TopAffy: Predicting transcription factors DNA-binding specificities using a general topological method

by

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

Department of *Computer Science*

Memorial University of Newfoundland

September 2021

St. John's Newfoundland

Abstract

Transcription factors (TFs) recognize and bind to specific DNA sequences. Knowing the binding specificity of TFs is crucial to understand gene regulation and how genetic differences in the DNA sequence of TF binding sites affect TF DNA binding activity. However, the transcription factor binding preferences of only 1% of all eukaryotic TFs are known. Computational prediction of TF binding preferences is an affordable and efficient way to increase the number of known binding preferences. Most bioinformatic tools for predicting the binding preferences of TFs require as input the binding preferences of related TFs. However, there are TF families for which very little experimental data is available. In this work, we present TopAffy, a new approach for predicting TF 8-mer binding profiles. TopAffy constructs a stochastic topological representation of DNA-binding domain sequences and learns a numerical representation of the binding preferences of neighbouring amino acid pairs. TopAffy's main contribution is to construct a family-independent model which can be used to predict the 8-mer binding profile for TF families for which no experimental data is yet available. TopAffy's predictive performance is comparable to the performance of state-of-the-art family-specific approaches. Our results demonstrate that it is possible to learn a general model of binding specificities suitable for predicting binding preferences for a number of TF families.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

A transcription factor (Figure 1.2) is a protein that binds to a specific DNA sequence and regulates the transcription of specific genes. Transcription is the process of copying a DNA segment into an RNA molecule. The DNA sequences that transcription factors bind to are called binding sites. Finding the patterns transcription factors bind to is one of the first steps to understand transcriptional regulation [2]. There are several approaches to experimentally determine these patterns, such as chromatin immunoprecipitation (ChIP)-based approaches [3] and Protein Binding Micro-array (PBM) [4].

See Table 1 of Lambert *et al.* [2] reproduced in Figure 1.1 for a summary of experimental methods for determining transcription factor binding specificities

Deciphering transcriptional regulation is essential because it leads to a better understanding of and can aid in research on genes that are involved in cancers [5,

		Method	Description	Features			
				Capability of de novo **motif discovery** (approx. length in base pairs with high information content)	Identifies **genomic binding locations** of a TF	Can measure effect of **CpG Methylation**	Can measure **cooperative binding and/or multimers**
High-throughput	*In Vitro Methods*	Protein Binding Microarray (PBM)	A GST-tagged TF is bound to a glass slide that has ~41,000 spots of short immobilized DNA sequences. Fluorescence-based detection of bound spots and k-mer enrichment analysis yields motifs.	✔ (< 12 bp)	✘	✔ Methyl-PBM	✔
		Bacterial one-hybrid	TF binding sites are selected in bacterial cells from a randomized library that is cloned in front of selectable marker genes. Can be reversed to select proteins able to bind a constant DNA sequence using a library of variant protein sequences.	✔ (< 14 bp)	✘	✘	✘
		SELEX-based methods	Systematic evolution of ligands through exponential enrichment (SELEX) involves adding TFs to a DNA pool containing many randomized sequences and selecting for binding in multiple rounds. Related methods include HT-SELEX, SELEX-seq, and Bind-n-Seq. Selection can be performed using affinity tags, or molecular trapping on a microfluidic platform (SMiLE-seq).	✔ (< 25 bp)	✘	✔ Methyl-HT-SELEX	✔ CAP-SELEX ✔ SMiLE-seq
Mid-throughput		DAP-seq	Single step SELEX using a library of fragmented genomic sequences. Sequence diversity is less than HT-SELEX, but genomic sequences that have co-evolved with the TF are included.	✔ Limited by skewed distribution of genomic sequences	✔ Peaks are not necessarily indicative of *in vivo* binding	✔ AmpDAP-seq	✘
		HiTS-FLIP	Uses an Illumina sequencer's flowcell as a PBM chip to measure binding to orders of magnitude more DNA sequences.	✔ (< 17 bp)	✘	✘	✘
		Spec-seq	Single step SELEX using a synthesized library of degenerate sequences of interest. The lower complexity library is useful for quantitatively measuring effects of binding site mutations using sequencing.	✔ Limited by number of sequences assayed	✘	✔ Methyl-Spec-seq	✘
		MITOMI	A microfluidic device is used to isolate DNA-protein complexes from free DNA instantaneously to accurately measure the relative binding affinities of TFs to ~10,000 individual sites.		✘	✘	✘
Low-throughput		EMSA	Tests if a DNA sequence is bound by a protein by observing a shift in the electrophoretic migration of DNA.	✘ Useful for validating known binding sites	✘	✔	✔ EMSA-FRET
		DNA footprinting	DNA is incubated with a TF and then degraded using DNase-I, resulting in cuts in all positions except those that were protected by the bound TF.		✘	✔	✔
		ITC, SPR, MSTP	Isothermal titration calorimetry (ITC), Surface plasmon resonance (SPR) and Microscale thermophoresis (MSTP) measure the binding affinity of TF-DNA interactions.		✘	✔	✔
	In Vivo Methods	ChIP-based assays	Proteins are crosslinked to DNA using formaldehyde and precipitated with an antibody. Bound DNA is detected with qPCR, microarray (ChIP-chip), or sequencing (ChIP-seq). The ChIP-Exo variant incorporates exonuclease treatment to enhance resolution.	✔ Limited by skewed distribution of genomic sequences, and inability to distinguish direct from indirect binding	✔	✔ ChIP + Bisulfite-sequencing	✔ Re-ChIP
		DamID-seq	A TF is expressed in mammalian cells as a fusion to bacterial Dam-methylase. The enzyme methylates a consensus sequence in close proximity to the TF's binding sites, which can be mapped using restriction enzymes and high-throughput sequencing.		✔	✘	✔ Split DamID-seq

Figure 1.1: Table 1 of Lambert *et al.* [2] summarizing experimental methods for determining transcription factor binding specificities.

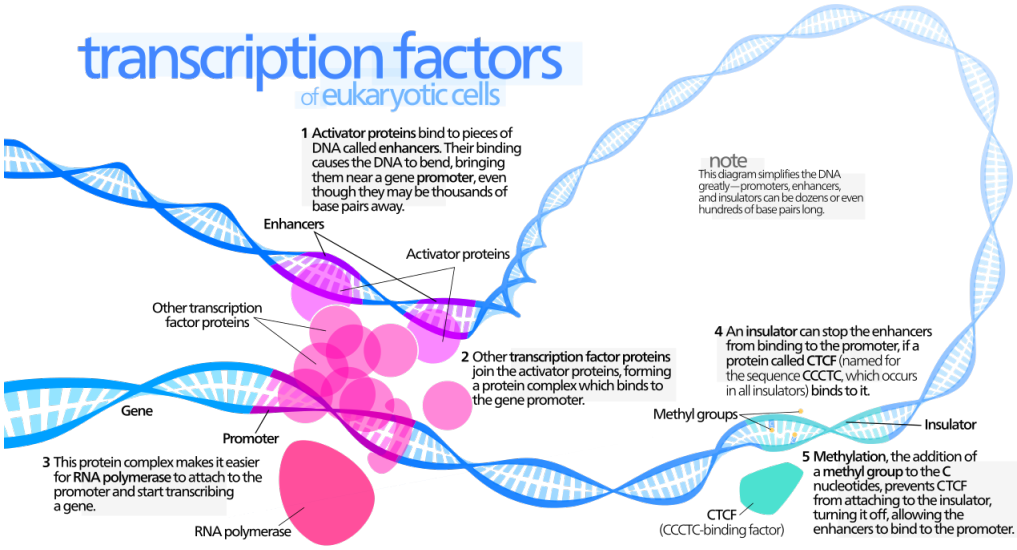
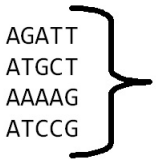


Figure 1.2: A diagram of a transcription factor. (From Kelvin13 - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=23272278).

6], other diseases [7] and virus virulence [8]. An accurate prediction algorithm is needed because transcription factor binding preferences are only known for $\sim 1\%$ of eukaryotic transcription factors [1]. Since prediction algorithms are faster and less expensive than experimental methods, a computational prediction is a more viable solution to increase the number of known transcription factor binding preferences for eukaryotic transcription factors.

Transcription factor binding preferences are represented in many models [9]. The most common model is the Position Weight Matrix (PWM), which is a model that represents the likelihood of a nucleotide being at a particular position in a sequence. A simple example is seen in Figure 1.3. The next representation of binding preferences is the 8-mer [10] profile, which is a vector representing how likely a transcription factor binds to all possible 8-mer sequences (Figure 1.4). Usually, this profile uses PBM

data to calculate it. Another representation is dinucleotides [11], which are a vector of the frequencies of two adjacent nucleotides in a sequence. Figure 1.5 shows an example of this. Some models also consider long-range dependencies. Such a model looks at how a nucleotide x positions away affects the binding of the nucleotide that is being considered [12].



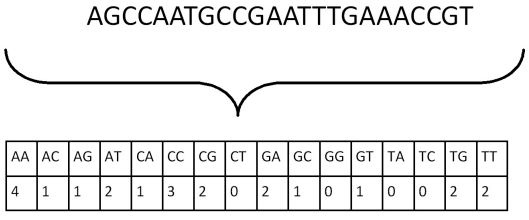
	1	2	3	4	5
A	1	.25	.5	.25	0
C	0	0	.25	.5	0
G	0	.25	.25	0	.5
T	0	.5	0	.25	.5

Figure 1.3: A sample position weight matrix. The four sequences on the left would make the matrix on the right.

Weirauch *et al.* [13] did a systematic comparison of models for transcription factor binding preferences. They compared 26 models on 66 mouse transcription factors from several families. The model types they looked at were PWM, k-mers, and dinucleotides. Some of these models were a combination of other models, and some had a machine learning algorithm built-in. After looking over the results Weirauch *et al.* concluded that models based on k-mers outperform other model types. Most models that are developed to predict a transcription factor binding preferences are built to predict one family of transcription factors [14, 15, 16, 17, 18, 19]. However, this type of model cannot be built for families with only a few members. Many models that exist require additional information such as multiple sequence alignment for the sequences or the structure of the sequences itself [20, 21, 22]. We give more detail in

8mer	8mer	Z-Score
AAAAAAAA	TTTTTTTT	2.3401
AAAAAAAC	GTTTTTTT	1.3416
AAAAAAAG	CTTTTTTT	1.3386
AAAAAAAT	ATTTTTTT	2.1334
.		
.		
.		
TTTCAAAA	TTTTGAAA	2.0584
TTTCCAAA	TTTGGAAA	0.5581
TTTCGAAA	TTTCGAAA	0.8202
TTTGAAAA	TTTTCAAA	1.6207
TTTGCAAA	TTTGCAAA	1.0773
TTTTAAAA	TTTTAAAA	1.9789

Figure 1.4: A sample of an 8-mer profile using Zscore for binding preferences.



AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
4	1	1	2	1	3	2	0	2	1	0	1	0	0	2	2

Figure 1.5: An example of dinucleotide representation. The sequence above makes the dinucleotide vector below.

Chapter 2.

1.2 Contributions Made in Thesis

The goal of this research was to build a family-independent model for predicting transcription factor binding preferences, which takes as input protein sequences of DNA binding domains (DBDs) and experimentally derived 8-mer profiles. Additionally, multiple sequence alignment (MSA) may introduce errors [23], if the sequences given do not correspond to the same region. Thus, we wanted a model that did not require MSA of the DBD sequences. Our method (called TopAffy for topological affinity) predicts the k-mer profile of any given transcription factor with an accuracy comparable to that of family-specific state-of-the-art methods.

1.2.1 Contribution Synopsis

- This work built a family-independent model for predicting transcription factor binding preferences (Section 3.2). Previous models for predicting transcription factor binding preferences are family-dependent (Chapter 2).
- Our model is able to match the prediction power of other models (Section 4.4.1).
- The performance of our model trained with multiple-family data is comparable to that of a model trained with family-specific data (Section 4.4.3).
- Our model can predict transcription factors for DBD sequences from families with few experimental data (Section 4.4.4).

1.3 Organization of Thesis

Past work on predicting transcription factor binding preferences is presented in Chapter 2. Chapter 2 hopefully illustrates the importance of a family-independent model. In Chapter 3, we describe TopAffy. We first discuss the construction of the structures used in TopAffy, the steps to make predictions. We then look at the implementation of TopAffy. Chapter 4 presents the results of evaluating TopAffy. We first look at the datasets that were tested. We then justify some of the design decisions made in the creation of TopAffy. We also examine the running time and space usage of TopAffy. We demonstrate that TopAffy can make predictions for transcription factor families for which no experimental data is available. Chapter 5 examines some biological insights that can be gathered based on the structures that make up TopAffy. Finally, we end with Chapter 6, discussing TopAffy's contributions and pointers for future work.

Chapter 2

Background

In this chapter we provide a brief description of previous computational methods for predicting transcription factor binding preferences. Our classification of these methods is like the one given in Table 1a of [24] reproduced in Figure 2.1.

2.1 Transcription Factor Binding Data

A resource paper by Weirauch *et al.* [1] tested transcription factor binding preferences based on PBM data on the most diverse data set yet. They looked at and built motifs for 1032 transcription factors binding preferences over 131 species and 54 DNA-binding domain (DBD) types. In this thesis, we use PBM data to construct our model.

(A) Computational models of protein–DNA binding specificity	
Model type	**Model description**
Position weight matrices (PWMs)	Simple probabilistic models that assume independence between positions in TF binding sites (TFBSs)
Dinucleotide weight matrices (DWMs)	Generalization of PWM models that incorporates frequencies of dinucleotides
Bayesian networks	Flexible probabilistic models that can incorporate dependencies between positions in TFBSs
Hidden Markov models	Probabilistic models that can incorporate dependencies between neighboring positions in TFBSs
High-order Markov models	Flexible probabilistic models that can incorporate high-order dependencies between neighboring positions in TFBSs
k-mer based regression models	Probabilistic models that predict the level of TF binding based on the frequencies of mono-, di-, and trinucleotides
Markov networks	Flexible probabilistic models that can incorporate high-order dependencies within TFBSs
Neural networks	Flexible probabilistic models that represent TF binding specificities using a system of interconnected, artificial 'neurons'
Random forest models	Flexible probabilistic models that represent TF binding specificities using a collection of decision trees
Support vector models	Probabilistic models that can incorporate complex patterns of similarities between TFBSs
Variable-order Bayesian networks	Flexible probabilistic models that can incorporate high-order dependencies within TFBSs
Thermodynamic/energy-based models	Models that infer DNA binding affinities by fitting thermodynamic equations to experimental data
Atomistic/structure-based models	Models based on known structures of TFs bound to DNA target sites
Probabilistic models that incorporate structural features	Models that incorporate DNA shape features such as groove geometries and helical parameters
Probabilistic models that incorporate *in vivo* data	Models that incorporate *in vivo* data such as DNA accessibility and histone modifications

Figure 2.1: Table 1a from Slattery *et al.* [24]

2.2 Early Models

One of the earliest approaches for computationally predicting transcription factor binding preferences was a model developed by Suzuki and Yagi [14], which looked at the structure of the zinc finger, probe helix, helix-turn-helix and C4 Zn-binding families and built a model based on the chemical and stereochemical merits of the structure. This model did not use machine learning but was developed based on knowledge about the structure itself. Suzuki and Yagi built a set of rules based on their research which they use to score DNA and protein sequences. Based on a measurement that they called the specificity index,

$$100 - n - \frac{m}{2} \tag{2.1}$$

where n is the percentage of the DNA sequence that score higher than the real binding sequence and m is the percentage of DNA sequences that score the same as the real binding sequence. Their algorithm was able to obtain an average specificity index above 90% for all the tested families.

Mandel-Gutfreund and Margalit [25] built a model that calculates their predictions by using a log odd likelihood of amino acid interaction with a nucleotide. They calculate this likelihood of interaction by looking at 218 different pairs of experimentally validated zinc finger protein-DNA interactions and built a 20 by 4 table.

