Cell Type Prediction of Transcription Factor Binding Sites using Machine Learning

Faizy Ahsan

Master of Science

School of Computer Science

McGill University

Montreal, Quebec

2015-06-12

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

This document is dedicated to my parents Rukhsana and Abdul Qaiyum.

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof. Mathieu Blanchette for introducing the wonderful field of bioinformatics. Also, I express my sincere gratitude to Prof. Doina Precup for providing machine learning concepts. I am blessed to have both of them as supervisors, they have always motivated and encouraged me through out the master program.

I am grateful to fellow bioinformatics and reasoning & learning lab members, especially Ayrin A. Tabibi, Rola Dali, Jimmy H.C. Chao, Vladimir Reinharz, Christopher J. F. Cameron and Emmanuel Bengio, for providing an enjoyable learning environment and to McGill university and its staff members for providing a comfortable stay.

Finally, I thank my parents and brothers Simaab Amir and Kashif Rizvee for their spirtual supports.

ABSTRACT

The cell type specific binding of transcription factors is known to contribute to gene regulation, resulting in the distinct functional behaviour of different cell types. The genome-wide prediction of cell type specific binding sites of transcription factors is crucial to understanding the genes regulatory network. In this thesis, a machine learning approach is developed to predict the particular cell type where a given transcription factor can bind a DNA sequence. The learning models are trained on the DNA sequences provided from the publicly available ChIP-seq experiments of the ENCODE project for 52 transcription factors across the GM12878, K562, HeLa, H1-hESC and HepG2 cell lines. Three different feature extraction methods are used based on k-mer representations, counts of known motifs and a new model called the skip-gram model, which has become very popular in the analysis of text. Three different learning algorithms are explored using these features, including support vector machines, logistic regression and k nearest neighbor classification.

We achieve a mean AUC score of 0.82 over the experiments for the best classifier and feature extraction combination. The learned models, in general, performed better for the pair of cell types that have a relatively large number of cell type specific transcription factor binding sites. We find that logistic regression and known motifs based combination detect cell-type specific signatures better than a previously published method with mean AUC improvement of 0.18 and can be used to identify the interaction of transcription factors.

ABRÉGÉ

Les facteurs de transcription spécifiques aux types de cellules contribuent à la régulation de l'expression des gènes, contribuant ainsi à des comportements distincts, propres à chaque type de cellule. La prédiction des sites de fixations des facteurs de transcription spécifiques aux types de cellules à travers le génome est cruciale pour comprendre le réseau régulatoire génétique. Dans cette thèse, une approche d'apprentissage machine est utilisée pour prédire le dans quel type cellulaire une région dADN donnée sera liée par un facteur de transcription spécifique. Les modèles d'apprentissage sont entrainés sur des séquences d'ADN provenant d'expériences Chip-Seq disponibles publiquement du projet ENCODE, pour 52 facteurs de transcription qu'on retrouve dans les types de cellules GM12878, K562, HeLa, H1-hESC et HepG2. Trois différentes méthodes d'extraction de caractéristiques sont utilisées, basées sur les représentations par k-mer, le comptage de motifs connus, ainsi qu'un nouveau modèle appellé *skip-gram*, qui est devenu très populaire dans l'analyse de texte. Trois algorithmes d'apprentissage sont explorés en utilisant ces méthodes d'extraction, notamment les machines à vecteurs de support, la régression logistique, ainsi que la classification par les k plus proches voisins.

Nous atteignons un score AUC de 0.82 pour les expériences de classification et d'extraction de caractéristiques. Les modèles appris, en général, performent mieux pour les paires de types cellulaires qui ont un nombre relativement élevé de sites de fixations des facteurs de transcription spécifiques à leur type. Nous trouvons que la

régression logistique combinée au comptage de motifs connus détecte mieux les signatures spécifiques aux types de cellules que des méthodes publiées précédemment avec une amélioration moyenne de l'AUC de 0.18. Finalement, cette méthode peut être utilisée pour identifier l'interaction des facteurs de transcription.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

The functional machinery of cells in an organism is encoded as genes present in the deoxyribonucleic acid (DNA) sequences. Genes store the biological information of all known living organisms and viruses in the form of genetic instructions. The gene products, called proteins, are synthesized from these instructions through the process of gene expression. The resulting gene products such as enzymes, hormones, receptors have important roles in the functioning of living organisms. Gene expression is a series of steps that include transcription, RNA splicing, translation and post-translational modification of a protein. In the transcription stage, a gene is transcribed from a DNA sequence when a particular type of protein, known as transcription factor, binds to a specific DNA sequence called the binding site. An interesting and intriguing phenomenon is that the same transcription factor might be responsible for the expression of different genes in different cell types, even though every cell type has essentially the same DNA sequence.

In this study, we explore the differential binding behavior of transcription factors through machine learning approaches to predict the cell type where a transcription factor can bind in the human genome. This would be useful in understanding the regulatory behavior of transcription factors and about the diseases that are associated with genetics disorders [67, 110].

1.1 Thesis Outline

In this thesis, chapter 1 gives the biological overview of transcription process and highlights the contributions made by this study. Chapter 2 reviews experimental and computational approaches to the prediction of TFBSs. Chapter 3 provides an overview of machine learning approaches and related works. Chapter 4 describes the methods developed in this study. Chapter 5 reports the data used for the experiments and the results. Chapter 6 concludes our study and discusses some possible extensions to the current work. Appendix A shows the learning parameter values and the data we use, Appendix B details the top ten and bottom ten AUC scores of our learned models. Appendix C lists the TF-TF interactions identified from our experiments.

1.2 Biology of Transcription

Transcription is the process of transcribing a particular DNA segment into a ribonucleic acid (RNA) molecule that has a major role in the synthesis, regulation and processing of proteins. The DNA segment being transcribed has a coding sequence that is used for transcribing RNA, and several regulatory sequences that help to direct and regulate the transcription. The regulatory sequences include several regions of the DNA, such as the promoter, enhancer and silencer.

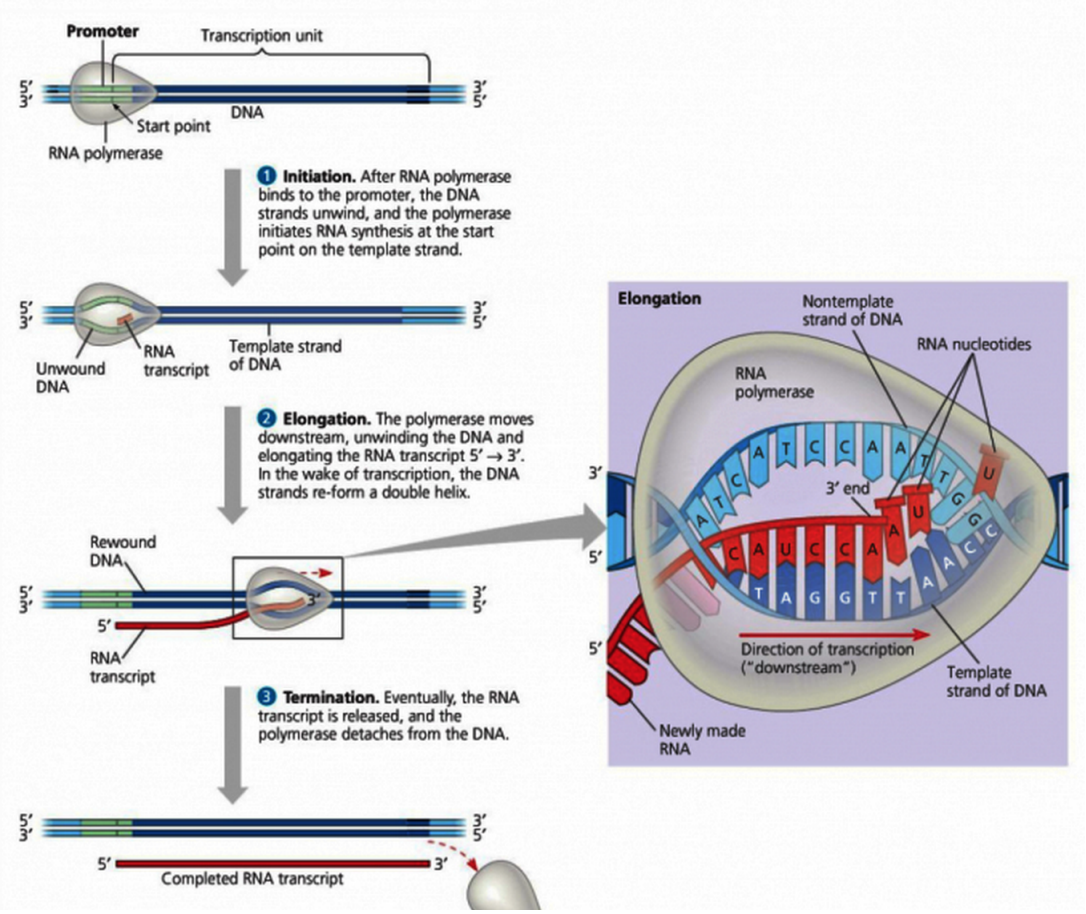


Figure 1–1: Description of transcription process. Figure is from § 7.3.3 of Tokresource [4]

The transcription start site (TSS) is the genomic region where the transcription process starts. The promoter region of a gene is located near the TSS and its length is in the range of 100-1000 base pairs [90]. The binding of particular transcription factors to the promoter region initiates the transcription process. The bound transcription factors help to bind the RNA polymerase enzyme to the promoter region, which creates a transcription bubble. Until the termination, the transcription bubble

moves along the DNA strand while dividing it and simultaneously forming the RNA strand (Fig 1-1).

A gene may have many enhancers and silencers. The enhancer region of the DNA is usually 50-1500 base pairs long [93]. It can be located in either direction from the TSS and it can be as far as 1000,000 base pairs away from the TSS [93]. The binding of some transcription factors to the enhancer region can increase the transcription rate [55]. The length and location of the silencer region of the DNA is similar to that of the enhancer region [102]. The binding of some transcription factors to the silencer region can decrease the transcription rate [74].

1.3 Biology of Transcription Factors

A protein that binds to a specific DNA sequence and regulates the transcription process is known as a transcription factor (TF). A TF binds to specific DNA sequences (such as promoters or enhancers) adjacent to the gene that they regulate. A study by Vaquerizas *et al.* [134] suggests that there are as many as 1700-1900 TFs in human genome. It should be noted that the number of genes observed in human genome are of one order of magnitude higher ($\sim$ 20,000 [94]) than the number of TFs. The reason is because its not just one TF, but several TFs that work together for the production of a gene.

1.3.1 Synthesis

Transcription factors (TFs) are transcribed from a DNA segment to an RNA molecule and then translated from the RNA molecule to a protein. In eukaryotes, the TFs are translated in the cell's cytoplasm. In order to regulate the transcription

process, the TFs need to be in the nucleus, where the DNA is located. The nuclear localization signals from many other active proteins direct them to the nucleus [135].

1.3.2 Functioning

A TF can be in an active or an inactive state. An active TF can take part in gene regulation and can even regulate itself by regulating the gene that encodes it. A TF has a signal sensing domain that can activate or deactivate it in several ways. For example, ligand binding can activate a TF. There are many TFs that need to be phosphorylated in order to bind with DNA sequences [21]. The interaction with other TFs and cofactors can activate a TF [84, 51].

1.3.3 Binding Domain

A TF has two main binding domains : the DNA-binding domain (DBD) and the trans-activating domain (TAD). The DBD is used to bind with specific DNA sequences such as promoter or enhancer. The TAD is used to bind with other proteins such as coactivators or corepressors.

1.4 Biology of Transcription Factor Binding Sites

Transcription factor binding sites (TFBSs) are the specific DNA sequences to which a TF binds. TFBSs are typically 5 to 15 base pairs long [25]. In most cases, a TF binds to a TFBS in a sequence specific way and may not be binding to all the bases of the TFBS. The binding strength of a TF with the different bases of a TFBS can differ. A TF can bind with a subset of closely related sequences. Thus, a TFBS can be degenerate i.e. the nucleotide occurring at a particular position in the TFBS is not fixed. It has been observed that TFs binding to DNA is highly clustered and TFBSs of many TFs are present in relatively same genomic regions

[137]. Due to these complexities, prediction of TFBSs is an active research area and numerous experimental and computational methods have been developed for it, which we review in Ch. 2 and § 3.4.

1.5 Cell Type Specificity of Transcription Factor Binding Sites

It is well known that a TF binds to a pre-determined pattern of a specific length in the TFBS, known as the motif. However, TF binding to a motif can be cell type specific i.e. a TF can bind to a particular genomic region in one cell type but not in the other. There are many factors that influence the binding of a TF to a motif in a cell type, such as:

1. Presence of cofactors such as coactivators and corepressors, which play role in the regulatory function of a TF [112, 111].

2. Cooperative binding with other TFs [85, 91, 131, 68].

3. Regulatory regions of DNA being accessible or inaccessible to TFs [97, 12].

4. Competition with other TFs and proteins [88, 89, 119].

5. Presence of pioneer TFs, which are capable of binding with nucleosomal DNA that are generally inaccessible and can make DNA accessible to other TFs [52, 14].

6. Thermal fluctuations, which can partially unwrap the nucleosome and make DNA temporarily accessible to TFs [34].

7. Histone methylation, i.e. the addition of methyl groups to histones; depending upon the number and the location of methyl groups attached, histone methylation can either increase or decrease the rate of binding of TFs. [71]

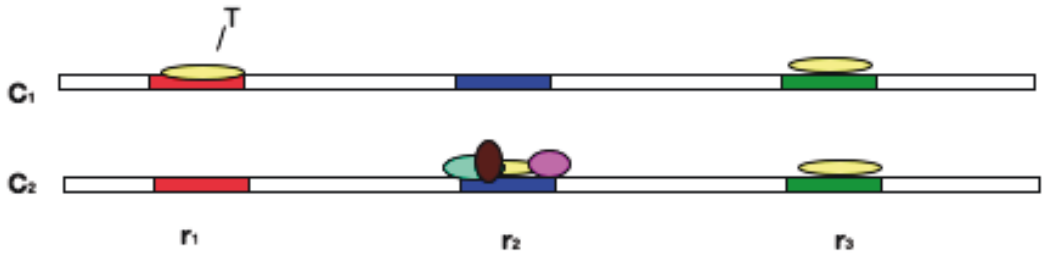


Figure 1–2: TF T selectively binds with region r_1, r_2 in cell type C_1, C_2, and non-selectively binds with region r_3 in both cell types. The regions r_1 and r_2 are known as cell type specific TFBS. The region r_3 is known as constitutive TFBS. In cell type C_2, the selective presence of other TFs might have helped T to bind with r_2.

1.6 Thesis Contributions

Consider two cell types C_1, C_2 and a TF T, s.t. T is expressed in both cell types (Fig 1-2). The DNA sequence of both cell types is essentially same. Thus, the regions r_1, r_2 and r_3 have same DNA sequence in both cell types. However, T binds with genomic regions r_1, r_3 in C_1 and r_2, r_3 in C_2. The regions r_1 and r_2 are called cell type specific TFBSs of T. This specificity could be due to several factors that are described in § 1.4, for example the selective presence of several other TFs in C_2 that could have helped T to bind with r_2. We develop a machine learning approach for analyzing the cell-type specific binding of a TF with a TFBS. For simplicity, we phrase this as a binary classification problem in which, given a TF, we want to know in which of two cell types it will bind to a given region. We build our learning models only on the cell type specific genomic regions bound by a TF T, ignoring sites where T binds constitutively i.e. regardless of the cell type. In order to perform the classification, three different feature extraction methods are used to generate features from

the given genomic regions: Known-motif, K-mer profile and Word2Vector methods (§ 4.2). The last two methods do not require any prior information about motifs of TFs, while the first one leverages known TFBS motifs. For each of the feature sets, we experiment with several learning algorithms. We use logistic regression as a linear classifier, and support vector machines and k-nearest neighbor classification as non-linear classifiers (§ 3.2). All classifiers were trained with 3-fold cross validation (§ 3.1) and evaluated using the AUC score (§ 3.3). For the k-fold cross-validation, we set $k = 3$ because higher values of k would require larger computation time. However, higher values of k should give better results because it allows for finer tuning of the parameters involved with the learning model. The proposed methods are able to capture cell-type specific signatures better than state-of-the-art for predicting cell-type specific TFBSs [10] (§ 5.2). The TF-TF interactions identified by one of our models are validated from existing biological results (§ 5.4).

CHAPTER 2
Review of Existing Approaches for TFBSs prediction

In this chapter, we summarize the main experimental approaches (§ 2.1), existing data sources available (§ 2.2) and computational methods (§ 2.3) proposed for TFBS identification.

2.1 Experimental Approaches

Several experimental techniques based on the measurements of protein-DNA interactions have been used to discover and analyze TFBSs [50, 80, 81]. Such experiments can be performed in controlled environment outside (*in vitro*) or within (*in vivo*) cells. The *in vitro* experiments produce PWM (§ 2.3.2) that can be used to identify the location of TFBSs in a DNA sequence. The *in vivo* experiments combined with computational techniques do not produce the exact locations of TFBSs, but the genomic regions where TFBSs can be present. We now describe the commonly used *in vitro* and *in vivo* techniques.

SELEX-seq: Systematic Evolution of Ligands by Exponential Enrichment followed by sequencing (SELEX-seq) is an *in vitro* method to identify DNA sequences for which a given TF has affinity [38] (Fig. 2-1 A). TFs and random DNA sequences are mixed in a solution in order to have binding reactions. Then, the high-affinity binding sites from the DNA sequences are selected. After several rounds of binding reactions, the bound DNA sequences get amplified and are then captured. Later,

these captured DNA sequences can be analyzed using sequencing techniques and the corresponding PWM (§ 2.3.2) is derived [96].

