

Aalto University
School of Science
Degree Programme in Computer Science and Engineering

Banafsheh Khakipoor

Integrated data analysis pipeline for whole human genome transcription factor binding sites prediction

Master's Thesis
Espoo, May 28, 2015

Supervisor: Professor Harri Lähdesmäki, Aalto University
Advisor: Professor Harri Lähdesmäki

Author:	Banafsheh Khakipoor	
Title:	Integrated data analysis pipeline for whole human genome transcription factor binding sites prediction	
Date:	May 28, 2015	Pages: 37
Major:	Bioinformatics	Code: T-61
Supervisor:	Professor Harri Lähdesmäki	
Advisor:	Professor Harri Lähdesmäki	
	<p>Transcription factors (TF) have a central role in regulating gene expression by binding to regulatory regions in DNA. Position weight matrix (PWM) model is the most commonly used model for representing and predicting TF binding sites. Consequently, several studies have been done on predicting TF binding sites using PWMs and many databases have been created containing large numbers of PWMs. However, these studies require the user to search for binding sites for each PWM separately, thus making it is difficult to get a general view of binding predictions for many PWMs simultaneously. In response to this need, this thesis project evaluates both individual and groups of PWMs and creates an effortless method to analyze and visualize the desired set of PWMs together, making it easier for biologist to analyze large amount of data in a short period of time. For this purpose, we used bioinformatics methods to detect putative TF binding sites in human genome and make them available online via the UCSC genome browser. Still, the sheer amount of data in PWM databases required a more efficient method to summarize TF binding prediction. Hence, we used PWM similarity measures and clustering algorithms to group together PWMs and to create one integrated database from four popular PWM databases: SELEX, TRANSFAC, UniPROBE, and JASPAR. All results are made publicly available for the research community via the UCSC genome browser.</p>	
Keywords:	Transcription Factor, PWM, TRANSFAC, JASPAR, PBM, SELEX	
Language:	English	

Acknowledgements

I wish to thank my supervisor Professor Harri Lähdesmäki for his support and guidance throughout this project.

The calculations presented in this project were performed using computer resources within the Aalto University School of Science "Science-IT" project.

Espoo, May 28, 2015

Banafsheh Khakipoor

Abbreviations and Acronyms

PWM	Position weight matrix
TF	Transcription factor
PBM	Protein binding microarray
UCSC	University of California Santa Cruz
HOMER	Hypergeometric optimization of motif enrichment
DBD	DNA domain binding

Contents

Abbreviations and Acronyms	4
1 Introduction	7
1.1 Transcription factors	7
1.2 Problem statement	9
1.3 Structure of the thesis	9
2 Databases	10
2.1 JASPAR	10
2.2 SELEX	11
2.3 UniPROBE	12
2.4 TRANSFAC	12
2.5 Classification	13
2.5.1 Rank definitions and contents	13
3 Methods	15
3.1 Representation of DNA binding sites	15
3.2 PWM	16
3.2.1 Modeling from qualitative binding site data	17
3.2.2 PWM threshold	18
3.3 Human DNA-binding sites prediction and visualization	19
3.4 Similarity algorithm	19
3.4.1 Gupta et al.	19
3.4.2 Habib et al.	20
3.4.3 Tanaka et al.	21
3.5 Hierarchical clustering	23
4 Implementation	24
5 Results	29

6 Conclusion	32
A UCSC genome browser links	37

Chapter 1

Introduction

For our bodies to function certain proteins need to be created and to do so our genetic material DNA needs to be converted into RNA and converted into proteins. The process in which DNA is converted into RNA is called transcription. DNA contains our genetic information and there are certain proteins which help in regulating the use of these genetic information. These transcription proteins or factors enable the usage of genetic information in the genome. By understanding the fundamental molecular mechanisms that control transcription in humans or in any other organism, we can gain a deeper understanding of what happens in our bodies and specifically what affects and causes diseases.