Kono and Sarai [15] also built a model based on the structure of a sequence. Their model uses statistical data on the complex structure of the DNA. Kono and Sarai looked at all zinc-finger 52 complex structures and counted the interactions between the protein and DNA sequences and classified them into four groups based on the

backbone and side-chains interacting with DNA sequences. Although this model does a good job finding the critical section of a zinc-finger, they also assessed performance on other transcription factor families such as homeodomain with little success.

Benos *et al.* [16] built a statistical model to predict the binding preferences of zinc fingers proteins. They built this model by looking at the interactions between proteins and nucleotides, dinucleotides and trinucleotides and assigning a value to these. This model predicts zinc finger transcription factor binding preferences with a correlation coefficient of over .95 on average. Zhou and Liu [26] made a model that uses a position-specific weight matrix and dinucleotide information to build a Markov chain to predict transcription factor binding sites. When tested against the standard PWM model, they found that it was less prone to making false positives (that is for 17 out of 22 transcription factors, it made less false positives than the standard PWM model). Kaplan *et al.* [17] like Benos *et al.* [16] built a statistical model for the sequence preferences of the zinc-finger family. This model looks at the probability of a nucleotide being at a certain position and interacting with a certain amino acid. This model was built for zinc-fingers because they have a particular structure. When compared to other methods at the time, it got a higher true positive/ lower false positive rate ($\sim$ 80%/1%). Like many of the other models above this model is built for the zinc finger family alone and may not work for other families.

2.3 Sequence Based Models

The models described in this section only use the amino acid sequence of transcription factors and binding data to generate their predictions. These models are classified

based on the machine learning method used to construct these models.

2.3.1 Nearest neighbour model

In a paper by Alleyne *et al.* [27] four machine learning algorithms (k-nearest neighbour, random forests, support vector machine and principal components regression) were comparatively evaluated on a mouse homeodomain dataset [28]. Out of the four approaches, the one that performed best in terms of Spearman's correlation, Root Mean Squared Error (RMSE) and number of predicted binding sequences in common, was the nearest neighbour algorithm.

2.3.2 Neural Network model

The next method is a neural network by Liu and Stormo [29] called Zifnet. Zifnet uses information about amino acid residues binding to the canonical recognition positions. Zifnet predicts on multi-finger information (2, 3 or 4 fingers) but predicts poorly when the number of fingers was greater than two. This model was compared to the method by Kaplan *et al.* [17] and Benos *et al.* [16] on 9 zinc finger datasets and found to outperform them on several of those nine datasets.

Shen *et al.* [30] developed a model that uses a Gated Recurrent Unit network with k-mer embedding. This model is a modified recurrent neural network that is good at learning features from a large dataset, and k-mer embedding learns long-range dependencies in the sequence. This model works on ChIP-seq data and tested on 125 transcription factors from 4 datasets (HESC, A549, HUVEC and MCF7). This model obtained AUC values of 0.9524, 0.9593, 0.9612 and 0.9649 respectively.

Rastogi *et al.* [31] made a statistical model for predicting transcription factor binding preferences based on SELEX [32] data. The model predicts sites with very low affinity. This model combines two interacting transcription factors (Hox and bZip) to build a model.

2.3.3 Support Vector Machines

Persikov *et al.* [18] used support vector machines on the family of Cys_2-His_2 zinc fingers. Persikov and Singh [33] expanded this algorithm by using information about amino acid and base pair combinations. As an input, the algorithm takes in a vector of 1280 elements, which represents all the information about amino acids, nucleotides, the canonical recognition positions and combinations of unique triplets of amino acids. Persikov and Singh [34] developed another algorithm that uses a support vector machine. Through experimental observation, they found that there may be three more contact positions that are important for the predicting of binding specificities of DNA. They built a model that uses this information. The model predicts the PWM of $\sim 80\%$ of the sequences in their dataset.

2.3.4 Random Forest

One algorithm that uses a random forest is ZFModels by Gupta *et al.* [35] which is an extension of a previous algorithm by Christensen *et al.* [19]. ZFModels works with a family of transcription factors Cys_2-His_2 zinc finger, which is one of the largest transcription factor families. This model can get a mean MSE of .017 and a median Mean Squared Error (MSE) of .009 between the prediction and observed position

frequency matrices.

2.3.5 Linear Model

Annala *et al.* [36] was the best performing method in the DREAM5 challenge [1]. This method uses a matrix based on PBM array and k-mers. Once the matrix is built, the model uses a conjugate gradient method to predict binding preferences. Pelossof *et al.* [37] developed a model called affinity regression that is a bilinear regression model for predicting transcription factor binding preferences from PBM datasets. The model only needs the sequences and PBM data and does not require a multiple sequence alignment or the motifs of the sequences. This model outperforms the k-nearest neighbour model for a dataset of 178 mouse homeodomains.

Lambert *et al.* [38] made a model named similarity regression, which uses transcription factor protein similarity to predict transcription factor sequence specificities. It does this by first aligning each transcription factor DBD sequence to Pfam HMM to get a global alignment. They use regression to train a matrix where each row is a pair of TF, and columns are positions in the sequence. Once the model trains, there is a value assigned to each position in the sequence based on its importance to produce a weight vector. They used this model to predict 8-mer profiles by using this weight vector to find the closest transcription factor or transcription factors in the dataset.

2.3.6 Hidden Markov Model

Dai *et al.* [12] created an algorithm that built a hidden Markov model based on the DNA sequences. The HMM allows the model to learn position information as well

as long-range dependency information. It does this with a message passing-like embedding algorithm. This model got tested on over 90 transcription factor datasets, one of which was the DREAM5 dataset. On the DREAM5 dataset, this model outperformed other algorithms when measuring, Pearson (0.741) and Spearman (0.765) correlation and Area under the ROC curve (AUC) (0.959).

2.4 Multiple Data Source Models

All the above methods for predicting transcription factor binding preferences made predictions based on the sequences and binding profiles alone. Methods by Andrabi *et al.* [20] and Li *et al.* [21] not only made predictions based on sequence but also uses the structure of the protein. If the structure is not known, Andrabi *et al.* [20] built an algorithm that predicts the structure from the sequence by using 65 support vector regression models, and Li *et al.* [21] uses DNAshape [22] which uses Monte Carlo simulation to predict structure. This method works better on datasets that contain more flanking base pairs. The method got tested on a gcPBM [22] dataset and a uPBM [13] dataset; the former got mean R^2 values around .8, and the latter got values around .4.

Chapter 3

Methods

In this chapter we present the methods used in designing TopAffy (short for Topological Affinity Prediction). This includes describing the structures that are used to make predictions, how the predictions are made, and discussing the implementation of TopAffy itself.

3.1 TopAffy Overview

TopAffy predicts transcription factor binding preferences (8-mer profiles). TopAffy uses a graph to represent transcription factor protein sequences, where vertices are amino acids in position i in a sequence, and edges connect neighbouring amino acids in the sequence. It also uses an emission matrix to represent the binding preferences of neighbouring amino acids.

TopAffy takes in as input a set of transcription factor sequences (just the DNA binding domain) and their corresponding 8-mer binding profiles, obtained from PBM data [4]. TopAffy does not require that the DNA binding domain (DBD) sequences

to be aligned. Once the model is built and trained, it takes in a sequence and outputs a predicted 8-mer profile for that sequence.

To predict the 8-mer profile of a sequence, TopAffy multiplies the weights on the graph corresponding to the DNA-binding of the sequence by the values on the emission matrix corresponding to the neighbouring amino acids. In this chapter, we explain how this is done in detail.

3.2 Model Construction

We built TopAffy to use two steps. The first step is to build a graph using all available DBD sequences, not only the training sequences but also the sequences TopAffy are predicting. The second step is to have the model learn values to fill the emission matrix.

The graph is a stochastic topological representation of the sequences. This structure is like a Hidden Markov Model (HMM) but instead on each vertex having an emission, this model has one emission matrix for all the vertices. HMM are a common method for representing sequences; it is the method used in Dai *et al.* [12], Team E in [13] and Keilwagen*et al.* [39].

This graph has a starting vertex and an ending vertex. To build the graph in-between the start and end, the model looks at all the amino acids in the first position of the DBD sequences. It finds which amino acids appear in the first position and creates a vertex for each of them. It then creates an edge between the start and all the new vertices. The weight of each edge is the frequency of each amino acid at position i (i.e. the number of occurrences of amino acid j at position i divided by the

total number of sequences). Once the model constructs the first level of the graph, it then builds the next levels. For all amino acids in the ith position of the sequences, a vertex is made. Then an edge is created between a vertex at position $i-1$ and a vertex at position i if an adjacent amino acid pair exist at positions $i-1$ and i in any of the sequences. The weight of each edge is the number of occurrences of each amino acid j at every position i divided by the maximum number of occurrences. A modified Gini impurity measure with a value in the range $[0, 1]$ is then used to get edge weight. TopAffy uses the Gini index so that edges connecting to rare AA pairs or almost invariant AA pairs receive the highest weight. We call each edge weight τ. This process continues until an edge is created between the amino acid at the end of each sequence to the end vertex. Every edge weight in the graph represents the likelihood of transitioning from vertex S_{i-1} to vertex S_i in the graph. See Figure 3.1 for an example of such a graph.

The next step is to calculate the emission matrix. This matrix quantifies the effect an amino acid pair has on the binding of the transcription factor to each k-mer, regardless of the pair position in the sequence. Each cell in the table is a vector of all possible 6-mers. Figure 3.2 shows an example of this. To reduce the number of parameters to fit, we use 6-mers instead of 8-mers in the emission matrix. TopAffy using a mapping table to look up for the frequency of each 6-mers contained in each 8-mer. Figure 3.3 shows a section of this table.

The emission matrix is populated with zeros, to begin with, and uses stochastic gradient descent [40] to fit. The sequences get shuffled into a random order then the model goes through all sequences in this shuffled order. For each sequence, the model picks a random selection of 8-mers. Then the model makes a prediction for each

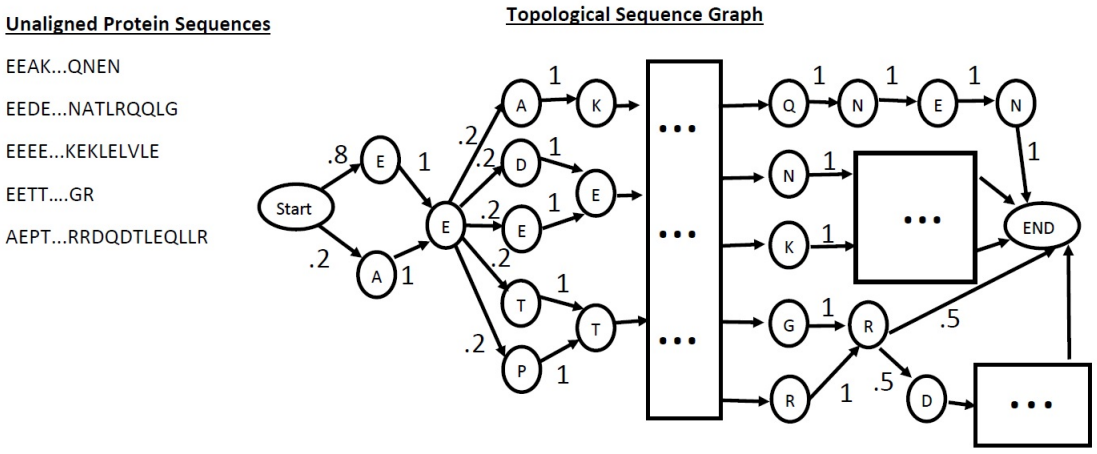


Figure 3.1: A representation of the topological stochastic graph. On the right, there is a list of unaligned protein sequences. On the left, there is a graphical representation of these sequences in a topological sequence graph. All vertices are the amino acid at that position, and edge weights are the probability from going for one amino acid at position i to amino acid $i+1$. TopAffy calculates probabilities by the number of amino acids x at position i followed by amino acid y at position $i+1$ divided by the maximum number of occurrences and uses a modified Gini impurity measure with a value in the range $[0, 1]$ to get the final edge weight. Note because the sequences are unaligned, at any point a vertex can go to the end vertex. This graph is calculated based on all available DBD sequences.

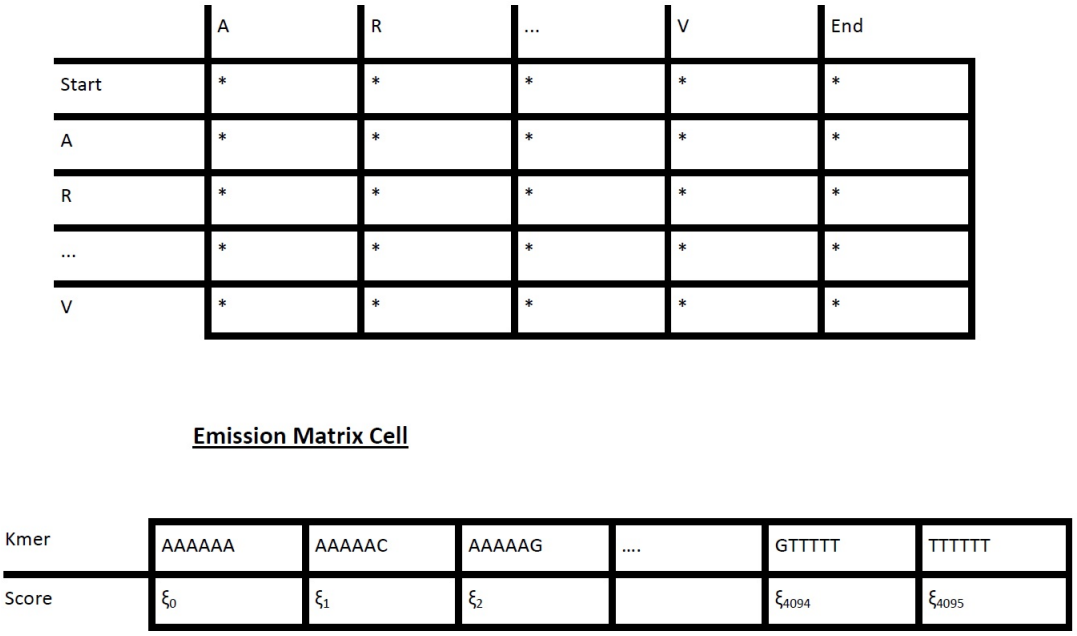


Figure 3.2: A representation of the emission matrix. The top matrix represents amino acid (plus start and end) pairs. Each cell in the top matrix contains a vector such as the one shown below the matrix. Each vector is calculated based on training sequences passed to the graph during the training phase. The ξ so the zscore portion for that 6-mer

8-MER to 6-MER Mapping Table

	AAAAAA	AAAAAC	AAAAAG	...	TTAAAA	...	TTTAAAA	...	TTTTAA	...	TTTTTG	TTTTTT
AAAAAAAA	3	0	0	...	0	...	0	...	0		0	0
AAAAAAAC	2	1	0	...	0	...	0	...	0		0	0
AAAAAAAG	2	0	1	...	0	...	0	...	0		0	0
AAAAAAAT	2	0	0	...	0	...	0	...	0		0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
TTTTAAAA	0	0	0	0	1	0	1	0	1	0	0	0

Figure 3.3: A break down of all the 6-mers that make-up 8-mers.

