and take part in the regulation of gene expression. In this work, our goal is to investigate the association of AS with chromatin structure, accessibility, and modifications. It has been reported that the speed of transcription (RNA polymerase II elongation) is affected by histone modifications and correlates with splicing patterns [72, 73]. The evidence for this phenomenon, called co-transcriptional splicing, has been previously reported [41–44].

2.2 Bioinformatics background

In this section, we describe the expression and chromatin profiling methodologies needed to quantify the genomic and epigenomic features involved in the regulation of AS.

2.2.1 Transcriptome analysis

The transcriptomic data relevant to the problem at hand targets the expression of the genes and their corresponding coding/non-coding parts. Over time, expression quantification has evolved; from Expressed Sequence Tags [74], to Serial Analysis of Gene Expression [75], DNA microarrays [76], and finally to high-throughput whole-transcriptome shotgun sequencing called RNA-Seq [77], which is the most popular technique used in expression studies these days. Next, we describe the major steps in transcriptome analysis: from RNA isolation to sequencing, and finally quantifying gene expression and alternative splicing.

Library preparation and sequencing

The first step in the quantification of expression is to isolate RNA from a group of cells or a tissue. The two most commonly used techniques for that are ribosomal RNA (rRNA) depletion [78]

and polyadenylated (poly-A) RNA enrichment [78]. The next steps in the standard RNA-Seq library preparation process are fragmentation of RNA, size selection, and complementary DNA (cDNA) synthesis, followed by amplification using Polymerase Chain Reaction (PCR). Note that here we focus on Illumina based short-read sequencing of RNA, where the sample is collected from a group of cells; this is in contrast to single-cell sequencing methods. Once prepared, the library for each sample in an experiment is sequenced using the massively parallel, high-throughput short-read sequencing platforms.

Alignment and expression quantification

The sequenced library consist of raw reads, usually 50 base pair (bp) to 150 bp long. In case of eukaryote species, the raw reads must be aligned to the reference genome using a splice-aware aligner. Most popular programs used for aligning RNA-Seq reads to the reference genome are *Tophat* [79] and *STAR* [80]. Once the reads have been aligned, we can quantify gene expression using reference genome annotations of the corresponding species. The expression is essentially a function of the number of reads aligned to the region of the genome where a specific gene resides. In practice, the aligned reads are normalized by the library size. Gene expression is commonly represented using Reads Per Kilobase of transcript per Million mapped reads (RPKM), which is given by:

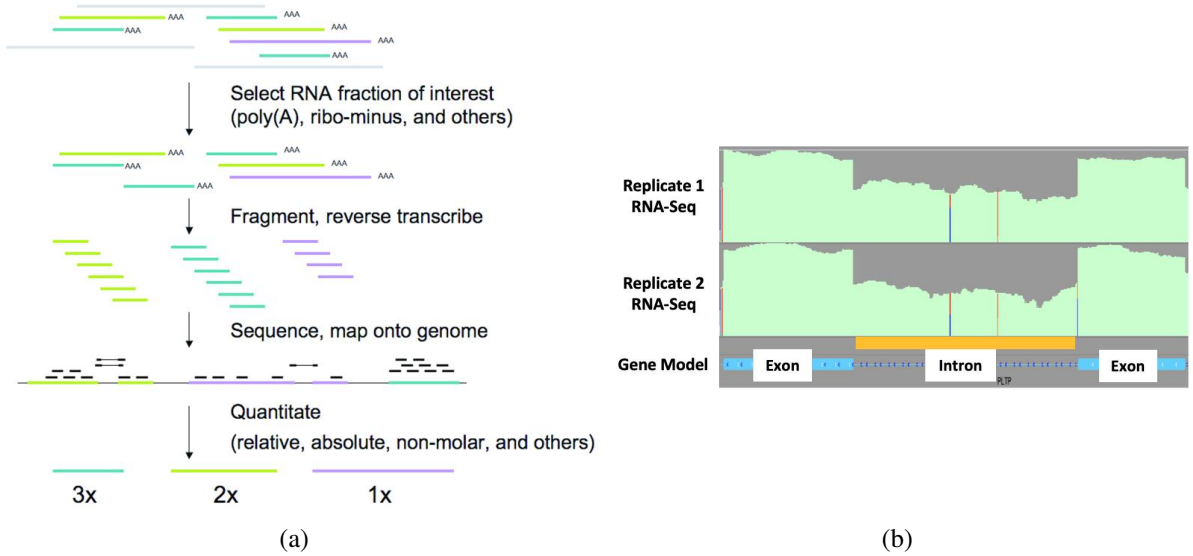
$$\mathrm{RPKM} = \frac{n}{\frac{L}{1000} \times \frac{N}{1{,}000{,}000}},$$

where n is the number of reads, L is the lenght of the gene, and N is the size of the library. The workflow of RNA-Seq experiment is depicted in figure 2.2(a).

Quantification of alternative splicing

AS events can be quantified using the RNA-seq data by counting the reads aligned to exons, introns, and across their junctions. However, in this case, the definition of exon and intron are based on the reference genome annotations. Some methods are geared toward identifying novel splice junctions by using machine learning methods. For example, Splicegrapher [1] can quantify both known and novel AS events while taking advantage of multiple expression profiling data

sources. In this work, our focus is primarily on intron retention, and to quantify IR events, we require that every position of the retained intron is covered by at least one aligned read. As an example, figure 2.2(b) depicts an intron retention event in K562 human cell-line where in both replicates, every position of the intron is covered by aligned reads.



(a) (b)

Figure 2.2: The steps of RNA-Seq workflow are shown in (a): RNA extraction, fragmentation, reverse transcription, sequencing, mapping to the genome, and finally quantification of expression (Adapted with permission from [5]. Copyright 2009, Springer Nature). An IR event is shown in (b) with evidence from RNA-Seq data across two biological replicates for human *K562* leukemia cell-line. The coverage plots in green are calculated using the number of reads aligned to that region of the genome. The high RNA-Seq coverage across the intron indicates its retention in both K562 replicates. This figure is generated using the Integrated Genome Viewer [6].

2.2.2 Chromatin profiling

Most of the work towards understanding alternative splicing has primarily used expression data, such as microarrays, and RNA-Seq. However, in recent years, evidence from chromatin profiling data has been included in several AS-related studies [51, 52, 81]. There are various sequencing techniques that probe different aspects of chromatin: its organization, accessibility, and modifications. Chromatin Immuno-Precipitation sequencing (ChIP-Seq) [82] reveals binding sites for specific transcription factors (TFs) as well as several histone modifications, such as methyla-

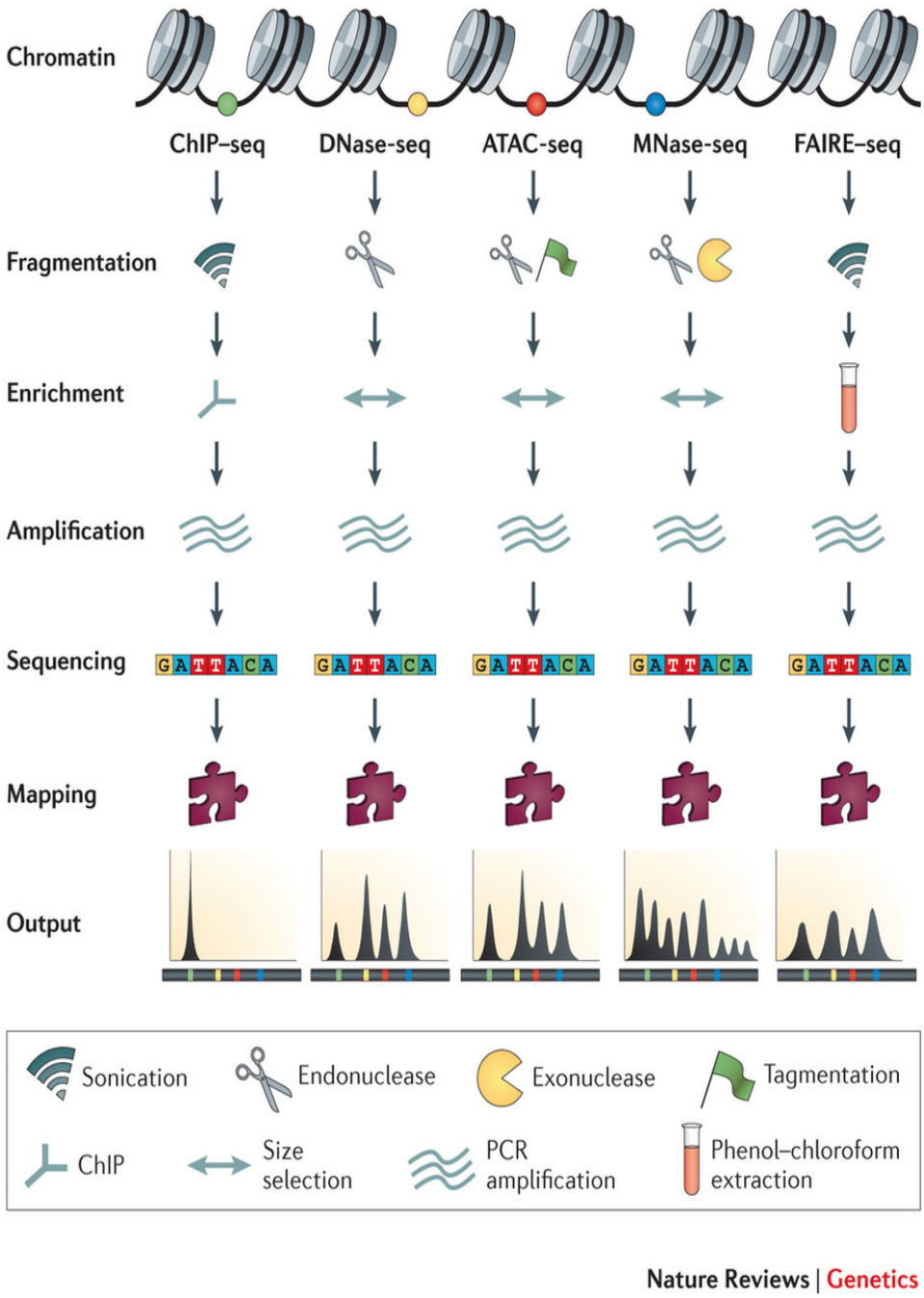


Figure 2.3: High throughput massively parallel sequencing experiments for profiling chromatin accessibility and modifications. Adapted with permission from [7]. Copyright 2014, Springer Nature.

tion (H3K4me3, H3K27me3, etc.) and acetylation (H3K9Ac, H4K16Ac, etc.). These modifications have been shown to affect the arrangement of chromatin by loosening or tightening the DNA strands around the histone proteins. Similarly, Micrococcal nuclease sequencing (MNase-Seq) [83] is used to quantify nucleosome occupancy and positioning.

To quantify chromatin accessibility, Deoxyribonuclease I based sequencing (DNase I-Seq) [84], Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIRE-Seq) [85], and Assay for Transposase-Accessible Chromatin sequencing (ATAC-Seq) [86] are used. For instance, in DNase I-Seq, an enzyme from the endonuclease family of proteins is used that cleaves DNA in the regions of open chromatin. These regions are deemed important in analyzing the activity of the regulatory proteins associated with gene expression and alternative splicing. Chromatin accessibility is discussed in Chapter 5 in greater detail. Finally, DNA methylation patterns are quantified using Bisulfite sequencing (BS-Seq) [87]. DNA methylation has been shown to affect chromatin accessibility [88] and more recently, alternative splicing [52].

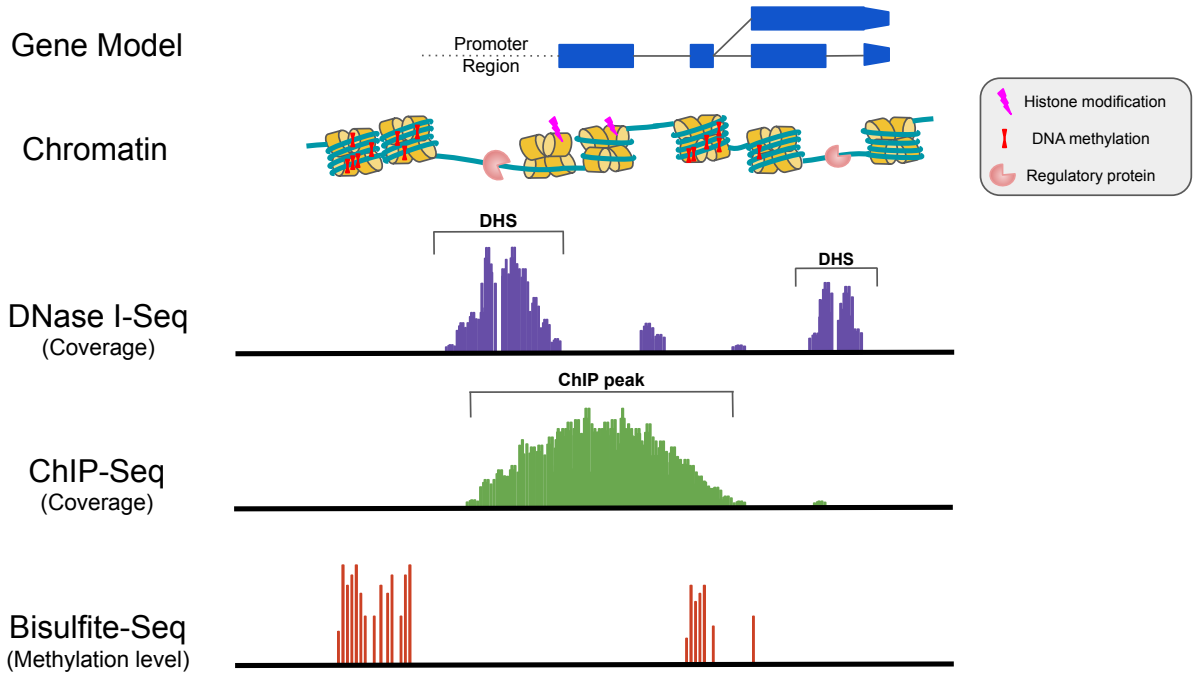


Figure 2.4: An example of chromatin profiling for a hypothetical gene. The gene model annotations are shown in the top row where the boxes represent exons/coding sequences. The corresponding chromatin state is depicted in the second row. Finally, the quantification of chromatin accessibility, histone modifications, and DNA methylation is shown in third, fourth, and fifth row, respectively.

Figure 2.3 summarizes several of the methods used for chromatin profiling. Most of the steps—library preparation and sequencing—are very similar to what we described in the previous section. However, to align the raw reads to the reference genome, an ungapped aligner, such

as bowtie [89] or BWA [90], is used. Note that this work is primarily focused on the association between AS and chromatin accessibility and modifications using DNase I-Seq, ATAC-Seq, and ChIP-Seq data. To identify the regions of open/modified chromatin—for instance, DNase I Hypersensitive Sites (DHSs), Transposase Hypersensitive Sites (THSs) in case of DNase I-Seq, ATAC-Seq respectively—a peak calling software is used after aligning the reads to the reference genome. For example, Hotspot [91] can be used to quantify peaks (DHSs) across the genome. To quantify DNA methylation in bisulfite sequencing data, bismark [92] can be used. Figure 2.4 depicts chromatin profile for a hypothetical gene. The accessible and modified chromatin is represented by the corresponding DHSs, ChIP peak, and DNA methylation levels.

Chapter 3

Deep Learning Background

In this work, we use different machine learning models, which take advantage of chromatin profiling data, to identify sequence elements and motifs associated with the regulation of AS. In our initial work on the association between intron retention and chromatin accessibility, we design a continuous Hidden Markov model to call footprints in the DNase I-Seq data across IR events in plants. These footprints represent potential binding sites for splicing regulating proteins. In Chapter 6, we use a deep learning model to predict chromatin accessibility in IR events. Several features and architectural elements are explored such as low-dimensional vector embeddings, convolutional, recurrent, and self-attention layers. One advantage of using a deep learning model with a CNN is that we don't need explicit feature engineering: weights of the first convolutional layer can be interpreted as DNA sequence motifs for the regulatory proteins and known transcription factors associated with intron retention in human. Finally in Chapter 7, we use a self-attention based model to infer regulatory protein interactions. The attention matrix of a self-attention layer can be used to quantify the influence of part of an input sequence on all other regions within that sequence. This leads us to identify interacting binding-sites for regulatory proteins within a genomic sequence.

Next, we describe the essential machine learning concepts and approaches that serve as building blocks for the aforementioned models. We mainly focus on the features and layers of a deep neural network: a fully connected network, convolution operation, recurrent layers, and self-attention mechanism etc.

3.1 Fully connected networks

A fully connected network is an important part of a deep learning model where a unit in one layer is connected to all other units in the subsequent layer (see Figure 3.1). In principal, such a network is basically a multi-layer perceptron [93]. A single fully connected layer—or a network of

such layers—serves as read-out layer or mechanism in a complex deep learning model: the outputs of convolutional and recurrent layers are passed through a network of fully connected layers which is eventually translated into a classification of two or more target labels. Besides, a cascade of such layers adds to the depth of the network and help the model capture complex underlying associations within the features.

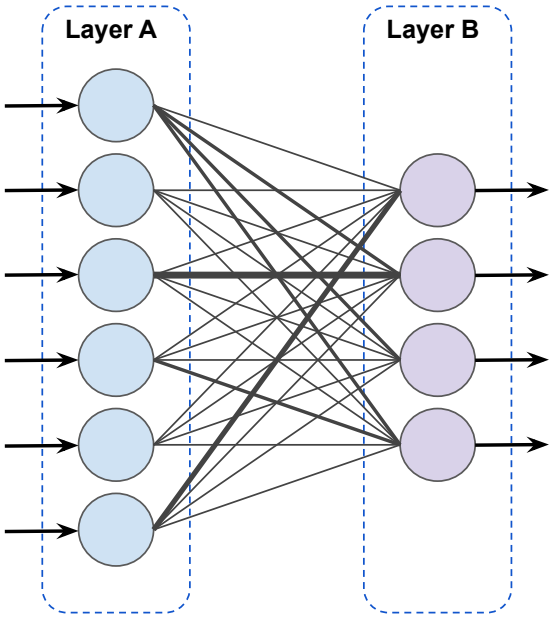


Figure 3.1: A simple fully connected network, part of a deep learning model [8]. Every unit in layer A is connected to all other units in layer B. Note that a fully connected network is usually followed by a read-out layer (not shown here).

3.2 Convolutional neural networks

Convolutional Neural Networks (CNNs) have been successfully used with a significant gain in performance over other machine learning methods in image classification [94], natural language processing [95], and computational biology [96]. A remarkable advantage of these models is their ability to capture complex underlying patterns in a given set of features. In problems involving the analysis of biological sequences, CNNs have exhibited remarkable success across multiple areas: gene expression analysis [97], TF binding prediction [98–101], chromatin acces-

sibility analysis [13, 102, 103], chromatin structure and its modifications [104], and identification of RNA-binding protein sites [105]. Besides providing improvement in accuracy over traditional machine learning models, CNNs can be used without explicit feature engineering, and can learn directly from sequence data.

The core process in a convolutional layer is the convolution operation. Since we are analyzing biological sequences in this work, for the rest of this section, we will focus on one-dimensional (1D) convolution, which can be written as:

$$X'_{i,j} = \sum_{a=0}^{A-1} \sum_{b=0}^{B-1} \omega^{j}_{a,b} X_{i+a,b}, \tag{3.1}$$

where X is the input matrix, i is the position at which convolution is performed, j is the index of the filter, and ω^j is the weight matrix of the filter with size $A \times B$ where A is the length of the filter (window size) and B is the number of input channels which is four in the case of one-hot encoded input sequence, because there are four letters in the DNA alphabet. Here the filters are equivalent to Positional Weight Matrices (PWMs) or sequence motifs. After the convolution operation, typically a non-linear function such as the Rectified Linear Unit (ReLU) is used, which is given by:

$$f(x) = \max(0, x). \tag{3.2}$$

ReLU is a standard activation function in deep learning which reduces the problem of vanishing gradients. Moreover, we reduce the output size by max-pooling by taking the maximum value in a window of a pre-determined size. This reduces the input size for the next layer and also achieves invariance to small shifts in the input sequence. The process of convolution, followed by a non-linear ReLU and max-pooling operations is depicted in figure 3.2.

3.3 Recurrent neural networks

Recurrent Neural Networks (RNNs) were first presented back in 1980s [106]. Unlike a regular neural network, an RNN has a feedback loop that allows it to accept a sequence of inputs:

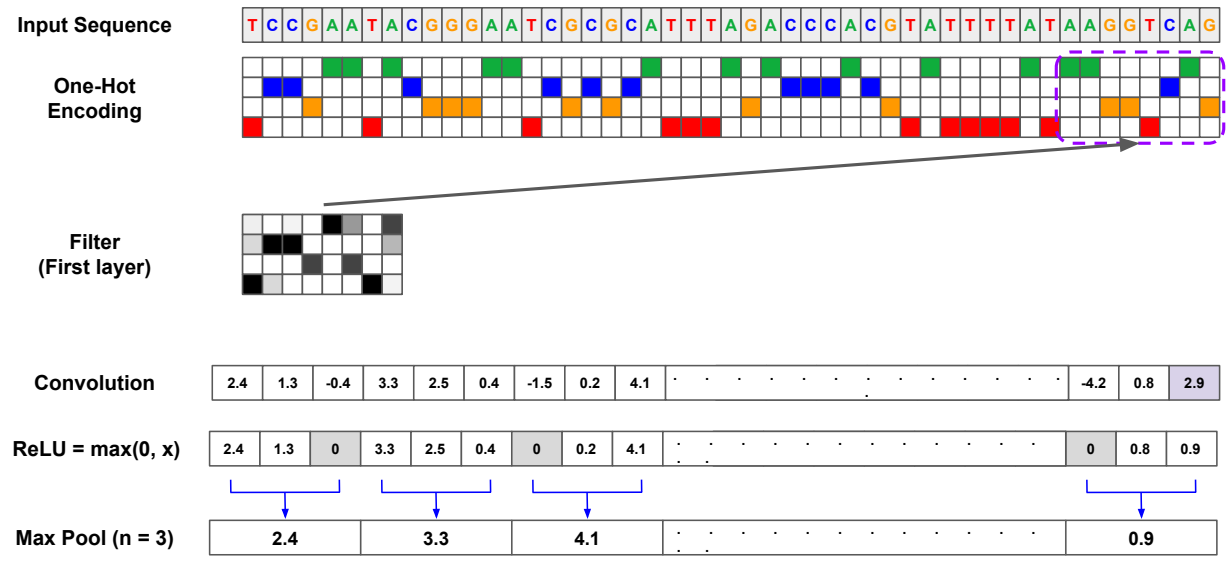
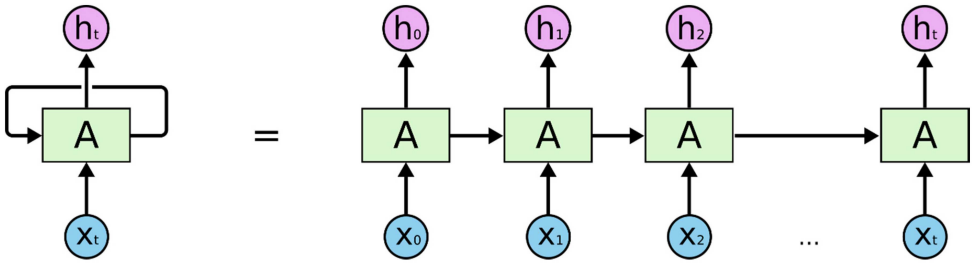


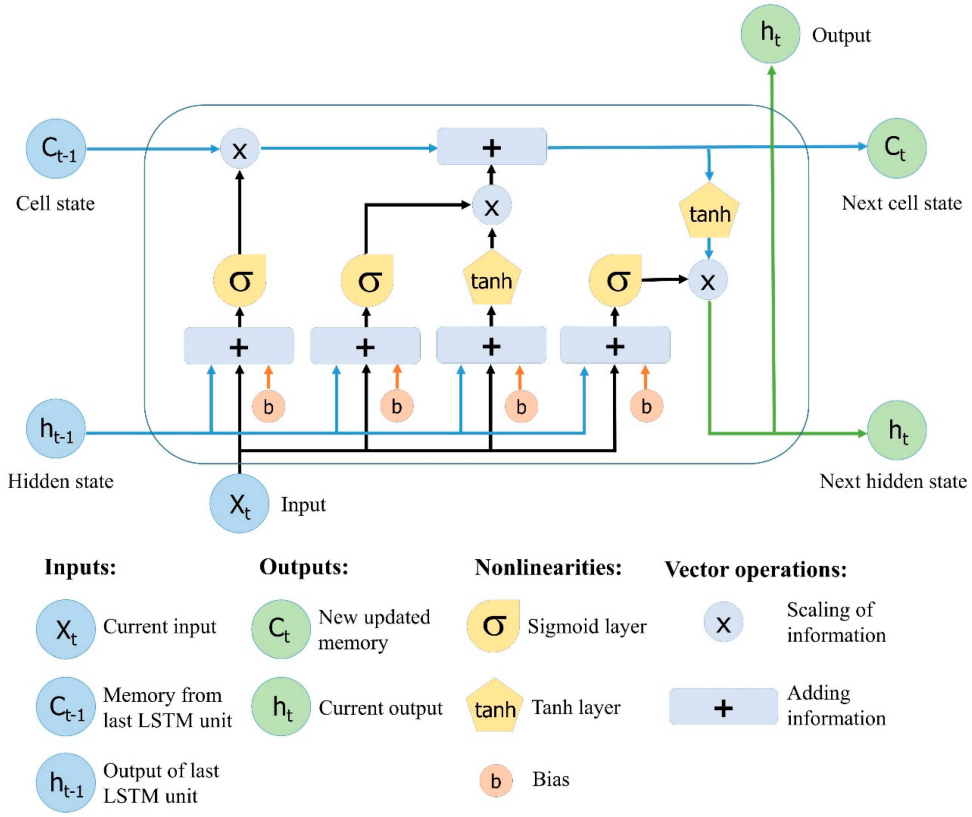
Figure 3.2: Summary of convolution operation followed by a non-linear ReLU function and max-pooling. In convolution, a filter of fixed size is scanned across the entire input matrix. Next, the ReLU operation gets rid of negative values. Finally, the output is reduced via max-pooling which takes a maximum value across a window of fixed size.

it follows that the output at step t is influenced by the output/state of the network at the previous step, $t-1$. This enables the network to maintain a memory of the previously *seen* inputs and capture long-term dependencies. Because of these properties, RNNs have pervasively been used in the area of natural language processing [95, 107] and other sequence prediction problems. Alongside convolutional neural networks, RNNs have also been used in the field of computational biology [100, 101]. Figure 3.3(a) shows the sequential processing of an RNN.

One disadvantage of an RNN is that while modelling long-term dependencies, it often suffers from the problem of vanishing and exploding gradients [108]. This make RNNs difficult to work with in a complex sequence prediction problem. To address this, Long Short-Term Memory (LSTM) [109] units can be used that controls the flow of input through a mechanism of multiple gates: input, forget, and output. A detailed diagram of an LSTM cell and its working are depicted in figure 3.3(b). The first step in an LSTM is to identify information that is not required and will be dropped from the cell in the corresponding step. This is achieved in the forget gate (f_t) of the LSTM which can be mathematically written as:



(a)



(b)

Figure 3.3: A simple recurrent neural network with sequential processing is shown in (a). In (b), the detailed structure of a long short-term memory unit is depicted. These figures were taken from [9], originally adapted from [10] and [11].

$$f_t = \sigma\left(W_f\left[h_{t-1}, X_t\right] + b_f\right), \tag{3.3}$$

where σ is the sigmoid function, X_t is the input to the current cell, h_{t-1} represents the output of the previous LSTM cell at step $t-1$, and W_f and b_f are the weight matrices and biases respectively. f_t is a vector with values ranging from 0 to 1, corresponding the each number in the cell state, C_{t-1}.

The next step deals with storing information from the current input X_t in the cell state as well as updating it. This is further divided into two parts: the sigmoid layer and the tanh layer. First, the sigmoid function decides whether to update or ignore the new information. Second, the tanh function weighs the values which pass through, deciding their level of importance. The values from these two steps are then multiplied to update the new cell state. Finally, this new memory is added to the previous cell state, C_{t-1}, to generate the next cell state, C_t. This whole process can be mathematically summarized using the following equations:

$$i_t = \sigma\left(W_i\left[h_{t-1}, X_t\right] + b_i\right), \tag{3.4}$$

$$N_t = tanh\left(W_n\left[h_{t-1}, X_t\right] + b_n\right), \tag{3.5}$$

$$C_t = C_{t-1} f_t + N_t i_t, \tag{3.6}$$

where W and b are the weights and biases, respectively, of the cell state.

In the final step, the output value, h_t, of the current cell is generated which is based on the output gate, O_t, but filtered using the current cell state, C_t. As shown in figure 3.3(b), the last sigmoid unit determines which parts of the cell state make it to the output. Essentially, the cell state is put through a tanh function to push the value to be between -1 and 1 and then multiplied by the output of the aforementioned sigmoid unit. This process can be mathematically expressed using the following equations:

$$O_t = \sigma\left(W_o\left[h_{t-1}, X_t\right] + b_o\right), \tag{3.7}$$

$$h_t = O_t tanh\left(C_t\right). \tag{3.8}$$

3.4 Self-attention

Recently, neural networks that use the concepts of *attention* and *self-attention* [110, 111] have achieved remarkable success in natural language processing tasks, specifically in machine translation [12]. One of the strengths of attention is that it can capture associations between features regardless of the distance between them, addressing a major shortcoming of convolutional and recurrent networks. This is particularly useful for tasks in computational biology where our goal is to identify regulatory elements and their associations/interactions in DNA or RNA sequences.

When it comes to machine translation, self-attention can successfully capture the long distance dependencies within an input sentence. For example, in figure 3.4(b), in a machine translation problem, the influence of all other words is shown for the verb "making" in a given sentence [12]. The top row shows the words for which attention is quantified with respect to all other words in the bottom row. Particularly, the word "making" is strongly influenced by the two words "more" and "difficult". The strength of the signal is indicated by the saturation of the colored boxes, each representing one of the multi-heads in the attention layer—the concept of multi-head attention is explained in a greater detail, later in this section. It follows that in this example, self-attention enables the model to capture the complete phrase: "making...more difficult".

In the field of computational biology, the value of attention for modeling transcription factor binding site prediction was recently demonstrated, and their work was motivated by the greater interpretability of the resulting networks [112]. As mentioned earlier, the attention mechanism can model dependencies within the input sequence regardless of their distance [12]. By doing so, it guides the network to focus on relevant features within the input and ignore irrelevant information. Pertinent to the problem at hand, a self-attention layer can help us identify interacting regions within the input sequences, for instance, binding site motifs of regulatory proteins. Consequently, we can capture interactions between regulatory events.

The mechanism of self-attention is summarized in figure 3.4(a). To formally define self-attention, consider the input X and two linear transformations of it, called the Query Q, and Key K which are defined by:

$$Q = W_{Q}^{\top} X, \tag{3.9}$$

$$K = W_{K}^{\top} X, \tag{3.10}$$

where W_{Q} and W_{K} are the corresponding weight matrices for the Query and Key, respectively. The attention matrix A is then computed using the following expression:

$$A(Q, K) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right), \tag{3.11}$$

where the softmax function is defined as:

$$\text{softmax(x)}_i = \frac{e^{x_i}}{\sum_j e^{x_j}}. \tag{3.12}$$

In Equation (3.11), d_k is the dimension of the Key K. The scaling by $\frac{1}{\sqrt{d_k}}$ ensures more stable gradients of the softmax function for large values of d_k [12]. The attention matrix A, defined in Equation (3.11), is a $d \times d$ matrix, which, for every position in a sequence of length d, summarizes the influence of all other positions on that position. This is crucial for capturing interactions among regulatory features, which is explained in greater detail in Chapter 7. To generate the output of the attention layer, we first define the Value matrix

$$V = W_{V}^{\top} X \tag{3.13}$$

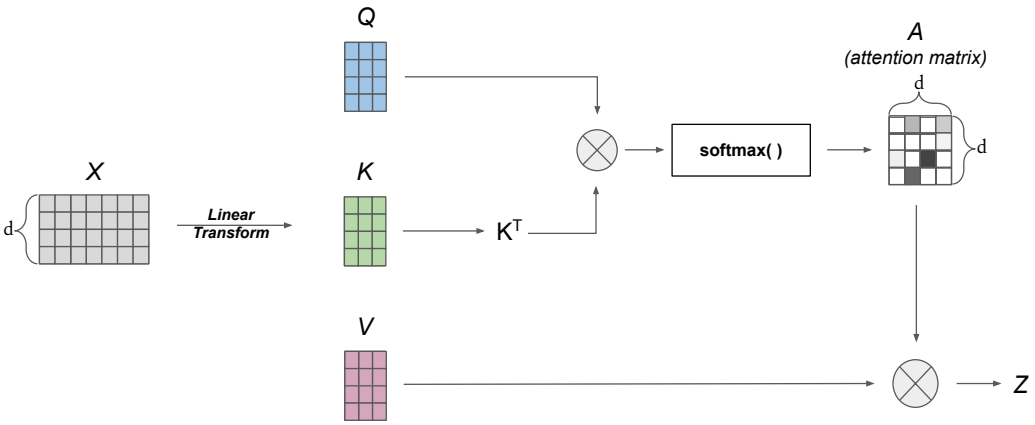
using the associated weight matrix W_{V}. Finally, we define the output of the attention layer as:

$$Z = AV. \tag{3.14}$$

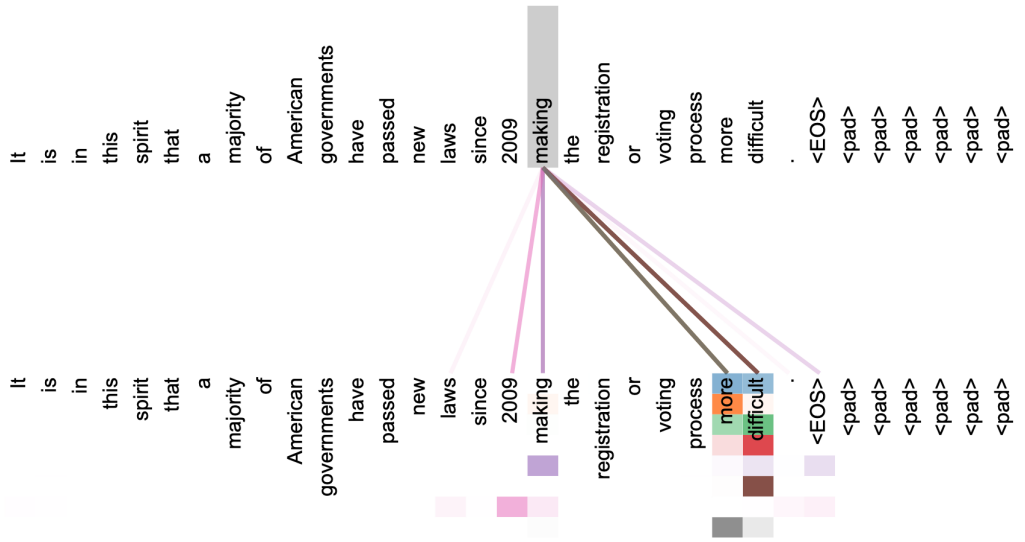
Intuitively, Equation (3.14) allows us to generate the output using the parts of the input that we want to focus on—those which exhibit strong inter-dependencies—and ignore irrelevant information.

