Modeling protein-DNA binding specificities with random forest

Anni Antikainen

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 4.1.2018

Thesis supervisor:

Prof. Harri Lähdesmäki

Thesis advisor:

PhD Markus Heinonen

Aalto University
School of Science

AALTO UNIVERSITY
SCHOOL OF SCIENCE

ABSTRACT OF THE
MASTER'S THESIS

Author: Anni Antikainen

Title: Modeling protein-DNA binding specificities with random forest

Date: 4.1.2018	Language: English	Number of pages: 7+86

Department of Computer Science

Professorship: Computational Systems Biology

Supervisor: Prof. Harri Lähdesmäki

Advisor: PhD Markus Heinonen

Protein-DNA binding specifities are modeled with random forest in this Master's thesis. Specific proteins called transcriptional factors are essential for gene expression regulation, since their binding on DNA can alter transcription initiation probability of target genes. Furthermore, transcriptional factors can bind DNA as dimers even though as individuals they would lack the required affinity for the binding site. Thus, models that predict individual protein and protein dimer binding sites, would be beneficial for deducing gene regulatory networks. In this Master's thesis HT-SELEX and CAP-SELEX data sets measured by Jolma et al. are utilized for modeling binding specificities. SELEX measurements yield large sets of DNA sequences, which are known to comprise a binding site. HT-SELEX measure individual transcriptional factor binding sites while CAP-SELEX measure binding sites of transcriptional factor dimers. Currently, position weight matrices (PWM) are most often utilized for modeling protein-DNA binding specifities even though they may be too simple and inflexible for accurate modeling. For instance a neural network model, DeepBind, have been shown to outperform PWM modeling significantly. In this Master's thesis, random forest, which is known to be well suited for high-dimensional and correlated data, is combined with PWMs to yield models for protein-DNA binding specifities. For individual transcriptional factor binding sites random forest perform almost equally to DeepBind and outperform PWM modeling significantly. In addition, random forest predict protein dimer binding sites significantly more accurately than position weight matrices. Furthermore, the difference between random forest and PWM modeling is greater for protein pairs than for individual proteins. In addition, DeepBind is not currently provided for transcriptional factor pairs. Thus, according to results represented in this Master's thesis, modeling protein-DNA binding specificities with random forest is beneficial in comparison to position weight matrices especially for protein dimers.

Keywords: transcriptional factor, motif, binding specificity, gene expression, position weight matrix, supervised learning, decision tree, random forest

AALTO-YLIOPISTO
PERUSTIETEIDEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Anni Antikainen

Työn nimi: Proteiini-DNA sitoutumisspesifisyyksien mallintaminen satunnaismetsällä

Päivämäärä: 4.1.2018 Kieli: Englanti Sivumäärä: 7+86

Tietojenkäsittelytieteen laitos

Professuuri: Laskennallinen systeemibiologia

Työn valvoja: Prof. Harri Lähdesmäki

Työn ohjaaja: FT Markus Heinonen

Diplomityössä mallinnetaan satunnaismetsällä proteiini-DNA sitoutumisspesifisyyksiä. Transkriptiotekijät ovat proteiineja, jotka säätelevät geenien ilmentymistä sitoutumalla DNA juosteelle ja täten laskemalla tai kasvattamalla kohdegeenien transkription todennäköisyytä. Lisäksi transkriptiotekijät voivat sitoutua DNA juosteelle dimeerisessä muodossa, vaikka yksittäisinä proteiineina näiden sitoutumisaffiniteetti ei olisikaan ollut riittävä kyseiselle sitoutumiskohdalle. Diplomityössä käytetään sitoutumisspesifisyyksien mallintamiseen Jolma et al. mittaamia HT-SELEX ja CAP-SELEX aineistoja. SELEX mittaukset tuottavat suuren joukon DNA juosteita, jotka sisältävät sitoutumiskohdan. HT-SELEX menetelmällä mitataan sitoutumiskohtia yksittäisille proteiineille ja CAP-SELEX menetelmällä proteiinipareille. Tällä hetkellä sitoutumisspesifisyyksiä mallinnetaan useimmiten positio paino matriiseilla (PPM), vaikka ne saattavat olla liian yksinkertaisia ja joustamattomia sitoutumiskohtien todenmukaiseen mallintamiseen. Esimerkiksi neuroverkkoihin perustuvan DeepBind mallin on näytetty ennustavan sitoutumiskohtia merkittävästi tarkemmin kuin positio paino matriisien. Diplomityössä mallinnetaan proteiinien sitoutumiskohtia yhdistämällä PPM malleja ja satunnaismetsä-mallinnusta, jonka tiedetään soveltuvan hyvin moniulotteiselle sekä korreloituneelle datalle. Työn tuloksista selvisi, että satunnaismetsä ennustaa yksittäisten proteiinien sitoutumiskohtia lähes samalla tarkkuudella kuin DeepBind ja että ennustustarkkuus on merkittävästi korkeampi kuin PPM malleilla. Satunnaismetsällä voi lisäksi mallintaa proteiiniparien sitoutumiskohtia merkittävästi tarkemmin kuin positio paino matriiseilla. Ero ennustustarkkuudessa satunnaismetsän ja PPM mallinnuksen välillä on suurempi proteiinipareilla kuin yksittäisillä proteiineilla. Lisäksi DeepBindia ei tarjota tällä hetkellä proteiinipareille. Täten Diplomityön tulosten perusteella satunnaismetsä on suositeltava menetelmä proteiini-DNA sitoutumisspesifisyyksien mallintamiseen erityisesti dimeeristä sitoutumista mallinnettaessa.

Avainsanat: transkriptiotekijä, motiivi, sitoutumisspesifisyys, geeniekspressio, positio paino matriisi, ohjattu oppiminen, päätöspuu, satunnaismetsä

Preface

I want to thank Professor Harri Lähdesmäki for offering me this interesting Master's thesis project and for being a very supportive instructor. In addition, I thank Markus Heinonen for all the help that I have received during this process. Your encouraging comments and advices have been vital for me. A big thank you also for everyone in the Computational systems biology group for creating a great working environment. Computational resources provided by Science-IT at Aalto University were used for the analysis conducted in this Master's thesis. In addition, I want to thank my family. I thank Ville for being my inspiration and for teaching me objective thinking. Merja, thank you for always being there for me and for all the guidance and support that you have given me. In addition, I want to thank Emma for teaching me not to take life too seriously and for laughing with me for the silliest things. I also thank my closest friends, Maustetytöt, for the unconditional support and funny moments that we have shared. Especially, I thank E. J. Eines and S. N. Pyttis for being my nerdy friends in high school. Finally, I thank Tuomas for encouraging me throughout my studies and for telling me your stories, which always light up my mood.

Otaniemi, January 4, 2018

Anni A. V. Antikainen

Contents

Symbols and abbreviations

DNA	Deoxyribonucleic acid
TF	Transcriptional factor
SELEX	Evolution of ligands by exponential enrichment
HT-SELEX	High-throughput SELEX
CAP-SELEX	Consecutive affinity-purification SELEX
RF	Random forest
PWM	Position weight matrix
DB	DeepBind
N_T	Number of trees in a random forest
N_S	Minimum number of instances in a decision tree node
N_V	Number of variables tried at a data partitioning in a decision tree
Strand+RF	Random forest with full length DNA strand
PWM+RF	Random forest with sites chosen by PWM
N+PWM+RF	Random forest with sites chosen by PWM from DNA sequences padded with 'N'
PWM+N+RF	Random forest with elongated sites chosen by PWM
RF+RF	Random forest with sites chosen by other random forests
ROC	Receiver operating characteristics
AUC	Area under ROC-curve
PWM1	PWM of TF1 in TF pair
PWM2	PWM of TF2 in TF pair
PWM1+RF	Random forest with sites chosen by PWM1
PWM2+RF	Random forest with sites chosen by PWM2
PWM1+PWM2+RF	Ensemble of random forests with best spacings between PWM1 and PWM2
PWM1+PWM2+N+RF	Random forest with elongated sites searched with PWM1 or PWM2

1 Introduction

Proteins regulate gene expression through interactions with DNA that further regulate cellular processes. In a cell, gene expression is initiated through binding of specific proteins, transcriptional factors (TF), on DNA sequence. Transcriptional factors increase chromatin accessibility locally and aid the assembly of transcription complex to the promoter sequence, which will initiate transcription. Furthermore, transcriptional factors can bind enhancer sequences far away from the promoter and increase or decrease the probability for transcription initiation to take place. Thus, TF binding to DNA can be understood as a mechanism to turn genes on and off. [1] Furthermore, transcriptional factor interactions with DNA can be represented as gene regulatory networks with direct relationships between target genes and transcriptional factors. In addition, gene regulatory networks could potentially be deduced from knowledge of TF binding sites. [2] Transcriptional factors bind DNA in a sequence specific manner. Binding specificity, whose representation is called a motif, describe the ability of a protein to distinguish between putative binding sites. [3] Currently, binding specifities of many transcriptional factors have not been measured. In addition, transcriptional factors can bind DNA in pairs with different spacings between the motifs. The dimeric binding of proteins increase complexity of gene regulatory networks. [4] Furthermore, in order to understand gene regulatory networks, models that describe all putative binding sites of transcriptional factors would be beneficial. In vitro measurements, referring to experiments conducted outside the cell, can be utilized for building such specificity models. [5]

In this Master's thesis protein-DNA binding specifities are modeled with random forest with data acquired by Jolma et al. with evolution of ligands by exponential enrichment (SELEX) in vitro measurements [5, 4]. SELEX is a method for selecting DNA sequences from a large DNA ligand library. Binding sites can be revealed by incubating the protein with a DNA library and separating bound DNA sequences from the non-bound sequences. [6] Furthermore, high-throughput (HT) SELEX is a method capable of measuring multiple transcriptional factor binding specifities in parallel [7]. Individual transcriptional factor binding specificities are modeled with HT-SELEX data sets [5]. The data comprise large amounts of sequenced DNA reads, which are known to comprise a bound protein. However, the exact position of the transcriptional factor on the DNA ligand and the length of the motif are unknown. [5] In addition, consecutive affinity purification (CAP) SELEX can be utilized for measuring binding specificities of transcriptional factor pairs. The method has a couple of adaptations to HT-SELEX that enable the selection of DNA ligands with both TFs bound to them. [4] Furthermore, CAP-SELEX data sets are utilized for modeling TF pair specifities in this Master's thesis [4].

Protein-DNA binding specifities can be modeled in different ways. Most often specificities are modeled with position weight matrices (PWM) that describe the probability of detecting a nucleic acid at certain position on the binding site. Position weight matrix modeling, although simple and intuitive, have inherent limitations. [8] Position weight matrices assume independence between nucleic acids, while the nucleic acids on a motif are often correlated [9]. Furthermore, more complex models may be better suited for modeling binding specifities. For instance, DeepBind, that utilize convolutional neural networks outperformed PWM models [10]. In this Master's thesis both HT-SELEX and CAP-SELEX data is utilized for modeling TF binding specifities with random forest. Models were trained on SELEX sequences and model performance was assessed by scoring unseen DNA ligands. Since the data sets only comprise sequences with proteins bound to them, background sequences for the classification task were constructed by shuffling the SELEX sequences. Shuffling was conducted such that dinucleotide counts, which are known occur hierarchically on genome, were preserved since negative sequences should resemble putative although negative binding sites [11, 10].

Random forest is an ensemble of multiple decision trees learned on randomly chosen subsets of the training data. Furthermore, a decision tree is a recursive partitioning of the data according to local models. The partitions thus form a tree structure where local models split the data at nodes and branches indicate the local model considered next. Finally, tree leaves define the class of the instances falling to each particular leaf. Aggregating multiple decision trees as in random forest increase modeling accuracy since decision trees are quite unstable models due to hierarchical tree growing process. [12] Random forest has an additional layer of randomness in comparison to other aggregation methods. At data partitions in decision tree nodes a random set of variables is chosen to be considered for the split. [13] Random forest is well suited for modeling high-dimensional and often correlated genetic data, since decision trees are able to select to a class entire data subsections with correlated variables [14, 15]. Furthermore, decision trees and random forest are able to utilize both numerical and categorical features [13].

It was discovered that random forest performed almost equally to DeepBind for modeling individual TF binding specificities. Thus, random forest outperformed scoring sequences with position weight matrices. Different random forest models were implemented and their performance was assessed. Since DNA ligands are double stranded, only one of the strands or even a shorter subsequence should be chosen for random forest training. The best model was obtained by combining PWM modeling with random forest. Thus, the most probable binding sites were searched with position weight matrices and either these sites or the entire DNA strands where the sites were located, were utilized for training the forests, depending on the properties of the SELEX experiment at hand. Furthermore, TF pair binding specificities were modeled with random forest. Different random forest model variants were implemented. Again the model with highest predictive accuracy was obtained by combining PWM and random forest modeling. Random forest outperformed significantly modeling with only PWMs. Furthermore, the difference was greater for pairs than it was for individual transcriptional factors. The increase in model complexity and flexibility due to random forest may be more significant for TF pairs, because of higher variation in binding specificities induced by alteration in TF pair spacings. DeepBind does not currently provide models for TF pairs. Thus, random forest is comparable to DeepBind and outperforms PWM models. Especially for TF pairs, models that provide higher flexibility are beneficial, and should be utilized for modeling motifs instead of position weight matrices.

2 Transcriptional regulation

Genes are expressed differently in differentiated cell types, which influences protein composition and determines the function of the cell [7]. Furthermore, regulation of gene expression is a an integral part of evolution and progression of diseases [16]. Initiation of transcription, the process of reading a gene into a messenger sequence, is to a large extent controlled by several associated proteins called transcriptional factors (TF) [3]. Transcriptional factors bind DNA in a sequence specific manner and has the ability to alter gene expression. Thus, understanding the binding specificities of transcriptional factors provide insight for unraveling the gene regulatory networks. A major goal in understanding gene regulatory networks is to be able to determine which sites at a genomic sequence are occupied by certain transcriptional factors. DNA binding specificities of proteins can be measured with inside the cell, in vivo, and outside the cell, in vitro, techniques. Binding profiles from in vitro measurements can be used for finding putative binding sites and target genes in the genome while in vivo methods measure binding specificities at living organisms at certain conditions. [7] In this Master's thesis binding specificities of human TFs measured with in vitro experiments using sequencing techniques are modeled. In this chapter, the significance of gene regulatory networks and transcriptional factor binding specificities are discussed. The experimental in vitro methods for which the models are constructed, are introduced and currently used modeling techniques for the quantitative data are represented.

2.1 Gene expression

The deoxyribonucleic acid (DNA) sequences that code for proteins are called genes. DNA comprises four bases, adenine (A), cytosine (C), guanine (G) and thymine (T), which are covalently linked into two nucleic acid chains. Hydrogen bonds between the two chains, or strands, hold them together in a DNA double helix. Furthermore, the hydrogen bonds form energetically favorably between adenine and thymine bases in addition to guanine and cytosine bases. Thus, genetic information is stored into complementary DNA double helix. Genes are read out into proteins through transcription and translation. Transcription produces a single-stranded ribonucleic acid (RNA) molecule complementary to the gene, which is finally translated into a protein at a ribosome in cytocol. Gene expression determines which genes are transcribed and thus, the protein content of the cell. Transcription of protein coding genes in eucaryotic cells is performed by RNA polymerase II with the help of additional proteins. General transcriptional factors aid in unwinding the double stranded DNA and positioning the polymerase on a promoter sequence. Furthermore, the assembly of general transcription factors and RNA polymerase begins with the binding of TFIID protein to the DNA sequence element upstream the transcription start site, which for many genes is the TATA box. The binding of TFIID leads to a significant change in DNA shape of the TATA sequence or an other element, which serves as a signal for other proteins. After the assembly of polymerase and additional proteins, the polymerase is released from the promoter for transcription and it continues to elongate messenger RNA sequence until it reaches a DNA terminator sequence. However, since DNA is packed into nucleosomes and other chromatin structures, the presence of specific transcriptional factors is needed for transcription initiation in vivo. These transcriptional factors bind to specific DNA regions and aid at attracting RNA polymerase II and general transcription factors to the promoter, which is packed in chromatin. In addition, a protein complex called mediator is needed for proper interactions between transcriptional factors and RNA polymerase. A cell can change its protein composition by controlling transcription rates of different genes. Even though regulation of gene expression can occur in each step of the protein synthesis, initiation of transcription is the most important point in gene expression control. Furthermore, gene expression is largely regulated by TF binding to DNA regulatory sequences. [1]

Transcriptional factors can be divided into general transcriptional factors, transcriptional activators and transcriptional repressors. All transcription factors include a DNA binding domain (DBD) while activators and repressors also contain a protein binding domain (PBD). Transcriptional activators function through positive control by increasing the probability for initiation of transcription of the target gene. Activator proteins increase the probability for transcription by promoting assembly of RNA polymerase and general transcriptional factors on the promoter. The activator binds to an enhancer sequence that can be far away from the promoter sequence with the DNA binding domain, and the DNA between enhancer and promoter loops out in order for the transcriptional activator to interact with other proteins at the promoter. The PBD of a transcriptional activator is also called activation domain. Transcriptional activators once bound to DNA may act in different ways. Often the mediator protein has already assembled general TFs and the RNA polymerase, and the activator proteins help this complex to bind the promoter sequence. However, sometimes the complex is missing some of the required general transcriptional factors. Activator proteins may help these proteins to assemble on the promoter. Transcription initiation is represented in Figure 1. In addition, some activators act by facilitating the stepwise assembly of general TFs on the promoters. [1]

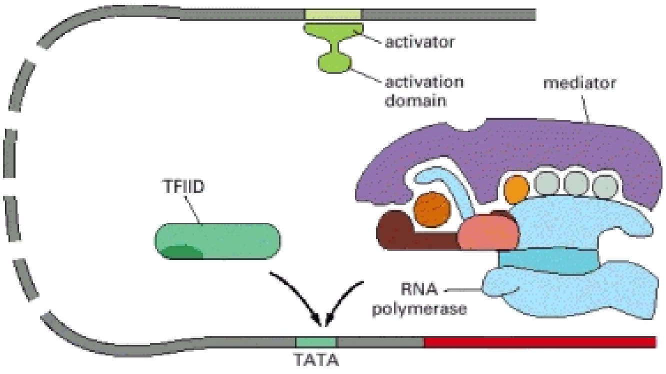


Figure 1: Transcription begins by binding of TFIID leading to assembly of transcription complex on the promoter aided by activators. Modified from [1].

Furthermore, transcriptional activators increase transcription of target genes by modifying DNA chromatin structure locally. Many transcriptional activators recruit histone acetyl transferases and ATP-dependent chromatin remodeling complexes, which in turn act by covalently modifying histones and remodeling nucleosomes. The alterations in the chromatin structure make DNA underneath more accessible. [1]

Transcriptional repressors on the other hand act through negative control, thus their binding to DNA decrease the probability for transcription initiation. The repressors also work by directly affecting transcription initiation and by modifying chromatin structure in various ways. Mainly repressors act by inhibiting the function of transcriptional activators. The inhibition may be achieved through competitive binding to DNA, masking the protein binding domain of the activator or by occupying the binding site of the activator at the mediator. Furthermore, transcriptional repressors can alter chromatin structure to make the DNA sequence at hand less accessible. Chromatin can be changed back to the pre-transcriptional state by certain types of chromatin remodeling complexes, which are attracted to the site by transcriptional repressors. In addition, repressors may recruit histone deacetylases that make the chromatin less accessible and reduce the affinity of TFIID towards the promoter. [1]

2.2 Complex gene regulatory networks

Gene expression is to a large extent regulated by transcriptional factors that bind DNA regulatory sequences, which can be understood as a mechanism of the cell for turning genes on and off. However, regulation of gene expression is highly complex, because of the large number of transcriptional activators and repressors and the numerous enhancer sequences that may be located far away from the promoter sequence. In addition, RNA polymerase II requires often multiple general transcriptional factors to initiate transcription. Thus, there are multiple steps for regulating the rate of transcription. Furthermore, transcriptional factors can alter DNA packaging into chromatin, which serve as the third mechanism of turning genes on and off, making gene expression regulation even more complex. It has been estimated that about 5-10 % of genes transcribe for proteins that regulate gene expression, which describes the complexity of gene expression networks. The expression of genes is regulated by multiple transcriptional factors whose expression in turn is regulated by other transcriptional factors and so on. In addition, cells adapt to environmental changes by altering their gene expression. The activity of transcriptional factors may be changed for example through protein synthesis, ligand binding, protein phosphorylation and unmasking the active site of a transcriptional factor. Each gene is regulated differently in a cell through synergistically functioning transcriptional factors. The joint effect of multiple transcriptional factors working together is the product of the effect of each transcriptional factor. Thus, the effect of multiple transcriptional activators on the transcription rate is much higher than the effect of one activator. [1]

A gene regulatory network denote the interactions between transcriptional factors and DNA regulatory sequences at a given time in various cellular contexts. In addition to the direct relationship between transcriptional factors and a target gene, a few studies have indicated that sometimes transcriptional factor binding do not correlate with target gene expression. Furthermore, it is not known whether this kind of binding is functional or not. The binding might be random or possibly conveying gene expression regulation at a distance for example by altering chromatin structure or looping chromatin. Thus, transcriptional factors binding to DNA may occur near the target gene in order to directly regulate the expression of the gene, or in a genome-wide manner regulating the chromatin structure. [2] For example in vivo experimental methods such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) may be used to determine TF binding sites in a genome and genes whose expression they are likely to regulate [3]. Thus, for transcriptional factors that regulate the expression of specific target genes, it could be possible to deduce their function from revealed binding sites. This might not be possible for transcriptional factors that bind to regulate chromatin structure and thus regulate gene expression more widely. However, the ultimate goal in studying gene regulatory networks would be to be able to directly infer them from the knowledge of the binding sites in the genome of a cell. [2] Different environments, cell types and many transcriptional factors make studying the gene regulatory networks complicated. Thus, the knowledge of all putative binding sites in a genome would be beneficial for understanding the function of transcriptional factors. [5]

Transcriptional factors may also bind DNA as complexes of multiple proteins, which increases the complexity of gene regulatory networks. In some cases individual gene regulatory proteins are not able to bind DNA on their own but as a dimer have the required affinity for binding specific DNA sequences. Furthermore, other transcriptional factors containing activator or repressor domains may assemble on the dimer to alter gene expression. Often, protein interactions are strong enough to occur only at their DNA binding site, which makes the DNA sequence a seed for assembly of the proteins. The fact that individual transcriptional factors may have different roles in different protein complexes, increases the complexity of gene regulatory networks. [1] Thus, many transcriptional factors bind DNA as homo- and heterodimers. A pair can possibly bind to multiple different DNA motif, if the spacing and orientation between the two proteins differ [17, 4]. Often, the transcriptional factor pairs form between the same structural family. However, pairing between different families have been identified. [4] Therefore, when gene expression regulation networks are studied it is important to also consider and model TF pair binding profiles.

2.3 Transcription factor binding specificities

As discussed the knowledge of TF binding sites can be used to infer gene regulatory networks. Transcriptional factors bind regulatory DNA sequences in a sequence-specific manner. The affinity of a protein for each potential DNA sequence can be determined separately. However, for modeling gene expression networks, the information that is needed is not the affinity of each potential binding site, but rather the affinity difference between probable binding sites at regulatory sequences and non-binding sites. Since the amount of genome is so high inside the nucleus, transcriptional factors will be constantly bound to DNA even if high affinity sites are not available. Thus, for transcriptional factors the ability to separate sequences at regulatory sites from other genomic regions is crucial. The difference in affinity of all the potential binding sites of a transcriptional factor is referred to as specificity, whose representation is called a motif. Specificity refers to how well a transcriptional factor distinguishes between different DNA sequences. [3]

Transcriptional factors recognize the functional binding sites according to base and shape information of the DNA. The interactions between the DNA sequence and transcriptional factor occur between the side chains of the protein aminoacids and accessible edges of the DNA bases. Readout of the transcriptional factor according to DNA base sequence is called base readout. Possible bonds include for example hydrogen bonds, hydrophobic interactions and water-mediated hydrogen bonds. In addition, transcriptional factors recognize shape features of the DNA including unwinding and bending. These features are dependent on DNA base sequence, thus they are also referred to as indirect readout. The DNA structure can be divided into major and minor groove, to which transcriptional factors bind slightly differently. Most transcriptional factors utilize both base and shape readouts when binding DNA. Figure 2 represent possible base interactions in major and minor groove in addition to the interplay of shape and base readout usage. [18]

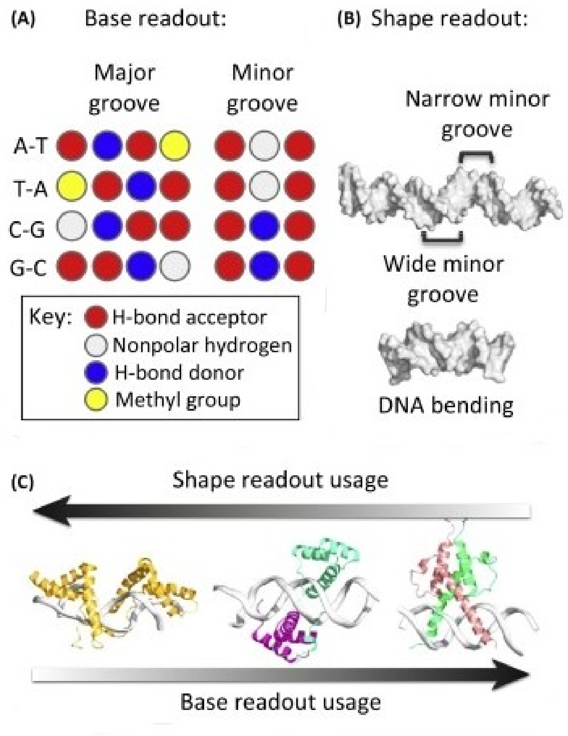


Figure 2: TF recognize binding sites through base- and shape readouts. Base readout describe the functional group that each base denote for protein binding while shape readout describe the shape features that affect binding. Base readout differs between major and minor grooves. Modified from [18].

In order for the transcriptional networks to function properly, the transcriptional factors need to separate functional binding sites from all putative binding sites. The specificity of a protein may be characterized differently depending on the experimental method used. Most often enrichment scores, which are based on rank median affinities, or position weight matrices, describing the probability for each base to occur in each position, are utilized. [3]

Transcriptional factor dimers bind DNA according to similar principles than single transcriptional factors. However, protein pairs may have multiple different binding specificities, since there might be fluctuation in the amount of gaps between the motifs for the two transcriptional factors. Furthermore, the orientation of the binding proteins may alter and binding may occur on opposite DNA strands. However, according to Jolma et al. most transcriptional factors exhibit one or two different binding orientations while gaps of one or two deoxyribonucleotides between the two motifs are most common. [4]

2.4 In vitro binding site determination

Currently, the DNA binding preferences are known only for a small fraction of human transcriptional factors. In vivo techniques can be utilized for discovering TF binding sites in a cell. However, in vivo measurement techniques do not reveal all potential TF binding sites. Developing models that describe transcriptional factor binding based on biochemical principles would be beneficial for understanding gene expression mechanisms. In addition, these models could possibly predict the effects of mutations, thus aid at understanding disease susceptibility. In vitro high-throughput techniques can measure protein-DNA binding specificities outside the cell, which makes the discovery of all putative binding sites possible, resulting in better sequence specificity models. [5] Currently, the most popular in vitro measurement techniques for protein binding affinities include protein binding microarray (PBM) and high-throughput SELEX. In PBM experiments, double stranded DNA probes are attached on a glass surface, which is washed with a solution containing epitope tagged protein and an antibody solution with fluorophore labeled antibody. Thus, binding intensities can be determined by measuring fluorescence intensity. [19] However, with PBM the amount of probes that can be placed on the array is restricted and the position of DNA probes on the array may cause disturbances. In addition, the studied proteins need to be purified, which makes it difficult to study transcriptional factors that need post-transcriptional modifications in order to function properly. The high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) is based on massively parallel sequencing and it has been utilized for studying large amounts of TF binding affinities in parallel. [7] Furthermore, the method was extended to measure DNA binding specificities of transcriptional factor pairs. PBM can also be modified to measure transcriptional factor pair binding affinities. However, CAP-SELEX enable measurements with many TF pairs in parallel and restrictions on DNA pool size are looser. [4] HT-SELEX and CAP-SELEX data is modeled in this Master's thesis. Thus, these methods are discussed in this chapter more deeply.

2.4.1 High-throughput SELEX

Evolution of ligands by exponential enrichment (SELEX) is a screening method that can be utilized for selection of specific DNA sequences that bind a TF. SELEX comprises repeated cycles of partition and amplification from a large nucleic acid sequence library. A random nucleotide pool is incubated with the protein and the bound sequences are separated from the non-bound sequences. Furthermore, bound sequences go through amplification for the next selection cycle. [6]

Jolma et al. have measured binding specificities for human transcriptional factors with high-throughput SELEX (HT-SELEX). The experiments begun by manufacturing the transcriptional factors or the DNA binding domains (DBDs). The DNA sequences encoding the proteins were cloned into a Gateway recombination cloning entry vector. Sequences were retained with streptaviding binding peptide (SBP) tagged Gaussia luciferase enzyme. Furthermore, the expression of the proteins was measured with a luciferase assay in primate cells, where the proteins had been transported. In addition, a library for double stranded DNA sequences, called ligands, was constructed. The DNA sequences were designed so that all possible nucleotide acid sequences were present. Furthermore, the sequences were attached to a barcode sequence, which indicate the identity of the ligand, in addition to a primer sequence, which is needed in PCR amplification. The barcode identification enables measurements of multiple TF binding specificities in parallel. After the TFs and DNA ligands are prepared, the HT-SELEX can begin. First the proteins and DNA ligands are mixed together in order for binding to take place. Then, the sample is washed and the TF bound DNA sequences are separated from the mixture. These DNA sequences are PCR amplified and utilized for the next HT-SELEX cycle. The cycle is repeated multiple times in order to verify that only the highest affinity sites remain in the DNA sequence pool. Figure 3 represents the HT-SELEX process. [7]

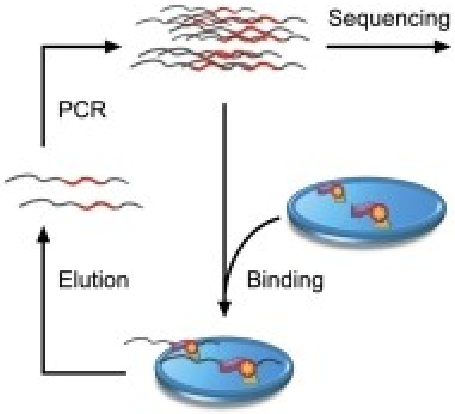


Figure 3: HT-SELEX with repeated cycles of elution, PCR amplification and sequencing. [7]

The DNA sequences are sequenced with massively parallel sequencing after each HT-SELEX cycle. Possible error sources in the HT-SELEX process include transcriptional factors having binding affinities towards primer sequences and having too small DNA sequence pools. Thus, a computational quality control pipeline can be utilized for detecting successful HT-SELEX cycles. Failed experiments could be detected by discovering low quantities of highly enriched subsequences. In addition, the positions of the TF binding sites at the ligands was assessed. Since real binding sites should be distributed quite evenly across the ligands, binding to barcodes or flanking constant regions can be detected by discovering binding events at the same sites in multiple ligands. Furthermore, the quality of individual cycles of HT-SELEX experiments could be assessed by assuring that the high affinity sequences are enriched exponentially. If enrichment is not observed at a SELEX cycle for some of the high affinity subsequences, the cycle is likely to be contaminated. [7]

Binding models can be constructed directly from the HT-SELEX data. Position weight matrices may be generated by counting the base occurrences in the most enriched 5-12 bases long subsequences and all the subsequences that differ from these enriched subsequences by one base. Jolma et al. compared HT-SELEX results for a specific mouse DNA binding domain to a similar experiment conducted with PBM. It was discovered that HT-SELEX yielded highly similar binding profiles than PBM. In addition, in vivo Chip-seq experiments with three different transcriptional factors gave similar profiles than found with HT-SELEX. High-throughput SELEX may be more suitable than PBM when DNA binding specificities of multiple transcriptional factors need to be assessed. [7]

2.4.2 Consecutive affinity-purification SELEX

A transcriptional factor pair can bind multiple different motifs due to different orientations and spacings between the two binding sites. Consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX) have been proposed for measuring TF pair binding specificities. The CAP-SELEX method is very similar to HT-SELEX. However, a few adjustments have been implemented in order to enable measurements of transcriptional factor pair binding specificities. Furthermore, CAP-SELEX is able to measure simultaneously DNA mediated TF pair binding events and binding of already dimerized TFs. [4]

In Jolma et al. CAP-SELEX experiments, Gateway recipient vectors were constructed with two different types of tags for the two different proteins. The first transcriptional factor was tagged with SBP and the second with 3 x Flag. Both types of proteins were expressed and purified from E. coli cells. Similarly as with HT-SELEX experimental procedure, double-stranded DNA ligands were barcoded. DNA ligands and both types of TFs were incubated together in a buffer that resemble the conditions of a cell. Furthermore, in order to sequence only those DNA ligands that have both transcriptional factors bound to them, two washing steps were introduced. Separation was performed through consecutive affinity purification first by the SBP tag and then by the 3 x Flag tag. The ligands, which had the protein dimer bound to them, were then washed and PCR amplified. These sequences were utilized for the next CAP-SELEX cycle. In CAP-SELEX procedure the cycle is repeated three times. [4] Figure 4 represent CAP-SELEX experimental procedure.