8-mer selected. Then the residuals are (the difference between the actual outputs y and the predicted outputs $\hat{y}$) calculated, and small changes are made to the cells used in the prediction to minimize the residuals. Each cell has a small number added to it. Equation 3.1 shows this calculation,

$$(y - \hat{y}) * \eta \tag{3.1}$$

where η is the rate of learning and, η shrinks over time by the decreasing learning factor λ. Figure 3.6 shows an example of modifying the emission matrix. When the algorithm starts, it makes more significant changes after each sequence is processed, and it starts to fine-tune as it runs with a decreasing learning factor. These steps repeat until the model converges, which means that several iterations of the algorithm have not yielded an overall positive change (more substantial than ϵ) to the predictive performance of the model. To check if an overall positive change has been made, the model predicts 8-mer binding profiles for all sequences and checks the Spearman correlations between all Y to $\hat{Y}$ where Y is the actual 8-mer binding profile and

$\hat{Y}$ is the predicted one. Once convergence has happened TopAffy outputs the final emission matrix. Table 3.1 gives the default values for the parameters mentioned above.

η	λ	ϵ
0.0001	.99	.05

Table 3.1: Parameter values for training that are used in this thesis.

3.3 Making a prediction

To predict a sequence with an unknown 8-mer binding profile, the model runs the sequence through the graph for each 8-mer. The model starts at the start vertex in the graph and moves to the end. For TopAffy to predict, the sequence has to traverse the graph and get the prediction for each 8-mer using Equation 3.2.

$$Zscore_z = \sum_{i=1}^{n+1} \sum_{j=1}^{m} (\tau S_{i-1}, S_i * e[S_{i-1}, S_i, j] * map[z, j]) \tag{3.2}$$

where z is an 8-mer to be predicted, sequence S is being considered, τ is the transition value from S_{i-1} to S_i which is the edge weight in the graph, e is the emission matrix. Term i represents a position in the sequence, j is one of the 6-mers, n is the length of the sequences and m is the number of 6-mers in the 8-mer. Figure 3.3 is a representation of the *map* (a way to break 8-mers into 6-mers) in the equation.

Let's look at an example. Suppose we have the sequence $AEPT$ and we are predicting the Zscore of $AAGTTGAA$ using the graph in Figure 3.4. The calculation of the predicted Zscore is illustrated in Figure 3.5.

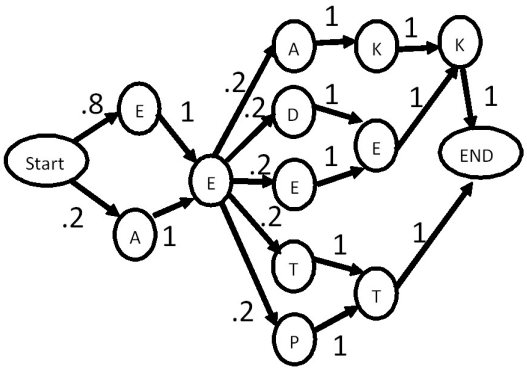


Figure 3.4: A simple graph for example prediction

Simplified Emission Matrix

	AAGTTG	AGTTGA	GTTGAA
Start,A	.009	.064	.066
A,E	.073	.071	.02
E,P	.084	.026	.031
P,T	.048	.017	.003
T,END	.083	.076	.008

X

Corresponding edge weights

t
.2
1
.2
1
1

=

Values From Inner Summation

AAGTTG	AGTTGA	GTTGAA	Total
0.0018	0.0128	0.0132	0.0278
0.073	0.071	0.02	0.164
0.0168	0.0052	0.0062	0.0282
0.048	0.017	0.003	0.068
0.083	0.076	0.008	0.167

Predicted Value

.455

Figure 3.5: Example of how TopAffy makes a prediction. This production uses the graph seen in Figure 3.4. The left matrix is the simplified emission matrix, only showing the relevant cells. The middle vector shows transition values of the sequence of an 8-mer profile (values are the paths edge weights). The last table is the value obtained by running the inner summation from Equation 3.2. The last column in that table is the sum of the rows. The value in the box to the right is the prediction for the 8-mer (*AAGTTGAA*) for the sequence (AEPT) obtained by adding up the Total column of the inner summation.

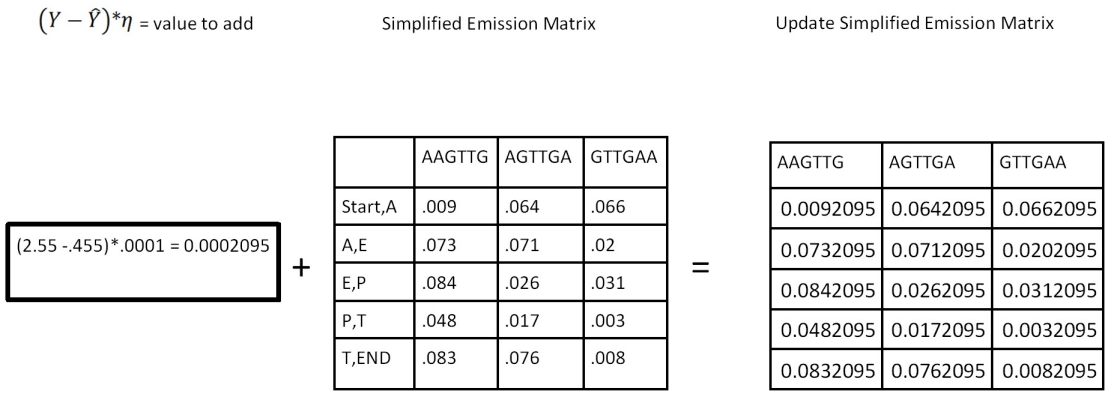


Figure 3.6: Example of how TopAffy makes changes to the emission matrix. The table gets updated by the value in the box to the left. This example uses the result from Figure 3.5; therefore it uses the 8-mer ($AAGTTGAA$) for the sequence (AEPT)

3.4 Implementation

See Algorithm 1 for the implementation of TopAffy. This code was written in Python3, and the code can be found at https://github.com/BioinformaticsLabAtMUN/TopAffy.

3.5 Summary

In this chapter, we described TopAffy. This method uses two structures to capture the information for predicting binding preferences as 8-mer profiles of given transcription factors. It uses a topological sequence graph to capture the sequences and an emission matrix to represent the effect of adjacent amino acids in the transcription factor DNA-binding domain on the binding of 6-mers. These two structures use Equation 3.2 to predict the binding preference of a transcription factor for each 8-mer.

Algorithm 1 TopAffy

inputs:

η = rate of learning

λ = decreasing learning factor

τ = sequence graph

e = emission matrix

map = kmer map

Y = 8-mer binding profile

X = Transcript factor sequences

output:

Updated emission matrix

function TopAffy($\eta, \lambda, \tau, e, map, Y, X$)
    for `not converged` **do**
        $\hat{X}$= shuffle X
        for `x in` $\hat{X}$ **do**
            $\hat{K}$= get random set of 8-mers
            for `k in` $\hat{K}$ **do**
                pred = score(x,k,τ,e)using Equation 3.2 see Figure 3.5
                error = Y_{xk} - pred
                update e using Equation 3.1 see Figure 3.6
        $\bar{X}$ = random set from X
        test for convergence using $\bar{X}$;
    return e

TopAffy has two advantages over existing methods to predict transcription factor binding preferences: 1) no multiple sequence alignment is required, and 2) it is family independent in the sense that a single model can predict for multiple transcription factor families. As we show in the next chapter, its prediction performance is comparable to that of state-of-the-art methods.

Chapter 4

Empirical Evaluation

In this chapter, we describe the empirical evaluation undertaken to 1) validate some of the design decisions made in TopAffy such as using 6-mers instead of 8-mers and constructing a general instead of a family specific model; 2) characterize the properties of TopAffy such as run time, memory usage and convergence; and 3) comparatively assess TopAffy's prediction performance.

4.1 Evaluation Measures

The first question that we answer in this chapter is whether using 6-mers is optimal. To answer this question in Section 4.3, we show the results of running TopAffy on different sizes of k-mers. In Section 4.4.1, we looked at the performance of TopAffy against a state-of-the-art model, Affinity Regression [37]. In Section 4.4.2, we evaluate whether TopAffy can predict binding preferences for transcription factors from families without data. We do this by seeing how well TopAffy would predict if we trained our model to make predictions on transcription factors from families we did not train

it on. We call this "Blind" predictions. In Section 4.4.3, we compare the prediction performance of a family specific model and a general model. Finally, we evaluate the predictive performance of general TopAffy on predicting binding preferences of a small family of transcription factors.

4.2 Datasets

Three families of transcription factors were used, namely Homeodomain [28], bZIP [41] and Myb/SANT [42]. We gathered the 8-mer binding profiles and DBD sequences of these families from CIS-BP version 1.01 [1]. We chose these families because they are families with the largest number of experimental determined k-mer binding profiles, with 218, 102, and 96, respectively. After filtering out transcription factors that had multiple protein domain sequences, we reduced the samples to those described in Table 4.1. Some of the transcription factors had replicate 8-mer profiles. For those, we calculated the average for each 8-mer and used the average value in their binding profile. Then the data was exponentially scaled to increase the importance of the top probes. See Equation 4.1 and 4.2 for the equation.

$$Y = \forall\ y \in Y\ (100 * y - 1) \tag{4.1}$$

$$Y = \forall\ y \in Y\ ((y * 1)/\sqrt{\sum Y^2}) \tag{4.2}$$

,where Y is a set of all the Zscores for a sample.

These are the same steps that Pelossof *et al.* [37] did while testing Affinity Regression. We also used the dataset used by Pelossof *et al.* [37], which had 218 Homeodomain transcription factors from diverse species found on CIS-BP [1].

Family	Number of Transcription Factors	Species
Homeodomain	172	diverse species
Bzip	68	diverse species
Myb/Sant	79	diverse species
Pelossof *et al.* [37]	218	diverse species

Table 4.1: The families of transcription factors and the number of sequences for each family.

4.3 Method Tuning

Deciding on the size of k-mers was an interesting problem to solve. We first thought about using the 8-mers and training each 8-mer independently. We quickly rejected this idea because of the simple fact that there are too many of them. This leads to each 8-mer only being updated a few times while training the model. We thought we would break the 8-mers in smaller k-mers. All 8-mers are a combination of a group of small k-mers. For example, the 8-mer AAGTGCAA is constructed with four 5-mers AAGTG, AGTGC, GTGCA, and TGCAA. Using a sliding window on the 8-mer of size 5. Figure 3.3 shows an example of a mapping.

The challenging part was figuring out the size of the k-mer. We performed an experiment to determine what k-mer size would lead to the best prediction. For this test, we ran TopAffy on the Bzip family with three different k-mer sizes (4, 5, and 6). Using 10-fold cross-validation, we trained the model and then computed the Spearman correlation between the predicted 8-mer binding profiles against the known ones. Figure 4.1 shows the results of this experiment. This test was done to Bzip

only because we found through the development of TopAffy that Bzip dataset was the most challenging dataset, and any improvements to Bzip predictions lead to a better prediction for all families. Based on the result shown in this chart, we found that there is only a small difference in the predictive performance for different sized k-mers. The difference between 4-mers and 5-mers is statistically significant (when running a Wilcoxon Signed-Rank Test, the p-value is 0.04538) with 5-mers leading to better predictions. The difference between 5-mers and 6-mers is not statistically significant (when running a Wilcoxon Signed-Rank Test, the p-value is 0.5719). With these results, we decided to use 6-mers in TopAffy's implementation.

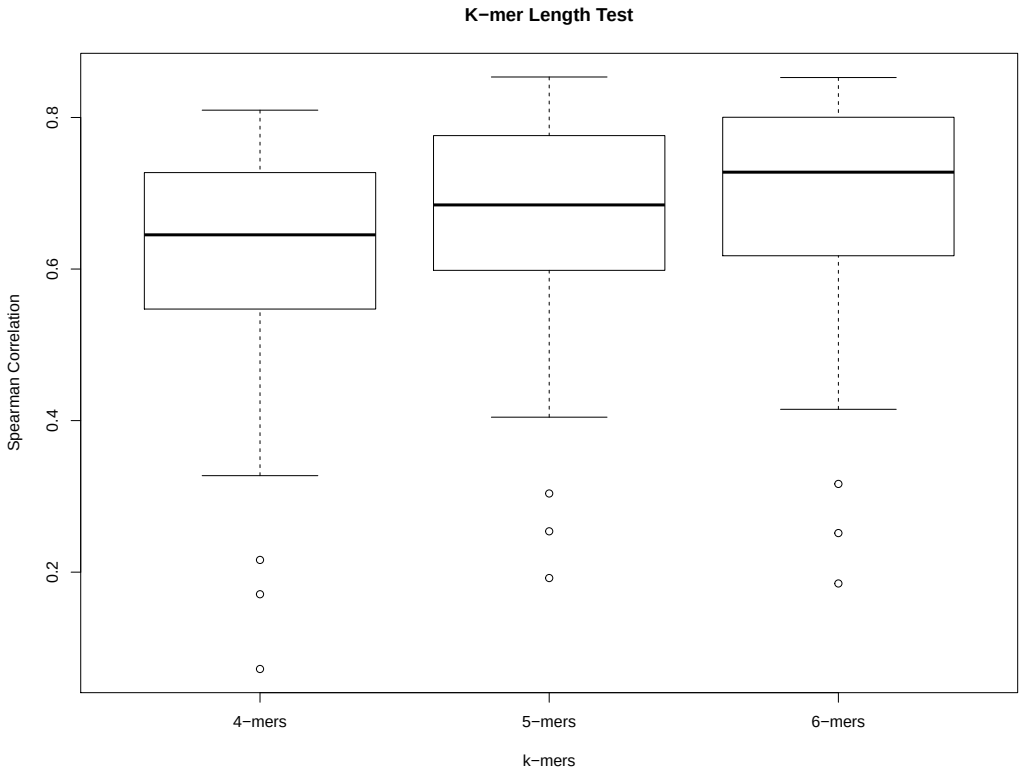


Figure 4.1: The effect of k-mer size on predictive performance using the Bzip family.

4.4 Performance Assessment

4.4.1 Comparative assessment

We compared TopAffy's performance to the performance of Affinity Regression [37]. We choose to compare to Affinity Regression because this model is predicting 8-mer binding profiles based on known profiles and DBD sequences only, which is similar to the way TopAffy works. To run TopAffy, we trained four models, one for each family and one for the dataset used for Affinity Regression [37]. To get the result for Affinity Regression, we ran Affinity Regression on the four datasets. Affinity regression version 1 directly ran from the software provided at https://bitbucket.org/leslielab/affreg/src/master/. We used the 8-mer binding profile output by affinity regression in our comparative assessment. The performance we obtained from that fourth dataset is very similar to the one reported by Pelossof *et al.* (see Fig 3.d in [37]). For the training of TopAffy, we used 10-fold cross-validation. We computed the Spearman correlation between the predicted 8-mer binding profile to the actual one. We did this for all three families of transcription factors we are testing plus the dataset used by Pelossof *et al.* [37]. These comparisons can be seen in Figures 4.2, 4.3, 4.4 and 4.5 and Table 4.2.

These results demonstrate that TopAffy has a performance comparable to that of Affinity Regression with both programs getting similar results for the Bzip and the Myb/Sant families, Affinity Regression outperforming TopAffy on the homeodomain dataset, and TopAffy outperforming Affinity Regression on the Pelossof *et al.* dataset.