The equations above as well as figure 3.4(a) summarize a single head attention layer. In practice, we use multiple instances of attention heads in a self-attention layer; this redundancy enables

the network to explore multiple sub-spaces while quantifying the attention profiles for every position in the sequence. To generate a final attention profile, we concatenate the outputs of the N single-heads followed by a linear transformation.



(a)



(b)

Figure 3.4: (a) The process of self-attention is summarized. The attention matrix A helps amplify the input signals that are relevant for the task at hand. (b) An example output, taken from a model that uses self-attention in a machine translation problem. In regards to self-attention, the top row shows the queries whereas the bottom row represents the keys. The results are highlighted for a specific query: the verb "making". Finally, the attention values are represented by the color boxes in the bottom row, each corresponding to one of the 8 single heads. This figure has been taken from [12], with permission from the author.

Chapter 4

Related Work: Machine Learning in Genomics

In this chapter, we review previously published machine learning methods in the domain of computational biology; relevant to the topic of this work, specifically those where genomic and epigenomic data are used. Note that many of the methods—particularly the deep learning based approaches—do not involve the prediction of alternative splicing. Nevertheless, their employment in related problems, such as predicting gene expression, TF binding sites, and chromatin state provides the foundation for a majority of this work.

Next, we categorize the related work into classical machine learning and deep learning based methods and discuss them in a greater detail.

4.1 Classical machine learning

In genomics and genetics, classical machine learning algorithms have been applied in a broad range of problems. These algorithms are particularly useful when inferring and annotating biologically relevant signals associated with various regulatory phenomena. For instance, in context of gene expression, these methods have been used to identify Transcription Start Sites (TSSs) in genomic sequences [113]. Similarly, numerous such methods have been employed to identify promoters [114], enhancers [115], splice sites [116], and nucleosome positioning [117]. Although the availability of complex high-throughput sequencing datasets has made deep learning a primary choice for a number of prediction tasks in genomics and genetics, nevertheless, traditional machine learning methods are still actively used by researchers in this area, particularly when dealing with smaller datasets. Next, we briefly review some of the classical machine learning methods relevant to the topic of this work.

4.1.1 Kernel based methods

In the field of computational biology, kernel based methods, particularly along with Support Vector Machines (SVMs) [118], gained remarkable popularity in the decade of 2000s. A kernel function is essentially an extension of the dot-product similarity between features, leading to more complex non-linear classification boundaries [119]. Besides their superior performance in various tasks, SVMs presented two major advantages in this area: first, these methods could handle noise and high-dimensional data, a typical characteristic of bioinformatics data. Second, SVMs were able to work with objects such as sequences and structures that don't have an obvious vector space representation, protein structures and gene networks, and with aided functionality using kernels, easily combine heterogeneous data [120]. Pertinent to the topic of this research, we will focus on the usage of sequence kernels with SVMs.

In genomic data, we are often interested in sequence motifs which have a biological significance; for instance, these motifs can be representative of the regulatory protein binding-sites. To extract such features from genomic sequences, the spectrum kernel was developed, and first used to classify protein sequences [121, 122]. Essentially, a spectrum kernel can quantify k-mer content within a given set of sequences where a k-mer is a substring of length k. The spectrum kernel has been widely used with SVMs in tasks involving genomic sequences. For instance, in a splice site recognition problem, the spectrum kernel based features significantly improved classifier's performance in comparison to when general sequence composition based features were used, such as GC-content [119]. An extension of the spectrum kernel, known as the Weighted Degree (WD) kernel, takes into account the positional information of individual k-mers within the genomic sequences. In aforementioned splice-site prediction task, where the positional information is important to capture the relevant splicing signals, using the WD kernel led to improved classifier's accuracy in contrast to the vanilla spectrum kernel [119]. It is worth mentioning that numerous studies have used kernel based SVMs to predict different forms of alternative splicing in genomic sequences [123–125].

Recently, to address the limitations of the spectrum kernel, a gapped *k*-mer based kernel for genomic sequences was introduced for SVMs [126]. Biological sequence motifs, particularly for transcription factors, are longer and not all positions within the motif have high information content. A spectrum kernel with longer *k*-mer size will lead to extremely sparse feature vectors and therefore, model overfitting. On the other hand, gapped *k*-mer SVM (gk-SVM) tend to model those properties of binding site motifs by specifying informative and non-informative regions within the *k*-mers. On several human ChIP-Seq datasets, gk-SVM has been shown to outperform the conventional string kernel based methods with a measurable improvement in accuracy [126]. Gapped *k*-mer SVM has been widely used as a standard baseline method in several genomics related studies that involve deep learning models—which will be discussed later in greater detail.

Besides kernel based SVMs, several other machine learning techniques have pervasively been used across different areas in computational biology. To stay within the scope of this work, in the following sections, we review some of these techniques used in the context of alternative splicing, utilizing genetic and/or epigenetic features.

4.1.2 Logistic regression

Due to the ease of implementation and high model interpretability, logistic regression has been employed in several studies involving AS prediction. Braunschweig et al. [51] reported the role of intron retention in genome-wide regulation of mRNA levels in general and turnover of non-physiologically relevant transcripts. Part of their research focused on generating an "IR code", a comprehensive set of sequence and chromatin state features that regulate intron retention. In a related study, Liu et al. [127] reported a close correlation between a few types of histone modifications and alternative splicing—exon skipping in that case. They modeled the effect of histone modifications on cassette exon inclusion using logistic regression.

4.1.3 Random forests

Another commonly used method in predicting alternative splicing is the random forest model: an ensemble classifier consisting of multiple independent decision trees [128]. Chen et al. [129]

employed a random forest model to classify alternatively spliced exons by exploiting differential evolutionary conservation between exons and introns. Similarly, in a recent study, Mao et al. [125] used random forest to differentiate retained introns from the excised ones. Compared to an SVM, their model exhibited higher accuracy in terms of Area Under the ROC curve.

4.2 Deep learning

In recent years, deep learning has made major breakthroughs in the field of machine learning. Particularly, CNN based models have been successfully used with a significant gain in performance over other machine learning methods in image classification [94], natural language processing [95], and computational biology [96]. A remarkable advantage of these models is their ability to capture complex underlying patterns without using feature engineering as a pre-processing step. Relevant to the problem at hand, string-matching approaches fail to accurately capture the full complexity of sequence data in a classification problem. For instance, the k-mer based approaches do not capture the positional information in a DNA/RNA sequence [119]. On the other hand, using k-mers with explicit positional information can significantly increase the size of the feature space, leading to model overfitting. Recently, it has been shown that in learning regulatory elements, deep neural networks can capture the complex dependencies between sequence positions [130–132]. Furthermore, these method can provide significant reduction in computational time by leveraging hardware accelerators.

Artificial Neural Networks (ANNs), particularly deep, multi-layered ANNs have previously been used in the prediction of tissue-specific alternative splicing, while taking into account numerous, handcrafted genomic and epigenomic features [133–135]. However, here we primarily focus on deep learning models in genomics that use CNNs to automatically infer features from sequence information. Towards that end, DeepBind has been one of the earliest—and perhaps the most influential—method that predicts protein binding site specificities within genomic sequences [98]. By current deep learning standards, DeepBind used a rather simple architecture with a single convolutional layer. Nevertheless, in contrast to its contemporary machine learning

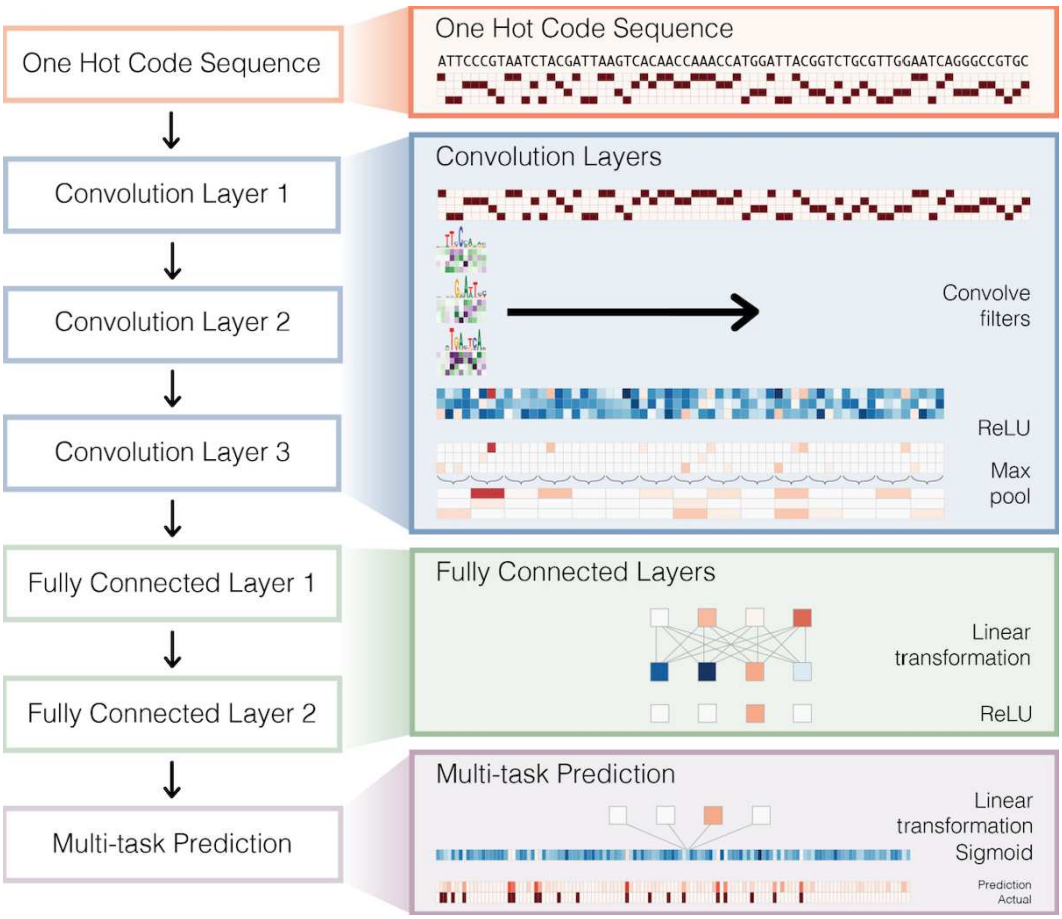


Figure 4.1: Summary of Basset architecture: to predict DHS occupancy across 164 human cell lines, Basset used three convolutional layers followed by multiple fully connected layers. This figure has been taken from [13].

methods, it exhibited superior performance while taking advantage of diverse experimental data. Moreover, DeepBind demonstrated the ability of CNNs to capture signal detectors that recapitulate known motifs. Similarly, another deep learning method, DeepSEA employed a multi-layer CNN to predict the chromatin effects of sequence alterations with single-nucleotide sensitivity [100]. Following the success of DeepBind and DeepSEA, a number of deep learning models have been developed that use RNNs in conjunction with CNNs: the addition of an RNN enables the models to capture long-term dependencies within sequence features, leading to an improved overall accuracy [99, 101].

A model that is worth discussing and has inspired part of this work, Basset, was developed by Kelley et al. to predict genome-wide chromatin accessibility across 164 human cell-lines and tis-

sues [13]. They used a deep neural architecture with three convolutional layers with max-pooling, followed by multiple fully connected layers. On the aforementioned task, Basset demonstrates greater predictive accuracy in contrast to a gapped k-mer SVM. Moreover, it can infer regulatory motifs within genomic sequences by interpreting weights of the first convolutional layer. In other words, Basset uses the activations of the first CNN layer filters in the input sequences and generates Position Weight Matrices (PWMs), that are representative of the sequence motifs. The architecture of Basset is summarized in figure 4.1. To predict chromatin accessibility in intron retention, we modified Basset's architecture and achieved measurable accuracy in terms of AUC and Area Under the Precision-Recall Curve (AUPRC). For more details, refer to Chapter 6.

Chapter 5

Intron Retention and Chromatin Accessibility in Plants

5.1 Introduction

We performed an initial study to explore the association between IR and chromatin accessibility in plants [52]. As mentioned in Chapter 1, in plants IR is the most prevalent form of alternative splicing. There is preliminary evidence in metazoans to suggest that chromatin structure may have an important role in the regulation of splicing; however, nothing is known about the role of chromatin structure in regulating IR in plants.

The fact that splicing can happen co-transcriptionally suggests that chromatin state is relevant for splicing [48, 136]. One of the primary tools for genome-wide exploration of chromatin is through exposure of DNA to Deoxyribonuclease I (DNAse I), which is an enzyme that cleaves DNA; sites that are sensitive to its action—DNase I hypersensitive sites (DHSs)—have been used as an indicator of regions in the DNA that are accessible *in-vivo*. DHSs have been used to identify several types of regulatory elements such as, promoters, silencers, enhancers, and insulators [137, 138]. It has been shown that when a protein binds a region of DNA, it protects it against the action of DNase I [139] and leaves a footprint which can be identified using DNase I-seq data [140, 141]. The ENCODE consortium has shown that DHSs identified in the human genome are robust markers for several genetic regulatory phenomena, including histone modifications, early replication regions, transcription factor binding sites, and transcription start sites [142].

When it comes to AS, Mercer et al. [81] have shown an association between DHSs and exon-skipping, reporting that higher numbers of DHS-containing exons are alternatively spliced. Furthermore, this study claims that DHS exons with promoter and enhancer-like features have a higher fractional overlap with AS. Specifically related to this work, the cross-talk between chromatin or-

ganization and IR has been studied in mammals [51]. They explore the co-transcriptional regulation of splicing reporting higher chromatin accessibility in retained introns and how polymerase II elongation speed affects IR and vice-versa. DNase I-seq has been used in plants [2, 143], but the data has not been analyzed in the context of AS.

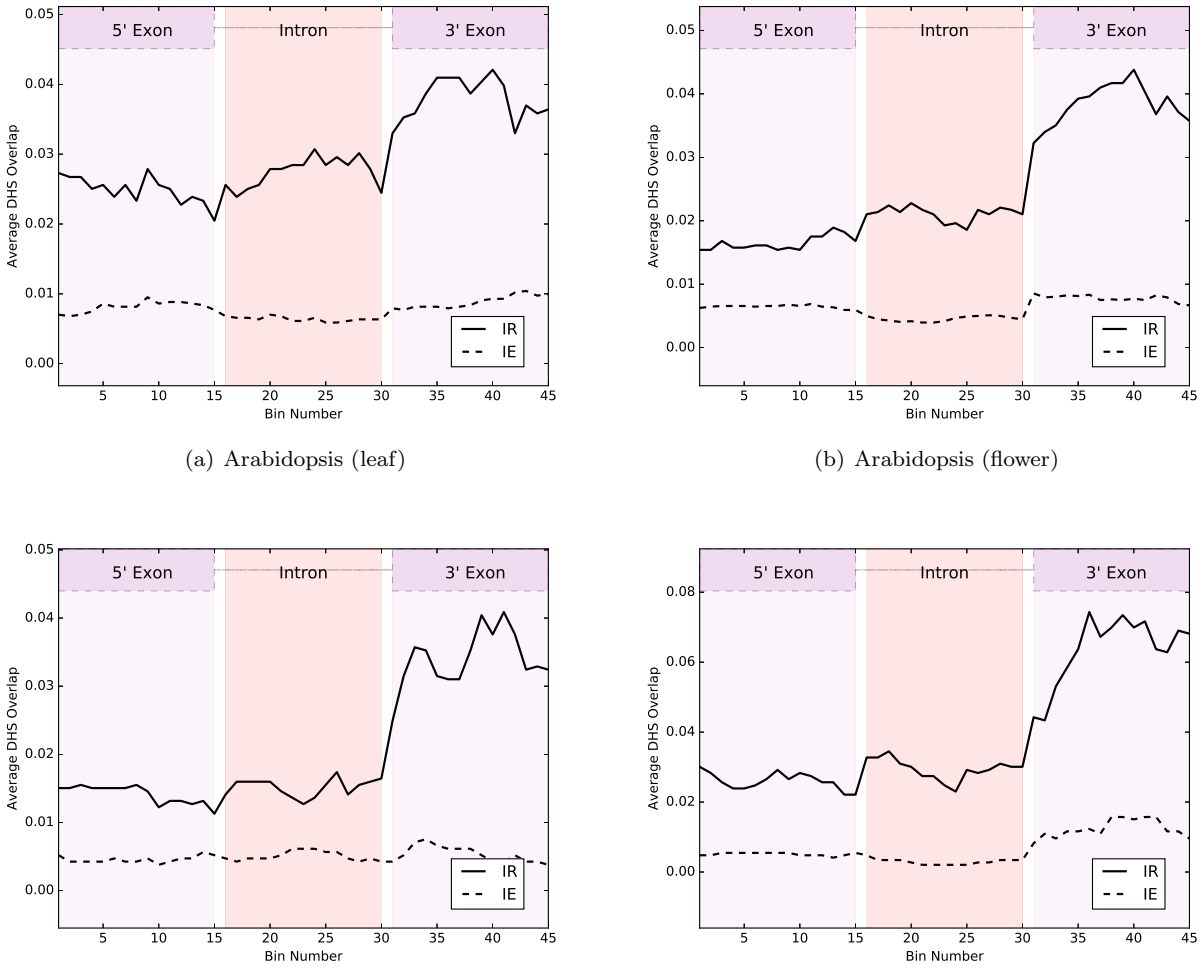
Our goal is to shed light on the regulation of IR from the perspective of chromatin organization. First we test the association between DHSs and IR using DNase I seq data in arabidopsis and rice, and find that DHSs have a highly significant association with IR; we then look for evidence at the DNA level for the footprints of protein binding and find a large collection of hexamers that are conserved across arabidopsis and rice, and likely function as SREs. Finally, we discuss how these observations are consistent with current models that describe the interaction between transcription, splicing, and chromatin organization.

5.2 Results

5.2.1 DHSs are enriched in IR events

Our first goal is to investigate the relationship between IR and chromatin accessibility. For this task we analyzed existing DNase I-seq data in both arabidopsis and rice for which RNA-seq data for the same samples is also available [2, 143]. First, we used the RNA-seq data to identify events where an intron is retained (IR), and events where there is no evidence for IR, which we refer to as intron excision (IE). Note that we do not use the term "constitutive splicing", as other alternative splicing events could be occurring. The DNase I-seq data associated with those samples were then used to identify peaks representing DHSs. We observe that IR events tend to overlap DHSs to a much greater degree than IE events: 13.3-26.5% of IR events overlap a DHS compared to 2.1-5.2% for IE, a difference that is highly statistically significant (see Table 5.1, Figure 5.1, and in Appendix A, Table A.2 and Table A.4 for details). Since expressed genes typically exhibit a large peak in DNAse I-seq coverage in their promoter region, we excluded IR/IE events in the first intron of a gene. Consistent with the above results and the higher chromatin accessibility of the first introns, they exhibit significantly higher rates of IR than other introns in both arabidopsis

and rice with a p-value of 5.90×10^{-89} in arabidopsis and a p-value of 8.93×10^{-25} in rice using the Fisher exact test.



(a) Arabidopsis (leaf)

(b) Arabidopsis (flower)

(c) Rice (leaf)

(d) Rice (callus)

Figure 5.1: DHS content profiles in IR and IE. For each sequence bin within an IR/IE event we show the frequency with which that bin overlaps a DHS. Profiles are computed for arabidopsis leaf samples (a), arabidopsis flower samples (b), rice leaf samples (c), and rice callus samples (d). In all samples, we see overall higher DHS occupancy across IR events compared to IE events, suggesting a more open chromatin in IR. Moreover, the DHS content is much higher in the 3' exons of IR events.

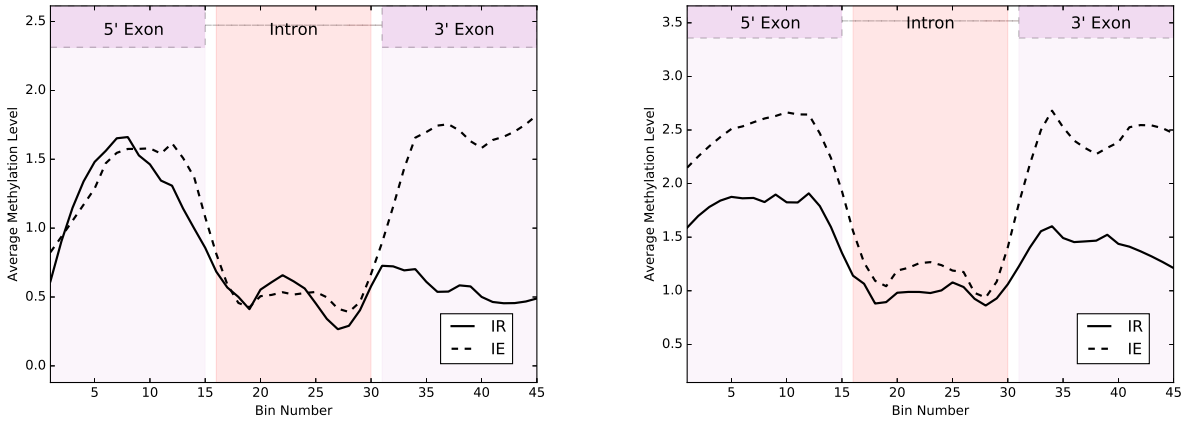
5.2.2 IR events exhibit higher chromatin accessibility than IE events

As a complement to the analysis of DHSs detected using peak calling, we compared IR and IE events on the basis of raw DNase I-seq read depth (see Figure A.1 in Appendix A). In agreement

Table 5.1: Enrichment of DHSs in IR and IE events. DHS content is the fraction of IR/IE events with an overlapping hypersensitive site. The significance of the difference is quantified by the Fisher exact test.

Data Source	**DHS Content**		**p-value**
	IR	***IE***	
Arabidopsis (leaf) [2]	15.24%	4.02%	1.07×10^{-66}
Arabidopsis (flower) [2]	13.28%	3.49%	9.43×10^{-93}
Rice (leaf) [143]	16.07%	2.13%	2.29×10^{-123}
Rice (callus) [143]	26.46%	5.21%	3.61×10^{-104}

with the higher proportion of DHSs associated with IR, we observe that IR events have a much higher mean DNase I-seq coverage than IE events (p-value of 1.22×10^{-56} in arabidopsis, and a p-value of 5.25×10^{-100} in rice using the Mann–Whitney U test [144]), demonstrating that chromatin is more open in IR events than in IE events. As further evidence we analyzed methylation profiling data in arabidopsis and rice, and found that IR events exhibit lower methylation levels in the 3' exon (see Figure 5.2). This is consistent with the results we reported using DNase I-seq data, as DNA methylation has been reported to have an inverse correlation with chromatin accessibility [88].

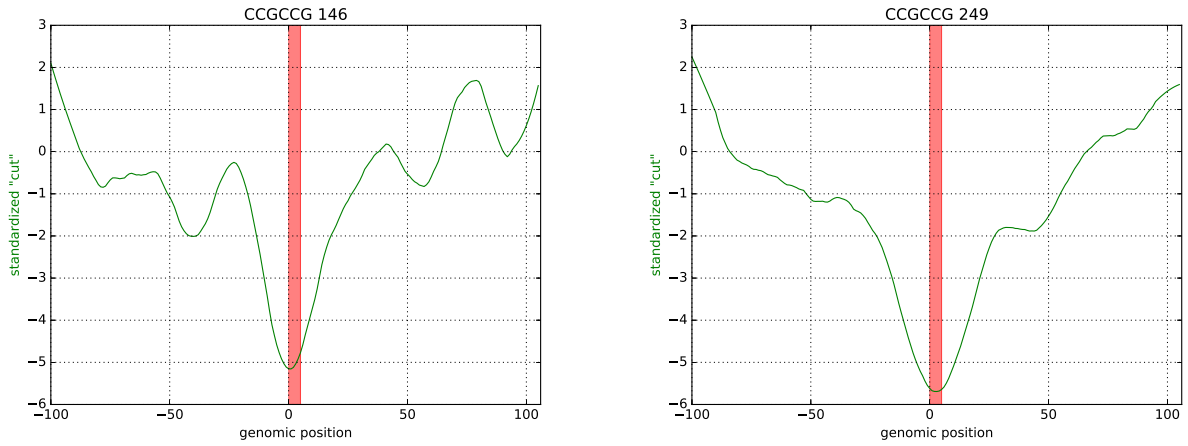


(a) Arabidopsis (b) Rice

Figure 5.2: Methylation profiles in IR and IE Methylation levels are shown across introns and their flanking exons in IR and IE events in arabidopsis (a) and rice (b).

5.2.3 Protein footprint analysis

Previous studies have used DNase I-seq data to detect potential transcription factor binding sites in promoter regions by searching for a dip in the DNase I-seq coverage [140]: a region of more accessible chromatin is interpreted as the "footprint" left by protein binding. Since splicing occurs co-transcriptionally, there is a potential for events at the DNA level to directly affect splicing, e.g. via recruitment of splicing factors through their interaction with DNA-binding proteins [48]. We used a continuous Hidden Markov Model (HMM) described in the Methods section to discover the footprints of protein binding by searching for a footprint in all occurrences of a given hexamer. A representative footprint is shown in Figure 5.3, which shows the DNase I-seq data profile for the hexamer `CCGCCG`, that was detected by our HMM to have a footprint in 3' exons, in both arabidopsis and rice. This hexamer is over-represented in IR events (p-value of 0.0008 in arabidopsis, and a p-value of 1.07×10^{-24} in rice, computed using the Fisher exact test).



(a) Arabidopsis (leaf)

(b) Rice (leaf)

Figure 5.3: HMM footprint detection. The hexamer `CCGCCG` was detected to have a footprint at the location of the hexamer (red bar) in the standardized DNase I-seq data profile in both arabidopsis (a) and rice (b). The number of occurrences of the hexamer in IR/IE events is shown next to the k-mer in the title of each sub-figure. The profile extends 100bp upstream and downstream of the hexamer location, and is used by our HMM to score the k-mer for a potential footprint. In both cases, we see a clear dip in coverage indicating a possible footprint at the hexamer location.

We performed a comprehensive analysis across all hexamers to detect those that have a footprint and exhibit an association with IR or IE in the arabidopsis and rice leaf data. Our first observation is that in IR events the majority of the hexamers come from the 3' exon, while for IE, all the hexamers are intronic (see Table 5.2 for details). In rice we identified a much larger number of hexamers in IR events, likely due to greater read coverage of the DNase I-seq data (see Table A.1 in Appendix A).

Table 5.2: Enriched hexamers exhibiting a footprint. For each of the four datasets we provide the number of hexamers that exhibit a footprint and are also enriched in either IR or IE events. The number of enriched footprint-hexamers are shown in each of the three regions of an event: 5' exon, intron and 3' exon. An HMM score cutoff of $S = 0.30$ was used to generate the footprint hexamers.

Sample	**IR Events**			**IE Events**		
	5' Exon	***Intron***	***3' Exon***	***5' Exon***	***Intron***	***3' Exon***
Arabidopsis (leaf) [2]	12	6	100	0	28	0
Arabidopsis (flower) [2]	4	3	105	0	27	0
Rice (leaf) [143]	88	75	262	0	14	0
Rice (callus) [143]	46	32	192	0	30	0

Many of the hexamers we identified are conserved in arabidopsis and rice: In the upstream 3' exon 246 hexamers were common between the two species, while 19 are conserved in the intronic region of IE events. This level of overlap is highly statistically significant (p-values of 2.25×10^{-165} and 2.10×10^{-32} respectively, in a hypergeometric test). This level of conservation is strong support for the functional importance of these hexamers. We note that for finding conserved hexamers we used a looser threshold for footprint calling, as the requirement of conservation provided an additional level of filtering of potential false positives. Manual inspection of the detected hexamers showed that all of them exhibited valid footprints.

The conserved hexamers in leaf tissue were clustered into motifs that are summarized in Table 5.3. In IE we detected motifs only in the intron; these motifs are `T`-rich with a few `A`s and no `G`s or `C`s. The converse holds for intronic motifs in IR: they are `GC` rich with few `A`s and no `T`s.

Furthermore, occurrences of the intronic IE motifs show a clear pattern in terms of their preferred position within the intron, with a very clear peak near the 3' of the intron, and are likely

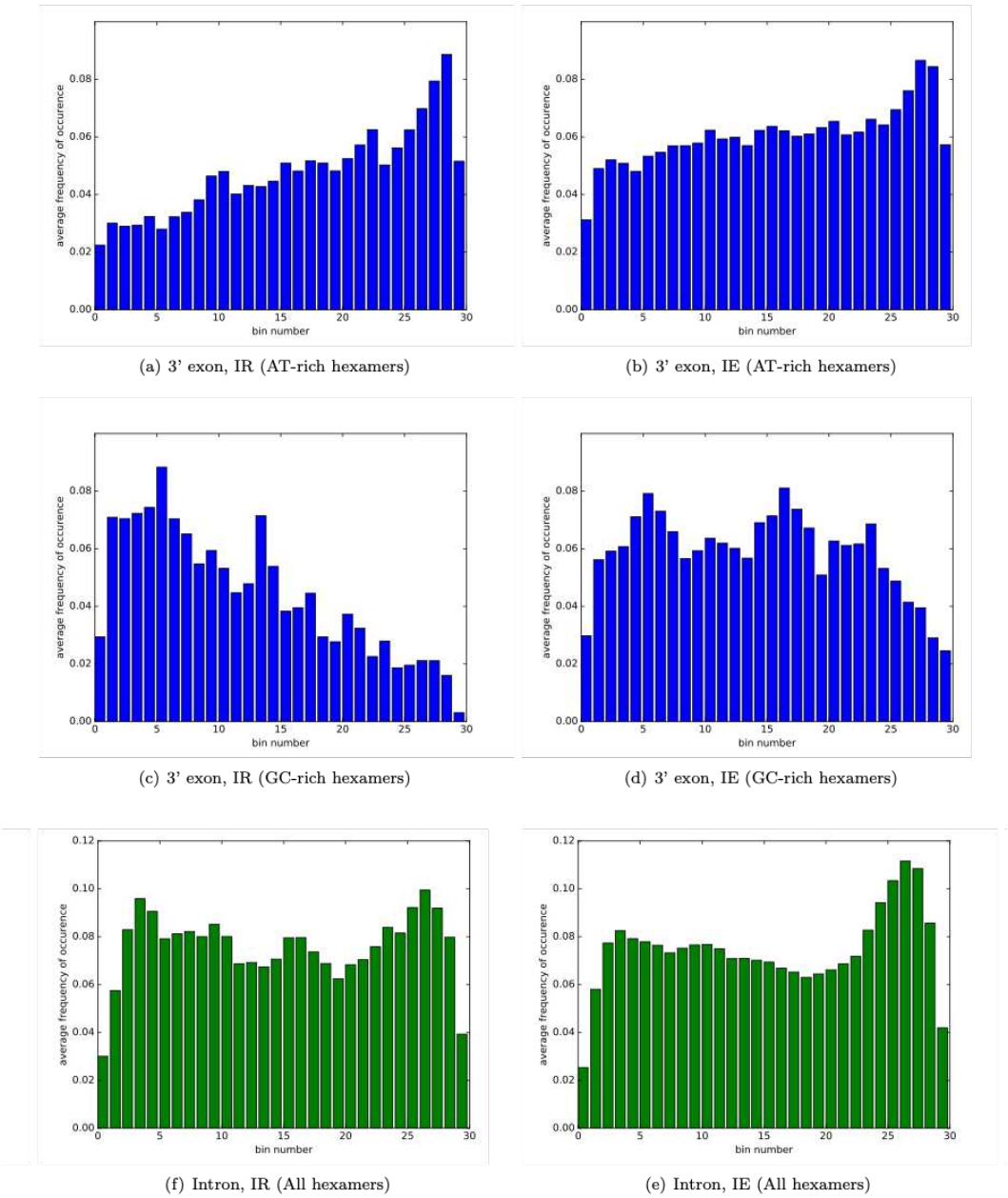
Table 5.3: Common enriched footprint-hexamers between arabidopsis and rice. The number of hexamers in common between the arabidiopsis and rice leaf samples, and the corresponding significance levels of the overlap are shown for all three regions of IR and IE events. The hexamers in each region were clustered, and motif consensus sequences are shown. When there is no clear consensus in a given position, that is denoted by an `x`. Leading or trailing positions without a clear consensus were omitted, so some consensus sequences are less than 6 nucleotides long. In the intronic region of IR events only 2 hexamers were detected so no clustering was performed. Here, an HMM score cutoff of $S = 0.20$ was used with manual verification of the footprints of the overlapping hexamers.

Event type	Region	hexamers	p-value	Motif consensus
IR	5' Exon	13	1.70×10^{-07}	`CGCCG,(G/C)(G/C)GCGG,(A/G)T(C/T)(G/T)(C/G)A`
	Intron	2	0.27	`AAGGAG,CGGCGG`
	3' Exon	246	2.25×10^{-165}	`AAAA,AAATT,CCGAC,CGCxCG,(C/A)TTT,GCGGC,GxTTT,(T/G)AAA,TTT(C/T),(G/T)T(C/T)(C/G)(G/A)`
IE	5' Exon	0	N/A	-
	Intron	19	2.10×10^{-32}	`TTAA(T/A)(T/A),T(T/A)TTT(A/T)`
	3' Exon	0	N/A	-

associated with the polypyrimidine tract (see Figure 5.4). No such pattern is observed for the IR intronic motifs.

Most of the hexamers and motifs associated with IR events occur in the 3' exon; the majority of them (6/10) are `AT`-rich, and some of the rest (3/10) are `GC`-rich. Both sets of motifs exhibit very different positional preferences: the `AT`-rich motifs tend to occur at the 3' end of the exon, while the `GC`-rich motifs tend to occur in the 5' end of the exon (see Figure 5.4 for the overall positional preferences of those motifs, and Figure A.4 in Appendix A for positional preferences of individual hexamers [52]). We believe that the positional preferences observed reflect different biological roles of these motifs in regulating IR and IE events, as discussed below.