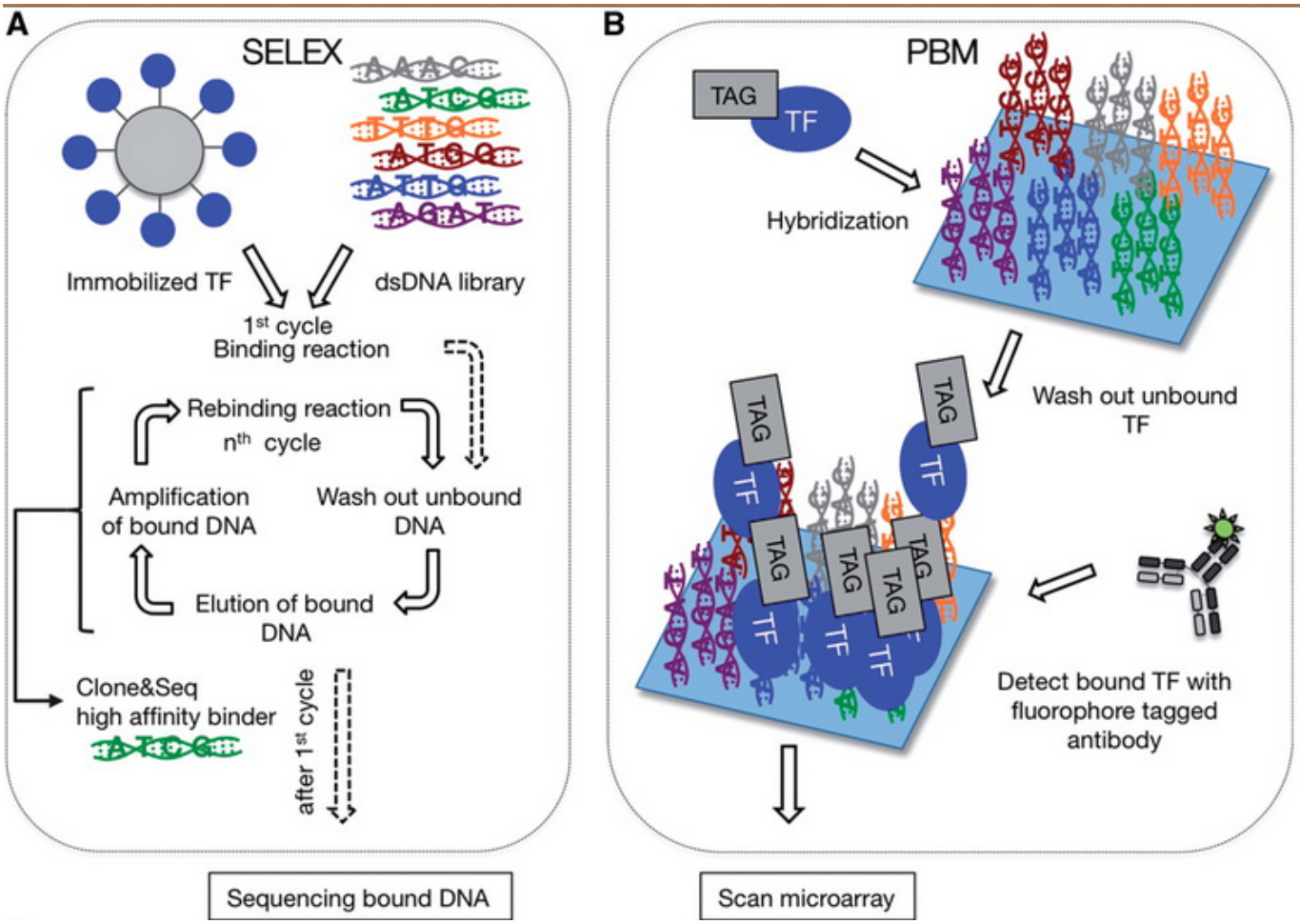


Figure 2–1: Workflow of (A) **SELEX** and (B) **PBM**. Figures are from page 4 of Geertz *et al.* [50]

PBM: The protein binding microarray (PBM) is an *in vitro* technique that uses DNA microarray-based technology [122] to identify short DNA fragments that can be bound by a given TF [18, 17] (Fig. 2-1 B). The TF of interest is injected into a

DNA microarray. The binding reaction of TFs in microarray can be detected using TF specific fluorescent antibodies. The sequences identified from such antibodies are transformed into PWM (§ 2.3.2).

ChIP: Chromatin immunoprecipitation (ChIP) is an *in vivo* technique used to measure protein-DNA interaction in cells. First, DNA is cross-linked with the TF of interest. Then, the DNA is fragmented into pieces of 500-700 base pairs by sonication [108]. This is followed by the immunoprecipitation with protein-specific antibodies of TF bound DNA fragments. After purification of the precipitated TF-DNA complex, the TFBSs can be identified using DNA microarray (ChIP-chip) [122] or sequencing techniques (ChIP-seq) [112, 140, 134], which are detailed below.

ChIP-chip: ChIP-chip is an *in vivo* technique that uses ChIP with microarray hybridization to discover the TFBSs [43]. The TF-DNA complex obtained from ChIP is purified and separated into single stranded DNA. Fluorescent dyes are used to label each of these strands, which are then poured into a DNA microarray [122] consisting of single stranded DNA oligos. The fluorescent tagged DNA strands hybridize with complementary microarray strands to form double stranded DNA. The intensity of the immunoprecipitated sample w.r.t. the DNA is used to locate the binding sites by various ChIP-chip peak calling programs [63, 19].

ChIP-seq: ChIP-seq is an *in vivo* technique that combines ChIP with high-throughput sequencing to discover the TFBSs [64, 43]. The precipitated TF-DNA complexes are used to create a library of fragments, which is treated with PCR amplification. From the amplified library, 200-300 bp DNA fragments are selected and sequenced, which produces sequences known as tags. These tags are aligned to a

reference genome. The regions of enriched tag counts indicate TFBSs and can be identified by peak-calling programs [103, 139, 45]. Fig. 2-2 describes the working of ChIP-seq. ChIP-seq technology has several advantages over ChIP-chip as it can detect repeated sequences that are usually undetectable in microarray and it is free from hybridization noise. In order to detect cell type specific binding sites, ChIP-chip and ChIP-seq are performed in a particular cell type chosen by the experimenter. Repeating these experiments in multiple cell types and comparing the results yields cell type specific and constitutive TFBS.

Some of the major drawbacks associated with experiments based on protein-DNA interactions are:

1. These experiments are expensive and time consuming. For example SELEX-seq require several repetition of TF-DNA binding reaction.

2. The TFBSs predicted from *in vitro* and *in vivo* experiments can differ from each other [50]. For example SELEX approaches do not identify genomic positions that are bound by a TF. They identify short DNA sequences (for example 8-mers) that can be bound. The results of these experiments are often used to create PWMs that can be used to scan actual genomic sequences (for example with HOMER [61]). On the other hand, *in vivo* experiments provide genomic regions that contain TFBSs.

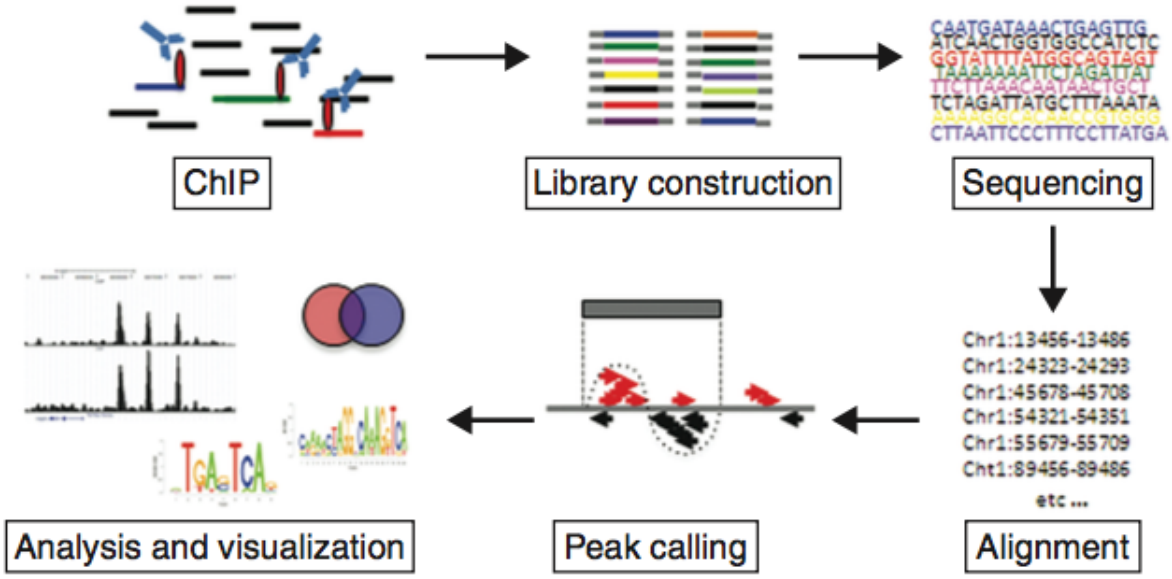


Figure 2–2: **ChIPseq** workflow. Figure is from page 2 of Liu *et al.* [76].

2.2 Database of TFBSs and TF-TF interactions

There are several public and private databases of different transcription factors along with their TFBSs in different organisms that are created from experimental reports and computational predictions of binding sites. TRANSFAC [136] is a private database of experimentally identified mammalian binding sites and PWMs built from them. RegulonDB [49] is a public database for Escherichia coli binding sites.

The Encyclopedia of DNA Elements (ENCODE) [31] is a public research project whose objective is to determine all functional elements in the human genome by systematically mapping regions of transcription, TF-DNA interactions, chromatin structure and histone modifications. The ENCODE project has led to significant

discoveries in biomedical research like novel DNA regulatory elements and relationships between diseases and differences in the DNA sequence [82]. The ENCODE project has associated 80% of human genome with at least one biochemical ***activity*** [31, 95].

The Biological General Repository for Interaction Datasets (BioGRID) [117] provides repositories of protein-protein interactions, genetic interactions, chemical interactions and post-translational modifications. BioGRID data are publicly available in various formats and include over 770,000 biological interactions that are derived from over 54,000 publications [2]. In our study, the predicted results of cell type specific TFBSs are validated with the TF-TF interactions dataset available from BioGRID.

2.3 Computational Approaches to Predict TFBS

Traditionally, computational approaches to predict TFBS use TFBS motif models that represent motifs using a consensus sequence or position weight matrix.

2.3.1 Consensus Sequence

A consensus sequence (CS) is formed by aligning all the related sequences obtained from experimental methods (§ 2.1) and retaining the predominant bases at each position (Fig. 2-3). The resulting CS can be degenerate i.e. positions with no predominant bases can have multiple possible base pairs. A CS does not inform about the relative frequency of nucleotides at each position [118]. Several methods of generating CS with their strengths and weaknesses have been discussed by Day and McMorris [35].

Figure 2–3: Consensus sequence (CS) for a given set of 6 sequences. Each position in CS corresponds to the predominant base pairs in the aligned sequences. Positions with no predominant base pairs can be represented by multiple possible alternative. Here, position 3 of the CS can be either C or G.

2.3.2 PWM

A position weight matrix (PWM) [118] represents motif as relative frequency of bases (A,C,G,T) at each position of the motif. First, a position frequency matrix (PFM) is created for the given set of related sequences that are obtained from experimental methods. The PFM shows the frequencies of each nucleotide at each position of the given sequences (Fig. 2-4(a)).The PWM W can be created from PFM F in the following way:

$$W_{i,j} = \log_2 \frac{p(i,j)}{p(i)}$$

$$p(i,j) = \frac{F(i,j) + s(i)}{n + \sum_{i'} s(i')}$$

where $i \in \{\text{A,C,G,T}\}$, $j \in \{1, ..., l\}$, l is the length of given sequences, $p(i,j)$ is the emperical probability of nucleotide i at position j, n is the number of sequences, $p(i)$ is the background probability of nucleotide i, $s(i)$ is the pseudocount of nucleotide

i (usually used when the set of sequences is small in order to have non-zero $p(i, j)$ values.)

Position	1	2	3	4	5	6
A	1	1	1	3	1	1
C	3	1	2	1	3	1
G	1	1	2	1	1	1
T	1	3	1	1	1	3

(a) Position Frequency Matrix (PFM)

(b) Sequence logo

Figure 2–4: (a) PFM showing frequency of base pairs at different positions for the 6 sequences of fig. 2-3. (b) Sequence logo for the given sequences, created using the weblogo tool [5]. The relative size of a base pair at a given position represents the relative occurrence of that base pair in the aligned sequences.

The PWM is used to discover candidate TFBSs in new sequences. Any DNA sequence S can be scanned and each of its positions can be scored in the following way:

$$\text{Score}[j] = \sum_{i=0}^{l-1} \text{PWM}_{S_{j+i},i}, \quad j \in \{1, ..., |S| - l\} \tag{2.1}$$

where S_{j+i} is the nucleotide at position $j + i$ in S, l is the length of PWM matrix and $\text{PWM}_{x,y}$ is the value of nuleotide x at position y in the PWM.

The list of scores for each position in S obtained from Eq. (2.1) can be filtered out using a threshold to obtain the required TFBSs. There are several approaches to create PWM that are discussed in [118]. There are several other efficient methods to compute the PWM score, which do not depend upon the PWM computation

[116, 29]. A PWM motif can be visualized as sequence logo [109] as shown in fig. 2-4(b), in which the size of the characters at a position indicates the relative frequency of the characters at that position in the given sequences.

The computational tasks to discover and analyse TFBSs can be divided into three problems: building a TFBS model, motif discovery and site search.

Building a TFBS model: DNA sequences of length k bound by a TF are derived from experiments (for example SELEX-seq, PBM) and are aligned to compute the corresponding consensus sequence or PWM.

Motif Discovery: This method starts with a set S^+ of larger DNA sequences ($\sim$ 200 bp) such that each contains a binding site for a TF T, at an unknown position, these are derived from experiments (for example ChIP-seq). A background set of sequences S^- is formed from regions having no binding sites for T. The task is to discover a motif of length k such that the corresponding consensus sequence or PWM of length k all (or most) sequence matches in S^+, and no (or few) matches in S^-. MEME (§ 2.3.3) and HOMER(§ 2.3.4) are motif model based tools that are commonly used for discovering motifs from given sets of sequences.

Site search: One long DNA sequence S and a PWM are given. The task is to find the set of all positions in S that match the PWM. A position is matched if its score (Eq. 2.1) is greater than a user defined threshold. There are several other methods to score a position in the given sequence [116, 29].

2.3.3 MEME

Multiple EM for Motif Elicitation (MEME) [13] is a motif discovery tool used with DNA and protein sequences. MEME searches for repeated, ungapped sequence

patterns in the DNA or protein sequences. MEME characterizes the motifs using PWM. It uses batch EM algorithms [37] to compute the PWM. Major drawbacks of MEME are that gaps/substitutions/insertions are not allowed in motifs and scales poorly to large datasets [98].

2.3.4 HOMER

HOMER (Hypergeometric Optimization of Motif EnRichment) [61] is a differential motif discovery algorithm used for analysing genomic regulatory elements. It distinguishes two sets of genomic sequences based on the relative enrichment of the regulatory elements. For given PWMs, it can scan a DNA sequence and locate the position of motifs. It is well suited for ChIP-Seq (§ 2.1.1) and promoter analysis and it can be applied to any nucleic acids motif finding problem. However, HOMER can miss many weak binding sites because of the threshold settings of the scoring function (Eq. 2.1).

The motif models based on PWM are simple to implement and can be used with various types of protein-DNA binding data [5,8]. However, these models assume that the positions in the binding site are independent of each other, which is not always true [26, 11, 66, 83, 121]. The computation of the PWM does not consider many other factors responsible for the cell type specific TFBSs that are discussed in § 1.5 and often results into false motif discovery and false prediction of TFBSs [118].

CHAPTER 3
Overview of Machine Learning Approaches

Machine learning is a collection of methods that can learn a function or model of interest from a given data set. A typical procedure is to collect data and divide it into disjoint training, validation and testing sets. Models are built by the selected algorithms on the training set. A loss function is associated with a model and the objective of the learning algorithms is to find a model that minimizes this loss function. The learned models are evaluated on the validation set and the best model is selected. Section 3.7 discusses this methodology further. Machine learning techniques are broadly divided into supervised and unsupervised learning. In the former, inputs associated with labels are provided and the learning task is to identify the general mapping from inputs to labels. In the latter, data sets without any labels are provided and the learning task is to identify the structure and to uncover the hidden patterns in the data sets. Classification and regression fall into the first category, while clustering and dimensionality reduction fall in the second. In this study, we use classification approach to solve the proposed machine learning problem (§ 4.1)

3.1 Bias vs Variance

Supervised learning algorithms are associated with two types of error: bias and variance. Erroneous assumptions in the class of models used to map inputs to their labels, result in systematic errors, which are called bias. Variance ocurs when the

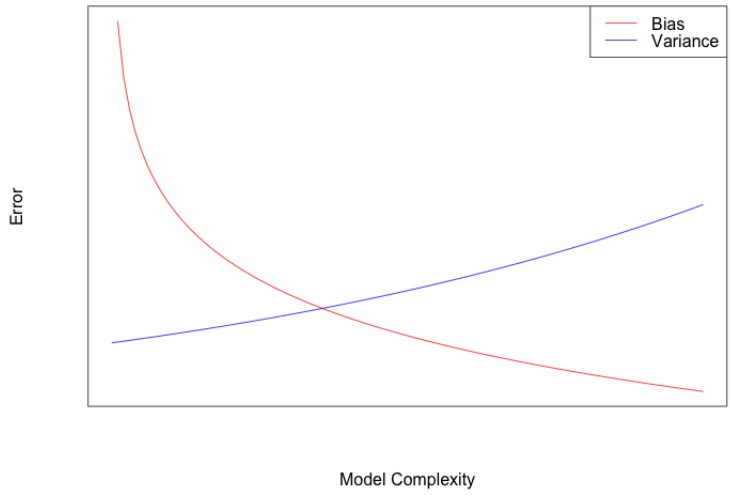


Figure 3–1: Bias-variance tradeoff for a model complexity, in terms of number of learning parameters involved.

learning algorithm is too sensitive to training set, so that small changes in the data lead to big changes in the model. Bias and variance can be traded off by varying the complexity of the model. Fig. 3-1 shows an illustration of the bias-variance tradeoff.

Underfitting is the phenomenon of ignoring relevant patterns existing in the dataset. A high bias leads to underfitting. Including more learning parameters into the model can avoid underfitting. Overfitting is the phenomenon of learning random noise in the data instead of the underlying pattern. In general, overfitting occurs when there are too many learning parameters compared to the amount of data available, leading to high variance. Regularization and cross validation are techniques used to avoid overfitting. Commonly used cross validation technique is k-fold cross validation, where data is partitioned into k equal parts. The models are

trained using $k-1$ parts and validated with the remaining k^{th} part. The model with best performance w.r.t. the accuracy measurement is selected as the best model.

Regularization techniques penalize complex models. Let $e(\mathbf{w})$ be the loss function of a learning algorithm that is fitting a model with parameter vector $\mathbf{w}$, so the objective is to minimize the loss function:

$$\arg \min_{\mathbf{w}} \; e(\mathbf{w})$$

The objective function with the regularization technique would be

$$\arg \min_{\mathbf{w}} \; \{ \; e(\mathbf{w}) + c||\mathbf{w}|| \; \} \tag{3.1}$$

where c is the regularization parameter that penalizes the models with extreme parameter values and can be tuned using cross validation techniques. $||\cdot||$ is either l_1-norm or squared l_2-norm. The l_1-norm gives sparse coefficients of $\mathbf{w}$ and serves as a feature selection method i.e. to identify the relevant components of $\mathbf{w}$. The l_2-norm can be computed efficiently and provides unique solution to Eq. (3.1).