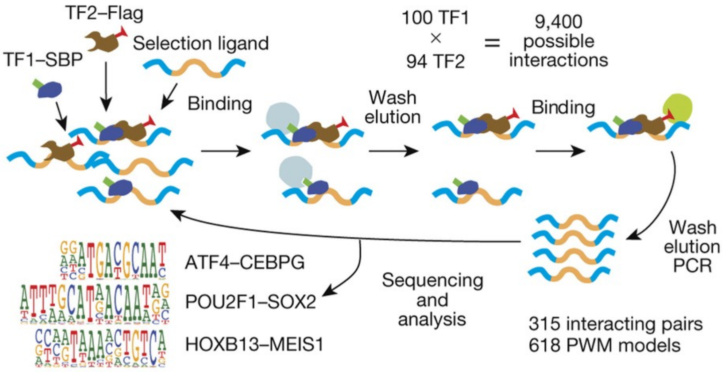


Figure 4: CAP-SELEX for revealing TF-TF-DNA interactions. [4]

Similarly as in HT-SELEX the binding specificities of the transcriptional factor dimers could be revealed by examining the most enriched sequences. However, the binding site could be identified as one combinatorial binding specificity or by two separate binding specificities of the two proteins. Furthermore, the gapped or ungapped enriched subsequences could be found with a de novo motif discovery algorithm, Autoseed. However, when the two proteins were spaced further apart on the ligands, a slightly different approach needed to be utilized. In these cases 6-mer representative sequences were defined for both TFs and for each experiment only those ligands were chosen for further examination that included both of the representative sequences. Furthermore, the spacing and orientation preferences were counted from the chosen sequences. [4] Thus, PWMs can be constructed also from CAP-SELEX data for protein dimers. The next section discusses position weight matrices more profoundly.

2.5 Modeling transcriptional factor binding sites

Transcriptional factor binding is modeled with binding specificities, which aim at describing how the protein differentiate between binding and non-binding sites. The most common and simplest model for specificities is the position weight matrix (PWM) that is constructed slightly differently for different types of experimental data. [3] The basic principles of position weight matrices is discussed in this chapter. In addition, a deep learning model, DeepBind, for predicting sequence specificities is introduced. The model constructs a binding motif matrix similar to PWM. Furthermore, it has been utilized for HT-SELEX data. [10] Random forest modeling implemented in this Master's thesis is compared to the performance of PWM models and DeepBind. In addition, random forest is combined with PWM by utilizing the most probable binding sites search from the SELEX ligands with PWMs for training the model.

2.5.1 Position weight matrices

Position weight matrices (PWM) may be utilized for scoring DNA sequences for being binding sites. PWM is defined as a matrix $W(b,i)$ where b refers to the base b={A, C, G, T} in position i of an L bases long binding site sequence. Each element in a PWM give a specificity score for a base at each position of the binding site. Thus, a PWM may be utilized for scoring candidate sequences of length L by summing the elements of the matrix that correspond to the sequence. [20] If the sequence S is represented as a binary matrix $S(b,i)$ containing the knowledge of which bases are present in the sequence, an additive score can be computed for a candidate sequence as [20],

$$Score(S|W) = \sum_{b,i} W(b,i)S(b,i). \tag{1}$$

SELEX data is qualitative in nature. The data comprises DNA reads, which contain TF binding sites. However, only a subsequence of each read might correspond to the binding site, which slightly complicates the PWM construction. [7] In general, probabilistic modeling may be utilized for constructing position weight matrices from known binding site sequences. First a position frequency matrix (PFM) is generated by determining the probability for each base. [20] Thus, position frequency matrix, F, can be constructed from N aligned sequences S_j as [20],

$$F(b,i) = \frac{1}{N}\sum_{j=1}^{N} S_j(b,i). \tag{2}$$

For instance Table 1 represent a PFM for Barhl1.

Table 1: PFM for Barhl1 [5]

Base/Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0.25	0.22	0.09	0.87	0.99	0.66	0.04	0.14	0.23	0.16
C	0.28	0.36	0.01	0.00	0.00	0.03	0.60	0.04	0.24	0.26
G	0.31	0.14	0.00	0.00	0.01	0.11	0.03	0.71	0.34	0.11
T	0.16	0.28	0.90	0.13	0.00	0.20	0.33	0.11	0.19	0.47

Alternatively to equation 1, PFM can be utilized for computing sequence scores for sequence S, through multiplication [20],

$$Pr(S|F) = Score(S|F) = \prod_{b,i} F(b,i)^{S(b,i)}. \tag{3}$$

The PFM can be transformed to PWM by taking logarithm of the matrix values [20]. In addition, information content (IC), which can be utilized for motif visualization, is computed from a PFM as [20],

$$IC(i) = \sum_{b} F(b,i)\log_2\left(\frac{F(b,i)}{0.25}\right). \tag{4}$$

For instance Barhl1 PFM is represented in Figure 5.

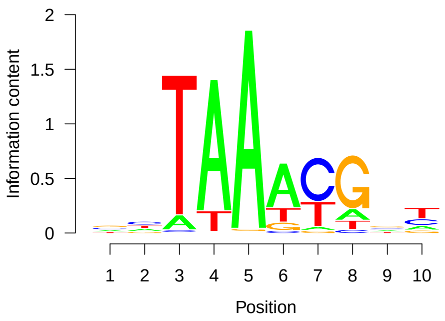


Figure 5: Barhl1 PWM from [5] and visualized using IC.

The binding site position on a DNA read in SELEX data is uncertain [7]. Thus, a motif discovery algorithm is utilized for finding the binding sites before constructing the PWM. Motif discovery algorithms are based on alignment of the DNA reads and finding the most enriched subsequences. [20] The most commonly used methods for motif discovery include the expectation maximization (EM) algorithm and Gibbs sampling [21].

2.5.2 Binding motif with deep learning

Although position weight matrices are functional and easy to use, recent advancements in the field suggest that more complex models capture sequence specificities more accurately. A deep learning algorithm, called DeepBind, has been proposed for constructing binding motifs from different types of experimental data including HT-SELEX sequencing data. The algorithm is based on convolutional neural networks. Thus, it can be utilized for HT-SELEX data even though the transcriptional factor binding site on the reads is unknown. The binding motif learned with DeepBind is similar in structure than the position weigh matrix. Therefore, the motif is easy to interpret and can be visualized similarly to PWM as represented in Figure 5. However, the motif values do not necessarily need to be probabilistic. [10]

Training with DeepBind is performed in four steps: convolution, rectification, pooling and learning with neural networks. In convolution step the sequenced reads, S, are scanned with a set of binding motifs, k, yielding scores for each subsequence, i, similarly as with PWMs with the exception of motif scores not having to be positive or summing up to one. Thus, the convolution step performs the same computation than the PWM in equation 1 so that $W_k(b, i)$ are the k DeepBind motif matrices. Next, in the rectification step the convolution scores are modified such that tunable thresholds, $t(k)$, are reduced from the scores yielding a rectified motif matrix $Y(k, i)$,

$$Y_k(i) = max(0, Score(S|W_k) - t(k)). \tag{5}$$

Furthermore, the scores are set to zero if they have negative values. Thus, the thresholds can be understood as activation thresholds setting those values to zero that are not greater than the thresholds. Pooling is performed by maximizing and averaging the rectified responses of each motif detector and subsequence. Maximizing gives information about longer motifs and averaging about the presence of multiple shorter motifs. For HT-SELEX experiments utilizing only the maximization step performed well enough. Thus, pooling was conducted with,

$$z_k = max(Y_k(1), ..., Y_k(n)), \tag{6}$$

where n is the total number of subsequences $i = 1...n$. These features are given to the neural network, which transforms them to output scores. The neural network is a vector of tunable weights and the output score is a linear combination of weights and features. The output score is 0 or 1 indicating a binding or non-binding site. Thus, output scores, p, are computed as,

$$p = w_{d+1} + \sum_{k=1}^{d} w_k z_k, \tag{7}$$

where d is the total number of motifs, w_k are the neural network weights and w_{d+1} is an additive bias term. Since the true target values are known in training, the prediction errors are utilized for tuning motif detectors, thresholds and weights. Figure 6 represent the four stages and work flow of the DeepBind algorithm. [10]

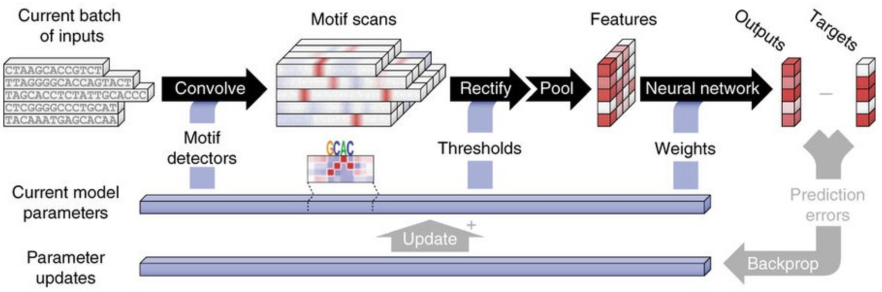


Figure 6: DeepBind comprise convolution, rectification, pooling and neural network. Motifs W_k, thresholds $t(k)$ and weights w_k are tuned in training. [10]

New sequences can be scored as binding or non-binding sites with DeepBind by utilizing the learned motif detectors, thresholds and weights for test sequences. [10]

3 Materials

Protein and DNA binding specificities are modeled in this Master's thesis. Binding specificities for individual human transcriptional factors are determined from HT-SELEX DNA sequences acquired by Jolma et al. [5]. Position weight matrices have been reconstructed from the data previously but adding complexity with random forest might yield a more accurate description of the binding specificity. In addition, Jolma et al. utilized CAP-SELEX method for finding position weight matrices for transcriptional factor pairs [4]. Similarly, these binding specificities can be modeled with higher complexity models. In this chapter, the HT-SELEX and CAP-SELEX data utilized in this thesis is represented.

3.1 HT-SELEX data

In [5], Jolma et al. measured binding specificities of human and mouse transcription factors with HT-SELEX method. Binding specificities of full length transcriptional factors and their DNA binding domains (DBD) were examined. Out of the 891 human DBDs and 444 mouse DBDs, 303 human DBDs and 84 mouse DBDs expressed binding affinity with enriched specific sequences. In addition, 151 full length human transcriptional factors were discovered to have binding specificities for specific subsequences, when 984 full length proteins were studied. It was discovered that especially proteins from high mobility group (HMG) and C2H2 zinc finger group did not show distinct binding specificities. This is in line with the knowledge of these proteins expressing unspecific binding. [5] The 538 data sets with significant binding specificities have been utilized also for assessing DeepBind model performance [10]. In this work only a subset of the transcriptional factors that have been found by Jolma et al. to bind to specific DNA sequences are modeled. The different ways of choosing features for the random forest is examined in a case study with Alx1 transcriptional factor. In addition, the best experimental schemes are studied with a data set comprising of 15 HT-SELEX experiments. Furthermore, the modeling schemes that perform the best on unseen testing data are utilized for a larger data set and compared to accuracies with DeepBind and Jolma et al. position weight matrices.

The HT-SELEX data for transcriptional factors is stored in European Nucleotide Archive (ENA) to Fastq file format. ENA is a database for high-throughput sequencing information and Fastq is a common file format for sharing sequences [22]. Fastq files contain the sequenced reads in addition to the per base quality scores that describe the estimated probability for error [22]. However, in this work only the sequenced reads are considered. The Fastq files are named so that first is the name of the protein, then the barcode that also indicates the length of the sequenced reads, then the name of the experimental batch and finally the experimental cycle.

Jolma et al. utilized a computational pipeline Inimotif to discover which experiments contained significantly enriched subsequences. The experiment quality was assessed by considering that the most enriched subsequences should be related to each other, the binding should occur evenly at all positions on the forward and reverse strands and that the high affinity sequences should be enriched exponentially against the experimental cycles as represented in Figure 7. Contrary, random subsequences do not express exponential enrichment against selection cycle. [7]

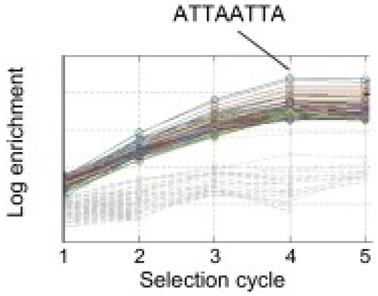
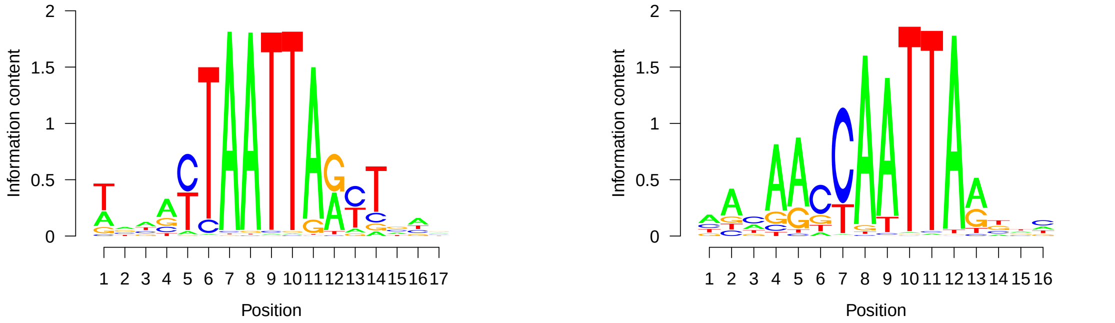


Figure 7: Enrichment increases exponentially against selection cycle for enriched subsequences contrary to random subsequences. [7]

Quality control was performed by computing the incidences of all 5 to 11 nucleic acids long subsequences and determining the most enriched one for each length. The Hamming distances were then computed for other subsequences. Hamming distance is the amount of base substitutions required to transform the subsequence into the most enriched one or its reverse complement. Thus, it could be inspected whether the experiment yielded sufficiently high incidence of the most enriched subsequence and that there were other subsequences closely related to it. Furthermore, the position of the most enriched sequences on the forward and reverse strands were computed in order to assure that binding does not occur on barcode sequences. The sequences in the ENA fastq-files does not contain the flanking regions including the barcode. Finally, the exponential enrichment of subsequences from cycle to cycle was assessed by calculating the incidence of 100 most enriched subsequences and 200 random subsequences in each cycle and assuring that the same subsequences are enriched in all cycles. [7] The experimental cycles studied in this Master's thesis are the same that have been selected by Jolma et al. for PWM construction and by Alipanahi et al. for comparison between DeepBind and the PWMs on unseen DNA sequences [5, 10]. However, a subset of 55 transcriptional factors were chosen randomly for random forest modeling.

Finally, position weight matrices could be derived from successful experiments. The most enriched subsequences for different lengths were identified and incidences were utilized for PWM construction. In addition, incidences for all subsequences that differed from the most enriched subsequence by at most one base were identified. The position weight matrices were computed from these subsequences as described in the previous chapter. Similar analysis was conducted on earlier cycle, which was utilized for background correction. The optimal position weight matrix was then selected for each protein by choosing the one with minimum length that still included all the highly specific positions. [7] The PWM positions were identified as sufficiently specific if the ratio between the most and least frequent base was greater than one [5]. Position weight matrices for two example proteins are represented such that the information content of each base is plotted in Figure 8.



(a) Alx3 PWM from [23] and visualized using information content.

(b) Barhl1 PWM from [23] and visualized using information content.

Figure 8: Position weight matrices for Alx3 and Barhl1.

3.2 CAP-SELEX data

Random forest modeling is utilized in this work also for transcriptional factor pair binding sites. Jolma et al. measured 9400 putative transcriptional factor pair binding specificities. Hundred proteins were marked with SBP and 94 proteins were flag-tagged. It was discovered that 315 TF pairs expressed significant DNA binding affinities. Furthermore, 162 protein pairs preferred only one site while 153 displayed many binding specificities. Most transcriptional factor pairs recognized DNA sequences with one negative spacing between the motifs of the two transcriptional factors. In addition, the transcriptional factors that bound DNA with positive spacing between the individual motifs had a more relaxed binding specificity, thus expressing multiple gap configurations. Furthermore, some transcriptional factors expressed both negative spacing binding and the more relaxed positive spacing binding tactics. However, the differences in the motif distances were mostly small. Out of the transcriptional factor pairs that expressed relaxed binding preferences 73 % showed binding preferences with only one base pair gap difference between the two configurations. Furthermore, there are four possible orientations of the proteins binding to the DNA ligand, since binding can occur on both strands. However, when multiple orientations were observed, two orientations was the most common case. Although, some proteins have palindromic binding sites, which refers to the sequence on reverse and forward strands of the binding site being similar, making the number of observed orientations likely an underestimate. Thus, binding preferences of transcriptional factor pairs are highly sensitive to orientation and distance between the proteins. Figure 9 represent the number of observed gaps between the individual motifs and the amount of orientations observed in all the CAP-SELEX assays. Furthermore, for many transcriptional factor pairs the two individual binding motifs were not distinguishable as expected. In these cases the observed binding sites differed from what would have been expected from individual PWMs and this occurred mostly with pairs that had overlap between their motifs. Thus, transcriptional factor pair binding is highly complex and to a large extent a DNA mediated process. [4]

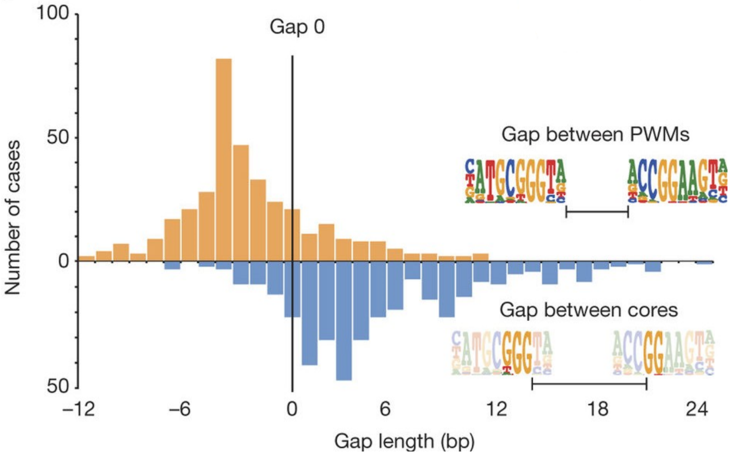


Figure 9: Gaps between TF motifs in CAP-SELEX experiments. Modified from [4].

Furthermore, PWMs could be recovered from each CAP-SELEX assay that showed enrichment of subsequences indicating significant transcriptional factor pair binding. Again, the sequenced reads were stored to European Nucleotide Archive (ENA) in Fastq file format. The files are named similarly as with HT-SELEX experiments so that the name of the two transcriptional factors are annotated first, then the experimental cycle and the name of the experimental batch. Finally, the length of the sequenced reads is represented with the barcode sequences. Sequencing depth was chosen so that on average each experiment yielded 250 000 sequenced reads. The experiments went through the same Inimotif quality control pipeline than the HT-SELEX experiments. Thus, experiments, which express most enriched subsequences related to each other evenly on all positions on the forward and reverse strands, are chosen for further analysis. The generation of position weight matrices was performed with Autoseed, which considers gapped subsequences in addition to ungapped ones. Thus, it can be utilized for finding PWMs for protein pairs that have gapped binding motifs. [4] Autoseed algorithm utilizes Hudding distance, which is a distance measure between aligned DNA sequences. Each subsequence is aligned against all other subsequences and the amount of subsequences that are within one Hudding distance away from each other are counted. Aligned sequences that have a Hudding distance of one, have n-1 perfectly matching bases, where n is the number of defined bases in the longer ungapped subsequence. If the count of a subsequence is higher than the count of any other related subsequence it is utilized as a seed for generating the PWM. [24]

Furthermore, Jolma et al. analyzed TF pairs that bind further apart on the DNA reads. This was performed by determining representative 6 bases long subsequences for each transcriptional factor. The reads that contained both of the representative subsequences were then chosen for further analysis in each experiment. The orientations and gaps between the binding sites were counted and represented with the maximum gap count for each orientation. The gap configuration was seen as significant if the count of it and the two neighboring gaps was higher than 30 % of the total amount spacings for that orientation. Finally, all the counts were mean normalized and the orientation and gap combinations were selected as preferred binding specificities if their count was higher than 50 % of all the spacing counts. The enriched subsequences were also utilized for PWM generation as a seed. [4] The seed matches and all subsequences that differed from the seed by one base were identified and the PWM was computed out of these sequences [7]. Jolma et al. provide the PWMs for transcriptional factor pairs [4]. In this Master's thesis the same experimental cycles chosen by Jolma et al. for PWM construction are utilized. However, out of the 315 TF pairs expressing DNA binding preferences, 50 were randomly selected for modeling.

4 Methods

Transcription factor binding sites measured with HT-SELEX and CAP-SELEX are modeled with random forest. Random forest is an ensemble method based on learning multiple decision trees on randomly chosen subsets of the data [12]. Furthermore, decision trees are defined by dividing the data recursively so that the partitioning can be represented as a tree with leaves corresponding to the final classes [12], which is this Master's thesis are binding site (1) and non-binding site (0). Genomic data is highly correlated and usually high-dimensional, which makes decision trees, and thus random forests, a suitable method for their modeling [14]. In this chapter, the theory behind random forest classification is discussed. Model learning and testing schemes are represented in addition to the introduction of background set construction and the different experimental settings. Experiments are conducted by combining position weight matrices with random forest and by utilizing only random forests. Slightly different approaches are needed for modeling individual TF motifs and TF pair motifs.

4.1 Supervised learning

In machine learning a predictive model is found by optimizing a performance criterion on training data set, which may be artificially constructed or experimental. Thus, machine learning may be utilized in situations where it is not possible to construct a model that perfectly resembles the data or the acquired knowledge for such a model is not accessible. However, a good approximation of a model may be found through optimization. In addition, machine learning is related to artificial intelligence in a way that models can learn features of the data and adapt to changes. Machine learning methods can be divided to supervised and unsupervised methods. In supervised learning the training data is used for constructing the model while the predictive accuracy of the model is assessed on a test data set. The data in a supervised learning task is divided into features and labels. Thus, the goal is to construct a mapping from features to known labels in the training data set. The model is tested on the test set by learning the labels from test set features and comparing the predicted labels to true labels. The error between true and predicted labels in the test set with unseen data is called generalization error. With categorical labels the supervised learning task is called classification and with continuous labels the task is called regression. Unsupervised learning differs from supervised learning by the absence of known labels. In unsupervised learning the aim is to discover interesting patterns in data. A central task in supervised learning is to find an optimal model complexity that is able to model all the true trends in the data but does not overfit. Overfitting may be caused by learning too flexible models that are in fact modeling the noise in the data. Such models do not perform well on unseen data sets. Regularization is a common way of preventing overfitting. Regularization refers to adding a regularization term to the loss function of the model that will penalize higher complexity models. In machine learning the task is often to choose the best model complexity and parameters through optimization with training data.

Often there is not enough data to make reliable estimates of model performance on the test set. Cross-validation is a learning scheme where the training data is divided into K folds, $k \in \{1,..,K\}$, of about the same size. The model is trained separately on each k fold and tested on all the other folds, which is called validation data set. The average error in validation set is utilized for selecting the optimal model. [12]

The goal in this Master's thesis is to model protein-DNA binding specifities. Thus, the modeling conducted is binary classification. In binary classification there are only two possible classes. When modeling SELEX data the two possible classes are binding site (1) and non-binding site (0). Features may be chosen in multiple ways. However, the simplest way is to use the nucleic acids sequence of the reads as categorical features. In addition, k-mer frequencies and different shape features of the DNA ligands may be utilized as features for learning the model. The sequenced DNA reads in each experiment are divided into training set of 75 % of the entire data and a test set of 25 % of the data with balanced classes. Thus, there are equal number of positive and negative instances in the training set. Furthermore, 3-fold cross-validation is utilized in the training for searching the optimal forest. Random forest is learned with the training data and the model performance is evaluated in the test data set.

4.1.1 Bias-variance decomposition

Model predictions differ from the true class labels with an estimated prediction error. The error is composed of bias and variance. Bias refers to the average error between true values and predicted values. Variance on the other hand is the difference between predictions with the same model in different data sets. In practice high complexity of the model yields low bias, since the model is able to discover all crucial patterns in the data, and high variance as the noise in the data might be modeled as well. This situation is referred to as over-fitting, which causes predictions on unseen observations to be inaccurate. Although, the aim should be to reduce both bias and variance, it is not possible, since there is a trade-off between them. The bias-variance decomposition can be demonstrated with single input regression. In regression a true value y from a data set D can be mapped from the features $\mathbf{x}$ with a regression function $h(\mathbf{x})$. However, regression function is not known exactly. Thus, model prediction is mapped with a prediction function $y(\mathbf{x};D)$. Furthermore, decomposition between bias and variance can be represented by considering the expectation of squared error between the true value $h(\mathbf{x})$ and model prediction $y(\mathbf{x};D)$,

$$\mathbf{E}_D[\{y(\mathbf{x};D)-h(\mathbf{x})\}^2] = \{\mathbf{E}_D[y(\mathbf{x};D)]-h(\mathbf{x})\}^2+\mathbf{E}_D[\{y(\mathbf{x};D)-\mathbf{E}_D[y(\mathbf{x};D)]\}^2]. \quad (8)$$

In equation 8, the first term is squared bias and the latter term is variance. [25] A similar reconstruction can be found for binary classification [26]. In general, the bias-variance decomposition shown in equation 8, can be represented for multiple input data with noise. Thus, the reconstruction is,

$$expected\ loss = bias^2 + variance + noise. \quad (9)$$

The aim in machine learning is to minimize the expected loss. However, there is a trade-off between bias and variance, decreasing the other will increase the other. Thus, optimal model complexity maintains a balance between bias and variance. [25] The bias variance trade-off is an important concept in tree based modeling, because decision trees are known to have high variance, which can be reduced by using an ensemble of multiple decision trees trained on independent training sets [27].

4.2 Background set

HT-SELEX and CAP-SELEX experimental data contain sequenced reads of the transcription factor binding sites [7, 5, 4]. Since information about non-binding sites is missing, the background set for the classification task has to be constructed. The artificial background set should resemble the real differences between specific binding sites and non-binding sites rather than the differences between binding sites and other regions of DNA [10]. Furthermore, dinucleotide frequencies are known to appear on a hierarchical manner in DNA [11]. For instance CG-content differs between promoters and coding regions [10]. Therefore, the background set should contain the same dinucleotide frequencies than the sequenced reads in order to resemble putative but negative binding sites. A background set with preserved dinucleotide counts ensures that the model does not rely on low-level statistics such as the CG-content difference between coding regions and protein-binding regions [10].

A tool proposed by Jiang et al. for generating uniform random permutations of DNA sequences while preserving dinucleotide counts is utilized for background set construction. The tool is based on Euler's algorithm and uses Wilson's algorithm at forming directed spanning trees. The underlying idea in Euler's algorithm is to form a directed graph of the DNA sequence so that each nucleotide is represented as a vertex and dinucleotides are represented with directed edges. The resulting graph may contain multiple edges at the same direction. The shuffled sequence can then be formed by visiting each edge in the graph exactly once, corresponding to an Eulerian walk, and choosing randomly the output edge when leaving each vertex. [11] Furthermore, it has been proven that if the Eulerian walk starts and ends at the same vertices as in the original sequence, the dinucleotide count is preserved [28]. A Eulerian walk on the graph corresponds to an uniform random directed spanning tree, which is rooted at the last vertex of the graph. Furthermore, Wilson's algorithm is utilized for generating the spanning tree. Thus, the tree is formed by simulating random walks that begin from each unvisited vertex until the walk encounters the growing tree that initially contains only the root. As an encounter occurs all the nodes along that walk are joined to the tree. However, if a walk encounters a previously visited node, it is erased, ensuring the formation of an Eulerian walk. [11] The background set is constructed by shuffling each sequenced read once so that the data contains binding and non-binding sequences with equal amount as it is a common practice with binary classification to generate balanced classes [10].

4.3 Random forest classification

Random forest can be utilized for modeling TF binding specificities. In random forest multiple decision trees are learned on random subsets of the training data. A decision tree is a nonparametric method that utilize divide-and-conquer strategy. Thus, a distance measure is used to split the input space into local regions according to training data. [29] Classification trees and random forests are common methods for modeling biological data. This is in part, because features in decision tree modeling can be a mixture of categorical and numerical, continuous and discrete variables. Thus, for instance a classification tree can be constructed for DNA sequences with nucleic acids as categorical features in combination with numerical variables such as DNA shape features. [12] In addition, decision trees carry out variable selection and perform well with large and correlated data sets [12, 14]. High-throughput biological data sets are usually large. Both HT-SELEX and CAP-SELEX data sets govern hundreds of thousands of DNA sequences [5, 4]. In this chapter, the theory behind random forest classification and the feature selection possibilities are discussed.

4.3.1 Decision tree

Decision trees are hierarchical data structures that can be implemented for non-parametric modeling. More precisely decision tree is a recursive partitioning of the data according to local models, which are defined at each partitioning. The partitioning can be represented as a tree where the splitting of data occur at nodes and branches indicate which local model is considered next for that subsection of the data. Thus, leaves of a tree will define the class of the test set observations that have been localized to that particular leaf. Furthermore, random forest is an ensemble of decision trees, in this case classification trees. [12] Figure 10 represent the structure of an example decision tree such that features **x** are nucleic acids.

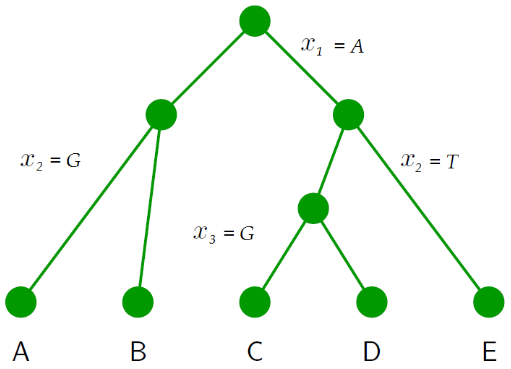


Figure 10: Decision tree partitions the data according to features $x_1, ..., x_3$ to leaves, which give probability of belonging to a class. Modified from [25].