The human genome has around 3 billion base pairs and that encodes roughly 22,000 genes. These are stretches of DNA sequence that encode, ultimately, a product that is a protein, which makes the cells function.

1.1 Transcription factors

Genetic information transfers from DNA to RNA to proteins. Interestingly, the transcription of genetic information from DNA to RNA is performed with the help of certain proteins. These proteins bind to specific locations of the DNA hence they are called sequence-specific DNA binding factors as well as Transcription Factors (TF).

Transcription factors either alone or with the help of other proteins are responsible for transcribing particular genes of DNA to primary RNA followed by a post-transcription processes such as RNA splicing and translation which leads to the creation of functional proteins [15]. Hence transcription factors help in gene expression as well as specificity of proteins produced in different tissues.

One main feature of transcription factors is existence of DNA binding domains (DBDs), these target specific sequences of DNA adjacent to the genes they regulate. There exist other proteins that have important role in gene regulation such as co-activators, chromatin remodelers, histone acetylases, deacetylases, kinases, and methylases. However these proteins do not have the DNA binding domains and hence are not considered as transcription factors [22].

Transcription factors can read and interpret DNA genetic blueprint as well as help to increase or decrease transcription of genes, which makes them crucial in many cellular processes. Here are some examples of some of TF functional roles; general transcription factors (GTFs) are an important part of the large transcription pre-initiation complex which directly interacts with RNA polymerase. The most common GTFs are TFIIA, TFIIB, TFIID (see also TATA binding protein), TFIIE, TFIIIF, and TFIIH [27]. Hox transcription factor family is another example of TF that helps in body pattern formation in vast majority of organisms from fruit flies to humans [16]. Some transcription factors act downstream of signaling cascades of environmental stimuli; heat shock factors (HSF), these up-regulates genes needed to survive in higher temperatures [22]. Or hypoxia inducible factor (HIF) that up-regulate genes needed for surviving low-oxygen environments [3], and SREBP (sterol regulatory element binding protein) helps in maintaining proper lipid level in the cells [29]. Other transcription factors that help in cell growth and apoptosis is Myc oncogenes which is a tumor suppressor that helps in regulating cell cycle and as well deciding on the growth of cell as well as the time of division to two daughter cells. [6]

There are 3 methods used in classifying transcription factors; 1) based on mechanisms of their actions, 2) based on regulatory function, and 3) based on sequence homology in their DNA binding domain. First method (mechanistic classification) divides the TFs into three groups of general transcription factors, upstream transcription factors that bind upstream of initiation sites to either stimulate or suppress transcription, and finally, specific transcription factors which look for recognition sequences in proximity of the genes. Functional classification categorizes the TFs based on their regulatory function into two main groups of constitutively active and conditionally active TFs. Finally, the one used in this thesis, is structural classification based on sequence similarity of transcription factors' domain binding sites. This classifies TFs into 9 super class, which is represented in more details in next chapter.[23]

1.2 Problem statement

PWMs contain informations about motifs that could help in predicting DNA regulatory binding sites. However, A PWM is only a matrix with different probabilities of nucleotides in each position, which can be used to scan human genome for possible binding sites. Hence, it would be good if we could use their information and visualize the transcription factor's binding sites on human genome.

Also, with many databases currently available with PWM data, it can be nerve-racking to analyze each database separately. A method that can combine these databases into one is preferable. Further, we can use this database to scan human genome for DNA binding sites.

To the best of our knowledge, there exists some programs that look into the first issue (such as; PWMTools, rVISTA), however, each PWM needs to be evaluated individually. In this project our main goal is comprehensive binding prediction for all TFs in commonly used databases, visualization of transcription factor binding sites on UCSC human genome browser, as well as, integration of some popular databases and visualization of their results on UCSC genome browser

1.3 Structure of the thesis

In chapter 2, we will give a brief description of four popular PWM databases, as well as a general overview of transcription factor classifications. Then, in chapter 3 we will go over the methods and algorithms used in this thesis. In Chapter 4, we will explain the data processing steps taken during this project, followed by results in chapter 5. Conclusion and Discussion is presented in last chapter.