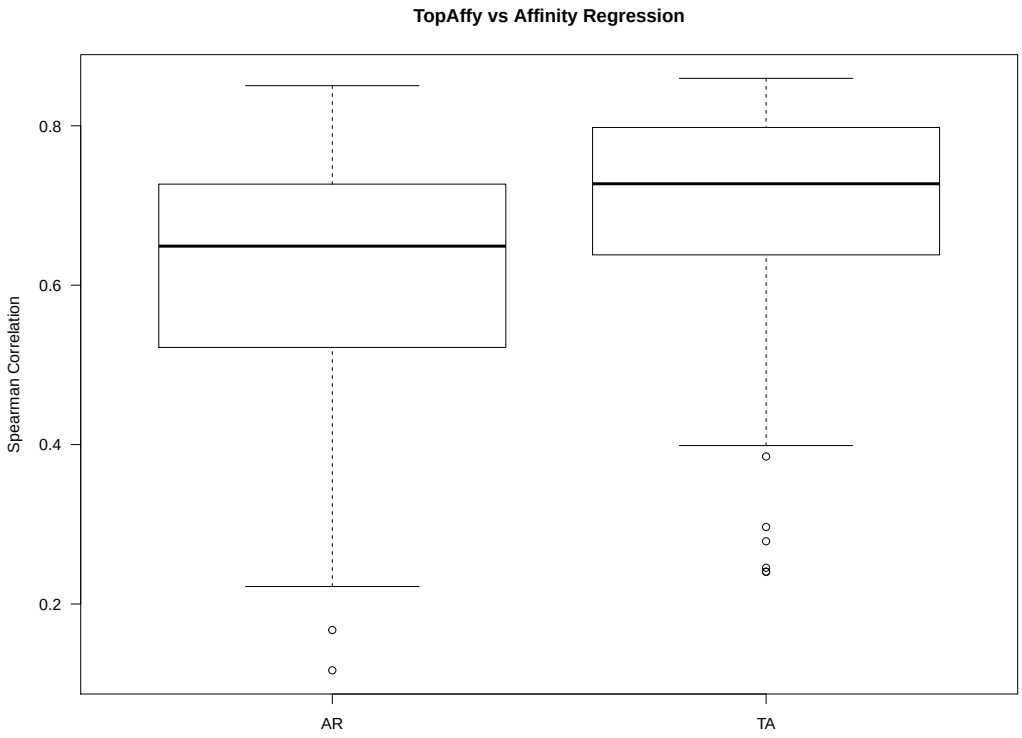


Figure 4.2: Spearman correlation for TopAffy (TA) and Affinity Regression (AR) for the Pelossof *et al.* data set.

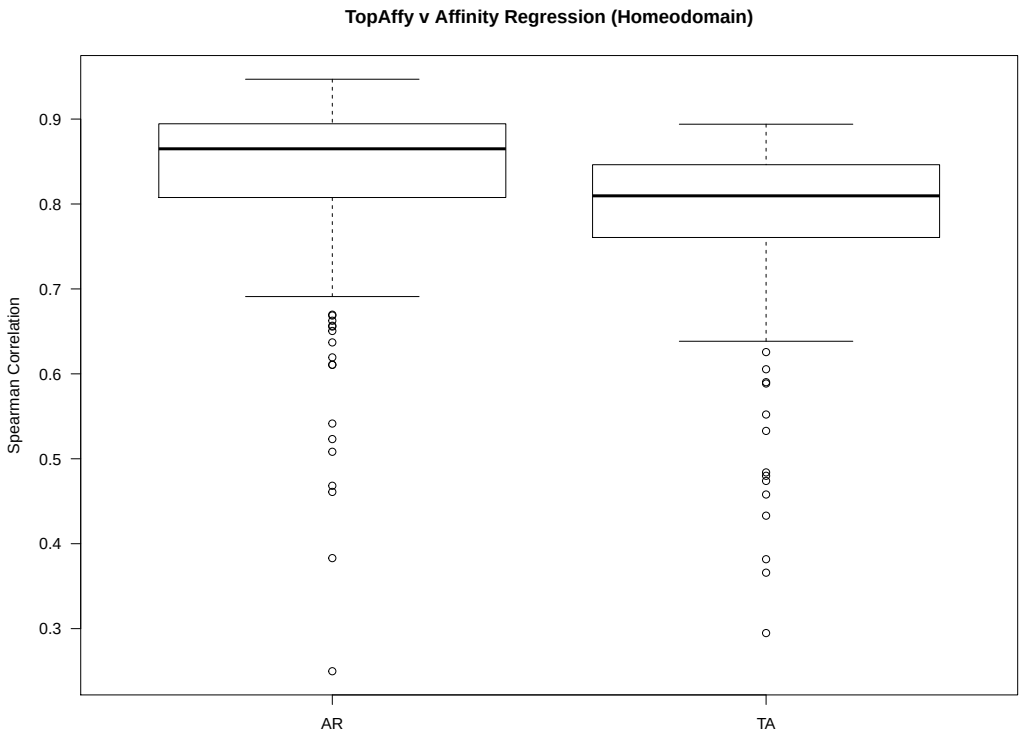


Figure 4.3: Spearman correlation for TopAffy (TA) and Affinity Regression (AR) for the Homeodomain family.

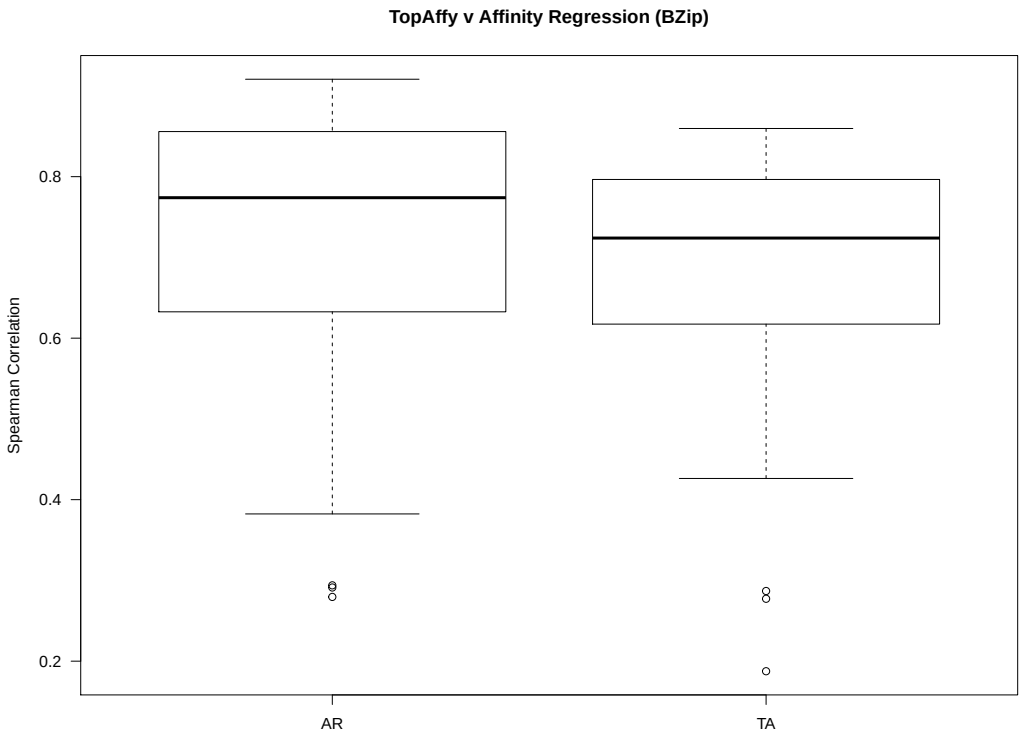


Figure 4.4: Spearman correlation for TopAffy (TA) and Affinity Regression (AR) for the Bzip family.

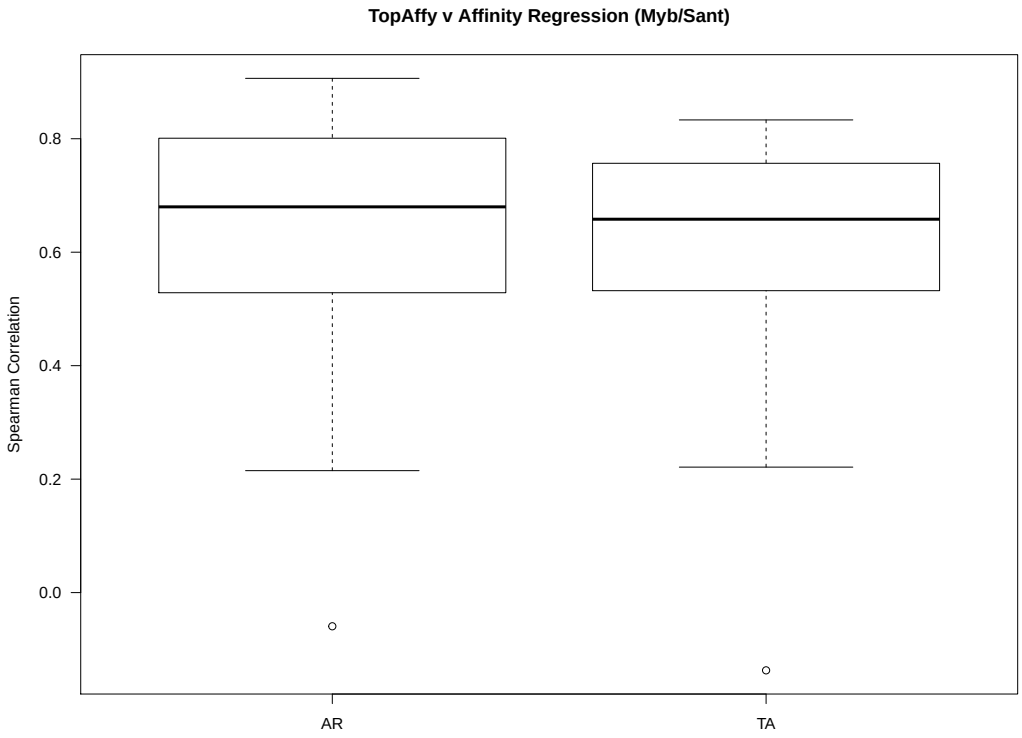


Figure 4.5: Spearman correlation for TopAffy (TA) and Affinity Regression (AR) for the Myb/Sant family.

Dataset	TopAffy (mean ± std)	Affinity Regression (mean ± std)	p-value
Homeodomain	0.7811 ± 0.1049	0.8284 ± 0.1083	1.196e-7
Bzip	0.6829 ± 0.1470	0.7285 ± 0.1624	0.0691
Myb/Sant	0.6186 ± 0.1794	0.6450 ± 0.1912	0.2769
Pelossof *et al.*Dataset	0.7010 ± 0.1272	0.6088 ± 0.1149	1.921e-14

Table 4.2: Mean ± standard deviation of the Spearman correlation and the p-value for a Wilcoxon Signed-Rank Test.

4.4.2 "Blind" prediction assessment

We assessed how well TopAffy, trained on the PBM data of two transcription factor families, could predict binding preferences for the transcription factors of a third family. We referred to this model as a "blind" model because we did not train on data from the family it is predicting. To do this, we constructed the topological graph using the DNA-binding domain sequences of all three families. Then we trained TopAffy using the 8-mer binding profiles of two of the families (e.g., Homeodomain and BZip) and predicted the 8-mer binding profile for the transcription factors of the left-out family (e.g., Myb/Sant). As a baseline, we used the average Spearman correlation value between transcription factors of the families used for training and those of the family predicting. Figures 4.6, 4.7, and 4.8 show that TopAffy was able to learn general binding relationships between DNA sequence and amino acids in the DNA binding domains. We used these relationships to obtain accurate predictions for a transcript factor family not seen during training without a major drop in predictive performance. The p-values for a Wilcoxon test between the family-specific and the

blind results are provided in Table 4.3.

Family	p-values
Homeodomain	$< 2.2e^{-16}$
Bzip	$3.73e^{-10}$
Myb/Sant	$8.04e^{-11}$

Table 4.3: The p-values for a Wilcoxon test between the family-specific and the blind results

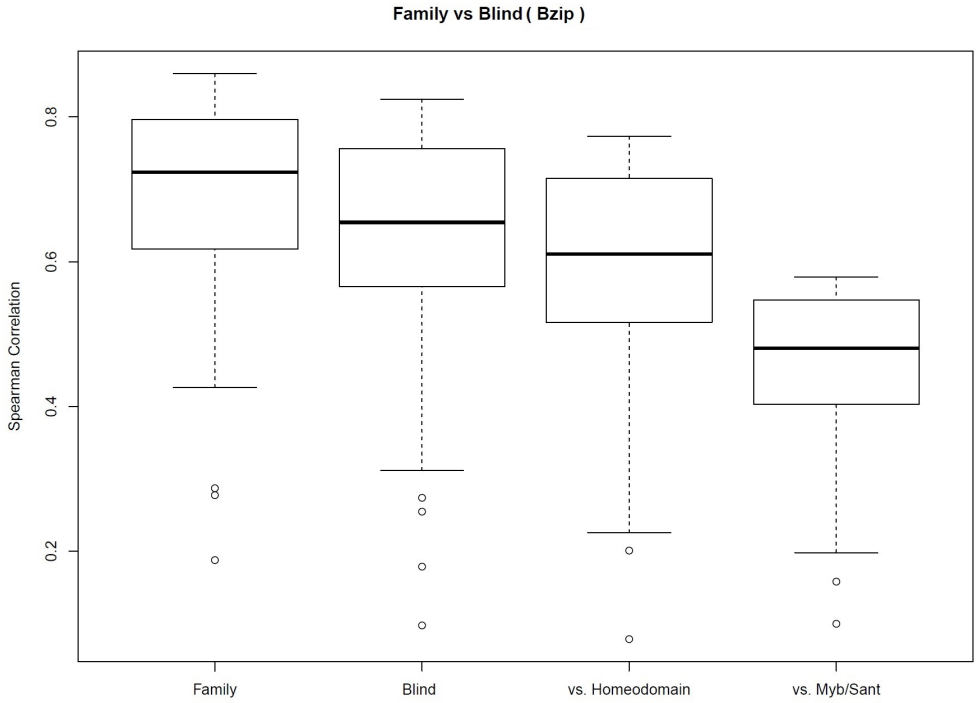


Figure 4.6: Spearman correlation for TopAffy family-specific and "blind" TopAffy, and the mean Spearman correlation between each Bzip transcription factor versus all Homeodomain transcription factors and the same for Bzip versus Myb/Sant.

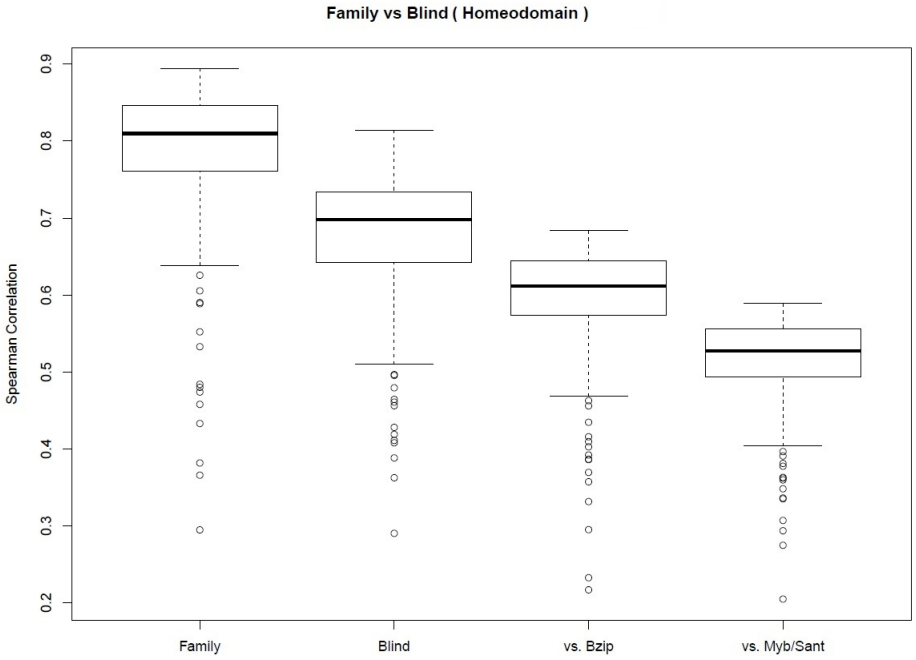


Figure 4.7: Spearman correlation for TopAffy family-specific and "blind" TopAffy and the mean Spearman correlation between each Homeodomain transcription factor versus all Bzip transcription factors and the same for Homeodomain versus Myb/Sant.