Classification tree is an adaptive basis function model where each model define the region of partitioning the data. In the case of classification the distribution of classes in each node is stored. Furthermore, the distribution give probabilities for each observation to belonging to each class according to features considered at the path from root to the node. Classification of DNA binding sites of proteins is a binary classification task. In classification, the basis functions $f(\mathbf{x})$ that map instances according to features $\mathbf{x}$ to label y, can be represented as,

$$f(\mathbf{x}) = E[y|\mathbf{x}] = \sum_{m=1}^{M} w_m I[\mathbf{x} \in R_m], \tag{10}$$

where R_m is the m'th region referring to the leaf with probabilities w_m of observations belonging to class m. [12]

The decision tree is often grown with a greedy approach. A split function chooses the best feature to consider at each node and the best value for that feature. This is performed by minimizing the cost of the data partitioning. The class-conditional probabilities are first estimated after the data split in each leaf for each class c as,

$$\pi_c = \frac{1}{|D|} \sum_{i \in D} 1(y_i = c), \tag{11}$$

where D is the data in a leaf and y_i are the labels of the samples in that leaf. After the class probabilities are evaluated, the cost function is used to assess the goodness of the partitioning. There are multiple possible cost functions for classification. The cost function utilized in this work is the Gini index, which is the expected error rate for each class c out of all classes C,

$$\sum_{c=1}^{C} \pi_c(1-\pi_c) = \sum_{c=1}^{C} \pi_c - \sum_{c=1}^{C} \pi_c^2 = 1 - \sum_{c=1}^{C} \pi_c^2. \tag{12}$$

Other options for cost functions would have included misclassification rate and entropy. The Gini index and entropy are more sensitive to changes in class probability than misclassification rate. Algorithm 1 represent the recursive procedure of growing a decision tree. [12]

```
function grow_tree(node, D, depth);
node_prediction = class label distribution;
{ $D_L$, $D_R$ } = split(D);
if not worth splitting then
    return node;
end
else
    node_left = grow_tree(node, $D_L$, depth+1);
    node_right = grow_tree(node, $D_R$, depth+1);
    return node;
end
```

Algorithm 1: Growing a decision tree. If according to Gini index the split is worthy, the data D in a node is split to two D_L and D_R. [12]

Thus, the split function chooses the optimal division of the data at a node. Then the algorithm examines whether the data at the node is worth splitting. The split is not fulfilled if the desired tree depth is obtained, the cost of the split is too small or the node is already pure or sufficiently homogeneous. [12]

A tree can also be pruned in order to prevent overfitting. Pruning refers to quitting the tree growing once the decrease in impurity measure is not great enough to explain the increase in tree complexity due to splitting the data. However, this might produce too little data partitioning if individual features do not explain the splitting very well but the split could be justified by considering multiple features. Thus, pruning is usually performed after the tree is fully grown by pruning back according to cross-validated error on each subtree. The subtree within one standard error of the minimum is chosen as the pruned tree. [12]

4.3.2 Aggregation methods

Single trees do not usually perform as well as other machine learning models due to the greedy nature of growing a tree. In addition, decision trees are relatively unstable models, because the trees are grown hierarchically. [12] Thus, small changes in the training data might cause large changes in the resulting classifier [27]. In the tree growing process, the instability manifests itself by small errors in the first splits causing larger errors at the resulting tree. Therefore, decision trees are high variance estimators. [12] However, the variance can be reduced by learning multiple weaker classifiers such as decision trees and combining them into a classifier with more predictive power. These methods are referred to as aggregation methods or ensemble methods. The most commonly used aggregation methods include bagging, boosting and random subspace methods (RSM). [27] Ensemble methods reduce variance, because statistically choosing the wrong model becomes less probable when multiple models are combined assuming that the predictors are uncorrelated. In addition, the decision tree growing algorithms may get stuck to local optima. Thus, building many models on different data sets may provide a model that more accurately finds the true trends in the data. [30]

Bagging is a combination of bootstrapping and aggregation. The training data is randomly divided into subsections with replacement, which is referred to as bootstrapping, and a base classifier is constructed with every subsection of the data. Thus, the subsections of the training set are independent of each other. Finally, the classifiers are aggregated in order to give the resulting decision rule. The aggregation of base classifiers is usually performed with simple majority voting, where the most often predicted label from the classifiers is chosen as the final decision. However, other combination rules may be utilized as well such as the mean of posterior probabilities given by the base classifiers. Since bootstrap samples contain only part of the data, the possible data outliers are not present in all of the samples. Therefore, some base classifiers perform better than a classifier constructed with the entire training data would perform. Furthermore, the base classifiers with more predictive power give more extreme posterior probabilities for data observations. Thus, they will have more decisive power in the aggregation decision, which results in higher predictive performance. [27] Figure 11 represent the basic principle of bagging.

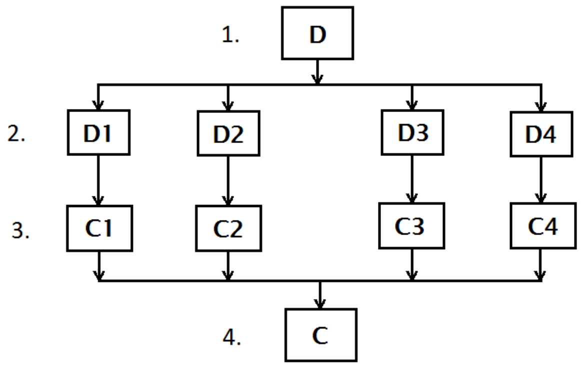


Figure 11: Bagging of weak classifiers. 1. Original training data 2. Training data randomly sampled to bootstrap samples. 3. Classifier trained on each set. 4. Ensemble classifier combining all classifiers.

Boosting on the other hand is a deterministic method for aggregating multiple base classifiers. The training set observations are assigned weights and classifiers are constructed sequentially. Initially all observations have same weights, which are altered according to the performance of the first classifier constructed on that data set. Observations that were classified incorrectly get higher weights for building the next classifier. The alterations in weights and sequential learning of classifiers is repeated until a sequence of different base classifiers is obtained. Final decision are made with simple majority voting or weighted majority voting from the classifiers. Random subspace methods are rather similar to bagging. However, instead of randomly selecting observations from the original data set for samples, the random subspace methods randomly selects features for the data subsections. [27]

4.3.3 Random forest

Bootstrap aggregation could be utilized as an ensemble of decision trees. However, bootstrap aggregation may cause highly correlated predictors limiting the amount of variance reduction that could be possible. [12] Random forest, on the other hand, is an ensemble method capable of even higher variance reduction than bagging [13]. In random forests a subset of variables is randomly chosen at each data partitioning at the decision tree nodes in addition to random selection of data sets for the construction of each tree. The random selection of input variables adds an other layer of randomness to bagging and aims at decorrelating the decision trees. [12] Thus, the best split, which include the feature and a value for it, is chosen among the randomly selected subset of variables. Furthermore, the random variable selection is repeated at each node. Therefore, in order to learn a random forest on a training data set, the number of variables considered at each data split has to be determined. If the number of variables at each split is set to the same as the maximum number of variables in the data set is, the task is the same as in bagging. [31] Too many variables considered at each split leads to too low variance reduction while too little variables considered yield increase in bias [30]. Random forest builds decision trees using CART (Classification and regression trees) methodology, described in chapter 4.3.1, to maximum size without pruning and is robust against overfitting. Furthermore, the accuracy of the random forest depends on the strength of individual classifiers and the correlation between them. [13] Thus, random forest, $f(\mathbf{x})$, is an ensemble of N individual classification trees, $f_n(\mathbf{x})$, which have been constructed by bootstrapping samples and randomly choosing variables for a data partitioning at each node [12],

$$f(\mathbf{x}) = \sum_{n=1}^{N} \frac{1}{N} f_n(\mathbf{x}). \tag{13}$$

The possibility to use left-out bootstrap samples to estimate important statistics about the decision trees are useful for random forest. Left-out samples are used to get out-of-bag predictions in order to evaluate the generalization error in the left-out samples. [30] The out-of-bag sample set size is about one third of all the instances and the out-of bag estimates are unbiased [13]. Furthermore, it has been shown that the out-of-bag estimates perform as well as or even better than the cross-validation estimates [32]. Thus, out-of-bag estimates provide an good alternative for cross-validation. However, since bootstrapping may cause differences in prediction accuracies [30], the out-of-bag generalization error estimation is combined with 3-fold cross-validation in this work.

Furthermore, one important feature of the random forest is the possibility to asses variable importance. Variable importance estimation helps to understand the modeled phenomena more deeply in addition to the possible predictions on unseen data. In random forests the importance of a variable X_j is assessed by averaging the sum of weighted decrease in impurity for all nodes t, where the variable is used for a split, over all decision trees N,

$$Imp(X_j) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t} 1(j_t = j)[p(t)\Delta i(s_t, t)], \tag{14}$$

where the weight $p(t)$ is the proportion of observations reaching the node t and $\Delta i(s_t, t)$ is the decrease in impurity in that node. This variable importance measure is referred to as Mean decrease Gini or Gini importance. [30]

The proximity measures for variables can also be assessed within random forests giving deeper insight into the modeled data. The aim is that the random forest could give a measure of closeness for observations in the data set. The proximity is defined as the number of times two observations, x_1 and x_2, fall to the same leaf in a single tree divided by the amount of trees,

$$prox(x_1, x_2) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t \in T_n} 1(x_1, x_2 \in X_t), \tag{15}$$

where T_n describe the leaf nodes t in decision tree n. The proximity measure can be utilized for visualization of data and for identification of outliers. [30]

Thus, the hyperparameters that are considered in this work include the number of variables considered at each node for a split and the number of decision trees that are grown. Furthermore, the minimum size of the tree nodes is considered in order to ensure the absence of overfitting. Although, random forests are often grown to purity, it has been shown that for large data sets larger leaf nodes usually perform better [14]. However, for genomic data with high feature space, near purity have been discovered to be more effective, since it lowers bias [14].

4.3.4 Random forest for genetic data

High-throughput measurement techniques yield large data sets with high complexity. Modeling of the type data is challenging due to the high dimensionality and the correlated nature of genomic data, since many standard statistical models rely on independence of variables. Regularized statistical learning enables the prevention of overfitting. With random forest the greedy procedure of growing the decision trees prevents overfitting. Furthermore, random forest can easily handle correlated variables, because of the grouping property of decision trees and thus also the forests. [14] The grouping property refers to the ability of the model to select entire data subsections that belong to a class for which there is a cluster of correlated variables. If the decision tree splits the data according to one of the correlated variables, the other variables are considered soon after. Thus, decision trees will give a small minimal depth for correlated set of variables and new observations fall to those leaves more probably. Furthermore, DNA sequences with similar functionality have often highly correlated sequences. [15] Thus, random forest is well suited for modeling high-dimensional and correlated data. In addition, feature selection can readily be handled by considering the variable importance measures given by the random forest, which can be utilized for ranking the variables [14]. However, the problem with ranking is that a rank is given to variables independently [14]. Thus, a combination of variables that would predict the class together but not independently might not be found [14]. In addition, random forest can handle a mixture of categorical and numerical input variables [13]. Thus, different genetic features can be easily combined for modeling the binding specificities.

The HT-SELEX and CAP-SELEX DNA ligands are utilized as features in this thesis for the random forest. However, only one of the DNA strands are sequenced and it is not known whether the sequenced strand contained the binding site or not. The optimal type of features derived from the ligands is searched in this Master's thesis. One option is that random forest is trained on categorical features such that each nucleic acid in each position is a variable. An other option is that k-mer frequencies of the DNA sequences are utilized as variables for random forest. A k-mer refer to a DNA sequence of length k [33]. However, the feature space increases quickly with higher order k-mers, which might lead to too high computational costs. Thus, only lower k-mers, 3-mers and 4-mers are considered. In addition, it is possible to combine DNA shape features with categorical nucleic acid features. A more detailed discussion of DNA shape features is provided next.

4.3.5 Combining sequence and shape information with random forest

Transcriptional factors recognize the DNA binding sites through sequence and shape readout. Thus, the combination of sequence and shape features for modeling the transcriptional factor binding sites might be beneficial. However, it remains unclear whether the shape readout is indirect or not [34]. It is possible that the DNA shape features that attract a transcriptional factor can be inferred from the sequence [34]. DNAshape is a model capable of estimating local DNA shape features from nucleic acid sequences and is provided as a R package DNAshapeR [35]. The method is based on Monte Carlo (MC) simulations, where the DNA conformations, which are based on twelve variables, are randomly sampled. The variables that they utilized for the model include translations and rotations in addition to internal variables such as bond angles. The MC simulations were analyzed and decomposed into overlapping pentamers. Then the average shape features, including minor groove width (MGW), DNA roll, helix twist and propeller twist, were computed for the central or two central base pairs. The central base pair was used for MGW and Roll, while the two central base pairs was used for helix and propeller twists. [36] These shape features can be separately or together utilized for training the random forest classifier. In this work they are combined with sequence variables separately in order to estimate their effect on model performance.

Previously the combination of shape features with sequence information have lead to more predictive accuracy of the model. For example Yang et al. showed that for 45 % of the tested proteins the addition of DNA shape features increased model performance significantly when modeling with position weight matrices [34]. In addition, Mathelier et al. showed that incorporating DNA shape features with sequence information improves accuracy when predicting the transcriptional factor bound sites in vivo [37].

4.4 Experimental setting

HT-SELEX measurements studied in this Master's thesis contain 14, 20, 30 or 40 nucleic acids long DNA sequences, most often 20 base long DNA reads. In addition, the CAP-SELEX measurements include 40 nucleic acids long DNA reads. The data sets comprise sequenced DNA reads for which the reverse complement must be considered as well. Since the data include only positive instances, the negative set for the classification task is constructed artificially. In this Master's thesis sequenced reads are dinucleotide shuffled so that each DNA read is shuffled once yielding a negative set equal in size to the positive set. Furthermore, 75 % of the positive DNA ligands are randomly sampled to the training set while similarly 75 % of the negative DNA ligands are sampled to the training set. Random forest modeling is conducted with R package 'randomForest' [31].

4.4.1 Random forest with full length DNA read

Random forest modeling experiments begin by considering full length DNA strands for training the forest (Strand+RF). Thus, a DNA strand has to be chosen inside each DNA ligand. It is possible to utilize only the sequenced DNA reads for training the random forest. However, random forest trained on the DNA strands that contain the most probable binding site is likely to perform better. Training with full length DNA strands can be perfromed similarly for HT-SELEX and CAP-SELEX data. Thus, for each ligand the strand that comprise the highest PWM score position was chosen for training the random forest. Position weight matrices were selected from the MotifDb database that is also a R package [38]. The MotifDb database include extensive amount of position frequency matrices from literature for multiple organisms [38]. Figure 12 represent the experimental setting and training with the most probable binding site strand.

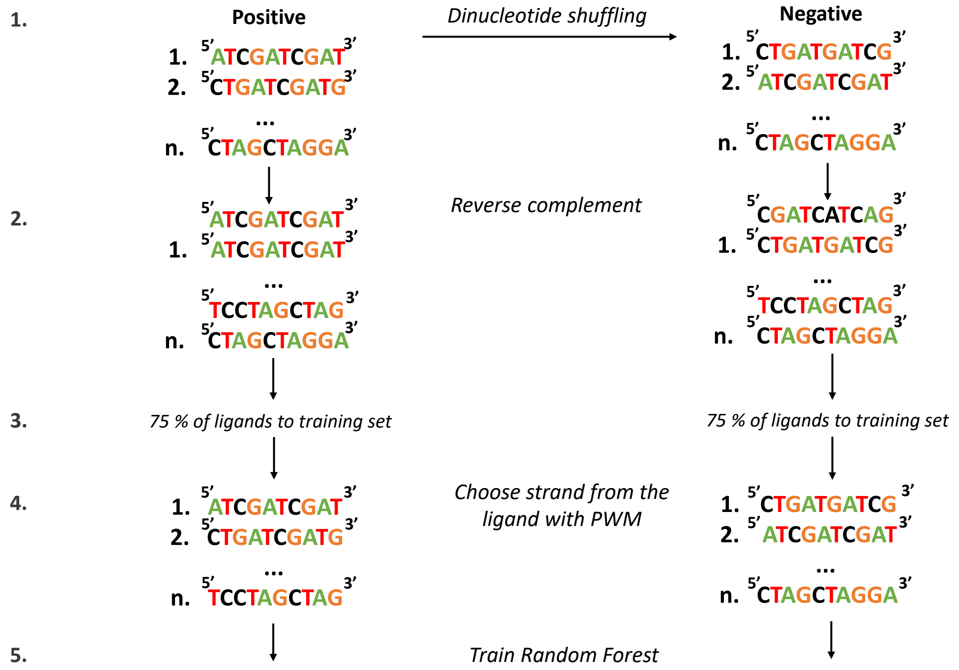
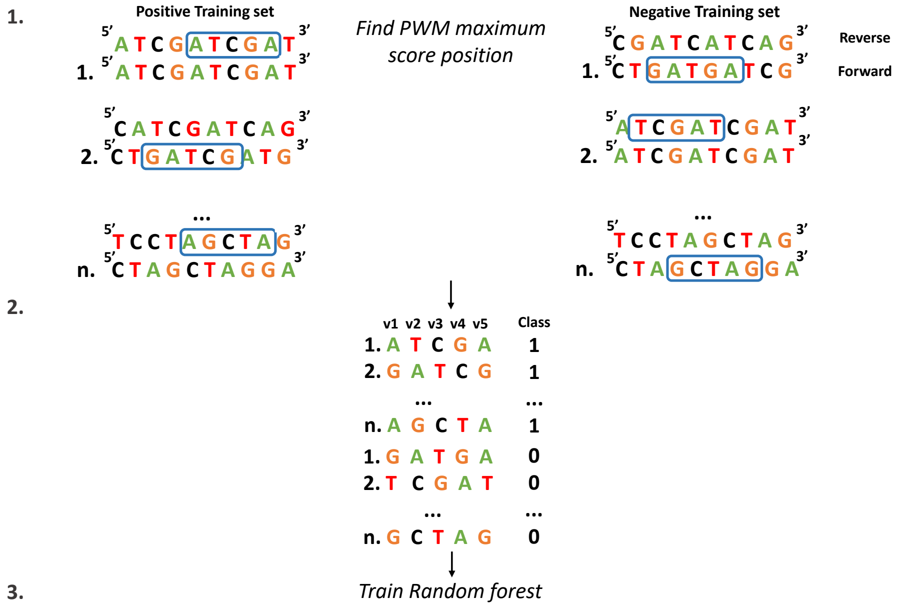


Figure 12: Experimental setting. **1**. Negative set is constructed by shuffling SELEX reads **2**. Reverse complement of each DNA sequence **3**.75 % of both positive and negative DNA ligands are randomly chosen for training **4**. The strand that include maximum PWM score is chosen out of each ligand **5**. Random forest is trained.

Random forest performance is evaluated on the test set, which include 25 % of the DNA ligands. Both of the DNA strands in each ligand are scored with the random forest and the maximum score out of the two is chosen as the final score for the ligand. The scores range between 0 and 1, thus yielding a probability for the ligand for containing a binding site.

4.4.2 PWM site and random forest for HT-SELEX

DNA ligands contain a binding site but the location and the length of this site is unknown. It is possible that binding sites are shorter than HT-SELEX sequences, which implicate that for random forest model with entire strands, excessive nucleic acids that are not related to binding would be utilized for training, which could disturb random forest performance. Therefore, random forest training could be more efficient by extracting the most probable binding sites inside each DNA ligand and utilizing only these sites for training the classifier (PWM+RF). The most probable binding sites can be found with PWMs from literature, which again were selected from the MotifDb database [38]. Homo sapiens PWMs from other studies than those modeled in this Master's thesis were chosen if possible to extract most probable binding sites in order to prevent learning with PWMs that have been trained with the same data. Figure 13 represent random forest training with the most probable binding sites. Experimental setting is similar than represented in Figure 12 for modeling with full length DNA reads until choosing the strand in step four.


1.
Positive Training set
5' A T C G A T C G A T 3'
1. 5' A T C G A T C G A T 3'
5' C A T C G A T C A G 3'
2. 5' C T G A T C G A T G 3'
...
5' T C C T A G C T A G 3'
n. 5' C T A G C T A G G A 3'
Find PWM maximum score position
Negative Training set
5' C G A T C A T C A G 3'
1. 5' C T G A T G A T C G 3'
5' A T C G A T C G A T 3'
2. 5' A T C G A T C G A T 3'
...
5' T C C T A G C T A G 3'
n. 5' C T A G C T A G G A 3'
Reverse
Forward
2.
3.
Train Random forest

	v1	v2	v3	v4	v5	Class
1.	A	T	C	G	A	1
2.	G	A	T	C	G	1
...						...
n.	A	G	C	T	A	1
1.	G	A	T	G	A	0
2.	T	C	G	A	T	0
...						...
n.	G	C	T	A	G	0

Figure 13: Random forest with PWM sites. **1**. Training data comprise positive and negative DNA ligands, for which each position is scored with PWM **2**. The subsequence with maximum PWM score is chosen inside each ligand **3**. Random Forest is trained with the most probable binding sites.

When training is performed with PWM sites, testing must be conducted by scoring subsequences of the same length than the PWM and thus the training sequences. Therefore, each subsequence on the forward and reverse strands are scored with the random forest. Final score for each ligand is then the maximum score out of all the per position scores assigned to that ligand. Figure 14 represent scoring one test set

DNA ligand with random forest trained with five nucleic acids long subsequences.

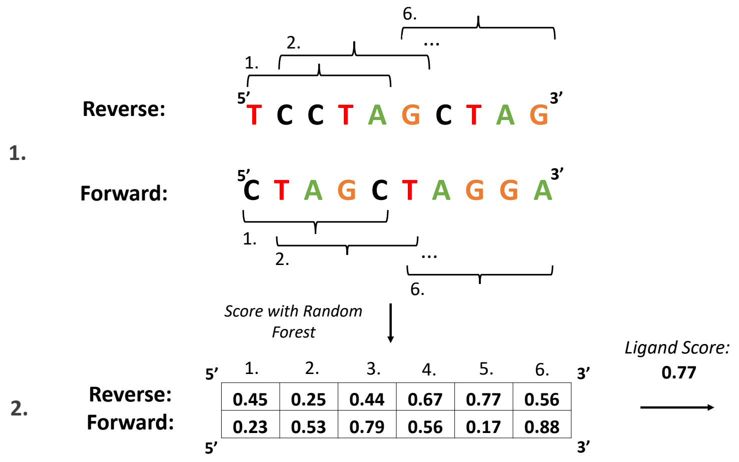


Figure 14: Scoring DNA ligand with random forest trained with PWM sites. 1. Each subsequence of the same length than the PWM is scored with random forest 2. Ligand score is the maximum score out of all per position scores.

Training and testing experimental scheme for random forest with PWM sites is represented in Algorithm 2 as well. Reverse complements are already computed. Thus, the represented algorithm takes the sequences in ligand form.

```
PWM_RF(Ligands, PWM);
1. Subsequences = $\emptyset$ ;
2. for each training ligand do
    s = find the subsequence with maximum PWM score;
    Subsequences = Subsequences $\cup$ s;
end
3. Train random forest with Subsequences;
4. Score = $\emptyset$;
5. for each test ligand do
    s = score all positions with random forest and choose the maximum score;
    Score = Score $\cup$ s;
end
6. Final score for test ligands = Score;
```

Algorithm 2: Training random forest with PWM sites and testing DNA ligands.

4.4.3 PWM site and random forest for CAP-SELEX

The sequences in CAP-SELEX data contain binding sites for two transcriptional factors [4]. Jolma et al. published combinatorial position weight matrices for the studied TF pairs [4]. These PWMs are utilized in this Master's thesis similarly as PWMs in HT-SELEX experiments to train random forest with full length DNA strands and with PWM sites only. Furthermore, random forests are trained with binding sites chosen by individual transcriptional factor PWMs, which were derived by Jolma et al. from HT-SELEX data sets [5, 4].

The two TFs can bind DNA with different spacings between them [4]. There are four possible orientations by which the TF pair can bind a DNA ligand, which are represented in Figure 15. Furthermore, the TF pair can bind DNA with overlapping binding motifs or with nucleic acid gaps between the two motifs [4]. PWMs of both TFs can be used for searching combination binding sites with different spacings between the motifs and training random forests with these sites (PWM1+PWM2+RF). In this model the training data is divided into training set and validation set. The two position weight matrices are attached to each other according all possible orientations and gaps yielding multiple different combination PWMs. In the training set, random forests are trained with subsequences chosen by these combination PWMs separately such that each PWM attachment configuration will yield one forest. The validation data set was extracted from the training set in order to estimate the performance of the random forests trained with different spacings between the two PWMs. Thus, the spacings whose forest give the best predictive accuracy on the validation set are chosen as the best TF pair spacings for the final model. Finally, random forests are trained with subsequences chosen by the best PWM attachment configurations using the entire training data. Test set DNA ligands are then scored as represented in Figure 14 with all the random forest separately. Furthermore, the average of these scores was assigned to each ligand as the final score.

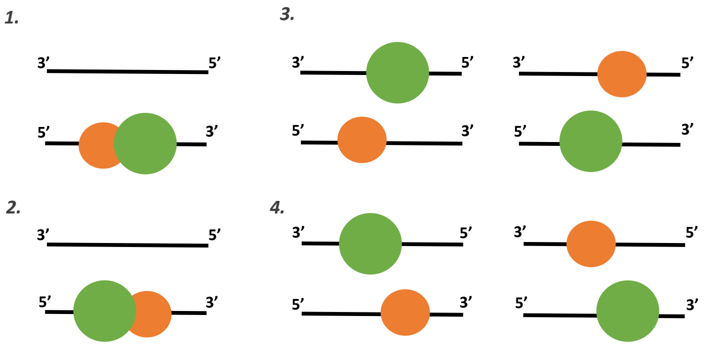


Figure 15: Four possible protein pair binding orientations.

Furthermore, the TF pair binding sites are modeled by searching the maximum PWM score position of one of the transcriptional factors and extending the site to cover also the binding site of the other protein (PWM1+PWM2+N+RF). The PWM extension is performed by adding [0.25, 0.25, 0.25, 0.25] columns to the other end of the PWM before searching for the binding site. However, the direction to which

PWM elongation should occur is unknown. Therefore, two random forests are trained separately such that both extension direction are covered. Binding sites for the first random forest are searched with PWM of TF1 elongated to the right and PWM of TF2 elongated to the left. Elongation is performed until the position weight matrix is 25 nucleic acids long. Furthermore, the binding site for random forest training will be the site with maximum score out of these two PWM matches. This way it is possible to find putative binding sites from both ends of the DNA ligands. Furthermore, since the elongation is conducted only until 25 nucleic acids, even shorter binding sites in the middle on the DNA ligands can be found. Although, it is possible that 25 nucleic acids binding motif is too short for some TF pairs. The second random forest is trained with subsequences that are given maximum PWM scores out of each ligand with either PWM of TF2 elongated to the right or PWM of TF1 elongated to the left. Therefore, the first random forest is trained with subsequences that resemble orientation one, three or four, which are represented in Figure 15, or their combination. The second random forest on the other hand is trained on subsequences searched with PWMs such that they capture either orientations two, three or four or their combination. DNA ligands are tested by scoring with both random forests as represented in Figure 14 and taking the average of the two scores as the final score for the ligand. This model is able to consider all possible gaps between the motifs simultaneously even though orientations between the proteins on the ligands need the be considered separately. Furthermore, Algorithm 3 represent the PWM1+PWM2+N+RF model training and testing.

```
PWM1_PWM2_N_RF(Ligands, PWM1, PWM2);
1. PWM1_x and x_PWM1 by elongating PWM1 to right and left;
2. PWM2_x and x_PWM2 by elongating PWM2 to right and left;
3. Subsequences_1 = $\emptyset$;
4. Subsequences_2 = $\emptyset$;
5. for each training ligand do
     s1 = find the subsequence with maximum PWM1_x or x_PWM2 score;
     s2 = find the subsequence with maximum PWM2_x or x_PWM1 score;
     Subsequences_1 = Subsequences_1 $\cup$ s1;
     Subsequences_2 = Subsequences_2 $\cup$ s2;
end
6. RF_1 = train random forest with Subsequences_1;
7. RF_2 = train random forest with Subsequences_2;
8. Score = $\emptyset$;
9. for each test ligand do
     s1 = score all positions with RF_1 and choose the maximum score;
     s2 = score all positions with RF_2 and choose the maximum score;
     Score = Score $\cup$ average(s1, s2);
end
10. Final score for test ligands = Score;
```

Algorithm 3: PWM1+PWM2+N+RF training and testing scheme.

4.4.4 Multiple random forests model

Furthermore, it is possible to find the binding sites with decision trees instead of the PWM before training the final random forest model with chosen sites (RF+RF). The learning problem with HT-SELEX and CAP-SELEX data can be considered as a multiple instance learning problem, where the sequences are divided into bags and it is known that in the positive set at least one of the bags comprise a binding site while information about which bag it is, is missing [39]. In the negative set all the bags are negative [39]. Thus, the binding sites can be discovered with multiple random forests by first dividing the DNA ligands into equally sized bags and then training a random forest on each of the bags. The bags are constructed by shifting along the DNA strand one nucleotide acid at a time similarly on forward and reverse strands. The chosen amount of bags depend on the length of the DNA reads in the SELEX data set. For instance for the 20 nucleic acids long DNA reads the bags are chosen to comprise 18 nucleic acids long subsequences. Figure 16 represent how the bags are constructed on training set and utilized for learning random forests.

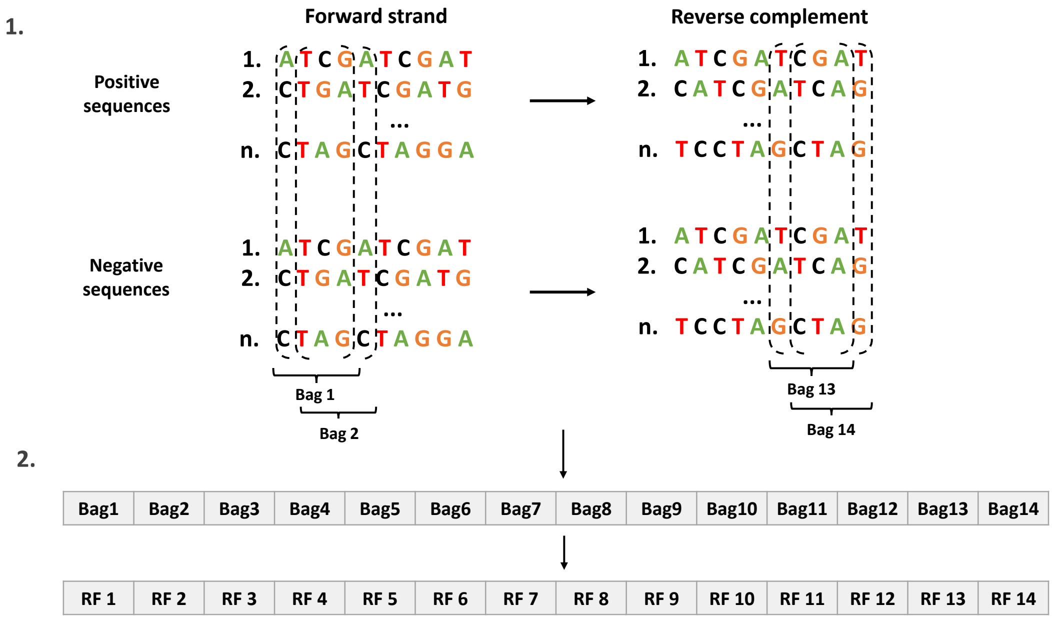


Figure 16: Searching binding sites with decision trees. **1**. Sequences are divided into bags such that both forward and reverse strands are considered **2**. Random forests are trained on each bag.

The sequences in each bag are scored with all random forests for which the training set did not comprise the sequence itself or the reverse complement. Final score for each subsequence is the average of all scores assigned to that subsequence. Then for each DNA ligand the subsequence with the maximum score will be chosen as the binding site and utilized for training the final random forest model.