Chapter 2

Databases

Discovery of potential transcription factor binding sites helps greatly in study of regulatory regions in human genome. By the time of this writing, position-specific scoring matrices, that will be explained in more details in following chapter, are the most commonly chosen method to represent these binding sites [21]. In this work, we are using 4 popular databases currently available; such as SELEX, JASPAR, PBM, and publicly free version of TRANSFAC.

2.1 JASPAR

JASPAR is one of the largest freely available databases that represent TF binding sites using PWM (position weight matrix). It has had five major releases by the time of this writing. JASPAR CORE is the most used JASPAR collection. JASPAR CORE is a collection of non-redundant profiles of transcription factor binding site for multicellular eukaryotes. There have been many efforts in its creation to provide the most appropriate binding site profiles for each TF. There are some exceptions where we would see more than one profile per TF, for example, when there exists a clear difference in sequence or length such as Nkx2-5 or JUND, respectively [17].

JASPAR supports GUI features and users can browse, search, subset, and download, as well as tools supporting sequences search, matrix clustering.[17] The newest edition of JASPAR has seen 30% increase with 135 new PFMs added to the database and 43 updated PFMs. Following steps has been performed by Mathelier, et. al [17] to create these updates: 1) Compilation of sequence specific DNA binding TF ChIP-seq data collection to PAZAR database as well as association of Homo sapiens, Mus musculus, Drosophila melanogaster, and Caenorhabditis elegans from TF ChIP-seq datasets from the ENCODE and modeENCODE. 2) Bound regions were taken from above

studies, also by use of MEME suit; over-represented motifs which are close to ChIP-seq peak max position were identified. 3) Position Frequency Matrix of TF binding site profiles were created.[17]

Subset	Number of non-redundant profiles in JASPAR 4.0	New non-redundant profiles in JASPAR 5.0	Updated profiles	Removed profiles	Total profiles (including older versions of profiles)	Total profiles (non-redundant)
Vertebrates	130	74	36	1	260	202
Plants	21	43	3		67	64
Insects	123	8	4	1	136	131
Nematodes	5	10			15	15
Fungi	177				177	177
Urochordata	1				1	1
Total	457	135	43	2	656	590

Figure 2.1: Summary of content and growth of JASPAR CORE database, http://www.researchgate.net/publication/258314698_JASPAR_2014_an_extensively_expanded_and_updated_open-access_database_of_transcription_factor_binding_profiles [17]

As previously mentioned, very few TFs have multiple binding profiles, such as JUND and JUN where two new profiles were added. These both are taken from same ChIP-set dataset. Also, a new profile for Nkx2-5 which is derived from ChIP-seq data as well is added. This new profile is distinct from a profile taken from in vitro SELEX experiment, however, includes the same binding features of Nkx2-5. It is important to note that the exceptional reason for this redundancy is the substantial deviation in binding abilities of TF which requires more than one PFM to represent it [17].

2.2 SELEX

SELEX stands for systematic evolution of ligands by exponential enrichment. It uses PCR amplification to enrich small groups of bound DNAs from a random sequence pool, which helps determining the binding capabilities of TFs in vitro[12]. SELEX uses affinity-tagged proteins, barcoded selection oligonucleotides, and multiplexed sequencing, which optimizes the parallel study of many transcription factors. Then, new bioinformatics tools help in analyzing hundreds of thousands of sequencing reads to maintain the quality of experiments and generate motifs for the transcription proteins [12]. To generate position weight matrix (PWM), SELEX firstly, assumes PWM is a representation of multinomial distribution. Hence, to estimate how this model is represented in site j of PWM, the consensus as well as the other

three sequences which can be obtained by replacing the j -th base with each of the other bases is taken. Then the frequencies of all the four nucleotides in DNA-reads results in an unbiased estimation of PWM of j -th site. This corresponds to maximum likelihood estimation of multinomial parameters [12].