4.4.3 TA-family-specific vs TA-general

We conceived TopAffy to construct a family-independent model that predicts binding preferences for transcription factors of families without PBM data. To test whether this hypothesis was true, we evaluated whether there was a difference in prediction performance when training TopAffy on data from a single transcription factor family and when training on data from the three families. For the family-specific models, we trained three models, one for each family of transcription factor (as in Section 3.4.1).

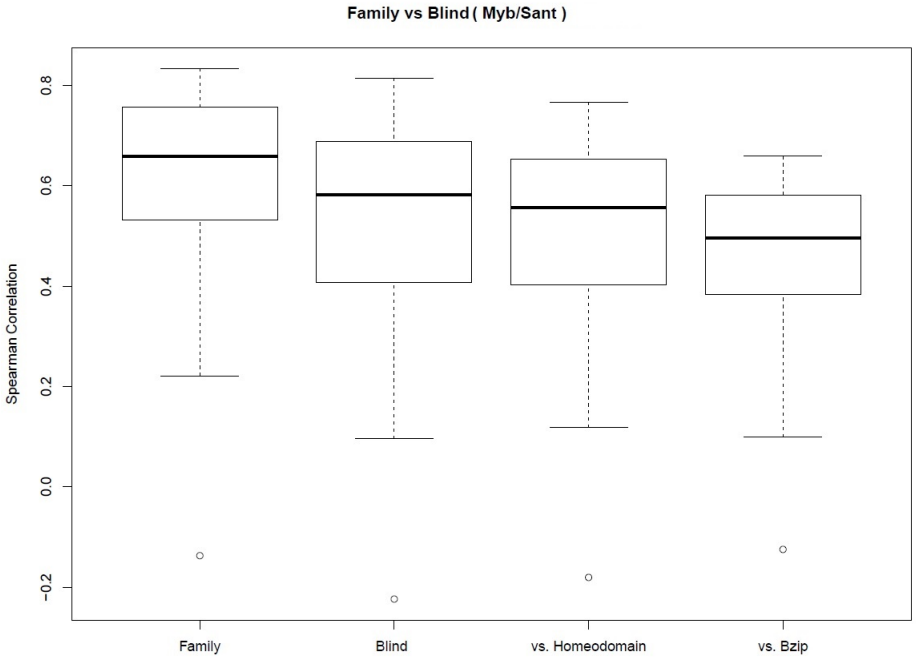


Figure 4.8: Spearman correlation for TopAffy family-specific and "blind" TopAffy, and the mean Spearman correlation between each Myb/Sant transcription factor versus all Homeodomain transcription factors and the same for Myb/Sant versus Bzip.

Then using 10-fold cross-validation, we computed the Spearman correlation between the predicted 8-mer binding profiles against the experimentally derived ones.

For the general model, to make the test set we combined all the transcription factor data we were using and randomly selected 50 of them. Then we randomly selected 100 different transcription factors as the training set. The intersection of these two sets was the empty set. We then trained a model using the 100 training transcription factors and used that model to predict the 50 test transcription factors. Once we predicted all 50 transcription factors, we computed the Spearman correlation between the predicted 8-mer binding profile to the actual ones.

Finally, we then matched the correlations of the transcription factor for the general model to the family model. These comparisons can been seen in Figures 4.9, 4.10 and 4.11. Figure 4.12 shows the overall performance for both a family trained model and a general trained model.

The results show there was no statistically significant difference between the two models (Table 4.4). Figure 4.13 shows that our general model was able to predict just as well as the family-specific one. Table 4.4 shows the detail of the performance. On average, the general model achieved a Spearman correlation, which is 0.024 below the corresponding family-specific model.

4.4.4 TA-general vs. AR to predict data on small families

We also assessed the predictive performance on a small transcription factor family not included in the training data. For this test, we compared TopAffy to AR. We ran this assessment using all mouse PBM data available on the CIS-BP version 2 [1]. This

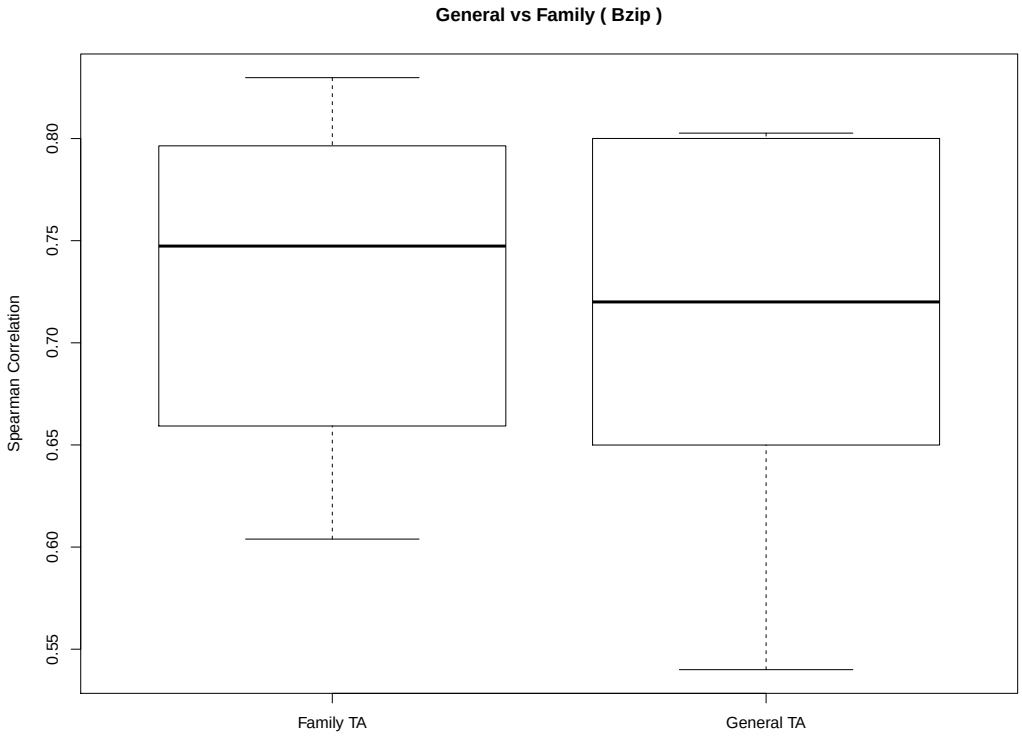


Figure 4.9: Spearman's correlation for the model trained on only Bzip transcription factors and a general model trained using 100 transcription factors taken randomly from the three families.

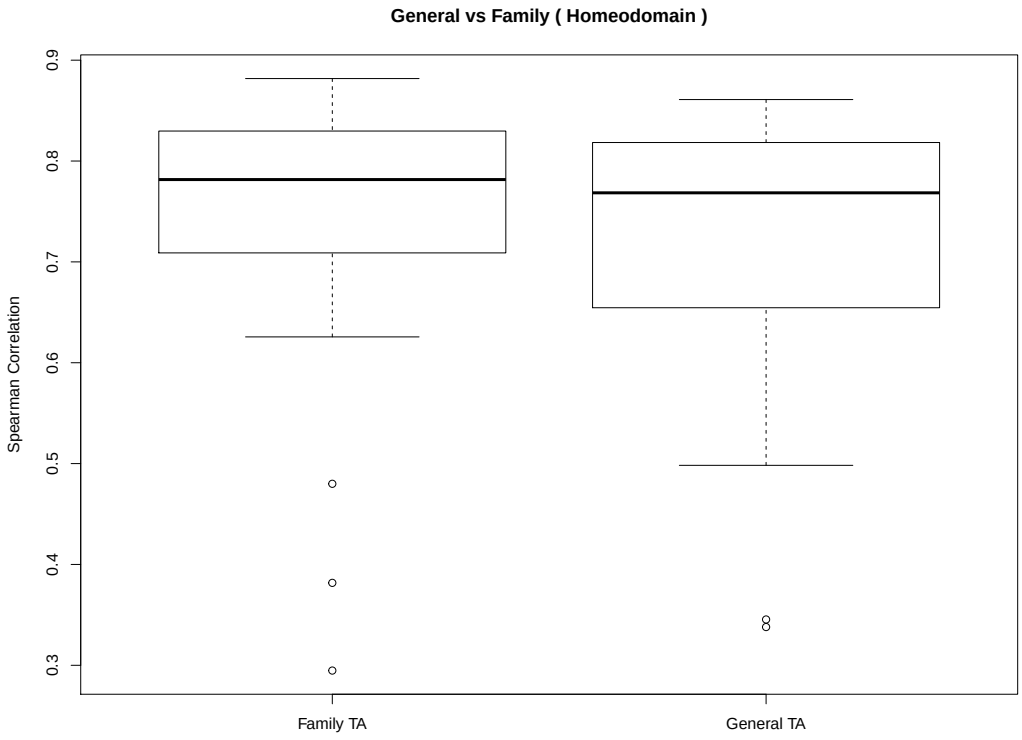


Figure 4.10: Spearman's correlation for the model trained on only Homeodomain transcription factors and a general model trained using 100 transcription factors taken randomly from the three families.

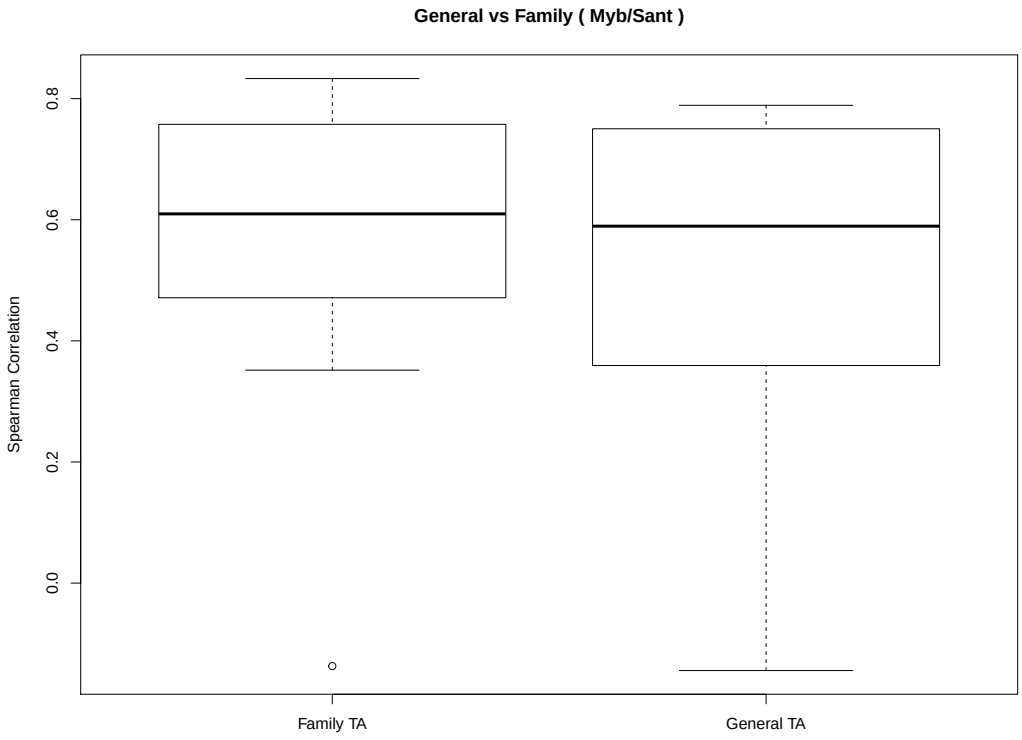


Figure 4.11: Spearman's correlation for the model trained on only Myb/Sant transcription factors and a general model trained using 100 transcription factors taken randomly from the three families.

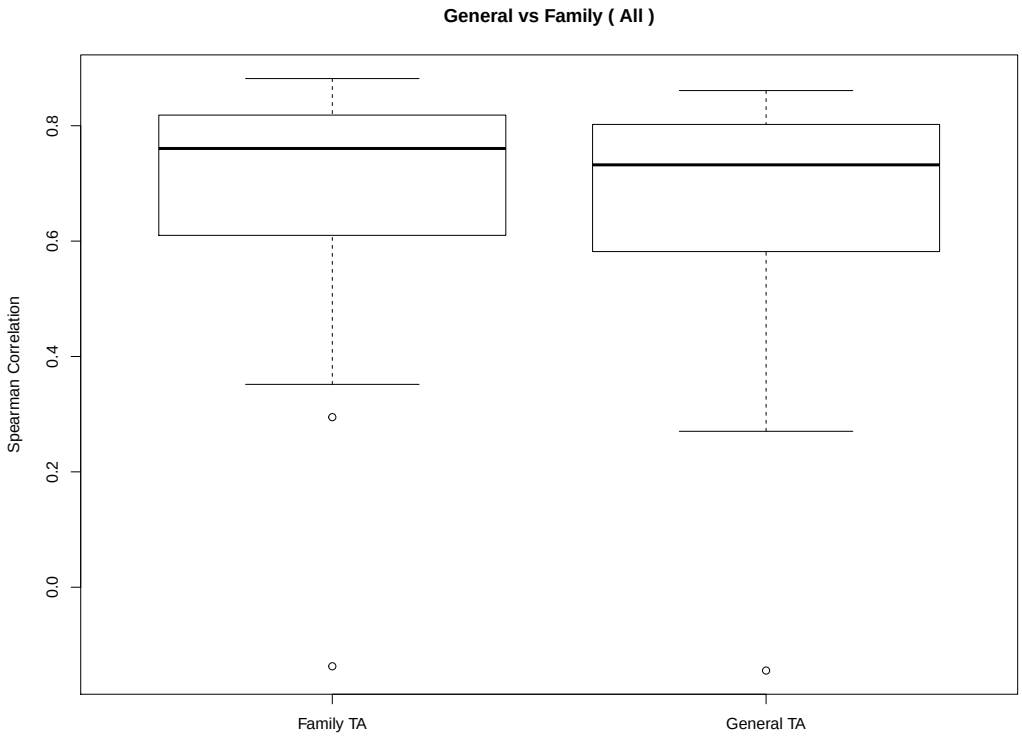


Figure 4.12: Spearman's correlation for the model trained on only the corresponding family of transcription factors and a general model trained using 100 transcription factors taken randomly from the three families.

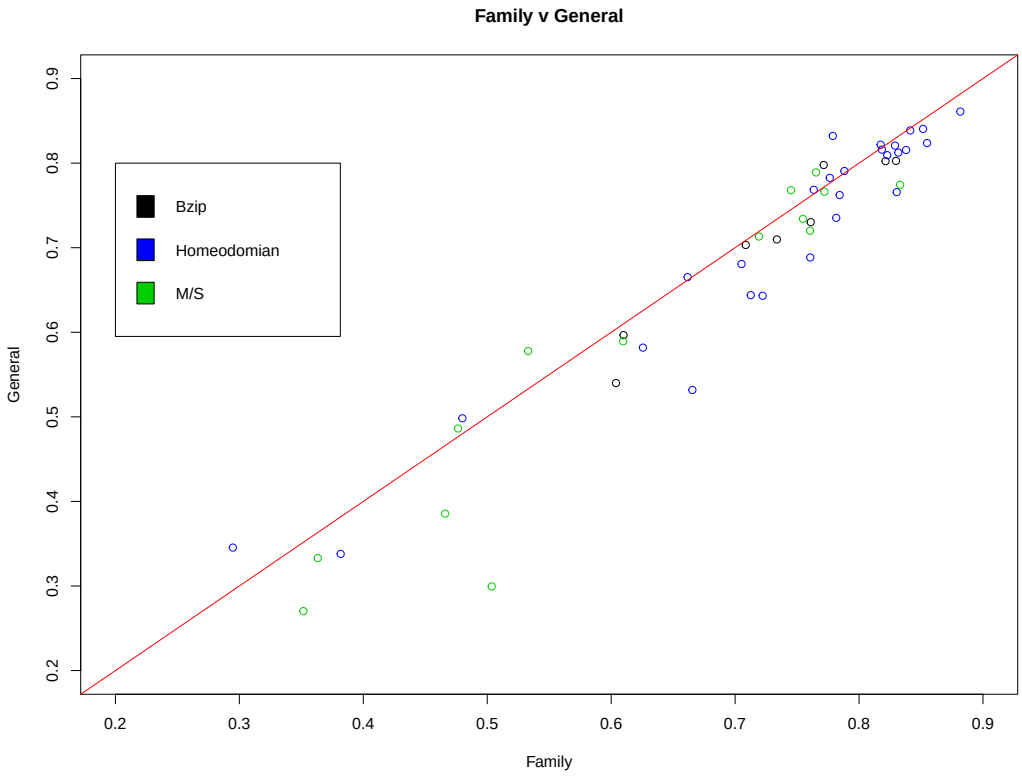


Figure 4.13: Performance in terms of Spearman correlation of TA-family specific and TA-general on a set of randomly selected transcription factors from the three transcription factor families.