In order to find potential proteins associated with our hexamers we searched all the arabidopsis hexamers against a collection of 410 transcription factor motifs from the Plant Cistrome [145] as described in the Methods section. Out of 280 enriched hexamers, 96 of them had at least one match. The breakdown into the different locations is found in Table A.7 in Appendix A. The matching motifs come from a variety of families of transcription factors. The largest number of matches was to the AP2/EREBP family, which is a plant-specific family of DNA-binding proteins [146].



(a) 3' exon, IR (AT-rich hexamers)

(b) 3' exon, IE (AT-rich hexamers)

(c) 3' exon, IR (GC-rich hexamers)

(d) 3' exon, IE (GC-rich hexamers)

(f) Intron, IR (All hexamers)

(e) Intron, IE (All hexamers)

Figure 5.4: Hexamer positional preference. Average positional preference is shown for AT-rich footprint-exhibiting hexamers in the 3' exon region of IR events (a), and for comparison, the same hexamers in the 3' exon region of IE events. Similarly, (c) and (d) show average positional preference for GC-rich hexamers in the 3' exon region of IR and IE events, respectively. To demonstrate the positional preference of footprint-exhibiting hexamers that are associated with IE events we show the average positional profile of those hexamers in IE events (f) and IR events(e).

The second-largest number of matches were to Dof proteins through hexamers in the 3' exon that contain mostly A or T nucleotides; this family of transcription factors is also plant-specific [147].

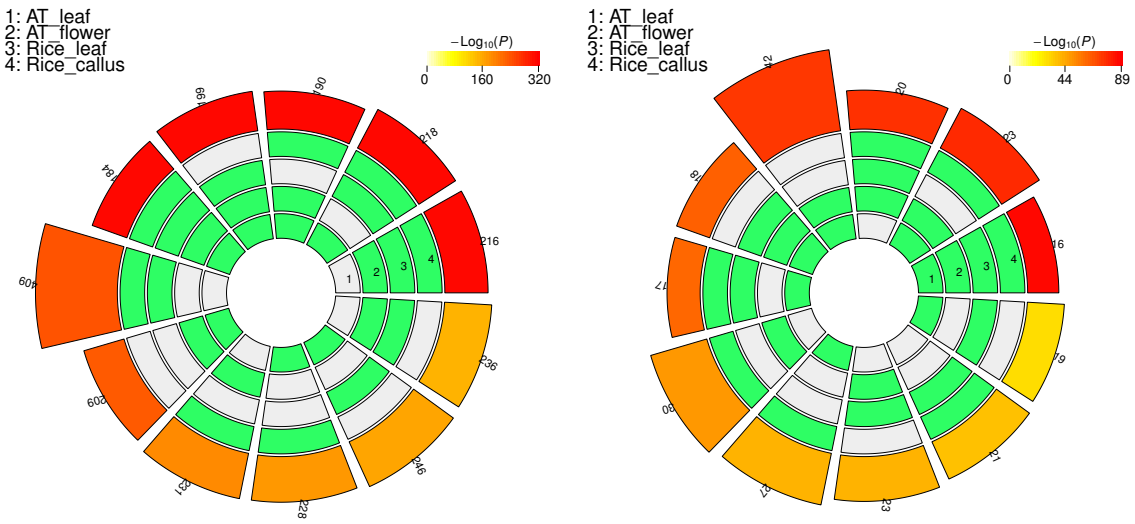
C2H2 DNA-binding proteins are also strongly represented. Interestingly, a vast majority (about 60%) of them have been shown to be involved in the regulation of AS in animals [148], although the effect could be either direct or indirect, through the regulation of splicing regulators. Some of these effects are likely to be direct since DNA-binding proteins, including transription factors, have been shown to bind in gene bodies [149]. These results implicate plant transcription factors in splicing regulation. This is in agreement with recent results in mammals that revealed that more than a third of splicing regulators detected in a high-throughput screen were transcription factors [148].

Next, we performed an additional enrichment analysis to test the significance of the overlap across all four datasets (arabidopsis leaf and flower tissue and rice leaf and callus). We used the *SuperExactTest* [150] to quantify the overlap between all subsets of samples simultaneously. Since the majority of hexamers occurred in the 3' exon of IR events and intronic part of IE events, we performed this analysis in those regions. The results shown in Figure 5.5 demonstrate a large and highly statistically significant overlap even when considering all combinations of samples.

5.3 Discussion

Splicing occurs co-transcriptionally, and there is increasing evidence indicating that chromatin organization involving epigenetic marks and rate of transcription regulate alternative splicing in mammalian systems [48]. However, in plants, virtually nothing is known in terms cotranscriptional regulation of alternative splicing. Here we investigate the role of chromatin architecture and potential DNA elements that may regulate IR.

In our data we observe a greater number of DHSs in IR events compared to IE events, and this is most prominent in the 3' exon. A similar pattern was observed in the raw DNase I-seq data as well. We present two possible hypotheses by which this increase in open chromatin contributes to IR. Splicing is a much slower process than transcription [48], and we hypothesize that the less open chromatin in IE events leads to more PolII pausing (the speed-bump model), which allows for a greater degree of recruiting of splicing factors and hence greater likelihood of intron



(a) 3' Exon (IR)

(b) Intron (IE)

Figure 5.5: Significance of overlap of enriched hexamers. The significance of overlap among enriched hexamers exhibiting a footprint is shown in the 3' exon region of IR events (a), and in the intronic region of IE events (b) for two or more samples. The overlap is shown in circular layout for all possible combinations of two or more of the four samples. The four inner sections of each slice represent the four samples and a sample is labeled in green if it is included in a particular combination. The right most slide provides the labeling of the samples. The size of the fifth section in each slice represents the number of hexamers in an intersection of the corresponding samples. The actual number of overlapping hexamers is also shown. Finally, the color of the fifth section indicates the significance of overlap (p-value). The intersections are sorted based on p-value, starting at the labelled segment in an anti-clockwise fashion.

recognition. Conversely, in retained introns, because of the higher elongation rates, there is less chance of recognizing the splice sites, leading to IR. The fact that retained introns have weaker splice sites [51,151], makes them more sensitive to the rate of elongation. However, this hypothesis does not take into account that the increased prevalence of DHSs could be due to binding of *trans*-factors, and also does not account for the much larger number of hexamers with footprints that are associated with IR. For example, in the arabidopsis leaf data we found 118 hexamers with footprints that are enriched in IR, and only 28 in IE.

The increased number of footprints that we observed in IR could be the result of one of two factors: 1. Increased PolII pausing and/or, 2. Binding of other chromatin/DNA-interacting proteins. Braunschweig et al. have recently shown that in mammalian systems retained introns are associated with increased PolII pausing [51]. This pausing may lead to recruitment of splicing

suppressors that compete or prevent splicing activators from binding, leading to IR. There is data supporting this hypothesis in non-plant systems [51], and this hypothesis is consistent with the observation that the high rate of DHS occurrence in the 3' exon is coupled with the occurrence of a much higher number of hexamers with footprints that are associated with IR. This suggests a key role for chromatin architecture in the 3' exon in regulating the splicing of the upstream intron. We believe the second mechanism is more likely; however additional work aimed at assaying PollII occupancy in retained vs excised introns is required to help distinguish between these two mechanisms.

Chromatin modifications have recently been associated with IR in humans: Braunschweig et al. have shown that the chromatin activation mark H3K27ac is enriched in retained introns [51]. This observation is consistent with our result showing greater DHS frequency in retained introns: this modification is associated with more flexible chromatin structure, which facilitates the interaction of proteins with IR regulatory elements.

The AT-rich hexamers in IE have a positional preference for the 3' end of the intron, which suggests they are likely associated with the polypyrimidine tract, which in plants is T-rich [16], leading to more efficient recognition of splice sites. In contrast, the hexamers we detected in the introns of IR events, show very different base composition, with virtually no Ts, likely resulting in poor recognition of these introns.

DNA methylation has been shown to regulate alternative splicing, including IR, in plants and animals [49, 50, 152, 153]. Part of this regulation could be due to reorganization of chromatin; in support of this, it has been shown that there is an inverse relationship between DNA methylation and open chromatin [88]. In our analysis we found a strong correlation between open chromatin and reduced methylation in IR vs IE events in both arabidopsis and rice. Open chromatin may make the DNA more available to binding by DNA-binding proteins. In our hexamer analysis we found that the majority of those hexamers occur in the 3' flanking exon, which demonstrated the highest level of open chromatin. Interestingly, the motifs in the introns of IR events are either CG- or AG-rich. Hence, it's possible that the hexamers enriched in CG di-nucleotides are the tar-

gets of methylation, which in turn could attract splicing suppressors, either directly, or through methylation-binding proteins [49]. Alternatively, proteins bound to methylated regions can modulate the rate of elongation of PolII [47, 49]. Further studies are required in order to confirm or exclude some of these possibilities.

In addition to the matches in the Plant Cistrome Database described above, we identified other transcription factors that have DNA binding motifs that match the hexamers discovered by our pipeline. These include Homeodomain-leucine zipper (HD-Zip) proteins, which are a family of transcription factors unique to plants [154] have DNA binding sequences that match some of the `AT`-rich hexamers that were detected in our analysis. For example, *ATHB9*, which is an HD-Zip class II protein, was shown to have affinity for the sequence `GTAAT(G/C)ATTAC`; the core `AAT(G/C)A` segment of this sequence matches multiple conserved hexamers detected in the 3' exon of retained introns. HD-Zip class IV proteins bind sequences containing a `TAAA` core, which is consistent with a large number of hexamers both in IR and IE events.

Although epigenetic changes, including DNA methylation and histone modifications have been shown to be important regulators of AS in animals [45, 49, 127], relatively little is known about their role in AS in plants. This work strongly indicates a role for chromatin organization and DNA methylation in IR. Recently Pajoro et al. [155] have shown that histone modifications alter AS in plants, supporting our conclusion that chromatin state is a critical regulator of AS.

5.4 Conclusions

In this work we established a clear correlation between IR and chromatin accessibility and DNA methylation in arabidopsis and rice. We found that chromatin is more open in retained introns, which can be explained using a kinetic model of the splicing process. The observed open chromatin in IR is consistent with the reduced methylation levels we observed in these regions. The more open chromatin in IR also suggests that IR is more highly regulated than constitutive splicing, which is supported by the large number of conserved sequence elements that were discovered in footprints associated with IR. A majority of the discovered sequence elements occur in exons

immediately downstream of retained introns, indicating its importance in regulating IR events. Further experiments are required in order to establish the biological function of these sequence elements and to experimentally verify the hypothesized connections between intron retention and chromatin organization.

5.5 Materials and methods

5.5.1 Data collection

For arabidopsis, the raw reads data from Zhang et al. [2] (GEO accession number $GSE53322$) was used. For rice, we used data from Wu et al. [143] (GEO accession number $GSE26610$); The corresponding RNA-seq was published elsewhere [156] (GEO accession number $GSE33265$). For rice, there were two samples coming from two tissues: leaf and callus. For bisulfite-seq, we used raw data from Zemach et al. [157] (GEO accession number $GSE41302$) and Chodavarapu et al. [158] (GEO accession number $GSE38480$), for arabidopsis and rice, respectively.

5.5.2 Alignment and processing

In case of data from Zhang et al. [2], we used their aligned DNase I-seq and RNA-seq files. For the rest of the data, the raw reads were first pre-processed using *FastQC* [159] and trimmed using *fastx-trimmer* [160] when required. Next, the processed reads were aligned to the corresponding reference genomes (*TAIR10* for arabidopsis and *MSU v7* for rice) using different alignment tools. All the RNA-seq samples were aligned using *Tophat2* [79] with default parameters. The *Tophat2* alignments were filtered to obtain only uniquely aligned reads. The arabidiopsis DNase I-seq data was aligned using *Bowtie* [89]. *Bowtie* was used with the command-line argument `-m 1` to suppress multiple alignments. For the rice DNase I-seq data, we used *STAR* [80] to align the reads with the parameters `outFilterMultimapNmax 1` and `alignIntronMax 1` to adjust for genomic data alignment. The bisulfite-seq data was quality- and adapter-trimmed using *Trim Galore!* [161]. For alignment and methylation calling, we used *bismark* [92]. Note that biological

and technical replicates—if there were any—were pooled together for each sample. The alignment statistics are summarized in Table A.1 in Appendix A.

5.5.3 Extraction of IR/IE events and peak calling

To extract IR and IE events we used annotated IR events from the gene models as well as evidence from the RNA-seq data found using SpliceGrapher [1], which is a tool that combines gene models and RNA-seq data to predict alternative splicing events. To avoid any ambiguity between IR and IE events, we used strict criteria to distinguish between the two on the basis of the RNA-seq data: exonic read depth of at least 20 was required for a gene to be considered in our analysis; full coverage across an intron was required for it to be considered retained, and no coverage for it to be considered an intron excision event. The choice of the exonic read depth threshold had little effect on our results (see Table A.4 in Appendix A). For DHS peak calling in the DNase I-seq data, in both arabidopsis and rice, we used the Hotspot [91] program with default parameters. Table A.2 summarizes the DHS peaks and the numbers of IR/IE events are provided in Table A.3 in Appendix A. When computing the DHS content profile and DNase I-seq coverage profiles across IR/IE events we excluded events involving the first intron of a gene, since the first intron often overlaps the DHS associated with the promoter region, and tends to exhibit higher DNase I-seq coverage than introns further downstream. As a further step for addressing the non-uniformity of DNase I-seq coverage across a gene, for each IR event, we selected IE events with similar relative positions within their genes.

5.5.4 Protein footprint analysis

Hexamer data generation

For the discovery of k-mers that exhibit footprints we chose to focus on hexamers since this provides a good balance of specificity and tractability of exhaustive search. We considered all possible hexamers coming from the three parts of an event: 5' exon, intron, and 3' exon. For every hexamer, we generated the DNase I-seq profile. For each occurrence of the hexamer we

extracted its DNase I cut at every nucleotide position of the hexamer as well as 100bp upstream and downstream of its location and then took the average over all positions. Note that in going 100bp upstream and downstream, we made sure not to go beyond the boundaries of the event parts: intron or the flanking exons. This was done to avoid introducing any bias coming from the properties of different segments of the event. In case of multiple instances of a hexamer in a sequence, we considered the one which had the lowest DNase I-seq coverage.

Footprint calling using continuous HMMs

We used the profile of DNase I-seq coverage to call footprints using a continuous HMM. Continuous HMMs are a good modeling tool for sequences of real values such as DNase I-seq coverage, and allow us to detect whether the observed profile contains a feature that can be identified as a protein footprint. Our model was inspired by a similar model [141] and the implementation uses SageMath [162]. As shown in Figure 5.6, our HMM has five core states: the leading background state (BG_1), the down state (DN), the footprint state (FP), the up state (UP) and finally, the trailing background state (BG_2).

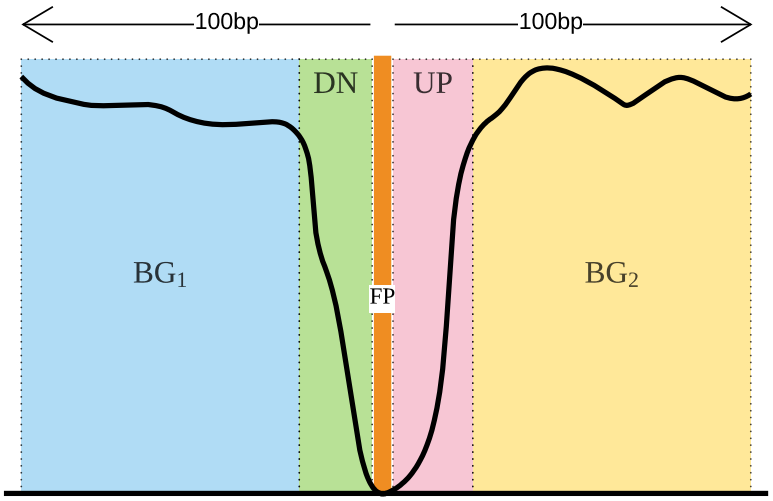


Figure 5.6: HMM Architecture The core continuous HMM states used to discover footprints are shown. The five states represent different regions of the DNase I-seq coverage profile: leading background (BG_1), down (DN), footprint (FP), Up (UP), and trailing background (BG_2). The footprint state is shown in the center, within the "dip" in the DNase I-seq coverage.

The HMM was trained on data profiles of hexamers with manually verified footprints and was used to score the rest of the hexamers. Note that all hexamer profiles were standardized to a background score calculated from the training set. To account for tandem motifs, we added additional states to the model to represent secondary footprints upstream or downstream of the primary footprint. The state diagram for the final HMM, which has 13 states, along with complete specification of the model (transition and emission probabilities), and the training and testing protocol, can be found in the Table A.5, Table A.6 and Figure A.3 in Appendix A.

Using the trained HMM we score hexamers as potential footprints using the following expression:

$$S = -\log\left[\frac{C_{FP}}{C_{BG}}\right],$$

where C_{FP} is the average standardized coverage at the footprint state and C_{BG} is the average coverage across the background states. A conservative threshold of $S = 0.30$ was used in the analysis of individual hexamers, and the cutoff was lowered to $S = 0.20$ in the cross-species analysis. To cluster the hexamers into motifs, we used complete linkage hierarchical clustering with a distance metric that assigns two k-mers a distance of 0 if they shared a 4-mer, and then edit distance was applied; clusters were cut at a depth of 4. We used *clustalw2* [163] to generate the multiple alignments which were then fed to *weblogo* [164] to generate motif logos. For positional preferences, when a hexamer occurred multiple times in an IR/IE event, we chose the one with lowest DNase I-seq read depth among all occurrences.

Motif matches in the plant cistrome database

All significantly enriched arabidopsis hexamers were searched against each motif from the Plant Cistrome Database [145] using their respective position weight matrices. A cistrome motif was considered a match for a given hexamer if the hexamer matched exactly the consensus sequence at some location, such that the information content in the positions covered by the hexamer consist of at least 50% of the overall information content of the motif.

5.5.5 Statistical tests

Whenever testing multiple hypotheses, the resulting p-values were adjusted using the Benjamini-Hochberg method [165]. All the statistical tests used in this work were performed in R; for the significance of multi-sample intersections, we used the R package for the super exact test [150] with population size of 4096.

Chapter 6

Predicting Intron Retention using Deep Learning

6.1 Introduction

A growing number of studies have shown the role of chromatin state in the regulation of alternative splicing. As mentioned in the previous chapter, Mercer et al. [81] showed an association between DHSs and exon-skipping, reporting that higher numbers of DHS-containing exons are alternatively spliced. Similarly, the cross-talk between chromatin organization and IR has been studied in mammals by Braunschweig et al. [51] where they explored the co-transcriptional regulation of splicing, reporting higher chromatin accessibility in retained introns and how polymerase II elongation speed affects IR and vice-versa. When it comes to identifying regulatory proteins associated with AS, Han et al. [148] reported a regulatory role of zinc finger transcription factors in exon skipping. This is in agreement with our work in plants (see Chapter 5) where we identified potential regulatory elements occurring primarily in the 3' flanking exon of IR events, several of which significantly match plant zinc finger transcription factor binding site motifs.

As further motivation for considering the role of Transcription Factors (TFs) in splicing regulation, we explored the frequency of motif matches for different TF families across regions of open chromatin in the human genome. We observe in Figure 6.1 that the prevalence of motif matches in the human intragenic regions is significantly higher than the promoter regions; a similar observation was also made in plants [166]. We validated this observation using ChIP-Seq data for five TFs from the ENCODE database [14], and observed similar behavior to that observed in Figure 6.1. This suggests a regulatory role of transcription factors beyond the regulation of gene expression.

Deep neural networks have become the tool of choice for exploring complex phenomena such as chromatin accessibility and structure [13, 96, 103]. A remarkable advantage of these models is their ability to capture the underlying patterns in large noisy datasets directly from sequence with minimal pre-processing, learning motifs of the regulatory proteins involved as part of the

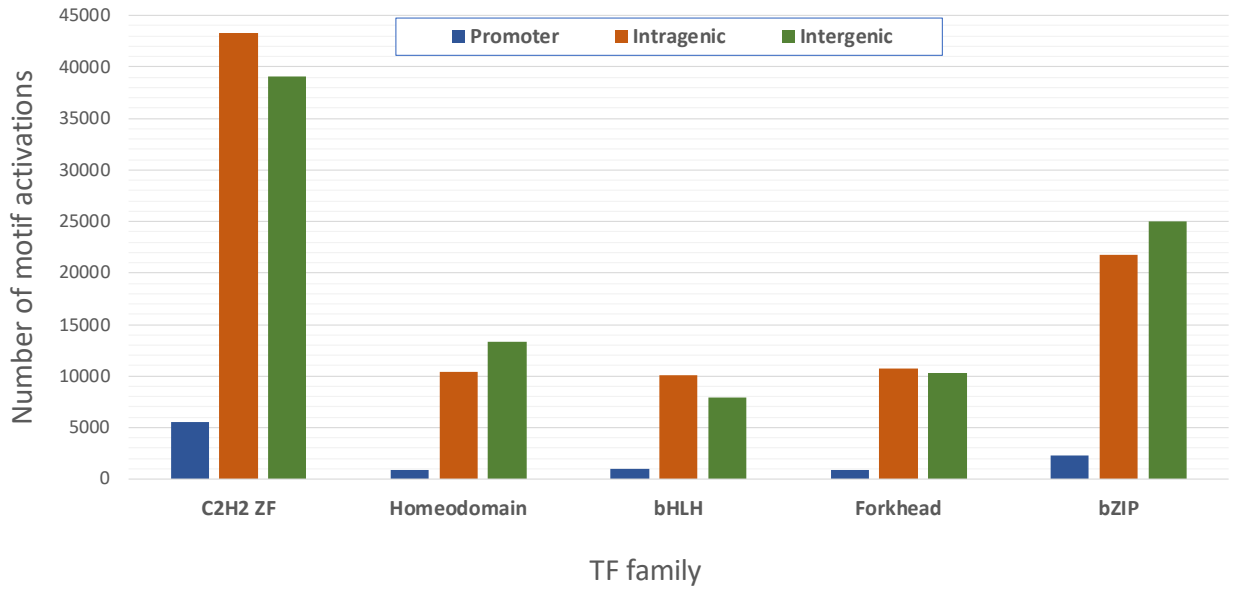


Figure 6.1: The distribution of different transcription factor families in the promoter, intragenic, and intergenic regions of the human genome. These statistics were obtained by training the Basset-like network [13] and analyzing the motifs learned by the network (see supplementary methods in Appendix B for more details).

training process. Deep learning has been used in genomics for gene expression analysis [97, 167], TF binding prediction [98–101, 168], chromatin accessibility analysis [13, 102, 103], prediction of chromatin structure and its modifications [104, 169], identification of RNA-binding protein sites [105, 167], and alternative splicing [170–172].

In this study we demonstrate that deep learning models can distinguish with good accuracy regions of open chromatin associated with IR from other intragenic regions of open chromatin using DNase I-Seq data across 164 different immortalized human cell-lines and tissues [14, 173]. The basis for this study lies in the fact that the proteins that regulate AS should bind in the vicinity of the splicing events. DHSs are regions of open and accessible chromatin where transcription factors and other regulatory proteins bind. Since a protein can only bind in accessible regions of the chromatin, DHSs that occur within or in the proximity of AS events are excellent candidates to look for the potential binding sites (and motifs). Our model is based on that knowledge in that if a DHS overlaps an IR event, and accommodates the IR-specific regulatory proteins, then the model should be able to discriminate it from a non-IR DHS.

By analyzing the motifs learned by the network, we find that specific families of TFs are associated with IR events, mostly members of the zinc finger family of TFs; results of ChIP-seq experiments for multiple zinc finger TFs in the K562 cell line support our findings for this association. Our work provides convincing evidence for a novel role of TFs in gene regulation, proposing a direction for further research.

6.2 Methods

6.2.1 Data collection, processing, and representation

We use DNase I-seq data from 125 human immortalized cell-lines and tissues from the ENCODE database [14] and 39 cell types from the Roadmap Epigenetics consortium [173] as processed by [13]: every DNAse I-seq peak is extended to a length of $600bp$ around its midpoint and adjacent peaks are greedily merged until no two peaks overlap by more than $200bp$. For our analysis we focus on over a million DHSs that occur within genes.

Next, we extracted IR events from the Ensembl GRCh37 (hg19) reference annotations, utilizing code from SpliceGrapher [1] and IdiffIR [174]. In total, we identified $58,305$ unique IR events out of which, $15,400$ had overlapping DHSs. These constitute our positive examples. We use a strict criterion requiring the DHS to overlap the retained intron, i.e., DHSs overlapping only the flanking exons do not qualify. All other intragenic DHSs that did not overlap an IR event are labelled as negative examples. The number of negative examples was twice the size of our positive set.

We use two methods to transform the sequences into input for the neural network: one-hot encoding and sequence embedding. For one-hot encoding a sequence is represented as a $4 \times N$ matrix where N is the length of the sequence. Each position in the sequence is represented by the columns of the matrix with a non-zero value at a position corresponding to one of the four DNA nucleotides. To represent a sequence using word embedding we first decompose it into overlapping *k*-mers of length *k*, and then train a word2vec model [175] to map each *k*-mer into an *m*-dimensional vector space. This gives us an embedding matrix of dimensions $(N - k + 1) \times m$. This representation is designed to preserve the context of the *k*-mers by producing similar

embedding vectors for k-mers that tend to co-occur. Recently in a TF binding site prediction task within genomic sequences, it has been shown that in contrast to one-hot-encoding, k-mer embedding representation of the input leads to improved model performance [176].

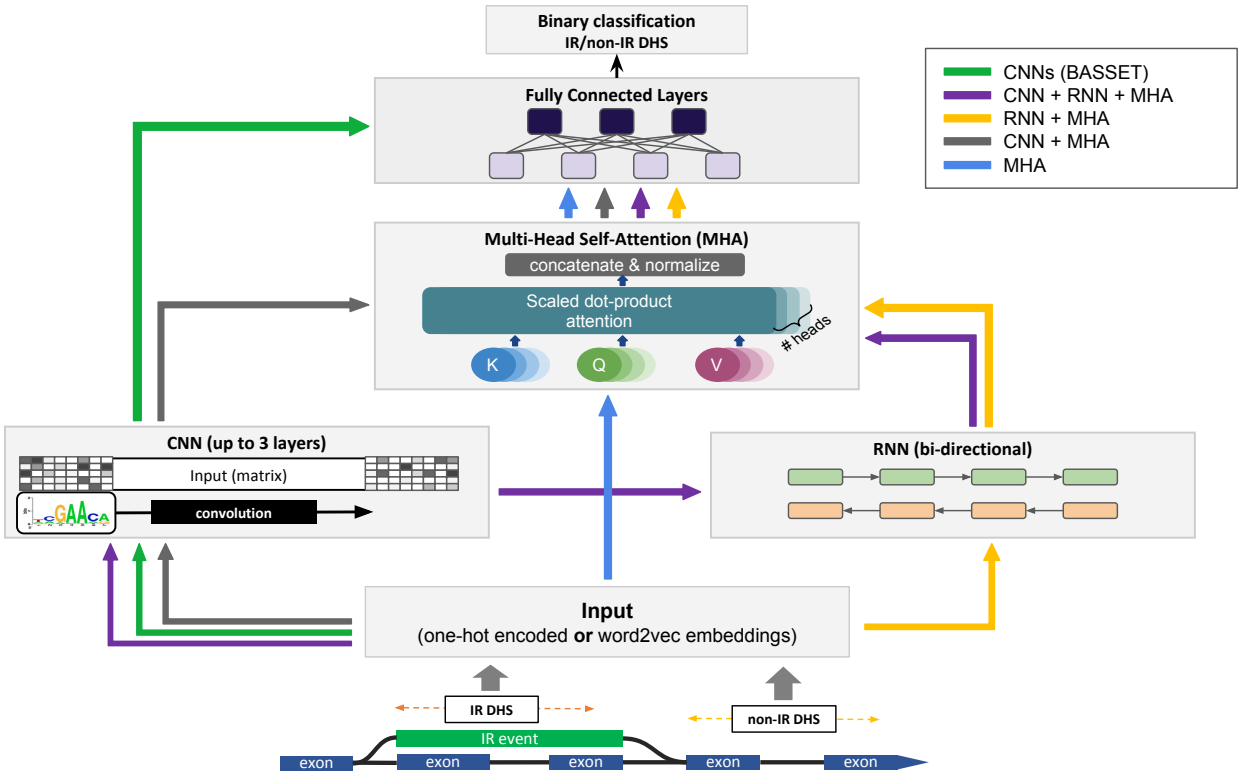


Figure 6.2: Summary of the different model variants explored when predicting DHS occupancy in IR events. Every architecture is represented by the corresponding colored arrows connecting different network components. The output represents a binary class prediction: IR vs. non-IR DHSs.

6.2.2 Network architecture

We investigate several network architectures to predict chromatin accessibility in IR events with the goal of understanding its chromatin-mediated regulation. The primary network element, a one-dimensional convolutional layer, scans a set of filters against the matrix representing the input sequence. As shown in equation 3.1, the number of input channels, B in our model are: 4 for DNA one-hot encoding input, the size of embedding, d in case of word2vec input, and *number*

of previous layer filters in case of higher convolutional layers. In the first layer, the filters are equivalent to PWMs or sequence motifs.

The output of a convolutional layer is produced by applying a non-linear activation function to the result of the convolution operation. In this work we use the Rectified Linear Unit (ReLU). This activation function addresses the problem of diminishing gradients in case of deeper networks, and has proven to be effective in genomics data [13]. Next, the output size is reduced by max-pooling where the maximum value in a window of a pre-determined size is selected. This reduces the input size for the next layer and also achieves invariance to small shifts in the input sequence.

Another feature that we explore in our model is recurrent layers. RNNs have an internal state that enables them to capture distant feature interactions in the input sequence. Specifically, we employ a bi-directional RNN with Long Short-Term Memory (LSTM) units [109]. In a bi-directional RNN, a forward and a backward layer are used that traverse the input in both directions, improving the model's performance.

We also incorporate a multi-head self-attention layer in our deep learning model. Attention is a powerful feature in that it can model dependencies within the input sequence regardless of their distances [12]. By doing so, it guides the network to focus on relevant features within the input and ignore irrelevant information. In case of multi-head self-attention, we concatenate the output of the H single-heads followed by a linear transformation. The final output is then collapsed along the hidden (attention) layer dimensions through addition and normalized by the mean and the standard deviation of the result. Empirically we find that this step is not only computationally efficient but also leads to better model accuracy in comparison to just flattening the attention layer output. The output of the attention layer is fed to one or more fully connected layers to generate the output of the network. In most architectures we employ a single fully connected layer; for Basset-like networks [13] we use three fully connected layers.

6.2.3 Network training and evaluation

As mentioned in the previous section, we explore several network variants with different layers and features. These architectures are summarized in figure 6.2. We tune the network hyperparameters using a semi-randomized grid search algorithm that employs a 5-fold cross validation strategy. In case of the Basset like model variant, we start with the hyperparaemters reported in [13] and fine-tune their values. This is because the problem at hand is similar to the one addressed using the Basset method [13]—predicting DHS occupancy from DNA sequence. The optimized hyperparameters are summarized in Table B.1 in Appendix B. We used two different schemes to train/test our model: 10-fold cross validation, and leave-one-chromosome-out cross validation. For the motif extraction analysis described later in this section, we used random train, test, and validation set splits with 80%, 10%, and 10% of the total data, respectively. To assess model performance, we use the area under the ROC curve (AUC) and the area under the Precision-Recall curve (AUPRC).

6.2.4 Gapped kmer SVM

We use the large-scale gapped kmer SVM (gkm-SVM), called the LS-GKM [177]. This version can handle bigger datasets (50k-100k examples) and exhibits better scalability. The LS-GKM is employed using both 10-fold and leave-one-chromosome-out cross validation strategies. We run the package with the following parameters: $-m$ 20000, $-x$ 10, and $-T$ 16 which specify the size of the memory cache in MB, number of cross validation folds (in case of 10-fold CV), and number of processing threads, respectively.

6.2.5 Motif extraction and analysis

To interpret the CNN based deep learning model, we extract sequence motifs using the weights (filters) of the first convolutional layer, similar to the methodology described in [13]. We select the positive examples (DHSs overlapping IR events) with the model prediction probability greater than 0.65. This cutoff is chosen as a best trade-off between the number of qualified examples and confidence in the prediction. For the negative examples, we used a cutoff value of less than 0.35.

Next, for each filter we identify regions in the set of sequences that activated the filter with a value greater than half of the filter's maximum score over all sequences. The highest scoring regions (sequence substrings) from all the sequences are stacked and for each filter, a position weight matrix is calculated using the nucleotide frequency and background information. We generate the sequence logos using the WebLogo tool [164]. The resulting PWMs are searched against the human CIS-BP database [178] using the TomTom tool [179] with distance metric set to euclidean.

6.2.6 TF ChIP-Seq analysis

We download the ChIP peaks of all the transcription factors that are enriched in IR events from the ENCODE database [14]. Next, we use our previously published pipeline [52] to test the enrichment of a given TF ChIP peaks in IR events. Briefly, we quantify the overlap of ChIP peaks with IR events and compare them to the overlap with non-IR events. The significance of overlap is tested using the Fisher exact test. To generate the profiles of TF occupancy across IR and non-IR events, we use the region of the ChIP peak where the PWM of the corresponding transcription factor has the highest score. This PWM scoring analysis is done using Biopython [180].

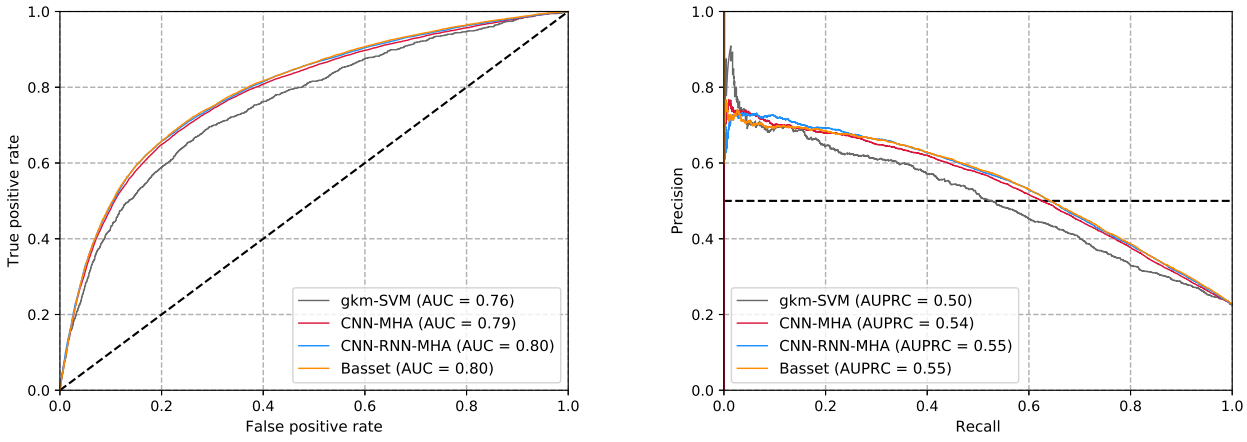
6.3 Results

6.3.1 Predicting DHSs associated with IR

We used the models described in figure 6.2 to distinguish DHSs associated with IR from non-IR DHSs. To assess the performance of each model variant, we used both 10-fold and leave-one-chromosome-out cross validation strategies. Figure 6.3 summarizes the results for several model architectures in the form of ROC and Precision-Recall curves. As expected, the deep learning based models exhibited improvement over the gkm-SVM in terms of AUC and AUPRC values. The architecture variants involving a multi-head self-attention layer exhibit accuracy on par with the Basset model. However, it should be noted that we fine-tuned Basset [13] hyperarameters as the default settings lead to results similar to that of the gkm-SVM. The reason for that is, unlike

the problem at hand, Basset was designed to predict DHS occupancy in 164 human cell-types in a multi-class, multi-label classification setting.