2.3 UniPROBE

UniPROBE stands for universal PBM Resource for oligonucleotide binding evaluation, and it is another open source database for in vitro capabilities of transcription factors binding sites on DNA. It uses protein-binding microarray technology and provides binding preferences for all variations of k -mers. Hence, this database main focus is on proteins and their binding capabilities on DNA, which it can be seen in forms of either k -mers, PWM, or graphical sequences logos. Some algorithms used to create PWM data for this database are Seed and Wobble, BEEML-PBM [11].

One interesting feature of UniPROBE is its new pipeline for depositing PBM dataset, compared to its old inefficient method of manually entering data to MySQL. This web based pipeline includes several scripts which automate the process. For example, the user can create a spreadsheet file to input the information into the database. Then he/she needs to create a folder for all their data files that are planned to be shown publicly and upload this folder as a zip file into UniPROBE server. Finally, the data will be integrated into web interface by creating sequence logos of proteins and making all the data searchable and available for download as well as some administrative work of putting the data into public interface.[11]

2.4 TRANSFAC

TRANSFAC is yet another transcription factor database for eukaryotes. In here, we use the publicly available free TRANSFAC database which was initially created a decades ago to model factor-site interaction [18]. TRANSFAC database is controlled using relational database and updates will be released through the web interface. It has six main files: 1) FACTOR contains TF interactions, 2) SITE includes DNA-binding sites of FACTOR: genomic sites, sites synthesized in the laboratory randomly with no previous knowledge about their association to a gene, and IUPAC consensus sequences, 3) GENE includes targeted genes regulated by SITE of FACTOR. 4) CELL contains the link to factor source. 5) MATRIX, hence, includes PWM, and 6) CLASS is a classification of transcription factors based on their DNA-binding

domain. Finally, it's worth mentioning the diversity of data in all databases mentioned here, where they mostly cover eukaryotic organisms from humans to yeast [18].

2.5 Classification

The different methods of classifying TF data was mentioned in chapter 1. Here, we explain classification based on DNA-binding domains. TFClass [30] is a comprehensive database classification of human genome TF by consideration of their DNA-binding domains. TFClass uses six levels for classifying the TF; four levels are based on different criteria used, level five deals with TF genes and level six represents individual gene products. It consists of nine superclasses, 40 classes, and 111 families. 1558 human TFs when counted by genes, or >2900 different TFs when including their isoforms have been classified.

UniProt is the source of domain assignments, protein sequences, as well as isoform information. TRANSFAC with 2012 update was also used for information about isoforms [30].

2.5.1 Rank definitions and contents

Ranking classification inspires from taxonomy of biological species as well as enzyme catalog. Similar to enzyme catalog, there is four levels of taxonomy in TFClass; including rank superclass, class, family, and subfamily, with subfamily being an optional section [30]. Superclasses are chosen based on general topology of their DBDs and how they interact with the targeted sequence of DNA. Classes however are looking at structural and similarities between sequences and was the primary level defined in TRANSFAC database. Finally families are differentiated based on their DNA-binding domains similarities, as it is a perceived assumption that DBDs with similar sequences may interact with more closely related DNA sequences.[30]

The aim of this classification was to create a comprehensive list with all human TFs with DBD or all that could possibly have DNA binding specific sites. Hence 10 superclasses including Superclass '0' , and 40 classes were defined. This results in 111 families with 336 subfamilies. Superclass 2 which represents zinc-coordinating DBDs, is greatly larger than other superclasses with having 53% of all transcription factors. The second largest superclass is helix-turn-helix by 26% of all TF genes, followed by 11% of basic domain factor genes [30].