Binding should occur evenly on all positions in order for the multiple random forest model to function properly. The Inimotif pipeline, which all the data sets studied in this Master's thesis have passed, should choose only experiments with evenly positioned binding sites on both strands [7]. Bags contain binding sites and random sequences on the positive set and random sequences on the negative set. As mentioned before decision trees can find correlated groups in the data and classify them to same class. Thus, randomness in the positive set does not necessarily disturb random forest significantly. The aim however is to construct the best possible model for protein-DNA interactions. Therefore, the most probable binding sites for training the random forest are chosen through the weaker random forests trained on each bag separately. Algorithm 4 describe how binding site sequence is chosen with random forests and utilized for training the final random forest model, which is eventually used for testing unseen DNA ligands. The represented algorithm performs computations for SELEX measurements such that the parameter shift describe bag positions as shown in Figure 16.

```
RF_RF(Ligands);
1. for each shift n do
    Bag_n = Subsequences in training ligands that belong to position n;
    RF_n = train random forest with sequences in Bag_n;
end
2. for each shift n do
    Score sequences in Bag_n with those random forests that were not trained
      with Bag_n or the reverse complement Bag;
    Average scores given by the random forests for each sequence in Bag_n;
end
3. Sequences = ∅;
4. for each training ligand do
    Seq = Choose sequence from Bag_n that has the maximum score;
    Sequences = Sequences ∪ Seq;
end
5. Train random forest with Sequences;
6. Score = ∅;
7. for each test ligand do
    s = score all positions with random forest and choose the maximum score;
    Score = Score ∪ s;
end
8. Final score for test ligands = Score;
```

Algorithm 4: Modeling with RF+RF.

4.4.5 Summary of models for HT-SELEX and CAP-SELEX data

Multiple random forest models are implemented and their performance is assessed. For HT-SELEX data the models include Strand+RF and modeling with most probable binding sites searched with either position weight matrices (PWM+RF) or decision trees (RF+RF). The models are summarized in Table 2.

Table 2: Random forest models for HT-SELEX

	Model	**RF model**
1.	Strand+RF	Full length DNA strands
2.	PWM+RF	Sites chosen by PWM
3.	RF+RF	Sites chosen by random forests

CAP-SELEX data is modeled with the same models represented in Table 2 by utilizing either individual position weight matrices of TF1 or TF2, or the TF pair combinatorial PWM. The ensemble model of multiple spacings (PWM1+PWM2+RF) is implemented in addition to the model with PWMs elongated to cover the other TF binding site (PWM1+PWM2+N+RF). These models are summarized in Table 3.

Table 3: Random forest models for CAP-SELEX

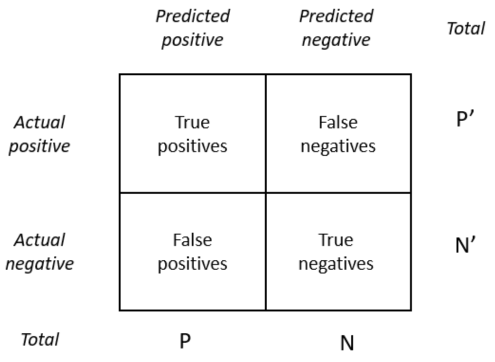
	Model	**RF model**
1.	Strand+RF	Full length DNA strands
2.	PWM1+RF	Sites chosen by PWM of TF1
3.	PWM2+RF	Sites chosen by PWM of TF2
4.	PWM+RF	Sites chosen by combinatorial PWM of TF1 and TF2
5.	RF+RF	Sites chosen by random forests
6.	PWM1+PWM2+RF	Ensemble of random forests with sites chosen by different TF1-TF2 spacing PWMs
7.	PWM1+PWM2+N+RF	Sites chosen by PWM of TF1 or TF2 and elongated to cover the other TF site

4.5 Evaluation of results

The performance of the models is assessed by comparing prediction accuracies. The simplest measure of prediction accuracy is the percentage of correctly classified cases among all predictions. The receiver operating characteristic (ROC) curve is an other way to visualize and evaluate model performance. For binary classification the area under ROC-curve yields a better estimate of the overall performance of the model than the simple prediction accuracy, since the discrimination threshold for classification to the two classes is chosen manually. Thus, in this Master's thesis, prediction accuracy is evaluated by assessing sensitivity and specificity of the model predictions and by computing the area under ROC-curve. [40]

4.5.1 Sensitivity and specificity

Sensitivity and specificity are measures needed for the construction of receiver operating characteristic (ROC) curve. True positive rate (TP) describes the percentage of correctly classified positive cases and true negative rate (TN) is the percentage of correctly classified negative cases. The four variables are often represented within a confusion matrix as shown in Figure 17. [40]



	Predicted positive	*Predicted negative*	*Total*
Actual positive	True positives	False negatives	P’
Actual negative	False positives	True negatives	N’
Total	P	N	

Figure 17: Confusion matrix.

In addition, model performance analysis may include inspection of variables false positive rate (FP) and false negative rate (FN). These measures describe the percentages of incorrectly classified positive and negative instances. Furthermore, the true positive rate is also called sensitivity, which is,

$$Sensitivity = TP\ rate = \frac{TP}{P}, \tag{16}$$

where P is the amount of positive samples as shown in Figure 17. [40] Specificity is an other important variable for ROC-curve analysis, which can be computed as [40],

$$Specificity = 1 - FP\ rate = \frac{TN}{FP + TN}. \tag{17}$$

4.5.2 Receiver operating curve

The ROC-graph is a graph where the TP rate is plotted against FP rate. Thus, the ROC-graph represents the benefits (TP) and costs (FP) of the classifier. Some classifiers such as the random forest yield a probability for each instance for belonging to one of the classes. [40] In addition, some classifiers might give general scores instead of probabilistic scores, as does DeepBind [10]. The predictions can be converted to classes by utilizing a threshold above which the sample is classified to class 1 and below to class 0. These predictions can be represented in ROC-curve so that each threshold value is utilized to give a pair of sensitivity and specificity values. All these plotted on the ROC-graph will yield a ROC-curve. ROC-curves are favored in binary classification evaluation, because they are insensitive to changes in class distributions, and they represent the ability of the classifier to predict positive samples relative to negative samples. The area under ROC-curve (AUC) can be used as one value to compare classifiers. The value will be between 0 and 1 since it is the area under a unit square. Although, AUC of 0.5 would be achieved by guessing the class. Therefore, reasonable classifiers have always AUC between 0.5 and 1. [40]

4.5.3 Statistical significance

AUC values are obtained with multiple different random forest models in addition to scoring DNA sequences with only the position weight matrices and for HT-SELEX data also with the neural network model DeepBind. The statistical significance of the differences in AUC values is assessed with Wilcoxon signed-rank test, which is a non-parametric statistical test [41]. Thus, the data does not need to be normally distributed as in paired t-test. The conducted tests are two-sided paired tests, which test whether population means differ significantly.

5 Results

In this chapter results for random forest modeling with single transcriptional factor HT-SELEX data and transcriptional factor pair CAP-SELEX data are represented. Multiple different models are implemented and their performances are tested. Furthermore, the best choice of features derived from SELEX sequences and values for random forest parameters are optimized with HT-SELEX assays. The best model was achieved by utilizing nucleic acids as categorical features for the random forest. Furthermore, random forests can be trained with full length DNA strands (Strand+RF), or with sites chosen by a PWM (PWM+RF) or decision trees (RF+RF) from the DNA ligands. Single TF binding specificities are modeled with these approaches using HT-SELEX data. TF pair binding specificities are modeled using CAP-SELEX data. The same experimental schemes are investigated for TF pairs. However, the PWM sites are searched with the individual TF position weight matrices in addition to the combinatorial position weight matrix of the TF pair. Furthermore, two proteins can bind a DNA ligand with four different orientations and the amount of gaps between their motifs can vary [4]. Thus, random forest models trained with TF pair binding sites found inside DNA ligands with individual position weight matrices of TF1 and TF2 are implemented (PWM1+PWM2+RF and PWM1+PWM2+N+RF). Random forest was discovered to perform almost equally to the DeepBind neural network model for HT-SELEX data and to outperform scoring sequences with position weight matrices for both HT-SELEX and CAP-SELEX data.

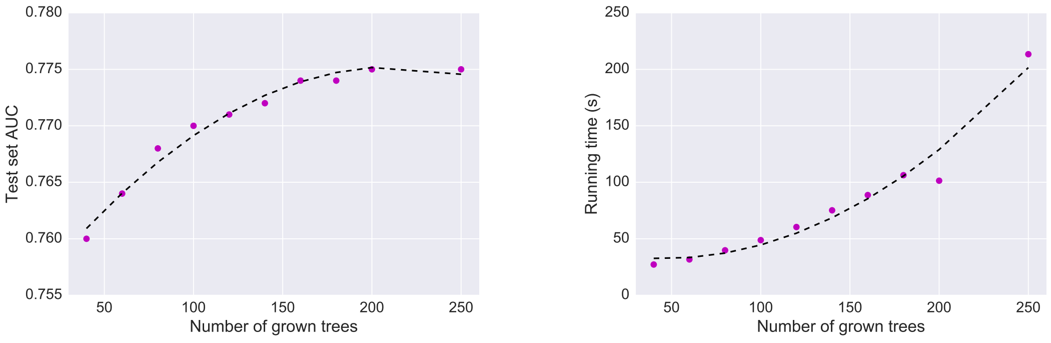
5.1 Random forest for HT-SELEX

In this Master's thesis DNA binding sites of TFs are modeled with random forest. Three different random forest models are compared to the neural network model, DeepBind [10]. In addition, the performance of the random forest models is assessed in comparison to scoring the sequences with PWMs by Jolma et al. [5]. The three best ways of modeling with random forest include modeling with the full length DNA strand, modeling with PWM sites extended with surrounding nucleic acids and modeling with subsequences chosen by random forests. In this chapter, parameter tuning is represented for random forest modeling with PWM sites and with the entire DNA strands separately. The rest of the experiments are performed with the optimized random forest parameters. In addition, a case study with Alx1 is represented, which demonstrate the selection of the best random forest models. Furthermore, experiments were performed on more TFs in order to validate the results and finally the performance of the best random forest models were compared to DeepBind and PWM scoring for a more comprehensive data set.

5.1.1 Tuning random forest parameters

Random forest can be tuned by altering decision tree and forest parameters. The parameters are tuned by considering subsets of HT-SELEX data for five proteins with different PWM lengths. One parameter at a time the others are set to random forest default values and the effects of altering the considered parameter on test set AUC is observed. The default value for minimum node size (N_S) is 1, for number of features considered at a split (N_V) it is $\sqrt{p}$, where p is the number of features, and for sample size it is the number of training data observations. The number of decision trees (N_T) is considered first and the rest of the experiments are conducted with that value. Only 50 % of the HT-SELEX data is utilized for model tuning in order to keep running time reasonable. Parameter tuning is performed for two different models separately: random forest with entire DNA reads and random forest with PWM sites. Optimization is performed with Alx1, Arx, Barhl2, Batf3 and Bhlhb3.

Increasing the number of bootstrap samples taken and thus the number of decision trees grown will improve modeling accuracy. However, computational complexity and running time increases accordingly. Figure 18 represent how increasing the number of decision trees grown affect the test set AUC and running time. The represented values are the mean values of the conducted experiments. Furthermore, the optimization of the number of decision trees is represented only for the random forest model with the entire DNA strand, because the most probable binding sites will perform at least as well due to the smaller number of features.



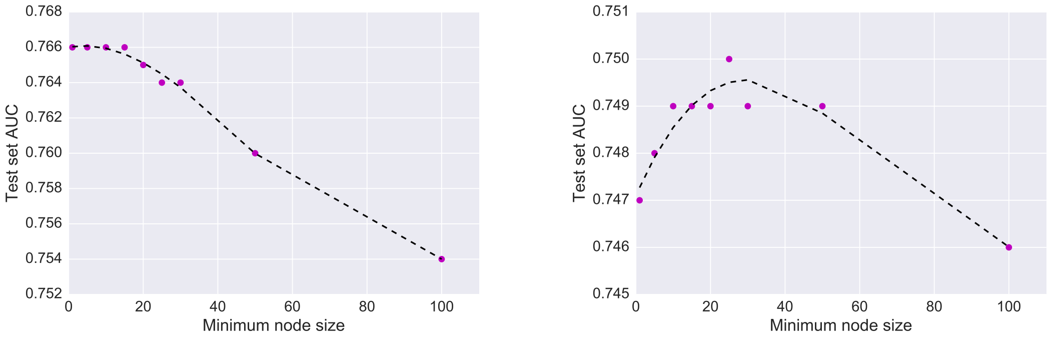
(a) AUC

(b) Running time

Figure 18: Optimizing number of decision trees grown.

The number of decision trees for further experiments is chosen according to Figure 18. The parameter should be large enough to assure that increasing it would not cause significant improvements in the test set AUC but at the same time to be small enough to minimize running time. It seems that growing 200 decision trees to a random forest is a good choice. The trade-off between consumed time and the obtained test set AUC is optimal. Thus, for further experiments with other proteins N_T is 200. However, when other random forest parameters are tuned N_T is set to 140 in order to keep running time minimal while performance is close to optimal.

The depth that the trees are grown can be altered through changing random forest parameters as well. In this Master's thesis tree depth is controlled with minimum size of terminal nodes. Thus, the trees are learned as deep as they can be unless the minimum node size is reached making the node a terminal node. The parameter is tuned with the same five proteins but this time also for the random forest model with PWM sites. The effects of altering minimum node size when the number of decision trees grown is set to 140 and other parameters are at default values is represented in Figure 19.



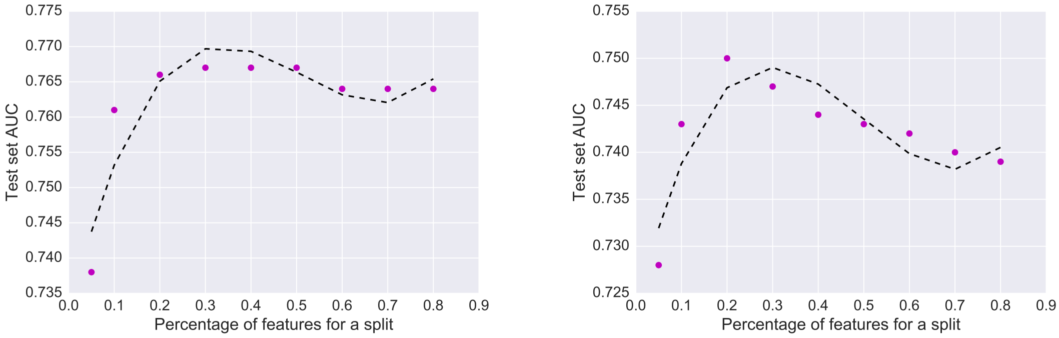
(a) Entire DNA strand

(b) Most probable binding site with PWM

Figure 19: Optimizing minimum node size.

Random forest modeling with full length DNA strand favor higher tree depths while for modeling with PWM site the best tree depth is lower. This might occur due to higher number of features in the full length DNA strand model. Furthermore, TFs that express more unspecific binding could favor lower tree depths while factors with highly specific binding possibly benefit from high tree depths. Therefore, N_S should be low enough to ensure optimal modeling for highly specific TFs and high enough to prevent overfitting especially for the more unspecific TFs. However, random forest does not seem to be very sensitive to the choice of N_S. Future experiments are performed with $N_S = 10$, which gives good accuracies for both models.

The number of variables tried at each split is probably the most crucial parameter random forest. Furthermore, the parameter depends upon the number of features in total. Thus, parameter N_V is represented as the the percentage of the variables tried at each split out of the total number of features. Parameter tuning similarly as before for Strand+RF and PWM+RF is represented in Figure 20.



(a) Entire DNA sequence

(b) Most probable binding site with PWM

Figure 20: Optimizing number of features tried at each split.

The optimal number of features chosen randomly for a split is different for modeling with the entire strand than with only the PWM site. In addition, model performance with the site chosen by PWM is slightly more sensitive to the choice of this parameter. It can be seen that when modeling with the entire DNA strand, 30%-50 % of the total number of features is the best choice for the number of variables sampled randomly for a split. Therefore, 40 % is chosen as the optimal value for N_V. However, for the random forest trained with PWM sites, the best amount of variables considered at each split is 20 % - 30 % of the length of the PWM. For shorter PWMs, which may be even shorter than ten nucleic acids, the 20 % might be too low. Therefore, the experiments in this Master's thesis for the more comprehensive data set are performed such that N_V is 30 % of the total number of features.

Furthermore, the size of the bootstrap samples can be altered. However, in this Master's thesis the default, which is the number of the instances N in the training set, is utilized. Thus, bootstrapping is performed. Bootstrapping will result in approximately 63.2 % of the training data set being in the sample for which the tree is grown while 37.7 % of the training data belongs to the out-of-bag sample [13]. In addition, prior probabilities for the two classes might influence random forest performance. However, in this thesis balanced classes are utilized for random forest training, which means that prior probability for both classes is 50 %. Thus, alterations in prior probabilities are not needed.

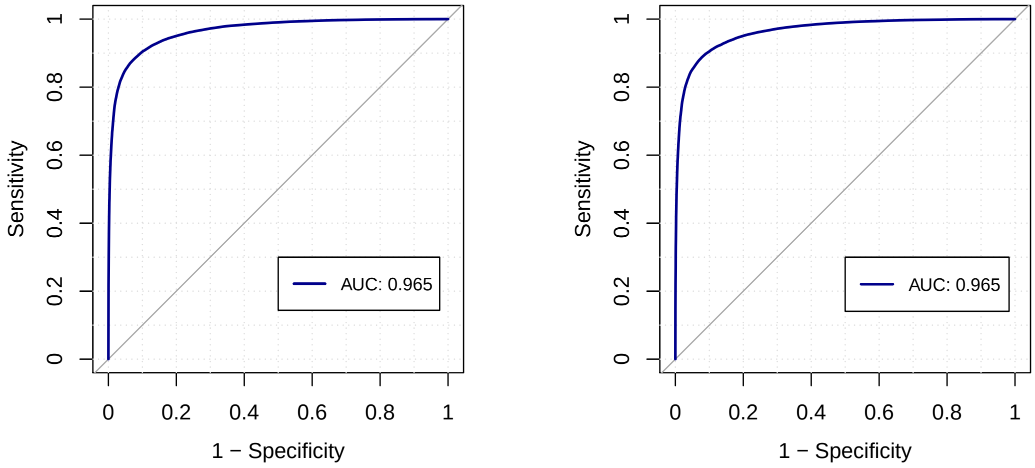
5.1.2 Transcriptional factor binding motif with random forest

In this chapter the best random forest models are searched with a case study with Alx1 transcriptional factor. Alx1 HT-SELEX assay cycle three was chosen for the case study, because the data set had an average size, with 20 nucleic acid long DNA reads, which is the most common experimental type in the Jolma et al. study [5]. Furthermore, other data sets are considered for some of the random forest models in order to validate the results. These include HT-SELEX assays with 20 nucleic acids long DNA reads for transcriptional factors Arx, Ar, Barhl2, Batf3, Bhlhb3 and Bhlhe41. In addition, experiments with 30 nucleic acids long DNA reads for TFs Atf7 and Dprx are considered as well as HT-SELEX data sets with 40 nucleic acids long DNA reads for Barhl2 and Cux1. The experiments begun by searching the best type of DNA features for random forest modeling with full length DNA reads. The options include the DNA sequence in a categorical form such that each feature is a nucleic acid $N \in \{\text{A, C, G, T}\}$ and the usage of k-mer frequencies in the DNA reads. Furthermore, the addition of the DNA shape features is considered. Performance of the random forest trained with only the PWM sites searched from each DNA ligand (PWM+RF) is assessed. Furthermore, certain modification to the PWM+RF model are implemented and their effect on predictive accuracy with unseen DNA ligands are examined. In addition, the RF+RF model performance is assessed.

5.1.2.1 Random forest with full length DNA strands

In this chapter random forest performance is assessed when the full length DNA strand is utilized for modeling. The HT-SELEX sequenced DNA reads are either

14, 20, 30 or 40 nucleic acids long. Thus, for example in the case of a 20 base long DNA sequence, the random forest model will get 20 variables $N_1, ..., N_{20}$ ordered according to the position of the nucleic acids in the DNA read, $i = 1, .., 20$, and each variable is a nucleic acid $N_i \in \{A, C, G, T\}$. Furthermore, only one of the strands of the DNA ligand is sequenced in HT-SELEX experiments. Thus, the strand where TF binds is unknown. Model training is performed either with the sequenced reads or with the DNA strand that contain the most probable binding site according to a PWM. Due to complementary nature of double stranded DNA, the binding site is reflected also on the other DNA strand. Thus, training the model with only the sequenced strand could work. However, it is likely that random forest can find the binding sites better when training is performed with the more probable binding site strand. When the model is tested with unseen DNA ligands, both of the strands are tested and probability for that ligand to contain a binding site is chosen to be the maximum probability out of the two strands. Random forest parameters were $N_T = 200$, $N_S = 10$ and N_V is 40 % of total number of features as optimized in previous chapter. These two modeling approaches are tested for Alx1. ROC-curves with AUC-values for Alx1 data set with both of the random forest training approaches that consider full length DNA strands are represented in Figure 21.



(a) Sequenced strands

(b) Strands with most probable binding sites

Figure 21: Training with the full length DNA strands in categorical form with Alx1.

Thus, random forest model with the sequenced DNA strand and with the most probable binding site strand perform equally at least for Alx1. Future experiments are performed by training the forest with the strand that contain PWM site, because this way of modeling is more reliable and should perform well also for other proteins.

5.1.2.2 Random forest model with k-mer frequencies

An other way of modeling is to utilize k-mer frequencies as features. Feature spaces with 1-mer, 2-mer, 3-mer and 4-mer vectors constructed from DNA reads are considered. Furthermore, full length DNA strands that contain PWM sites are utilized for k-mer random forest modeling. It can be expected that 1-mer and 2-mer counts do not perform very well as random forest features since the 1-mer frequencies are not usually very informative for binding specificities and the background DNA set was constructed so that dinucleotide count was preserved. Therefore, 3-mer and 4-mer frequency features were expected to be give more accurate random forest models for Alx1 data set. The 4-mer frequency features already produced so time consuming training that the number of trees grown on random samples had to be decreased to 20, while 1-mer and 2-mer random forests were trained with $N_T = 200$ and $N_T = 100$ respectively. Furthermore, the 3-mer random forest had $N_T = 50$. Thus, k-mers higher than four would lead to computationally too heavy problems. In addition, 1-mer random forest was trained with 2 and other k-mer random forest with 8 variables chosen randomly for each split. For all k-mer models $N_S = 10$. Figure 22 represent the k-mer random forest model ROC-curves and AUC values.

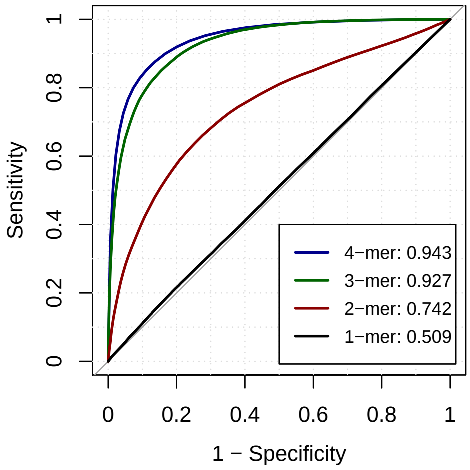
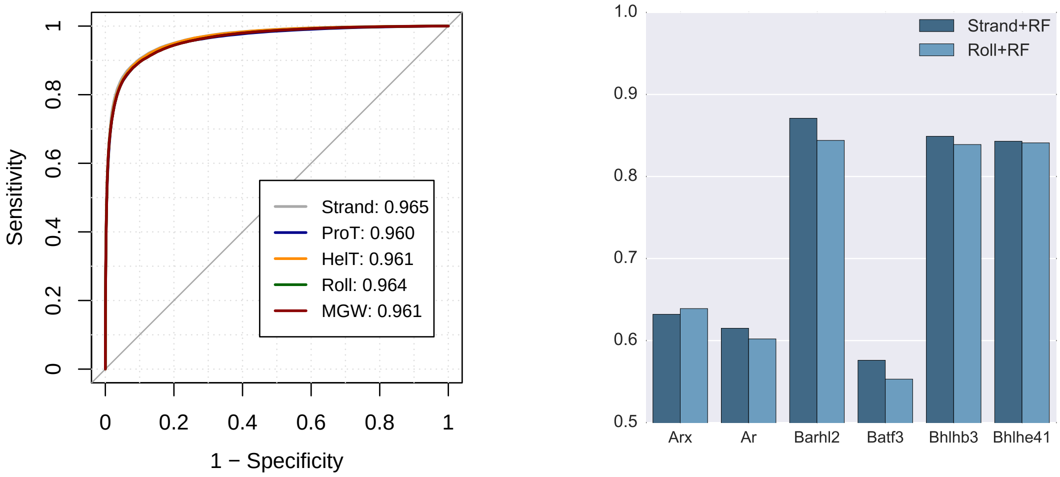


Figure 22: ROC-curves and AUC for Alx1 with different k-mer frequencies.

As expected the 1-mer feature random forest model is not able to distinguish between binding and non-binding sites. However, similar results were expected for 2-mer model as well. One explanation for the ability of random forest to classify with 2-mer features is the fact that the training set does not contain the corresponding negative 2-mer feature vectors shuffled from each positive sequence since the training set is constructed by sampling randomly 75 % of the positive reads and similarly 75 % of the negative reads to the training set. In fact, when the training and testing sets were divided so that the training set contained 75 % of the positive reads and the corresponding shuffled negative reads, test set AUC was 0.500, while 3-mer (AUC = 0.924) and 4-mer (AUC = 0.943) test set AUC values remained almost unchanged.

5.1.2.3 Combining sequence and shape information

Categorical features for modeling the binding sites perform better than k-mers and is considered further. It has been shown previously that combining DNA shape features to 1-mer sequence variables increase modeling accuracy significantly [36]. Thus, the shape features are combined with the categorical features. The four shape features are minor groove width (MGW), DNA roll (Roll), propeller twist (ProT) and helix twist (HelT). The number of decision trees was set to 200, minimum node size to 10 and the number of variables chosen for a split was 40 % of total number of features as optimized for the full length DNA strand categorical model. Figure 23 represent the four shape features independently combined with the sequence variables for Alx1. The shape features are utilized independently of each other in order to keep the feature space and thus the computational cost smaller. Figure 23 represent also test set AUC values for other HT-SELEX data sets when shape feature Roll is combined with sequence features. Random forest parameter values were the same as for Alx1, $N_V = 40\%$, $N_T = 200$ and $N_S = 10$. However, as the feature space increases with the addition of the Roll features, running time increases as well. Therefore, for the largest data set of Arx protein, N_T was set to 100.



(a) Different shape features for Alx1

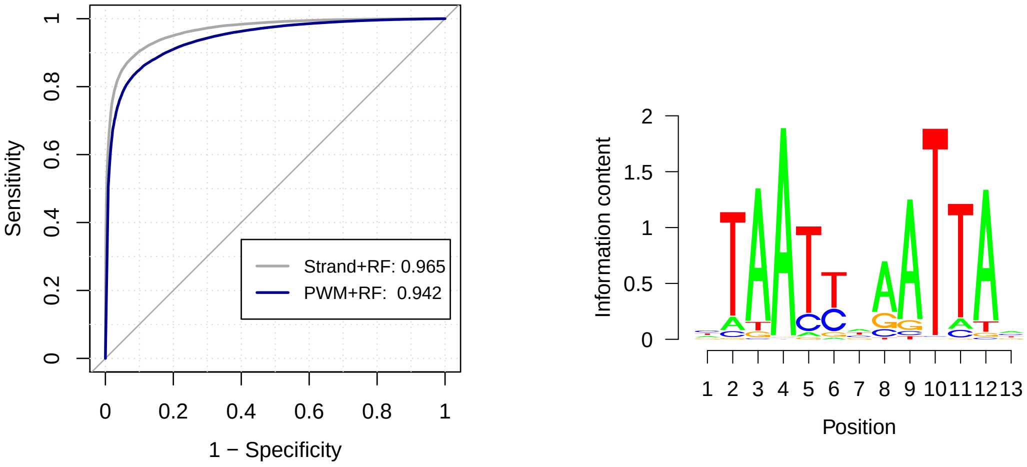
(b) Roll feature for multiple TFs

Figure 23: Training with the combination of sequence and shape features.

Shape features do not increase modeling accuracy for Alx1. Out of the four different features, DNA Roll feature is the best shape feature for Alx1. Thus, DNA Roll feature was tested for other TFs as well. However, for most proteins adding shape feature Roll did not increase modeling accuracy. Although for Arx data set the test set AUC with Roll feature is probably an underestimate due to the smaller N_T, it can be concluded that the addition of DNA Roll features do not improve modeling accuracy significantly. Shape features are generated through k-mer feature encoding from the sequences [35]. Thus, the result in Figure 23 is as expected since k-mer information is already present in the sequence features. Furthermore, lower feature space of the model with only sequence features might be beneficial for random forest.

5.1.2.4 Random forest model with PWM

If the binding site is shorter than the sequenced DNA read in HT-SELEX experiment, random forest might have problems building an accurate model. Since true binding sites are located in different positions in the DNA reads and each nucleic acid position is represented as a feature, there will be unnecessary dispersion in the features that makes uncovering the true binding specificity difficult. The information about the binding site characteristics will not be only at certain features. Therefore, it could be beneficial to first extract the most probable binding site with a PWM and use only those subsequences for training the random forest. As nucleic acid positions relative to each other would be similar across all instances, learning the binding specificities might be easier for the forest. Therefore, position weight matrices are utilized for finding the most probable binding sites inside DNA ligands before the random forest is trained with these subsequences. Furthermore, the unseen DNA ligands are tested with the model by going through each possible binding site within the sequenced strand and its reverse complement. In addition, random forest could learn a binding specificity model that is more accurate than the position weight matrix alone, since the decision trees can increase model complexity. Figure 24 represent the ROC-curve for Alx1 transcriptional factor model that combines position weight matrices and random forest modeling. The random forest parameters were as optimized before, $N_T = 200$, $N_S = 10$, and $N_V = 20\%$.



(a) ROC-curves and AUC with Strand+RF and PWM+RF.

(b) Alx1 PWM from [5] and visualized with information content.

Figure 24: Random forest modeling with PWM sites with Alx1 data set.

Utilizing full length DNA strands for random forest training performs better at least for Alx1 than using only the most probable binding sites found with PWM. In order to validate whether training random forest with PWM sites is inefficient also for other TFs the same comparison is conducted for multiple HT-SELEX data sets with 20 nucleic acids long DNA reads. Furthermore, model performances are assessed with HT-SELEX data sets comprising 30 nucleic acids long DNA reads in addition to data sets comprising 40 nucleic acids long DNA reads in order to examine whether predictive accuracy between the two random forest models depend on the DNA read length. For all the experiments $N_V = 20\%$ for the PWM+RF model and $N_V = 40\%$ for the Strand+RF model. In addition, $N_T = 200$ and $N_S = 10$. Figure 25 represent the test set AUC values.