The results shown in figure 6.3 are generated using the different model variants with one-hot encoded input. Nevertheless, we also used word2vec embeddings and reported a measurable boost in accuracy, as shown in Figure B.2 (Appendix B). This is particularly evident in the case of the Basset-like model when used with one-hot encoded input (Basset) vs. low-dimensional word2vec embeddings (Basset-E). It follows that using word2vec embeddings improves overall model performance but at the cost of lower network interpretability; this is discussed at length in the next section.



(a) (b)

Figure 6.3: ROC and Precision-Recall curves are shown for the different deep learning architectures as well as the gkm-SVM in (a) and (b) respectively. The median AUC and AUPRC values are also provided in the legends. These results were generated using a 10 fold cross validation strategy.

6.3.2 Embeddings lead to poor interpretability

As mentioned in the previous section, the deep learning based models predicted DHS occupancy in IR and non-IR events with measurably higher accuracy than the gkm-SVM. However, it is not trivial to interpret the rules that the neural networks are learning. One way to extract information from the model is by examining its parameters, modulating the flow of information through

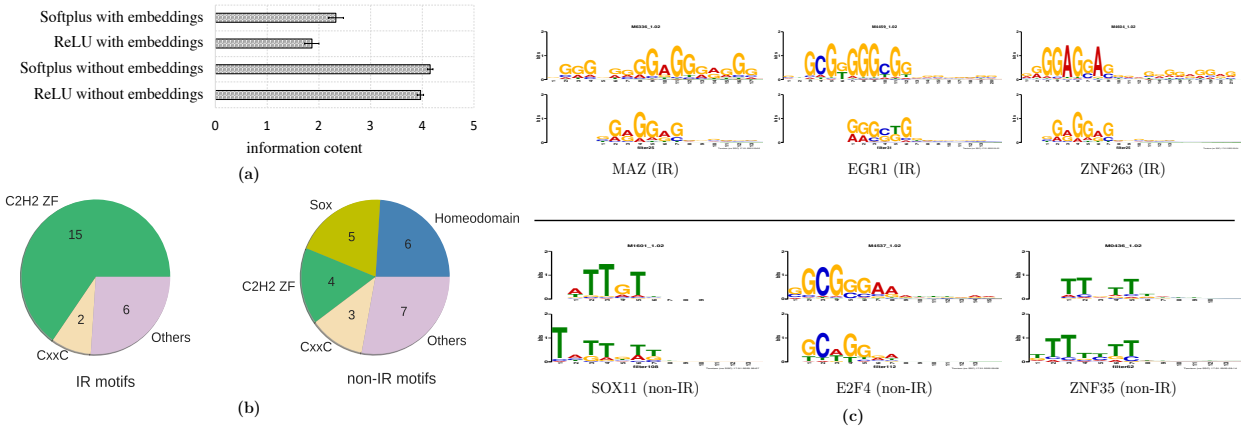


Figure 6.4: In (a), the mean information content is summarized for different cases: whether we use word2vec embeddings and exponential activations in the first convolutional layer. The distribution of TF families enriched in IR vs non-IR events are summarized in (b). Finally, the top 3 matches (based on the adjusted p-value) for the IR and non-IR convolutional layer filters against the CISBP database are shown in (c). In each match, the target transcription factor motif in the database is shown in the top row whereas the bottom row shows the actual CNN filter/motif.

layers of the network and analyzing its prediction on a specific set of sequences. These sequences can be specific dataset examples—for instance, DHSs overlapping IR events—that are predicted by the deep CNN with higher confidence. To achieve this, we used top positive (and negative) predictions of our model and implemented the strategy described in [98] and [13] (see Methods section for more details).

We chose the *CNN-MHA* model variant (see Figure 6.2) to do the motif analysis. This is because we need the convolutional layer to infer regulatory motifs. Besides, this architecture is not as complex as the other variants, yet exhibits comparable accuracy. We used this model with both one-hot and word2vec representations of the input sequences. Interestingly, the average information content (IF) of enriched motifs [181] significantly varied with the two input representations. When using the regular one-hot encoding, we find the motifs to be more informative and useful (mean IF = 4.0). The same is not true for word2vec embeddings where we get motifs with far lower information content (mean IF = 1.8). Recently, it was reported that, in contrast to ReLU activation, exponential functions in the first convolutional layer lead to more informative motifs [182]. By using Softplus as the activation in the convolutional layer, we report a slight improvement in information content when using the one-hot encoded input. Nevertheless, when we use the word2vec

input representation, there is a measurable improvement in average motif information content, as shown in Figure 6.4(a). Note that we did try different exponential functions but the Softplus activation gives the highest average information content. These findings are summarized in figure 6.4(a).

6.3.3 The zinc finger transcription factor family is enriched in IR events

We analyzed the motifs that were derived from the CNN filters for both top positive and top negative examples (see Methods). Next, we searched both sets of motifs against the Human CIS-BP transcription factor database [178] using the TomTom tool [179]. In case of IR DHSs, 23 motifs have significant hits against multiple known human TFs (q-value < 0.01). In comparison, 25 of the non-IR motifs have significant matches. Figure 6.4(c) shows the top hits reported for both IR and non-IR motifs. In the figure, a match is represented by the gold-standard CISBP human TF motif at the top, and the CNN filter motif at the bottom. We also observe that most of the IR motifs have significant hits in the Zinc Finger (ZF C2H2) super-family of transcription factors whereas the non-IR motifs are predominantly matched to the Homeodomain and Sox families of transcription factors. The distribution of the top four most frequent families in IR vs non-IR events are shown in figure 6.4(b). C2H2 ZF is the largest family of transcription factors and is highly active in the promoter, intragenic, and intragenic regions of the human genome (see figure 6.1). However, it is highly significant to observe that it is far more enriched in IR events compared to non-IR events. Zinc finger transcription factors have previously been implicated in the regulation of alternative splicing [148], particularly exon skipping. Here we report a role of this family in the regulation of intron retention.

6.3.4 Evidence from Chip-Seq data

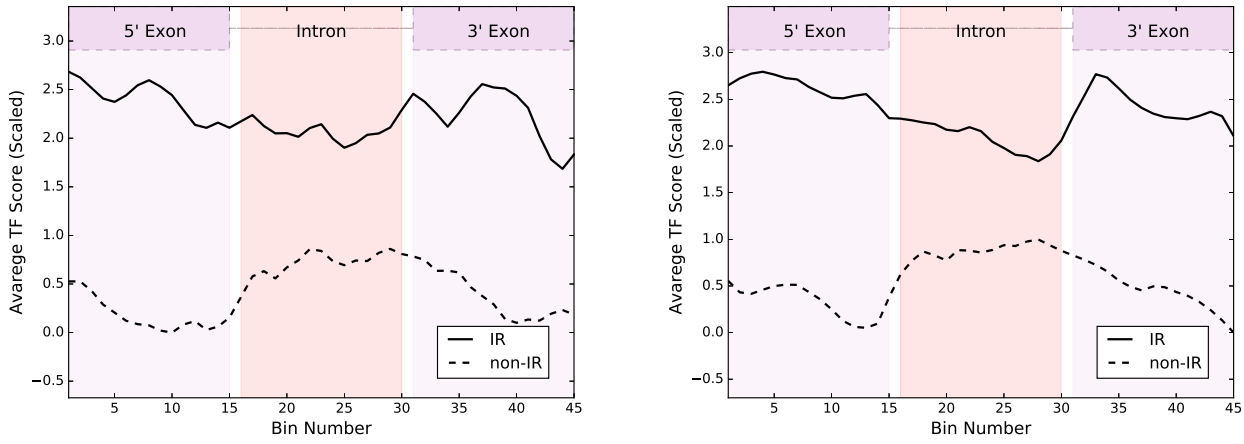
To validate our findings using experimental data, we downloaded available ChIP-Seq peaks of all zinc finger transcription factors in the human K562 cell line from the ENCODE database [14]. To test their enrichment in IR vs. non-IR events, we followed a strategy similar to [52]: For each transcription factor, we measured the overlap of its ChIP-Seq peaks with IR and non-IR events and tested its significance using the Fisher-exact test. All five TFs demonstrated highly significant

Table 6.1: Enrichment of C2H2 ZF transcription factors binding in IR vs non-IR events quantified using ChIP-Seq peaks of the corresponding TF.

TF	IR TF occupancy (%)	non-IR TF occupancy (%)	p-value
MAZ	14.85	7.66	3.58E-65
EGR1	14.52	8.22	4.11E-60
ZNF263	1.31	0.7	1.68E-06
SP1	4.43	2.21	6.98E-21
SP2	2.04	1.13	2.12E-08

enrichment in IR events (see Table 6.1), validating our in-*silico* findings that the C2H2 ZF family plays a role in the regulation of IR.

As table 6.1 suggests, in terms of the overlap, MAZ and EGR1 are far more frequent in IR events. Therefore, we picked those TFs and generated their binding affinity profiles across IR and non-IR events. To do that, we scored the ChIP-Seq peaks with the PWM of the corresponding transcription factor and picked the location with the highest score. Next, that particular location of the peak was used to determine where exactly it overlapped the IR/non-IR event. The binding affinity profiles are depicted in figure 6.5 for both EGR1 and MAZ, normalized by the over-all CHIP-peak occupancy in IR and non-IR events, respectively.



(a) EGR1 (b) MAZ

Figure 6.5: TF occupancy profiles across IR and non-IR events are shown for two transcription factors, (a) EGR1 and (b) MAZ. To generate the profile, PWM of the corresponding transcription factor was used to score the actual ChIP-Seq peaks (their DNA sequences). The regions with the highest score were then used to determine the TF occupancy within the events.

Interestingly, we observed that for both transcription factors, the binding affinity is stronger in the flanking exons of the retained introns. In contrast, in non-IR events, both EGR1 and MAZ are preferentially bound in the intronic regions. These unique TF-occupancy profiles suggest a role of zinc finger TFs in regulating intron retention, and AS in general.

6.3.5 Experimental validation

To experimentally validate the influence of C2H2 ZF transcription factors in regulating intron retention in a specific human cell-line, we picked human MAZ TF and a candidate IR event: intron 1 of the SCAND1 gene. The candidate event were picked based on the following criteria:

- The prediction score assigned by our deep learning model: we selected the top ranked examples in the test set.
- The experimental evidence of retention of the candidate introns from the RNA-Seq data in the corresponding cell-line (K562 in this case).
- The evidence for an overlap of MAZ ChIP-Seq peaks with the candidate introns (or their flanking exons) in K562 cell-line.

Our goal was to measure the effect of silencing MAZ–a human zinc finger TF–on splicing of the aforementioned introns. The results are summarized in figure 6.6.

The wet-lab experimentation was performed by our collaborator, Maayan Salton, at The Hebrew University of Jerusalem. In summary, short interfering—or, also referred to as silencing—RNA (siRNA) were used to silence the target transcription factor, represented as siMAZ in figure 6.6. As a control, the siRNA silenced Green Fluorescent Protein (siGFP) was used and the relative expression of the intron was measured. As shown in the figure, the level of retention inversely correlates with the silencing of MAZ gene: with knocked-down MAZ, the retention level drops as shown in figure 6.6. Note that for the given IR event, the difference in relative expression is significant (t-test p-value < 0.05).

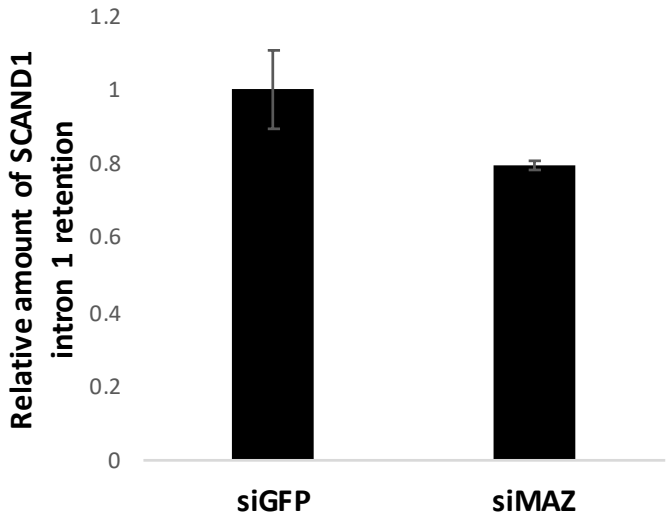


Figure 6.6: Evidence of MAZ, a C2H2 ZF transcription factor, regulating intron retention in the human K562 cell line.

6.4 Discussion

We explore several deep learning architectures to accurately predict chromatin accessibility in IR events. Because it solves a similar problem, we first redesign the Basset method [13] to discriminate IR overlapping DHSs from those that occurred elsewhere within genes. Note that in our problem formulation, the out of the box Basset doesn't perform so well; the accuracy in terms of AUC/AUPRC scores is similar to that of a gapped kmer SVM (baseline). Therefore, we fine-tune the default parameters of the model which leads to a significant improvement in accuracy, as depicted in figure 6.3. Besides Basset, in other architecture variants, we explore multi-head self-attention alongside convolutional and recurrent layers. One of the strengths of self-attention is that it can guide the model to focus on relevant information by quantifying inter-feature dependencies. Essentially, for every input feature, self-attention determines how much it is influenced by the rest of the features. That way, the model focuses on the relevant and important parts of the input and discards irrelevant information. Similarly, a recurrent layer can help capture long-term dependencies within the input sequence. Overall, the self-attention and recurrent layers improves model performance in terms of AUC/AUPRC values.

In addition to network features and layers, we also explore input transformation using word2vec embeddings [175]. The embedding representation is superior to a simple one-hot encoding in that it learns statistical information of k-mer co-occurrence relationships in the input. We find that, in contrast to one-hot encoding, these embeddings improve model performance in every architecture variant (see Figure B.1 in Appendix B). However, this transformation comes with an inherent drawback: poor network interpretability. It is worth mentioning that for the problem at hand, accuracy is undeniably helpful, nevertheless, model interpretability is of paramount importance. Our goal is to understand how chromatin accessibility can help us elucidate the regulation of intron retention. To this end, we convert the weights/filters of the convolutional layer to potential binding site motifs by using their activation within the input sequences. Unfortunately, in case of embeddings, the average information content is significantly lower in contrast to the regular one-hot encoded input (see figure 6.4(a)). Therefore, for all downstream motif analysis, we use the model with the input represented as a one-hot encoded matrix.

In the motif analysis, we found that the zinc finger (C2H2 ZF) family of transcription factors has a strong association with IR events: More than 50% of all motifs associated with IR have significant hits to C2H2 ZF transcription factors. This is consistent with previous work reporting that zinc finger transcription factors influence exon skipping [148]. Overall, this suggests that the C2H2 ZF family plays an important role in the regulation of alternative splicing in general. Non-IR events on the other hand exhibit enrichment of the Homoedomain and Sox families (see Figure 6.4(b)).

To validate our predictions on the association of these TFs with IR, we used experimental ChIP-Seq data for multiple zinc finger transcription factors: EGR1, MAZ, SP1, SP2, and ZNF263. We observe much higher occupancy of these transcription factors in IR events in the K562 human cell line, validating the model's predictions. It is also useful to analyze where in an event these regulatory proteins preferentially bind. Therefore, we generate binding affinity profiles for MAZ and EGR1 across IR events. We observe that both transcription factors have stronger binding preference in the flanking exons of retained introns. In contrast, in non-IR events, MAZ and EGR1

exhibit higher affinity in the intronic region. This suggest a regulatory role of zinc finger TFs by preferentially binding at specific regions of IR and non-IR events. MAZ4 elements—an element that contains four copies of the MAZ protein binding sequence—have previously been reported to influence alternative splicing [183]. This might be true for EGR1 as well; both MAZ and EGR1 have a decent binding site sequence similarity (see figure 6.4(c)).

In the next chapter, we analyze the cooperativity and interactions among the IR-enriched motifs by employing a self-attention based deep learning model. We report numerous statistically significant TF interactions in IR events, multiple of which have previously been reported in the scientific literature. It follows that intron retention is regulated by a complex orchestration of transcription factor interactions. Further wet-lab experiments are needed to validate these findings and provide a solid foundation on how these proteins regulate intron retention, and AS in general.

Chapter 7

A Self-Attention Model for Inferring Regulatory Interactions

7.1 Introduction

In the previous chapter, we identified and validated the role of transcription factors—particularly the C2H2 zinc finger family—in the regulation of intron retention in human. In order o capture a comprehensive landscape of the regulation of alternative splicing, in this chapter, our goal is infer interactions between transcription factors by interpreting the values learnt by self-attention layer of our deep learning model.

The discovery that TFs work in tandem to regulate the expression of their targets [184] has sparked the development of a variety of computational methods for predicting cooperativity among TFs and other regulatory proteins by looking at regulatory element co-occurrences [185–191]. Despite the demonstrated ability of deep neural networks to extract regulatory signals directly from sequence, there are very few studies that explore cooperativity between regulatory features in genomic data using these methods. Deep Feature Interaction Maps (*DFIM*) uses a network attribution method called DeepLIFT [192] to estimate interactions between regulatory elements, tested for one pair at a time [193]. The major drawback of DFIM is that it is computationally expensive: the interactions are inferred in a separate post-processing step and involves recalculation of network gradients. We note that the recent *DeepResolve* method infers feature importance and whether a feature participates in interactions with other features, but does not infer pairs of interacting features explicitly [194].

Recently, neural networks that use the concepts of *attention* and *self-attention* [110, 111] have achieved remarkable success in natural language processing tasks, specifically in machine translation [12]. One of the strengths of attention is that it can capture associations between features

regardless of the distance between them, addressing a major shortcoming of convolutional and recurrent networks. This is particularly useful for tasks in computational biology where our goal is to identify regulatory elements and their associations/interactions in DNA or RNA sequences. The value of attention for modeling transcription factor binding site prediction was recently demonstrated, and their work was motivated by the greater interpretability of the resulting networks [112]. However, to the best of our knowledge, it has not been employed for inferring regulatory interactions between TFs and other regulatory elements. To this end, we propose **SATORI**, a Self-ATtentiOn based deep learning model to capture Regulatory element Interactions in genomic sequences. The primary components of the architecture of our model are a CNN layer and a multi-head self-attention layer. Optionally, we also incorporate an RNN layer between the two primary layers. The convolutional layer discovers features/motifs in the input sequences. The self-attention layer then captures potential interactions between those features without the need for explicitly testing all possible combinations of motifs. That enables us to infer a global landscape of interactions in a given genomic dataset, without a computationally-expensive post-processing step.

We test SATORI on several simulated and real datasets, including data on chromatin accessibility in 164 cell lines in all human promoters and genome-wide chromatin accessibility data across 36 samples in Arabidopsis. Moreover, pertinent to this work, we use SATORI to infer TF interactions involved in the regulation of intron retention. To compare our method to DFIM, we incorporate their Feature Interaction Scores (FIS) [193] into our framework. In all our experiments, SATORI and FIS scoring return highly consistent sets of interactions, with SATORI returning a much larger number of biologically confirmed interactions. Due to the relative paucity of experimentally determined TF-TF interactions it is important to have multiple independent methods for this task. We believe this work will assist researchers in improving the interpretability of complex deep learning methods and providing actionable hypotheses for follow up experiments.

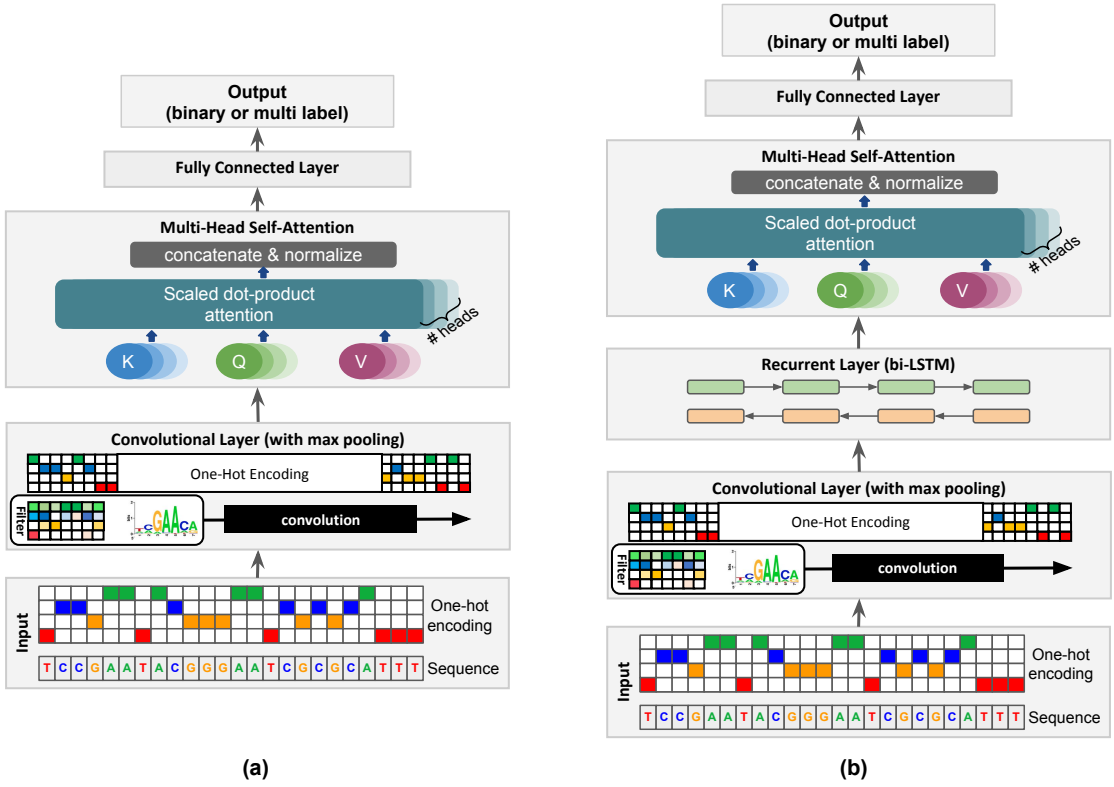


Figure 7.1: Model architecture variants. We use a convolutional layer followed by a multi-head self-attention layer (a); optionally, we add a recurrent layer between the two (b). The input in both cases is a one-hot encoding of the DNA sequence. The output of the model is either be a binary or multi-label prediction.

7.2 Methods

7.2.1 Model architecture

We present a self-attention based deep neural network to capture interactions between regulatory features in genomic sequences. Figure 7.1 depicts the two main architectures we explored in our work. Note that the input to the model—DNA/RNA sequences—was represented as a one-hot encoding: a sequence of length L is transformed into a matrix of size $4 \times L$ where each position in the sequence is represented by a column in the matrix as a non-zero value at location corresponding to one of the four DNA nucleotides.

The first component of our model is a CNN layer where a finite set of filters are scanned against the input sequence/matrix. For more information on the convolutional layer, please refer to

Chapter 3. Similar to the model architecture describe in the previous chapter, after the convolution operation, we use ReLU as the activation function, followed by a max-pooling operation.

Optionally, as shown in Figure 7.1(b), we use an RNN layer following the CNN layer. RNNs have an internal state that enables them to capture distant feature interactions in the input sequence. Specifically, we employ a bi-directional RNN with Long Short-Term Memory (LSTM) units [109].

The core component of our network is a multi-head self-attention layer. Attention mechanism can model dependencies within the input sequence regardless of their distance [12]. By doing so, it guides the network to focus on relevant features within the input and ignore irrelevant information. Pertinent to the problem at hand, a self-attention layer can help us identify interacting regions within the input sequences. Consequently, we can capture interactions between regulatory events. For more technical details on self-attention, refer to Chapter 3.

For multi-head self-attention, we concatenate the output of the N single-heads followed by a linear transformation. The final output is then collapsed along the hidden (attention) layer dimensions through addition and normalized by the mean and the standard deviation of the result. Empirically we find that this step is not only computationally efficient but also lead to better model accuracy in comparison to just flattening the attention layer output. The final fully connected readout layer outputs the model's prediction: either a binary or multi-label classification, depending on the experiment. For binary classification, we use the standard cross entropy loss function. For multi-label classification, we use the binary cross entropy with logits loss function:

$$L(y, \hat{y}) = -\frac{1}{C} \sum_{i=1}^{C} \left[y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i)) \right],$$

where y is the vector of ground truth labels, $\hat{y}$ are the network predictions, C is the number of classes, and σ is the sigmoid function.

Table 7.1: Summary of the datasets used in the four experiments we designed to test and analyze SATORI. The first two datasets have binary labels whereas the last two experiments deal with a multi-label, multi-calss problem.

Experiment	Dataset	Description	Size and Labels
1	Simulated	Simulated DNA sequences ELF1 & SIX5 embedded in pos. examples ELF1 or SIX5 embedded in neg. examples Random embeddings of AP3 and TAL1	120,000 (total) 40,000 (+), 80,000 (−)
2	TAL-GATA ChIP-Seq	DNA sequences for TAL1, GATA1, and GATA2 ChIP-peaks in K562. Positive examples were peaks that overlapped DHSs. Negative examples were all other DHSs.	105,134 (total) 25,134 (+), 80,000 (−)
3	Human Promoter open chromatin	DHSs overlapping the human promoters	20,613, across 164 cell types
4	Arabidopsis open chromatin	DHSs/THSs across arabidopsis genome	88,245, across 36 samples

7.2.2 Network training and evaluation

For model selection and optimization, we employ a random search based algorithm to tune the network's hyperparameters. For the convolutional layers we considered filter size, number of filters, and size of window over which pooling is performed. For the multi-head attention layer we tuned the dimensionality of the features generated, and the size of the output of the multi-head attention layer. Details are provided in Table C.1 (Appendix C). To evaluate the model, we use a simple strategy of splitting the data into 80%, 10%, and 10% for train, test, and validation sets, respectively. To assess the model's performance, Area Under the ROC Curve (AUC) is used. The hyperparameters for the two architectures (see figure 7.1) are summarized in Table C.1 in Appendix C. The package was implemented in PyTorch [195] and all the experiments were ran on a Ubuntu server with a 12 GB TITAN V GPU.

7.2.3 Motif extraction

To interpret the deep learning model, we extract sequence motifs from the weights matrices (filters) of the first convolutional layer, similarly to the methodology used in [13]. For binary classification problems, we use the positive test set examples that achieve a probability score greater than 0.70. This cutoff was chosen as a good trade-off between the number of qualifying examples and confidence in the prediction. We use all test set examples when dealing with a multi-class or multi-label problems. Next, for each filter we identify regions in the set of sequences that activate the filter with a value greater than half of the filter's maximum score over all sequences. The resulting substrings are stacked and for each filter, a PWM is calculated using the nucleotide frequency and background information. Sequence logos are generated using the WebLogo tool [164]. The PWMs are searched against appropriate TF databases using the TomTom tool [179] with distance metric set to Euclidean. For searching we use the human CISBP [178] and arabiodpsis DAP [196] databases. In the benchmark experiments, we use custom TF databases, details of which are provided in Appendix C.

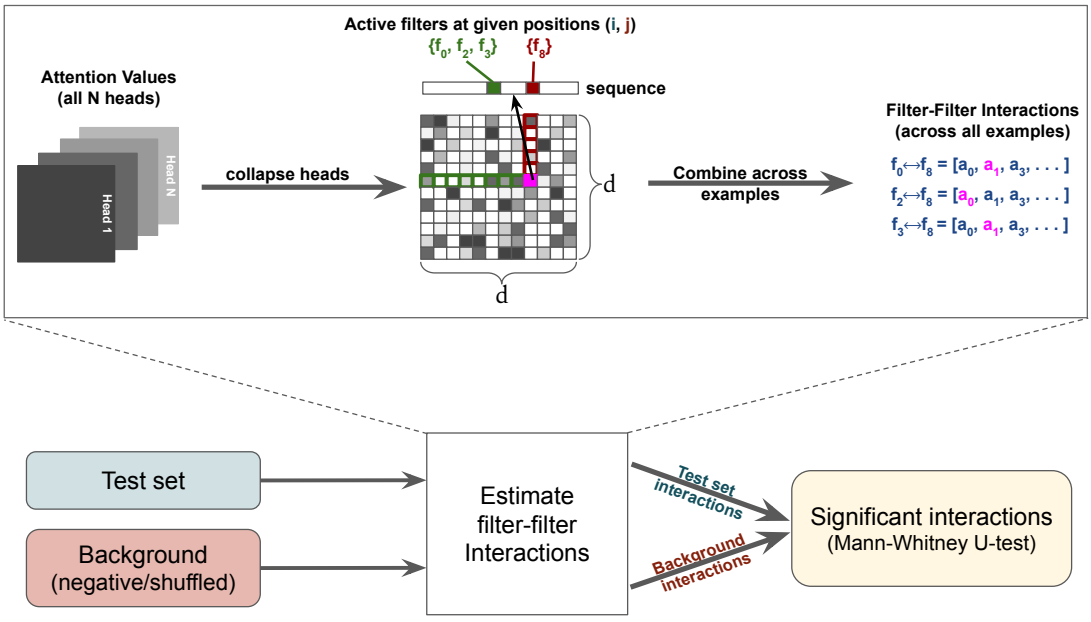


Figure 7.2: Summary of the process of inferring interactions from self-attention layer values. For a given example, we collapse the attention heads into a single matrix. Next, at each pair of positions, the corresponding active CNN filters are identified and the attention value is assigned to the interacting pair. This is repeated for all examples to generate interaction profiles for all filter-pairs. Finally, we use a background set to test the significance of filter-filter interactions.

Recently, it has been reported that exponential activation functions such as Softplus in the convolutional layer improve motif information content [182]. In our experiments this led to a slight improvement in the information content of the CNN filters when the Softplus function was used instead of ReLU activation (median information content = 4.12 with Softplus compared to 4.00 with ReLU).

7.2.4 Quantifying feature interactions

In this section we describe the process of inferring motif interactions from the self-attention layer. The attention matrix for each head is calculated using Equation (3.11). Next, we collapse the N heads to a single $d \times d$ matrix by taking the maximum at each position (see Figure 7.2). This step summarizes the self-attention values from multiple subspaces associated with the corresponding single-heads to a single, attention profile.

The attention matrix provides information about interactions between positions in the sequence. That is next converted into interactions between filters by retrieving the filters that are active in those positions. Finally, for each identified filter-filter interaction, we generate the attention profile across all testing examples. For a given pair, this profile consists of a vector of its attention values at positions where the corresponding filters were active. An interaction pair is discarded if its maximum attention value is below a certain threshold. By default we used the value 0.10; in the human promoter data we used 0.08 to increase sensitivity. We note that the distribution of attention values tends to be bi-modal, with most of the values close to 0 or 1 (see Figure C.1, Appendix C).

Filter-filter interactions are then translated to motif interactions by picking the most significant TomTom hits in the appropriate TF database. Note that we might not find significant matches for every CNN filter in the database; it can be expected that our model is capturing interactions of un-characterized regulatory elements. However, in this paper we focus on the interactions between known TFs. To test the statistical significance of motif interactions, we first generate their attention profiles in the background data (described next). Then the non-parametric Mann-Whitney U Test is used to calculate their significance. All the p-values are adjusted for multiple hypothesis testing using Benjamini-Hochberg method [165].

Background selection

As mentioned above, to test the statistical significance of regulatory interactions, we need to compare them to a background. We use a biologically relevant background depending on the experiment:

- For binary classification problems, the negative test set is used as the background.
- For multi-label, multi-class or regression problems we generate a background set by shuffling the test set sequences while preserving their di-nucleotide frequencies. Next, in the shuffled sequences, we randomly embed motifs that are generated based on the CNN filters, interpreted as probability distributions, taking into account the number of times a filter is active in the original test sequences.

Quantifying interactions using FIS scoring

To infer interactions between motifs using the FIS method, we closely follow the strategy described in [193]. Given a test sequence and all of its activated first layer CNN filters, first a source motif is selected. The remaining filters serve as the target motifs for the given source motif. Using Integrated Gradients [197], we calculate the attribution scores for all the target motifs in the given test sequence. The attribution score determines how important a motif is for the model to accurately predict the test example. Next, the source motif is mutated based on the GC-content of the given sequence and the attribution scores are recalculated for all targets. Finally, to infer interactions, for each source and target pair, the FIS score is calculated as the difference between the attribution scores for each target motif, before and after mutating the source motif. Intuitively, if modifying the source motif affects the attribution of the target motif, this suggests a potential interaction, and the magnitude of the change in attribution scores is used to quantify this potential. We compute FIS scores for all unique pairs of source and target motifs across all test sequences and identify statistically significant interactions using the the same approach used in SATORI (see Figure 7.2).

Selecting test examples

To quantify interactions using SATORI or the FIS-based approach, we use the high-confidence predictions of the model. For binary classification, we pick all positive examples that are assigned prediction confidence above a specified threshold. We use a threshold of $p = 0.70$ in our experiments. For the background examples, we pick all the negative test examples that score below $1 - p$. In case of the multi-label classification problem, we pick our test examples based on the precision of the model's prediction probabilities: for a test example to qualify, the precision value—calculated using the given labels and their model assigned probabilities—must be above a specified threshold (default precision threshold $= 0.50$). We note that for FIS scoring of multi-label classification problems, we only use the attribution values of the true positive predictions. These values are summed and used in calculating the final FIS score.

7.2.5 Data collection and processing

We use multiple datasets to test our model's ability to capture interactions between regulatory elements. The datasets are summarized in Table 7.1, and specific details are provided in the Methods section of the supplementary material. As mentioned earlier, we also used SATORI with the IR DHSs; more information on the processing of that dataset can be found in the Methods section of Chapter 6.