3.2 Supervised Machine Learning Algorithms

In this section we describe the learning algorithms used for our experiments. In order to explain the algorithms, consider a data set D, where data items are divided into two groups of positive and negative examples s.t.

$$D = \{(\mathbf{x}_i, y_i) \; | \; \mathbf{x}_i \in \mathbb{R}^p, \;\; y_i \in \{-1, 1\} \; \}_{i=1}^{n} \tag{3.2}$$

where $\mathbf{x}_i$ is a p-dimensional real vector, y_i designates the label of $\mathbf{x}_i$, and n is the total number of data items. The group with positive examples is called the positive class and the group of negative examples is called the negative class.

3.2.1 k-Nearest Neighbor Classification

k-Nearest Neighbor classification (KNN) [9] is a non-parametric prediction algorithm that predicts the class label of an object as the most common class among its k nearest neighbors. Neighbors of an object are computed using a distance metric such as the Euclidean or Manhattan distance. In general, higher values of k can improve classification by reducing the effect of noise. However, the class boundaries become less distinct with the higher values of k [41]. In other words, high k increases bias and decreases variance.

3.2.2 Logistic Regression

Logistic regression [48] is a probabilistic classification model that uses the logistic function for classification. The logistic function is defined as:

$$f(t) = \frac{1}{1 + e^{-t}}, \;\; t \in \mathbb{R} \tag{3.3}$$

Given a dataset D (3.2), logistic regression computes the probability of a data item $\mathbf{x}_i$ belonging to positive class as follows:

$$P(\; y_i \;=\; 1 \mid \mathbf{x}_i,\; \mathbf{w}) = \frac{1}{1 + e^{-(\; \mathbf{w}^T \cdot \mathbf{x}_i)}} \tag{3.4}$$

$$P(\; y_i \;=\; -1 \mid \mathbf{x}_i,\; \mathbf{w}) = 1 - p \tag{3.5}$$

where $P(.)$ is a probability space, and $\mathbf{w}$ is the model parameter.

The parameter $\mathbf{w}$ for data set D (4.1), is estimated by solving the following optimization problem

$$\arg \max_{\mathbf{w}} \sum_{i=1}^{n} \log\ p(y_i \mid \mathbf{x}_i,\ \mathbf{w}) \ -\ c\ ||\mathbf{w}||_2^2 \tag{3.6}$$

where c is the regularization parameter and $||\cdot||_2^2$ denoting the squared l_2-norms. Eq. (3.6) can be solved using Newton's method [106]. The estimated model parameters show the linear relationship between input features and the labels. For example the log odds ratio of probabilities (Eq. 3.4 and 3.5) is linearly related to $\mathbf{x}_i$.

$$\ln\left(\frac{P(\ y_i\ =\ 1 \mid \mathbf{x}_i,\ \mathbf{w})}{P(\ y_i\ =\ -1 \mid \mathbf{x}_i,\ \mathbf{w})}\right) = w_0 + \mathbf{w}_1 \cdot \mathbf{x} \tag{3.7}$$

If the dot product between $\mathbf{w}_1$ and $\mathbf{x}_i$ in eq. 3.10 is expanded, i.e.

$$\mathbf{w}_1 \cdot \mathbf{x}_i = w_{11}x_{i1} + w_{12}x_{i2} + ... + w_{1j}x_{ij} + ... + w_{1p}x_{ip}$$

then the value of parameter w_{1j} determines the direction and strength of the linear relation between x_{ij} and y_i. If $w_{1j} > 0$, then $|w_{1j}|$ dictates the strength of linear relation between x_{ij} and $y_i = 1$. Similarly, if $w_{1j} < 0$, then $||w_{1j}||$ dictates the strength of linear relation between x_{ij} and $y_i = -1$.

3.2.3 Support Vector Machines

Support Vector Machines (SVMs) [32] are a class of supervised learning algorithms that perform binary classification. Given a dataset D, a linear SVM finds a hyperplane that maximally separates the data items having $y_i = 1$ from those having

$y_i = -1$. Any hyperplane h separating the two classes of D could be represented as

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \tag{3.8}$$

where $\mathbf{w}$ is the normal vector to h and b is a bias term.

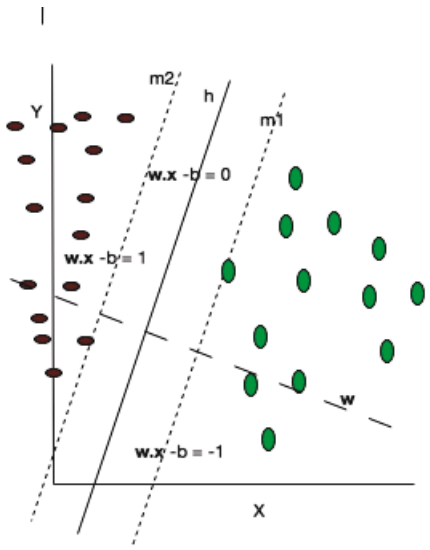


Figure 3–2: SVM finds the maximal-separating hyperplane h. Here, m1 and m2 are hyperplanes that are parallel to h, **w** is a normal vector to h and b is a bias term. The region formed by h, m1 and m2 is known as margin.

The goal is to find a hyperplane h that separates the data items and has a maximum distance to them. Then, the data items are labelled according to which side of h they are. To achieve h, consider two other hyperplanes m_1 and m_2 parallel to h such that the region R called the margin bounded by m_1 and m_2 contains no data items and R is the widest possible (Fig. 3-2). The hyperplanes m_1 and m_2 can

be represented as

$$\mathbf{w} \cdot \mathbf{x} - b = 1 \tag{3.9}$$

$$\mathbf{w} \cdot \mathbf{x} - b = -1 \tag{3.10}$$

The goal of finding the maximum margin can be formulated as the following optimization problem

$$\arg \min_{(\mathbf{w},b)} \frac{1}{2} ||\mathbf{w}||^2 \tag{3.11}$$

$$\text{subject to, } y_i \ (\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 \ \forall \ i \in \{1, ..., n\}$$

SVM solves the dual form of the maximum margin problem (Eq. 3.11) that depends only on the support vectors, the data items that lie only on margins. The dual form can be obtained using the lagrangian multipliers and KKT conditions [40] :

$$\arg \max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{3.12}$$

subject to

$$\textstyle\sum_{i=1}^{n} \alpha_i y_i = 0, \ \ \alpha_i \geq 0 \ \forall \ i = 1, ..., n$$

where α_i's are lagrangian multipliers.

Eq. (3.12) can be solved using quadratic programming techniques such as the interior point method [22] to get the value of α_i's, which are non-zero only for support vectors. Then, $\mathbf{w}$ can be calculated from,

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \tag{3.13}$$

Using Eqs. (3.9), (3.10) and (3.13), the label of any input data $\mathbf{x}$ is given by,

$$\text{class}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} - b) \tag{3.14}$$

Now, there can be cases where no hyperplane exists that can separate the data items. Even if the hyperplane exists, SVM may overfit the data items (for example due to outliers [56]). To tackle these situations, SVM uses soft margin to find the separating hyperplane [32]. The idea is to include some data items inside the margin and to find a hyperplane that is at maximum distance possible from rest of the data items. Slack variables ξ_i's are used to determine the data items that should be included within the margin. Slack variables meausure the degree of violation of linear constraint of the data items mentioned in Eq. (3.11). The optimization problem of finding the maximum margin is extended with a penalty function applied to slack variables. If the penalty function is linear, then the optimization problem is,

$$\arg \min_{(\mathbf{w},\xi,b)} \left\{ \frac{1}{2} \ ||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i \right\} \tag{3.15}$$

subject to,

$$y_i \ (\mathbf{w} \cdot \mathbf{x}_i \ - \ b) \geq 1 - \xi_i \quad , \quad \xi_i \geq 0 \quad \forall \ i \ \in \ \{1, ..., n\}$$

where ξ_i is a slack variable that indicates the amount by which data item $\mathbf{x}_i$ violates the linear constraint of $y_i \ (\mathbf{w} \cdot \mathbf{x}_i \ - \ b) \geq 1$ and C is a constant that penalizes the slack variables and determines the model complexity. For larger values of C, margin would have smaller number of data items. The model would overfit the data items and would be more complex. Similarly, for smaller values of C, margin would have larger number of data items and the model would be less complex.

The corresponding dual problem of Eq. (3.15) that the SVM solves is,

$$\arg \max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{3.16}$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \;\; 0 \leq \alpha_i \leq C \;\; \forall \;\; i = 1, ..., n$$

SVM can be used to perform non linear classification using what is known as the 'kernel trick'. A data set with non linear decision boundary can be mapped into a high-dimensional feature space where a suitable separating hyperplane can be found. Let $\phi(\cdot)$ be such a mapping for these datasets. Then, the inner product of data items in Eq. (3.12) becomes $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. Instead of computing high dimensional representation of data items, SVM computes a kernel function, $k(\cdot)$ that satisfies the Mercer's theorem [8] and $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$. Thus, Eq. (3.12) using a kernel function, $k(\cdot)$, becomes,

$$\arg \max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{3.17}$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \;\; \alpha_i \geq 0 \;\; \forall \;\; i = 1, ..., n$$

Some of the examples of kernel functions are radial basis function [24], polynomial [33] and hyperbolic tangent [30].

3.3 Area Under Curve

Accuracy is a measure to evaluate the predictive ability of learned models. For a bianry classification task, accuracy could be computed from how often a model

(in)correctly predicts the outcome. Sensitivity or true positive rate (TPR) and specificity or true negative rate (TNR) are commonly used accuracy measures that give complete picture of the prediction errors. Let TP, TN, FP and FN be the number of true positive, true negative, false positive and false negative predictions of the classifier. Then,

$$\text{Sensitivity} = \text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{(\text{FP} + \text{TN})}$$

The Receiver Operating Characteristic (ROC) curve is a Sensitivity vs Specificity plot, which is used to evaluate the performance of a binary classification algorithm. For example, fig 5-2 compares the performance of SVM, KNN and logistic regression at different settings of the threshold used to determine the positive and negative class. The area under the ROC curve (AUC) [60], is widely used in machine learning for performance comparisons [58]. A higher AUC score means a better classification algorithm. For our experiments, we report the AUC score of the learned models.

3.4 Related Work

Traditionally, machine learning techniques are applied to the prediction of TFBS by discriminating k-mer profile (§ 4.2.1) patterns. This approach combined with a SVM (§ 3.2.3) is used to solve and analyze different problems associated with the TFBSs, which we discuss in this section.

Agius *et al.* [7] used a SVM with k-mer profiles to learn *in vitro* and *in vivo* TF binding preferences. For *in vitro* experiments, authors predict binding intensities from probe sequences of *in vitro* protein binding microarray data. The learned

model outperformed motif based models for 81% of mouse TF data [11]. For *in vivo* experiments, authors classify peak and non peak regions of ChIP-chip and ChIP-seq experiments obtained from three cell types: ES cells of mouse and GM12878 and HepG2 cell lines of human. The learned model performed better than motif based models with sometimes improvement of 0.1 AUC score.

Arvey *et al.* [10] examined a SVM with k-mer profile, chromatin signatures and Dnase signatures for the prediction of TFBSs. The authors classified peak and non peak regions of 238 ChIP-seq experiments consisting of 67 transcription regulators for three human cell types, GM12878, K562 and HeLa using SVM and k-mer profile. The learned model performed better than motif based models for 90% of TFs with mean AUC improvement of 0.07. The authors found that k-mer profile with chromatin signatures and Dnase signatures improved prediction by mean AUC of 0.04 and 0.08 respectively. The previous model was trained on one cell line, used to predict TFBSs on new cell line and the mean AUC improvement reported was 0.05. The authors presented a machine learning approach to predict cell type specific TFBSs. Two specific models, GM12878- and K562-specific were trained using multi task learning [27]. GM12878- and K562-specific models would predict the cell type specific TFBSs in GM12878 cell type and K562 cell type respectively. The peak regions of ChIP-seq experiments specific to GM12878 and K562 were used for simultaneously training the GM1278- and K562-specific models.

Previous studies related to TFBSs prediction have been conducted largely with *in vitro* experiments and databases consisting of yeast and mouse transcription factors. However, *in vivo* protein-DNA interactions differ significantly from *in vitro* and

such interactions could be distinct among different species. Moreover, a very little amount of work is done on developing machine learning models that could detect cell-type specific TFBSs. In our study, we focus on these shortcomings by developing machine learning techniques to identify cell-type specific TFBSs across multiple cell types and to detect factors for cell-type specificity such as TF-TF interactions.

CHAPTER 4
Methods

We represent the prediction of cell type for TFBSs as a binary classification machine learning task as discussed in § 4.1. We assume that TFBSs have the same binding affinity in different cell types. Given two sets of genomic sequences (one bound in cell type A and the other bound in cell type B), we construct feature vectors using three different feature extraction methods, namely K-mer Profile, Known-Motif and Word2Vector methods. Then we use three types of classifiers (logistic regression, SVM and KNN, § 3.2) to build our TFBS predictor models. We report the Area Under Curve (§ 3.3) score as evaluation metric for the learned models. The machine learning work flow used in our experiments is summarized in fig. 4-1. The following sections state the machine learning problem and describe the feature extraction, model selection and model evaluation stages.

4.1 Problem Definition

Consider cell types C_1 and C_2, as well as a particular TF T. A question of interest to biologists is whether a given genomic region R will be bound by T in

(1) C_1 only,

(2) C_2 only,

(3) both C_1 and C_2, or

(4) neither C_1 nor C_2.

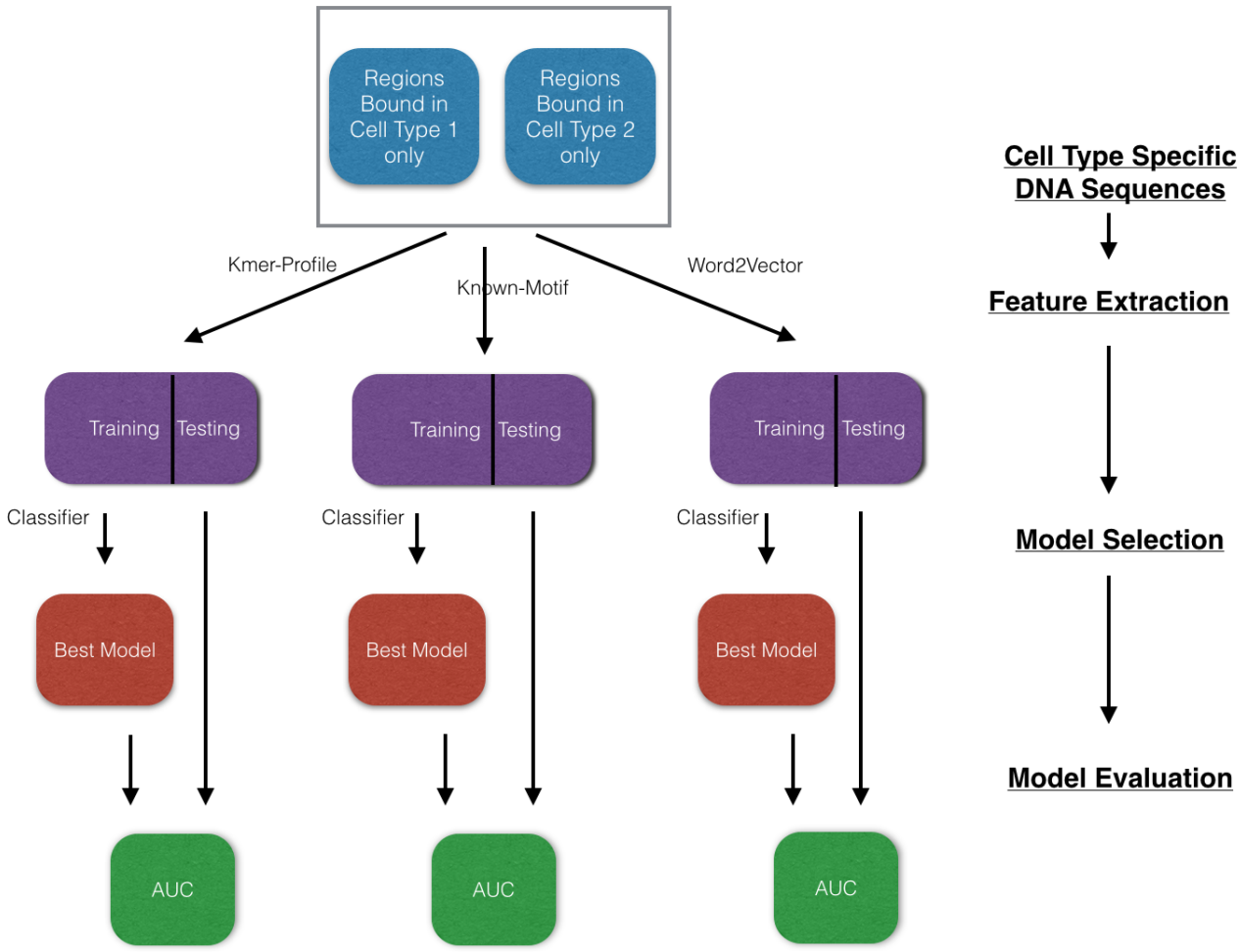


Figure 4–1: Machine learning Work Flow. Three different methods - Kmer-Profile, Known-Motif and Word2Vector, are used to extract features from the given DNA sequence data. The extracted feature data is divided into training and testing. Three different classifiers - SVM, KNN and logistic regression, are used to select best model using 3-fold cross validation. Each model has same set of training and testing DNA sequences. Best models are evaluated with AUC score on the test set.

The question naturally generalizes to more than two cell types: Given $C_1, ..., C_k$ predict in what subset of cell types, the genomic region R would be bound by T. To keep the classification task as simple as possible, we concentrate on a version of the $k = 2$ problem where region R under consideration is expected to be bound in exactly one of the two cell types. This allows us to focus on the features that determine cell type specific binding. More formally,

Definition 4.1.1. A cell type specific sequence S, is bound by a transcription factor T in either cell type C_1 or C_2, but not in both.

Definition 4.1.2. A constitutive cell type sequence S, is bound by a transcription factor T in both cell type C_1 or C_2.