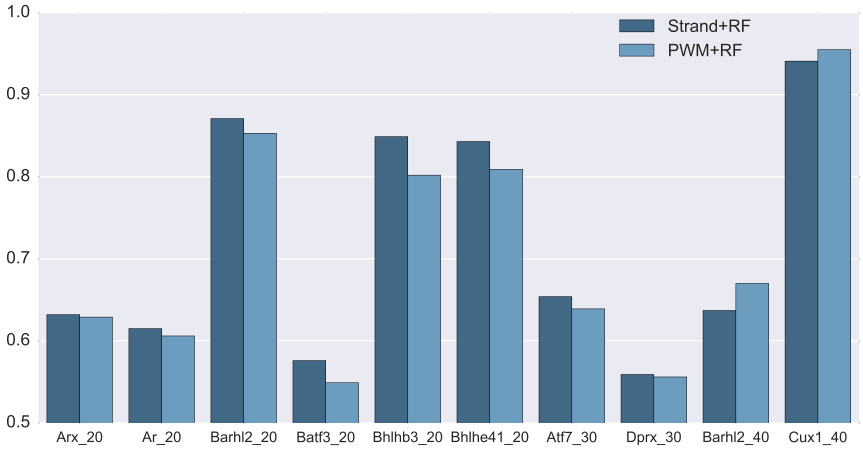
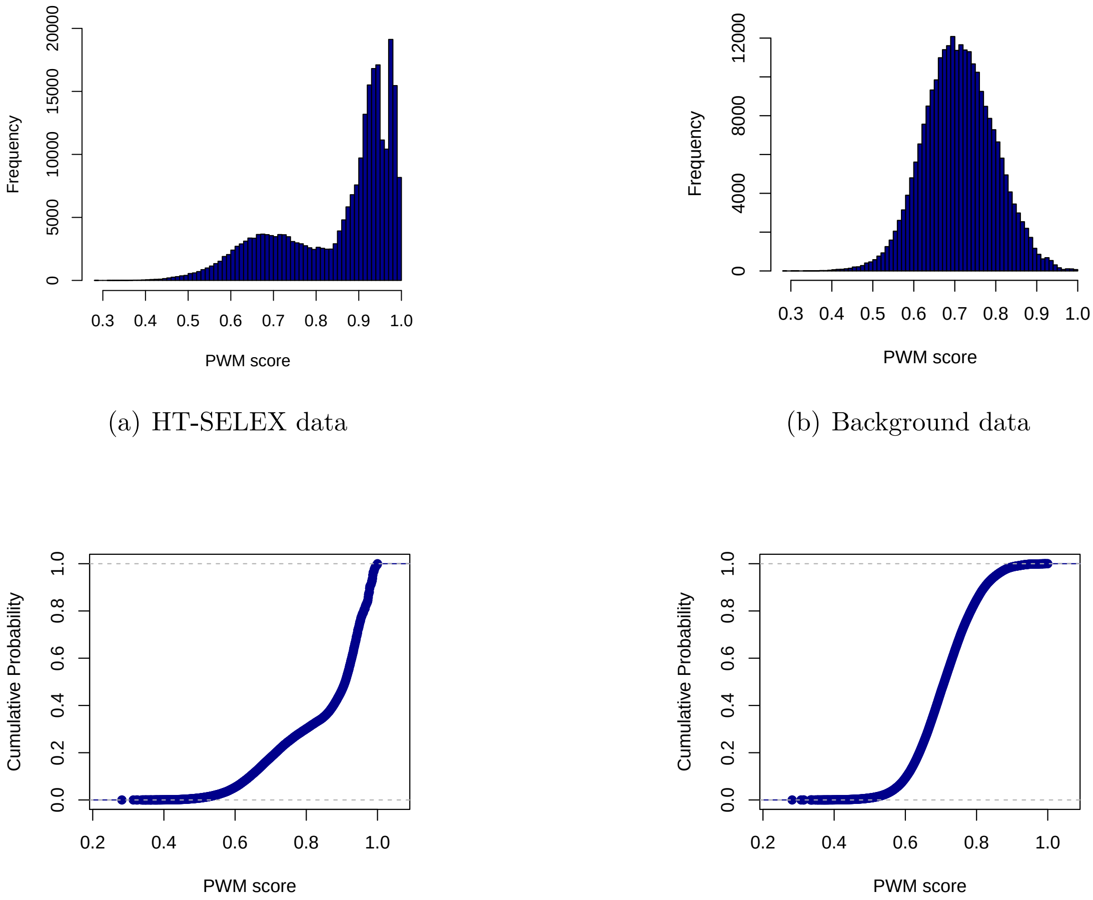


Figure 25: Test set AUC with random forests trained with full length DNA strands and PWM sites for multiple TFs. DNA read length is noted after TF name.

Thus, the Strand+RF model perform better for most data sets. Although, it can be noted that for HT-SELEX measurements with 40 nucleic acids long DNA reads, the extraction of the PWM sites before random forest training is beneficial unlike for HT-SELEX data sets with shorter DNA reads. Thus, it is possible that the excess nucleic acids and variation in binding site positions become significant for random forest performance only after DNA read length exceed 30 nucleic acids. Furthermore, random forest might find information regarding binding specificities outside the PWM sites, which would make modeling with the entire DNA strands beneficial for the HT-SELEX data sets with shorter sequences. Binding site length is assessed more profoundly in the next chapter.

One possible reason for modeling with PWM sites not increasing predictive accuracy in comparison to modeling with the entire DNA sequences, is that the PWM might not find the binding site inside the DNA ligand accurately. If the PWM cannot find the most probable binding site accurately before training the random forest, the forest will be learning DNA features, which do not actually resemble the true differences between binding and non-binding sites. In order to figure out whether this is the case, PWM score histograms on the HT-SELEX sequences and the background sequences for Alx1 are compared to each other. PWM scores are computed with matchPWM R-function such that the maximum score is chosen for each ligand as the final score [42]. In addition, empirical cumulative density functions with the data set that is known to include binding sites and with the shuffled background set are considered. The empirical cumulative density function show the proportion of instances that fall below a certain PWM score. Figure 26 represent the histograms and cumulative density functions.



(a) HT-SELEX data

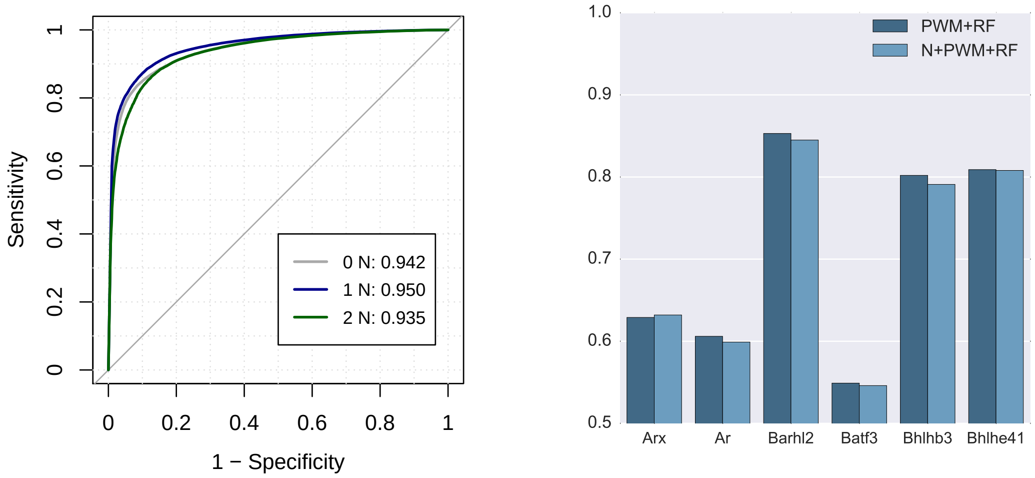
(b) Background data

(c) HT-SELEX data

(d) Background data

Figure 26: The position weight matrix scores and cumulative density functions on Alx1 data set.

It seems that the PWM can distinguish between positive and negative instances quite accurately at least with Alx1. However, for some DNA strands the PWM might choose the wrong site. One explanation for the problem with the position weight matrix is the fact that proteins might bind DNA ligands at their ends so that part of the protein is not attached to the ligand. Figure 24 represent PWM of Alx1 transcriptional factor. For Alx1 the bases farthest apart from the middle point at both ends of the binding site are the most unspecific. Furthermore, the binding sites usually comprise a core with highly specific binding and a few nucleic acids around it with more unspecific binding preferences. Therefore, it is likely that the proteins sometimes bind the DNA ligands at their ends with the core binding motifs so that the more unspecific positions are not attached to the ligand. For example the motif in Figure 24 could bind to a DNA sequence so that the position 2 binds to the first position in the DNA sequence. Thus, utilizing the PWM as it is to find the most probable binding site will not result in the correct binding site. This could be solved by adding letters of N to both ends of the DNA reads before searching the PWM sites (N+PWM+RF). The PWM would then score the N categories with a probability of 0.25. Figure 27 shows ROC-curves when one or two N letters are added to both ends of the strands with the Alx1 data set. In addition, AUC values for other HT-SELEX data sets with 20 nucleic acids long DNA reads, which are padded with one N letters at both ends before finding the PWM site, are represented.



(a) ROC-curves and AUC with DNA strands padded with 0, 1 and 2 N for Alx1.

(b) AUC values with DNA strands padded with 0 and 1 N for multiple TFs.

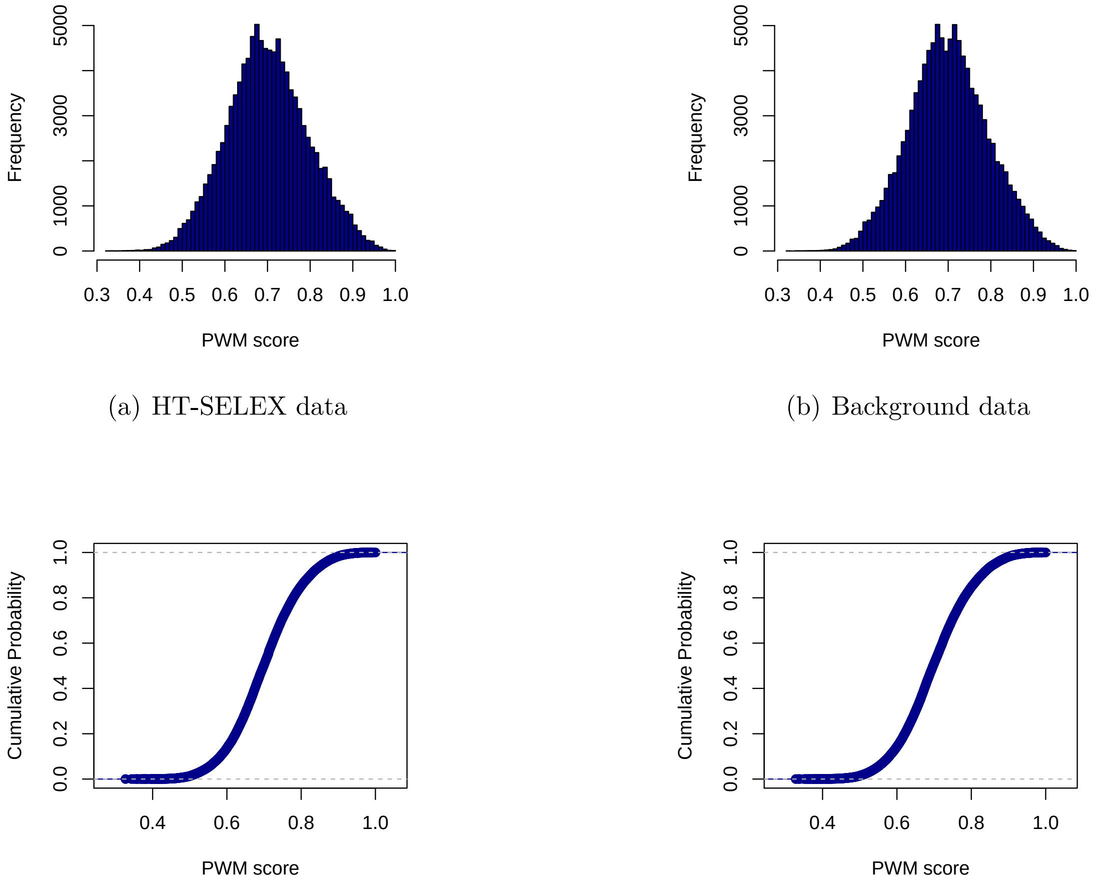
Figure 27: N letters added to both ends of the DNA strands before finding PWM sites (N+PWM+RF).

It can be concluded form Figure 27 that adding N letters to the ends of the DNA ligand might be beneficial for some transcriptional factors although the amount should be kept small. Thus, one nucleic acid at both ends of the position weight matrix do not necessarily affect binding. Although, adding one N categories at both ends of the ligands for modeling binding specificities with HT-SELEX might be beneficial for some transcriptional factors, the effect is not significant and for most of the proteins padding sequences with N letters result in lower modeling accuracy. Perhaps adding a new category of N to {A, C, G, T} disturb random forest performance.

5.1.2.5 Modifications to PWM

Analysis with multiple proteins demonstrate that finding the most probable binding site with position weight matrix before training the random forest does not perform as well as utilizing the full length DNA strand chosen by the same position weight matrix. In this chapter two modifications to PWM+RF model are implemented and modeling accuracy is assessed. First there is a possibility that PWM chooses the binding sites incorrectly. Therefore, a modeling scheme where position weight matrix is optimized simultaneously with random forest is examined. Furthermore, the results in Figure 25 insinuated that the issue with utilizing the PWM site for random forest training might be related to binding site length. Therefore, extending the found PWM sites with surrounding nucleic acids before utilized for training the random forest is considered as well.

Although, it was shown in Figure 26 that the position weight matrix is in fact able to distinguish between the sequences containing a binding site and sequences not containing one, it is likely that for some proteins other than Alx1, the PWM is not able to perform as expected. As an example, the position weight matrix score distributions and empirical cumulative density functions of positive and negative data sets separately are represented for Batf3 in Figure 28. The PWM scores are computed again so that the score for each ligand is the maximum score of all the position scores.



(a) HT-SELEX data

(b) Background data

(c) HT-SELEX data

(d) Background data

Figure 28: Histograms and cumulative density functions of position weight matrix scores for Batf3 data set.

According to Figure 28, position weight matrix scores the HT-SELEX sequences and the shuffled background sequences similarly. Thus, the Batf3 PWM might be bad and it is likely that the PWM finds the most probable binding sites incorrectly. It can be concluded that for Batf3 and possibly for other proteins as well, the position weight matrix is not capable of differentiating the positive and negative DNA reads. Although, one possibility is that Batf3 express so unspecific binding that there might be true binding sites in the negative set as well. However, random forest modeling with the entire DNA strand for Batf3 outperform the random forest with the PWM sites. Thus, it is possible that the position weight matrix simply does not perform as well as it should.

Thus, position weight matrices are probably at least with some HT-SELEX data sets choosing the most probable binding sites from the ligands incorrectly and crucial information for the classifier is lost while non-significant information is preserved. However, this problem could potentially be solved by optimizing the PWM simultaneously with the random forest in the training set. To test this idea, an iterative approach was implemented. After each iteration a new position weight matrix was constructed from the true positive sequences according to confusion matrix of the out-of-bag predictions of the random forest. Figure 29 illustrate how the test set AUC change within ten iterations of tuning PWM for transcriptional factor Barhl1.

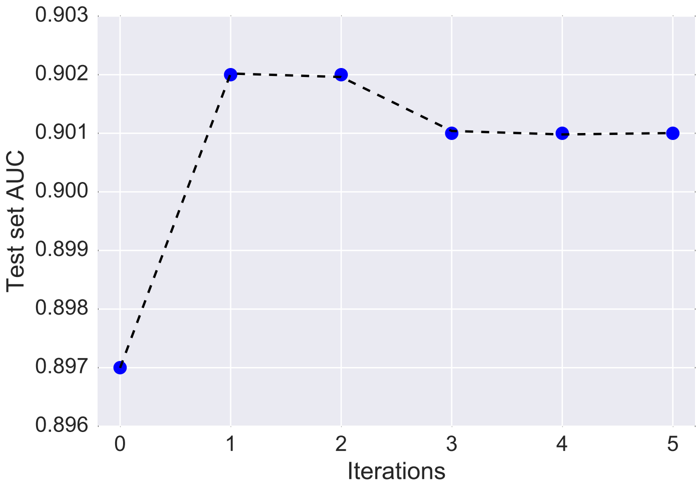
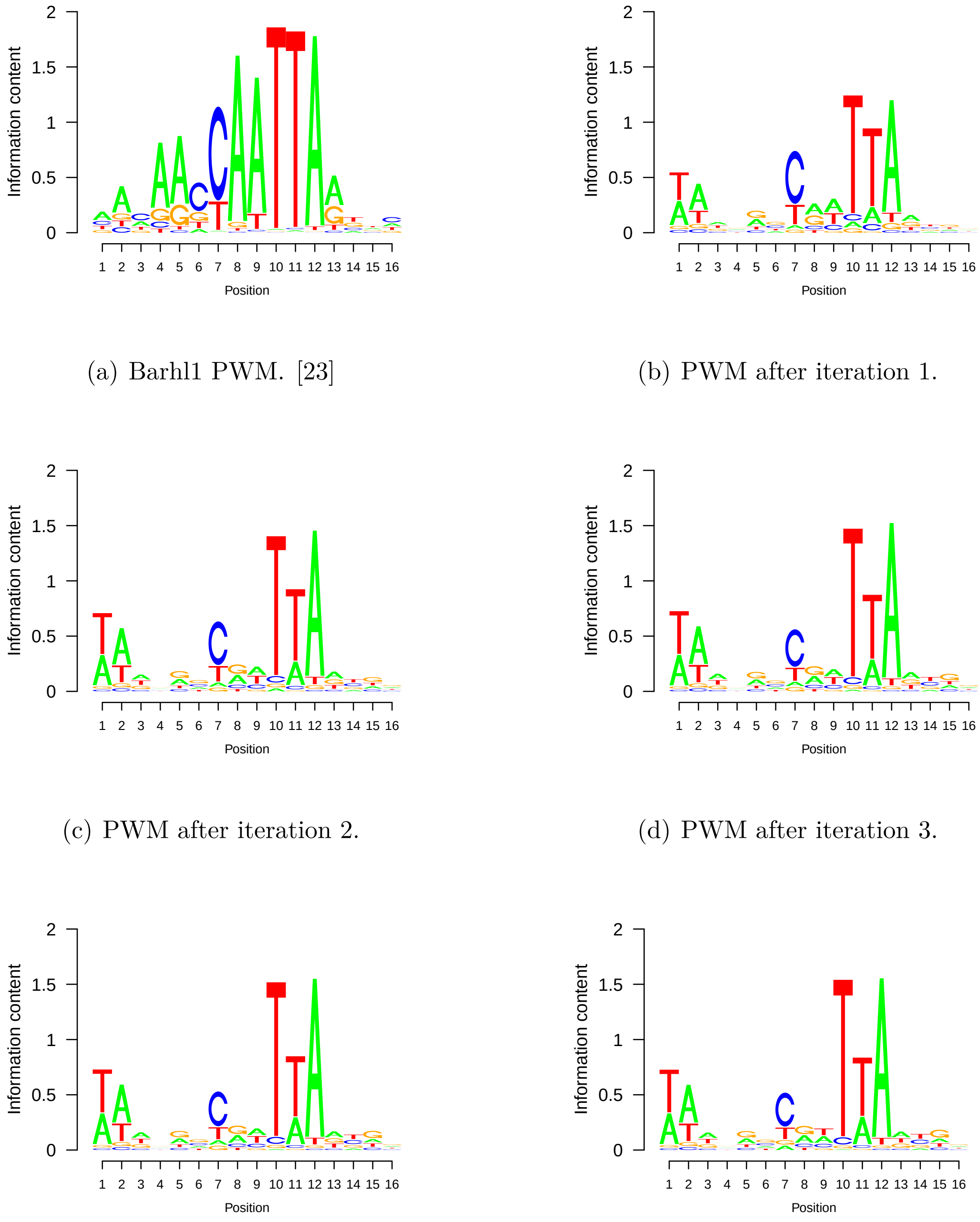


Figure 29: Test set AUC with random forests trained with subsequences chosen by PWM tuned after each iteration for Barhl1.

Thus, position weight matrix modification simultaneously with random forest seems to increase modeling performance, although the effect is moderate. Tuning PWM according to out-of-bag predictions might not be beneficial due to the increase in model training time caused by learning of multiple random forests. Furthermore, the change in PWM after the iterations might reveal something about the binding specificity. The change in Barhl1 position weight matrix after each iteration is represented in Figure 30. It can be seen that the specificity of the PWM decrease especially for some of the positions while a binding preference at the beginning of the PWM emerge. Since the increase in test set AUC is moderate and the increase in running time is significant, the PWM modification simultaneously with random forest training is not considered further.



(a) Barhl1 PWM. [23]

(b) PWM after iteration 1.

(c) PWM after iteration 2.

(d) PWM after iteration 3.

(e) PWM after iteration 4.

(f) PWM after iteration 5.

Figure 30: PWM modification simultaneously with random forest training.

Furthermore, it is possible that the binding sites are longer than the PWM insinuates or that the protein actually binds to the full length DNA sequence. In addition, results in Figure 25 suggest that utilizing the PWM sites for random forest training for HT-SELEX experiments with 40 nucleic acids long DNA reads is beneficial in comparison to modeling with the entire DNA strands. Thus, it is possible that transcriptional factors bind to the entire DNA strands in case of 20 nucleic acids long reads, while the 40 nucleic acids long reads are longer than the binding site. To test this hypothesis, the found PWM sites were extended by one, two or three nucleic acids at both ends for Alx1. Furthermore, if the binding site was found from an end of the DNA ligand then the PWM site was extended with the required amount of N letters so that each binding site utilized for random forest has the same length. In addition, the testing sequences were padded with one, two or three N features accordingly in order for the testing sequences to resemble the training scheme. The model with elongated PWM sites is referred to as PWM+N+RF. Figure 31 shows how adding nucleic acids at both ends of the most probable binding site affects ROC-curve and the AUC-value of Alx1.

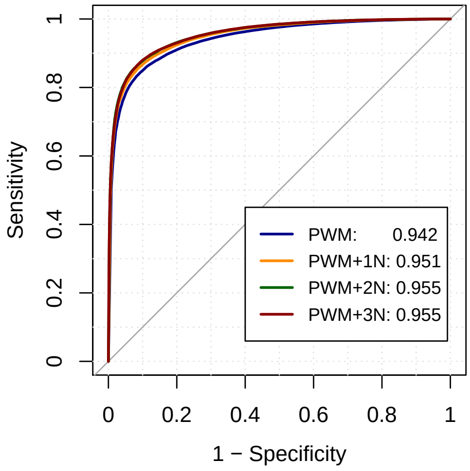
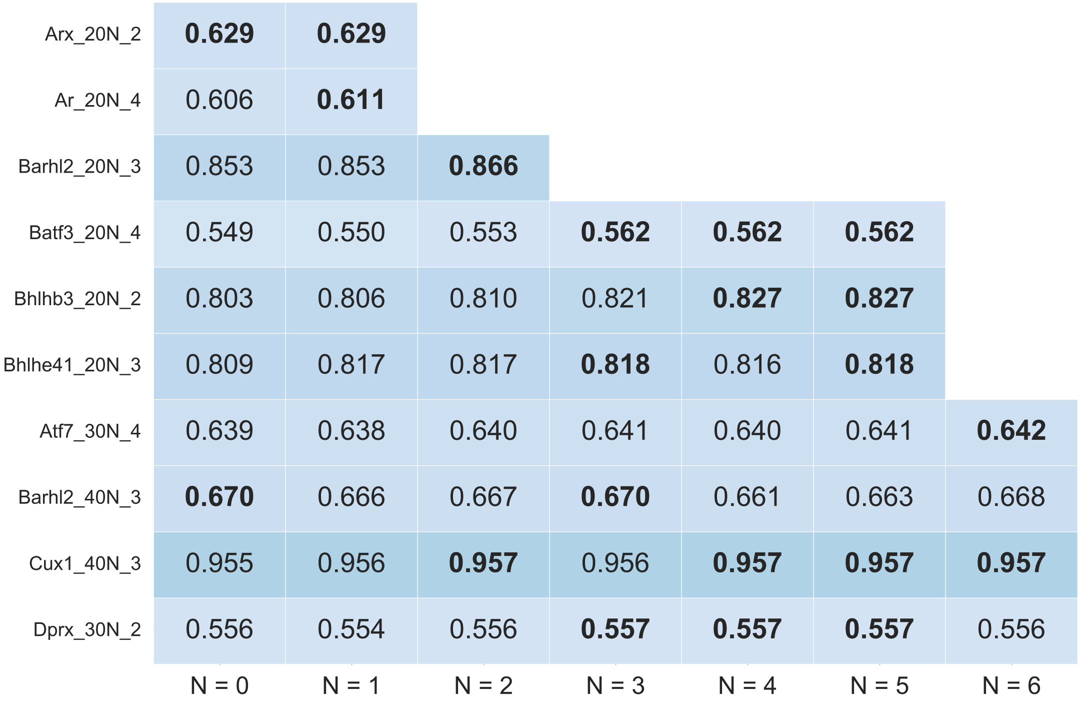


Figure 31: ROC-curve and AUC with PWM+N+RF for Alx1 data set with 0, 1, 2, and 3 N.

Thus, elongating the found PWM sites with surrounding nucleic acids seems to increase modeling accuracy. Perhaps the binding site of Alx1 is longer than the position weight matrix insinuates. The extended PWM sites as it grows to the maximum becomes a centralized version of the entire DNA strand. However, modeling with the entire DNA strand performs still better than modeling with the PWM site only. This might occur, because the addition of multiple N categories decreases information content in spite of the sequence length remaining similar.

Furthermore, since Figure 31 insinuated that extending the found PWM sites will increase modeling performance, it is tested whether extending the PWM sites by more than one nucleotides at both ends increases the model accuracy with other proteins as well. Table 4 summarizes these findings with the same 10 TFs studied previously: Arx, Ar, Barhl2, Batf3, Bhlhb3, Bhlhe41, Atf7, Dprx, Barhl2 and Cux1. The most probable binding sites found with PWM were extended as far as they could without exceeding the length of the sequenced read in the case of 20 nucleic acid long DNA read experiments. For experiments with 30 and 40 nucleic acid long sequences the extension were performed until extension of 6 bases to both ends was reached.



Arx_20N_2	**0.629**	**0.629**					
Ar_20N_4	0.606	**0.611**					
Barhl2_20N_3	0.853	0.853	**0.866**				
Batf3_20N_4	0.549	0.550	0.553	**0.562**	**0.562**	**0.562**	
Bhlhb3_20N_2	0.803	0.806	0.810	0.821	**0.827**	**0.827**	
Bhlhe41_20N_3	0.809	0.817	0.817	**0.818**	0.816	**0.818**	
Atf7_30N_4	0.639	0.638	0.640	0.641	0.640	0.641	**0.642**
Barhl2_40N_3	**0.670**	0.666	0.667	**0.670**	0.661	0.663	0.668
Cux1_40N_3	0.955	0.956	**0.957**	0.956	**0.957**	**0.957**	**0.957**
Dprx_30N_2	0.556	0.554	0.556	**0.557**	**0.557**	**0.557**	0.556
	N = 0	N = 1	N = 2	N = 3	N = 4	N = 5	N = 6

Table 4: Test set AUC for random forest with PWM sites extended by N nucleotides to both directions

Table 4 suggest that elongating the PWM sites to maximum length at least with 20 nucleic acid long HT-SELEX experiments is beneficial. Thus, PWM sites seem to be too short for the studied transcriptional factors. For the experiments with 30 and 40 nucleic acid long DNA reads, the best random forest model is achieved with PWM sites extended with a few nucleic acids to both ends. PWM length for Atf7 is 14 bases, for Barhl2 16, for Cux1 18 and for Dprx 11 bases. Respectively, the length of the subsequences utilized for random forest training after optimum number of nucleic acid extensions to the PWM is 26 for Atf7, 16 or 22 for Barhl2, between 20 and 30 for Cux1 and between 17 and 21 for Dprx. Therefore, it seems that the optimal model is achieved by extending the found PWM site to be a slightly over 20 bases long binding site. The AUC values in Table 4 are achieved with the same random forest parameters as optimized previously for the PWM site with random forest model. However, as the feature space for random forest increased, it was discovered that increasing the amount of variables for a data partitioning at each node, improved model performance slightly. Thus, the number of variables tried at a split N_V will be 30 % of the total amount of features for the final PWM and random forest combination model (PWM+N+RF).

In summary, the best PWM and random forest combination model is achieved by finding the PWM sites inside the ligands and elongating the sites with surrounding nucleic acids. Thus, the PWM and random forest model is implemented by searching for the most probable binding site with PWM, extending the site for 19 or 20 nucleic acids depending whether the PWM has odd or even length and utilizing that information for random forest training. In addition, modeling HT-SELEX experiments with 30 and 40 bases long DNA sequences is performed by extending the PWM site to 23 or 24 nucleic acids long sequences and utilizing these for the random forest.

5.1.2.6 Random forest with subsequences found by decision trees

The transcriptional factor binding sites could also be modeled by utilizing subsequences, that are found by decision trees to be binding sites. The forward and reverse strands are divided into equally sized bags such that the subsequences in each DNA ligand at certain positions belong to the same bag. Furthermore, random forests are trained on each bag. Subsequences in each bag are scored with these random forests excluding the forest trained with the same bag or the corresponding reverse complement bag. The final score for each subsequence is the mean of the assigned scores. The subsequence with the highest score inside each ligand is chosen as the binding site for training the final random forest. In order for this model to function properly the binding should occur evenly on all positions so that the different random forests would predict the binding site accurately. The Inimotif pipeline should choose only experiments with evenly positioned binding sites on both strands [7]. In order to validate this, the start positions of maximum PWM score sites inside the DNA ligands is represented in Figure 32 for Alx1.

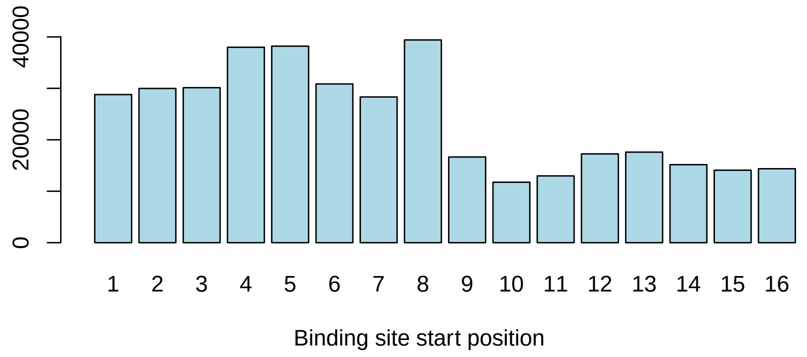


Figure 32: PWM positions for Alx1 indicated as start positions such that both forward and reverse strands are covered.

Thus, Figure 32 shows that according to PWM the transcriptional factor binds quite evenly on all positions at the DNA forward and reverse strands. Alx1 HT-SELEX data set include 20 nucleic acids long DNA reads. Previously, it has been discovered that extending the PWM sites improve modeling accuracy. Thus, for the RF+RF model, 18 nucleic acid long subsequences are considered, which are quite long but sufficient amount of bags would still be formed. Therefore, the procedure yields six different forests, which are trained with $N_V = 40\%$, $N_S = 100$ in order to prevent overfitting and $N_T = 100$ in order to keep the running time moderate. Out of each DNA ligand the subsequence with the maximum score given by these random forests is chosen as the most probable binding site sequence for that read and utilized for training the final random forest. The final forest has parameter values $N_T = 200$, $N_S = 10$ and $N_V = 40\%$. Figure 33 represent the ROC-curve and AUC values for Alx1 data set with RF+RF model in addition to Strand+RF and PWM+N+RF models.

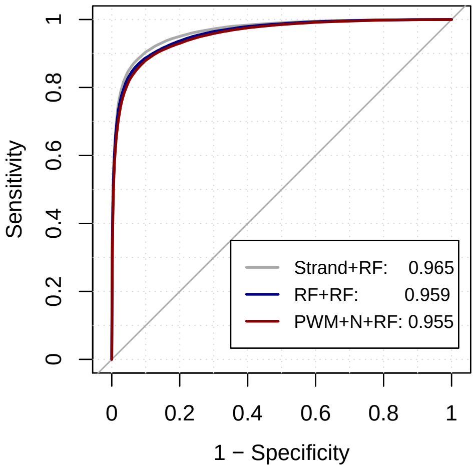


Figure 33: ROC-curve and AUC value with RF+RF model with 18 nucleotide binding site for Alx1.