Family	Mean	Max	Min	p-value
Homeodomain	-0.0217 ± 0.0398	0.0533	-0.1338	0.4496
Bzip	-0.0196 ± 0.0254	0.0263	-0.0639	0.5737
Myb/Sant	-0.0303 ± 0.0604	0.0448	-0.2041	0.7748
All	-0.024 ± 0.0446	0.0533	-0.2041	0.434

Table 4.4: Spearman correlation differences between the general model and the family model. The last column show the p-value obtained by a WMW test comparing the correlations of the general model and the family specific model.

dataset has 435 transcription factor from 31 families, a break down of the families is seen in Figure 4.14. From this dataset, we picked one of the families (T-Box) [43], and we left it out of the training. As Affinity Regression was designed to run with data from a single-family, Affinity Regression results correspond to leave-one-out cross-validation results using only the T-Box data. For this assessment, we constructed a topological graph using the 435 DNA-binding domain sequences and trained TopAffy using all available 8-mer profiles except those for the eight T-box transcription factors. TopAffy achieved a mean correlation of 0.456 ± 0.064, while Affinity Regression had a mean correlation of -0.0194 ± 0.0502. The Spearman correlation obtained by both programs for the eight members of the T-box family is seen in Figure 4.15. Note that Affinity Regression was designed to predict binding preferences for transcription factors of the same family as those included in the training data. Affinity Regression was restricted to make predictions about the T-box family using a minimal data set with only data for seven transcription factors. This case study demonstrates the benefits of having a general model, such as TopAffy, able to make predictions even

for transcription factors from unseen families or families with a minimal number of transcription factors (i.e., less than ten members).

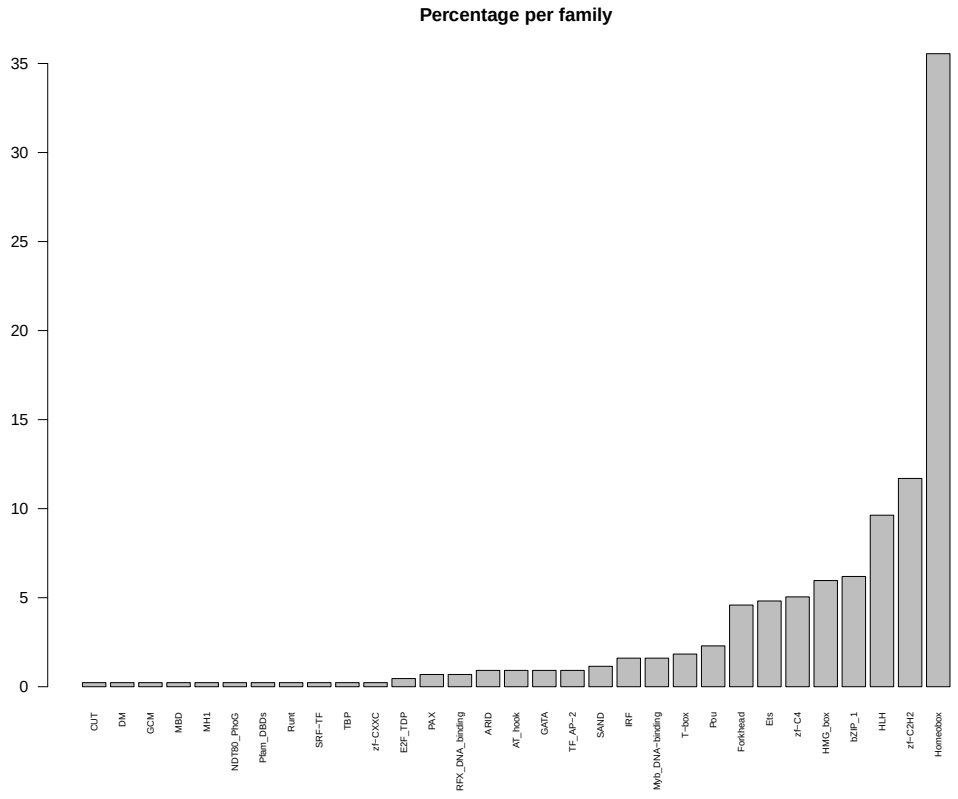


Figure 4.14: The transcription factor family composition (in percentage) for mouse data available on CIS-BP version 2 [1].

4.5 TopAffy properties

To quantify TopAffy's runtime, we ran TopAffy on a Windows 10 64-bit Operating System with an Intel Core i7-6500U CPU 2.50GHz with 12 GB of RAM. Table 4.5 shows the time in seconds to took to build the structures for the TopAffy and the

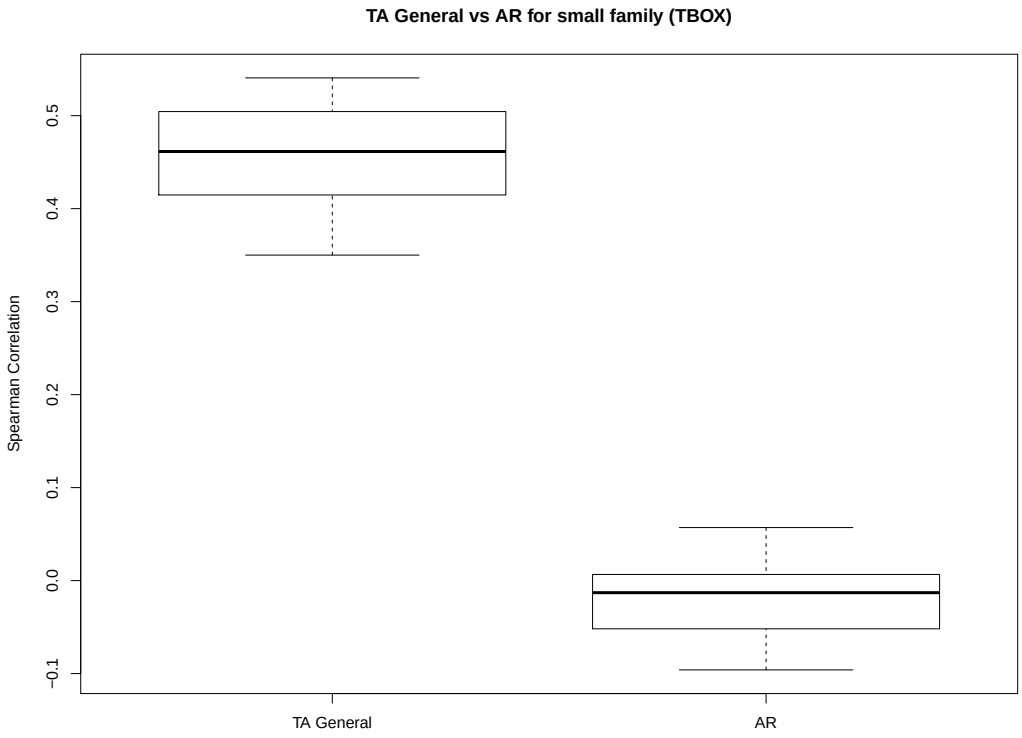


Figure 4.15: Performance of TA-General and Affinity Regression for predicting the Z-score 8-mer profile of the mouse T-Box transcription factor family. We trained TopAffy without T-Box data, and AR results correspond to leave-one-out cross-validation using T-Box data.

time it took TopAffy to make predictions for two runs of different datasets.

Graph building Total (s)	Emission Training Total (s)	Predicting Total (s)	Number of Epochs	Number of Training Sequences	Number of Test Sequences
0.12	12006	1503	6590	100	50
0.22	28377	2519	16542	251	68

Table 4.5: Time in seconds to train TopAffy and make predictions.

The memory usage of TopAffy is allocated mainly on the four data structures of 64-bit floating numbers and one of strings. The first structure stores the binding preference profiles for the training set in a matrix of $n \times 32896$, where n is the number of sequences in the training set. The next structure stores the list of all the sequences as strings. The data structure the keeps the topological graph is a matrix of $441 \times m$, where m is the length of the longest sequence in the dataset. We store the emission matrix as a 3D matrix of $22 \times 22 \times 4096$. The last structure is the map table, with a size of 32896×3. An example of memory usage to train on a dataset with n=100 and m=70 is, $(100 \times 32896 + 441 \times 70 + 22 \times 22 \times 4096 + 32896 \times 3) \times 64 + 5674 \times 8 = 345749200 bits = 43.2 MB$.

Figure 4.16 shows the reduction of the absolute difference over the number of epochs. As expected, the absolute error decreases dramatically during the first few iterations, and then the error reduction slows down until the test for convergence is met.

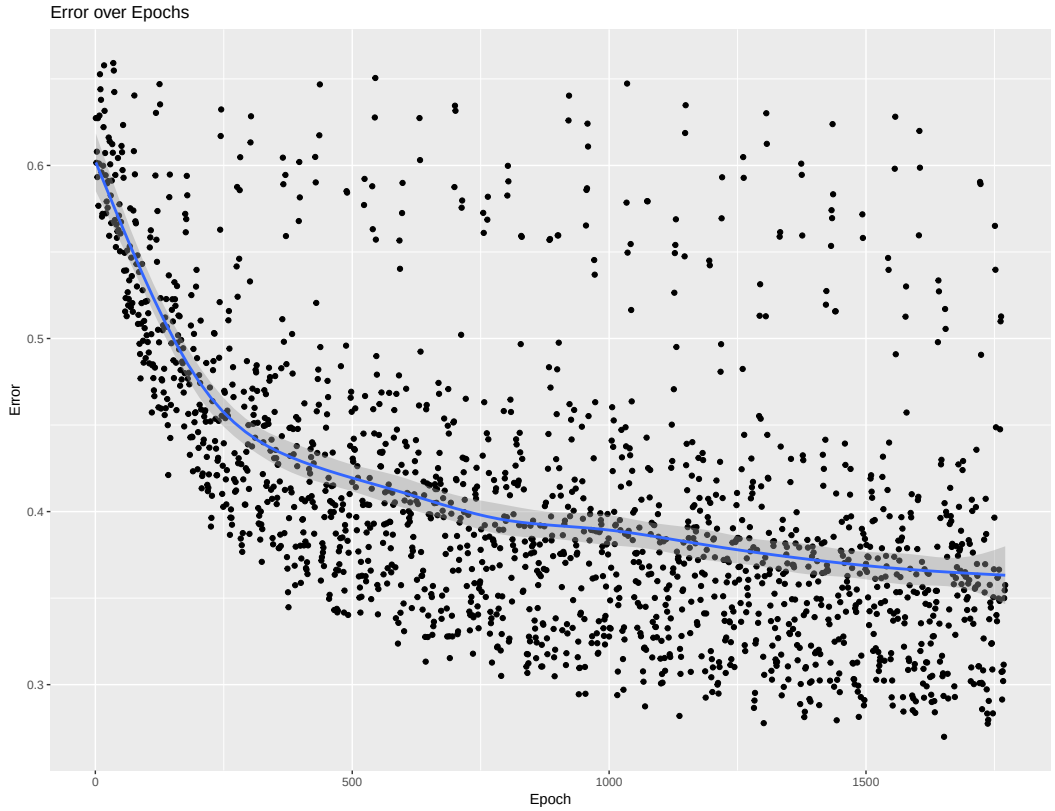


Figure 4.16: Decrease of absolute difference during training. Every data point is the mean absolute error between the real and predicted Z-score for 1248 random 8-mer.

4.6 Summary

In this chapter, we compared the performance of TopAffy to Affinity Regression [37] on four datasets (Homeodomain, BZip, Myb/Sant, and Pelossof*et al.* [37] dataset). We showed that the performance of these two models is statistically comparable for medium to larger families containing at least 50 members. After showing that TopAffy is comparable to Affinity Regression, we then assessed how effective our model would be making predictions for transcription factors whose family was not represented in the training data. With only this limited information, TopAffy was still able to make reasonable predictions for most families. The Homeodomain family saw the worst performance decrease, and we believe this is the case because of how conserved the Homeodomain family is across all samples. Nevertheless, the other two families saw statistically comparable results. TopAffy shows that it is possible to obtain a family-independent model of transcription factor binding preferences.

We evaluated a general model by taking all the three families and combining them into a large sample. Then we partitioned the transcription factors into a training set and a test set. We showed that the general model did a statistically equivalent job at predicting 8-mer profiles as the three family-specific models that we trained. The general model got a mean difference in Spearman correlation of -0.024 $\pm$ 0.0446 from the family-specific TopAffy model.

Finally, we evaluated the performance of TopAffy on predicting the binding preferences of transcription factors from a small transcription factor family (8 members). To run this test, we created a general model that would pull information for many families. We downloaded all the PBM data from mouse (*Mus musculus*) transcription

factors available on CIS-BP version 2 [1]. We selected these species because of the large number of transcription factors it has from a diverse selection of families. Our model was able to get a mean Spearman correlation of 0.456 ± 0.064 for the T-box family (8 members).

In sum, we have demonstrated that TopAffy is the first family-independent approach for predicting transcription factor binding preferences with comparable performance as state-of-the-art family-specific approaches requiring only DBD sequences and 8mer profiles.

Chapter 5

Biological Insights

In this chapter, we look into the insight that we can gather based on the two structures that TopAffy generates. We show that not only TopAffy is a good predictor of transcription factor binding preferences, but also the patterns in the binding preferences matrix are valuable as well to understand binding specificities at the amino-acid pair level.

5.1 Most frequent top-ranked 6-mers are contained in known transcription factor binding motifs

First, we looked at the highest-ranked 6-mer for each AA pair in the binding preferences matrix learned by TopAffy using as input the mouse dataset. The dataset is 427 homeodomains transcription factors from 30 families, which is all mouse PBM data available on the CIS-BP version 2 [1] minus the T-Box family of transcription factors. We saw 73 distinct 6-mers among the top-ranked 6-mers, and out of the

438 amino acid pairs (i.e., 20 amino acids, plus the start and end of the sequence minus two pairs, WC and WEND, which were absent from all DBD sequences on this dataset), 303 (or 69%) had one of ten 6-mers as their top-ranked 6-mer. All of these ten 6-mers are in the consensus sequences of known motifs (Table 5.1).

6-mer	Number of occurrences	Motif	Reference
CACGTG	94	G-box	[44]
AGGTCA	52	DR4	[45]
ACGTAC	29	GMEB2	[46]
CACCTG	26	E-box	[47]
TAAACA	25	Core Forkhead	[48]
TAATTA	22	Homeodomain	[49]
ACGTAA	14	A-box-related	[50]
CCGTTA	14	Ovol1	[51]
CTGTCA	14	Meis/Pknox	[52]
CGCGCG	12	CpGs	[53, 54]

Table 5.1: Ten most frequent top ranked 6-mers.