Problem 4.1.1. *Let T be a transcription factor expressed in cell types C_1 and C_2. Given a set of cell type specific sequences of base pairs S_1 and S_2, bound by T in cell types C_1 and C_2 respectively, the task is to find a predictor, $f : S \rightarrow \{C_1, C_2\}$, that predicts whether a cell type specific sequence S will be bound by T in cell type C_1 or C_2.*

4.2 Feature Extraction

We develop three different methods to extract features from genomic sequences: K-mer Profile, Known-Motif and Word2Vector methods. The genomic sequences provided from lab experiments can be of varying length with the exact location of TFBS unknown. As evident from ChIP-Seq data, for a given genomic sequence, the

TFBS is most likely to be present somewhere in the middle of the region. We consider the given genomic sequence is of 400 base pairs. The given genomic sequence is reformed into a sequence of 400 base pairs, either by trimming or expanding from both ends, so that the learning of side effects of lab experiments could be avoided. We apply our feature extraction methods on these processed sequences, which return fixed-length input vectors.

4.2.1 K-mer Profile Method

A k-mer is a contiguous sub-sequence of length k from a string or sequence. In the context of predicting TFBS, k-mers are used to represent motifs. Certain k-mers would be pre dominant in the sequences having TFBS as compared to the sequences not having TFBS. The k-mer patterns have been used with machine learning techniques to explain the cell type specific binding of TFs [10].

The reverse complement of a DNA k-mer is the same k-mer in reverse order with each of the base pair replaced with its complement (A,T and C,G are complement of each other). For example, **ATG** is the reverse complement of **CAT**. In K-mer Profile method, we use the number of occurrences of k-mers and its reverse complements in a given sequence S as the feature vector for S. We consider reverse complements because the TF can bind to either strand of DNA in TFBS.

If there are n possible characters for each position in the k-mer, then there are n^k possible k-mers. We typically consider DNA strings for which $n = 4$ (A,C,G,T), giving 4^k possible k-mers. We use the vector of length 4^k representing the number of occurrences of these k-mers in a given genomic sequence S as a feature vector

for S. As mentioned in § 1.3, TFBS are typically of length 5 to 15 base pairs long, which means motifs length would be in this range. The higher values of k require more computation power. Therefore, we set $k = 6$ for our experiments. The K-mer Profile method gives a feature vector of length ($4^6 =$) 4096 for every given genomic sequence.

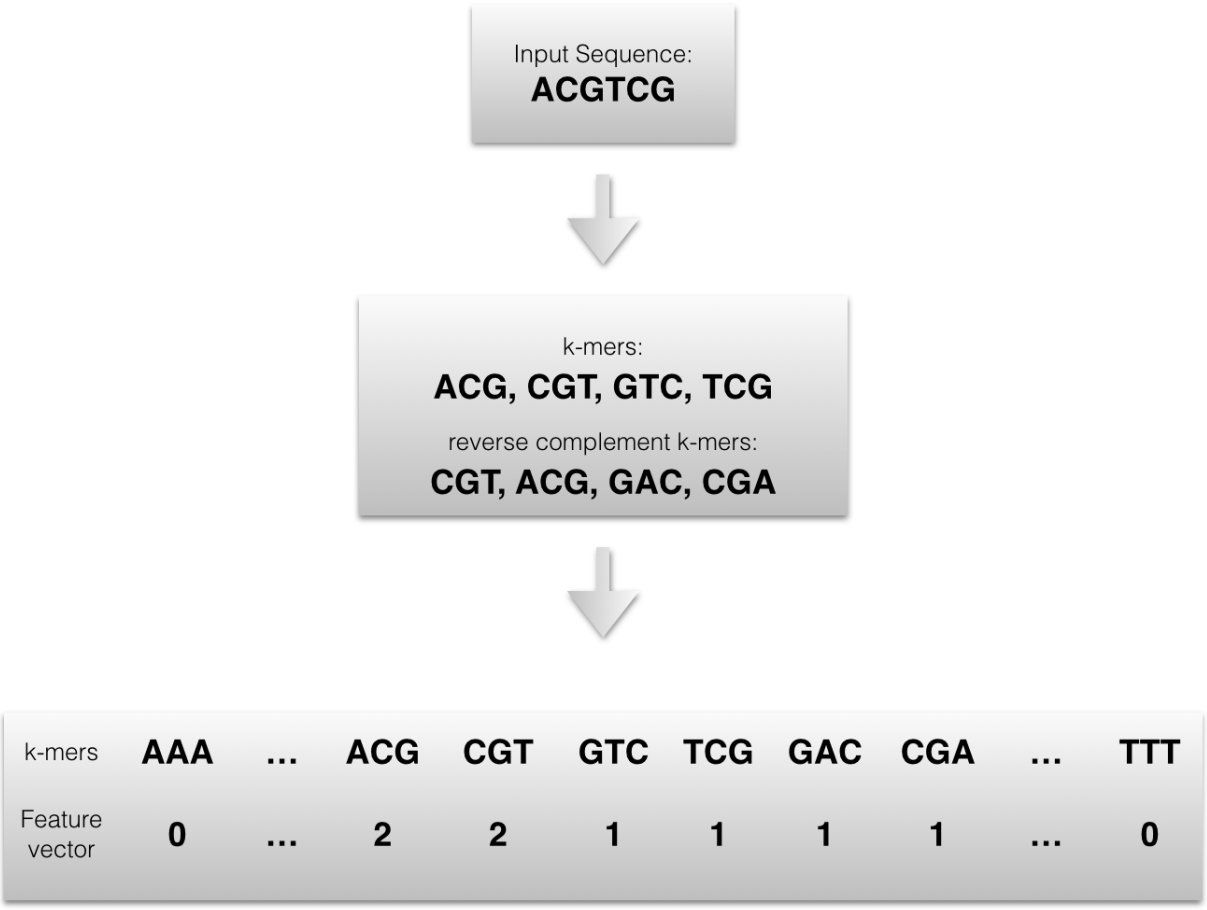


Figure 4–2: Work flow of K-mer Profile method, with $k = 3$. It produces feature vector of length 64, where only eight 3-mers have non-zero values.

Fig 4-2 demonstrates the working of K-mer Profile method with a sample sequence S and k is set to 3. The input vector would be of length $4^3 = 64$. Out of 64 k-mers, only 4 are present in S. K-mer Profile method counts the presence of k-mers as well as their reverse complements. As a result, 8 k-mers are present in S, whose counts are shown in last flow of Fig 4-2. The other 56 k-mers are mapped with 0.

4.2.2 Known-Motif Method

Known-Motif method uses the HOMER tool (§ 2.3.4) to extract a feature vector from a given genomic sequence. HOMER tool has a database of 321 TFs along with their motifs and PWMs [1]. These motifs are termed as known motifs. HOMER tool can scan through a given sequence and return the number of occurrences of these known motifs. The vector of length 321 representing the number of occurrences of these known motifs is used as a feature vector for a given sequence. We use the following command with HOMER tool:

findMotifs.pl <target> fasta <output-dir> -fasta <background> -find <knownmotif>

<target> is the input sequence in fasta file format. <output-dir> is the output directory for the results to be stored. <background> is the sequence file in fasta file format that HOMER uses for differential motif finding. Here, we use both target and background as same input sequence file. <knownmotif> is the list of 321 known motifs.

4.2.3 Word2Vector Method

The Natural Language Processing (NLP) deals with the interaction of computer and human languages. One of the major NLP task is to make computer understand the textual data, TD. NLP community often incorporates machine learning techniques to solve the problem of computational understanding of TD [53, 114, 125, 124, 132]. In order to utilize machine learning techniques with TD, various feature extraction methods have been developed. These feature extraction methods would process TD and produce fixed-length input vectors. The Bag-of-words is one such feature extraction method that is widely used in NLP community [59], owing to its simplicity and efficiency. The Bag-of-words represents TD with the pair of unique words and their frequencies in TD.

However, Bag-of-words can not reveal the semantic relations existing among the words of TD efficiently, as the word order is not maintained. Similar Bag-of-words representations are possible for different TD. The Bag-of-words is not compatible with large sized TD and produces sparse feature matrix that degrades the performance of learning algorithms. There exist many other techniques for feature extraction of TD, but to some extent, all of them suffer from similar disadvantages[16].

Mikolov *et al.* [86, 87] developed skip gram model to address these problems associated with the feature extraction methods . In the skip-gram architecture, a word w of TD is used to predict its surrounding word. In the process, this prediction produces a continuous real vector, which is used as a feature vector for w. These feature vectors are known as word vectors. It has been shown that such feature

vectors preserve the semantic relations involved among the words and has been used to solve various NLP tasks efficiently [86, 87].

Fig 4-3 demonstrates the working of skip-gram architecture. Consider a text sentence of n words, $w_1, w_2, ..., w_i, ..., w_n$. The word vector of w_i can be computed from its surrounding, determined by a context size c. Here, the surrounding words of w_i are $w_{i-c}, w_{i-c+1}, ..., w_{i-1}, w_{i+1}, ..., w_{i+c-1}, w_{i+c}$. A neural network model [57] is build from w_i as an input layer, surrounding words as a output layer, and one hidden layer with user-defined p number of nodes. The weights of this neural network can be computed using back-propagation algorithm [105]. The weights connecting w_i to the hidden layer nodes, gives word vector of length p for w_i.

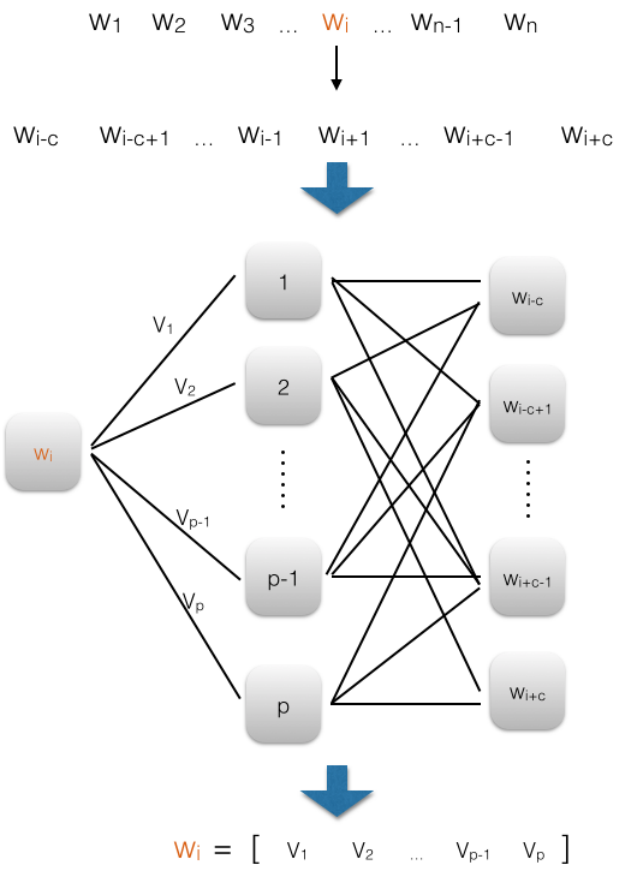


Figure 4–3: $w_1, w_2, ..., w_i, ..., w_{n-1}, w_n$ is a text sentence of n words. For a context size of c, surrounding words of w_i are $w_{i-c}, w_{i-c+1}, ..., w_{i+c-1}, w_{i+c}$. A neural network [57] is build with w_i as input layer, surrounding words as output layer and a hidden layer with p number of nodes. For clarity, not all weight edges are shown. The weight edges $v_1, ...v_p$ that connects input layer to hidden layer nodes, provide word vectors for w_i after the execution of back-propagation algorithm [105].

As mentioned in § 1.4, TFBSs can be clustered and degenerate. DNA sequences involved in binding with TF can be lying close to each other having some relations among themselves. Interestingly, this scenario is similar to that of continuous skip

gram model where words in a sentence have relations with surrounding words (contexts). Therefore, we would like to use a model similar to the continuous skip gram model to compute 'word' vectors representing the subsequences involved in binding.

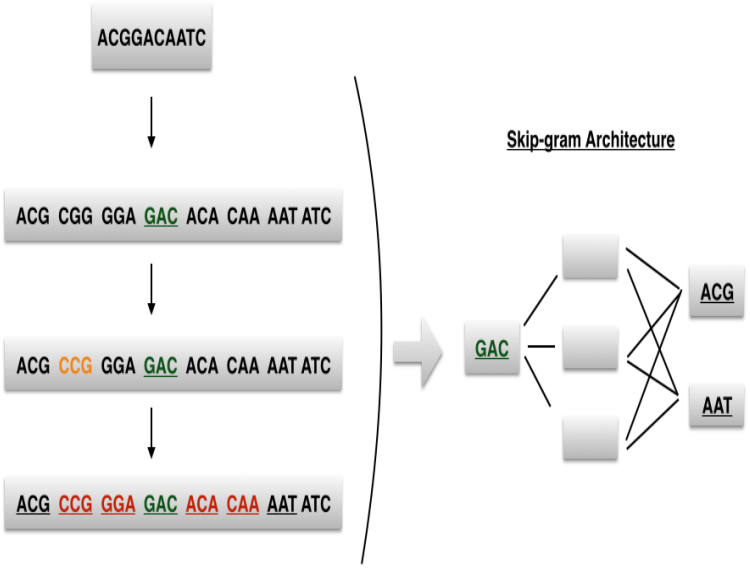


Figure 4–4: Work flow of Word2Vector method. Construction of word vectors for a 3-mer, GAC is shown. In the second step, all the alphabetically higher ranked 3-mers w.r.t. their reverse complements are replaced with their lower ranking counterpart, for example CGG is replaced with CCG. Context window is set to 3, so the surrounding k-mers of GAC are ACG, CCG, GGA, ACA, CAA and CAA. All the immediate four neighbors of GAC are excluded in skip-gram training as they overlap with GAC.

However, there are stark differences between text data and DNA sequence data. Text involves sentences having clear boundaries between words making them suitable for the skip gram architecture. On the other hand, DNA sequences are contiguous sequences of only four letters (A,C,G,T), with no clear boundaries. Therefore, we

represent a DNA sequence as a sequence of overlapping k-mers. Fig 4-4 demonstrates the computation of word vectors for a 3-mer **GAC**, present in the sequence **ACGGACAATC**. Here, context size of 3 is used. So, the output layer of skip-gram architecture consists of the 6 neighbors of **GAC**. However, the immediate four neighbors (two on each side) are excluded, as they overlap **GAC**, originally. The 3-mers alphabetically ranking higher than their reverse complements are replaced with their reverse complements. In the final step, the skip-gram model with 3 nodes in the hidden layer is used to compute the word vector of length 3. Similarly, word vectors of other 3-mers from the same input sequence are computed. The average of these word vectors results into feature vector for the given sequence.

The Word2Vector method uses the above formulation to compute word vectors for k-mers. We then obtain a feature vector for a sequence by averaging these word vectors. Considering general range of motif lengths and computational power required, we set $k = 6$ for our experiments. After several experimental trials, we use the context size of 200 and number of nodes in hidden layer as 500 for the skip-gram architecture. This results into a feature vector of length 500 for each sequence. The word2vec code available at [6] were used with the modifications discussed above, to compute the word vectors.

4.3 Model Selection and Evaluation

Our proposed problem (§ 4.1.1) of cell type prediction is a supervised binary classification machine learning problem. The decision boundary between the two

classes could be either linear or non-linear. We develop our model for the classification task using three different classifiers: logistic regression with ℓ_1 and ℓ_2 penalty, SVM and KNN classifier and the three different feature extraction methods discussed in § 3.2 and § 4.2. Logistic regression and SVM can learn linear and non linear decision boundaries respectively. We use KNN because it can learn a non-linear decision boundary and it is a non-parametric classifier, which makes it less complex model than SVM. Each combination of a classifier and a feature extraction method represents a model, so we have 12 models in total.

We keep 33% of data as testing and use the rest as training sets. We select the best learning parameter values for each classifier using 3-fold cross validation over training set and performing grid search over learning parameters. We use logistic regression and SVM with a regularization parameter to avoid overfitting. After several trials with different kernels and regularization techniques, we use SVM with a radial basis kernel [24] and logistic regression with ℓ_1 and ℓ_2 penalties. The selected parameter settings for the classifiers used are shown in table A1. The classifiers available from scikit package are used [92]. The learned models are evaluated with AUC score.

CHAPTER 5
Data and Results

In this chapter, we discuss about the data used and the results obtained from our experiments. We compare the learned models consisting of different classifiers and feature extraction methods in terms of the AUC score and running time. We find that the learned models perform better for cell-type pairs with relatively large number of cell-type specific binding sites. Moreover, one of the model can be used to identify TF-TF interactions associated with the cell type specific TFBSs.

5.1 Data

We use ChIP-seq data available from ENCODE to examine TFs across five human cell-lines: GM12878 (lymphoblastoid cells), H1-hESC (embryonic stem cells), HeLa-S3 (epithelial cancerous cells), K562 (myelogenous leukemia cells) and HepG2 (human liver carcinoma cells). For each TF, we consider every pair of these five cell lines to solve the proposed machine learning problem (Problem 4.1.1). Only 52 TFs of ENCODE bind in our chosen cell-lines pair. The combination of a cell type pair $\{C_1, C_2\}$ and a TF T constitute a machine learning experiment. In such an experiment, C_1 is chosen as a positive class and C_2 as a negative class. The genomic regions bounded by T in both C_1 and C_2 are excluded because we want to study the differential binding behavior of TFs accross different cell-lines. This would guide the learning algorithms to look for the patterns that caused T to bound

at a particular location in C_1, but not at the same location in C_2 and vice versa. For every experiment, in order to have an unbiased classification, both class sizes are kept same by randomly downsampling the larger class to the size of the smaller one. For computational efficiency reasons, the combinations that give large number of positive and negative examples, we retain 6000 examples randomly sampled. This led us to 165 machine learning experiments. Table A2 displays the combinations of cell types and TFs used in these experiments along with the number of cell type specific and constitutively bounded sequences.

5.2 Analysis of the accuracy of predictors

We evaluate each of classifiers (SVM, logistic regression with penalty ℓ_1 and ℓ_2, and KNN) and each type of feature extraction methods using the AUC score (Fig. 5-1). The AUC score for each model varied across the datasets. On an average, the learned models give better AUC score with logistic regression and SVM as compared to KNN (Table 5-2). We find that the model m with the combination of Known-Motif method and logistic regression (ℓ_1 penalty) gave the best mean AUC score (0.8186). All learned models, except for two cases, perform much better than a random classifier. We observe a weak correlation between size of training examples and AUC scores. The top ten and bottom ten performances for each model are reported in Tables B1-B12.