Thus, at least for Alx1 the random forest trained with the full length DNA strand performs the best. However, the random forest trained with subsequences chosen by decision trees outperforms the random forest trained with the elongated position weight matrix sites. Furthermore, predictive accuracies of Strand+RF, PWM+N+RF and RF+RF models are assessed with the 10 transcriptional factors studied before. For all HT-SELX data sets N_T is set to 200 and N_S to 10. In addition, N_V is set to 40 % of the total number of features for Strand+RF and RF+RF models and to 30 % for PWM+N+RF model. With RF+RF model, 18 nucleic acids long subsequences are considered for constructing the bags with all HT-SELEX data sets with 20 nucleic acids long DNA reads. Furthermore, for data sets with 30 and 40 nucleic acid long DNA sequences, 25 and 32 nucleic acid long binding sites are searched with decision trees correspondingly. The test set AUC values for the 10 transcriptional factor data sets are represented in Figure 34.

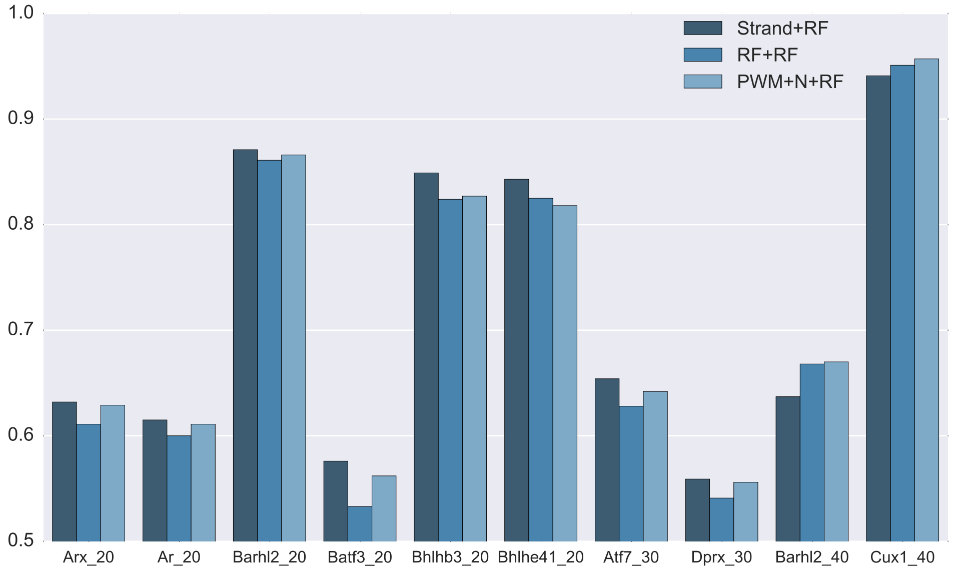


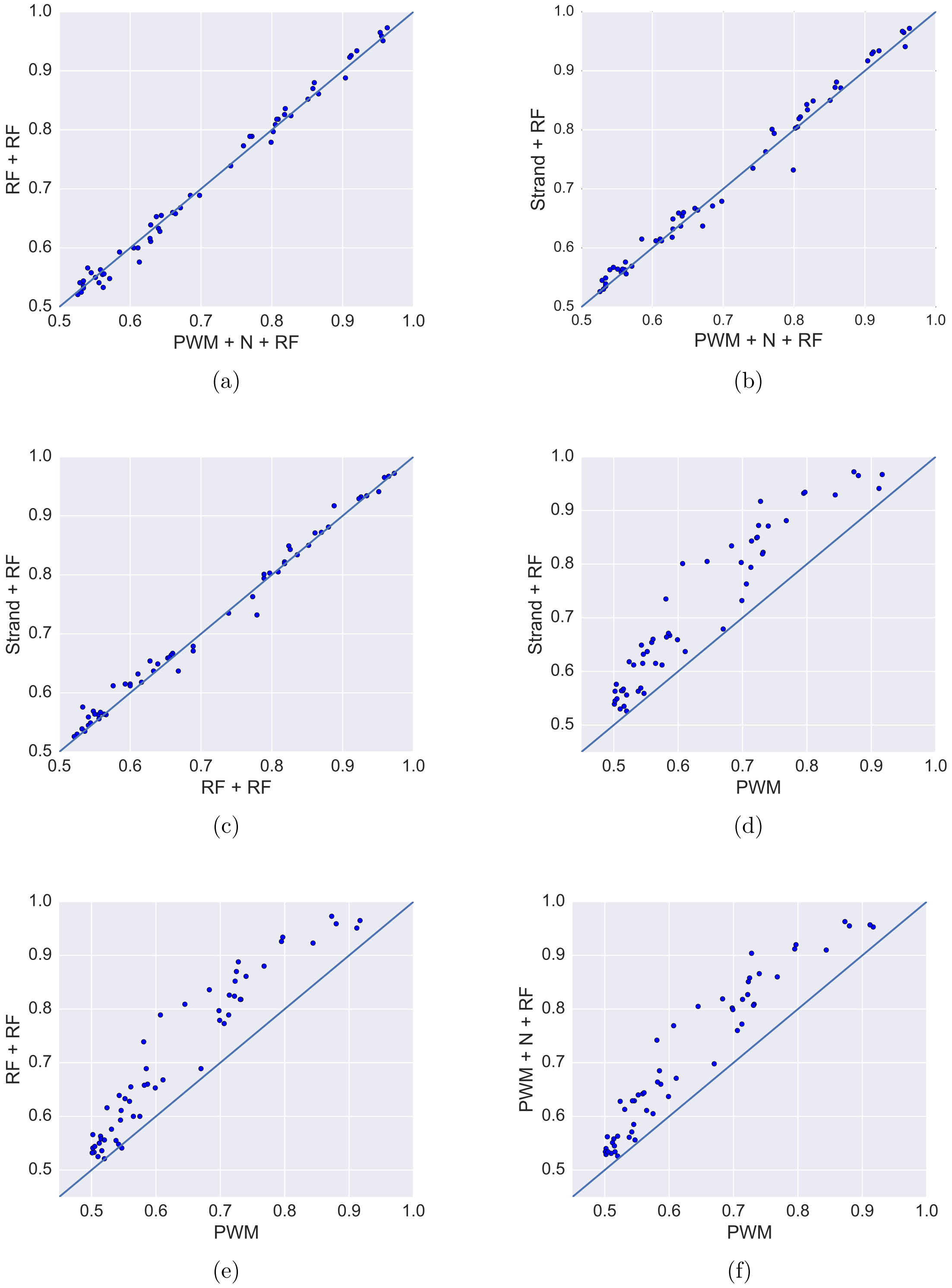
Figure 34: Test set AUC for multiple TFs.

Thus, RF+RF does not perform as well as Strand+RF model. However, RF+RF model predicts binding almost equally to PWM+N+RF. Therefore, three models, Strand+RF, PWM+N+RF and RF+RF, are chosen for further analysis with a more comprehensive data set.

5.1.3 Comparison to DeepBind and position weight matrices

Three random forest models are compared to the neural network model DeepBind and to Jolma et al. PWMs [10, 5]. In addition, the performance of random forest with full length DNA strands (Strand+RF), with elongated PWM sites (PWM+N+RF) and with sites chosen by decision trees (RF+RF), are compared to each other. Some HT-SELEX measurements were conducted on 14 bases long DNA sequences. For these data sets, PWM sites are extended to 13 or 14 nucleic acids for the PWM+N+RF model and RF+RF model is implemented by searching for 12 bases long binding sites. In addition, scoring sequences with only the Jolma et al. PWMs is compared to the random forest models. PWM scoring (PWM) was performed by scoring all positions on a ligand with PWM and averaging these scores to give final score for the ligand. In addition, DNA strands were padded with four 'N' features at both ends, because this improved model performance. If Jolma et al. published multiple PWMs only the best AUC is reported. It was discovered by Alipanahi et al. that padding with 'N' improved PWM model performance and that taking the mean score of all position scores for a ligand performed better than utilizing the maximum score [10].

Experiments are conducted with 55 HT-SELEX data sets in order to carry out the comparison. Figure 35 represent test set AUC values with all the 55 data sets.

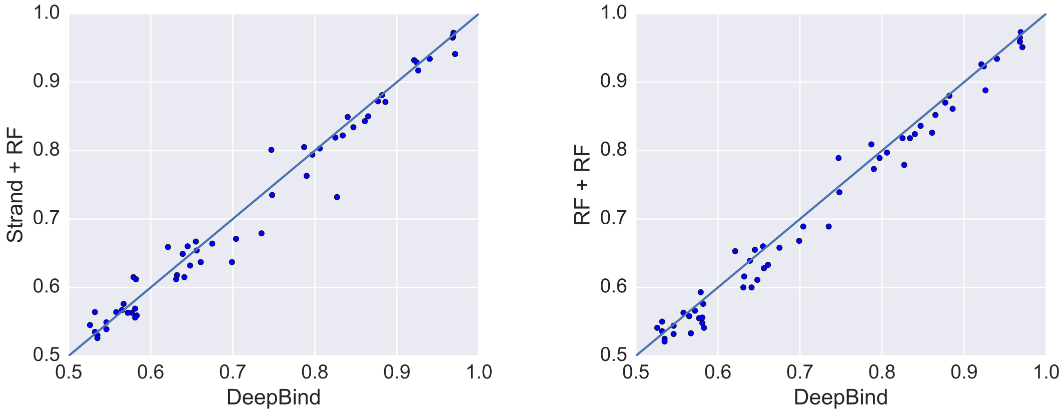


(a) (b) (c) (d) (e) (f)

Figure 35: AUC comparison of PWM, PWM+N+RF, RF+RF and Strand+RF

PWM+N+RF and RF+RF models perform with almost equal accuracy and AUC values do not differ significantly (p-value = 0.46, two-sided Wilcoxon signed rank test, n = 55). However, Strand+RF model has higher test AUC (0.713) than the two other random forest models. Furthermore, the difference is statistically significant in comparison to RF+RF model (average = 0.708, p-value = 0.00030, two-sided Wilcoxon signed rank test, n = 55) and PWM+N+RF model (average = 0.707, p-value = 0.00027, two-sided Wilcoxon signed rank test, n = 55). All three random forest models perform significantly better than scoring sequences with only the PWMs. For instance Strand+RF model has significantly higher AUC than scoring with PWMs (average=0.633, p-value = $1.14 * 10^{-10}$, two-sided Wilcoxon signed rank test, n = 55). From Figure 35 it can be seen that PWM+N+RF model perform slightly better for HT-SELEX data sets that yield lower AUC values in general, while the RF+RF model predict higher AUC values even more accurately.

Furthermore, comparison of the random forest models to DeepBind is conducted. DeepBind software can be utilized for testing new sequences since Alipanahi et al. offer the trained motifs in addition to the neural network package for scoring the sequences [10]. However, the motifs have been trained partly with the same HT-SELEX sequences that are scored in this thesis. Therefore, some of the DeepBind AUC values that are reported here are possibly overestimates of the DeepBind model performance. DeepBind achieved higher test set AUC (mean AUC = 0.719) than the random forest models, RF+RF (average = 0.708, p-value = $5.79 * 10^{-5}$, two-sided Wilcoxon signed rank test, n = 55), PWM+N+RF (average = 0.707, p-value = $3.53 * 10^{-6}$, two-sided Wilcoxon signed rank test, n = 55) and Strand+RF (average = 0.713, p-value = 0.041, two-sided Wilcoxon signed rank test, n = 55). Figure 36 represent how RF+RF and Strand+RF compare to DeepBind.



(a) (b)

Figure 36: AUC comparison of RF+RF and Strand+RF to DeepBind.

Increasing modeling complexity of TF binding specificities with random forest is beneficial in comparison to utilizing only PWMs. Furthermore, DeepBind outperform the different random forest models. However, a final random forest model can be constructed differently for different types of HT-SELEX data sets. For measurements with 14, 20 or 30 nucleic acids long DNA reads, the random forest can be trained with full length DNA strands, while HT-SELEX data sets with 40 nucleic acids long DNA reads can be trained with elongated PWM sites. Test set AUC values for this random forest model are represented in Figure 37 in comparison to DeepBind.

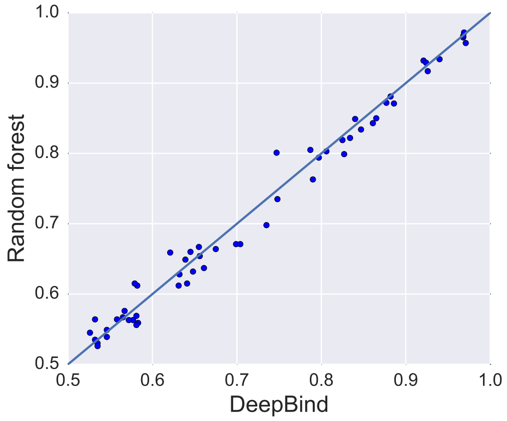


Figure 37: AUC comparison of final random forest and DeepBind.

DeepBind with mean AUC of 0.719 outperforms final random forest model (average=0.716, p-value = 0.063, two-sided Wilcoxon signed rank test, n = 55). However, means do not differ statistically significantly. Therefore, a relatively competitive model can be provided with random forest also for TF pairs whose binding specificities have been measure with CAP-SELEX. Figure 38 summarizes test set AUC values for the final random forest model, DeepBind and PWM. The TFs are sorted according to test set AUC values computed with random forest.

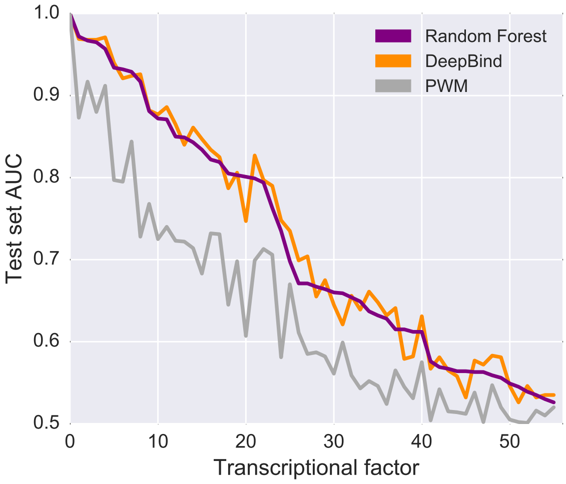


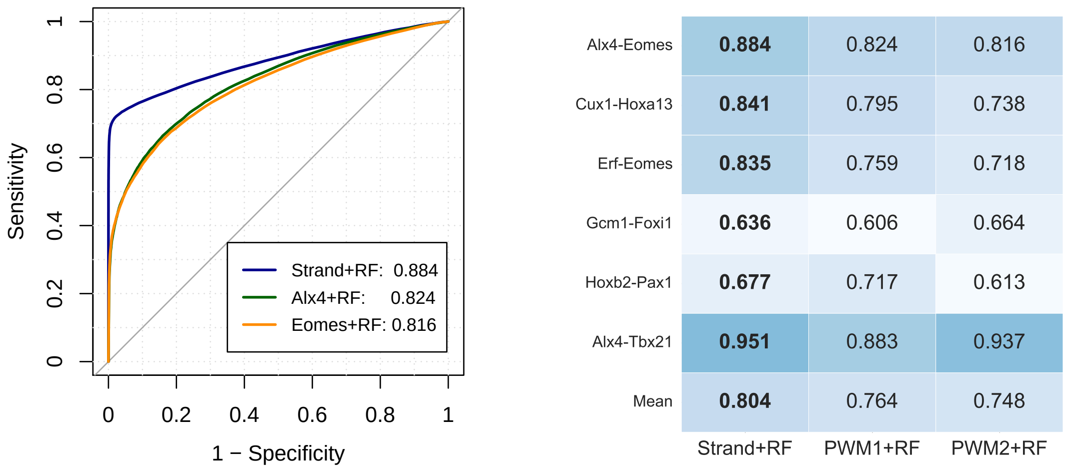
Figure 38: Test set AUC of 55 TFs.

5.2 Random forest for CAP-SELEX

In this chapter random forest is applied for modeling transcriptional factor pair binding specificities measured with CAP-SELEX by Jolma et al. in [4]. The sequenced DNA reads in CAP-SELEX are 40 nucleic acids long. The experimental schemes that performed the best for individual TF motifs measured with HT-SELEX are utilized for TF pair modeling as well. The models include training random forest with the full length DNA strand (Strand+RF) and training with the individual transcriptional factor PWM sites (PWM1+RF and PWM2+RF). Furthermore, Jolma et al. published position weight matrices for the TF pair combinatorial binding sites [4]. Thus, random forests are trained additionally with the binding sites found with these combinatorial PWMs (PWM+RF) and extending the found sites by one nucleic acids at both ends is considered as well (PWM+N+RF). Random forests are also trained with subsequences inside the CAP-SELEX DNA ligands found by decision trees trained in bags containing subsequences starting at different positions (RF+RF). Transcriptional factor pairs can in addition be modeled by considering each possible orientation and gap between the two individual TF binding motifs. Random forests can be trained with subsequences inside the DNA ligands found with PWMs, which combine the individual TF motifs according to the best spacings (PWM1+PWM2+RF). Furthermore, binding sites could be found by choosing the maximum score position with the other PWM and extending the chosen site to the other direction in order to cover binding site of the other TF as well (PWM1+PWM2+N+RF). Since the direction to which the site should be extended is not known, two random forest are trained separately with both extension directions. Test DNA sequences are scored with both random forests and the average score is assigned to each ligand. Models for CAP-SELEX except the combinatorial PWM elongation model were summarized in Table 3. Random forest parameters are as optimized for HT-SELEX data in the previous chapter. For all CAP-SELEX experiments $N_S = 10$ and $N_T = 200$. For the forests trained with entire DNA strand (Strand+RF) or subsequences chosen by decision trees (RF+RF) N_V is 40 % of total number of features, while for the forests that are trained on subsequences chosen by PWMs N_V is 30 % out of total amount of features. The results are evaluated with a case study with Alx4-Eomes transcriptional factor pair and five other TF pairs including Cux1-Hoxa13, Erf-Eomes, Gcm1-Foxi1, Hoxb2-Pax1 and Alx4-Tbx21. The models are trained on CAP-SELEX data sets selected by Jolma et al. in [4]. Thus, these data sets passed the quality control pipeline and were utilized for PWM construction. In addition, results with 50 CAP-SELEX data set are represented for different random forest models and compared to each other and scoring with only the combinatorial PWMs.

5.2.1 Modeling with full length DNA reads and individual PWM sites

Random forest modeling is first performed by modeling with the full length DNA strands, which according to the Jolma et al. combinatorial PWM include the binding site (Strand+RF). In addition, performance of random forests trained with the individual TF binding sites are evaluated (PWM1+RF and PWM2+RF). The position weight matrices published by Jolma et al. derived from HT-SELEX measurements were utilized [5]. ROC-curves and AUC values for Alx4-Eomes with these three random forest models in addition test set AUC values obtained with five other CAP-SELEX data sets are represented in Figure 39.



	Strand+RF	PWM1+RF	PWM2+RF
Alx4-Eomes	**0.884**	0.824	0.816
Cux1-Hoxa13	**0.841**	0.795	0.738
Erf-Eomes	**0.835**	0.759	0.718
Gcm1-Foxi1	**0.636**	0.606	0.664
Hoxb2-Pax1	**0.677**	0.717	0.613
Alx4-Tbx21	**0.951**	0.883	0.937
Mean	**0.804**	0.764	0.748

(a) Alx4-Eomes

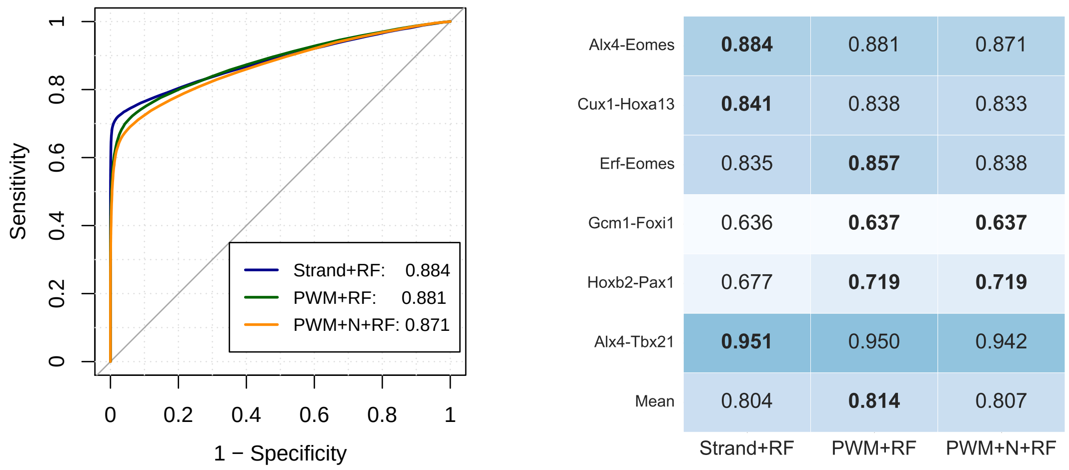
(b) AUC values with 6 TF pairs

Figure 39: Random forest model with DNA strand and individual PWM sites

Thus, the random forest model with the full length DNA strand seems to perform already quite accurately. Although the true binding site is likely to be shorter than the 40 nucleic acid long strand, the random forest is able to reveal binding specificities inside the strand despite variation in binding site position. Modeling with the entire DNA strands perform better than utilizing the individual TF binding motifs for random forest learning as could have been expected. The models trained with only one transcriptional factor PWM site loose information about the other PWM site which should be present on the ligand. Random forest with the entire DNA strand on the other hand can utilize the information of both of the binding motifs even though the excess nucleic acids and differences in binding site positions may disturb the random forest.

5.2.2 Modeling with combinatorial PWM sites

Jolma et al. derived combinatorial position weight matrices for transcriptional factor pairs in [4]. These position weight matrices can be utilized for finding the most probable binding sites inside DNA ligands, and training the forest with them. Furthermore, it was discovered in HT-SELEX experiments that extending the found PWM sites to both directions and utilizing these longer PWM sites for random forest training increased modeling performance. Thus, it is tested for the six TF pairs whether extending the combinatorial position weight matrix sites by one nucleic acid to both directions increases modeling accuracy. ROC-curves and AUC values for Alx4-Eomes and test set AUC values for five other TF pair data sets are represented in Figure 40 when modeling is perfromed with full length DNA strands and with subsequence chosen by the combinatorial PWM (PWM+RF). In addition, performance of PWM+N+RF model with the elongated PWM sites is assessed.



	Strand+RF	PWM+RF	PWM+N+RF
Alx4-Eomes	**0.884**	0.881	0.871
Cux1-Hoxa13	**0.841**	0.838	0.833
Erf-Eomes	0.835	**0.857**	0.838
Gcm1-Foxi1	0.636	**0.637**	**0.637**
Hoxb2-Pax1	0.677	**0.719**	**0.719**
Alx4-Tbx21	**0.951**	0.950	0.942
Mean	0.804	**0.814**	0.807

(a) Alx4-Eomes

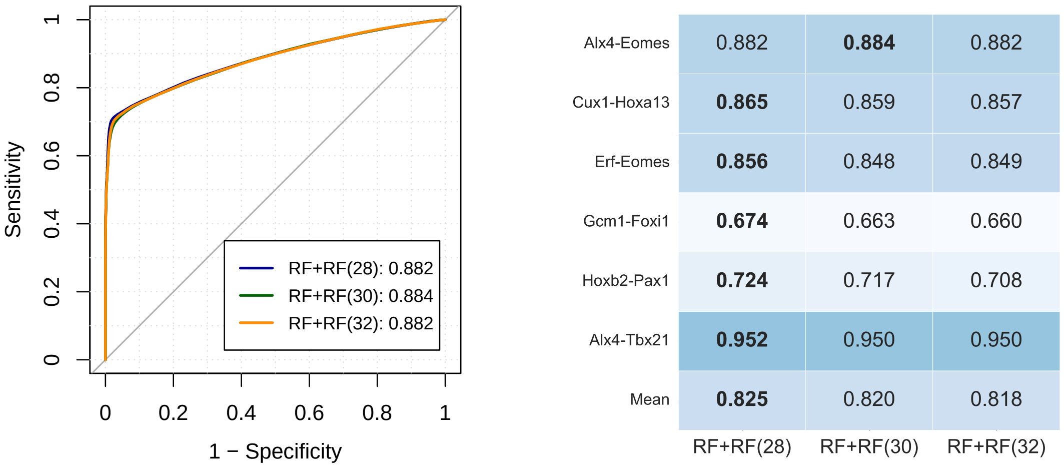
(b) AUC values with 6 TF pairs

Figure 40: Random forest model with TF1-TF2 combinatorial PWM sites

Utilizing PWM sites for training the random forest outperforms learning with the entire DNA strand even though for Alx4-Eomes these two models predict binding with almost equal accuracy. The DNA reads are 40 nucleic acids long in CAP-SELEX measurements, while the TF pair binding motifs are usually about 20 nucleic acids long [4]. Thus, with the Strand+RF model random forest receives nucleic acid features that do not concern binding, increasing randomness in the training data and complicating learning. In addition, with the PWM+RF model, random forest should receive nucleotide acids at specific positions on the binding motif always as certain variables, which should aid random forest in classification. Furthermore, for CAP-SELEX data it seems that extending the found PWM sites do not increase modeling accuracy. Thus, PWM length seems to be sufficient for TF pairs.

5.2.3 Random forest model

Random forest modeling could also be conducted by training with subsequences chosen by decision trees to be the binding site inside each DNA ligand. DNA ligands are divided into equally sized bags and random forests with $N_T = 100$ and $N_S = 100$ are trained on each of the bags. Furthermore, all subsequences are scored with those random forests that were not trained on the bag that contained the subsequence or the bag with the reverse complement of the sequence. The final score for a subsequence is then the average of the scores assigned to it. The subsequence with the highest score inside each ligand is chosen as the binding site for training the final random forest with $N_T = 200$ and $N_S = 10$. For CAP-SELEX data subsequence lengths of 28, 30 and 32 bases are tested. Figure 41 represent the ROC-curves with the different subsequence length RF+RF models for Alx4-Eomes in addition to test set AUC values for five other transcriptional factor pair data sets.



	RF+RF(28)	RF+RF(30)	RF+RF(32)
Alx4-Eomes	0.882	**0.884**	0.882
Cux1-Hoxa13	**0.865**	0.859	0.857
Erf-Eomes	**0.856**	0.848	0.849
Gcm1-Foxi1	**0.674**	0.663	0.660
Hoxb2-Pax1	**0.724**	0.717	0.708
Alx4-Tbx21	**0.952**	0.950	0.950
Mean	**0.825**	0.820	0.818

(a) Alx4-Eomes

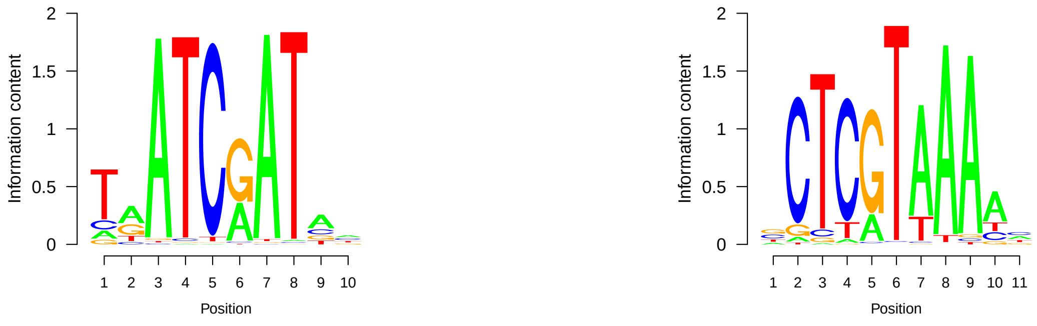
(b) AUC values with 6 TF pairs

Figure 41: Test set AUC with RF+RF models.

The RF+RF model performs quite well for TF pair binding site modeling. In fact, the model outperforms modeling with the entire DNA strands and modeling with the subsequences chosen by combinatorial PWMs. For Alx4-Eomes the optimal length for the subsequences and thus for the bags is 30 nucleic acids. However, when multiple TF pairs are considered it can be seen that the optimal length for subsequences is 28 nucleic acids.

5.2.4 Modeling with binding sites of both transcriptional factors

Individual PWMs could also be combined in an optimal manner to search for binding sites inside the DNA ligands and utilizing the found subsequences for random forest training. Thus, the two position weight matrices are attached to each other according to the knowledge about the orientation of the two transcriptional factors on the DNA ligand and the amount of nucleic acid gaps between their motifs. It is assumed that the TFs bind DNA according to a fixed spacing. According to Jolma et al. those transcriptional factor pairs that bind DNA with overlapping motif have fixed spacings between the motifs and those pairs that bind DNA with multiple gaps between the two motifs have more relaxed binding preferences [4]. Since the true spacing between the transcriptional factor binding motifs is unknown, the two PWMs are attached to each other with all four possible orientations and amounts of gaps between them. In this Master's thesis the amount of gaps is assumed to range between negative 10 to positive 5 between the individual motifs. Position weight matrices are combined according to all four possible transcriptional factor pair orientations, which are represented in Figure 15. Spacings are demonstrated with Cux1-Hoxa13 transcriptional factor pair. Figure 42 represent the Cux1 and Hoxa13 position weight matrices.

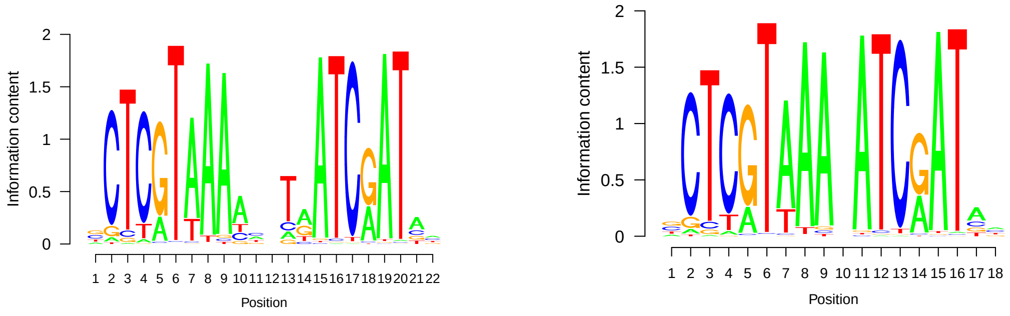


(a) Cux1 PWM from [5] and visualized with information content.

(b) Hoxa13 PWM from [5] and visualized with information content.

Figure 42: Cux1 and Hoxa13 PWMs

Therefore the PWM combinations become Cux1-Hoxa13, Hoxa13-Cux1, Cux1-reverse(Hoxa13) and Hoxa13-reverse(Cux1), where reverse refers to the reverse complement of a PWM, which is thus able to search for the binding motif when the two TFs are bound to different DNA strands. Furthermore, positive gaps are added between the PWMs with [0.25, 0.25, 0.25, 0.25] columns while negative gaps are obtained by overlapping the two PWMs. If the negative spacing is uneven, the PWMs are overlapped with an additional nucleotide acid and a [0.25, 0.25, 0.25, 0.25] column is added between the two motifs. Figure 43 represent two examples of Cux1 and Hoxa13 PWM attachments with orientation Hoxa13-Cux1.