8 Superclass: β -Sheet binding to DNA		Class descr
8.1 Class: TATA-binding proteins		Class descr ?
8.1.1 Family: TBP-related factors		TATAAAA
8.1.1.0.1	TBP	U T00794
8.1.1.0.1.1	TBP, isoform 1	U T08520
8.1.1.0.1.2	TBP, isoform 2	U T28040
8.1.1.0.2	TBP-L1 (TRF2)	U T10643
8.1.1.0.3	TBP-L2 (TRF3)	U T20139
8.2 Class: A.T hook factors		Class descr ?
8.2.1 Family: HMGA factors		GGAAATT
8.2.1.0.1	HMGA1 (HMGI(Y))	U T02368
8.2.1.0.1.1	HMGA1a, isoform HMGI-I	U T01851
8.2.1.0.1.2	HMGA1b, isoform HMGI-Y	U T01980
8.2.1.0.1.3	HMGA1c, isoform HMGI-R	U T19005
8.2.1.0.2	HMGA2 (HMGI-C)	U T24219
8.2.1.0.2.1	HMGA2, isoform 1 (HMGA2a)	U T01859
8.2.1.0.2.2	HMGA2, isoform 2 (HMGA2f)	U T26144
8.2.1.0.2.3	HMGA2, isoform 3 (HMGA2d')	U T30273
8.2.1.0.2.4	HMGA2, isoform 4 (HMGA2d)	U T30274
8.2.1.0.2.5	HMGA2, isoform 5 (HMGA2c')	U T30275
8.2.1.0.2.6	HMGA2, isoform 6 (HMGA2c)	U T30276
9 Superclass: β -Barrel DNA-binding domains		Class descr
9.1 Class: Cold-shock domain factors		TFC ?
9.1.1 Family: Dbp factors		CCAATCAG
9.1.1.1 Subfamily: A (DbpA-like)		
9.1.1.1.1	DbpA (YBX3)	U T10811
9.1.1.1.1.1	DbpA, isoform 1	U T02498
9.1.1.1.1.2	DbpA, isoform 2	U T18886
9.1.1.1.1.3	DbpA, isoform 3	U T00185
9.1.1.2 Subfamily: B (DbpB/YB-1-like)		
9.1.1.2.1	YB-1 (DbpB, YBX1)	U T10348
9.1.1.3 Subfamily: C (FRG Y2-like)		
9.1.1.3.1	YBX2 (DbpC, MSY2, FRG Y2)	U T23997
0 Superclass: Yet undefined DNA-binding domains		Class descr
0.1 Class: AXUD/CSRNP domain factors		Class descr
0.1.1 Family: CSRNP factors		AGAGZG
0.1.1.0.1	CSRNP-1 (AXUD1, TAIP3)	U T27220
0.1.1.0.2	CSRNP-2 (TAIP12)	U T27231
0.1.1.0.3	CSRNP-3 (TAIP2)	U T27291
0.1.1.0.3.1	CSRNP-3, isoform 1	U T27290
0.1.1.0.3.2	CSRNP-3, isoform 2	U T27292
0.2 Class: NonO domain factors		Class descr
0.2.1 Family: NonO-related factors		
0.2.1.0.1	NONO (NMT55, p54(nrb), NRB54)	U T08172
0.2.1.0.1.1	NONO, isoform 1	U T08892
0.2.1.0.1.2	NONO, isoform 2	U T28015
0.2.1.0.2	SFPQ (PSF)	U T08171
0.2.1.0.2.1	SFPQ, isoform Long (A)	U T08891
0.2.1.0.2.2	SFPQ, isoform Short (F)	U T25770
0.2.1.0.3	PSPC1 (PSP1, Paraspeckle component 1)	U T27237
0.2.1.0.3.1	PSPC1, isoform 1 (PSP1- α)	U T27235
0.2.1.0.3.2	PSPC1, isoform 2 (PSP1- β)	U T27236

Figure 2.2: Sample of TFClass, http://www.edgar-wingender.de/huTF_classification.html [5]

Chapter 3

Methods

3.1 Representation of DNA binding sites

One of the problems leading to development of computer algorithms for prediction of DNA binding sites is discovery of new binding sites using a representation of known binding sites. One general method that is widely used for the this, is consensus sequence. Consensus sequence as its name suggests is a sequence that matches all the bases from various examples closely but not exactly. The compromise exists between number of mismatches versus the ambiguity of consensus sequences. A higher number of mismatches allows for identification of more sites in price of accuracy. Hence consensus sequence might not be the best method to represent a model for predictin new binding sites [24].