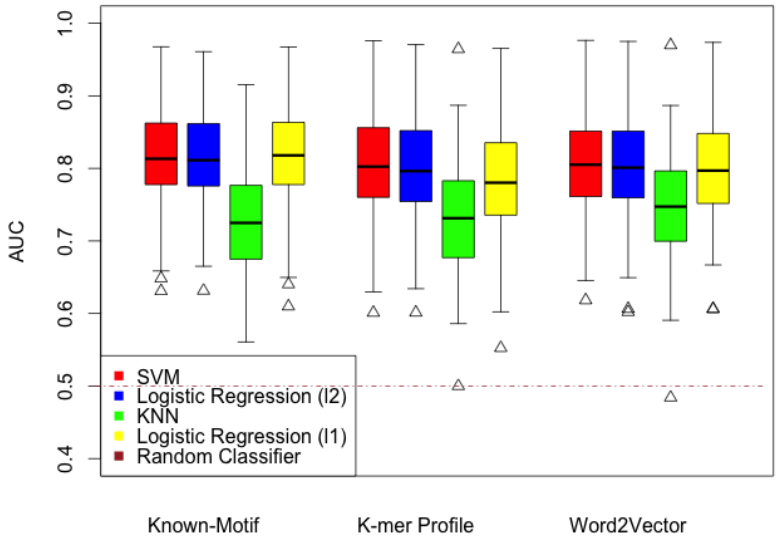


Figure 5–1: Models performance. Each box represents the AUC score measures for a classifier and a feature extraction method across the 165 experiments. The AUC score of models varies across the dataset. Logistic regressions and SVM perform relatively better than KNN. Logistic regression with ℓ_1 penalty and Known-Motif method give best mean AUC score of 0.8186.

We conduct anova analysis [46] to determine whether the AUC scores are statistically significant or just random results. The AUC score is used as a response variable and TFs, cell-type pairs, feature extraction methods and classifiers as independent variables. Table 5-1 reports the F ratio [77] and their associated p-values [70] of the independent variables. F ratio indicates the relevance of variance in response variable due to an independent variable. Larger value of F ratio shows the significant effect of an indpendent variable on the response variable. We find that

all of the independent variables are significant for the resultant AUC score. In particular, classifiers have the most effect on AUC score.

The following command is used in RStudio [104] for anova analysis:

summary(aov (AUC ∼ TF + Cell Type Pair + Feature Extraction Method + Classifier, data=data))

Table 5–1: Anova Analysis

Independent Variable	**F value**	**p-value**
TF	46.44	< 2e-16 ***
Cell Type Pair	28.01	< 2e-16 ***
Feature Extraction Method	14.31	6.8e-07 ***
Classifier	283.08	< 2e-16 ***

We perform the Mann-Whitney-Wilcoxon Test (MWWT) [44] of the best model, m with other models. Table 2 reports the mean AUC score of models and the associated p-values of the MWWT test w.r.t. m. Based on the p-values, we observe that m is comparable with models consisting of SVM, logistic regression (1, 2, 5, 9) and perform much better than the models consisting of KNN (3, 6, 7, 8, 10, 11, 12) .

Table 5–2: Model Comparison

	Model	**mean AUC**	**p-value**
1	KnownMotif with SVM	0.8179	0.882
2	KnownMotif with Log. Reg. (ℓ_2 penalty)	0.8161	0.696

3	KnownMotif with KNN	0.7254	1.27×10^{-24}
4	KnownMotif with Log. Reg. (ℓ_1 penalty)	0.8186	NA
5	K-Mer with SVM	0.8057	0.0788
6	K-mer with Log. Reg. (ℓ_2 penalty)	0.8013	0.0184
7	K-Mer with KNN	0.7329	4.0×10^{-21}
8	K-Mer with Log. Reg. (ℓ_1 penalty)	0.7837	1.81×10^{-5}
9	Word2Vec with SVM	0.8056	0.0688
10	Word2Vec with Log. Reg. (ℓ_2 penalty)	0.8015	0.0199
11	Word2Vec with KNN	0.7493	3.73×10^{-16}
12	Word2Vec with Log. Reg. (ℓ_1 penalty)	0.7982	0.00664

Ideally, we would like to compare the performance of our models with the state-of-the-art techniques. Traditionally, PWM-based approaches are used for the prediction of TFBSs. However, these approaches do not take the cell-types where the TF binds as input, thus, can not be used for the problem 4.1.1. The methodology proposed by Arvey *et al.* [10] is the closest approach that could be adapted to our task. Arvey *et al.* [10] proposed a Kmer based SVM model to predict TFBSs in a given cell type, which is shown to outperform the traditional motif based approaches. In order to evaluate our methods, we compare the AUC score of m with the Kmer SVM model (Fig. 5-2). We use the code and datasets available from [3] to train the Kmer SVM model. We select nine combinations of TF and cell-type pairs that are common to datasets (§ 5.1) and [3]. In order to form Kmer SVM model for the

proposed problem 4.1.1, we made several additions. For example, to predict one of the cell-types C_1 and C_2, where a TF T could bind in a given DNA sequence, C_1- and C_2-specific Kmer SVM models are used. C_1- and C_2-specific models are trained on binding and non-binding genomic regions of T in C_1 and C_2 respectively. We test such cell-type specific models on our testing set for the common nine combinations. Let s_1, s_2 be the score of C_1-, C_2-specific models for a given test sequence. Then, $(s_1 - s_2)$ is used as the final score. As the Kmer SVM models are trained on DNA sequences of 100 bp length, we partition our test DNA sequences of 400 bp length in four parts. The scores s_1, s_2 are the aggregate scores on these four parts. We find that m gives better AUC score, with the mean improvement of 0.18 (p-value 0.0027, MWWT test).

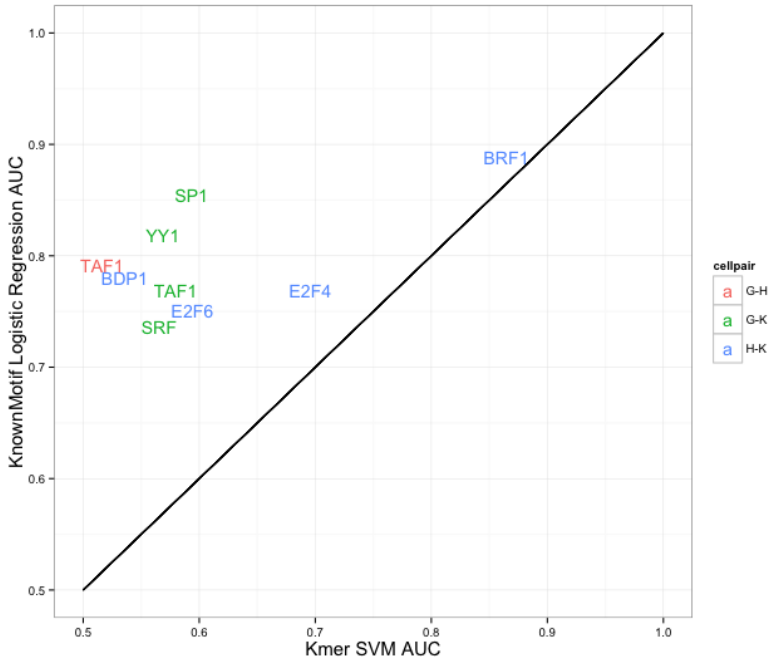
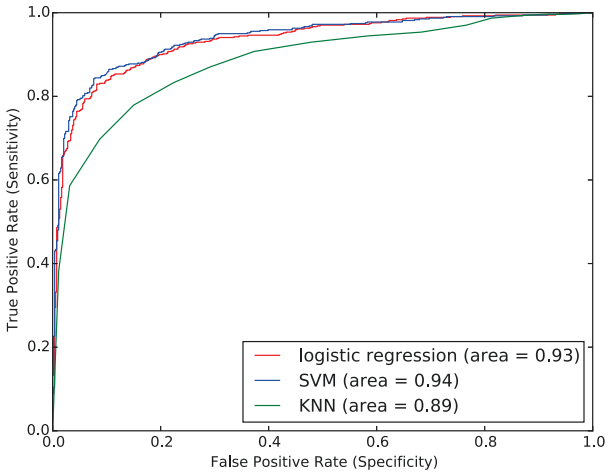
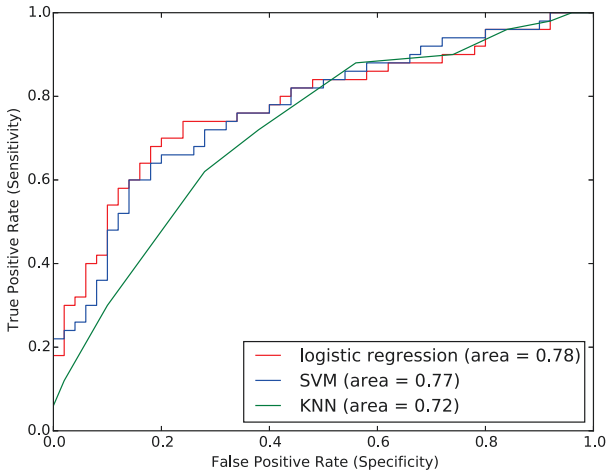


Figure 5–2: Comparison of AUC scores between the Kmer SVM model as proposed in [10] and the model based on KnownMotif method and logistic regression with ℓ_1 penalty.

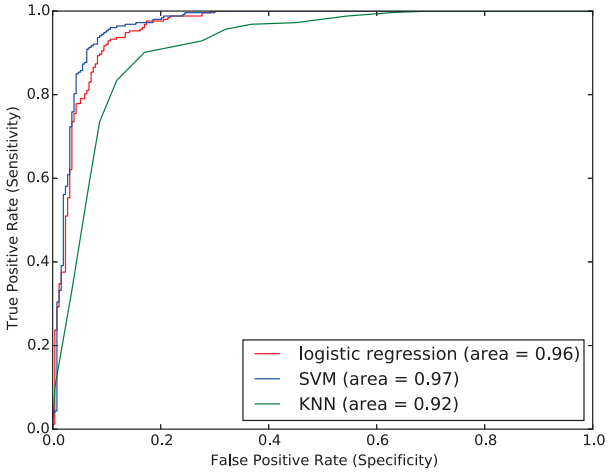
A particular classifier can perform better than another depending upon sensitivity and specificity thresholds. The probability measures obtained from the classifiers can be used to plot the ROC curves, and the models can be evaluated based on the specificity and sensitivity values. Fig 5-3 shows four of the ROC curves from our experiments. Due to space constraint, we do not report all of the 165 ROC plots.



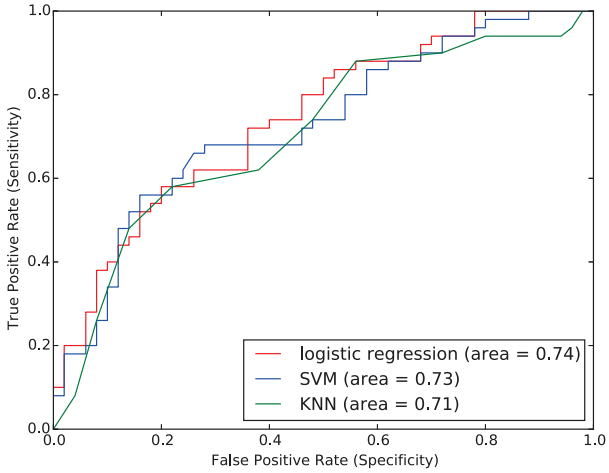
(a) TF: EZH2, pair: {H1-hESC, HeLa}



(b) TF: IRF3, pair: {HeLa, HepG2}



(c) TF: FOSL1, pair: {H1-hESC, K562}



(d) TF: ZZZ3, pair: {GM12878, HeLa}

Figure 5–3: ROC curves

5.3 Cases of Variability in Predictor's Accuracy

We observe that the learned models do not perform consistently on the datasets. For certain cases, the learned models detect cell-type specific signatures strongly,

while weakly for several other cases. To understand the role of cell-type specific content in model performance, we compute 'disjointness' of the cell-type pairs.

Definition 5.3.1. For a cell-type pair C_1, C_2 and a TF T,

$$\text{disjointness} = \log_{10} \frac{\#\text{ cell type specific TFBSs of } T}{\#\text{ cell type constitutive TFBSs of } T}$$

Disjointness reflects the amount of differential binding of a TF across a cell-type pair. Fig. 5-4 shows the relation between disjointness and AUC score of the model m. We observe that m discriminates cell-type specific signatures much better for the combinations of TF, cell-type pair having larger disjointness. With few exceptions, we can conclude that the prevalent differential binding of TF across the cell-type pair would result into easier detection of cell-type specific signatures.

5.4 TF-TF interaction

We use the model with Known-Motif feature extraction method and logistic regression with ℓ_1 penalty to identify putative TF-TF interactions. For a particular experiment with cell type pairs C_1, C_2 and TF T, the 321 features used with this model are motifs that represent TFBSs. Thus, the weights assigned to these features by the logistic regression with ℓ_1 penalty would indicate their relevance in classifying C_1 and C_2 w.r.t. T. If C_1 and C_2 are assigned as positive and negative class, then a feature f_1 with a positive weight may suggest that T is interacting with the TF binding f_1 in C_1 and vice versa. We select only those features that are assigned weights of absolute magnitude greater than or equal to 0.2. This results in 1451 TF-TF putative interactions from 165 experiments. Of those, 64 TF-TF interactions are present

Figure 5–4: Effect of relative number of cell type specific binding sites on the performance of model with Known-Motif method and SVM classifier. Each TF text symbol represents one of the 165 experiments involving that TF and a cell type pair.

in the BioGRID database, which contains experimentally validated protein-protein interactions (§ 2.2). The strongest TF-TF interaction observed is in between SP1 and HNF4A in cell type HepG2. The weight of the motif of HNF4A has the most negative value (- 0.337) for classifying H1-hESC as positive cell type and HepG2 as negative cell type w.r.t. SP1. Thus, the strongest TF-TF interaction observed from our experiments is in between SP1 and HNF4A in cell type HepG2. It has been shown that the interaction between HNF4A and SP1 in HepG2 has an impact on differential transcription regulation of human eosinophil RNases [128]. Moreover, several of the identified TF-TF interactions belong to the same TF family (Table 5-3), which shows the correctness of our model. Often, TFs from same family would

form a complex and initiate the transcription process. Finally, we should note that the BioGRID database is still not complete and many other of our identified TF-TF interactions could be putative. Thus, our model can be used to search for TF-TF interactions, which can be later experimentally verified and studied for biological interests.

Table 5–3: TF-TF interaction

TF Family	**Observed in Cell Type**
ATF	GM12878, HepG2
CEBP	H1-hESC, HeLa-S3, K562, HepG2
E2F	HeLa-S3
STAT	HeLa-S3, K562
USF	H1-hESC, K562

In order to verify that the matches with BioGRID are not random, we create a graph of 1451 TF-TF interactions identified by our experiments, where nodes are TFs and edges are TF-TF interactions. Then we use algorithm 1 to randomize this graph while preserving the degree of nodes.

Algorithm 1 Randomization of Existing Graph with Preservation of node degree. Steps 2-4 are repeated ten thousand times in order to have sufficient randomization.

1: Repeat steps 2-4 ten thousand times

2: Randomly select nodes n_1, n_2 s.t. $n_1 \neq n_2$

3: Randomly select edges (n_1, u), (n_2, v) s.t. $u \neq v$ and edges (n_1, v), (n_2, u) do not exist in the graph

4: Remove edges (n_1, u), (n_2, v) and create edges $(n_1, v), (n_2, u)$ in the graph

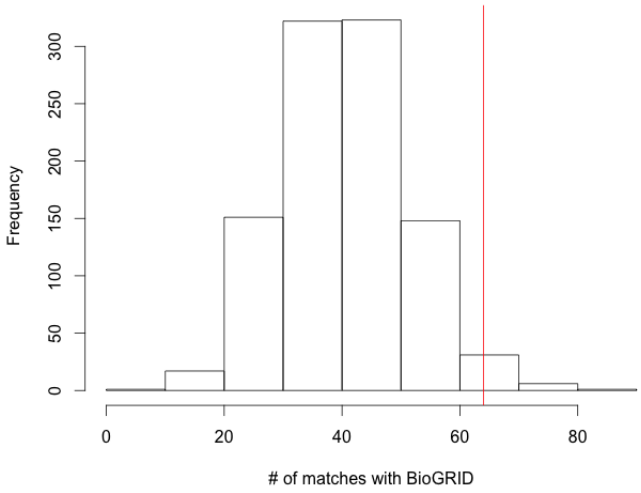


Figure 5–5: Histogram of the number of TF-TF interactions matched with the BioGRID database. The TF-TF interactions identified by the model with the Known-Motif method and the logistic regression with ℓ_1 penalty are randomly shuffled as per algorithm 1 and are matched with the BioGRID database. The red line shows the number of interactions i.e. 64 that are identified by the model without any shuffling.

The TF-TF interactions from the graph obtained from algorithm 1 are matched with the BioGRID database. We repeat this procedure thousand times. Fig. 5-5 shows the histogram of the number of TF-TF interactions matched with the BioGRID

database. The p-value of the number of TF-TF interactions i.e. 64 matched from the original graph is 0.0153, which is statistically significant. Thus, the observed number of matched TF-TF interactions is unlikely to have occurred by chance. We report these interactions in table C1.

5.5 Comparison of Running Time

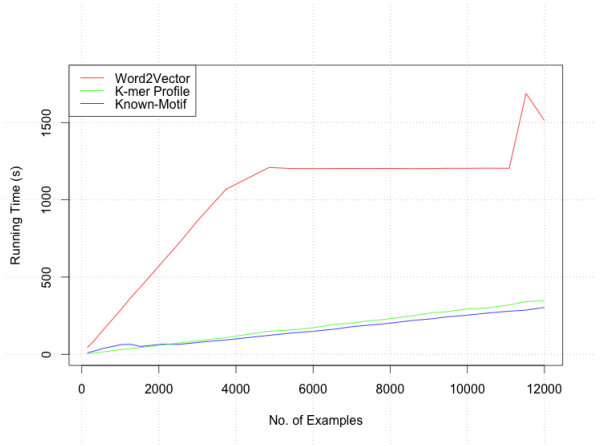


Figure 5–6: Running time of feature extraction methods

We compare the running time of feature extraction methods to produce the input vectors on 64 bit Intel machine with 16 processors (Fig 5-6). The skip-gram model requires additional time for the training of word vectors that results into high running time for Word2Vector method for the given sequences. The k-mer profile and Known-Motif method build feature vectors of length 4096 and 321 respectively. For smaller set of sequences k-mer profile runs fastest, but for larger set of sequences it runs slower than Known-Motif, due to processing required for generating longer feature vector.

CHAPTER 6
Conclusion

Transcription Factors play essential role in gene regulatory networks and their functional behavior can be understood through their binding with genomic regions [113]. For this reason, prediction of TFBSs has become an important research area in bioinformatics. With the advent of experimental methods, such as ChIP-seq, research has been focussed on genome wide mapping of TFBSs, an effort led in particular by the ENCODE Consortium. However, these experiments are expensive and time consuming. For example, ChIP-seq experiments need to be repeated for each TF and in each cell type of interest in order to determine cell-type specific TFBSs. Computational approaches to predict TFBSs have largely been depended upon traditional motif based methods that have several drawbacks leading to high false predictions and are unsuitable for making predictions about cell-type specific TFBSs.