7.3 Results and Discussion

7.3.1 Benchmark 1: embedded motif interactions in simulated sequences

In this experiment we used SATORI to test if it can recover interactions embedded in simulated DNA sequences. This test served as a benchmark in order to compare our model to the recently published DFIM method [193]. We used a very similar approach to theirs when creating the simulated dataset: $120,000$ random DNA sequences were generated; in $40,000$ sequences we embedded motifs of the transcription factors SIX5 and ELF1. This simulated an interaction between the two TFs since we required both to be present in every sequence. We labelled these sequences as positive examples. In the remaining sequences that serve as negative examples, we embedded only one of the two motifs in each sequence. In addition, motifs of TAL1 and AP3 were embedded at random across the whole dataset.

Not surprisingly, both variants of our model achieved perfect classification accuracy on the test set for this data. We then analyzed the attention layer weights and inferred statistically significant motif interactions and found that all the significant interactions returned by our model involve SIX5 and ELF1 as expected. These interactions are summarized in Table C.2 (Appendix C).

7.3.2 Benchmark 2: Inferring TAL-GATA motif interactions from ChIP-Seq data

The TFs TAL1 and GATA1 have been reported to interact: GATA1 requires a prior or simultaneous binding of TAL1 before it can bind DNA [198]. To investigate these interactions in this

Table 7.2: The most frequent interacting families of human transcription factors in the TAL-GATA ChIP-peaks in human K562 cell-line. All interactions are significant with adjusted p-value < 0.05.

TF Family Interaction	Frequency	Percent of total interactions
C2H2 ZF $\longleftrightarrow$ CxxC	170	14.36%
Homeodomain $\longleftrightarrow$ C2H2 ZF	121	10.22%
C2H2 ZF $\longleftrightarrow$ C2H2 ZF	90	7.60%
GATA $\longleftrightarrow$ C2H2 ZF	79	6.67%
Homeodomain $\longleftrightarrow$ CxxC	72	6.08%
C2H2 ZF $\longleftrightarrow$ Sox	62	5.24%
C2H2 ZF $\longleftrightarrow$ bHLH	53	4.48%
GATA $\longleftrightarrow$ CxxC	51	4.31%
Sox $\longleftrightarrow$ CxxC	39	3.29%
C2H2 ZF $\longleftrightarrow$ THAP finger	21	1.77%
CxxC $\longleftrightarrow$ bHLH	19	1.60%
Nuclear receptor $\longleftrightarrow$ C2H2 ZF	19	1.60%
GATA $\longleftrightarrow$ bHLH	18	1.27%
Homeodomain $\longleftrightarrow$ bHLH	17	1.44%
SAND $\longleftrightarrow$ C2H2 ZF	17	1.44%

experiment, which follows a similar experiment performed by the authors of DFIM, we formulated a binary classification problem where the positive set consisted of sequences of the TAL1, GATA1, and GATA2 ChIP-Seq peaks that overlapped regions of open chromatin (DHSs) in the human K562 cell-line. For the negative set, sequences of all other K562 DHSs that didn't overlap any of the ChIP-Seq peaks were used. This experiment serves as another benchmark for our model, and was also used by Greenside et al. to test their model's ability uncover interactions between TAL1 and GATA1/GATA2 [193]. Further details on the dataset are provided in Appendix C.

We trained both variants of our model on this dataset; in this harder dataset the variant with an RNN layer performed much better with an AUC of 0.94 on the test set compared to 0.85 for the model without an RNN. The authors of DFIM achieved similar accuracy using five layers of convolution. Please note that we do not seek to demonstrate better model accuracy. Our focus is on model interpretability, which we seek to achieve without compromising in that regard. We recovered multiple significant filter-filter interactions that mapped to TAL1 and GATA motifs with highly significant p-values (see Table C.3 in Appendix C), demonstrating the ability of our model to recover in-*vivo* interactions between TFs.

Since the ChIP-Seq peaks of TAL and GATA transcription factors are occurring in regions of open chromatin, it is highly likely that in those regions, other regulatory elements/TFs are active and interacting with each other. Therefore, we also let SATORI search for interactions among all known human transcription factors and found numerous other interactions. Table 7.2 summarizes the 15 most frequent TF family-family interactions, consisting predominantly of interactions between members of the C2H2 ZF, Homeodomain, CxxC, and GATA families (see Table C.7 in Appendix C for the list of individual TF interactions). An interesting observation here is that the interactions between the GATA and bHLH (TAL1) families are not the most frequent, despite the fact that the model used their ChIP-Seq peaks. This is likely because of the differences in size of these TF families: the C2H2 ZF and Homeodomain families are the largest TF families in humans, whereas bHLH and particularly GATA, are much smaller in size.

Because of the similarity between some TF motifs, the matching between filters and motifs is not without errors. In this second experiment GATA was predicted to interact with TCF15, which has a motif that closely resembles that of TAL1. In fact, both of them belong to the same bHLH family. Figure C.2 in Appendix C shows the similarity between these motifs.

7.3.3 The TF interaction landscape across human promoters

In this experiment we investigated regulatory interactions between TFs in all human promoter regions using DNase I hypersentivity data (DHSs) across 164 immortalized cell lines and tissues. This experiment was based on Kelley et al's work where they predicted chromatin accessibility from sequence information alone across the entire human genome [13]. The labels in this data represent presence/absence of a given DHS across each of the 164 cell lines, and is a multi-label classification problem. We trained both network variants (see figure 7.1) and observed that with the optional RNN layer, the network performed better in terms AUC scores. In Appendix C, Figure C.3 compares the accuracy for each of the 164 cell lines for both variants of the architecture.

The trained network yielded filters that matched 93 TFs with known motifs (counting only filters that had information content greater than 3.0). Among those 93 TFs, our model identified

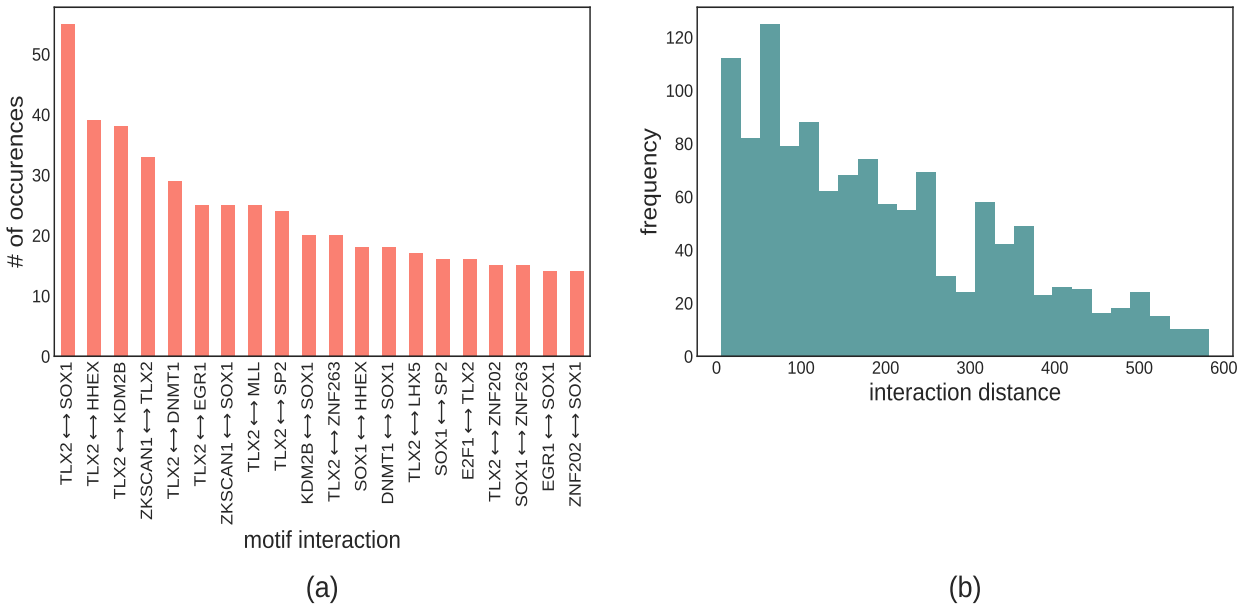


Figure 7.3: The most frequent TF interactions in human promoters (a). The distribution of TF-TF interaction distances (b).

234 pairs of motifs that interact, with a total of 1250 statistically significant interactions. The 20 most frequent interactions are shown in Figure 7.3(a). For the complete list, refer to Table C.8 (Appendix C). We also looked at the distribution of the distances between interacting motifs, and observed that, as expected, interactions tend to occur in close proximity with a median distance of interaction of 168 bp (see Figure 7.3(b)). Overall, the Homeodomain, C2H2 ZF, and CxxC families were the most frequent families of interacting TFs (see Figure C.5 in the appendix). Finally, it is worth mentioning that for twelve interactions out of the total of 234, we found evidence in the TRRUST database [3] which lists only 58 interactions among the 93 TFs (see Table C.4 in Appendix C for details). This overlap is statistically significant with a p-value of 4.66×10^{-7} using the hypergeometric test.

We note that in this work we only analyzed interactions between motifs of known TFs. Not all filters can be mapped to characterized regulatory proteins. In this dataset, TomTom returned no significant matches for 80 out of the CNN filters with motif information content greater than 3.0. The interactions of such filters require further investigation to discover the regulatory molecules associated with them.

As mentioned above, the motif matching results returned by TomTom are noisy and imperfect. For example, some of the statistically significant matches are clearly incorrect, as shown in Figure C.4 (Appendix C). However, this is not a shortcoming of our model, but rather a limitation of the interpretation of filter-filter interactions.

7.3.4 Genome-wide regulatory interactions in arabidopsis

We extended our analysis to plants by designing a similar experiment as described in the previous section. More specifically, we predict chromatin accessibility using sequence information alone across 36 arabidopsis samples from recently published arabidopsis DNase I-Seq and ATAC-Seq studies (GEO accession numbers provided in the Supplementary Methods section in Appendix C). Like the previous dataset, this too is a multi-label prediction problem, where the labels indicates whether a given region has a peak in each of the 36 samples of DNase I-Seq and ATAC-Seq. We trained both variants of our model and similarly to the other datasets, the network that included an RNN layer performed slightly better in terms of median AUC across samples (0.86 compared to 0.85).

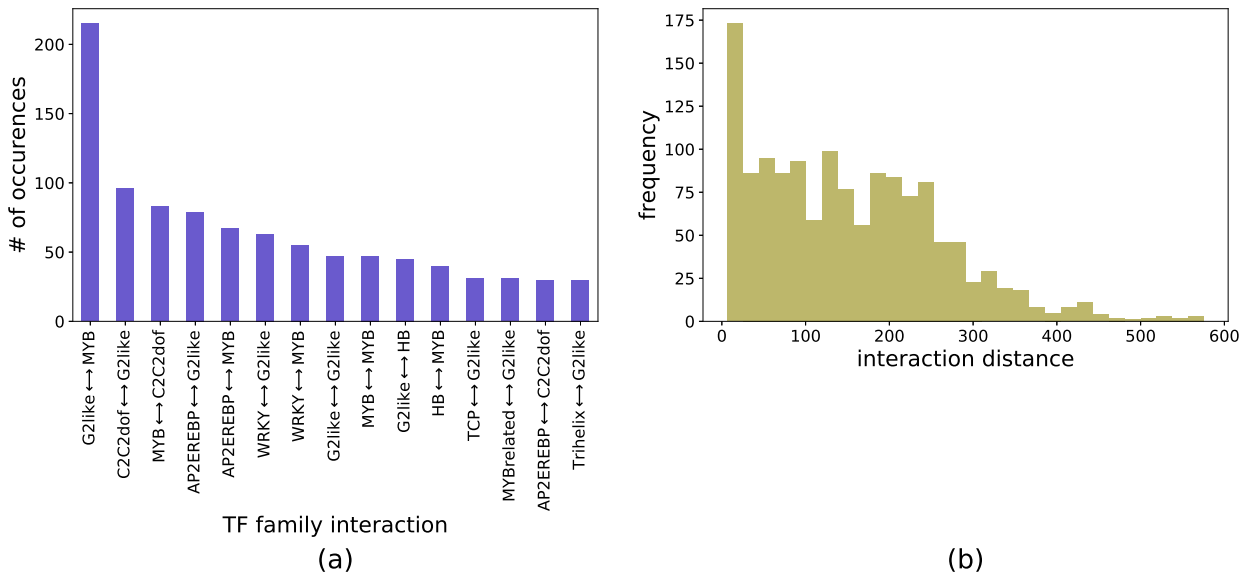


Figure 7.4: The regulatory interaction landscape in accessible chromatin in the arabipdosis genome. The most frequently interacting families of plant transcription factors (a). The distribution of distances between inferred TF-TF interactions (b).

In the next step, we investigated genome-wide regulatory interactions in those regions of open-chromatin. The trained network yielded 189 filters with information content above 3.0, and we obtained 100 unique matches for those filters in the DAP-Seq arabidopsis TF database [196]. Among these 100 TFs, our model identified interactions between 230 pairs of TFs with a total of approximately 1400 statistically significant interactions involving diverse plant transcription factors (see Figure 7.4(a)). G2like, MYB, C2C2dof, and AP2 were the most frequently represented TF families in those interactions. Similarly to our findings in human, plant TF interactions tend to occur in relative proximity (median distance = 138 bp) as shown in Figure 7.4(b). Arabidopsis does not have a database of known interactions between TFs, so our results could not be validated. Targeted experimental validation of these predictions can thus significantly enrich our knowledge of the combinatorial regulation of gene expression in plants.

7.3.5 Comparison: SATORI and FIS-based interactions

To compare our model to DFIM [193], we incorporated its FIS scoring method as a feature in our framework and tested it on the three real-world datasets. A key observation is that among the top scoring interactions predicted by the two methods there is very high overlap: In the TAL-GATA dataset the top 15 interactions predicted by FIS scoring were also predicted by SATORI; for the human promoter dataset 14 out of the top 15 FIS predictions were detected by SATORI; finally, for the arabidopsis genome-wide dataset, nine out of the top FIS predictions were predicted by SATORI. At the TF family level, we observed perfect agreement in the top ten predictions. These results are summarized in Figure 7.5. The agreement on the top predictions suggests their high likelihood of being biologically relevant, and make them promising candidates for experimental validation.

Among the three real-world datasets the lowest agreement was observed for the arabidopsis dataset. This is likely due to its complexity, as it probes interactions on a genome-wide scale across all accessible regions of the arabidopsis genome. Nevertheless, there is a high level of agreement at the level of TF families. We also compared the computation times for the two meth-

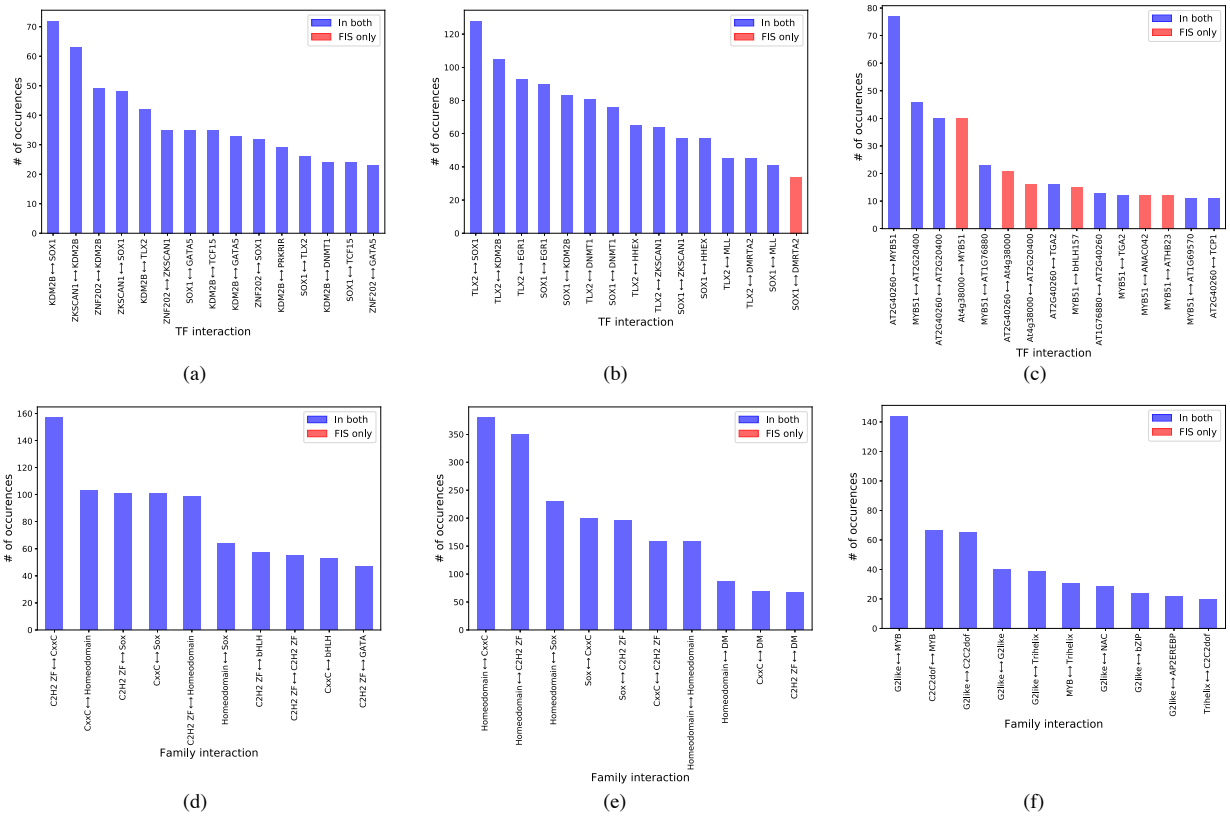


Figure 7.5: Common interactions in the top predictions of SATORI and FIS. Interactions predicted by FIS are sorted by frequency. Those predicted by both methods are shown in blue, and ones predicted only by FIS are shown in red. Top predictions are shown for the TAL-GATA dataset (a) the human promoter dataset (b), and the genomewide arabidopsis dataset (c). For each experiment, the 10 most frequent TF family interactions are shown in (d), (e), and (f) respectively.

ods. As discussed earlier, unlike the FIS method, SATORI does not require re-calculation of the gradients to estimate the interactions, leading to much faster computation times: it processed all motif interactions 8 to 20 times faster than FIS (see Figure 7.6).

In our experiments SATORI reported more interactions for the human and arabidopsis chromatin accessibility datasets, while the FIS method identified more motif interactions for the TAL-GATA experiment (see Table C.5 in Appendix C). For the human promoter dataset, we searched the interactions identified by the two methods in the TRRUSTv2 database. Interestingly, for SATORI we found matches for 12 TF-TF interactions; in comparison only a single FIS interaction was found in the TRRUSTv2 database. Due to the small number of experimentally verified interactions, more extensive wet-lab validation is needed to test the quality of the reported interactions by

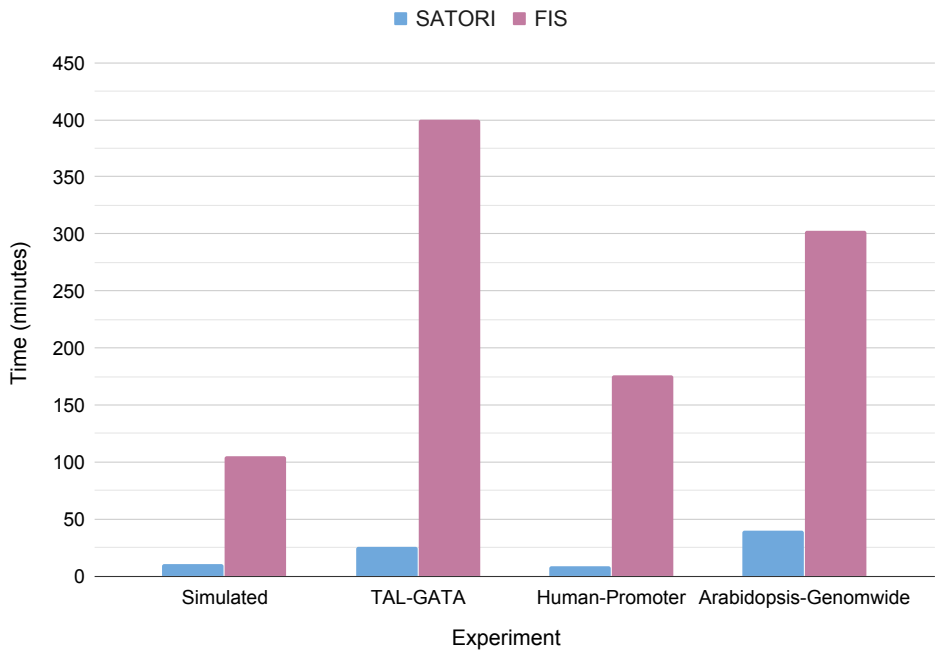
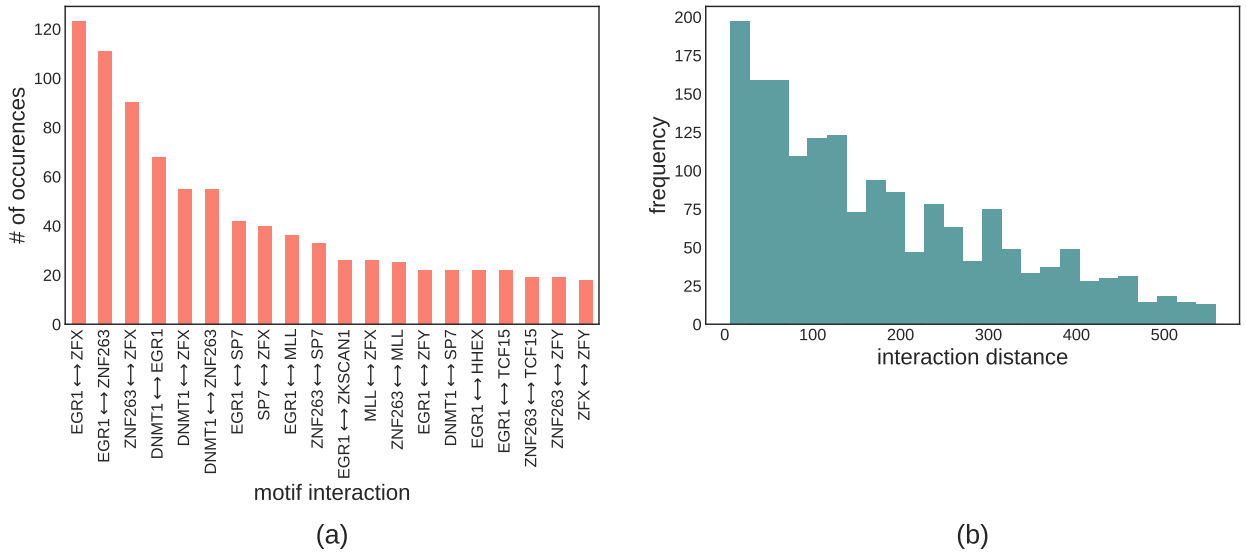


Figure 7.6: Run time in minutes for SATORI and FIS-based interaction estimation for the four datasets.

the two methods. High-frequency interactions consistently detected by both methods can be used as the most promising candidates for experimental follow-up.

7.3.6 Regulatory interactions in IR events in human

After thoroughly testing our method on multiple datasets, we used SATORI to predict interactions involved in the regulation of intron retention. We were able to identify 241 unique interactions in the DHSs occurring in IR events. The top 20 most frequent interactions are summarized in figure 7.7(a). The complete list of significant TF interactions can be found in Table C.9 in the appendix. As expected, most of the interacting transcription factors belong to the C2H2 ZF family. We also observe interactions between C2H2 ZF and CxxC, Nuclear receptor, and bHLH families of transcription factors. In Appendix C, Figure C.6 shows the most frequently interacting TF families. Moreover, we looked at the average interaction distance and found out that TF motifs preferentially interact in proximity. This is evident from the distribution of interaction distances in figure 7.7(b); the median interaction distance was 164bp.



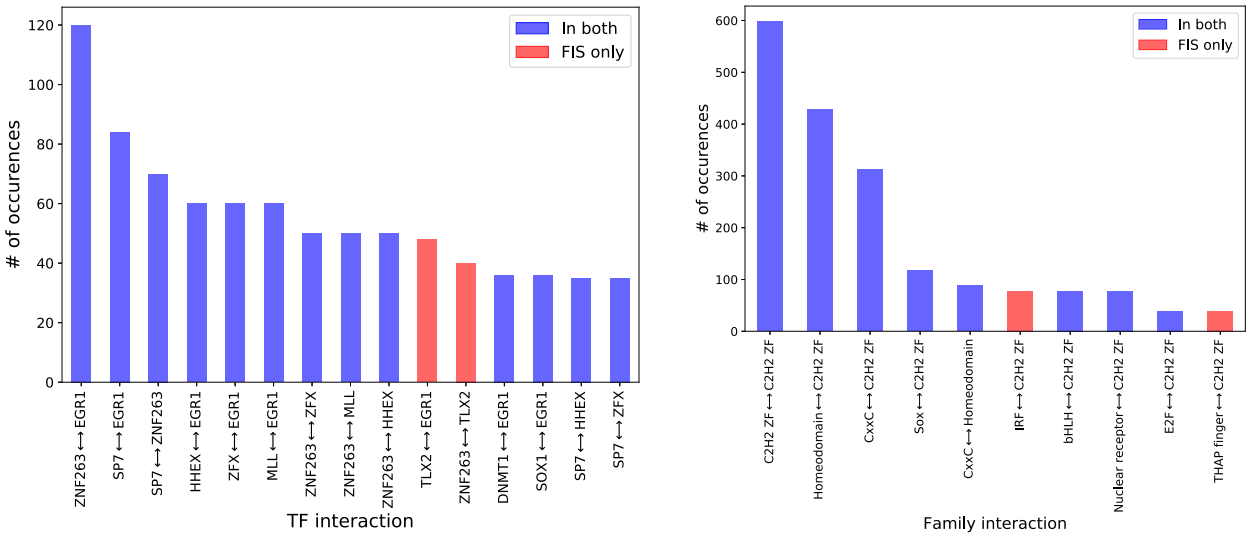
(a) (b)

Figure 7.7: The most frequent transcription factor interactions in intron retention events are depicted in (a). A majority of these interactions involve C2H2 ZF family. In (b), the distribution of distances is shown for all the statistically significant interactions.

Besides SATORI, we used FIS based scoring method and identified 253 unique interactions in the DHSs occurring in IR events. Interestingly, out of the 15 most frequent interactions, 13 were also predicted by SATORI, as shown in figure 7.8(a). Finally, we searched the significant interactions in TRRUSTv2 [3], a database that annotates TF regulatory roles and their interactions by text-mining previously published literature. As summarized in Table C.6 (Appendix C), 14 of the TF interactions predicted by SATORI were annotated in the database. In contrast, only 7 FIS interactions were found in the TRRUSTv2 database, indicating higher accuracy of SATORI in identifying relevant TF-TF interactions. As previously mentioned, for experimental follow-up, high-frequency interactions that are consistently detected by both methods can be used as the most promising candidates.

7.4 Conclusions and Future Work

In this work we presented SATORI — a method for extracting interactions between the learned features of an attention-based deep learning model. Unlike existing methods, it does not require any post-processing and uses the sparsity of the attention matrix to infer the most salient interac-



(a) (b)

Figure 7.8: Common interactions in the top predictions of SATORI and FIS for the DHS occupancy in IR dataset are shown in (a). The 10 most frequent TF family interactions are shown in (b).

tions. We compared SATORI to the FIS interaction estimation method and reported a 10x speed-up in its computation time in most cases. Furthermore, the top predictions made by both methods show very high overlap, suggesting such interactions as promising targets for follow-up biological experiments. This high overlap, despite the big difference in the approach provides good evidence for their potential biological relevance.

Pertinent to the problem at hand, SATORI identified numerous TF interactions with a potential role in the regulation of intron retention. A majority of these interactions were between the members of C2H2 zinc finger family of transcription factors. It follows that regulation of IR is orchestrated by complex interactions among transcription factors, predominantly from the zinc finger family. Further wet-lab experimentation is needed to validate these findings.

The proposed method can be extended in several ways. In this work we focused on globally scoring interactions between TFs with known PWMs. This is in contrast to feature attribution methods that score the contribution of features in genomic regions of interest. We believe that the sparsity of the attention matrix could make it useful as an attribution method as well, but further experiments are required in order to validate that. SATORI is able to detect interactions between

filters, even if they do not correspond to known TFs. Furthermore, the proposed methodology is flexible enough to be applied to deep networks that integrate multiple data modalities, and has potential applications outside of computational biology. For example, it can allow discovery of interactions between different characteristics of chromatin structure to provide a better understanding of the relationship between epigenetic markers such as histone modifications, DNA methylation, and nucleosome positioning and their contribution to the regulation of alternative splicing.

Chapter 8

Conclusions

In this work, we investigated the regulation of alternative splicing—primarily, intron retention—in both plants and animals, by leveraging from various epigenetic data sources. Our contributions are summarized below:

- We investigated the role chromatin accessibility in the regulation of intron retention in two plant species: arabidopsis and rice. Our findings suggested a more open and accessible chromatin in IR events compared to the non-IR events. Moreover, using a continuous HMM, we identified potential footprints in the regions of open chromatin. By quantifying the enriched sequence elements in the aforementioned footprints, we compiled a list of binding site motifs associated with the regulation of intron retention in both plant species.

- In human, we successfully predicted chromatin accessibility in IR events using a deep learning model. By analyzing the first convolutional layer filters, we identified numerous human transcription factors involved in the regulation of intron retention, a majority of which belonged to the C2H2 ZF family. These findings were validated using publicly available TF ChIP-Seq data and in wet lab experiments using candidate IR events.

- We developed a self-attention based model (SATORI) to infer cooperativity and interactions between regulatory proteins. SATORI successfully identified significant interactions between transcription factors across multiple datasets. We also used SATORI to identify TF interactions involved in the regulation of intron retention in human.

8.1 Open Problems

Next we discuss the open problems in this area and towards that end, the tentative experimentation that we have conducted which show promising results. We report these findings in this chapter as well as the accompanying supplementary material in Appendix D.

8.1.1 Predict chromatin accessibility in AS in plants

It is also crucial to understand how AS is regulated in plants. However, in the paradigm of deep learning, one of the problems associated with plant datasets is the far fewer number of AS events to begin with. For instance, in arabidopsis (TAIR10 genome annotations), there are roughly 3500 IR events and around 1500 of them have a DHS/THS overlap. We trained our deep learning model using that dataset and observed poor performance, slightly better than random guessing ($AUC \approx 0.55$). Nevertheless, using the same data with a gkm-SVM model gave an AUC score of 0.68. This shows that in plants, the overlapping DHSs/THSs potentially have IR specific regulatory elements (binding sites). Therefore, with sufficiently large dataset, we believe the deep learning model can successfully predict chromatin accessibility in AS events in plants. There are different ways to address this issue:

- Use RNA-Seq data along with gene models to find novel splicing events. Splicegrapher [1], for example, can take advantage of different expression datasets along with the genome annotations and call significantly more AS events.

- Design a multi-task deep learning model that use data from several plant species. In Chapter 5, We reported a significant number of IR hexamers that exhibited a footprint and were common between rice and arabidopsis. It follows that for AS in plants, the regulatory elements might not be too different. Therefore, by sharing datasets from different plant species in a multi-task learning setting will not only improve the model's performance but also capture the overall plant splicing code.

8.1.2 Investigate epigenetic regulation of other forms of AS

In this work, we primarily focused on intron retention for two reasons: first, it is the most prevalent form of AS in plants and second, it is not well studied and has remained greatly underappreciated. Nevertheless, it is also crucial to understand how chromatin plays its role in the regulation of other forms of AS: exon skipping, alternative 3' and 5' splicing. We have already run experiments using our deep learning model and the preliminary results are promising for the

aforementioned types of AS, particularly alternative 5' splicing. Table 8.1 summarize the results in terms of AUC scores.

Table 8.1: Preliminary results in terms of AUC scores for the three types of alternative splicing (ES, A3, and A5) data used with our deep CNN model. The table also shows the number of positive examples for each dataset. Note that the number of negative examples were roughly twice the size of the positive set.

AS type	No. of positive examples	AUC
Exon skipping	33705	0.7
Alternative 5' splicing	11646	0.83
Alternative 3' splicing	15570	0.73

From the table, we can see that the model exhibits decent AUC of 0.83 while predicting DHS occupancy in alternative 5' site splicing. In order to improve the model's performance in exon skipping and alternative 3' site splicing, one future direction is to extract the biologically relevant AS events. For example, in case of exon skipping, we used DHSs that overlapped the alternatively spliced exon. However, it has been shown that the regulatory elements can be found not only in the flanking introns but even the neighboring exons, upstream and downstream of the exon of interest [40]. We believe by using DHSs that occur in the extended regions around the alternatively spliced exons will improve the model performance and give us better information (in terms of binding site motifs) to understand the regulation of exon skipping.

8.1.3 Use evidence from other chromatin marks

Our findings suggest that DHS occupancy is crucial in understanding how AS is regulated in both plants and animals. Nevertheless, in Chapter 5 we also reported an association between intron retention and DNA methylation. Similarly, Braunschweig et al. [51] observed enrichment of certain histone modifications in the retained introns. It follows that it is also useful to incorporate evidence from several other chromatin marks (histone modifications, nucleosome occupancy, DNA methylation etc.) to better model the regulation of AS in both plants and animals. It is worth mentioning that we successfully predicted histone modification (*H3K4me3*) from sequence information alone, in a Basset [13] like experiment that used data from four cultivars of sorghum

under control and stress conditions, with an average AUC of 0.83. It follows that these chromatin markers can be used to assist our model better capturing the regulation of AS.

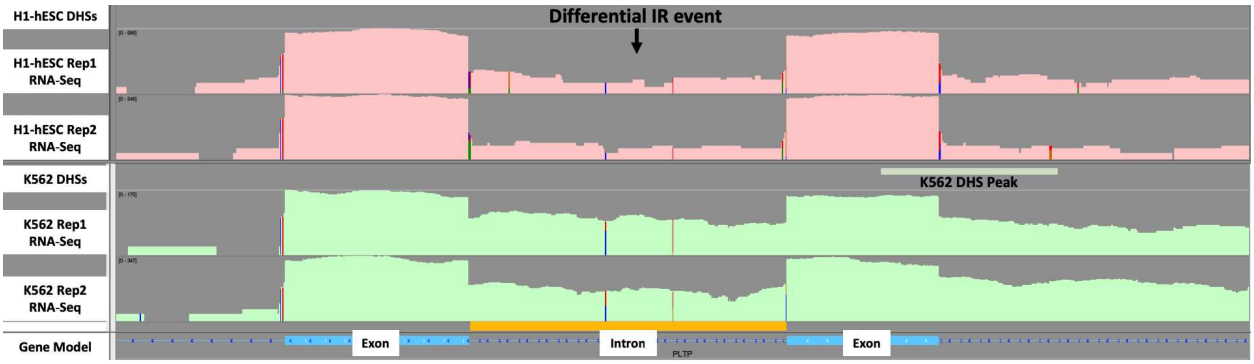
One way to incorporate data from chromatin marks directly into the deep learning model is by using the method described by Hiranuma et al. [199]. To predict transcription factor occupancy in the promoter region of a gene, they used chromatin accessibility data from an ATAC-Seq experiment and added it to the one-hot encoded representation of the input, at per-base level. This is particularly useful since to learn the underlying rules, their model had access to extra information: how "accessible" is the chromatin at each of the nucleotides in the DNA sequence. A significant improvement in performance was reported over DeepSea [100] which, in a similar problem, uses sequence information alone.