CACGTG was the most frequent top-ranked 6-mer, a G-box motif bound by transcription factors in the basic helix-loop-helix (bHLH) and basic-leucine zipper (bZIP) families [44]. The second most frequent top 6-mer is AGGTCA, which is a direct repeat (DR) element bound by nuclear receptors (NRs) [45]. Using Tomtom [55], we found a significant match (p-value = 0.0009) for ACGTAC, the third most frequent

top 6-mer, to a motif identified by [46] (accession GMEB2_DBD_1 in footprintDB [56]) bound by transcription factors in the SAND family. The fourth most common top-ranked 6-mer is CACCTG, which is an E-box bound by members of the bHLH transcription factor family [47]. The fifth most common top 6-mer is part of a binding site preferred by the forkhead family, while the homeodomain family contains binding sites that preferred the sixth most common top 6-mer TAATTA [49]. ACGTAA is a 6-mer that is related to A-box [50] and CCGTTA is found in the binding of Ovol1 [51]. The CTGTCA Motfit is seen in Meis/Pknox [52] and lastly CGCGCG is in CpGs [53, 54]. This indicates that TopAffy is learning AA pairs binding preferences related to actual transcription factor binding motifs.

5.2 TopAffy learns distinct binding preferences for amino acid pairs

TopAffy can learn distinct binding preferences, as 22 out of the 73 most preferred 6-mers are also among the 32 least preferred 6-mers (Figure 5.1). We expected this because TopAffy learns a distinct binding preference for each amino acid pair. To visualize the similarity among the AA pairs' inferred binding profiles, we generated a heatmap of the pairwise Pearson correlation coefficient between predicted binding profiles of AA pairs (Figure 5.2). The binding profile of each amino acid pair is only modified by TopAffy if it appears in the transcription factor DBD sequence corresponding to the binding profile it is learning. We looked at whether or not each AA pair appears on the DBD sequences of a given transcription factor family. It

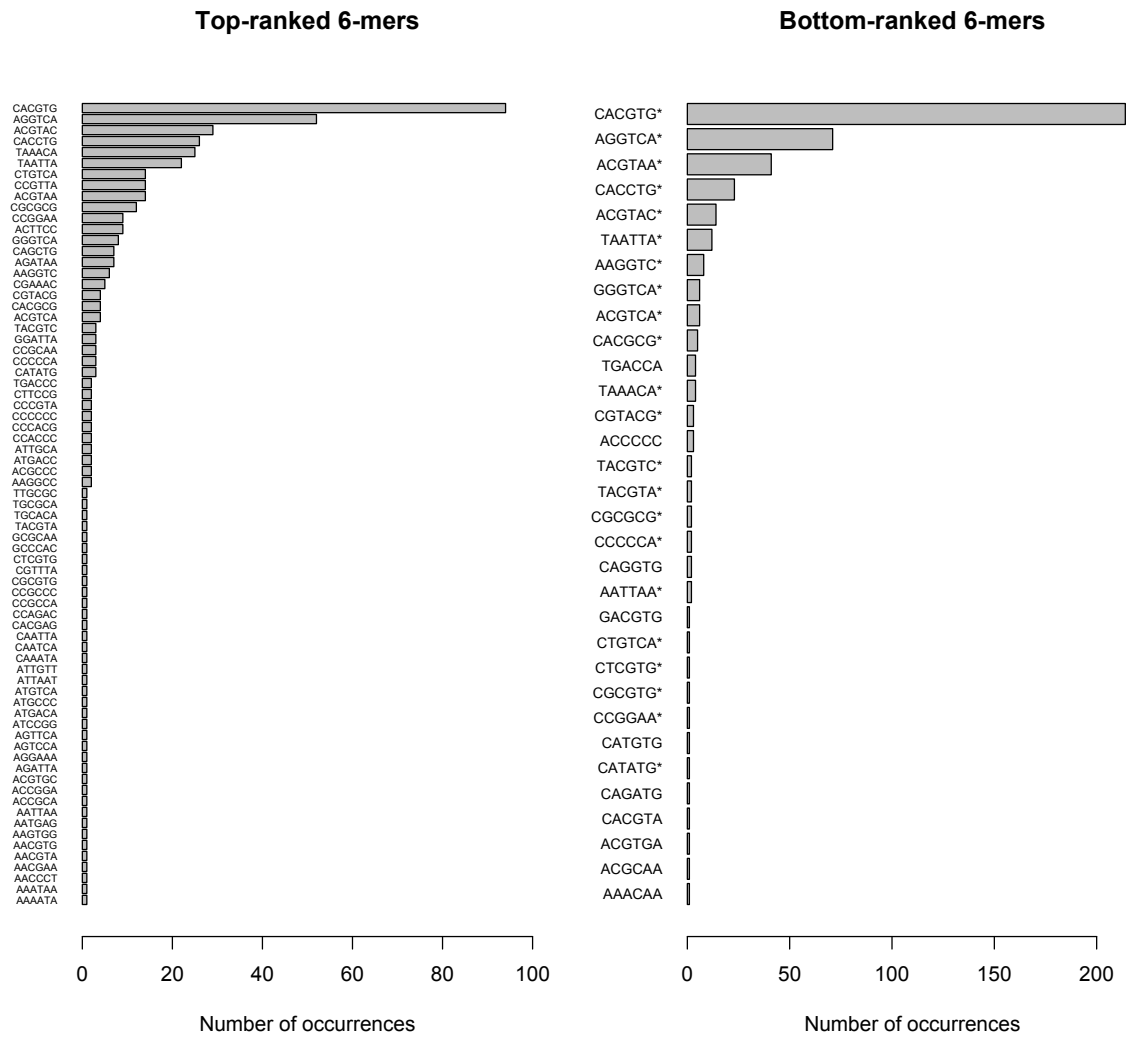


Figure 5.1: Top and Bottom ranked 6-mers. Left: number of occurrences of highest ranked 6-mers. Right: number of occurrences of lowest ranked 6-mers. An asterisk indicate a 6-mer that is a top 6-mer for some AA pairs and a bottom 6-mer for other AA pairs.

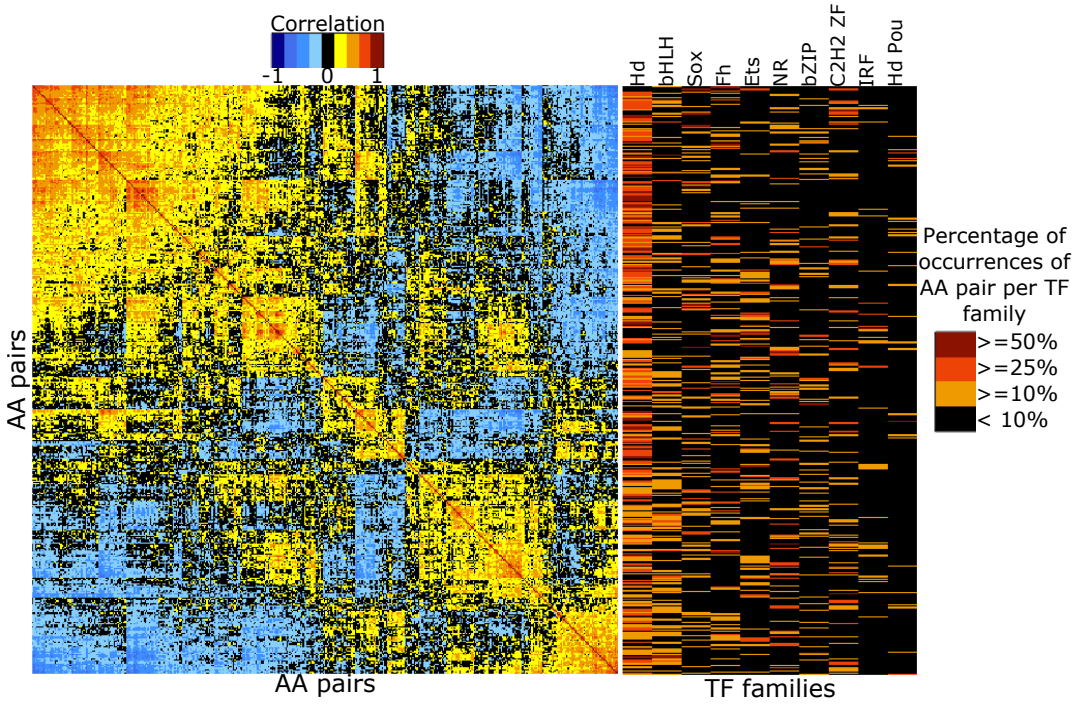


Figure 5.2: Heatmap of similarities among predicted amino acid pairs binding preference profiles. Left: The Pearson correlation coefficients were calculated between the predicted binding preferences of amino acid pairs, and amino acid pairs were ordered using average hierarchical clustering. Right: For each AA pair the percentage of occurrences of this AA pair in DBDs from a specific transcription factor family is shown. Only the ten transcription factor families (Homeodomain - Hd, bHLH, sox, forkhead - Fh, ets, nuclear receptor -NR, bZIP, C2H2 zinc fingers - ZF, IRF, and homeodomain pou - Hd Pou) with the highest percentage of AA pairs occurrences are shown.

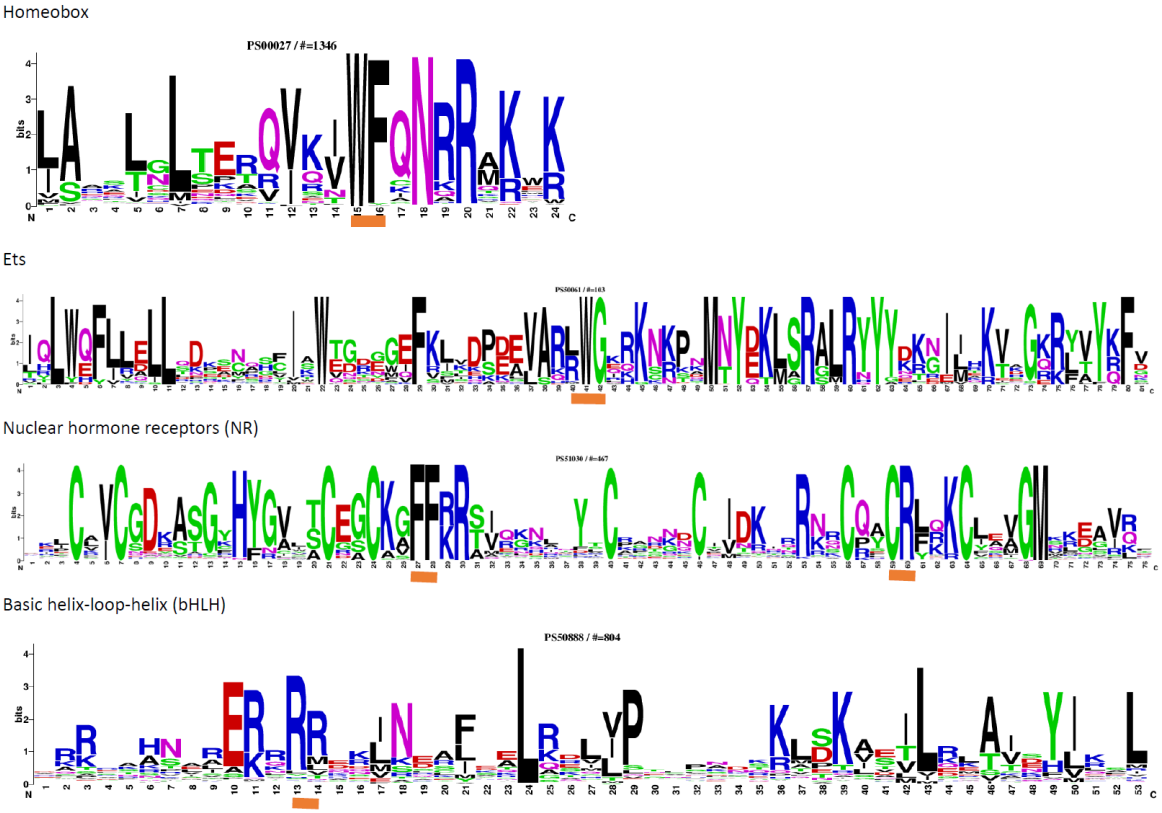


Figure 5.3: Sequence logos from alignments of the DNA binding domains. Amino acid pairs chosen for Figure 5.2 in the manuscript are underlined with a red line. Sequence logos available in PROSITE [57].

turned out that 35% of transcription factors for the mouse dataset belong to the homeodomain family. This reflects the fact that the homeodomain family has a high percentage of AA pair occurrences. There are groups of amino acid pairs with highly similar predicted binding profiles as well as with negatively correlated binding profiles, which suggests that TopAffy learns distinct binding preferences for AA pairs.

We also looked at whether the binding preference profile of AA pairs strongly conserved in a specific transcription factor family reflects the known binding preferences of that family. To do this, we obtained the sequence logo for the homeodomain, Ets, Nuclear receptor (NR), and basic helix-loop-helix (bHLH) DBD from PROSITE [57]. We identified highly conserved AA pairs for each of these families (Figure 5.3). Three of these AA pairs, namely WF, WG and FF appear mostly in the DBD of the associated family (Figure 5.4). We observed that the top-ranked 6-mers for WF, a signature AA pair at position 15 of the homeodomain DBD, are indeed contained in known homeodomain motifs [49] (Figure 5.4). Similarly, the top-ranked 6-mers for WG are part of known Ets motifs [58]. From the NR sequence logo, we selected the AA pairs FF and CR. These two AA pairs, in addition, to have top-ranked 6-mers that are part of known NR motifs [58, 45] and have similar binding profiles. The binding profiles for the 41 6-mers shown in Fig. 5.4 and for all possible 6-mers have a Spearman correlation of 0.93 and 0.58, respectively. Finally, RR, which has a strong signal at position 13 of the bHLH sequence logo, strongly prefers CACGTG, which is known to be bound by bHLH transcription factors [44]. These results suggest that TopAffy is indeed inferring biologically relevant binding profiles for AA pairs.

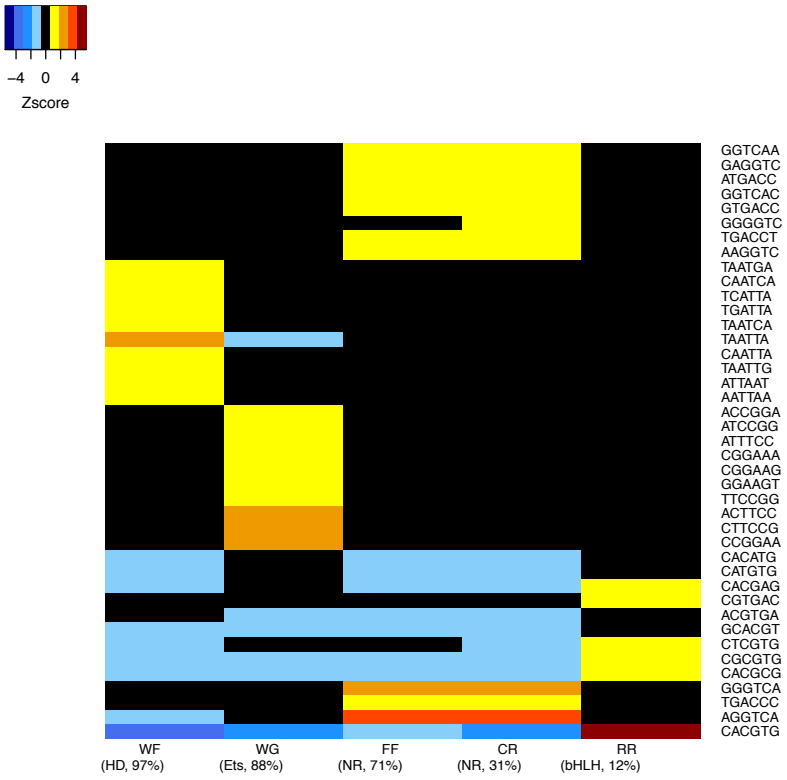


Figure 5.4: Most preferred 6-mers for specific AA pairs highly conserved in certain DBDs. The ten top-ranked 6-mers for each of the AA pairs shown were selected. AA pairs were selected by looking at sequence logos (Figure 5.3) corresponding to the following families: homeodomain (HD), Ets, Nuclear receptor (NR), and basic helix-loop-helix (bHLH). Between brackets below each AA pair, their associated transcription factor family and the percentage of occurrences of that AA pair that happen in the DBDs of the corresponding family are provided. To obtain these binding preferences and percentage of occurrences, we used as input for TopAffy the PBM binding data and DBD sequences of 427 murine transcription factors from 31 families.