Our study is motivated by the remarkable results from the application of machine learning to TFBSs prediction [7, 10, 141, 62] and we emphasize on developing models oriented toward cell-type specific TFBSs. For a range of cell type and TF combinations, we are able to predict with the mean AUC score of 0.82. Our model works well with the cell types that have relatively large number of cell-type specific

TFBSs for a given TF. In comparison with state-of-the-art prediction of TFBS cell-type specificity [10], our model captures the cell-type specific signatures better with mean AUC score improvement of 0.18.

It has been shown that TF-TF interactions have an impact on transcription process [85, 91, 131, 68]. Our predictive model with Known Motif as features and logistic regression as predictor can be used effectively to identify TF-TF interactions. Our results detect both previously known TF-TF interactions and putative known ones.

There are many databases on TF and TFBSs (§ 2.2) that can be used with our Known-Motif method for the TFBSs prediction. With the recent advances in technology, many species are being sequenced and their genomes are becoming available. The binding behavior of TFs in these situations can be studied with K-Mer and Word2Vector methods. While the former method represents motifs as k-mers, the latter looks for the semantic context of k-mers in the genomic regions.

We should note that all of our machine learning methods developed for the proposed problem (§ 4.1) are comparable. For most of the cases, SVM and logistic regression performs better than KNN, but there is no clear winner.

There are many other factors that are responsible for specific binding of TFs such as histone modifications and DNase accessibility, which are not considered here. The methods developed in this study form a strong foothold for more generative approaches that could include such factors and comprehend the transcription process.

6.1 Future Work

Although we made significant efforts in our work to maximize the accuracy and usefulness of the predictors we proposed, additional avenues could be explored to obtain further improvements. In order to counter lab experiment bias, we restricted every sequence in our training and testing data to be 400 base pairs long. An interesting potential alternative to deal with sequences of unequal lengths could be to generate random sequence of same length as of the given sequence. Then, the number of occurrences of a k-mer in the given sequence can be normalized w.r.t. the number of occurrences of the same k-mer in the random sequence, which effectively normalizes for sequence length and composition bias.

Second, in the Known-Motif method, the HOMER tool requires background sequences to find the instances of known motifs in the given sequences. Currently, we are using same input sequences as the background sequences. There are several combinations that should be tried for background sequences, for example sequences generated from a particular distribution can act as background sequences.

In our current K-Mer profile approach, the word vector of a given sequence is computed by taking the average of the word vectors of the k-mers present in the sequence. Two different sequence with same k-mers, but present in different order, can result into the same word vector. For example, consider two sequences **AACAA** and **CAACA**. Both have same set of 3-mers i.e. **AACAA** ={AAC, ACA, CAA} and **CAACA** = {CAA, AAC, ACA}. Our current approach would assign same word vectors to these different sequences. A possibly better approach could be to concatenate the word vector of k-mers in the order of their presence in the sequence.

This would result into a high-dimensional feature space for the given sequences. We can use dimensionality reduction techniques such as principle component analysis [65] to retrieve the desired number of features from this high-dimensional concatenated word vectors.

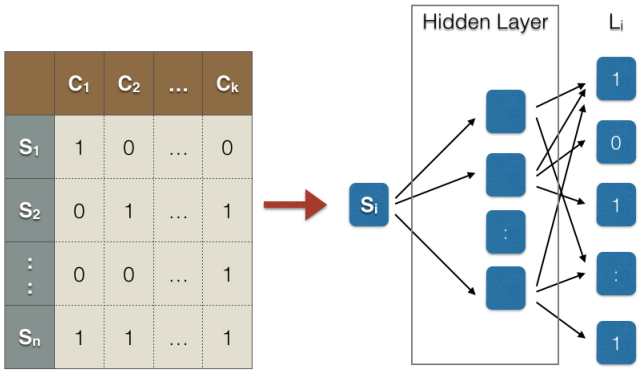


Figure 6–1: Extension of current model for the prediction of cell types for TFBSs of a particular TF T involving more than two cell types. $C_1, C_2, ..., C_k$ are k cell types and $S_1, S_2, ..., S_n$ are n genomic regions. The binary value indicates whether the region is bound by T or not with 0 and 1 being not bound and bound respectively. The required model can be developed by learning a neural network [57] with genomic region S_i as the input layer and its k-length label vector L_i as the output layer.

Finally, our machine learning approach is designed for the binary version ($k = 2$) of the problem mentioned in § 4.1. We can extend our approach to k (> 2) number of cells (Fig 6-1). We need to assign a binary vector L_i of length k to a genomic region i obtained from ENCODE. Value at position j of L_i would indicate whether the region i is bound in cell type j or not. Then we can train a neural network [57] to find the function that maps the region i to their label L_i. For the desired neural network, region i and its label L_i would be the input and output layer respectively.

Appendix A

Table A1: Parameter Settings

Known-motif model	regularization parameter	gamma	k
logistic regression with ℓ_2 penalty	0.001	NA	NA
svm	1	0.001	NA
knn	NA	NA	20
logistic regression with ℓ_1 penalty	0.1	NA	NA
k-mer profile model	cost	gamma	k
logistic regression with ℓ_2 penalty	0.0001	NA	NA
svm	1	0.0001	NA
knn	NA	NA	200
logistic regression with ℓ_1 penalty	0.1	NA	NA
word2vector model	cost	gamma	k
logistic regression with ℓ_2 penalty	0.01	NA	NA
svm	10	0.0001	NA
knn	NA	NA	200
logistic regression with ℓ_1 penalty	0.1	NA	NA

Table A2: Dataset

TF	Cell Type 1 C_1	Cell Type 2 C_2	# C_1 specific TFBSs	# C_2 specific TFBSs	# constitutive TFBSs
ARID3A	K562	HepG2	6896	15340	2071
ATF2	GM12878	H1-hESC	20371	3886	1769
ATF3	H1-hESC	K562	1560	12698	3237
ATF3	H1-hESC	HepG2	2810	1298	1987
ATF3	GM12878	H1-hESC	515	3642	1155
ATF3	GM12878	K562	443	14708	1227
ATF3	GM12878	HepG2	609	2224	1061
ATF3	K562	HepG2	13566	916	2369
BCL11A	GM12878	H1-hESC	17768	2428	90
BCL3	GM12878	K562	14893	1252	331
BCLAF1	GM12878	K562	4610	2937	1396
BDP1	HeLa-S3	K562	221	283	287
BHLHE40	GM12878	K562	8379	16801	5521
BHLHE40	GM12878	HepG2	9655	10313	4245
BHLHE40	K562	HepG2	15781	8017	6541
BRCA1	H1-hESC	HeLa-S3	773	6837	1246
BRCA1	H1-hESC	HepG2	1162	635	857
BRCA1	HeLa-S3	HepG2	7029	438	1054
BRF1	HeLa-S3	K562	72	98	121

BRF2	HeLa-S3	K562	281	1064	22
CEBPB	H1-hESC	HeLa-S3	7047	52336	8495
CEBPB	H1-hESC	K562	6769	29853	8773
CEBPB	GM12878	H1-hESC	5230	15156	386
CEBPB	GM12878	HeLa-S3	4487	59702	1129
CEBPB	GM12878	K562	4845	37855	771
CEBPB	GM12878	HepG2	4960	55925	656
CEBPB	HeLa-S3	K562	42872	20667	17959
CEBPB	HeLa-S3	HepG2	39307	35057	21524
CEBPB	K562	HepG2	17979	35934	20647
CHD1	H1-hESC	K562	4141	6124	1957
CHD1	GM12878	H1-hESC	5103	5010	1088
CHD1	GM12878	K562	4562	6452	1629
CHD2	H1-hESC	HeLa-S3	2567	15669	4183
CHD2	H1-hESC	K562	3277	4199	3473
CHD2	H1-hESC	HepG2	4056	2402	2694
CHD2	GM12878	H1-hESC	11332	2951	3799
CHD2	GM12878	HeLa-S3	8416	13137	6715
CHD2	GM12878	K562	10922	3463	4209
CHD2	GM12878	HepG2	11923	1888	3208
CHD2	HeLa-S3	K562	15199	3019	4653
CHD2	HeLa-S3	HepG2	16227	1471	3625

CHD2	K562	HepG2	4602	2026	3070
E2F4	GM12878	HeLa-S3	1934	1163	1470
E2F4	GM12878	K562	1142	5719	2262
E2F4	HeLa-S3	K562	667	6015	1966
E2F6	HeLa-S3	K562	779	19653	3482
ELF1	GM12878	K562	9280	14549	12790
ELF1	GM12878	HepG2	12620	8265	9450
ELF1	K562	HepG2	16345	6721	10994
ELK1	GM12878	HeLa-S3	2901	2173	2571
ELK1	GM12878	K562	3608	1041	1864
ELK1	HeLa-S3	K562	2948	1109	1796
ETS1	GM12878	K562	1452	7593	2638
EZH2	H1-hESC	HeLa-S3	3009	1631	30
EZH2	H1-hESC	HepG2	2384	1871	655
EZH2	GM12878	HeLa-S3	2184	1629	32
EZH2	HeLa-S3	HepG2	1593	2458	68
FOSL1	H1-hESC	K562	764	10824	349
GTF2F1	H1-hESC	HeLa-S3	1974	9718	1519
GTF2F1	H1-hESC	K562	2316	2322	1177
GTF2F1	HeLa-S3	K562	9692	1954	1545
HDAC2	H1-hESC	K562	5058	6080	581
HDAC2	H1-hESC	HepG2	4558	17586	1081

HDAC2	K562	HepG2	5446	17452	1215
IRF3	HeLa-S3	HepG2	1040	149	533
MAZ	GM12878	HeLa-S3	10739	5586	6866
MAZ	GM12878	K562	6713	20904	10892
MAZ	GM12878	HepG2	11071	5034	6534
MAZ	HeLa-S3	K562	3301	22645	9151
MAZ	HeLa-S3	HepG2	6578	5694	5874
MAZ	K562	HepG2	23515	3287	8281
MEF2A	GM12878	K562	16250	4295	1330
NFIC	GM12878	HepG2	23903	13071	1877
PML	GM12878	K562	8794	9059	5425
RBBP5	H1-hESC	K562	8091	5762	5352
RFX5	H1-hESC	HeLa-S3	786	18285	897
RFX5	H1-hESC	K562	1048	1544	635
RFX5	H1-hESC	HepG2	704	5000	979
RFX5	GM12878	H1-hESC	3589	956	727
RFX5	GM12878	HeLa-S3	1940	16806	2376
RFX5	GM12878	K562	3414	1277	902
RFX5	GM12878	HepG2	2322	3985	1994
RFX5	HeLa-S3	K562	17882	879	1300
RFX5	HeLa-S3	HepG2	16242	3039	2940
RFX5	K562	HepG2	1031	4831	1148

RPC155	HeLa-S3	K562	1662	573	571
RXRA	H1-hESC	HepG2	916	16611	389
RXRA	GM12878	H1-hESC	1605	1208	97
RXRA	GM12878	HepG2	1337	16635	365
SIN3A	GM12878	H1-hESC	3871	13913	5278
SIX5	H1-hESC	K562	883	1634	2503
SIX5	GM12878	H1-hESC	2244	863	2523
SIX5	GM12878	K562	1459	829	3308
SMC3	GM12878	HeLa-S3	7870	16918	22623
SMC3	GM12878	K562	11615	4705	18878
SMC3	GM12878	HepG2	10137	10425	20356
SMC3	HeLa-S3	K562	20323	4365	19218
SMC3	HeLa-S3	HepG2	17901	9141	21640
SMC3	K562	HepG2	5704	12902	17879
SP1	H1-hESC	K562	10972	3122	4033
SP1	H1-hESC	HepG2	10463	20678	4542
SP1	GM12878	H1-hESC	12596	9571	5434
SP1	GM12878	K562	13605	2730	4425
SP1	GM12878	HepG2	13027	20217	5003
SP1	K562	HepG2	3504	21569	3651
SP2	H1-hESC	K562	456	1178	1913
SP2	H1-hESC	HepG2	1667	1919	702

SP2	K562	HepG2	2384	1914	707
SRF	H1-hESC	K562	3026	2638	2074
SRF	H1-hESC	HepG2	3340	3549	1760
SRF	GM12878	H1-hESC	6432	2991	2109
SRF	GM12878	K562	6076	2247	2465
SRF	GM12878	HepG2	6544	3312	1997
SRF	K562	HepG2	3005	3602	1707
STAT1	GM12878	HeLa-S3	1404	14734	307
STAT1	GM12878	K562	1661	2143	50
STAT1	HeLa-S3	K562	14682	1109	359
STAT3	GM12878	HeLa-S3	5782	13226	491
STAT5A	GM12878	K562	6401	9122	597
TAF1	H1-hESC	K562	8497	4595	8869
TAF1	H1-hESC	HepG2	7782	4863	9584
TAF1	GM12878	H1-hESC	5008	9221	8145
TAF1	GM12878	HeLa-S3	5847	7150	7306
TAF1	GM12878	K562	5420	5731	7733
TAF1	GM12878	HepG2	5488	6782	7665
TAF1	HeLa-S3	HepG2	6521	6512	7935
TAF1	K562	HepG2	5028	6011	8436
TAF7	H1-hESC	K562	8225	1512	1843
TBP	H1-hESC	HeLa-S3	9800	11265	6496

TBP	H1-hESC	K562	8169	8350	8127
TBP	H1-hESC	HepG2	9241	6102	7055
TBP	GM12878	H1-hESC	8675	10724	5572
TBP	GM12878	HeLa-S3	8824	12338	5423
TBP	GM12878	K562	8543	10773	5704
TBP	GM12878	HepG2	9076	7986	5171
TBP	HeLa-S3	K562	11080	9796	6681
TBP	HeLa-S3	HepG2	11487	6883	6274
TBP	K562	HepG2	9564	6244	6913
TCF12	H1-hESC	HepG2	7713	1950	111
TCF12	GM12878	H1-hESC	18693	6209	1615
TCF12	GM12878	HepG2	20128	1881	180
TCF7L2	HeLa-S3	HepG2	2601	2101	231
TEAD4	H1-hESC	K562	16719	27459	3108
TEAD4	H1-hESC	HepG2	17159	12071	2668
TEAD4	K562	HepG2	27923	12095	2644
USF2	H1-hESC	HeLa-S3	3729	9058	3217
USF2	H1-hESC	K562	4941	1074	2005
USF2	H1-hESC	HepG2	4180	3519	2766
USF2	GM12878	H1-hESC	5877	3815	3131
USF2	GM12878	HeLa-S3	5544	8811	3464
USF2	GM12878	K562	7108	1179	1900

USF2	GM12878	HepG2	6265	3542	2743
USF2	HeLa-S3	K562	10336	1140	1939
USF2	HeLa-S3	HepG2	9152	3162	3123
USF2	K562	HepG2	1186	4392	1893
YY1	H1-hESC	K562	9914	14941	8192
YY1	H1-hESC	HepG2	10992	9952	7114
YY1	GM12878	H1-hESC	20613	8613	9493
YY1	GM12878	K562	18628	11655	11478
YY1	GM12878	HepG2	19651	6611	10455
YY1	K562	HepG2	13607	7540	9526
ZBTB33	GM12878	K562	962	2101	1164
ZBTB33	GM12878	HepG2	942	1670	1184
ZBTB7A	K562	HepG2	15994	4580	4489
ZZZ3	GM12878	HeLa-S3	604	151	97

Appendix B

Table B1: Model: Known-Motif + SVM

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
FOSL1	H1-hESC	K562	506	321	0.97
SP2	K562	HepG2	1264	321	0.94
SP2	H1-hESC	HepG2	1102	321	0.94
EZH2	H1-hESC	HeLa-S3	1078	321	0.94
CEBPB	GM12878	HepG2	3274	321	0.94
CEBPB	GM12878	H1-hESC	3452	321	0.94
TCF12	H1-hESC	HepG2	1288	321	0.93
TCF12	GM12878	HepG2	1242	321	0.93
CEBPB	GM12878	K562	3198	321	0.93
BCL11A	GM12878	H1-hESC	1604	321	0.93
ELK1	HeLa-S3	K562	732	321	0.71
SP2	H1-hESC	K562	302	321	0.7
RFX5	GM12878	HepG2	1534	321	0.7
MAZ	HeLa-S3	HepG2	3760	321	0.7
ELK1	GM12878	HeLa-S3	1436	321	0.7
E2F4	GM12878	HeLa-S3	768	321	0.69
GTF2F1	H1-hESC	K562	1530	321	0.66
BRF2	HeLa-S3	K562	186	321	0.66
BRCA1	H1-hESC	HepG2	420	321	0.65
ELK1	GM12878	K562	688	321	0.63

Table B2: Model: Known-Motif + logistic regression with ℓ_2 penalty

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
FOSL1	H1-hESC	K562	506	321	0.96
CEBPB	GM12878	HepG2	3274	321	0.94
CEBPB	GM12878	H1-hESC	3452	321	0.94
EZH2	H1-hESC	HeLa-S3	1078	321	0.93
BCL11A	GM12878	H1-hESC	1604	321	0.93
TCF12	H1-hESC	HepG2	1288	321	0.92
TCF12	GM12878	HepG2	1242	321	0.92
SP2	K562	HepG2	1264	321	0.92
SP2	H1-hESC	HepG2	1102	321	0.92
CEBPB	GM12878	K562	3198	321	0.92
SIX5	GM12878	K562	548	321	0.71
RFX5	HeLa-S3	K562	582	321	0.71
SP2	H1-hESC	K562	302	321	0.7
MAZ	HeLa-S3	HepG2	3760	321	0.7
ELK1	GM12878	HeLa-S3	1436	321	0.7
E2F4	GM12878	HeLa-S3	768	321	0.7
GTF2F1	H1-hESC	K562	1530	321	0.67
BRF2	HeLa-S3	K562	186	321	0.67
BRCA1	H1-hESC	HepG2	420	321	0.67
ELK1	GM12878	K562	688	321	0.63

Table B3: Model: Known-Motif + logistic regression with ℓ_1 penalty

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
FOSL1	H1-hESC	K562	506	321	0.97
SP2	H1-hESC	HepG2	1102	321	0.96
CEBPB	GM12878	HepG2	3274	321	0.95
SP2	K562	HepG2	1264	321	0.95
EZH2	H1-hESC	HeLa-S3	1078	321	0.94
CEBPB	GM12878	H1-hESC	3452	321	0.94
BCL11A	GM12878	H1-hESC	1604	321	0.93
CEBPB	GM12878	K562	3198	321	0.93
TCF12	GM12878	HepG2	1242	321	0.93
TCF12	H1-hESC	HepG2	1288	321	0.93
ELK1	GM12878	HeLa-S3	1436	321	0.7
ELK1	HeLa-S3	K562	732	321	0.7
MAZ	HeLa-S3	HepG2	3760	321	0.7
E2F4	GM12878	HeLa-S3	768	321	0.69
SIX5	GM12878	K562	548	321	0.68
SP2	H1-hESC	K562	302	321	0.67
GTF2F1	H1-hESC	K562	1530	321	0.66
BRCA1	H1-hESC	HepG2	420	321	0.65
BRF2	HeLa-S3	K562	186	321	0.64
ELK1	GM12878	K562	688	321	0.61