8.1.4 Towards a comprehensive epigenetic splicing code: tissue and condition specific splicing

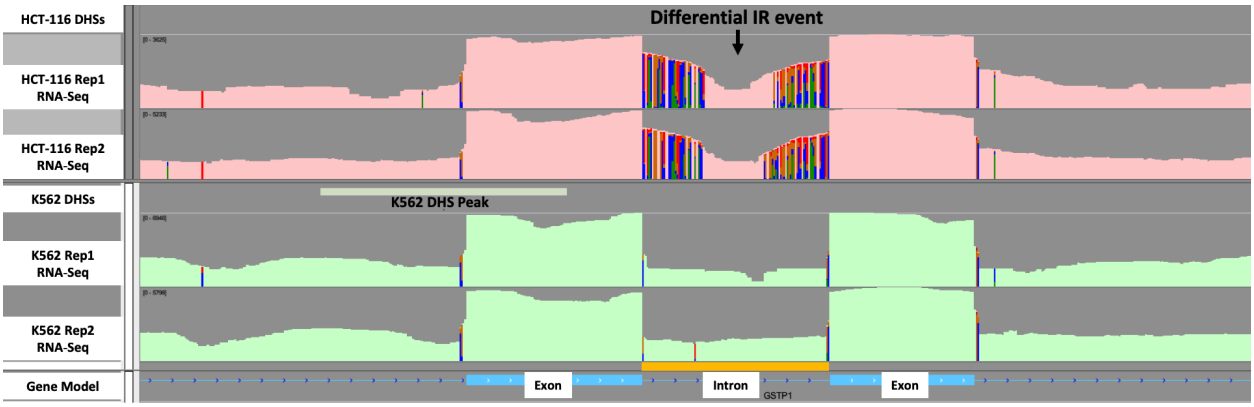
To compile a universal code—and set of rules—that govern the regulation of alternative splicing, it is important to understand how AS is regulated across different cell-lines/tissues (eg. liver vs. brain) or samples under different conditions (eg. heat stressed vs. control samples). In a way, this is an extension of differential gene expression which has been widely studied in both plants and animals. Differential AS has been reported in several studies; the major type of AS investigated are exon skipping in animals and intron retention in plants [174, 200]. Recently, we explored differential AS across two drought resistant and drought susceptible lines (cultivars) of sorghum [201]. In Appendix D, Figure D.1 shows the genome-wide occurrence of differential IR and ES events across control and treated samples in all four lines of sorghum. It shows that in different cultivars of sorghum, under regular and stress conditions, there is a strong evidence of differential AS.

To this end, our tentative experimentation suggest a strong association between differential intron retention and chromatin accessibility in human. Figure 8.1 depicts this association in two differential intron retention events in human cell lines. For a complete summary across eight

different cell-lines, refer to the Table D.1 in Appendix D. These preliminary results signify a need for further experimentation in order to elucidate the influence of chromatin state (accessibility, modifications, and arrangement) on the differential regulation of alternative splicing.



(a) Same Direction



(b) Opposite Direction

Figure 8.1: Differential occurrence of a DHS in a DIR event when comparing (a) the cell-lines *K562* and *H1-hESC*. The DIR event is evident from the RNA-seq coverage plots in two biological replicates of the corresponding cell-lines. The DHS is overlapping the up-regulated IR event (*K562*) while entirely absent in the down-regulated event (*H1-hESC*). That is, the differential DHS and IR event are in the same direction. The opposite direction DHS and IR event are shown in (b) for *K562* and *HCT-116* cell-lines. This plot is generated using Integrated Genome Viewer [6].

Bibliography

[1] M F Rogers, J Thomas, A S N Reddy, and A Ben-Hur. Splicegrapher: Detecting patterns of alternative splicing from RNA-seq data in the context of gene models and EST data. *Genome Biology*, 13, 2012.

[2] W Zhang, T Zhang, Y Wu, and J Jiang. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in arabidopsis. *Plant Cell*, 24:2719–31, 2012.

[3] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2017.

[4] Joseph R Ecker, Wendy A Bickmore, Inês Barroso, Jonathan K Pritchard, Yoav Gilad, and Eran Segal. Genomics: ENCODE explained. *Nature*, 489(7414):52, 2012.

[5] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature methods*, 6(11s):S22, 2009.

[6] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative Genomics Viewer. In *Nature Biotechnology*, 2011.

[7] Clifford A. Meyer and Xiaole Shirley Liu. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15:709–721, 2014.

[8] Bharath Ramsundar and Reza Bosagh Zadeh. *TensorFlow for deep learning: from linear regression to reinforcement learning*. " O'Reilly Media, Inc.", 2018.

[9] Xuan-Hien Le, Hung Viet Ho, Giha Lee, and Sungho Jung. Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, 11(7):1387, 2019.

[10] Christopher Olah. Understanding lstm networks, 2015. *URL http://colah.github.io/posts/2015-08-Understanding-LSTMs*, 2015.

[11] Shi Yan. Understanding LSTM and its diagrams. *MLReview.com*, 2016.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[13] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[14] Ian Dunham, Ewan Birney, Bryan R Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Troy W Whitfield, Jiali Zhuang, et al. An integrated encyclopedia of dna elements in the human genome. 2012.

[15] Martin Krzywinski, Jacqueline E. Schein, Inanç Birol, Joseph M. Connors, Randy D. Gascoyne, Douglas E. Horsman, Steven J. Jones, and Marco A. Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19 9:1639–45, 2009.

[16] A S N Reddy. Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu. Rev. Plant Biol.*, 58:267–94, 2007.

[17] A Kalsotra and TA Cooper. Functional consequences of developmentally regulated alternative splicing. *Nature Rev. Genetics*, 12:715–29, 20011.

[18] A S N Reddy, M F Rogers, D N Richardson, M Hamilton, and A Ben-Hur. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. *Front. Plant Sci.*, 3:18, 2012.

[19] M A Schuler. Plant pre-mRNA splicing. In J Bailey-Serres and D R Gallie, editors, *A look beyond transcription: Mechanisms determining mRNA stability and translation in plants*, pages 1–19. 1998.

[20] A S N Reddy. Nuclear pre-mRNA splicing in plants. *Crit. Rev. in Plant Sci.*, 20:523–71, 2001.

[21] Eddo Kim, Alon Magen, and Gil Ast. Different levels of alternative splicing among eukaryotes. In *Nucleic acids research*, 2007.

[22] Heebal Kim, Robert Klein, Jacek Majewski, and Jurg Ott. Estimating rates of alternative splicing in mammals and invertebrates. *Nature genetics*, 36(9):915, 2004.

[23] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.

[24] Nuno L Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Y Xiong, Serge Gueroussov, Leo J Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Çolak, et al. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.

[25] Richard Treisman, Stuart H Orkin, and Tom Maniatis. Specific transcription and RNA splicing defects in five cloned β-thalassaemia genes. *Nature*, 302:591–596, 1983.

[26] Richard Treisman, Nicholas J Proudfoot, Monica Shander, and Tom Maniatis. A single-base change at a splice site in a β-thalassemic gene causes abnormal RNA splicing. *Cell*, 29(3):903–911, 1982.

[27] Tsuyoshi Kashima and James L Manley. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nature Genetics*, 34:460–463, 2003.

[28] Luca Cartegni and Adrian R. Krainer. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nature Genetics*, 30:377–384, 2002.

[29] Michael T. Lovci, Dana Ghanem, Henry L Marr, Justin M Arnold, Sherry L. Gee, Marilyn G Parra, Tiffany Y. Liang, Thomas J. Stark, Lauren T. Gehman, Shawn Hoon, Katlin Brauer Massirer, Gabriel A. Pratt, Douglas L. Black, Joe W. Gray, J. G. Conboy, and Gene W. Yeo. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural Molecular Biology*, 20:1434–1442, 2013.

[30] Luca Cartegni, Shern L. Chew, and Adrian R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, 3:285–298, 2002.

[31] Charles J. David and James L Manley. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes development*, 24 21:2343–64, 2010.

[32] Jian Zhang and James L Manley. Misregulation of pre-mRNA alternative splicing in cancer. *Cancer discovery*, 3 11:1228–37, 2013.

[33] Richard A. Padgett. New connections between splicing and human disease. *Trends in genetics : TIG*, 28 4:147–54, 2012.

[34] Seishi Ogawa. Splicing factor mutations in myelodysplasia. *International journal of hematology*, 96 4:438–42, 2012.

[35] Hideki Makishima, V Visconte, Hirotoshi Sakaguchi, Anna M. Jankowska, Sarah M Abu Kar, Andrés Jerez, Bartlomiej Przychodzen, Manoj Kumar Bupathi, Kathryn M Guinta, Manuel G. Afable, Mikkael A. Sekeres, Richard A. Padgett, Ramón V. Tiu, and Jaroslaw Maciejewski. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. *Blood*, 119 14:3203–10, 2012.

[36] Christopher N Hahn and Hamish S. Scott. Spliceosome mutations in hematopoietic malignancies. *Nature Genetics*, 44:9–10, 2011.

[37] Bartlomiej Przychodzen, Andrés Jerez, Kathryn M Guinta, Mikkael A. Sekeres, Richard Padgett, Jaroslaw Maciejewski, and Hideki Makishima. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood*, 122 6:999–1006, 2013.

[38] Marcin Imielinski, Alice H. Berger, Peter S. Hammerman, Bryan Hernandez, Trevor J. Pugh, Eran Hodis, Jeonghee Cho, James Suh, Marzia Capelletti, Andrey Sivachenko, Carrie L. Sougnez, Daniel Auclair, Michael Sohn Lawrence, Petar Stojanov, Kristian Cibulskis, Kyusam Choi, L. F. de Waal, Tanaz Sharifnia, Angela N. Brooks, Heidi Greulich, Shantanu Banerji, Thomas Zander, Danila Seidel, Frauke Leenders, Sascha Ansén, Corinna Ludwig, Walburga Engel-Riedel, Erich Stoelben, Juergen Wolf, Chandra Goparju, Kristin M. Thompson, Wendy Winckler, David Kwiatkowski, BruceE. Johnson, Pasi Antero Jänne, Vincent A. Miller, William Pao, William D Travis, Harvey I. Pass, Stacey Bolk Gabriel, Eric S. Lander, Roman K. Thomas, Levi A. Garraway, Gad Getz, and Matthew Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150:1107–1120, 2012.

[39] Angela N. Brooks, Peter S Choi, Luc de Waal, Tanaz Sharifnia, Marcin Imielinski, Gordon Saksena, Chandra Sekhar Pedamallu, Andrey Sivachenko, Mara W. Rosenberg, Juliann Chmielecki, Michael Sohn Lawrence, David S. DeLuca, Gad Getz, and Matthew Meyerson. A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. In *PloS one*, 2014.

[40] Y Barash, J A Calarco, W Gao, Q Pan, X Wang, O Shai, B J Blencowe, and B J Frey. Deciphering the splicing code. *Nature*, 465:53–59, 2010.

[41] Evan C. Merkhofer, Peter H. Hu, and Tracy L. Johnson. Introduction to cotranscriptional RNA splicing. *Methods in molecular biology*, 1126:83–96, 2014.

[42] Yvonne N. Osheim, Jr. O.L. Miller, and A. L. Beyer. RNP particles at splice junction sequences on drosophila chorion transcripts. *Cell*, 43:143–151, 1985.

[43] A. L. Beyer, Amy H. Bouton, and Jr. O.L. Miller. Correlation of hnRNP structure and nascent transcript cleavage. *Cell*, 26:155–165, 1981.

[44] Zhengan Wu, Carol Murphy, Harold G. Callan, and Joseph G Gall. Small nuclear ribonucleoproteins and heterogeneous nuclear ribonucleoproteins in the amphibian germinal vesicle: loops, spheres, and snurposomes. *The Journal of Cell Biology*, 113:465 – 483, 1991.

[45] Yuanpeng Zhou, Yulan Lu, and Weidong Tian. Epigenetic features are significantly associated with alternative splicing. *BMC genomics*, 13(1):123, 2012.

[46] Reini F. Luco, Mariano Alló, Ignacio E. Schor, Alberto R. Kornblihtt, and Tom Misteli. Epigenetics in alternative pre-mRNA splicing. *Cell*, 144:16–26, 2011.

[47] Reini F Luco, Qun Pan, Kaoru Tominaga, Benjamin J Blencowe, Olivia M Pereira-Smith, and Tom Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000, 2010.

[48] Shiran Naftelberg, Ignacio E Schor, Gil Ast, and Alberto R Kornblihtt. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annual Review of Biochemistry*, 84:165–198, 2015.

[49] Galit Lev Maor, Ahuvi Yearim, and Gil Ast. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31(5):274–280, 2015.

[50] Xutong Wang, Lanjuan Hu, Xiaofei Wang, Ning Li, Chunming Xu, Lei Gong, and Bao Liu. DNA methylation affects gene alternative splicing in plants: an example from rice. *Molecular plant*, 9(2):305–307, 2016.

[51] U Braunschweig, N L Barbosa-Morais, Q Pan, E N Nachman, B Alipanahi, T Gonatopoulos-Pournatzis, B Frey, M Irimia, and B J Blencowe. Widespread intron retention in mammals functionally tunes transcription. *Genome Research*, 24:1774–86, 2014.

[52] Fahad Ullah, Michael Hamilton, Anireddy SN Reddy, and Asa Ben-Hur. Exploring the relationship between intron retention and chromatin accessibility in plants. *BMC genomics*, 19(1):21, 2018.

[53] F.H. Crick. On protein synthesis. *In Symp Soc Exp Biol*, 12(138-63):8, 1953.

[54] L A Chasin. Searching for splicing motifs. *Adv. Exp. Med. Biol.*, 623:85–106, 2007.

[55] I S Day, M Golovkin, S G Palusa, A Link, G S Ali, J Thomas, D N Richardson, and A S N Reddy. Interactions of SR45, an SR-like protein, with spliceosomal proteins and an intronic sequence: insights into regulated splicing. *The Plant Journal*, 71:936–47, 2012.

[56] R Xiao, Y Sun, S H Ding, S Lin, D W Rose, M G Rosenfeld, X D Fu, and X Li. Splicing regulator SC35 is essential for genomic stability and cell proliferation during mammalian organogenesis. *Mol. Cell. Biology*, 27:5393–5402, 2007.

[57] Yoseph Barash, John A Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J Blencowe, and Brendan J Frey. Deciphering the splicing code. *Nature*, 465(7294):53, 2010.

[58] J C Long and J F Caceres. The SR protein family of splicing factors: master regulators of gene expression. *Biochem. Journal*, 417:15–27, 2009.

[59] Douglas L Black. Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes & development*, 5(3):389–402, 1991.

[60] Donald J Patterson, Ken Yasuhara, and Walter L Ruzzo. Pre-mRNA secondary structure prediction aids splice site prediction. In *Biocomputing 2002*, pages 223–234. World Scientific, 2001.

[61] Sayed-Amir Marashi, Changiz Eslahchi, Hamid Pezeshk, and Mehdi Sadeghi. Impact of RNA structure on the prediction of donor and acceptor splice sites. *BMC Bioinformatics*, 7:297 – 297, 2006.

[62] Peter J. Shepard and Klemens J. Hertel. Conserved RNA secondary structures promote alternative splicing. *RNA*, 14 8:1463–9, 2008.

[63] Emanuele Buratti and Francisco E. Baralle. Influence of RNA secondary structure on the pre-mRNA splicing process. *Molecular and cellular biology*, 24 24:10505–14, 2004.

[64] M. Bryan Warf and J Andrew Berglund. Role of RNA structure in regulating pre-mRNA splicing. *Trends in biochemical sciences*, 35 3:169–78, 2010.

[65] Auinash Kalsotra, Xinshu Xiao, Amanda J. Ward, John C Castle, Jason M. Johnson, Christopher B. Burge, and Thomas A Cooper. A postnatal switch of CELF and MBNL proteins reprograms alternative splicing in the developing heart. *Proceedings of the National Academy of Sciences of the United States of America*, 105 51:20333–8, 2008.

[66] Eric T. Wang, Neal A.L. Cody, Sonali P. Jog, Michela Biancolella, Thomas T. Wang, Daniel J. Treacy, Shujun Luo, Gary P. Schroth, David Housman, Sita Reddy, Eric Lécuyer, and Christopher B. Burge. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, 150:710–724, 2012.

[67] Debashish Ray, Hilal Kazan, Esther T. Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J. Blencowe, Quaid D Morris, and Timothy R. Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27:667–670, 2009.

[68] Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Fei Ji, Young-Soo Kwon, Chao Zhang, G. K. Yeo, Douglas L. Black, Hui Sun, Xiang-Dong Fu, and Yi Zhang. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Molecular cell*, 36 6:996–1006, 2009.

[69] Luciana I Gómez Acuña, Ana Fiszbein, Mariano Alló, Ignacio E. Schor, and Alberto R. Kornblihtt. Connections between chromatin signatures and splicing. *Wiley interdisciplinary reviews. RNA*, 4 1:77–91, 2013.

[70] Sérgio F de Almeida and Maria Carmo-Fonseca. Design principles of interconnections between chromatin and pre-mRNA splicing. *Trends in biochemical sciences*, 37 6:248–53, 2012.

[71] Camilla Iannone and Juan Valcárcel. Chromatin's thread to alternative splicing regulation. *Chromosoma*, 122:465–474, 2013.

[72] Gwendal Dujardin, Celina Lafaille, Ezequiel Petrillo, Valeria Buggiano, Luciana I Gómez Acuña, Ana Fiszbein, Micaela A Godoy Herz, Nicolás Nieto Moreno, Manuel Javier Muñoz, Mariano Alló, Ignacio E. Schor, and Alberto R. Kornblihtt. Transcriptional elongation and alternative splicing. *Biochimica et biophysica acta*, 1829 1:134–40, 2013.

[73] Manuel de la Mata, Claudio R Alonso, Sebastián Kadener, Juan P Fededa, Matıas Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R Kornblihtt. A slow RNA Polymerase II affects alternative splicing in vivo. *Molecular Cell*, 12(2):525 – 532, 2003.

[74] Mark D. Adams, Jenny M. Kelley, Jeannine D. Gocayne, Mark Dubnick, Mihael H. Polymeropoulos, Huan juan Xiao, C. R. Merril, Ai min Wu, Bjorn Olde, and Rita Moreno. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252 5013:1651–6, 1991.

[75] Victor E. Velculescu, Lan Zhang, Bert Vogelstein, and Kenneth W. Kinzler. Serial analysis of gene expression. *Science*, 270 5235:484–7, 1995.

[76] Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 5235:467–70, 1995.

[77] Ryan Lister, Ronan O'Malley, Julian Tonti-Filippini, Brian D. Gregory, Charles C. Berry, A. Harvey Millar, and Joseph R. Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133:523–536, 2008.

[78] Wei Zhao, Xiaping He, Katherine A Hoadley, Joel S Parker, David Neil Hayes, and Charles M Perou. Comparison of RNA-seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC genomics*, 15(1):419, 2014.

[79] D Kim, G Pertea, C Trapnell, H Pimentel, R Kelley, and S L Salzberg. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14:R36, 2013.

[80] A Dobin, Davis C A, F Schlesinger, J Drenkow, C Zaleski, S Jha, P Batut, M Chaisson, and T R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15–21, 2013.

[81] T R Mercer, S L Edwards, M B Clark, S J Neph, H Wang, A B Stergachis, S John, R Sandstrom, G Li, K S Sandhu, Y Ruan, L K Nielsen, J S Mattick, and J Stamatoyannopoulos. DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genetics*, 45:852–59, 2013.

[82] D. S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara J. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316 5830:1497–502, 2007.

[83] Carlo Nobile, Joanne M. Nickol, and Robert G. Martin. Nucleosome phasing on a DNA fragment from the replication origin of simian virus 40 and rephasing upon cruciform formation of the DNA. *Molecular and cellular biology*, 6 8:2916–22, 1986.

[84] A P Boyle, S Davis, H P Shulha, P Meltzer, E H Margulies, Z Weng, T S Furey, and G E Crawford. High resolution mapping and characterization of open chromatin across genomes. *Cell*, 132:311–22, 2008.

[85] Paul G. Giresi, Jonghwan Kim, Ryan McDaniell, Vishwanath R. Iyer, and Jason D. Lieb. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome research*, 17 6:877–85, 2007.

[86] Jason D Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10:1213–1218, 2013.

[87] Christian Ibarra, Xiaoqi Feng, Vera Karolina Schoft, Tzung-Fu Hsieh, Rie Uzawa, Jessica Astrid Rodrigues, Assaf Zemach, Nina Chumak, Adriana Machlicová, Toshiro Nishimura, Denisse Rojas, Robert L. Fischer, Hisashi Tamaru, and Daniel Zilberman. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, 337 6100:1360–1364, 2012.

[88] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75, 2012.

[89] B Langmead, C Trapnell, M Pop, and S L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, 2009.

[90] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[91] S John, P J Sabo, R E Thurman, M H Sung, S C Biddie, T A Johnson, G L Hager, and J A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43:264–68, 2011.

[92] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *bioinformatics*, 27(11):1571–1572, 2011.

[93] F Rosenbaltt. The perceptron–a perceiving and recognizing automation. *Report 85-460-1 Cornell Aeronautical Laboratory, Ithaca, Tech. Rep.*, 1957.

[94] Geoffrey E Hinton, Alex Krizhevsky, and Ilya Sutskever. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1106–1114, 2012.

[95] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[96] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[97] Ramzan Kh Umarov and Victor V Solovyev. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, 12(2):e0171410, 2017.

[98] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831, 2015.

[99] Hamid Reza Hassanzadeh and May D Wang. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 178–183. IEEE, 2016.

[100] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931, 2015.

[101] Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research*, 44(11):e107–e107, 2016.

[102] Nicholas E Banovich, Yang I Li, Anil Raj, Michelle C Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E Burnett, Marsha Myrthil, Samantha M Thomas, et al. Impact of regulatory variation across human iPSCs and differentiated cells. *Genome research*, 28(1):122–131, 2018.

[103] Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome biology*, 18(1):67, 2017.

[104] Pang Wei Koh, Emma Pierson, and Anshul Kundaje. Denoising genome-wide histone ChIP-seq with convolutional neural networks. *Bioinformatics*, 33(14):i225–i233, 2017.

[105] Xiaoyong Pan and Hong-Bin Shen. RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18(1):136, 2017.

[106] Paul J Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356, 1988.

[107] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[108] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[109] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[110] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

[111] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[112] Chen Chen, Jie Hou, Xiaowen Shi, Hua Yang, James A Birchler, and Jianlin Cheng. Interpretable attention model in transcription factor binding site prediction with deep neural networks. *bioRxiv*, page 648691, 2019.

[113] Uwe Ohler, Guo-chun Liao, Heinrich Niemann, and Gerald M Rubin. Computational analysis of core promoters in the Drosophila genome. *Genome biology*, 3(12):research0087–1, 2002.

[114] Philipp Bucher. Weight matrix descriptions of four eukaryotic RNA polymerase ii promoter elements derived from 502 unrelated promoter sequences. *Journal of molecular biology*, 212(4):563–578, 1990.

[115] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith 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.

[116] Sven Degroeve, Bernard De Baets, Yves Van de Peer, and Pierre Rouzé. Feature subset selection for splice site prediction. *Bioinformatics*, 18(suppl_2):S75–S83, 2002.

[117] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K Moore, Ji-Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, 2006.

[118] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[119] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS computational biology*, 4(10):e1000173, 2008.

[120] Jean-Philippe Vert. Kernel methods in genomics and computational biology. In *Kernel methods in bioengineering, signal and image processing*, pages 42–63. IGI Global, 2007.

[121] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific, 2001.

[122] Christina S Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.

[123] Gideon Dror, Rotem Sorek, and Ron Shamir. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, 21(7):897–901, 2005.

[124] Gunnar Rätsch, Sören Sonnenburg, and Bernhard Schölkopf. RASE: recognition of alternatively spliced exons in C. elegans. *Bioinformatics*, 21(suppl_1):i369–i377, 2005.

[125] Rui Mao, Praveen Kumar Raj Kumar, Cheng Guo, Yang Zhang, and Chun Liang. Comparative analyses between retained introns and constitutively spliced introns in arabidopsis thaliana using random forest and support vector machine. *PLoS One*, 9(8), 2014.

[126] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad Noori, and Michael A. Beer. Enhanced regulatory sequence prediction using gapped k-mer features. In *PLoS Computational Biology*, 2014.

[127] Hui Liu, Ting Jin, Jihong Guan, and Shuigeng Zhou. Histone modifications involved in cassette exon inclusions: a quantitative and interpretable analysis. *BMC genomics*, 15(1):1148, 2014.

[128] Alexander Statnikov and Constantin F Aliferis. Are random forests better than support vector machines for microarray-based cancer classification? In *AMIA annual symposium proceedings*, volume 2007, page 686. American Medical Informatics Association, 2007.

[129] Liang Chen and Sika Zheng. Identify alternative splicing events based on position-specific evolutionary conservation. *PLoS One*, 3(7), 2008.

[130] Sean R Eddy. What is a hidden markov model? *Nature biotechnology*, 22(10):1315, 2004.

[131] Byung-Jun Yoon. Hidden markov models and their applications in biological sequence analysis. *Current genomics*, 10(6):402–415, 2009.

[132] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[133] Hui Yuan Xiong, Yoseph Barash, and Brendan J Frey. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, 27(18):2554–2562, 2011.

[134] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[135] Anupama Jha, Matthew R Gazzara, and Yoseph Barash. Integrative deep models for alternative splicing. *Bioinformatics*, 33(14):i274–i282, 2017.

[136] Fernando Carrillo Oesterreich, Lydia Herzel, Korinna Straube, Katja Hujer, Jonathon Howard, and Karla M Neugebauer. Splicing of nascent RNA coincides with intron exit from RNA polymerase ii. *Cell*, 165(2):372–381, 2016.

[137] G Felsenfeld and M Groudine. Controlling the double helix. *Nature*, 421:448–53, 2003.

[138] D S Gross and W T Garrard. Nuclease hypersensitive sites in chromatin. *Annual Rev. Biochem.*, 57:159–197, 1988.

[139] D J Galas and A Schmitz. DNase footprinting: A simple method for detection of protein-DNA binding specificity. *Nucleic Acids Res.*, 5:3157–70, 1978.

[140] J R Hesselberth, X Y Chen, Z H Zhang, P J Sabo, R Sandstrom, A P Reynolds, R E Thurman, S Neph, M S Kuehn, W S Noble, S Fields, and J A Stamatoyannopoulos. Global mapping of protein-DNA interactions in-vivo by digital genomic footprinting. *Nature Methods*, 6:283–89, 2009.

[141] A P Boyle, L Y Song, B K Lee, D London, D Keefe, E Birney, V R Iyer, G E Crawford, and T S Furey. High-resolution genome-wide in-vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21:456–64, 2011.

[142] E Birney et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

[143] W Zhang, Y Wu, J C Schnable, Z Zeng, M Freeling, G E Crawford, and J Jiang. High-resolution mapping of open chromatin in the rice genome. *Genome Research*, 22:151–62, 2012.

[144] H B Mann and D R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[145] Ronan C O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G Lewsey, Anna Bartlett, Joseph R Nery, Mary Galli, Andrea Gallavotti, and Joseph R Ecker. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 165(5):1280–1292, 2016.

[146] Karl-Josef Dietz, Marc Oliver Vogel, and Andrea Viehhauser. Ap2/erebp transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. *Protoplasma*, 245(1-4):3–14, 2010.

[147] Shuichi Yanagisawa and Robert J Schmidt. Diversity and similarity among recognition sequences of dof transcription factors. *The Plant Journal*, 17(2):209–214, 1999.

[148] Hong Han, Ulrich Braunschweig, Thomas Gonatopoulos-Pournatzis, Robert J Weatheritt, Calley L Hirsch, Kevin CH Ha, Ernest Radovani, Syed Nabeel-Shah, Tim Sterne-Weiler, Juli Wang, et al. Multilayered control of alternative splicing regulatory networks by transcription factors. *Molecular cell*, 65(3):539–553, 2017.

[149] Tim R Mercer, Stacey L Edwards, Michael B Clark, Shane J Neph, Hao Wang, Andrew B Stergachis, Sam John, Richard Sandstrom, Guoliang Li, Kuljeet S Sandhu, et al. DNase I–hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature genetics*, 45(8):852, 2013.

[150] M Wang, Y Zhao, and B Zhang. Efficient test and visualization of multi-set intersections. *Scientific Reports*, 5:16923, 2015.

[151] Adam Labadorf, Alicia Link, Mark F Rogers, Julie Thomas, Anireddy SN Reddy, and Asa Ben-Hur. Genome-wide analysis of alternative splicing in *Chlamydomonas reinhardtii*. *BMC genomics*, 11(1):114, 2010.

[152] Mélanie Rigal, Zoltán Kevei, Thierry Pélissier, and Olivier Mathieu. DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. *The EMBO journal*, 31(13):2981–2993, 2012.

[153] Juan I Young, Eugene P Hong, John C Castle, Juan Crespo-Barreto, Aaron B Bowman, Matthew F Rose, Dongcheul Kang, Ron Richman, Jason M Johnson, Susan Berget, et al. Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-cpg binding protein 2. *Proceedings of the National Academy of Sciences*, 102(49):17551–17558, 2005.

[154] Mohamed Elhiti and Claudio Stasolla. Structure and function of homodomain-leucine zipper (HD-Zip) proteins. *Plant signaling & behavior*, 4(2):86–88, 2009.

[155] A Pajoro, E Severing, GC Angenent, and RGH Immink. Histone H3 lysine 36 methylation affects temperature-induced alternative splicing and flowering in plants. *Genome biology*, 18(1):102, 2017.

[156] Y Wu, S Kikuchi, H Yan, W Zhang, et al. Euchromatic subdomains in rice centromeres are associated with genes and transcription. *Plant Cell*, 23:4054–64, 2011.

[157] Assaf Zemach, M Yvonne Kim, Ping-Hung Hsieh, Devin Coleman-Derr, Leor Eshed-Williams, Ka Thao, Stacey L Harmer, and Daniel Zilberman. The arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 153(1):193–205, 2013.

[158] Ramakrishna K Chodavarapu, Suhua Feng, Bo Ding, Stacey A Simon, David Lopez, Yulin Jia, Guo-Liang Wang, Blake C Meyers, Steven E Jacobsen, and Matteo Pellegrini. Transcriptome and methylome interactions in rice hybrids. *Proceedings of the National Academy of Sciences*, 109(30):12040–12045, 2012.

[159] FastQC toolkit.

[160] Gregory J Hannon. FASTX-toolkit: FASTQ/A short-reads pre-processing tools, 2012.

[161] F Krueger. Trim galore!: a wrapper tool around cutadapt and fastqc to consistently apply quality and adapter trimming to fastq files. babraham bioinformatics, cambridge, united kingdom, 2015.

[162] W A Stein et al. *SageMath, the Sage Mathematics Software System (Version 7.1)*, 2016. `http://www.sagemath.org`.

[163] M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and clustal X version 2.0. *Bioinformatics*, 23:2947–48, 2007.

[164] G E Crooks, G Hon, J M Chandonia, and S E Brenner. Weblogo: A sequence logo generator. *Genome Research*, 14:1188–90, 2004.

[165] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[166] Steven J Burgess, Ivan Reyna-Llorens, Sean R Stevenson, Pallavi Singh, Katja Jaeger, and Julian M Hibberd. Genome-wide transcription factor binding in leaves from C3 and C4 grasses. *The Plant Cell*, 31(10):2297–2314, 2019.

[167] Sai Zhang, Jingtian Zhou, Hailin Hu, Haipeng Gong, Ligong Chen, Chao Cheng, and Jianyang Zeng. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic acids research*, 44(4):e32–e32, 2015.

[168] Qian Qin and Jianxing Feng. Imputation for transcription factor binding predictions based on deep learning. *PLoS computational biology*, 13(2):e1005403, 2017.

[169] Jacob Schreiber, Maxwell Libbrecht, Jeffrey Bilmes, and William Noble. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv*, page 103614, 2018.

[170] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[171] Anupama Jha, Matthew R Gazzara, and Yoseph Barash. Integrative deep models for alternative splicing. *Bioinformatics*, 33(14):i274–i282, 2017.

[172] Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. Predicting splicing from primary sequence with deep learning. *Cell*, 2019.

[173] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J. Ziller, Viren Amin, John W. Whitaker, Matthew D. Schultz, Lucas D. Ward, Abhishek Sarkar, Gerald Quon, Richard S. Sandstrom, Matthew L. Eaton, Yi-Chieh Wu, Andreas R. Pfenning, Xinchen Wang, Melina Claussnitzer, Yaping Liu, Cristian Coarfa, R. Alan Harris, Noam Shoresh, Charles B. Epstein, Elizabeta Gjoneska, Danny Leung, Wei Xie, R. David Hawkins, Ryan Lister, Chibo Hong, Philippe Gascard, Andrew J. Mungall, Richard A. Moore, Eric Chuah, Angela Tam, Theresa K. Canfield, R. Scott Hansen, Rajinder Kaul, Peter J. Sabo, Mukul S. Bansal, Annaick Carles, Jesse R. Dixon, Kai-How Farh, Soheil Feizi, Rosa Karlic, Ah-Ram Kim, Ashwinikumar Kulkarni, Daofeng Li, Rebecca F. Lowdon, GiNell Elliott, Tim R. Mercer, Shane J. Neph, Vitor Onuchic, Paz Polak, Nisha Rajagopal, Pradipta Ray, Richard C. Sallari, Kyle T. Siebenthall, Nicholas A. Sinnott-Armstrong, Michael C. Stevens, Robert E. Thurman, Jie Jayne Wu, Bo Zhang, Xin Zhou, Arthur E. Beaudet, Laurie A. Boyer, Philip L. De Jager, Peggy J. Farnham, Susan J. Fisher, David Haussler, Steven J. M. Jones, Wei Li, Marco A. Marra, Michael T. McManus, Shamil R. Sunyaev, James A. Thomson, Thea D. Tlsty, Li-Huei Tsai, Wei Wang, Robert A. Waterland, Michael Q. Zhang, Lisa H. Chadwick, Bradley E. Bernstein, Joseph F. Costello, Joseph R. Ecker, Martin Hirst, Alexander Meissner, Aleksandar Milosavljevic, Bing Ren, John A. Stamatoyannopoulos, Ting Wang, and Manolis Kellis. Integrative analysis of 111 reference human epigenomes. In *Nature*, 2015.