A	-38	19	1	12	10	-48
C	-15	-38	-8	-10	-3	-32
G	-13	-48	-6	-7	-10	-48
T	17	-32	8	-9	-6	19

Figure 3.1: consensus sequence of TATAAT <http://bigscience.uncc.edu/bioinformatics-seminar-november-14-2pm/Bioinf-00.pdf> [24]

Position weight matrix is an alternative method to represents sites. This is

a matrix containing an element for each base in every position of the site. We can calculate the score of sequence by summation of its site sequence. Other sequences which are different from consensus sequence will also get a score lower than original score, hence, more conserved areas would have higher scores and the assumption is that, these areas are more important for the given site activity. For example, in figure 3.1, consensus of TATAAT is given in boxes and represents score of 85, and any sequence with different consensus will have lower score. Some noteworthy points are; 1) consensus sequence can always be transformed to weight matrix, however the converse is not possible. 2) although we calculated the position weight matrix, a threshold of the PWM to detect binding sites is needed, and 3) how does one go about choosing the elements in the weight matrix for the site representation [24]. There has been large amount of work done in scientific community in relation with these weight matrix, and in previous chapter, we mentioned four databases that could be used for the third point. We will explain a solution for the second point in this chapter.

3.2 PWM

Consider a matrix $W(b, i)$, where b represents all bases ($b = A, C, G, T$) and i is the position ($i = 1 \dots L$), where L is the length of protein binding sites. Hence, one can easily sum elements of W to get a score for any sequence of L length. For example, consider a L -length sequence S_j which is represented as a matrix such that it has 1 at each position for the base that occurs there and 0 for the rest. The score of S_j is the following: [25]

$$Score(S_j|W) = \sum_{i=1}^L \sum_b W(b, i) S_j(b, i) \quad (3.1)$$

PWM is a generalized form of consensus sequence, but also PWM can offer an advantage to consensus sequence by providing penalties for each position individually, based on their distance to consensus sequence instead of a general score for all mismatches. Which is important, as some positions often are more relevant to a binding sites either positively or negatively and PWM enables a simple model to represent these differences. However it is important to note that PWM does not record the cause of base preferences in each position rather it only contains the quantitative differences of each contributing base at each position [25].

3.2.1 Modeling from qualitative binding site data

Three main types of classification exists for estimating PWM based on collected sites. One considers positive and negative examples and hence discriminant models can be found where positive sequences have higher score in comparison to every negative sequence. The other approach is when we only have positive examples (this is the method used in this work) and one can use probabilistic approach for score detection. These scores represent an estimate of the probability that one sequences represents a set of sites or non-sites (the background) [25]. Probabilistic model started by aligning a collection of many binding sites for a specific transcription factor, then their alignment is used for constructing position frequency matrix, which as its name suggests, for each position gives the frequency of each base from the alignments. For N sequences, we will have:

$$F(b, i) = 1/N \sum_{j=1}^N S_j(b, i) \quad (3.2)$$

These alignments, can also be used to create a sequence logo for graphical representation of consensus sequence. For example, figure 3.2 is the graphical representation of BRCA1 taken from Jaspard database.[1]