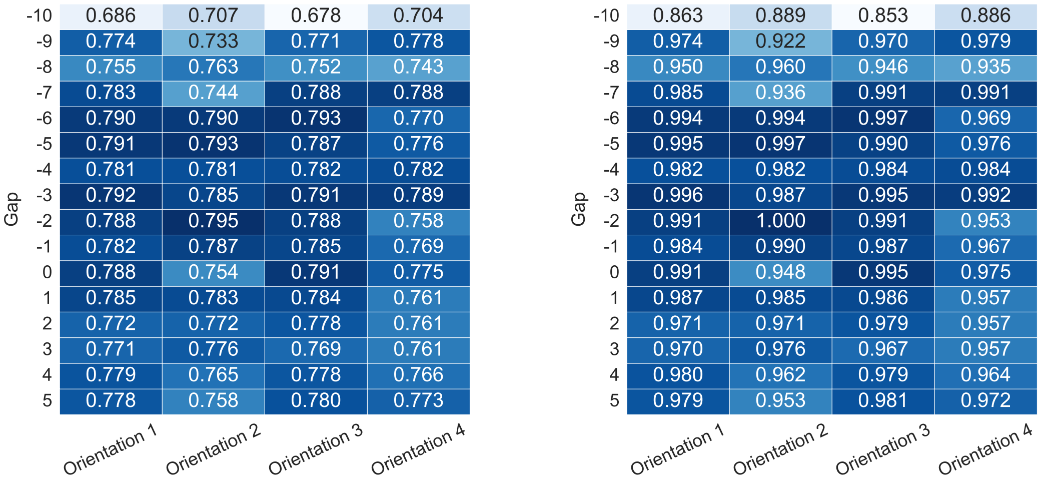


(a) Gap 1

(b) Gap -3

Figure 43: Two possible spacings of Cux1 and Hoxa13 in orientation Hoxa13-Cux1.

The training data set is divided into training and validation sets such that 20 % of the training set will belong to validation set. Random forests are trained in the training set with subsequences searched from the DNA ligands with all possible spacings separately. Furthermore, the performance of the forests trained with different spacings is assessed in the validation data set. Figure 44 represent the validation set AUC values for Cux1-Hoxa13 and the percentage of each spacing model from the spacing that yielded the maximum AUC value.



Gap	Orientation 1	Orientation 2	Orientation 3	Orientation 4
-10	0.686	0.707	0.678	0.704
-9	0.774	0.733	0.771	0.778
-8	0.755	0.763	0.752	0.743
-7	0.783	0.744	0.788	0.788
-6	0.790	0.790	0.793	0.770
-5	0.791	0.793	0.787	0.776
-4	0.781	0.781	0.782	0.782
-3	0.792	0.785	0.791	0.789
-2	0.788	0.795	0.788	0.758
-1	0.782	0.787	0.785	0.769
0	0.788	0.754	0.791	0.775
1	0.785	0.783	0.784	0.761
2	0.772	0.772	0.778	0.761
3	0.771	0.776	0.769	0.761
4	0.779	0.765	0.778	0.766
5	0.778	0.758	0.780	0.773

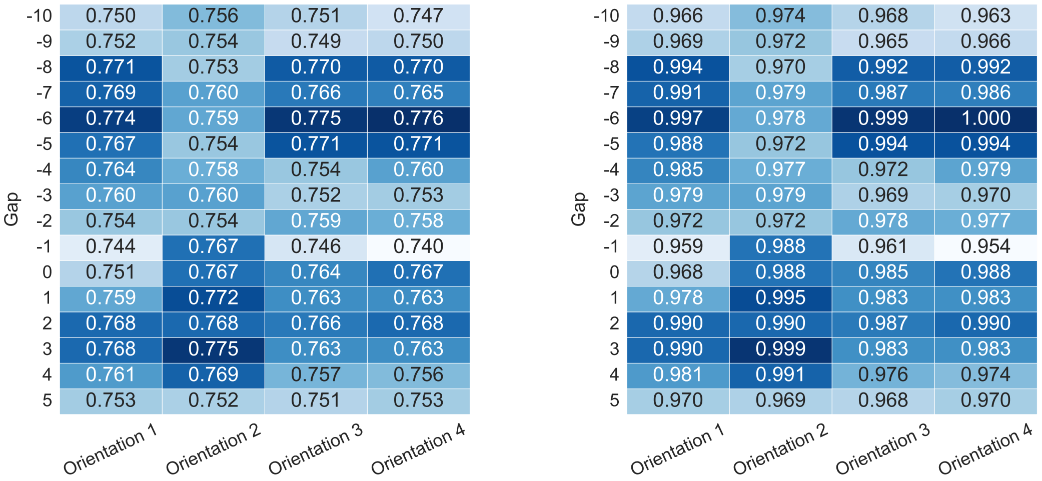
(a) Validations set AUC

Gap	Orientation 1	Orientation 2	Orientation 3	Orientation 4
-10	0.863	0.889	0.853	0.886
-9	0.974	0.922	0.970	0.979
-8	0.950	0.960	0.946	0.935
-7	0.985	0.936	0.991	0.991
-6	0.994	0.994	0.997	0.969
-5	0.995	0.997	0.990	0.976
-4	0.982	0.982	0.984	0.984
-3	0.996	0.987	0.995	0.992
-2	0.991	1.000	0.991	0.953
-1	0.984	0.990	0.987	0.967
0	0.991	0.948	0.995	0.975
1	0.987	0.985	0.986	0.957
2	0.971	0.971	0.979	0.957
3	0.970	0.976	0.967	0.957
4	0.980	0.962	0.979	0.964
5	0.979	0.953	0.981	0.972

(b) Difference in percent of the maximum AUC

Figure 44: Cux1-Hoxa13 validation set AUC

Thus, random forest can quite accurately predict binding with most of the spacings. A similar analysis is represented for Alx4-Eomes in Figure 45.



Gap	Orientation 1	Orientation 2	Orientation 3	Orientation 4
-10	0.750	0.756	0.751	0.747
-9	0.752	0.754	0.749	0.750
-8	0.771	0.753	0.770	0.770
-7	0.769	0.760	0.766	0.765
-6	0.774	0.759	0.775	0.776
-5	0.767	0.754	0.771	0.771
-4	0.764	0.758	0.754	0.760
-3	0.760	0.760	0.752	0.753
-2	0.754	0.754	0.759	0.758
-1	0.744	0.767	0.746	0.740
0	0.751	0.767	0.764	0.767
1	0.759	0.772	0.763	0.763
2	0.768	0.768	0.766	0.768
3	0.768	0.775	0.763	0.763
4	0.761	0.769	0.757	0.756
5	0.753	0.752	0.751	0.753

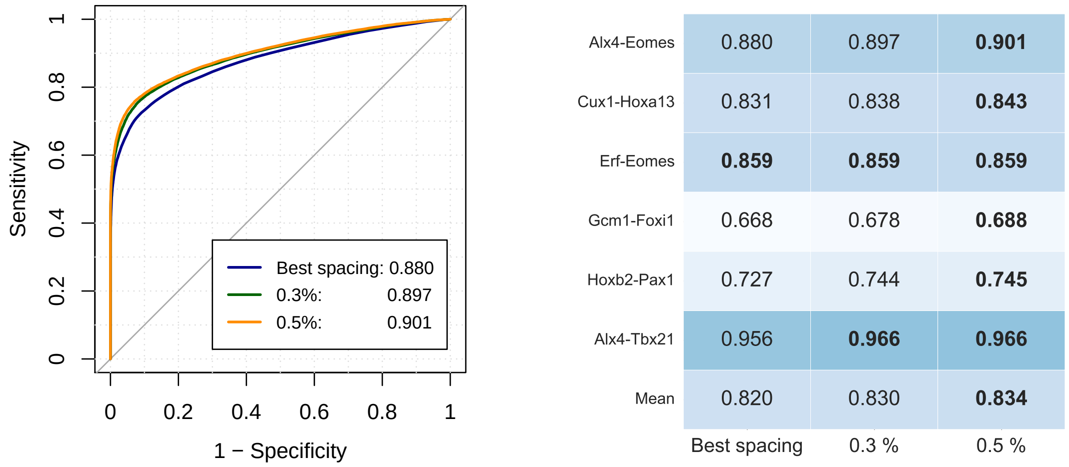
(a) Validations set AUC

Gap	Orientation 1	Orientation 2	Orientation 3	Orientation 4
-10	0.966	0.974	0.968	0.963
-9	0.969	0.972	0.965	0.966
-8	0.994	0.970	0.992	0.992
-7	0.991	0.979	0.987	0.986
-6	0.997	0.978	0.999	1.000
-5	0.988	0.972	0.994	0.994
-4	0.985	0.977	0.972	0.979
-3	0.979	0.979	0.969	0.970
-2	0.972	0.972	0.978	0.977
-1	0.959	0.988	0.961	0.954
0	0.968	0.988	0.985	0.988
1	0.978	0.995	0.983	0.983
2	0.990	0.990	0.987	0.990
3	0.990	0.999	0.983	0.983
4	0.981	0.991	0.976	0.974
5	0.970	0.969	0.968	0.970

(b) Difference in percent of the maximum AUC

Figure 45: Alx-Eomes validation set AUC

Alx4-Eomes seem to express more specific binding preferences than Cux1-Hoxa13. However, it seems that random forest can perform quite well with all possible spacings even though binding is assumed to occur with only certain orientations and gap configurations. However, random forest can already predict binding relatively accurately when trained with only the subsequences search with the other PMW. Even when searching subsequences with incorrect PWM spacings, the different spacing models include the other PWM site and at least sometimes by change the extension to the correct direction. Thus, there is probably crucial binding motif information in subsequences searched with different PWM spacings even though they were possibly incorrect. Therefore, most probably the correct spacing between the two transcriptional factors is given by the random forest with the highest validation set AUC. However, the spacings which yield AUC values close to the forest with the most probably correct spacing should perhaps be considered as well in the final model. The spacings may be combined by scoring DNA ligands in test set with multiple random forests trained with subsequences chosen by the best motif spacings. Furthermore, a DNA ligand would get a final score through averaging over the various random forest scores or by selecting the maximum score to be the final score. It was discovered that averaging produced higher test set AUC values. Experiments are conducted with three settings. First only the best spacing between the TF motifs chosen in the validation data set is utilized for scoring test set DNA ligands with the corresponding random forest. Then random forests with motif spacings of at most 0.3 % difference in validation set AUC to the the best spacing are utilized for scoring test ligands and finally all forests within 0.5 % of difference to the best spacing. Thus, the model becomes an ensemble of random forests. Test set AUC values for six TF pair data sets and ROC-curves for Alx4-Eomes are represented in Figure 46.



	Best spacing	0.3 %	0.5 %
Alx4-Eomes	0.880	0.897	**0.901**
Cux1-Hoxa13	0.831	0.838	**0.843**
Erf-Eomes	**0.859**	**0.859**	**0.859**
Gcm1-Foxi1	0.668	0.678	**0.688**
Hoxb2-Pax1	0.727	0.744	**0.745**
Alx4-Tbx21	0.956	**0.966**	**0.966**
Mean	0.820	0.830	**0.834**

(a) Alx4-Eomes

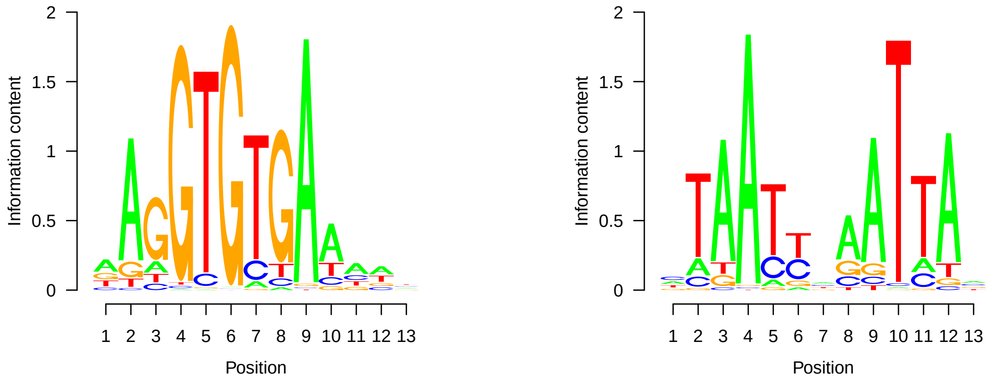
(b) AUC values of 6 TF pairs

Figure 46: Test set AUC with PWM1+PWM2+RF.

Thus, a higher classification accuracy is achieved by averaging the scores given by multiple random forests trained with different motif spacings than by considering only the best spacing between the TF motifs. Furthermore, the ensemble model performs better with more different spacings considered. It could be possible to achieve even higher test set AUC values if more random forests were added to the ensemble. However, this would increase running time significantly.

5.2.5 Modeling by utilizing binding specificity of one protein

A model that considers all the possible gaps at once between the two transcriptional factor motifs could be constructed as well. This model searches the maximum PWM score sites with only the other transcriptional factor PWM such that the matrix is extended with N = [0.25, 0.25, 0.25, 0.25] columns to the extent of 25 positions long PWM. However, the extension is performed only to the other direction to which the binding of the other motif could have occurred on the forward or the reverse strands. Thus, two random forest are trained with both of the possible orientations separately. Test set DNA ligands are then scored with both of the random forests and the final score for the ligand is obtained by averaging the two scores or by choosing the maximum score. Figure 47 shows Alx4 and Eomes position weight matrices for which the PWM extensions are represented later as an example.

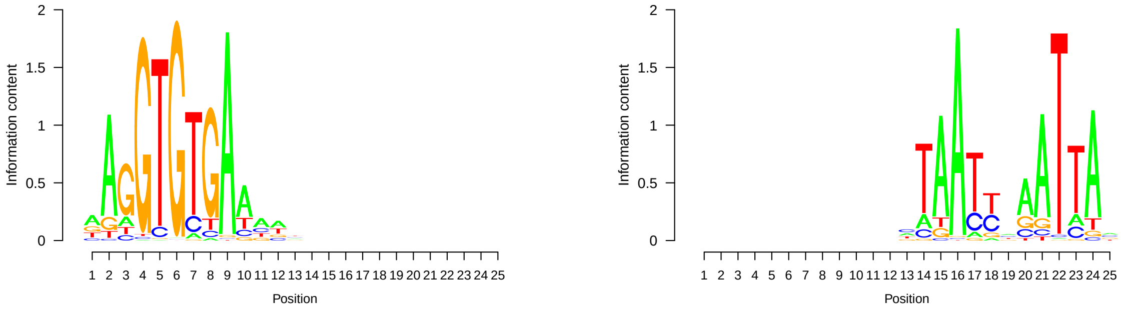


(a) Alx4 PWM from [5] and visualized with information content.

(b) Eomes PWM from [5] and visualized with information content.

Figure 47: Alx4 and Eomes PWMs

The two individual PWMs can bind to DNA in orientations Alx4-Eomes or Eomes-Alx4. When the other binding motif is masked with N columns, these two orientations automatically consider all possible binding scenarios on opposite strands. Furthermore, as the amount of gaps is unknown the length of the true combinatorial PWM site is unknown. Therefore, both ends of the DNA ligands should be searched with a true PWM, not only with the N extension. Thus, searching for the maximum PWM score sites is performed by searching with two extended PWMs for both of the orientations. For instance, for orientation Alx4-Eomes, the maximum score site is searched with a PWM where Alx4 position weight matrix is extended to 25 bases to the right and with a PWM where Eomes position weight matrix is extended to 25 bases to the left. Then the subsequence that has the highest score out of the sites searched with both of these PWMs is chosen as the maximum score site for training the random forest. The extended PWMs for orientation Alx4-Eomes are represented in Figure 48. Furthermore, the length of the PWM extensions was chosen to be 25 in order for all the possible sites in the middle of the DNA ligands to be scored with the PWMs as well, since the binding motif is very unlikely to be shorter than 10 nucleic acids. However, this binding motif length might be too short for some of the TF pairs. Thus, part of the information might be missing. Random forest should however model the binding motif with 25 bases already rather accurately.

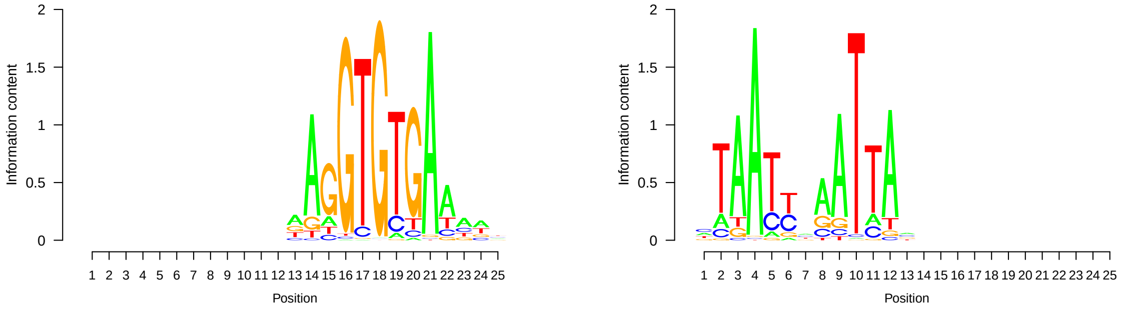


(a) Alx4 + N PWM

(b) N + Eomes PWM

Figure 48: Alx4-Eomes PWM extensions

Furthermore, Figure 49 represent the PWM extensions for orientation Eomes-Alx4. Thus, Eomes PWM is extended to 25 nucleic acids to the right and Alx4 is extended to 25 nucleic acids to the left.

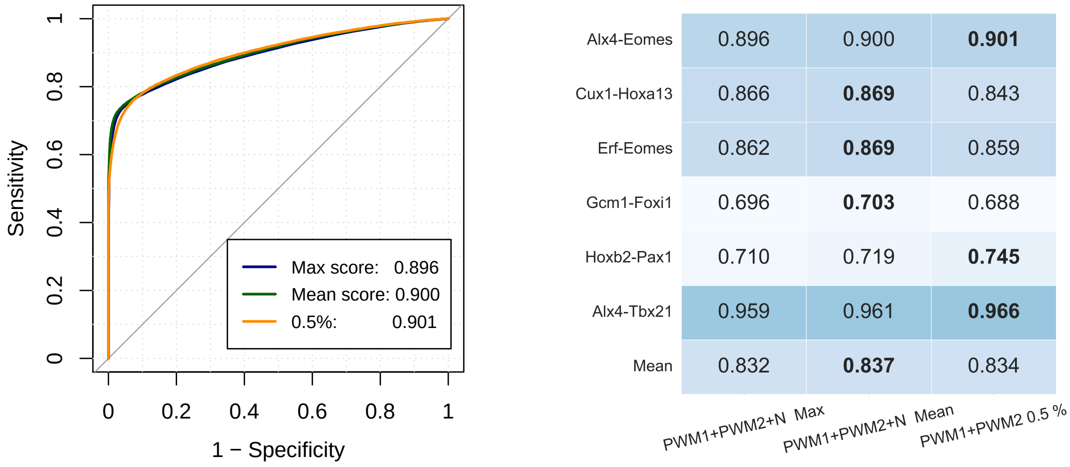


(a) N + Alx4 PWM

(b) Eomes + N PWM

Figure 49: Eomes-Alx4 PWM extensions

Thus, two random forests are trained with subsequences chosen by PWMs covering the two orientations. It is tested for Alx4-Eomes and for the five other TF pairs whether taking the maximum score out of the two random forest scores for each DNA ligand in test set or averaging them produce higher test set AUC. In addition, the performance of these two approaches (PWM1+PWM2+N+RF Max) and (PWM1+PWM2+N+RF Mean) are compared to the random forest ensemble with spacings within 0.5 % of best spacings according to validation set AUC (PWM1+PWM2+RF). The results are represented in Figure 50 for six data sets.



	PWM1+PWM2+N Max	PWM1+PWM2+N Mean	PWM1+PWM2 0.5 %
Alx4-Eomes	0.896	0.900	**0.901**
Cux1-Hoxa13	0.866	**0.869**	0.843
Erf-Eomes	0.862	**0.869**	0.859
Gcm1-Foxi1	0.696	**0.703**	0.688
Hoxb2-Pax1	0.710	0.719	**0.745**
Alx4-Tbx21	0.959	0.961	**0.966**
Mean	0.832	**0.837**	0.834

(a) Alx4-Eomes

(b) AUC values with 6 TF pairs

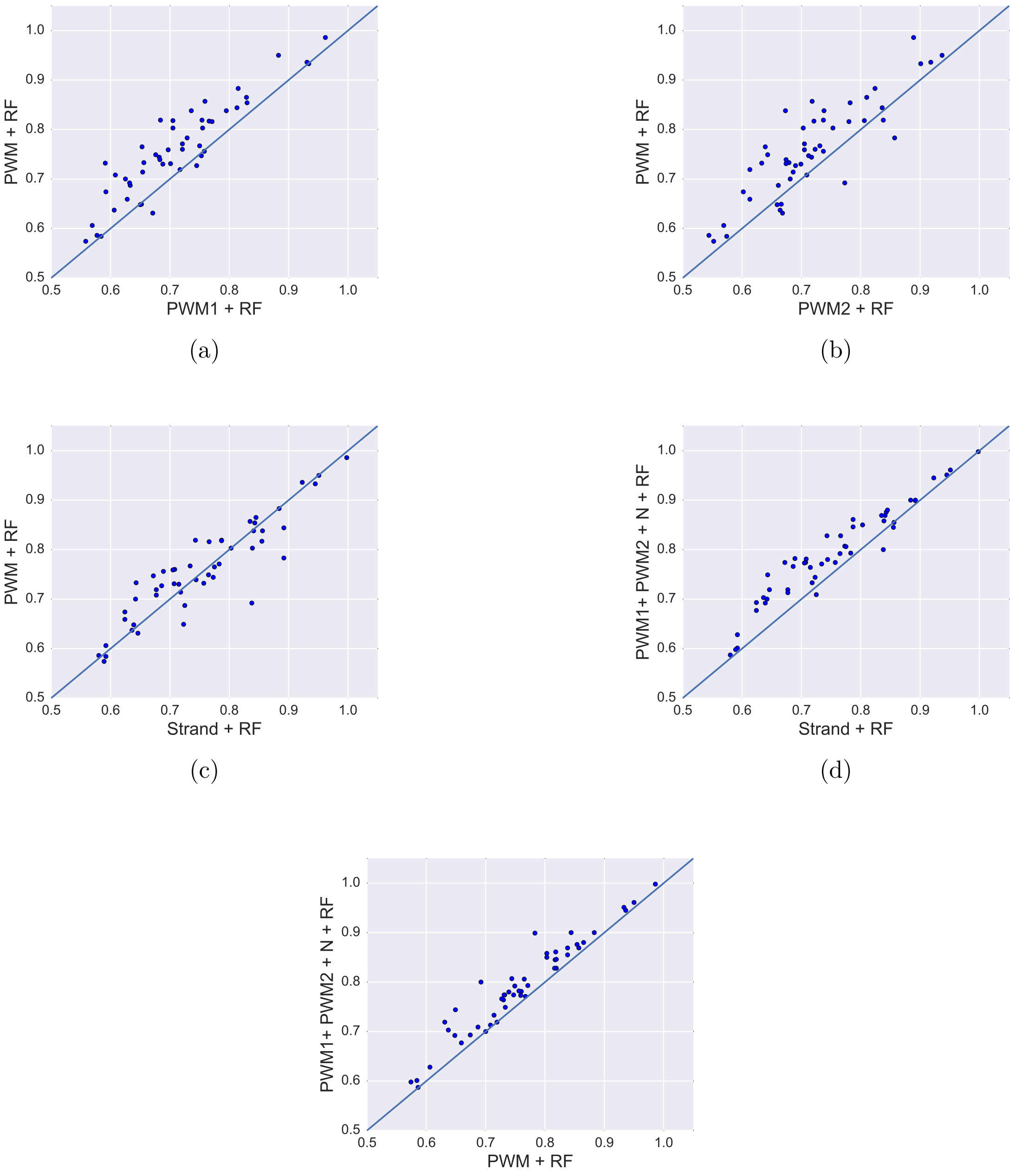
Figure 50: Test set AUC with PWM1+PWM2+N+RF model in comparison to PWM1+PWM2+RF.

The results indicate that finding subsequences from DNA ligands for random forest training with the extended PWMs (PWM1+PWM2+N+RF) performs equally or slightly better than the random forest ensemble with all the best spacings. This might occur, because all the gaps are considered simultaneously. Thus, choosing the PWM site incorrectly is less likely than when the two PWMs are attached to each other according to one of the spacings. In addition, running time for this model is lower. Thus, random forest modeling with sites found with extended PWMs is considered further. However, it is surprising that averaging the scores of the two random forest for each DNA ligand in test set outperforms taking the maximum, since only one orientation should occur on each DNA ligand. Ensemble of the two random forests increase predictive accuracy even though one of the forests contain incorrect information about the motif at least partly. Since the PWM sites are searched with only the other PWM, even the forest trained on the incorrect orientation include correct information about the binding motif.

5.2.6 Results with multiple experiments

The performance of random forest trained with 25 nucleic acids long subsequence chosen by matching only one of the transcriptional factor pair position weight matrices (PWM1+PWM2+N+RF) is assessed with a more comprehensive set comprising 50

CAP-SELEX data sets. Furthermore, random forests trained with the entire DNA reads (Strand+RF) in addition to training with only individual PWM sites (PWM1+RF and PWM2+RF) and with Jolma et al. combinatorial PWM sites (PWM+RF) are considered. Figure 51 shows the comparisons.



(a)

(b)

(c)

(d)

(e)

Figure 51: Comparison of Strand+RF, PWM1+RF, PWM2+RF, PWM+RF and PWM1+PWM2+N+RF.

Random forest modeling with the combinatorial PWM sites with a mean AUC of 0.759 outperforms modeling with the individual PWM sites: PWM1+RF (average=0.712, p-value=$2.39 * 10^{-8}$, two-sided Wilcoxon signed rank test, n=50) and PWM2+RF (average=0.716, p-value=$1.39 * 10^{-6}$, two-sided Wilcoxon signed rank test, n=50). However, the full length DNA strand model and random forest modeling with the combinatorial PWM sites perform equally. Random forest trained with 25 nucleic acids long subsequence chosen by matching only one of the TF pair PWMs (PWM1+PWM2+N+RF) outperforms with a mean AUC of 0.790 the PWM+RF model (average=0.759, p-value=$1.68 * 10^{-9}$, two-sided Wilcoxon signed rank test, n=50) and the Strand+RF model (average=0.753, p-value=$2.09 * 10^{-8}$, two-sided Wilcoxon signed rank test, n=50)

PWM1+PWM2+N+RF and PWM+RF models are compared to the prediction accuracies obtained by scoring sequences with only the Jolma et al. TF pair position weight matrices (PWM). Similarly as with HT-SELEX experiments, the PWM scores of each position in a ligand are averaged, which gives the final score to the DNA ligand. The DNA reads are padded with four 'N' values at each end, which are scored with probability 0.25. Furthermore, if there is multiple PWMs published for a TF pair, the DNA ligands are scored with all of them and the maximum score for each ligand is chosen as the final score, because TF pairs can bind DNA with different spacings. However, sometimes utilizing only one of the PWMs yielded highest AUC value. For each CAP-SELEX data set the highest obtained AUC is reported for scoring with PWMs. Figure 52 represents how the two random forest models that perform the best on test data compare to scoring sequences with TF pair PWMs.

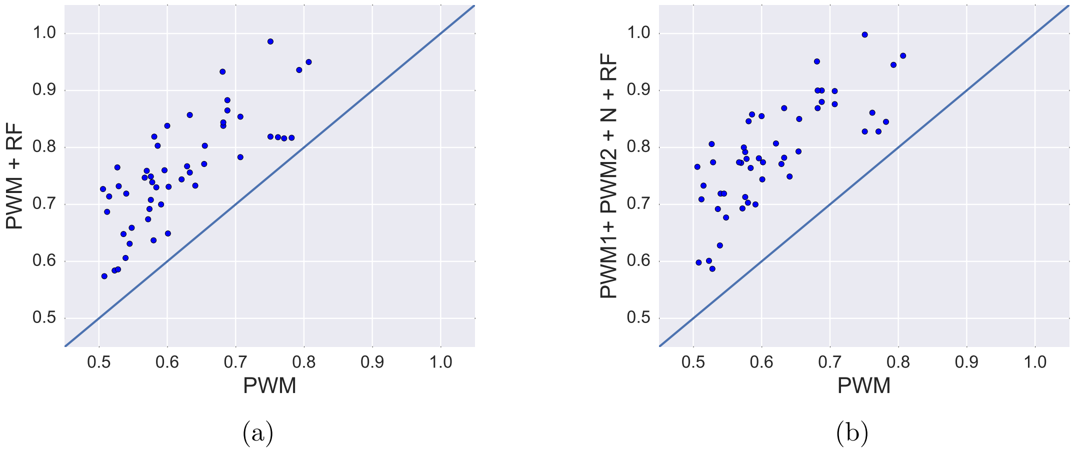


Figure 52: Comparison of PWM1+PWM2+N+RF and PWM+RF to PWM.

Thus, random forest increase classification accuracy in comparison to scoring sequences only with the position weight matrices. Random forest trained with the combinatorial PWM sites outperforms scoring sequences with only the PWM (average=0.618, p-value=$7.78 * 10^{-10}$, two-sided Wilcoxon signed rank test, n=50). Furthermore, random forest model with elongated PWM sites of the individual transcriptional factors (PWM1+PWM2+N+RF) with mean test set AUC of 0.790 outperforms the TF pair PWMs (average=0.618, p-value=$7.79 * 10^{-10}$, two-sided

Wilcoxon signed rank test, n=50). In addition, the difference between PWM and random forest seems to be more significant for transcriptional factor pairs than for individual transcriptional factor motifs as represented in Figure 35 for HT-SELEX data. Figure 53 represent test set AUC values for the 50 studied transcriptional factor pair data sets with the two random forest models in addition to scoring sequences only with the position weight matrices. The transcriptional factors are sorted according to AUC values with PWM1+PWM2+N+RF model.

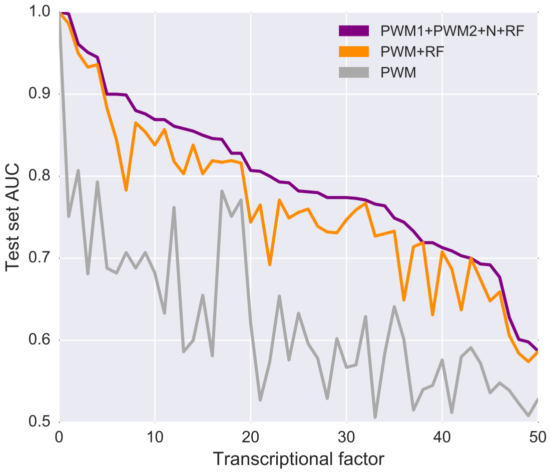


Figure 53: Test set AUC for 50 TFs.

6 Discussion

Protein-DNA binding specificities measured with HT-SELEX by and CAP-SELEX by Jolma et al. were modeled with random forests in this Master's thesis [5, 4]. SELEX measurements yield DNA sequences, which are known to contain a binding site although the exact position is unknown. In addition, binding might have occurred on the complementary DNA strand. [7] Binding specificities are usually modeled with position weight matrices, which describe the probability for seeing a certain nucleic acid at specific positions on the motif [3]. Recently, a neural network model, DeepBind, have been proposed to model TF binding motifs [10].