[174] Sergei A Filichkin, Michael Hamilton, Palitha Dharmawardhana, Sunil Kumar Singh, Christopher W. Sullivan, Asa Ben-Hur, A. S. N. Reddy, and Pankaj Jaiswal. Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching. In *Front. Plant Sci.*, 2018.

[175] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[176] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific reports*, 8(1):1–10, 2018.

[177] Dongwon Lee. LS-GKM: a new gkm-svm for large-scale datasets. *Bioinformatics*, 32 14:2196–8, 2016.

[178] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate B. Cook, Hong Yuan Zheng, Alejandra Goity, Harm van Bakel, Javier Fernández Lozano, Mary Galli, Mathew G Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J M Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158:1431–1443, 2014.

[179] Shobhit Gupta, John A. Stamatoyannopoulos, Timothy L. Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8:R24 – R24, 2006.

[180] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[181] Thomas D Schneider, Gary D Stormo, Larry Gold, and Andrzej Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431, 1986.

[182] Peter K Koo and Matthew Ploenzke. Improving convolutional network interpretability with exponential activations. *bioRxiv*, page 650804, 2019.

[183] Nicole D Robson-Dixon and Mariano A Garcia-Blanco. MAZ elements alter transcription elongation and silencing of the fibroblast growth factor receptor 2 exon IIIb. *Journal of Biological Chemistry*, 279(28):29075–29084, 2004.

[184] Wyeth W Wasserman and James W Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *Journal of molecular biology*, 278(1):167–181, 1998.

[185] Sridhar Hannenhalli and Samuel Levy. Predicting transcription factor synergism. *Nucleic acids research*, 30(19):4278–4284, 2002.

[186] Debraj GuhaThakurta and Gary D Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–621, 2001.

[187] Yitzhak Pilpel, Priya Sudarsanam, and George M Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, 29(2):153, 2001.

[188] Priya Sudarsanam, Yitzhak Pilpel, and George M Church. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome research*, 12(11):1723–1731, 2002.

[189] Hernan Roca and Renny T Franceschi. Analysis of transcription factor interactions in osteoblasts using competitive chromatin immunoprecipitation. *Nucleic acids research*, 36(5):1723–1730, 2008.

[190] Stephen Safe. MicroRNA-specificity protein (Sp) transcription factor interactions and significance in carcinogenesis. *Current pharmacology reports*, 1(2):73–78, 2015.

[191] Gaia Ceddia, Liuba Nausicaa Martino, Alice Parodi, Piercesare Secchi, Stefano Campaner, and Marco Masseroli. Association rule mining to identify transcription factor interactions in genomic regions. *Bioinformatics (Oxford, England)*, 2019.

[192] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

[193] Peyton Greenside, Tyler Shimko, Polly Fordyce, and Anshul Kundaje. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics*, 34(17):i629–i637, 2018.

[194] Ge Liu, Haoyang Zeng, and David K Gifford. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC bioinformatics*, 20(1):1–14, 2019.

[195] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[196] Ronan C O'Malley, Shao-shan Carol Huang, Liang Song, Mathew G Lewsey, Anna Bartlett, Joseph R Nery, Mary Galli, Andrea Gallavotti, and Joseph R Ecker. Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, 165(5):1280–1292, 2016.

[197] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017.

[198] Mira T Kassouf, Jim R Hughes, Stephen Taylor, Simon J McGowan, Shamit Soneji, Angela L Green, Paresh Vyas, and Catherine Porcher. Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome research*, 20(8):1064–1083, 2010.

[199] Naozumi Hiranuma, Scott Lundberg, and Su-In Lee. DeepATAC: A deep-learning method to predict regulatory factor binding activity from ATAC-seq signals. *bioRxiv*, page 172767, 2017.

[200] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome research*, 22 10:2008–17, 2012.

[201] Anireddy Reddy and Asa Ben-Hur. Global analysis of epigenetic regulation of gene expression in response to drought stress in sorghum. Technical report, Colorado State Univ., Fort Collins, CO (United States), 2017.

[202] Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin C. Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William Stafford Noble. MEME Suite: tools for motif discovery and searching. In *Nucleic Acids Research*, 2009.

[203] Pouya Kheradpour and Manolis Kellis. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research*, 42(5):2976–2987, 2014.

[204] Kevin Streit, Clemens Hammacher, Andreas Zeller, and Sebastian Hack. Sambamba: runtime adaptive parallel execution. In *ADAPT '13*, 2013.

[205] Kevin C. Potter, Judy Wang, G Eric Schaller, and Joseph J Kieber. Cytokinin modulates context-dependent chromatin accessibility through the type-B response regulators. *Nature Plants*, 4:1102–1111, 2018.

[206] Denghui Xing, Yajun Wang, Michael Hamilton, Asa Ben-Hur, and Anireddy S. N. Reddy. Transcriptome-wide identification of RNA targets of arabidopsis SERINE/ARGININE-RICH45 uncovers the unexpected roles of this RNA binding protein in RNA processing. *The Plant cell*, 27 12:3294–308, 2015.

Appendix A

Chapter 5 Supplementary Material

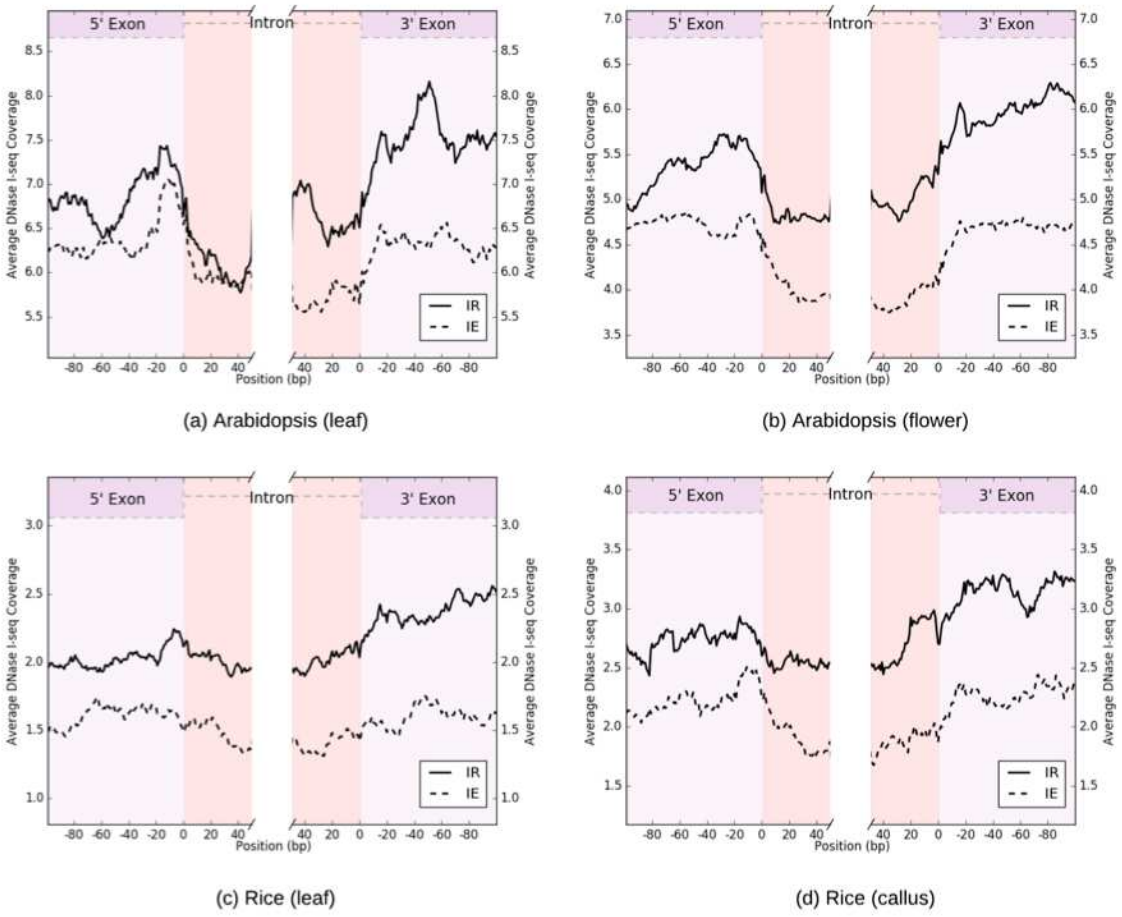


Figure A.1: Average DNase I-seq coverage profile is shown across IR and IE events in the four samples: Arabidopsis leaf (a) and flower (b); rice leaf (c) and callus (d). The profile is centered at the 5' and 3' splice sites (indicated by "0" on x-axis in the split figure), and goes 50bp into the intron and 100bp into the flanking exons. Note that we chose all three parts of an IR/IE event to be at least 100bp. These profiles do not include events that come from the first intron of a gene. Moreover, to avoid bias, for each IR event, we selected IE events with similar relative positions within the gene.

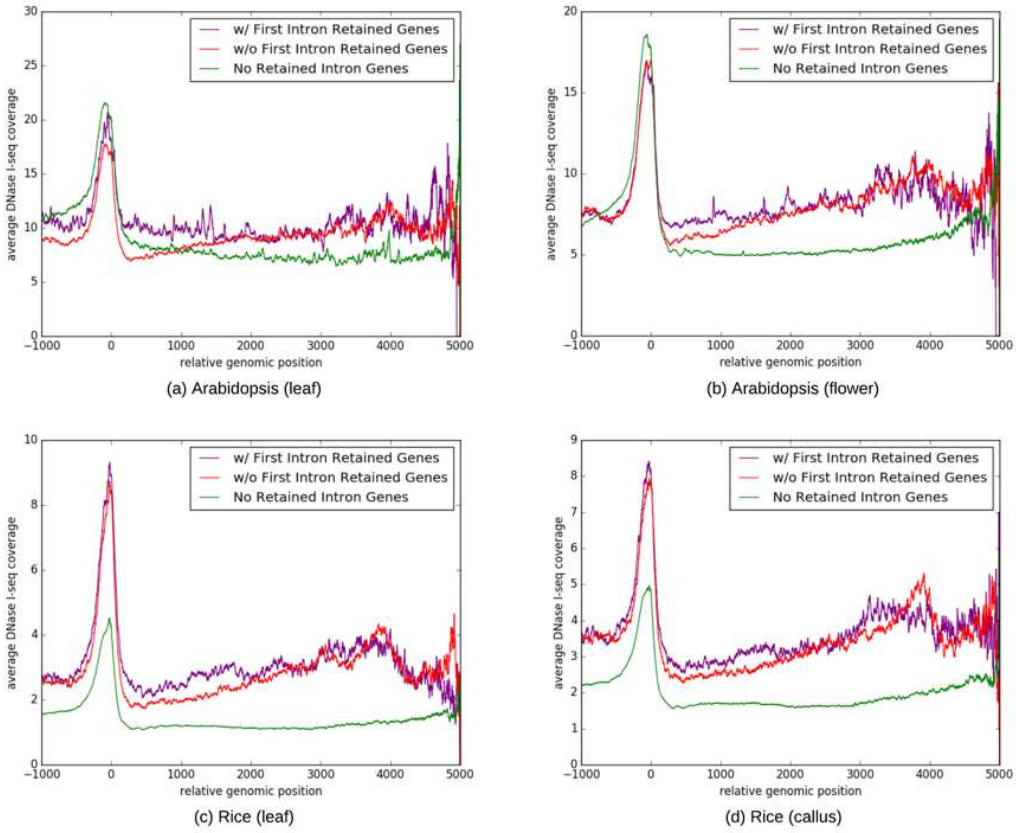


Figure A.2: Average DNase I-seq coverage profile for the four samples: Arabidopsis leaf (a) and flower (b); rice leaf (c) and callus (d). In each case, a pool of genes up to 5000bp in length are used (roughly 95% of total genes). The profile encompasses the gene body and 1000bp upstream of the transcription start site represented by '0' on the x-axis. Each figure shows the profile for three sub-categories: genes with first intron retained (purple), genes with intron(s) retained anywhere else but the first one (red), and genes without any retained intron (green).

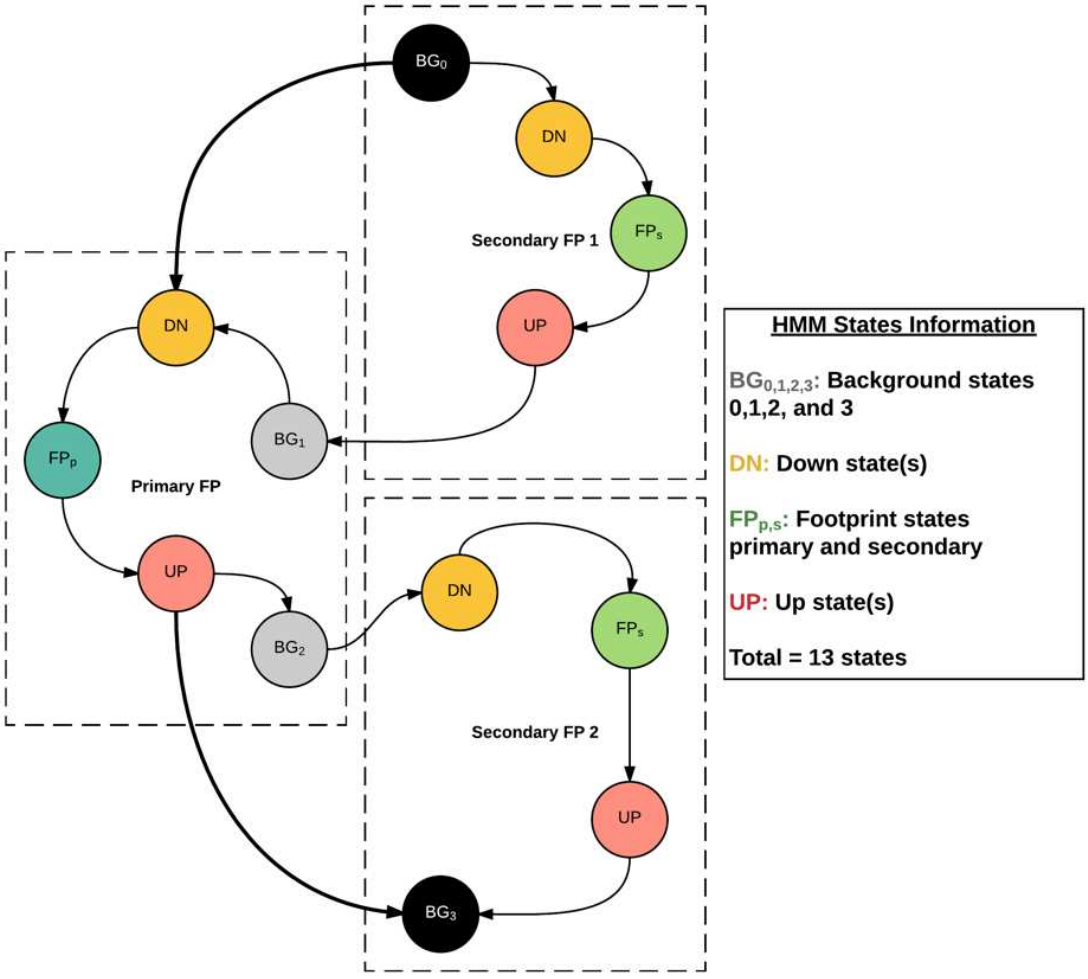


Figure A.3: The complete state diagram for the continuous HMM used to predict hexamers with potential footprints. The diagram shows all 13 states. The HMM consists of three modules, to enable us to model leading/trailing footprints in addition to the primary footprint. Each module has copies of the five core states. The size of the arrow (transition) as in $BG_0 \longrightarrow DN$ and $UP \longrightarrow BG_3$ represents higher probabilities than the other transitions from the same state. These probabilities are highlighted in the supplementary table 4. This is used to emphasize the primary footprint detection by our model in all cases. The figure also summarizes the HMM states in the rectangular box to the right.

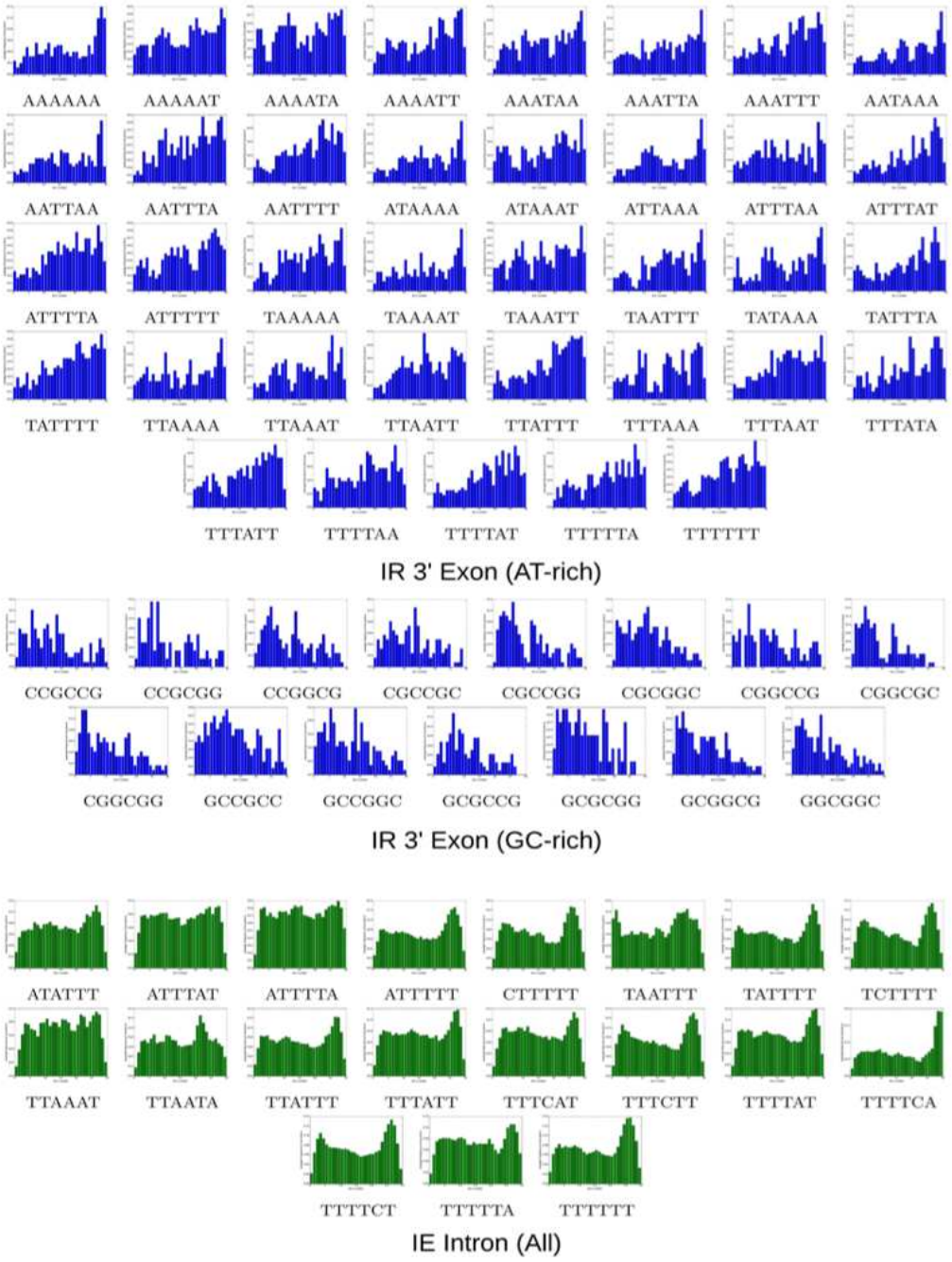


Figure A.4: Positional preference is shown for AT-rich hexamers (top), GC-rich hexamers (middle) in 3' exon region of IR events, and all hexamers in intron region of IE events (bottom). All hexamers mentioned in the figure exhibit a footprint.

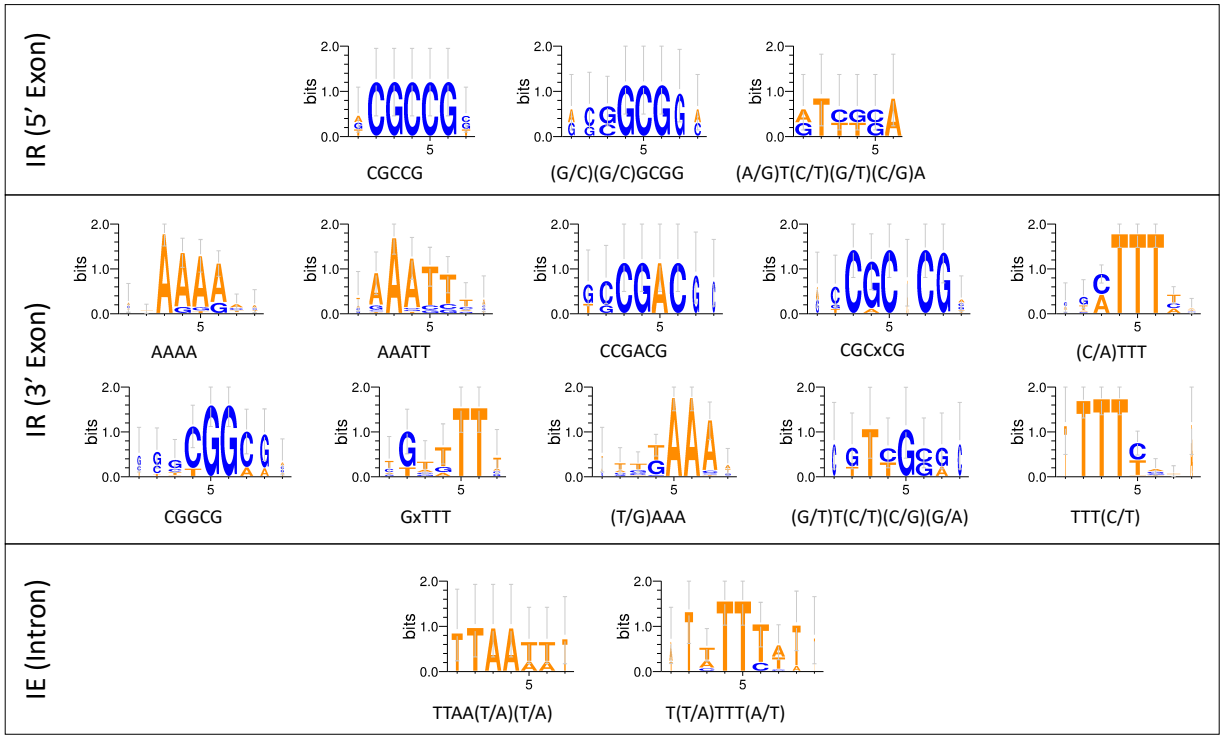


Figure A.5: Motifs generated after clustering the IR and IE enriched hexamers exhibiting a footprint across in leaf samples in both species. Motif logos were generated using the weblogo tool. In the table, these motifs are grouped based on the type of event (IR and IE) they are enriched in and part of the event from which their respective hexamers were found (5' exon, intron, and 3' exon).

Table A.1: Alignment statistics for different Arabidopsis thaliana (AT) and rice samples. Note that the aligned reads went through preprocessing and then aligned using tophat2 for RNAseq and bowtie/STAR for DNase I-seq (see Methods in the main text). The reads in both cases (DNase I-seq and RNA-seq) were filtered for multiple alignments and filtered for spurious junctions for the RNAseq. Also, in all samples, biological and technical replicates were pooled. As mentioned in the main text, we used pre-aligned DNase I-seq and RNA-seq from [2].

Type of data	Sample	Total Reads	Aligned reads (Unique/Filtered)
DNase I-seq	Rice (leaf, control) [143]	42593905	29260669 (68.70%)
	Rice (callus, control) [143]	57037438	39867789 (69.90%)
RNA-seq	Rice (leaf, control) [143]	40206025	37364769 (92.93%)
	Rice (callus, control) [143]	29634838	27100117 (91.45%)
Bisulfite-seq	Arabidopsis [157]	41177470	16559509 (40.20%)
	Rice [158]	130128482	62386292 (47.90%)

Table A.2: Hotspot was used to call DHS peaks in all DNase I-seq samples. Rice samples, on average, had more DHS peaks identified. Since hotspot can't handle replicates, we pooled DNase I-seq libraries.

S.No.	Sample	# of DHSs
1	AT (leaf) [2]	45,665
2	AT (Flower) [2]	42,782
3	Rice (Leaf) [143]	69,277
4	Rice (Callus) [143]	107,092

Table A.3: Number of IR and IE events extracted at different coverage levels are listed below. Evidence from known gene models and RNA-seq data was used to extract the events as described in the Methods section in the main paper.

Sample	Expression Level	IR Events	IE Events
AT (Leaf) [2]	1	3804	63538
	5	3599	50666
	10	3196	38568
	20	2397	21426
AT (Flower) [2]	1	5007	64005
	5	4856	54665
	10	4568	47229
	20	3811	33936
Rice (Leaf) [143]	1	3882	33945
	5	3579	24089
	10	3254	17850
	20	2619	10522
Rice (Callus) [143]	1	2758	40757
	5	2426	30514
	10	1980	23189
	20	1399	13446

Table A.4: DHS overlap statistics are shown for the four samples in both IR and IE events at the four levels of read coverage. For both IR and IE events, the number of DHS (peaks) overlapping the events is shown at both individual parts (5' exon, Intron, and 3' Exon) and the whole event (shown in the column titled "All"). Finally, the fisher exact test p-value is shown for each case, indicating that the overlap is significant in IR events in contrast to IE events.

Sample	Exp. Level	IR Events				IE Events				Fisher Pval
		5' Exon	Intron	3' Exon	All	5' Exon	Intron	3' Exon	All	
AT (Leaf) [2]	1	61	61	224	346	498	614	1102	2214	9.55E-69
	5	59	61	220	340	384	448	839	1671	9.57E-76
	10	56	59	202	317	288	331	630	1249	1.92E-76
	20	45	51	172	268	173	189	381	743	1.07E-66
AT (Flower) [2]	1	70	70	308	448	475	484	1348	2307	1.44E-78
	5	66	69	300	435	354	363	1121	1838	2.41E-84
	10	63	68	285	416	282	281	946	1509	3.36E-88
	20	56	64	259	379	192	194	653	1039	9.43E-93
Rice (Leaf) [143]	1	66	58	308	432	182	172	413	767	3.00E-143
	5	62	57	297	416	114	96	286	496	6.77E-148
	10	58	57	290	405	75	66	219	360	2.92E-147
	20	44	50	248	342	37	33	130	200	2.29E-123
Rice (Callus) [143]	1	67	61	339	467	404	378	1404	2186	3.43E-112
	5	54	54	319	427	244	191	1030	1465	1.40E-120
	10	42	45	290	377	164	120	766	1050	4.77E-122
	20	34	40	225	299	99	68	466	633	3.61E-104

Table A.5: The HMM's transition probabilities for all 13 states. The probabilities were derived from the training data (8 hexamers that were manually detected to have a footprint). Some of the probabilities were manually tweaked to adjust for the noise in our data. The highlighted probabilities (as described in figure 2) are relatively higher than the other transition from the same state. This is to force our HMM to prioritize detection of the primary footprint.

States	BG_0	UP	FP_S	DN	BG_1	UP	FP_P	DN	BG_2	UP	FP_S	DN	BG_3
BG_0	0.996	0.000037239	0	0	0	0.004	0	0	0	0	0	0	0
UP	0	0.042	0.958	0	0	0	0	0	0	0	0	0	0
FP_S	0	0	0.879	0.121	0	0	0	0	0	0	0	0	0
DN	0	0	0	0.036	0.964	0	0	0	0	0	0	0	0
BG_1	0	0	0	0	0.990	0.010	0	0	0	0	0	0	0
UP	0	0	0	0	0	0.042	0.958	0	0	0	0	0	0
FP_P	0	0	0	0	0	0	0.879	0.121	0	0	0	0	0
DN	0	0	0	0	0	0	0	0.036	0.001	0	0	0	0.963
BG_2	0	0	0	0	0	0	0	0	0.990	0.010	0	0	0
UP	0	0	0	0	0	0	0	0	0	0.042	0.958	0	0
FP_S	0	0	0	0	0	0	0	0	0	0	0.879	0.121	0
DN	0	0	0	0	0	0	0	0	0	0	0	0.036	0.964
BG_3	0	0	0	0	0	0	0	0	0	0	0	0	1

Table A.6: Emissions for the all HMM's 13 states are listed. These emissions are modeled by Gaussian distributions with the corresponding mean and standard deviation (std) shown. Note that these values are derived after standardization of raw hexamer profile coverage to the background score calculated from the training data. The BG_1 and BG_2 (intermediary/secondary backgrounds) were calculated (and tweaked) based on the measured BG_0 and BG_3 values (somewhere in between the two).

States	**Arabidopsis**		**Rice**	
	mean	***std***	***mean***	***std***
BG_0	0.370681027	0.145492001	0.530290538	0.231679565
UP	-1.103074051	0.903519929	-1.149144567	0.856341604
FP_S	-2.726769331	0.034938222	-2.638014215	0.03536976
DN	-1.359386756	0.969790822	-1.290317173	0.832150009
BG_1	0.490681027	0.145492001	0.480290538	0.231679565
UP	-1.103074051	0.903519929	-1.149144567	0.856341604
FP_P	-2.726769331	0.034938222	-2.638014215	0.03536976
DN	-1.359386756	0.969790822	-1.290317173	0.832150009
BG_2	0.490681027	0.145492001	0.480290538	0.231679565
UP	-1.103074051	0.903519929	-1.149144567	0.856341604
FP_S	-2.726769331	0.034938222	-2.638014215	0.03536976
DN	-1.359386756	0.969790822	-1.290317173	0.832150009
BG_3	0.629707395	0.228739766	0.476770207	0.27267674

Table A.7: The overlap stats between all significantly enriched arabidopsis IR/IE hexamers and transcription factor motifs from Plant Cistrome Database are summarized below. The actual overlaps are provided in the Additional file 3.

AS Type	**Part**	**Total Motifs**	**Total Hexamers**	**Overlapping Hexamers**
IR	5' Exon	410	13	6
	Intron		2	1
	3' Exon		246	80
IE	Intron		19	9

Appendix B

Chapter 6 Supplementary Material

B.1 Generating transcription factor family distributions

To generate figure 1 in the main text, we used the original Basset [13] with all DHSs (2 million) across 164 human cell lines. The dataset was split into 80%, 10%, and 10% for training, validating, and testing the model. Once the model was trained, we followed the motif analysis pipeline, as described in [13]. We analyzed first CNN layer filters to generate motifs for three different test sets: DHSs that overlapped the human promoter, intragenic, and intergenic regions, separately. Next, for each CNN filter, we inferred enrichment by counting all of its activations across the sequences coming from one of the three aforementioned regions, separately. Finally, the enriched CNN motifs in each set were mapped to human CISBP database [178] using TomTom tool from MEME suite [202]. In the final figure, we only used families of those TF motifs which had a significant match with adjusted p-value < 0.05.

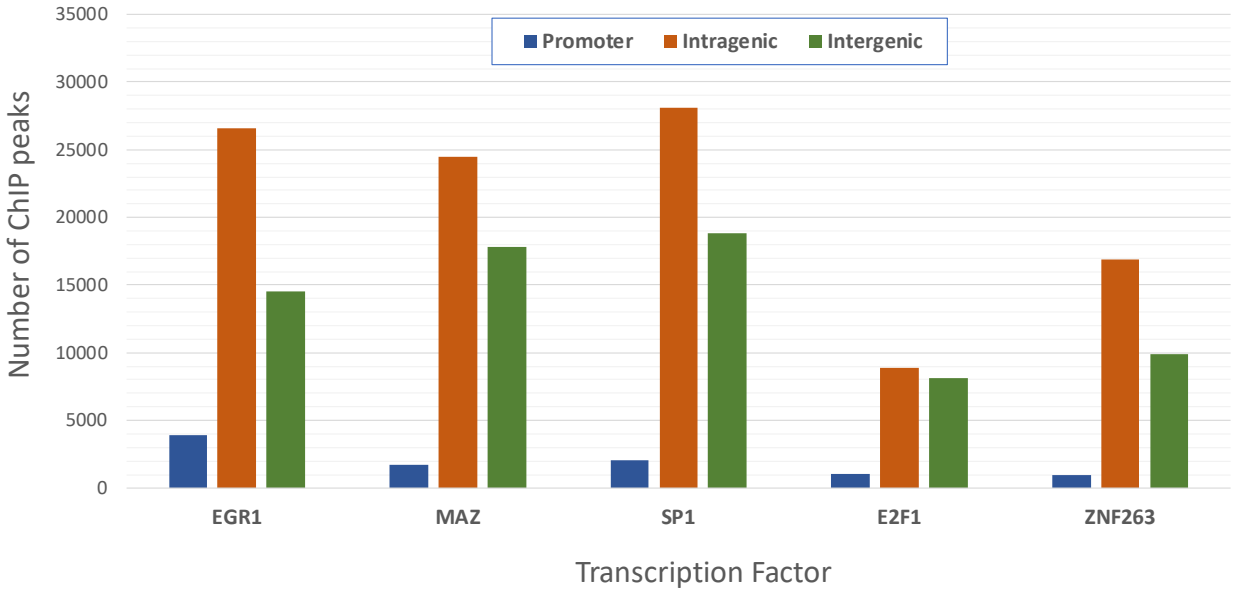


Figure B.1: The distribution of different transcription factor ChIP-Seq peaks in the promoter, intragenic, and intergenic regions of the human genome. The ChIP-Seq peaks for the corresponding TFs were downloaded from the ENCODE database [14].