Table B4: Model: Known-Motif + KNN

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
FOSL1	H1-hESC	K562	506	321	0.92
EZH2	H1-hESC	HeLa-S3	1078	321	0.89
SIX5	GM12878	H1-hESC	570	321	0.87
SP2	K562	HepG2	1264	321	0.86
SP2	H1-hESC	HepG2	1102	321	0.86
TBP	GM12878	H1-hESC	3960	321	0.85
TCF12	GM12878	HepG2	1242	321	0.84
TAF1	HeLa-S3	HepG2	3960	321	0.84
STAT1	GM12878	K562	1098	321	0.84
SIX5	H1-hESC	K562	584	321	0.84
CHD2	H1-hESC	HepG2	1586	321	0.62
BRF2	HeLa-S3	K562	186	321	0.61
ELK1	GM12878	K562	688	321	0.6
CEBPB	H1-hESC	K562	3960	321	0.6
EZH2	H1-hESC	HepG2	1236	321	0.59
ELK1	GM12878	HeLa-S3	1436	321	0.59
E2F4	GM12878	HeLa-S3	768	321	0.59
MAZ	HeLa-S3	HepG2	3760	321	0.58
GTF2F1	H1-hESC	K562	1530	321	0.56
BRCA1	H1-hESC	HepG2	420	321	0.56

Table B5: Model: K-mer Profile + SVM

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	4096	0.98
SP2	K562	HepG2	1264	4096	0.94
SP2	H1-hESC	HepG2	1102	4096	0.94
FOSL1	H1-hESC	K562	506	4096	0.94
CEBPB	GM12878	H1-hESC	3452	4096	0.93
TCF12	H1-hESC	HepG2	1288	4096	0.92
TCF12	GM12878	HepG2	1242	4096	0.92
SIX5	GM12878	H1-hESC	570	4096	0.92
CEBPB	GM12878	HepG2	3274	4096	0.92
BCL11A	GM12878	H1-hESC	1604	4096	0.92
MAZ	HeLa-S3	HepG2	3760	4096	0.69
ELK1	HeLa-S3	K562	732	4096	0.69
ELK1	GM12878	HeLa-S3	1436	4096	0.69
USF2	K562	HepG2	784	4096	0.68
RFX5	HeLa-S3	K562	582	4096	0.68
BRCA1	H1-hESC	HepG2	420	4096	0.68
E2F4	GM12878	HeLa-S3	768	4096	0.67
BRF2	HeLa-S3	K562	186	4096	0.67
GTF2F1	H1-hESC	K562	1530	4096	0.63
ELK1	GM12878	K562	688	4096	0.6

Table B6: Model: K-mer Profile + logistic regression with ℓ_2 penalty

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	4096	0.97
CEBPB	GM12878	H1-hESC	3452	4096	0.93
SP2	K562	HepG2	1264	4096	0.92
SP2	H1-hESC	HepG2	1102	4096	0.92
FOSL1	H1-hESC	K562	506	4096	0.92
BCL11A	GM12878	H1-hESC	1604	4096	0.92
TEAD4	H1-hESC	K562	3960	4096	0.91
TBP	GM12878	H1-hESC	3960	4096	0.91
NFIC	GM12878	HepG2	3960	4096	0.91
CEBPB	GM12878	HepG2	3274	4096	0.91
RFX5	GM12878	HepG2	1534	4096	0.69
ELK1	HeLa-S3	K562	732	4096	0.69
RFX5	HeLa-S3	K562	582	4096	0.68
ELK1	GM12878	HeLa-S3	1436	4096	0.68
USF2	K562	HepG2	784	4096	0.67
E2F4	GM12878	HeLa-S3	768	4096	0.67
BRF2	HeLa-S3	K562	186	4096	0.67
BRCA1	H1-hESC	HepG2	420	4096	0.67
GTF2F1	H1-hESC	K562	1530	4096	0.63
ELK1	GM12878	K562	688	4096	0.6

Table B7: Model: K-mer Profile + logistic regression with ℓ_1 penalty

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	4096	0.97
FOSL1	H1-hESC	K562	506	4096	0.96
SP2	K562	HepG2	1264	4096	0.95
SP2	H1-hESC	HepG2	1102	4096	0.95
SIX5	GM12878	H1-hESC	570	4096	0.92
CEBPB	GM12878	H1-hESC	3452	4096	0.92
TCF12	GM12878	HepG2	1242	4096	0.92
BCL11A	GM12878	H1-hESC	1604	4096	0.91
CEBPB	GM12878	HepG2	3274	4096	0.91
TEAD4	H1-hESC	K562	3960	4096	0.91
ATF3	GM12878	HepG2	402	4096	0.66
RFX5	HeLa-S3	K562	582	4096	0.65
USF2	K562	HepG2	784	4096	0.65
ELK1	HeLa-S3	K562	732	4096	0.65
ELK1	GM12878	HeLa-S3	1436	4096	0.64
ZZZ3	GM12878	HeLa-S3	100	4096	0.64
BRCA1	H1-hESC	HepG2	420	4096	0.63
GTF2F1	H1-hESC	K562	1530	4096	0.61
BRF2	HeLa-S3	K562	186	4096	0.6
ELK1	GM12878	K562	688	4096	0.55

Table B8: Model: K-mer + KNN

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	4096	0.96
TBP	GM12878	H1-hESC	3960	4096	0.89
SP2	K562	HepG2	1264	4096	0.89
FOSL1	H1-hESC	K562	506	4096	0.89
SIX5	GM12878	H1-hESC	570	4096	0.88
SP2	H1-hESC	HepG2	1102	4096	0.87
SIX5	H1-hESC	K562	584	4096	0.87
TCF12	H1-hESC	HepG2	1288	4096	0.86
BCL11A	GM12878	H1-hESC	1604	4096	0.86
TBP	H1-hESC	HeLa-S3	3960	4096	0.85
SIN3A	GM12878	H1-hESC	2556	4096	0.62
RFX5	HeLa-S3	K562	582	4096	0.62
MAZ	HeLa-S3	HepG2	3760	4096	0.62
E2F4	GM12878	HeLa-S3	768	4096	0.62
RBBP5	H1-hESC	K562	3804	4096	0.61
ATF3	GM12878	H1-hESC	340	4096	0.61
USF2	K562	HepG2	784	4096	0.6
GTF2F1	H1-hESC	K562	1530	4096	0.59
ELK1	GM12878	K562	688	4096	0.59
ZZZ3	GM12878	HeLa-S3	100	4096	0.5

Table B9: Model: Word2Vector + SVM

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	500	0.98
SP2	K562	HepG2	1264	500	0.95
SP2	H1-hESC	HepG2	1102	500	0.94
FOSL1	H1-hESC	K562	506	500	0.94
TCF12	GM12878	HepG2	1242	500	0.92
SIX5	GM12878	H1-hESC	570	500	0.92
CEBPB	GM12878	H1-hESC	3452	500	0.92
BCL11A	GM12878	H1-hESC	1604	500	0.92
TCF12	H1-hESC	HepG2	1288	500	0.91
CEBPB	GM12878	HepG2	3274	500	0.91
USF2	HeLa-S3	K562	754	500	0.69
ELK1	HeLa-S3	K562	732	500	0.69
E2F4	GM12878	HeLa-S3	768	500	0.69
BRCA1	H1-hESC	HepG2	420	500	0.69
USF2	K562	HepG2	784	500	0.68
RFX5	GM12878	HepG2	1534	500	0.68
ELK1	GM12878	HeLa-S3	1436	500	0.68
BRF2	HeLa-S3	K562	186	500	0.67
ELK1	GM12878	K562	688	500	0.65
GTF2F1	H1-hESC	K562	1530	500	0.62

Table B10: Model: Word2Vector + logistic regression with ℓ_2 penalty

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	500	0.97
SP2	K562	HepG2	1264	500	0.95
SP2	H1-hESC	HepG2	1102	500	0.94
FOSL1	H1-hESC	K562	506	500	0.94
SIX5	GM12878	H1-hESC	570	500	0.93
TCF12	GM12878	HepG2	1242	500	0.92
CEBPB	GM12878	H1-hESC	3452	500	0.92
BCL11A	GM12878	H1-hESC	1604	500	0.92
TCF12	H1-hESC	HepG2	1288	500	0.91
CEBPB	GM12878	HepG2	3274	500	0.91
USF2	K562	HepG2	784	500	0.68
USF2	HeLa-S3	K562	754	500	0.68
ELK1	HeLa-S3	K562	732	500	0.68
E2F4	GM12878	HeLa-S3	768	500	0.68
BRCA1	H1-hESC	HepG2	420	500	0.68
RFX5	GM12878	HepG2	1534	500	0.67
ELK1	GM12878	HeLa-S3	1436	500	0.67
BRF2	HeLa-S3	K562	186	500	0.65
GTF2F1	H1-hESC	K562	1530	500	0.61
ELK1	GM12878	K562	688	500	0.6

Table B11: Model: Word2Vector + logistic regression with ℓ_1 penalty

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	500	0.97
SP2	K562	HepG2	1264	500	0.94
SP2	H1-hESC	HepG2	1102	500	0.94
FOSL1	H1-hESC	K562	506	500	0.93
SIX5	GM12878	H1-hESC	570	500	0.92
CEBPB	GM12878	H1-hESC	3452	500	0.92
BCL11A	GM12878	H1-hESC	1604	500	0.92
TCF12	GM12878	HepG2	1242	500	0.91
TEAD4	H1-hESC	K562	3960	500	0.91
CEBPB	GM12878	HepG2	3274	500	0.91
USF2	HeLa-S3	K562	754	500	0.68
BRF2	HeLa-S3	K562	186	500	0.68
E2F4	GM12878	HeLa-S3	768	500	0.67
ATF3	GM12878	HepG2	402	500	0.67
RFX5	GM12878	HepG2	1534	500	0.67
ELK1	GM12878	HeLa-S3	1436	500	0.67
ELK1	HeLa-S3	K562	732	500	0.67
USF2	K562	HepG2	784	500	0.67
GTF2F1	H1-hESC	K562	1530	500	0.61
ELK1	GM12878	K562	688	500	0.61

Table B12: Model: Word2Vector + KNN

TF	**Cell type 1**	**Cell type 2**	**Testset Size**	**Input Size**	**AUC**
EZH2	H1-hESC	HeLa-S3	1078	500	0.97
SP2	K562	HepG2	1264	500	0.89
FOSL1	H1-hESC	K562	506	500	0.89
TBP	GM12878	H1-hESC	3960	500	0.88
TAF1	HeLa-S3	HepG2	3960	500	0.88
SIX5	GM12878	H1-hESC	570	500	0.88
BCL11A	GM12878	H1-hESC	1604	500	0.88
SP2	H1-hESC	HepG2	1102	500	0.87
SIX5	H1-hESC	K562	584	500	0.87
BRF1	HeLa-S3	K562	48	500	0.87
USF2	HeLa-S3	K562	754	500	0.64
CHD2	HeLa-S3	HepG2	972	500	0.64
BRCA1	H1-hESC	HepG2	420	500	0.64
RFX5	GM12878	HepG2	1534	500	0.63
ELK1	GM12878	K562	688	500	0.63
USF2	K562	HepG2	784	500	0.62
MAZ	HeLa-S3	HepG2	3760	500	0.62
ELK1	GM12878	HeLa-S3	1436	500	0.62
GTF2F1	H1-hESC	K562	1530	500	0.59
ZZZ3	GM12878	HeLa-S3	100	500	0.48

Appendix C

Table C1: TF1, Cell Type 1 and Cell Type 2 form a machine learning experiment and TF 2 is identified as an important feature for prediction by the model consisting of known method and logistic regresion with l_1 penalty. The TF 1 and TF 2 interactions are matched with the BioGRID database.

TF 1	**Cell Type 1** (+)	**Cell Type 2** (-)	**TF 2**	**Interaction Observed In**	**Reference**
ATF2	GM12878	H1-hESC	ATF3	+	[100]
ATF3	H1-hESC	HepG2	ATF4	-	[130]
ATF3	H1-hESC	HepG2	CEBP	-	[101]
ATF3	GM12878	H1-hESC	ATF4	-	[130]
ATF3	GM12878	K562	CHOP	-	[28]
ATF3	GM12878	HepG2	ATF4	-	[130]
BRCA1	HeLa-S3	HepG2	GATA3	+	[120]
CEBPB	H1-hESC	HeLa-S3	CEBP	+	[75]
CEBPB	GM12878	H1-hESC	CEBP	-	[75]
CEBPB	GM12878	HeLa-S3	CEBP	-	[75]
CEBPB	GM12878	K562	CEBP	-	[75]
CEBPB	GM12878	HepG2	CEBP	-	[75]
CEBPB	HeLa-S3	K562	CEBP	-	[75]
CEBPB	HeLa-S3	HepG2	CEBP	-	[75]
CEBPB	K562	HepG2	CEBP	-	[75]

E2F4	GM12878	HeLa-S3	E2F1	-	[20]
E2F4	HeLa-S3	K562	E2F1	+	[20]
FOSL1	H1-hESC	K562	ATF3	-	[101]
HDAC2	H1-hESC	K562	GATA4	-	[123]
HDAC2	H1-hESC	K562	GFI1B	-	[107]
HDAC2	H1-hESC	HepG2	CEBP	-	[39]
HDAC2	K562	HepG2	CEBP	-	[39]
HDAC2	K562	HepG2	GFI1B	+	[107]
RXRA	H1-hESC	HepG2	CTCF	+	[133]
RXRA	GM12878	H1-hESC	THRA	-	[23]
RXRA	GM12878	HepG2	CTCF	+	[133]
RXRA	GM12878	HepG2	THRA	-	[23]
SIN3A	GM12878	H1-hESC	CTCF	-	[79]
SP1	H1-hESC	K562	GATA4	-	[47]
SP1	H1-hESC	HepG2	HNF4A	-	[128]
SP1	GM12878	HepG2	HNF4A	-	[128]
SP1	GM12878	HepG2	MEF2C	+	[69]
SP1	K562	HepG2	GATA4	+	[47]
SP1	K562	HepG2	HNF4A	-	[128]
SRF	GM12878	K562	GATA4	-	[15]
STAT1	GM12878	HeLa-S3	STAT1	-	[72]
STAT1	GM12878	HeLa-S3	STAT3	-	[115]

STAT1	GM12878	K562	STAT1	-	[72]
STAT1	GM12878	K562	STAT3	-	[115]
STAT1	HeLa-S3	K562	STAT1	+	[72]
STAT1	HeLa-S3	K562	STAT3	+	[115]
STAT5A	GM12878	K562	STAT5	-	[129]
TAF7	H1-hESC	K562	GATA1	-	[100]
TBP	H1-hESC	HepG2	HNF4A	-	[54]
TBP	GM12878	H1-hESC	SPIB	+	[99]
TBP	GM12878	HeLa-S3	SPIB	+	[99]
TBP	GM12878	HepG2	HNF4A	-	[54]
TBP	HeLa-S3	K562	ATF4	+	[73]
TBP	HeLa-S3	K562	SP1	-	[78]
TBP	HeLa-S3	HepG2	HNF4A	-	[54]
TBP	K562	HepG2	HNF4A	-	[54]
TCF12	GM12878	H1-hESC	RUNX1	+	[138]
USF2	H1-hESC	HeLa-S3	USF1	+	[126]
USF2	H1-hESC	HepG2	USF1	+	[126]
USF2	GM12878	H1-hESC	USF1	-	[126]
USF2	GM12878	HepG2	USF2	-	[42]
USF2	HeLa-S3	K562	USF1	-	[126]
USF2	HeLa-S3	HepG2	USF2	-	[42]
USF2	K562	HepG2	USF2	-	[42]

YY1	H1-hESC	HepG2	YY1	+	[127]
YY1	GM12878	H1-hESC	YY1	-	[127]
YY1	GM12878	K562	YY1	-	[127]
YY1	K562	HepG2	YY1	+	[127]
ZBTB33	GM12878	HepG2	TCF4	-	[36]

REFERENCES

[1] http://homer.salk.edu/homer/custom.motifs. [Online; accessed 10-August-2015].

[2] About the biogrid. http://wiki.thebiogrid.org/doku.php/aboutus. [Online; accessed 04-August-2015].

[3] Leslie lab. http://cbio.mskcc.org/leslielab/TFcelltype/. [Online; accessed 09-November-2015].

[4] Transcription and translation. http://www.tokresource.org/tok_classes/biobiobio/biomenu/transcription_translation/. [Online; accessed 11-August-2015].

[5] Weblogo - create sequence logos. http://weblogo.berkeley.edu/logo.cgi. [Online; accessed 13-August-2015].

[6] Word2vec - tool for computing continuous distributed representations of words. - google project hosting. https://code.google.com/p/word2vec/. [Online; accessed 08-August-2015].

[7] Phaedra Agius, Aaron Arvey, William Chang, William Stafford Noble, and Christina Leslie. High Resolution Models of Transcription Factor-DNA Affinities Improve In Vitro and In Vivo Binding Predictions. *PLoS Comput Biol*, 6(9):e1000916, 09 2010.

[8] A Aizerman, Emmanuel M Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[9] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[10] Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome research*, 22(9):1723–1734, 2012.

[11] Gwenael Badis, Michael F Berger, Anthony A Philippakis, Shaheynoor Talukder, Andrew R Gehrke, Savina A Jaeger, Esther T Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, et al. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, 2009.

[12] Lu Bai and Alexandre V Morozov. Gene regulation by nucleosome positioning. *Trends in genetics*, 26(11):476–483, 2010.

[13] Timothy L Bailey. Discovering novel sequence motifs with MEME. *Current Protocols in Bioinformatics*, pages 2–4, 2002.

[14] Iros Barozzi, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular cell*, 54(5):844–857, 2014.

[15] Narasimhaswamy S Belaguli, Jorge L Sepulveda, Vishal Nigam, Frédéric Charron, Mona Nemer, and Robert J Schwartz. Cardiac tissue enriched factors serum response factor and GATA-4 are mutual coregulators. *Molecular and cellular biology*, 20(20):7550–7558, 2000.

[16] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[17] Michael F Berger and Martha L Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols*, 4(3):393–411, 2009.

[18] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11):1429–1435, 2006.

[19] Mark Bieda, Xiaoqin Xu, Michael A Singer, Roland Green, and Peggy J Farnham. Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome research*, 16(5):595–605, 2006.