5.3 Summary

We showed that the emission matrix provides insights into binding preferences at the amino-acid pair level. The 6-mers that ranked highest for the majority of AA pairs are contained in known transcription factors motifs. Additionally the binding preferences of AA pairs strongly conserved in a specific transcription factor family reflect the known binding preferences of this family. This suggests that TopAffy is learning binding preferences associated with actual transcription factors binding motifs.

Chapter 6

Conclusions and Future Work

In this thesis, we presented a new model for predicting binding preferences for transcription factors called TopAffy. This model uses two structures to make predictions. One is a topological sequence graph, and the other is an emission matrix. The topological sequence graph is based on the DBD sequences, and TopAffy trains an emission matrix based on 8-mer PBM binding profiles.

Our model was compared to the Affinity Regression method created by Pelossof *et al.* [37]. We showed that our model has comparable performance in terms of Spearman correlation to Affinity Regression for medium to large families. We also showed that our model outperforms Affinity Regression for a small family. Our model has also comparable performance when comparing a general model (trained with data from several families) and a "blind" model (trained with the family to be predicted left out). Making TopAffy amongst the first family-independent approach for predicting binding preferences for transcription factors.

We showed that not only is TopAffy a good predictor of transcription factor bind-

ing preference, but also the emission matrix provides insights into binding preferences at the amino-acid pair level. We showed that top-ranked 6-mers are contained in known transcription factors motifs and that binding preferences of AA pairs strongly conserved in a specific transcription factors family reflect the known binding preferences of this family. This suggests that TopAffy is indeed learning binding preferences associated to actual transcription factors binding motifs.

As a family independent predictor of transcription factor binding preferences, TopAffy can predict binding preferences for small families or families without PBM data available. This is essential because 67 out of 85 families with PBM data on CIS-BP version 2 [1] have less than 50 members, and 41 have less than 10 members.

As future work, one could use the emission matrix and topological graph to predict how genetic variations in a DBD sequence or transcription factor binding site affects the transcription factor binding activity. As these two structures together capture the information about how transcription factor binding preferences are affected by the transcription factor sequence. Hence, one could use the two structures that TopAffy generate to predict binding profiles of transcript factors with genetic variants, similar to what was done by Zhou and Troyanskaya [59] and Barrera [60].

Since TopAffy can make reasonable general predictions, we could train a model using all DBD sequences and all PBM data available and use it to predict the binding preference profiles for any new DBD sequence. To facilitate the addition of new data to TopAffy, TopAffy could be modified to allow for online learning or for using batch training where the already pre-trained model gets further trained with n new observations. This would allow us to overcome the biggest downside of TopAffy, which is the time it takes to train the emission matrix.

One more aspect that can be looked at is whether the model always converges and what the optimal values for the training parameters (η , λ, ϵ) are.

Bibliography

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

[2] Samuel A Lambert, Arttu Jolma, Laura F. Campitelli, Jussi Taipale, Timothy R. Hugnes, and Matthew T. Weirauch. The human transcription factors. *Cell*, 172(4):P650–665, 2018.

[3] Philippe Collas. The current state of chromatin immunoprecipitation. *Molecular Biotechnology*, 45(1):87–100, May 2010.

[4] Michael F. Berger and Martha L. Bulyk. Protein binding microarrays (PBMs) for the rapid, high-throughput characterization of the sequence specificities of

DNA binding proteins. *Method in Molecular Biology*, 338:245–260, 2006.

[5] Andrew J. Colebatch, Leon Di Stefano, Stephen Q. Wong, Ross D. Hannan, Paul M. Waring, Alexander Dobrovic, Grant A. McArthur, and Anthony T. Papenfuss. Clustered somatic mutations are frequent in transcription factor binding motifs within proximal promoter regions in melanoma and other cutaneous malignancies. *Oncotarget*, 7(41):66569–66585, 2016.

[6] James Darnell. Transcription factors as targets for cancer therapy. *Nature Reviews Cancer*, (2):740–749, October 2002.

[7] Natthapol Songdej and A. Koneti Rao. Hematopoietic transcription factor mutations: important players in inherited platelet defects. *Blood*, 129(21):2873–2881, 2017.

[8] Jin-Yan Liu, Wen-Jing Li, Ce Shi, Ying Wang, Yue Zhao, and Ming-Jie Xiang. Mutations in the flo8 transcription factor contribute to virulence and phenotypic traits in *Candida albicans* strains. *Microbiological Research*, 178:1–8, 2015.

[9] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. Transcription factor-DNA binging: Beyond binding site motifs. *Current Opinion in Genetics and Development*, 43:110–119, April 2017.

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

[11] Yue Zhao, Shuxiang Ruan, Manishi Pandey, and Gary D Stormo. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, 191:781–790, July 2012.

[12] Hanjun Dai, Ramzan Umarov, Hiroyuki Kuwahara, Yu Li, Le Song, and Xin Gao. Sequence2vec: A novel embedding approach for modeling transcription factor binding affinity landscape. *Bioinformatics*, 33(22):3575–3583, 2017.

[13] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, DREAM5 consortium, Harmen J. Bussemaker, Quaid D. Morris, Martha L. Bulyk, Gustavo Stolovitzky, and Timothy R. Hughes. Evaluation of methods for modeling transcription-factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, 2013.

[14] Masashi Suzuki and Naoto Yagi. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proceedings of the National Academy of Science of the United States of America*, 91(26):12357–12361, 1994.

[15] Hidetoshi Kono and Akinori Sarai. Structure-base prediction of DNA target sites by regulatory proteins. *PROTEINS: Strcture, Function, and Genetics*, 35:114–131, 1999.

[16] Panayiotis V. Benos, Martha L. Bulyk, and Gary D. Stormo. Additivity in protein-DNA interactions: How good an approximation is it? *Nucleic Acids Research*, 30(20):4442–4451, 2002.

[17] Tommy Kaplan, Nir Friedman, and Hanah Margalit. *Ab initio* prediction of transcription factor targers using structural knowledge. *PloS Computational Biology*, 1(1), 2005.

[18] Anton V Persikov, Robert Osada, and Mona Singh. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, 25(1):22–29, 2009.

[19] Ryan G. Christensen, Metewo Selase Enuameh, Marcus B. Noyes, Michael H. Brodsky, Scot A. Wolfe, and Gary D. Stormo. Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics*, 28:i84–i89, 2012.

[20] Munazah Andrabi, Andrew Paul Hitchins, Diego Miranda-Saavedra, Hidetoshi Kono, Ruth Nussinov, Kenji Mizuguchi, and Shandar Ahmad. Prediction conformational ensembles and genome-wide transcription factor binding sites from DNA sequences. *Scientific Reports*, 7(4071), 2017.

[21] Jinsen Li, Jared M. Sagendorf, Tsu-Pei Chiu, Marco Pasi, Alberto Perez, and Remo Rohs. Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Research*, 45(22):12877–12887, 2017.

[22] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research*, 41:W56–W62, 2013.

[23] Kiyoshi Ezawa. Characterization of multiple sequence alignment errors using complete-likelihood score and position-shift map. *BMC Bioinformatics*, 17, 2016.

[24] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordan, and Remo Rohs. Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences*, 39, 2014.

[25] Yael Mandel-Gutfreund and Hanah Margalit. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Research*, 26(10):2306–2312, 1998.

[26] Qing Zhou and Jun S. Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916, 2004.

[27] Trevis M. Alleyne, Lourdes Peña-Castillo, Gwenael Badis, Shaheynoor Talukder, Michael F. Berger, Andrew R. Gehrke, Anthony A. Philippakis, Martha L. Bulyk, Quaid D. Morris, and Timothy R. Hughes. Predicting the binding preference of transcription factors to individual DNA k-mers. *Bioinformatics*, 25(8):1012–1018, 2009.

[28] Michael F. Berger, Gwenael Badis, Andrew R. Gehrke, Shaheynoor Talukder, Anthony A. Philippakis, Lourdes Peña-Castillo, Trevis M. Alleyne, Sanie Mnaimneh, Olga B. Botvinnik, Esther T. Chan, Faiqua Khalid, Wen Zhang, Daniel Newburger, Savina A. Jaeger, Quaid D. Morris, Martha L. Bulyk, and Timothy R. Hughes. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266—1276, 2008.

[29] Jiajian Liu and Gary D. Stormo. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, 24(17):1850–1857, June 2008.

[30] Zhen Shen, Wenzheng Bao, and De-Shuang Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific Reports*, 8(15270), 2018.

[31] Chaitanya Rastogi, H. Tomas Rube, Judith F. Kribelbauer, Justin Crocker, Ryan E. Lokere, Gabriella D. Martini, Oleg Laptenkoc, William A. Freed-Pastor, Carol Prives, David L. Stern, Richard S. Mann, and Harmen J. Bussemaker. Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences of the United States of America*, 115(16):E3692–E3701, 2018.

[32] Yue Zhao, David Granas, and Gary D. Stormo. Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5(12), December 2009.

[33] Anton V Persikov and Mona Singh. An expanded binding model for Cys2His2 zinc finger protein DNA interfaces. *Physical Biology*, 8(3), 2011.

[34] Anton V. Persikov and Mona Singh. *De novo* prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Research*, 42(1):97–108, 2014.

[35] Ankit Gupta, Ryan G. Christensen, Heather A. Bell, Mathew Goodwin, Ronak Y. Patel, Manishi Pandey, Metewo Selase Enuameh, Amy L. Rayla, Cong Zhu, Stacey Thibodeau-Beganny, Michael H. Brodsky, J. Keith Joung, Scot A.

Wolfe, and Gary D. An improved predictive recognition model for Cys2-His2 zinc finger proteins. *Nucleic Acids Research*, 42(8):4800—-4812, February 2014.

[36] Matti Annala, Kirsti Laurila, and Matti Nykter Harri Lähdesmäki. A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS One*, 6(5):e20059, May 2011.

[37] Raphael Pelossof, Irtisha Singh, Julie L Yang, Matthew T Weirauch, Timothy R Hughes, and Christina S Leslie. Affinity regression predicts the recognition code of nucleic acid–binding proteins. *Nature Biotechology*, 33(1126):1242–1250, 2015.

[38] Samuel A. Lambert, Ally W. H. Yang, Alexander Sasse, Gwendolyn Cowley, Mihai Albu, Mark X. Caddick, Quaid D. Morris, Matthew T. Weirauch, and Timothy R. Hughes. Similarity regression predicts evolution of transcription factor sequence specificity. *Nature Genetics*, (51):981–989, 2019.

[39] Jens Keilwagen, Jan Grau, Ivan A. Paponov, Stefan Posch, Marc Strickert, and Ivo Grosse. De-novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Computational Biology*, 7(2):e1001070, February 2011.

[40] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[41] Grigoris Amoutzias, Amelie Veron, III Weiner, January, Marc Robinson-Rechavi, Erich Bornberg-Bauer, Stephen Oliver, and David Robertson. One billion years of bZIP transcription factor evolution: Conservation and change in dimerization

and DNA-binding site specificity. *Molecular Biology and Evolution*, 24(3):827–835, 2006.

[42] Asmaa M. Baker, Qiang Fu, William Hayward, Stuart M. Lindsay, and Terace M. Fletcher. The Myb/SANT domain of the telomere-binding protein TRF2 alters chromatin structure. *Nucleic Acids Research*, 37(15):5019–5031, 2009.

[43] Virginia E. Papaioannou. The t-box gene family: emerging roles in development, stem cells and cancer. *Development*, 141(20):3819–3833, October 2014.

[44] Daphne Ezer, Samuel J K Shepherd, Anna Brestovitsky, Patrick Dickinson, Sandra Cortijo, Varodom Charoensawan, Mathew S Box, Surojit Biswas, Katja E Jaeger, and Philip A Wigge. The G-Box transcriptional regulatory code in Arabidopsis. *Plant physiology*, 175(2):628–640, Oct 2017.

[45] Ashley Penvose, Jessica L Keenan, David Bray, Vijendra Ramlall, and Trevor Siggers. Comprehensive study of nuclear receptor DNA binding provides a revised framework for understanding receptor specificity. *Nature Communications*, 10(1):2514, 06 2019.

[46] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39, Jan 2013.

[47] Hiroshi Kataoka, Toshinori Murayama, Masayuki Yokode, Seiichi Mori, Hideto Sano, Harunobu Ozaki, Yoshifumi Yokota, Shin-Ichi Nishikawa, and Toru Kita. A novel snail-related transcription factor smuc regulates basic helix-loop-helix transcription factor activities via specific E-box motifs. *Nucleic Acids Research*, 28(2):626–33, Jan 2000.

[48] Xi Chen, Zongling Ji, Aaron Webber, and Andrew D Sharrocks. Genome-wide binding studies reveal DNA binding specificity mechanisms and functional interplay amongst Forkhead transcription factors. *Nucleic Acids Research*, 44(4):1566–78, Feb 2016.

[49] 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, Hanna Kuznetsov, Chi-Fong Wang, David Coburn, Daniel E Newburger, Quaid Morris, Timothy R Hughes, and Martha L Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–3, Jun 2009.

[50] Pavel Cherenkov, Daria Novikova, Nadya Omelyanchuk, Victor Levitsky, Ivo Grosse, Dolf Weijers, and Victoria Mironova. Diversity of cis-regulatory elements associated with auxin response in *Arabidopsis thaliana*. *Journal of Experimental Botany*, 69(2):329–339, 01 2018.

[51] Mahalakshmi Nair, Andy Teng, Virginia Bilanchone, Anshu Agrawal, Baoan Li, and Xing Dai. Ovol1 regulates the growth arrest of embryonic epidermal

progenitor cells and represses c-myc transcription. *The Journal of Cell Biology*, 173(2):253–64, Apr 2006.

[52] Joseph Martin Grice. *The role of vertebrate conserved non-coding elements in hindbrain development and evolution*. PhD thesis, UCL (University College London), 2016.

[53] Dominik Hartl, Arnaud R Krebs, Ralph S Grand, Tuncay Baubec, Luke Isbel, Christiane Wirbelauer, Lukas Burger, and Dirk Schübeler. CG dinucleotides enhance promoter activity independent of DNA methylation. *Genome Research*, 29(4):554–563, 04 2019.

[54] Aimée M Deaton and Adrian Bird. CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010–22, May 2011.

[55] Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome Biology*, 8(2):R24, 2007.

[56] Alvaro Sebastian and Bruno Contreras-Moreira. footprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, 30(2):258–65, Jan 2014.

[57] Christian J A Sigrist, Edouard de Castro, Lorenzo Cerutti, Béatrice A Cuche, Nicolas Hulo, Alan Bridge, Lydie Bougueleret, and Ioannis Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(Database issue):D344–7, Jan 2013.

[58] Yimeng Yin, Ekaterina Morgunova, Arttu Jolma, Eevi Kaasinen, Biswajyoti Sahu, Syed Khund-Sayeed, Pratyush K Das, Teemu Kivioja, Kashyap Dave, Fan Zhong, Kazuhiro R Nitta, Minna Taipale, Alexander Popov, Paul A Ginno, Silvia Domcke, Jian Yan, Dirk Schübeler, Charles Vinson, and Jussi Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337), 05 2017.

[59] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning—based sequence model. *Nature Methods*, 12(10):981–934, October 2015.

[60] Luis A. Barrera, Anastasia Vedenko, Jesse V. Kurland, Julia M. Rogers, Stephen S. Gisselbrecht, Elizabeth J. Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, Trevor Siggers, Leila Shokri, Raluca Gordân, Nidhi Sahni, Chris Cotsapas, Tong Hao, Song Yi, Manolis Kellis, Mark J. Daly, Marc Vidal, David E. Hill, and Martha L. Bulyk. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science*, 351(6280):1450–1454, September 2016.