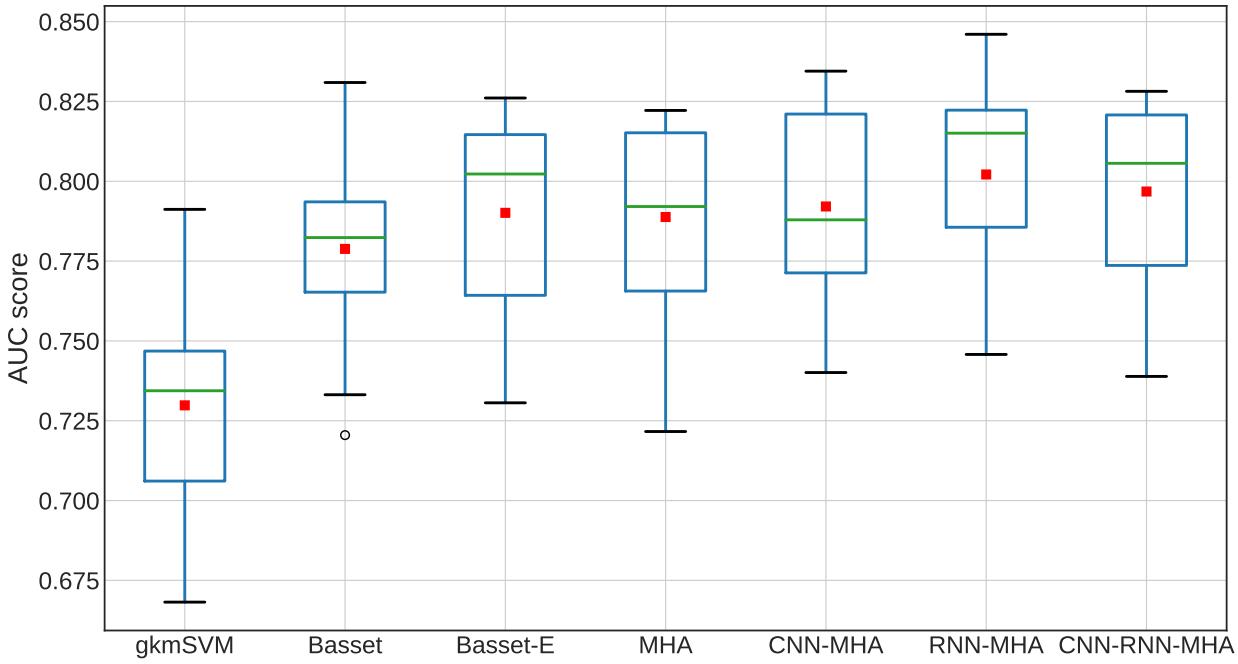


Figure B.2: AUC box and whiskers plot is shown for the different network architectures and gapped kmer SVM, using the leave-one-chromosome-out strategy. For each model, the green line in the box represents median AUC across the 22 chromosomes whereas average AUC value is represented by the red marker. All deep learning methods use embedded representation of the input except Basset.

Table B.1: List of neural network hyperparameters.

Hyperparameter	type	Description
use_embd	bool	Whether to use the word2vec embeddings [default: False]
embd_size	int	Size of the word2vec embedding vectors [default: 50]
embd_window	int	Size of the word2vec embedding window [default: 5]
embd_kmer	int	Length of the kmer (for word2vec embeddings) [default: 3]
singlehead_size	int	Size of the attention single head [default: 32]
num_heads	int	Number of heads in multi-head self-attention layer [default: 8]
multihead_size	int	Output size of the multi-head after concatenation [default: 100]
batch_size	int	Batch size in training/testing the model [default: 172]
use_RNN	bool	Whether to use the RNN layer. [default: based on model variant]
RNN_hidden_size	int	Size of the RNN layer. [default: 100]
CNN_filters	int	Number of CNN filters to use. [default: 200]
CNN_filter_size	int	Size of each CNN filter. [default: 13]
use_CNN_pool	bool	Use max pooling in the CNN layer. [default: True]
CNN_pool_size	int	Size of the max pooling window in CNN layer. [default: 6]
input_channels	int	Number of input channels. [default: 4 (for DNA sequences)]
num_epochs	int	Number of training epochs. [default: 30]
readout_strategy	string	Normalize the MHA output or flatten it. [default: "normalize"]

Appendix C

Chapter 7 Supplementary Material

C.1 Data collection and processing

C.1.1 Experiment 1: simulated dataset

In this experiment, we simulated DNA sequences using random sampling from a distribution of [0.27, 0.23, 0.23, 0.27] for A, C, G, and T respectively as used for a similar dataset generated by Greenside et al. [193]. We generated 120,000 sequences each with a length of 200 bp. Similar to Greenside et al. [193], we randomly embedded instances of the motifs of both ELF1 and SIX5 transcription factors in 40,000 of the total sequences. This was our positive set of examples where we essentially simulated interactions between the aforementioned motifs. In the negative set (80,000 sequences), we embedded instances of either ELF1 or SIX5 in a sequence (but not both). Moreover, we embedded instances of the AP1 and TAL1 motifs across all examples. The motifs for the four transcription factors were obtained from Kheradpour et al. [203].

TF Database information

To map CNN filters to motifs of known TFs, we used TomTom with a custom TF database (MEME format) containing PWMs of the four transcription factors: SIX5, ELF1, AP1, and TAL1.

C.1.2 Experiment 2: TAL-GATA ChIP-peaks

Here we followed the same strategy described in DFIM [193]: ChIP-Seq peaks were downloaded for the three TFs TAL1, GATA1, and GATA2 from the ENCODE [14] database in the K562 cell line (hg19 genome assembly and annotations). For the chromatin accessibility data, we downloaded processed DNase I Hypersensitive Sites (DHSs) from the ENCODE database for the corresponding cell line. Next, every ChIP-Seq peak for the three transcription factors was searched for an overlap with DHSs in the K562 cell line. If an overlap was found, the sequence of the ChIP-

Seq peak was extended 500 bp upstream and downstream from its center. This served as a positive set in our binary classification problem. For the negative set, we randomly sampled 80,000 examples from all K562 DHSs that didn't overlap a ChIP-Seq peak for any of the three transcription factors.

TF Database information

In this experiment, we used two TF databases: the first one was a custom motif file with PWMs of TAL1, GATA1, and GATA2 transcription factors. This was because we wanted to directly compare our model to DFIM [193] where Greenside et al. measured interactions between the aforementioned transcription factors. The second reference was the entire CISBP TF database [178] that we used in order to infer other TF interactions within the ChIP-Seq peaks.

C.1.3 Experiment 3: human promoter DHSs

In this experiment, we used DHSs overlapping gene promoter regions across the entire human genome. We used the pipeline described by Kelley et al. in Basset [13]: DHSs were downloaded for 164 human immortalized cell lines from the ENCODE [14] and ROADMAP [173] consortia. These regions of open chromatin were merged if they overlapped more than 200 bp. Finally, every DHS was extended to a length of 600 bp around its center. Kelley et al. [13] used DHSs across the entire genome however, we selected only those which overlapped the human promoter regions. To do that, we defined promoter as a region of 1000 bp upstream of the transcription start site (TSS) of a gene—Ensemble based hg19/GrCh37 reference and annotations were used. The final dataset had 20,613 genomic sequences of the corresponding DHSs (that overlapped the human promoters). The targets in this case were either a single or multiple labels, corresponding to the164 cell lines in which the DHSs were observed.

TF Database information

In motif analysis (and later in the TF interactions), we used the human CISBP transcription factor database [178].

C.1.4 Experiment 4: genome-wide arabidopsis regions of open chromatin

Here we designed a similar experiment as described in the previous section. The dataset was constructed using the same procedure described above as used by Kelley et al. [13]. We used regions of open chromatin: DHSs and ATAC-Seq based Transposase Hypersensitive Sites (THSs), across the entire arabidopsis genome using TAIR10 annotations. We used the following publicly available datasets (GEO accession numbers provided):

- For DHSs: GSE53322, GSE53324, GSE53323, GSE46987, GSE34318
- For THSs: GSE89346, GSE85203, GSE101940, GSE116287, GSE101482

We ended up with 88,245 examples in our final dataset across 36 different samples. Note that peaks occurring in multiple biological samples were merged.

TF Database information

Here we used the DAP-Seq based arabidopsis transcription factors database [196].

C.2 Limitations of the TomTom motif comparison tool

We used the TomTom tool from the MEME suite [202] to map CNN filters to motifs of known transcription factors. In some cases, we observed the match to be dubious despite the tool assigning it a significant p-value. This is shown in Supplementary Figure F4 for two of our CNN filters matching the known TF motifs of HOXA2 and ZNF263 in the human CISBP database [178]. By default, TomTom uses Pearson correlation for comparing motifs. However, we obtained better results using the Euclidean distance.

C.3 Additional tables

Table C.7: A list of significant interactions in the TAL-GATA ChIP-Seq.

(available: https://github.com/fahadahaf/SATORI/blob/master/process_results/tables/C7.xlsx)

Table C.8: A list of significant interactions in the human promoters.

(available: https://github.com/fahadahaf/SATORI/blob/master/process_results/tables/C8.xlsx)

Table C.9: A list of significant interactions in the intron retention

(available: https://github.com/fahadahaf/SATORI/blob/master/process_results/tables/C9.xlsx)

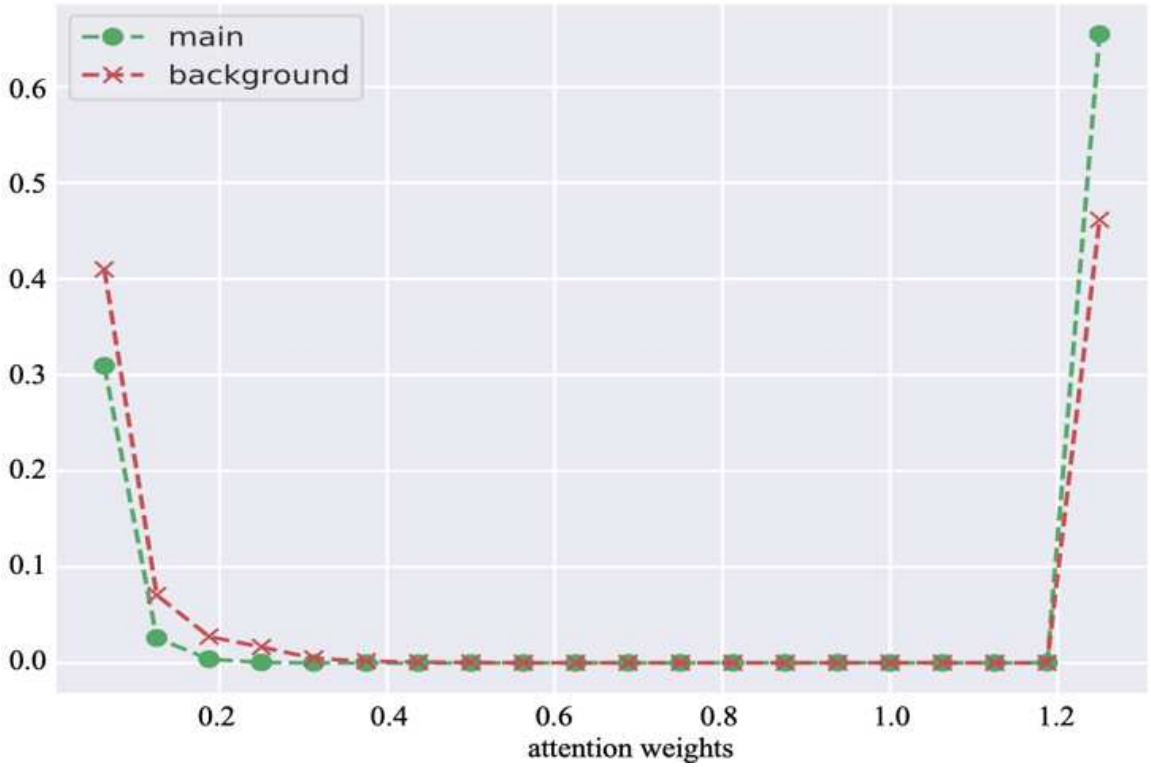


Figure C.1: Distribution of the attention weights for the main test and the background sets. The actual frequencies (y-axis) are normalized by total sizes of the test and background sets. This figure helps in selecting the appropriate attention cutoff, one of the parameters of SATORI. We use a default value of 0.10.

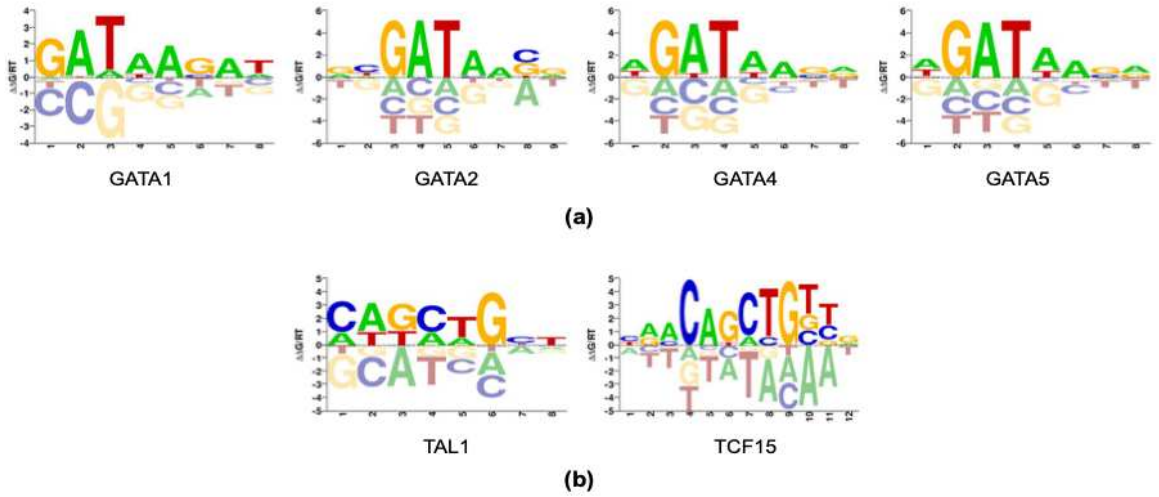


Figure C.2: Similarities between motifs of GATA variants (a). Similarly, TAL1 and TCF15, both belonging to the bHLH family, have very similar motifs (CAGCTG consensus) (b).

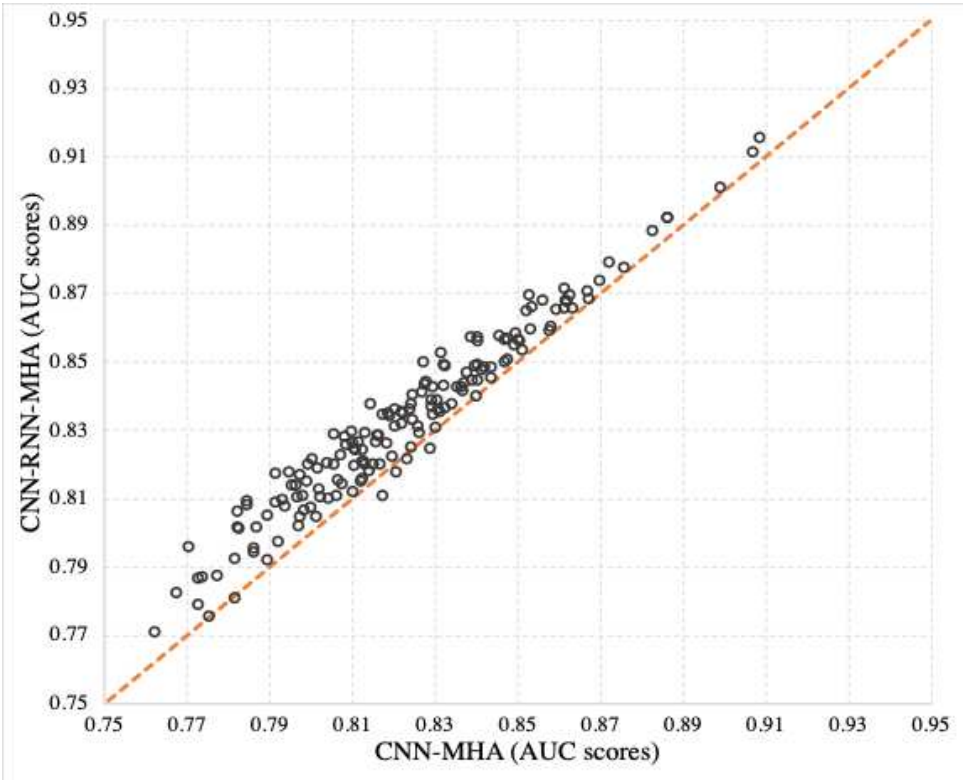


Figure C.3: AUC scores for DHSs in human promoters across 164 cell types, achieved by the two model variants. Each circle represents performance on detecting DHSs in that cell line.

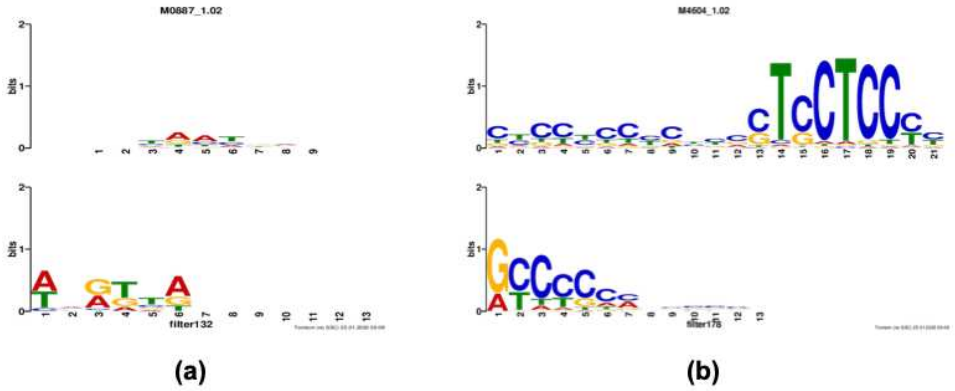


Figure C.4: Limitations of the TomTom motif comparison tool. Matches shown here are statistically significant (q-value < 0.01) for both (a) HOXA2 and (b) ZNF263. The top row depicts the gold standard motif in the CISBP database and the bottom row shows the CNN filter/motif. .

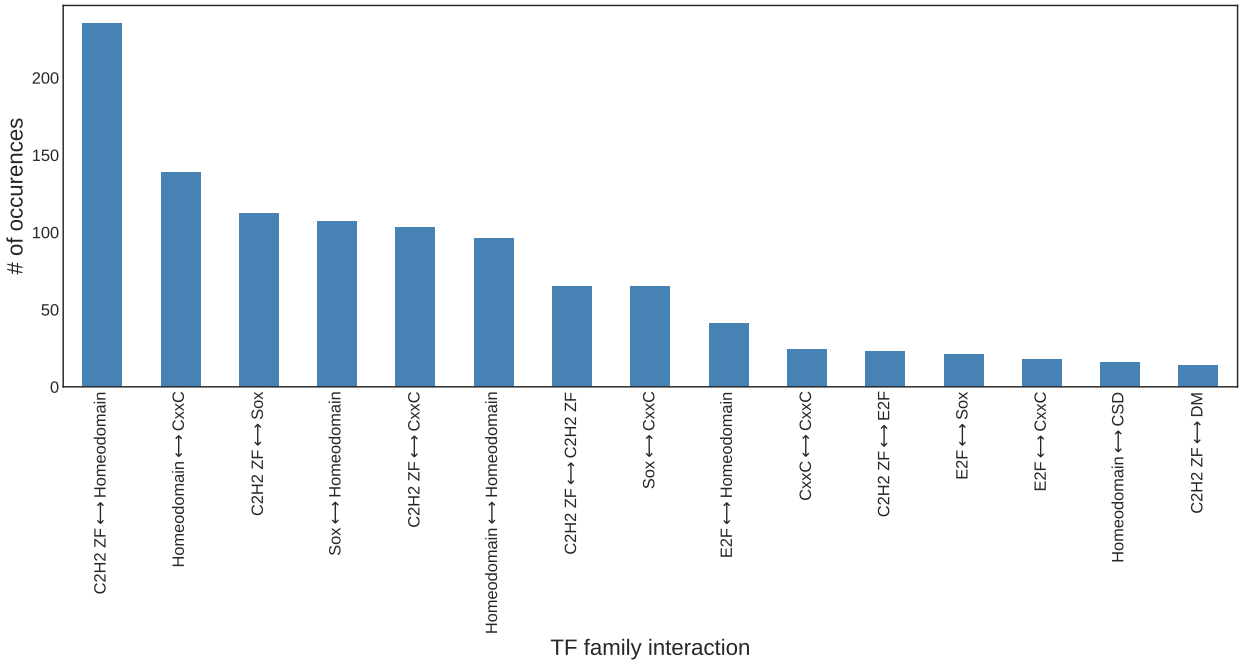


Figure C.5: The most frequent interacting transcription factor families in human promoter regions.

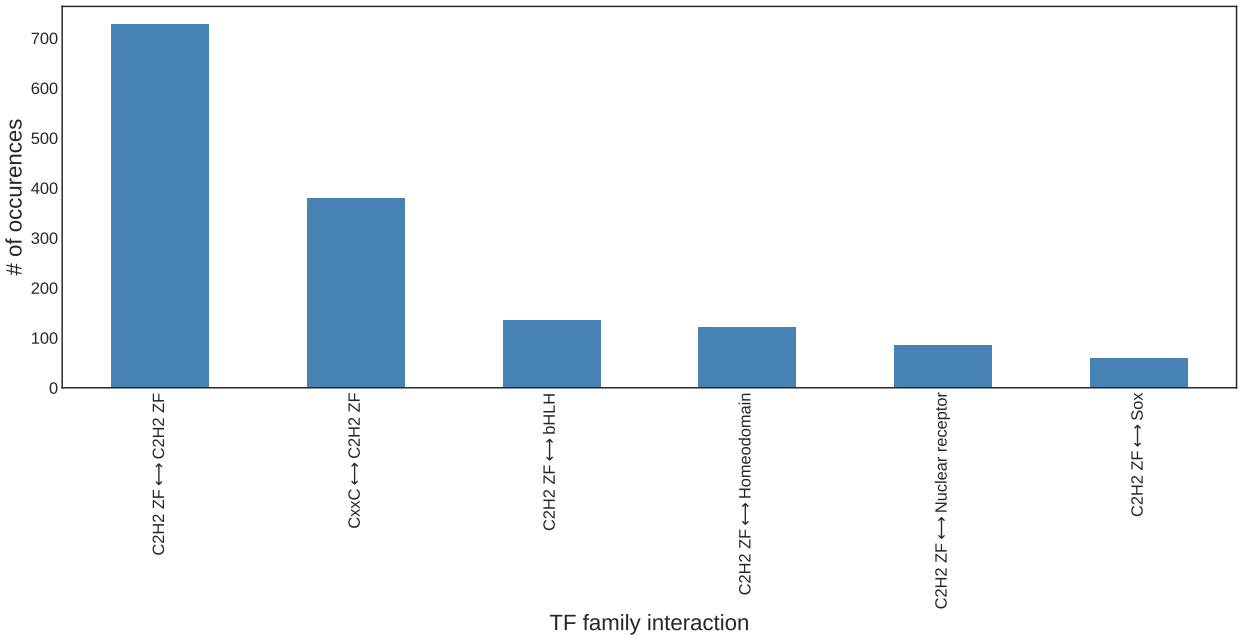


Figure C.6: The most frequent interacting transcription factor families in the intron retention events.

Table C.1: List of neural network hyperparameters.

Hyperparameter	type	Description
singlehead_size	int	Size of the attention head [default: 32]
num_heads	int	Number of heads in multi-head self-attention layer [default: 8]
multihead_size	int	Output size of the multi-head after concatenation [default: 100]
batch_size	int	Batch size in training/testing the model [default: 172]
use_RNN	bool	Whether to use the RNN layer. [default: based on model variant]
RNN_hidden_size	int	Size of the RNN layer. [default: 100]
CNN_filters	int	Number of CNN filters to use. [default: 200]
CNN_filter_size	int	Size of each CNN filter. [default: 13]
use_CNN_pool	bool	Use max pooling in the CNN layer. [default: True]
CNN_pool_size	int	Size of the max pooling window in CNN layer. [default: 6]
input_channels	int	Number of input channels. [default: 4 (for DNA sequences)]
num_epochs	int	Number of training epochs. [default: 30]
readout_strategy	string	Normalize or flatten the MHA output. [default: "normalize"]

Table C.2: Summary of all significant interactions in the simulated/toy dataset. Our model is able to recover multiple interactions involving SIX5 and ELF1 TF motifs. We also provide the actual CNN filter interactions in the first column, named based on the total number of filters in the convolutional layer.

Filter interaction	TF motif interaction	Adjusted p-value
filter033⟷filter170	SIX5⟷ELF1	1.47E-42
filter170⟷filter181	ELF1⟷SIX5	1.50E-39
filter033⟷filter111	SIX5⟷ELF1	1.12E-20
filter111⟷filter181	ELF1⟷SIX5	3.17E-22
filter033⟷filter091	SIX5⟷ELF1	2.44E-18
filter091⟷filter181	ELF1⟷SIX5	2.14E-12
filter033⟷filter055	SIX5⟷ELF1	2.64E-08
filter055⟷filter181	ELF1⟷SIX5	3.84E-04
filter019⟷filter033	ELF1⟷SIX5	1.35E-07
filter019⟷filter181	ELF1⟷SIX5	2.48E-12
filter091⟷filter199	ELF1⟷SIX5	2.82E-03

Table C.3: All significant interactions between TAL1 and GATA transcription factors. Note that in this case, a custom TF database was used containing motifs for TAL1, GATA1, and GATA2. In case of TAL1, other TFs (LYL1, NHLH2, and TAL2) also shared the same binding site motif and hence are mentioned here.

TF A	TF B	Adjusted p-value
GATA2	LYL1,**TAL1**,NHLH2,TAL2	4.37E-24
GATA2	LYL1,**TAL1**,NHLH2,TAL2	8.18E-21
LYL1,**TAL1**,NHLH2,TAL2	GATA2	5.71E-19
LYL1,**TAL1**,NHLH2,TAL2	GATA2	1.54E-16
LYL1,**TAL1**,NHLH2,TAL2	GATA2	8.68E-11
GATA2	LYL1,**TAL1**,NHLH2,TAL2	3.92E-10
LYL1,**TAL1**,NHLH2,TAL2	GATA2	3.27E-08
LYL1,**TAL1**,NHLH2,TAL2	GATA2	3.28E-08
GATA2	LYL1,**TAL1**,NHLH2,TAL2	5.66E-05
LYL1,**TAL1**,NHLH2,TAL2	GATA2	6.76E-04
LYL1,**TAL1**,NHLH2,TAL2	GATA2	4.51E-03
LYL1,**TAL1**,NHLH2,TAL2	GATA2	4.66E-03
LYL1,**TAL1**,NHLH2,TAL2	GATA2	4.72E-03
LYL1,**TAL1**,NHLH2,TAL2	GATA2	5.41E-03

Table C.4: A list of known TF interactions identified by our model in the human promoter regions. TRRUSTv2 [3] database was used as a reference of all known interactions. The level of significance (adjusted p-value) assigned by our model to each interaction is provided in the last column.

motif interaction	TF1 family	TF2 family	adjusted p-value
EGR1⟷LCOR	C2H2 ZF	Pipsqueak	1.84E-27
E2F1⟷LCOR	E2F	Pipsqueak	1.32E-23
DNMT1⟷LCOR	CxxC	Pipsqueak	1.72E-16
EGR1⟷SRF	C2H2 ZF	MADS box	1.75E-14
SRF⟷SP2	MADS box	C2H2 ZF	4.12E-10
E2F1⟷SRF	E2F	MADS box	5.96E-09
EGR1⟷E2F4	C2H2 ZF	E2F	9.64E-07
E2F1⟷DNMT1	E2F	CxxC	4.36E-05
E2F1⟷E2F4	E2F	E2F	1.64E-04
DNMT1⟷E2F4	CxxC	E2F	1.64E-04
E2F1⟷EGR1	E2F	C2H2 ZF	1.99E-03
DNMT1⟷EGR1	CxxC	C2H2 ZF	2.33E-03

Table C.5: Summary of the number of unique statistically significant TF interactions reported by SATORI and FIS for the three real-world datasets.

Experiment	SATORI	FIS
TAL-GATA ChIP-Seq	152	235
Human promoters	234	184
Arabidopsis genome-wide	230	224

Table C.6: A list of known TF interactions in the IR events. TRRUSTv2 [3] database was used as a reference of all known interactions. The level of significance (adjusted p-value) assigned by SATORI to each interaction is provided in the last column.

motif interaction	TF1 family	TF2 family	adjusted p-value
DNMT1⟷E2F1	E2F	E2F	4.15E-04
DNMT1⟷EGR1	CxxC	CxxC	4.94E-02
DNMT1⟷ESR1	Nuclear receptor	Nuclear receptor	5.55E-03
DNMT1⟷SP4	CxxC	CxxC	6.19E-04
E2F1⟷ESR1	E2F	Nuclear receptor	8.25E-07
E2F1⟷SP4	E2F	C2H2 ZF	1.47E-10
EGR1⟷E2F1	E2F	E2F	6.54E-04
EGR1⟷ESR1	C2H2 ZF	Nuclear receptor	5.49E-05
EGR1⟷MAZ	C2H2 ZF	C2H2 ZF	2.65E-03
EGR1⟷RREB1	C2H2 ZF	C2H2 ZF	4.25E-02
EGR1⟷SP4	C2H2 ZF	C2H2 ZF	3.27E-03
ESR1⟷RARG	Nuclear receptor	Nuclear receptor	1.67E-04
ESR1⟷SP4	Nuclear receptor	C2H2 ZF	7.72E-10
MAZ⟷SP2	C2H2 ZF	C2H2 ZF	2.06E-04

Appendix D

Chapter 8 Supplementary Material

D.1 Methods: Differential AS and chromatin accessibility

D.1.1 Data collection and processing

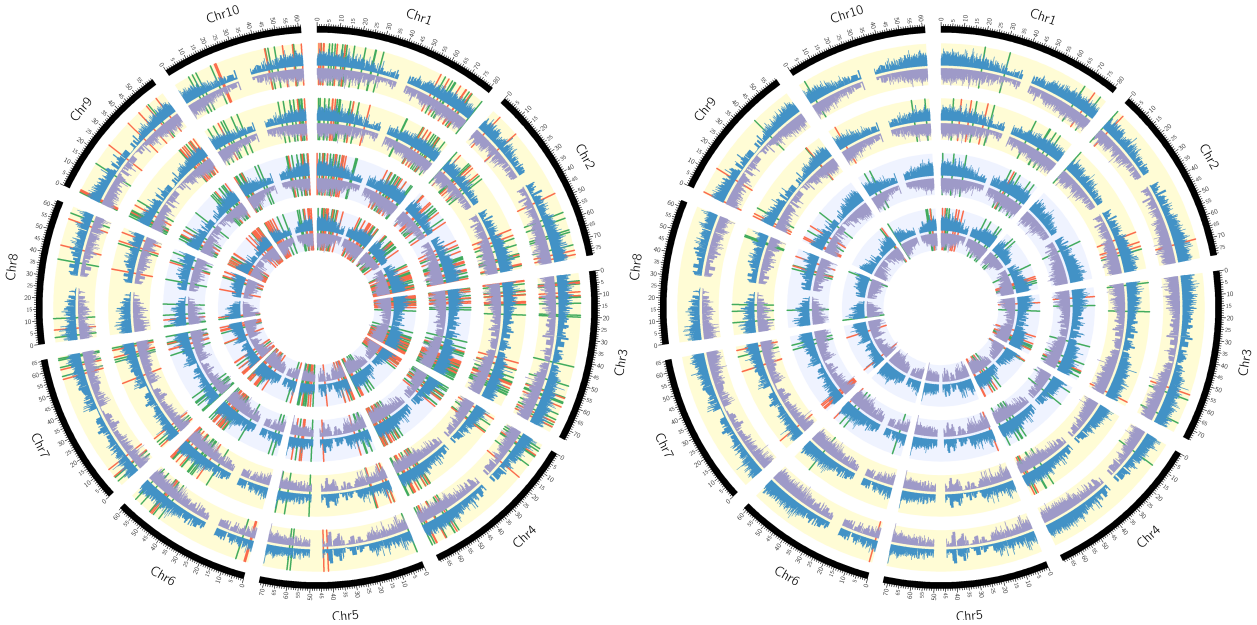
We downloaded the processed RNA-Seq and DNase I-Seq peaks (DHSs) for the *K562* and the other eight cell lines from Encode database [14]. In case of RNA-Seq data, every cell type had at least two biological replicates. We noticed that the library sizes varied significantly across cell types: the smallest library had around 70 million reads where the largest one had over 250 million reads. To adjust for the variation, we used Sambamba [204] to sample reads from every library such that they all had roughly the same sizes.

In case of arabidopsis, raw RNA-Seq reads and processed ATAC-Seq peaks (THSs) from [205] were used(GEO accession number *GSE116287*). The data consisted of multiple biological replicates for arabidopsis root and shoot tissues under control and treatment conditions. The ATAC-Seq peaks were merged across biological replicates. In case of RNA-Seq, the reads were first pre-processed using FastQC [159] and trimmed using fastx-trimmer [160]. Reads were aligned to the TAIR10 reference genome using STAR [80] with parameter `outFilterMultimapNmax 1` to get uniquely aligned reads.

D.1.2 Differential IR and chromatin accessibility analysis

To get differential IR events, we used idiffIR [206] using the default parameters. In case of human data, $K562$ was compared against the rest of the eight cell-lines. Since the data was aligned using hg19 reference genome, we used the corresponding genome annotations with idiffIR. In case of arabidopsis, the treated samples were compared to the control samples in both root and shoot tissues. Finally, we selected an adjusted p-value cutoff of 0.05 for a differential IR event to be used in the downstream analysis.

Next, we used the chromatin accessibility peaks (DNase I-Seq in case of human and ATAC-Seq in case of arabidopsis) to analyze their occupancy in the differential IR events. For each event, we checked if a given peak overlapped its coordinates: from the start of the upstream exon to the end of the downstream exon. We checked that using peaks from both control and treated samples. A stringent requirement was used for differential peak occurrence. For instance, for a differential IR event to qualify, we required that it must had overlapping peak(s) from only one of the two conditions. To check the significance of overlap between differential IR events and differentially occurring peaks, we used fisher exact test. Finally, to visualize the events and overlapping peaks, Integrated Genome Viewer was used [6].



(a) IR (b) ES

Figure D.1: Genome-wide differential AS in four lines (cultivars) of sorghum is shown in (a) for IR events and (b) for ES events. Each slice represents one of the 10 sorghum chromosomes. From the center, the first four concentric circles represent sorghum lines 1, 2, 3, and 7 respectively. The outer most circle shows the genomic coordinates with a step size of 10 million bp. The gene expression levels are shown by purple and blue coverage plots for the treated and control samples, respectively. Finally, across the coverage plots, the up-regulate and down-regulated differential IR events are marked by green and red lines, respectively. This figure is generated using CIRCOS [15].

Table D.1: The overlap between differential IR events with the differentially occuring DHSs in *K562* vs. eight other human cell-lines. The significance of overlap is shown in the last column in terms of p-value (Fisher test).

K562 vs.	Total events	Events with Differential DHSs	Fisher p-value
GM12878	191	60	9.32×10^{-06}
H1-hESC	163	42	0.032164027
HCT-116	137	29	0.00098362
HepG2	119	23	0.065757989
HSMM	77	15	0.236222736
MCF-7	145	35	0.001566634
NHEK	82	24	0.012541207
NHLF	88	22	0.008227137