[20] Ranjit S Bindra, Shannon L Gibson, Alice Meng, Ulrica Westermark, Maria Jasin, Andrew J Pierce, Robert G Bristow, Marie K Classon, and Peter M

Glazer. Hypoxia-induced down-regulation of BRCA1 expression by E2Fs. *Cancer research*, 65(24):11597–11604, 2005.

[21] D Bohmann. Transcription factor phosphorylation: a link between signal transduction and the regulation of gene expression. *Cancer cells (Cold Spring Harbor, NY: 1989)*, 2(11):337–344, 1990.

[22] Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006.

[23] Thomas H Bugge, Jens Pohl, O Lonnoy, and Hendrik G Stunnenberg. RXR alpha, a promiscuous partner of retinoic acid and thyroid hormone receptors. *The EMBO journal*, 11(4):1409, 1992.

[24] Martin D Buhmann. *Radial basis functions: theory and implementations*, volume 12. Cambridge university press, 2003.

[25] Martha L Bulyk et al. Computational prediction of transcription-factor binding site locations. *Genome biology*, 5(1):201–201, 2004.

[26] Martha L Bulyk, Philip LF Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255–1261, 2002.

[27] R Caruana. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Citeseer.

[28] BP Chen, Curt D Wolfgang, and Tsonwin Hai. Analysis of ATF3, a transcription factor induced by physiological stresses and modulated by gadd153/Chop10. *Molecular and cellular biology*, 16(3):1157–1168, 1996.

[29] Jean-Michel Claverie and Stephane Audic. The statistical significance of nucleotide position-weight matrix matches. *Computer applications in the biosciences: CABIOS*, 12(5):431–439, 1996.

[30] B Collins. Collins Concise Dictionary, 1999.

[31] ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

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

[33] D. Cox, J. Little, and D. O'Shea. *Using Algebraic Geometry*. Springer-Verlag, 2005.

[34] Santiago Cuesta-López, Hervé Menoni, Dimitar Angelov, and Michel Peyrard. Guanine radical chemistry reveals the effect of thermal fluctuations in gene promoter regions. *Nucleic acids research*, 39(12):5276–5283, 2011.

[35] William HE Day and FR McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic acids research*, 20(5):1093–1099, 1992.

[36] Beatriz del Valle-Pérez, David Casagolda, Ero Lugilde, Gabriela Valls, Montserrat Codina, Natàlia Dave, Antonio García de Herreros, and Mireia Duñach. Wnt controls the transcriptional activity of Kaiso through CK1ε-dependent phosphorylation of p120-catenin. *Journal of cell science*, 124(13):2298–2309, 2011.

[37] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[38] Marko Djordjevic. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular engineering*, 24(2):179–189, 2007.

[39] Hong Duan, Caroline A Heckman, and Linda M Boxer. Histone deacetylase inhibitors down-regulate bcl-2 expression and induce apoptosis in t (14; 18) lymphomas. *Molecular and cellular biology*, 25(5):1608–1619, 2005.

[40] Joseph G. Ecker and Michael. Kupferschmid. *Introduction to operations research / Joseph G. Ecker, Michael Kupferschmid*. Wiley New York, 1988.

[41] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. Miscellaneous clustering methods. *Cluster Analysis, 5th Edition*, pages 215–255, 2011.

[42] Rob M Ewing, Peter Chu, Fred Elisma, Hongyan Li, Paul Taylor, Shane Climie, Linda McBroom-Cerajewski, Mark D Robinson, Liam O'Connor, Michael Li, et al. Large-scale mapping of human protein–protein interactions by mass spectrometry. *Molecular systems biology*, 3(1):89, 2007.

[43] Peggy J Farnham. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, 10(9):605–616, 2009.

[44] Michael P Fay and Michael A Proschan. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, 4:1, 2010.

[45] Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008.

[46] Andy P. Field. *Analysis of Variance (ANOVA)*. SAGE Publications, Inc., 0 edition, 2007.

[47] Christa E Flück and Walter L Miller. GATA-4 and GATA-6 modulate tissue-specific transcription of the human gene for P450c17 by direct interaction with Sp1. *Molecular Endocrinology*, 18(5):1144–1157, 2004.

[48] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[49] Socorro Gama-Castro, Heladia Salgado, Martin Peralta-Gil, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Hilda Solano-Lira, Verónica Jimenez-Jacinto, Verena Weiss, Jair S García-Sotelo, Alejandra López-Fuentes, et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic acids research*, 39(suppl 1):D98–D105, 2011.

[50] Marcel Geertz and Sebastian J Maerkl. Experimental strategies for studying transcription factor–DNA binding specificities. *Briefings in functional genomics*, page elq023, 2010.

[51] Christopher K Glass and Michael G Rosenfeld. The coregulator exchange in transcriptional functions of nuclear receptors. *Genes & development*, 14(2):121–141, 2000.

[52] Sebastian Glatt, Claudio Alfieri, and Christoph W Müller. Recognizing and remodeling the nucleosome. *Current opinion in structural biology*, 21(3):335–341, 2011.

[53] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the Twenty-eight International Conference on Machine Learning, ICML*, 2011.

[54] Victoria J Green, Efi Kokkotou, and John AA Ladias. Critical structural elements and multitarget protein interactions of the transcriptional activator AF-1 of hepatocyte nuclear factor 4. *Journal of Biological Chemistry*, 273(45):29950–29957, 1998.

[55] Anthony JF Griffiths. *An introduction to genetic analysis*. Macmillan, 2005.

[56] Frank E Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[57] Kevin Gurney. *An introduction to neural networks*. CRC press, 1997.

[58] James A Hanley and Barbara J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[59] Zellig S Harris. Distributional structure. *Word*, 1954.

[60] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[61] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–589, 2010.

[62] Bart Hooghe, Stefan Broos, Frans Van Roy, and Pieter De Bleser. A flexible integrative approach based on random forest improves prediction of transcription factor binding sites. *Nucleic acids research*, 40(14):e106–e106, 2012.

[63] David S Johnson, Wei Li, D Benjamin Gordon, Arindam Bhattacharjee, Bo Curry, Jayati Ghosh, Leonardo Brizuela, Jason S Carroll, Myles Brown, Paul Flicek, et al. Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. *Genome research*, 18(3):393–403, 2008.

[64] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

[65] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[66] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. DNA-binding specificities of human transcription factors. *Cell*, 152(1):327–339, 2013.

[67] Frederick Kinyua Kamanu, Yulia A Medvedeva, Ulf Schaefer, Boris R Jankovic, John AC Archer, and Vladimir B Bajic. Mutations and binding sites of human transcription factors. *Frontiers in genetics*, 3, 2012.

[68] Malka Kitayner, Haim Rozenberg, Remo Rohs, Oded Suad, Dov Rabinovich, Barry Honig, and Zippora Shakked. Diversity in DNA recognition by p53 revealed by crystal structures with hoogsteen base pairs. *Nature structural & molecular biology*, 17(4):423–429, 2010.

[69] Dimitri Krainc, Guang Bai, Shu-ichi Okamoto, Maria Carles, John W Kusiak, Roger N Brent, and Stuart A Lipton. Synergistic Activation of the N-Methyl-D-aspartate Receptor Subunit 1 Promoter by Myocyte Enhancer Factor 2C and Sp1. *Journal of Biological Chemistry*, 273(40):26218–26224, 1998.

[70] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[71] Allan Lazarovici, Tianyin Zhou, Anthony Shafer, Ana Carolina Dantas Machado, Todd R Riley, Richard Sandstrom, Peter J Sabo, Yan Lu, Remo Rohs, John A Stamatoyannopoulos, et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proceedings of the National Academy of Sciences*, 110(16):6376–6381, 2013.

[72] Xiaoxia Li, Stewart Leung, Sajjad Qureshi, James E Darnell, and George R Stark. Formation of STAT1-STAT2 heterodimers and their role in the activation of IRF-1 gene transcription by interferon. *Journal of Biological Chemistry*, 271(10):5790–5794, 1996.

[73] Guosheng Liang and Tsonwin Hai. Characterization of human activating transcription factor 4, a transcriptional activator that interacts with multiple domains of cAMP-responsive element-binding protein (CREB)-binding protein (CBP). *Journal of Biological Chemistry*, 272(38):24088–24095, 1997.

[74] Michael Lieberman and Allan D Marks. *Marks' basic medical biochemistry: a clinical approach*. Lippincott Williams & Wilkins, 2009.

[75] Lin Lin, Yong Qian, Xianglin Shi, and Yan Chen. Induction of a cell stress response gene RTP801 by DNA damaging agent methyl methanesulfonate through CCAAT/enhancer binding protein. *Biochemistry*, 44(10):3909–3914, 2005.

[76] Edison T Liu, Sebastian Pott, and Mikael Huss. Q&A: ChIP-seq technologies and the study of gene regulation. *BMC biology*, 8(1):56, 2010.

[77] Richard G Lomax and Debbie L Hahs-Vaughn. *Statistical concepts: a second course*. Routledge, 2013.

[78] Arianna Loregian, Katia Bortolozzo, Silvia Boso, Antonella Caputo, and Giorgio Palù. Interaction of Sp1 transcription factor with HIV-1 Tat protein: looking for cellular partners. *FEBS letters*, 543(1):61–65, 2003.

[79] Marcus Lutz, Les J Burke, Guillermo Barreto, Frauke Goeman, Heiko Greb, Rüdiger Arnold, Holger Schultheiß, Alexander Brehm, Tony Kouzarides, Victor Lobanenkov, et al. Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic acids research*, 28(8):1707–1713, 2000.

[80] Sebastian J Maerkl and Stephen R Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007.

[81] Sebastian J Maerkl and Stephen R Quake. Experimental determination of the evolvability of a transcription factor. *Proceedings of the National Academy of Sciences*, 106(44):18650–18655, 2009.

[82] Brendan Maher. ENCODE: The human encyclopaedia. *Nature*, 489(7414):46, 2012.

[83] Tsz-Kwong Man and Gary D Stormo. Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence

relative affinity (QuMFRA) assay. *Nucleic acids research*, 29(12):2471–2478, 2001.

[84] Joan Massagué, Joan Seoane, and David Wotton. Smad transcription factors. *Genes & development*, 19(23):2783–2810, 2005.

[85] Sebastiaan H Meijsing, Miles A Pufall, Alex Y So, Darren L Bates, Lin Chen, and Keith R Yamamoto. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324(5925):407–410, 2009.

[86] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[87] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *CoRR*, abs/1310.4546, 2013.

[88] Joanna A Miller and Jonathan Widom. Collaborative competition mechanism for gene activation in vivo. *Molecular and cellular biology*, 23(5):1623–1632, 2003.

[89] Leonid A Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences*, 107(52):22534–22539, 2010.

[90] José Carlos Ribeiro Pacheco. PGP: prokaryote gene prediction software. 2013.

[91] Daniel Panne. The enhanceosome. *Current opinion in structural biology*, 18(2):236–242, 2008.

[92] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[93] Len A Pennacchio, Wendy Bickmore, Ann Dean, Marcelo A Nobrega, and Gill Bejerano. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4):288–295, 2013.

[94] E Pennisi. Genomics. encode project writes eulogy for junk dna. *Science*, 2012.

[95] Elizabeth Pennisi. ENCODE project writes eulogy for junk DNA. *Science*, 337:1159, 2012.

[96] Erik Pettersson, Joakim Lundeberg, and Afshin Ahmadian. Generations of sequencing technologies. *Genomics*, 93(2):105–111, 2009.

[97] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3):447–455, 2011.

[98] Daniel Quang and Xiaohui Xie. EXTREME: an online EM algorithm for motif discovery. *Bioinformatics*, 30(12):1667–1673, 2014.

[99] Sridhar Rao, Amy Matsumura, Jung Yoon, and M Celeste Simon. SPI-B activates transcription via a unique proline, serine, and threonine domain and exhibits DNA binding affinity differences from PU.1. *Journal of Biological Chemistry*, 274(16):11115–11124, 1999.

[100] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

[101] Aaron W Reinke, Jiyeon Baek, Orr Ashenberg, and Amy E Keating. Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science*, 340(6133):730–734, 2013.

[102] Jean-Jack M Riethoven. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. In *Computational Biology of Transcription Factor Binding*, pages 33–42. Springer, 2010.

[103] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–657, 2007.

[104] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.

[105] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5:3, 1988.

[106] Victor S Ryaben'kii and Semyon V Tsynkov. *A theoretical introduction to numerical analysis*. CRC Press, 2006.

[107] Shireen Saleque, Jonghwan Kim, Heather M Rooke, and Stuart H Orkin. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Molecular cell*, 27(4):562–572, 2007.

[108] Joseph Sambrook and David W Russell. Fragmentation of DNA by sonication. *Cold Spring Harbor Protocols*, 2006(4):pdb–prot4538, 2006.

[109] 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.

[110] Gregg L Semenza. *Transcription factors and human disease*. Number 37. Oxford University Press, 1998.

[111] Trevor Siggers, Michael H Duyzend, Jessica Reddy, Sidra Khan, and Martha L Bulyk. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Molecular systems biology*, 7(1):555, 2011.

[112] Matthew Slattery, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen J Bussemaker, et al. Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*, 147(6):1270–1282, 2011.

[113] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences*, 39(9):381–399, 2014.

[114] Richard Socher, Cliff C. Lin, Andrew Y. Ng, and Christopher D. Manning. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2011.

[115] K Spiekermann, S Biethahn, S Wilde, W Hiddemann, and F Alves. Constitutive activation of STAT transcription factors in acute myelogenous leukemia. *European journal of haematology*, 67(2):63–71, 2001.

[116] Rodger Staden. Methods for calculating the probabilities of finding patterns in sequences. *Computer applications in the biosciences: CABIOS*, 5(2):89–96, 1989.

[117] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[118] Gary D Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[119] Vladimir B Teif and Karsten Rippe. Statistical–mechanical lattice models for protein–DNA binding in chromatin. *Journal of Physics: Condensed Matter*, 22(41):414105, 2010.

[120] D Tkocz, NT Crawford, NE Buckley, FB Berry, RD Kennedy, JJ Gorski, DP Harkin, and PB Mullan. BRCA1 and GATA3 corepress FOXC1 to inhibit the pathogenesis of basal-like breast cancers. *Oncogene*, 31(32):3667–3678, 2012.

[121] Andrija Tomovic and Edward J Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, 2007.

[122] Victor Trevino, Francesco Falciani, and Hugo A Barrera-Saldaña. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Molecular Medicine*, 13(9-10):527, 2007.

[123] Chinmay M Trivedi, Wenting Zhu, Qiaohong Wang, Cheng Jia, Hae Jin Kee, Li Li, Sridhar Hannenhalli, and Jonathan A Epstein. Hopx and Hdac2 interact to modulate Gata4 acetylation and embryonic cardiac myocyte proliferation. *Developmental cell*, 19(3):450–459, 2010.

[124] Peter D. Turney. Distributional semantics beyond words: Supervised learning of analogy and paraphrase. *CoRR*, abs/1310.5042, 2013.

[125] Peter D. Turney and Patrick Pantel. From Frequency to Meaning: Vector Space Models of Semantics. *CoRR*, abs/1003.1141, 2010.

[126] Benoît Viollet, Anne-Marie Lefrançois-Martinez, Alexandra Henrion, Axel Kahn, Michel Raymondjean, and Antoine Martinez. Immunochemical characterization and transacting properties of upstream stimulatory factor isoforms. *Journal of Biological Chemistry*, 271(3):1405–1415, 1996.

[127] Chi-Chung Wang, Meng-Feng Tsai, Ting-Hao Dai, Tse-Ming Hong, Wing-Kai Chan, Jeremy JW Chen, and Pan-Chyr Yang. Synergistic activation of the tumor suppressor, HLJ1, by the transcription factors YY1 and activator protein 1. *Cancer research*, 67(10):4816–4826, 2007.

[128] Hsiu-Yu Wang, Po-Chun Ho, Chung-Yu Lan, and Margaret Dah-Tsyr Chang. Transcriptional regulation of human eosinophil RNase2 by the liver-enriched hepatocyte nuclear factor 4. *Journal of cellular biochemistry*, 106(2):317–326, 2009.

[129] Jian Wang, Keke Huo, Lixin Ma, Liujun Tang, Dong Li, Xiaobi Huang, Yanzhi Yuan, Chunhua Li, Wei Wang, Wei Guan, et al. Toward an understanding of the protein interaction network of the human liver. *Molecular systems biology*, 7(1):536, 2011.

[130] Qiuyan Wang, Helena Mora-Jensen, Marc A Weniger, Patricia Perez-Galan, Chris Wolford, Tsonwin Hai, David Ron, Weiping Chen, William Trenkle, Adrian Wiestner, et al. ERAD inhibitors integrate ER stress with an epigenetic mechanism to activate BH3-only protein NOXA in cancer cells. *Proceedings of the National Academy of Sciences*, 106(7):2200–2205, 2009.

[131] Todd Wasson and Alexander J Hartemink. An ensemble model of competitive multi-factor binding of the genome. *Genome research*, 19(11):2101–2112, 2009.

[132] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling Up to Large Vocabulary Image Annotation. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2764–2770. AAAI Press, 2011.

[133] Oliver Weth, Christine Weth, Marek Bartkuhn, Joerg Leers, Florian Uhle, Rainer Renkawitz, and L Tora. Modular insulators: genome wide search for composite CTCF/thyroid hormone receptor binding sites. *PLoS One*, 5(4):e10119, 2010.

[134] Michael A White, Davis S Parker, Scott Barolo, and Barak A Cohen. A model of spatially restricted transcription in opposing gradients of activators and repressors. *Molecular systems biology*, 8(1):614, 2012.

[135] Simon T Whiteside and Stephen Goodbourn. Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. *Journal of Cell Science*, 104(4):949–955, 1993.

[136] Edgar Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in bioinformatics*, 9(4):326–332, 2008.

[137] Jian Yan, Martin Enge, Thomas Whitington, Kashyap Dave, Jianping Liu, Inderpreet Sur, Bernhard Schmierer, Arttu Jolma, Teemu Kivioja, Minna Taipale, et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–813, 2013.

[138] Jinsong Zhang, Markus Kalkum, Soichiro Yamamura, Brian T Chait, and Robert G Roeder. E protein silencing by the leukemogenic AML1-ETO fusion protein. *Science*, 305(5688):1286–1289, 2004.

[139] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of ChIP-seq (MACS). *Genome biology*, 9(9):R137, 2008.

[140] Yue Zhao, David Granas, and Gary D Stormo. Inferring binding energies from selected binding sites. *PLoS Comput Biol*, 5(12):e1000590–e1000590, 2009.

[141] Qing Zhou and Jun S Liu. Extracting sequence features to predict protein–DNA interactions: a comparative study. *Nucleic acids research*, 36(12):4137–4148, 2008.