Experiments with HT-SELEX data sets were conducted by training random forests with full length DNA reads or with binding sites inside the DNA ligands search with either PWM or with decision trees. Modeling binding motifs with only the PWM is inaccurate in comparison to all the proposed random forest models and the neural network model, Deepbind. It has been shown that with in vivo experiments the independent normalization of each position on the binding sites may skew the probabilities, which might lead to incorrect PWM models [8]. Thus, it is possible that there are problems in PWM construction that impair predictive accuracy. In addition, modeling TF binding motifs with position weight matrices is inflexible. Higher complexity may be added to modeling the motif with random forest. This could be observed particularly as training random forest with the PWM sites yielded higher AUC values than simply utilizing the PWM. Random forest is a suitable method for modeling TF binding specificities as it can process DNA sequences with correlated neighboring nucleic acids as well as correlated nucleic acids with gaps. Due to the grouping property of random forests, the binding sites could be modeled such that DNA sequences with correlated variables get small minimal depths and thus high probabilities are given to new similar sequences [15]. One of the biggest limitations with PWMs is the assumed independence of nucleic acid positions on the motif [9]. In addition, it was shown by Rahul Siddharthan that although the nearest neighbor dinucleotide correlations are observed most commonly in yeast transcriptional factor binding sites, correlations of gapped dinucleotides are observed as well [9]. Since random forest is able to exploit these correlations, it is obvious that random forest would outperform PWM models. Furthermore, the instability of decision trees can be reduced by learning multiple decision trees and combining them [27]. Thus, random forests, as an ensemble of decision trees, will achieve high modeling complexity. Furthermore, when decision trees are constructed, the variables, which are considered for a data partitioning, are chosen randomly at each node [13]. In addition to improving variance reduction, this may enable the decision trees to discover groups of data with correlated variables at alternating positions. Therefore, random forest may be able to model binding even if there is variation in the exact location of the motif on the DNA sequence.

Random forest can increase model complexity in comparison to utilizing only the position weight matrices. Furthermore, one model was constructed by first searching for the PWM sites and utilizing these for training the random forest. This model give nucleic acids on the motifs positioned on certain features for random forest training, which could aid decision trees to partition training data correctly. However, since the PWM does not perform very well for all proteins, it might be problematic to utilize the PWM before training the random forest. However, even if random forest was trained partly on wrong subsequences, the forest would probably be able to discover true binding sites due to the grouping property of the decision trees. It was discovered for HT-SELEX data that training random forest with full length DNA strands outperformed training with binding motifs found either with PWM or decision trees. However, with HT-SELEX measurements conducted with 40 nucleic acids long DNA sequences random forest trained on subsequences outperformed forests trained on the entire DNA strands. This suggest that the TF motifs might actually be longer than the Jolma et al. position weight matrices insinuates [5]. Furthermore, it was revealed that elongating the found PWM sites with surrounding nucleic acids increased classification accuracy. If the PWM sites were found from the end of the DNA reads, they were elongated with the corresponding number of 'N' letters denoting nucleotides that are not determined. However, even the random forests with PWM sites elongated to the length of the entire DNA sequence did not predict binding sites as accurately as forests trained with the strands. Thus, it is possible that the 'N' letters disturb model performance. Since the amount of features is the same but for a part of the DNA sequences nucleic acids are replaced with 'N' features, information is lost, which seem to result in lower test set AUC values. Although, it could be assumed that centralizing specific nucleic acids of the TF motif positions to certain random forest features would increase the model performance, the information loss due to introduction of 'N' categories is more significant.

Furthermore, a model where the binding motifs were searched from DNA ligands with other decision trees before training the final random forest was implemented. This model performed almost equally to the forest trained with elongated PWM sites. The searched subsequences were long in order to ensure training the forests with entire motifs. Thus, the problem with PWM should not be only the inability to select subsequences correctly. Rather, the issue is probably with the length of the subsequences. Furthermore, Alipanahi et al. utilized the entire 20 nucleic acids long DNA strands for training the neural network model DeepBind as well [10]. In addition, for HT-SELEX measurements with 30 or 40 nucleic acids long DNA reads, the searched motifs were over 20 nucleic acids long [10]. Thus, the results in this Master's thesis are in line with the optimal motif lengths according to Alipanahi et al. for DeepBind. In addition, it is possible that a transcriptional factor can bind differing DNA motifs that should be modeled with more than one PWM [43]. Thus, searching with only one PWM, may result in incorrect motifs to be chosen for forest training. Contrary, random forest should be able to classify sequences even if there were different motifs present.

The optimal random forest model for HT-SELEX data is obtained by utilizing full length DNA strands chosen by a position weight matrix for measurements with 14, 20 or 30 nucleic acids long DNA reads, while for measurements with 40 nucleic acids long DNA reads the random forest is trained with elongated PWM sites. The PWM sites were elongated to 23 or 24 nucleic acids depending on whether the PWM has an odd or even length. The DeepBind outperformed random forest slightly but the difference is not statistically significant according to a two-sided Wilcoxon signed-rank test. Thus, random forest is a promising method for modeling transcriptional factor pair binding motifs as well.

Random forest was applied for CAP-SELEX data for modeling TF pair DNA binding motifs. Again random forest increased classification accuracy in comparison to only scoring sequences with PWMs. Thus, random forest is able to increase model complexity and flexibility also for transcriptional factor pair motifs. Furthermore, the difference between PWM scoring and random forest is more significant for TF pairs than for individual TFs in HT-SELEX data. It is possible that TF pairs have more variation in their binding motifs than individual TFs. Thus, the increased model complexity achieved with random forest is even more beneficial for TF pairs than for transcriptional factors that bind DNA individually. Modeling TF pair motifs with single PWMs might not be sufficient, since Jolma et al. noticed that some TF pairs bind DNA with relaxed gap spacings between the two individual TF motifs [4]. Random forest on the other hand does not seem to be highly sensitive to the location of the core motif on the features, since rather high classification accuracies can be achieved even with the full length DNA sequences. Thus, a motif that comprise different possible spacings of the two TFs can be constructed with random forest.

Random forests trained with the full length DNA strands chosen by the combinatorial PWM perform almost equally to the forests trained with the sites chosen by Jolma et al. position weight matrices published in [4]. Therefore, random forest seems to be able to discover the motif information even if the exact position of the motif on the DNA reads is uncertain. However, the test set AUC values with random forests trained on the sites chosen by Jolma et al. PWMs are slightly better, which might be due to DNA sequences of 40 bases being too long just as shown with HT-SELEX. However, the difference between these two approaches is not as significant with TF pairs as it is with individual TFs. One possibility is simply that motifs are longer for TF pairs, which improves performance of the full length DNA strand model. However, it was tested whether elongating the searched Jolma et al. PWM sites would increase classification accuracy and for TF pairs it seemed that the combinatorial PWMs were long enough. Thus, it is possible that TF pair motifs are not significantly longer than motifs of individual TFs. An other option is that the combinatorial PWMs does not choose the correct subsequences as accurately as the individual TF PWMs. Thus, random forest trained with the PWM sites would not perform as well as it could for TF pairs. Furthermore, random forests were trained on subsequences chosen by PWMs of the individual transcriptional factors separately. Testing unseen DNA ligands with the two random forest would indicate whether accurate models could be constructed already with information of either one of the transcriptional factor motifs. It was discovered that random forest can in fact distinguish between ligands that contain a binding site and those that do not quite accurately when either one of the TF motifs were considered. Thus, there is information about individual TF motifs present on the DNA reads as well.

Furthermore, experiments were conducted by attaching the two transcriptional factor PWMs to each other according to all four different orientations with varying amount of gaps. These combination PWMs were then utilized for searching for subsequences corresponding to the assumed motifs and used for training random forest. Surprisingly it was discovered the most of the spacings yielded good classification accuracies on unseen DNA ligands even though it was shown by Jolma et al. that transcriptional factors preferred certain orientations over others [4]. One explanation for this is the ability of random forest to predict binding already with either one of the transcriptional factor PWM sites out of the pair. In addition, random forest is not extremely sensitive to excessive features not pertaining to binding. Thus, as the combination PWMs are likely to choose sites that comprise at least one of the two TF motifs even if the spacing between the PWMs was incorrect, there will be information considering the motifs in random forest training. Furthermore, in the case of incorrect spacing chosen for PWM attachment, the subsequences are chosen by the more specific binding motif on the combination PWM. However, some spacing result in higher test set AUC values than others. These spacing may be considered as the spacings choosing subsequences from the ligands most often correctly. Separate random forest trained with the best spacings may be combined by scoring test ligands with each of them and choosing the maximum score or the average of the scores as the final score. As expected averaging produced higher test set AUCs. Although, only one spacing should be present on each DNA ligand, the ensemble of random forests combines the predictive power of multiple forests and thus is able to predict TF pair binding sites more accurately.

Finally, a model that incorporate multiple spacings at once was implemented. However, since the orientation of the transcriptional factors on the ligand is unknown, all possible spacing were not combined for training a single random forest. Rather, the individual PWMs were elongated to 25 nucleic acids to both directions separately and utilized for searching subsequences corresponding to certain orientations. The model yields two random forest such that all possible orientations were covered. Furthermore, since the individual PWMs were elongated, all possible gaps between the motifs could be considered simultaneously. Unseen test DNA ligands were scored with both of the random forests and the final score could be obtained by either maximizing or averaging the two scores. Averaging again produce higher AUC values, although only one orientation would be present on each ligand. If the TFs are bound to opposite strands both of the forests should be capturing to a large extent the same binding preferences in which case averaging will yield a higher AUC due to the combination of more decision trees increasing modeling accuracy. When binding have occurred on the same strand, only the other forest should be scoring binding correctly. However, the forest trained with the incorrect binding orientation should still include the other TF binding site. Therefore, the forest should be able to predict binding to some extent as well according to previous results with scoring CAP-SELEX sequences with only the other TF binding motif. Thus, combining the two forest scores through averaging yield higher AUC values. Furthermore, since the individual PWMs were extended only to 25 nucleic acids, binding motif information might be missing partly for some pairs, although for most pairs 25 should be sufficient. The elongation to 25 nucleic acids was chosen so that shorter motifs positioned on the middle of the strands could be found as well. Furthermore, random forest is not particularly sensitive to positive training set containing partly random sequences, since decision trees are able to find the correlated groups such that new sequences will fall to these leaves more likely [15]. Thus, the forests should model binding specificities accurately even if for some DNA ligands the wrong sites were chosen, such as if the true motif was shorter than 10 nucleic acids and in the middle of the DNA strand. However, TF pairs motifs should be longer than 10 nucleic acids. The model with elongated transcriptional factor PWMs outperformed slightly the model with the ensemble of best spacings search with attached PWMs. Although, the spacing ensemble model could probably be improved further by averaging over more forests. However, the computational cost and running time would increase greatly.

7 Conclusions

In this Master's thesis random forest was implemented for modeling protein-DNA binding specificities. The studied proteins were transcriptional factors, thus responsible for regulating gene expression in cells. Currently, binding specificities are most often modeled with position weight matrices even though they are likely to be too simple for modeling DNA binding motifs. Furthermore, a neural network model, DeepBind, have been proposed for modeling binding specificities and it was shown by Alipanahi et al. that DeepBind outperformed PWM models [10]. In this Master's thesis random forest models were trained with HT-SELEX and CAP-SELEX data sets measured by Jolma et al. [5, 4]. Thus, models were implemented for individual transcriptional factor binding specificities in addition to transcriptional factor pair binding specificities. The best performing models combined position weight matrices with random forest. Random forest models outperformed scoring sequences with position weight matrices with both individual and dimer transcriptional factor motifs. Furthermore, the difference between PWM models and random forest models was greater for transcriptional factor pairs than for individual transcriptional factors. Thus, the possible variations in spacing between the two transcriptional factors in the dimer may increase the benefits of utilizing more flexible models. Furthermore, for individual transcriptional factor motifs DeepBind outperformed random forest slightly. However, the difference is not significant and for transcriptional factor pairs DeepBind models are not provided. Thus, modeling transcriptional factor pair DNA binding specificities with random forest instead of PWM models is advantageous.

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell*. Garland, 4th edition, 2002.

[2] Kyle L MacQuarrie, Abraham P Fong, Randall H Morse, and Stephen J Tapscott. Genome-wide transcription factor binding: beyond direct target regulation. *Trends in Genetics*, 27(4):141–148, 2011.

[3] Gary D Stormo and Yue Zhao. Determining the specificity of protein–dna interactions. *Nature Reviews Genetics*, 11(11):751–760, 2010.

[4] Arttu Jolma, Yimeng Yin, Kazuhiro R Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. Dna-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 534(7607):1–2, 2016.

[5] 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.

[6] Yan Yang, Dongliang Yang, Hermann J Schluesener, and Zhiren Zhang. Advances in selex and application of aptamers in the central nervous system. *Biomolecular engineering*, 24(6):583–592, 2007.

[7] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanpää, et al. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome research*, 20(6):861–873, 2010.

[8] Shuxiang Ruan and Gary D Stormo. Inherent limitations of probabilistic models for protein-dna binding specificity. *PLoS computational biology*, 13(7):e1005638, 2017.

[9] Rahul Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS one*, 5(3):e9722, 2010.

[10] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[11] Minghui Jiang, James Anderson, Joel Gillespie, and Martin Mayne. ushuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, 9(1):192, 2008.

[12] Kevin Murphy. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012.

[13] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[14] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

[15] Hemant Ishwaran, Udaya B Kogalur, Eiran Z Gorodeski, Andy J Minn, and Michael S Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010.

[16] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, 2014.

[17] R Kurokawa, VC Yu, A Näär, S Kyakumoto, ZHIHUA Han, S Silverman, MG Rosenfeld, and CK Glass. Differential orientations of the dna-binding domain and carboxy-terminal dimerization interface regulate binding site selection by nuclear receptor heterodimers. *Genes & development*, 7(7b):1423–1435, 1993.

[18] 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.

[19] Martha Bulyk. Protein binding microarrays for the characterization of dna–protein interactions. *Analytics of Protein–DNA Interactions*, pages 65–85, 2007.

[20] Gary D Stormo. Modeling the specificity of protein-dna interactions. *Quantitative biology*, 1(2):115, 2013.

[21] Patrik D'haeseleer. How does dna sequence motif discovery work? *Nature biotechnology*, 24(8):959–961, 2006.

[22] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2009.

[23] Daniel E Newburger and Martha L Bulyk. Uniprobe: an online database of protein binding microarray data on protein–dna interactions. *Nucleic acids research*, 37(suppl_1):D77–D82, 2008.

[24] Kazuhiro R Nitta, Arttu Jolma, Yimeng Yin, Ekaterina Morgunova, Teemu Kivioja, Junaid Akhtar, Korneel Hens, Jarkko Toivonen, Bart Deplancke, Eileen EM Furlong, et al. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife*, 4:e04837, 2015.

[25] Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

[26] Giorgio Valentini and Thomas G Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *Journal of Machine Learning Research*, 5(Jul):725–775, 2004.

[27] Marina Skurichina and Robert PW Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis & Applications*, 5(2):121–135, 2002.

[28] D Kandel, Yossi Matias, Ron Unger, and Peter Winkler. Shuffling biological sequences. *Discrete Applied Mathematics*, 71(1-3):171–185, 1996.

[29] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[30] Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[31] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[32] David H Wolpert and William G Macready. An efficient method to estimate bagging's generalization error. *Machine Learning*, 35(1):41–55, 1999.

[33] Benny Chor, David Horn, Nick Goldman, Yaron Levy, and Tim Massingham. Genomic dna k-mer spectra: models and modalities. *Genome biology*, 10(10):R108, 2009.

[34] Jichen Yang and Stephen A Ramsey. A dna shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites. *Bioinformatics*, 31(21):3445–3450, 2015.

[35] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. Dnashaper: an r/bioconductor package for dna shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, 2015.

[36] 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(W1):W56–W62, 2013.

[37] Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W Wasserman. Dna shape features improve transcription factor binding site predictions in vivo. *Cell systems*, 3(3):278–286, 2016.

[38] P Shannon. Motifdb: An annotated collection of protein-dna binding sequence motifs. *R package version*, 1(0), 2014.

[39] Hendrik Blockeel, David Page, and Ashwin Srinivasan. Multi-instance tree learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 57–64. ACM, 2005.

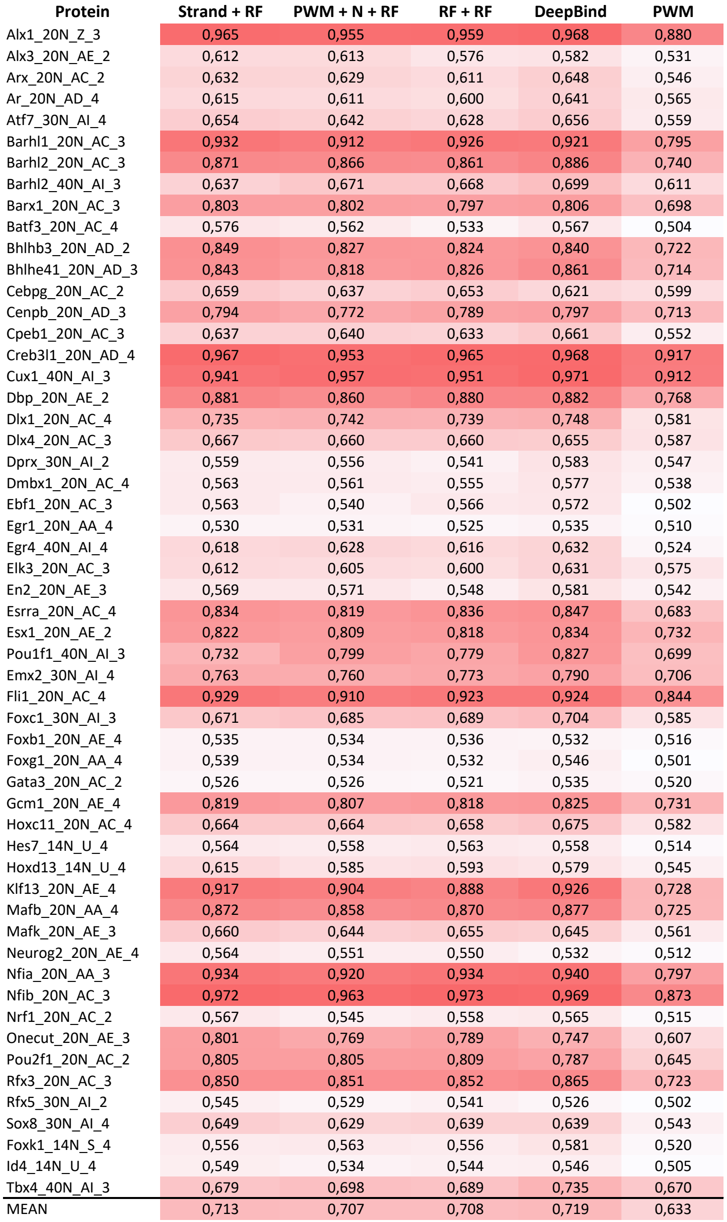
[40] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[41] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[42] H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy. Biostrings: String objects representing biological sequences, and matching algorithms. 2017. R package version 2.44.2.

[43] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. Transcription factor–dna binding: beyond binding site motifs. *Current Opinion in Genetics & Development*, 43:110–119, 2017.

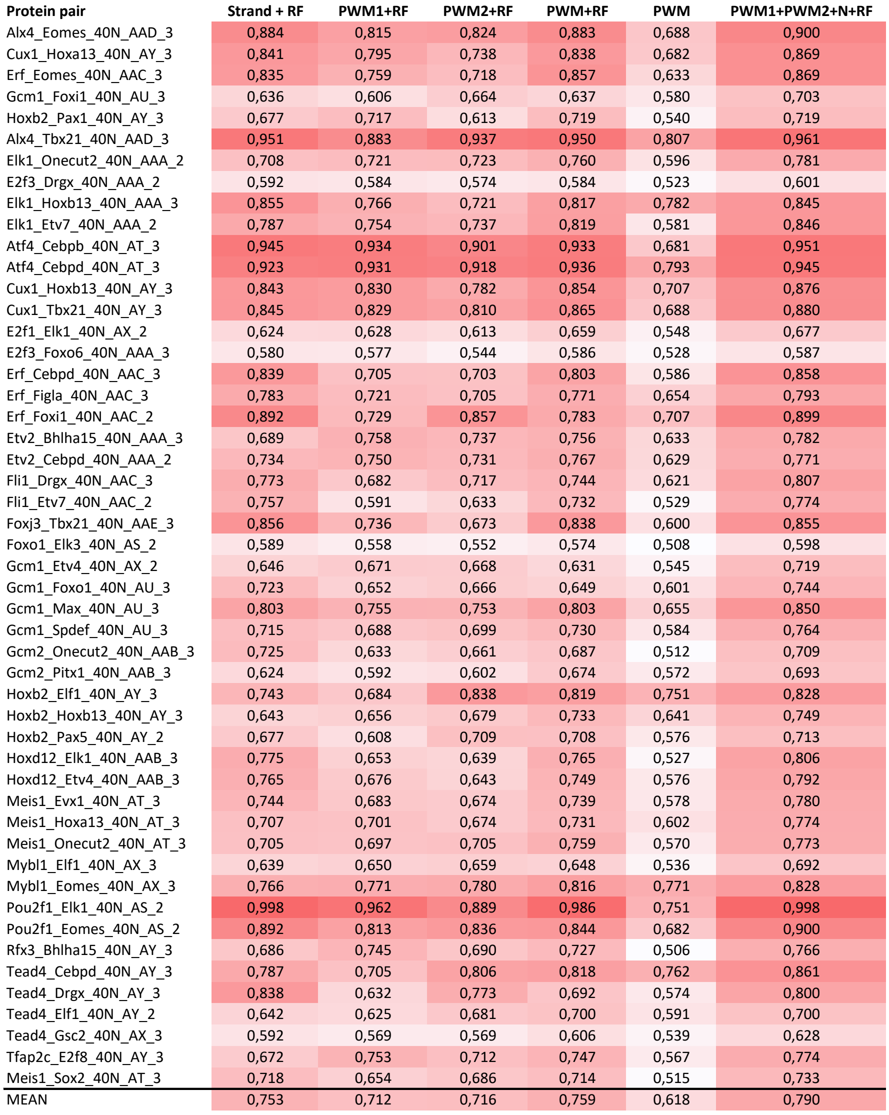
A HT-SELEX test set AUC for TFs



Protein	Strand + RF	PWM + N + RF	RF + RF	DeepBind	PWM
Alx1_20N_Z_3	0,965	0,955	0,959	0,968	0,880
Alx3_20N_AE_2	0,612	0,613	0,576	0,582	0,531
Arx_20N_AC_2	0,632	0,629	0,611	0,648	0,546
Ar_20N_AD_4	0,615	0,611	0,600	0,641	0,565
Atf7_30N_AI_4	0,654	0,642	0,628	0,656	0,559
Barhl1_20N_AC_3	0,932	0,912	0,926	0,921	0,795
Barhl2_20N_AC_3	0,871	0,866	0,861	0,886	0,740
Barhl2_40N_AI_3	0,637	0,671	0,668	0,699	0,611
Barx1_20N_AC_3	0,803	0,802	0,797	0,806	0,698
Batf3_20N_AC_4	0,576	0,562	0,533	0,567	0,504
Bhlhb3_20N_AD_2	0,849	0,827	0,824	0,840	0,722
Bhlhe41_20N_AD_3	0,843	0,818	0,826	0,861	0,714
Cebpg_20N_AC_2	0,659	0,637	0,653	0,621	0,599
Cenpb_20N_AD_3	0,794	0,772	0,789	0,797	0,713
Cpeb1_20N_AC_3	0,637	0,640	0,633	0,661	0,552
Creb3l1_20N_AD_4	0,967	0,953	0,965	0,968	0,917
Cux1_40N_AI_3	0,941	0,957	0,951	0,971	0,912
Dbp_20N_AE_2	0,881	0,860	0,880	0,882	0,768
Dlx1_20N_AC_4	0,735	0,742	0,739	0,748	0,581
Dlx4_20N_AC_3	0,667	0,660	0,660	0,655	0,587
Dprx_30N_AI_2	0,559	0,556	0,541	0,583	0,547
Dmbx1_20N_AC_4	0,563	0,561	0,555	0,577	0,538
Ebf1_20N_AC_3	0,563	0,540	0,566	0,572	0,502
Egr1_20N_AA_4	0,530	0,531	0,525	0,535	0,510
Egr4_40N_AI_4	0,618	0,628	0,616	0,632	0,524
Elk3_20N_AC_3	0,612	0,605	0,600	0,631	0,575
En2_20N_AE_3	0,569	0,571	0,548	0,581	0,542
Esrra_20N_AC_4	0,834	0,819	0,836	0,847	0,683
Esx1_20N_AE_2	0,822	0,809	0,818	0,834	0,732
Pou1f1_40N_AI_3	0,732	0,799	0,779	0,827	0,699
Emx2_30N_AI_4	0,763	0,760	0,773	0,790	0,706
Fli1_20N_AC_4	0,929	0,910	0,923	0,924	0,844
Foxc1_30N_AI_3	0,671	0,685	0,689	0,704	0,585
Foxb1_20N_AE_4	0,535	0,534	0,536	0,532	0,516
Foxg1_20N_AA_4	0,539	0,534	0,532	0,546	0,501
Gata3_20N_AC_2	0,526	0,526	0,521	0,535	0,520
Gcm1_20N_AE_4	0,819	0,807	0,818	0,825	0,731
Hoxc11_20N_AC_4	0,664	0,664	0,658	0,675	0,582
Hes7_14N_U_4	0,564	0,558	0,563	0,558	0,514
Hoxd13_14N_U_4	0,615	0,585	0,593	0,579	0,545
Klf13_20N_AE_4	0,917	0,904	0,888	0,926	0,728
Mafb_20N_AA_4	0,872	0,858	0,870	0,877	0,725
Mafk_20N_AE_3	0,660	0,644	0,655	0,645	0,561
Neurog2_20N_AE_4	0,564	0,551	0,550	0,532	0,512
Nfia_20N_AA_3	0,934	0,920	0,934	0,940	0,797
Nfib_20N_AC_3	0,972	0,963	0,973	0,969	0,873
Nrf1_20N_AC_2	0,567	0,545	0,558	0,565	0,515
Onecut_20N_AE_3	0,801	0,769	0,789	0,747	0,607
Pou2f1_20N_AC_2	0,805	0,805	0,809	0,787	0,645
Rfx3_20N_AC_3	0,850	0,851	0,852	0,865	0,723
Rfx5_30N_AI_2	0,545	0,529	0,541	0,526	0,502
Sox8_30N_AI_4	0,649	0,629	0,639	0,639	0,543
Foxk1_14N_S_4	0,556	0,563	0,556	0,581	0,520
Id4_14N_U_4	0,549	0,534	0,544	0,546	0,505
Tbx4_40N_AI_3	0,679	0,698	0,689	0,735	0,670
MEAN	0,713	0,707	0,708	0,719	0,633

Figure A1: AUC with HT-SELEX. Labeling: TF_Read length_Batch_Cycle

B CAP-SELEX test set AUC for TF pairs



Protein pair	Strand + RF	PWM1+RF	PWM2+RF	PWM+RF	PWM	PWM1+PWM2+N+RF
Alx4_Eomes_40N_AAD_3	0,884	0,815	0,824	0,883	0,688	0,900
Cux1_Hoxa13_40N_AY_3	0,841	0,795	0,738	0,838	0,682	0,869
Erf_Eomes_40N_AAC_3	0,835	0,759	0,718	0,857	0,633	0,869
Gcm1_Foxi1_40N_AU_3	0,636	0,606	0,664	0,637	0,580	0,703
Hoxb2_Pax1_40N_AY_3	0,677	0,717	0,613	0,719	0,540	0,719
Alx4_Tbx21_40N_AAD_3	0,951	0,883	0,937	0,950	0,807	0,961
Elk1_Onecut2_40N_AAA_2	0,708	0,721	0,723	0,760	0,596	0,781
E2f3_Drgx_40N_AAA_2	0,592	0,584	0,574	0,584	0,523	0,601
Elk1_Hoxb13_40N_AAA_3	0,855	0,766	0,721	0,817	0,782	0,845
Elk1_Etv7_40N_AAA_2	0,787	0,754	0,737	0,819	0,581	0,846
Atf4_Cebpb_40N_AT_3	0,945	0,934	0,901	0,933	0,681	0,951
Atf4_Cebpd_40N_AT_3	0,923	0,931	0,918	0,936	0,793	0,945
Cux1_Hoxb13_40N_AY_3	0,843	0,830	0,782	0,854	0,707	0,876
Cux1_Tbx21_40N_AY_3	0,845	0,829	0,810	0,865	0,688	0,880
E2f1_Elk1_40N_AX_2	0,624	0,628	0,613	0,659	0,548	0,677
E2f3_Foxo6_40N_AAA_3	0,580	0,577	0,544	0,586	0,528	0,587
Erf_Cebpd_40N_AAC_3	0,839	0,705	0,703	0,803	0,586	0,858
Erf_Figla_40N_AAC_3	0,783	0,721	0,705	0,771	0,654	0,793
Erf_Foxi1_40N_AAC_2	0,892	0,729	0,857	0,783	0,707	0,899
Etv2_Bhlha15_40N_AAA_3	0,689	0,758	0,737	0,756	0,633	0,782
Etv2_Cebpd_40N_AAA_2	0,734	0,750	0,731	0,767	0,629	0,771
Fli1_Drgx_40N_AAC_3	0,773	0,682	0,717	0,744	0,621	0,807
Fli1_Etv7_40N_AAC_2	0,757	0,591	0,633	0,732	0,529	0,774
Foxj3_Tbx21_40N_AAE_3	0,856	0,736	0,673	0,838	0,600	0,855
Foxo1_Elk3_40N_AS_2	0,589	0,558	0,552	0,574	0,508	0,598
Gcm1_Etv4_40N_AX_2	0,646	0,671	0,668	0,631	0,545	0,719
Gcm1_Foxo1_40N_AU_3	0,723	0,652	0,666	0,649	0,601	0,744
Gcm1_Max_40N_AU_3	0,803	0,755	0,753	0,803	0,655	0,850
Gcm1_Spdef_40N_AU_3	0,715	0,688	0,699	0,730	0,584	0,764
Gcm2_Onecut2_40N_AAB_3	0,725	0,633	0,661	0,687	0,512	0,709
Gcm2_Pitx1_40N_AAB_3	0,624	0,592	0,602	0,674	0,572	0,693
Hoxb2_Elf1_40N_AY_3	0,743	0,684	0,838	0,819	0,751	0,828
Hoxb2_Hoxb13_40N_AY_3	0,643	0,656	0,679	0,733	0,641	0,749
Hoxb2_Pax5_40N_AY_2	0,677	0,608	0,709	0,708	0,576	0,713
Hoxd12_Elk1_40N_AAB_3	0,775	0,653	0,639	0,765	0,527	0,806
Hoxd12_Etv4_40N_AAB_3	0,765	0,676	0,643	0,749	0,576	0,792
Meis1_Evx1_40N_AT_3	0,744	0,683	0,674	0,739	0,578	0,780
Meis1_Hoxa13_40N_AT_3	0,707	0,701	0,674	0,731	0,602	0,774
Meis1_Onecut2_40N_AT_3	0,705	0,697	0,705	0,759	0,570	0,773
Mybl1_Elf1_40N_AX_3	0,639	0,650	0,659	0,648	0,536	0,692
Mybl1_Eomes_40N_AX_3	0,766	0,771	0,780	0,816	0,771	0,828
Pou2f1_Elk1_40N_AS_2	0,998	0,962	0,889	0,986	0,751	0,998
Pou2f1_Eomes_40N_AS_2	0,892	0,813	0,836	0,844	0,682	0,900
Rfx3_Bhlha15_40N_AY_3	0,686	0,745	0,690	0,727	0,506	0,766
Tead4_Cebpd_40N_AY_3	0,787	0,705	0,806	0,818	0,762	0,861
Tead4_Drgx_40N_AY_3	0,838	0,632	0,773	0,692	0,574	0,800
Tead4_Elf1_40N_AY_2	0,642	0,625	0,681	0,700	0,591	0,700
Tead4_Gsc2_40N_AX_3	0,592	0,569	0,569	0,606	0,539	0,628
Tfap2c_E2f8_40N_AY_3	0,672	0,753	0,712	0,747	0,567	0,774
Meis1_Sox2_40N_AT_3	0,718	0,654	0,686	0,714	0,515	0,733
MEAN	0,753	0,712	0,716	0,759	0,618	0,790

Figure B1: AUC with CAP-SELEX. Labeling: TF pair_Read length_Batch_Cycle