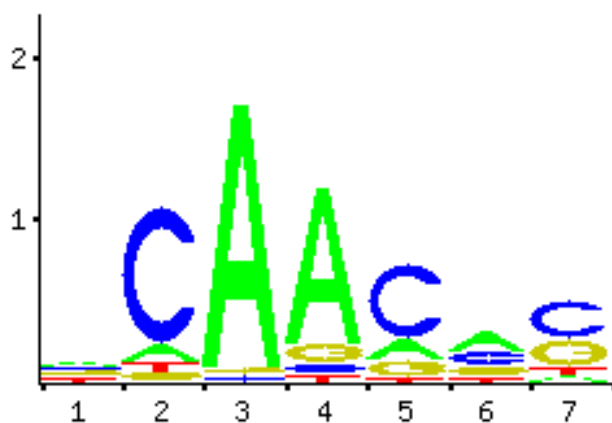


Figure 3.2: BRCA1 sequence logo http://jaspar.genereg.net/cgi-bin/jaspar_db.pl[1]

PWM is then created by dividing each base probability by background probability and transforming the value to log-scale:[28]

$$W_{bi} = \log_2 p(b, i) / p(b) \quad (3.3)$$

Summation of these W_{bi} elements, results in the score for a given sequence, as shown in equation 3.1[28].

3.2.2 PWM threshold

One of the PWM challenges is detection of a good threshold for its binding capabilities. We use a inferential statistics model to detect these thresholds as it enable us to use a sample of data to represent the behavior in whole population. For binding site prediction we use hypothesis testing (significance testing). Below are the steps taken in this work to detect the threshold:

1. Hypothesis or claim is that there is no binding sites for a PWM
2. We used equation given in 3.3 with uniform distribution as background probability, to calculate PWM thresholds for one million random sequences taken from each chromosome of human genome, resulting in 24 million threshold scores for each PWM.
3. Finally, we used a cut off P-value= 0.0001 to detect the likely threshold.

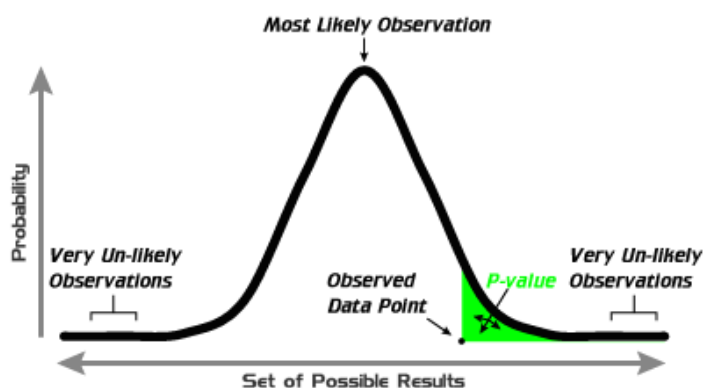


Figure 3.3: p-value is the probability of an observation happening by chance http://www.scottbot.net/HIAL/wp-content/uploads/2013/04/P-value_Graph1.png[2]

3.3 Human DNA-binding sites prediction and visualization

HOMER (Hypergeometric Optimization of Motif EnRichment), is a software used in Motif discovery and next generation sequence studies [10]. In this work, Homer is used to predict instances of a motif on human genome. Homer creates Bed files containing locations of motif instances on the whole human genome which can be converted to bigbed files and shown on UCSC genome browser[14]. Hence, a user can easily observe binding-site predictions by exploring the human genes visually using the UCSC genome browser.

3.4 Similarity algorithm

In this project, we have used PWMs as the representation of motifs which are simple yet flexible methods to represent motifs. After choosing the method used for representing motifs, one needs to consider ways to analyze the PWM information for motif studies. One question often appears when working with motifs discovery is whether the new motif is similar to any known motifs. Hence many works have focused on this topic. In this work, however, our question was; how we can integrate the information collected from four previously mentioned databases into one. Here we will explore three methods for that purpose.

3.4.1 Gupta et al.

Gupta et al. [8] did a comprehensive study on different motif vs motif similarity algorithm and created an algorithm called Tomtom. It works by defining a comparison function $S(Q, T)$ where Q and T are two motifs, and smaller $S(Q, T)$ shows the higher similarity between Q and T . Given a comparison function for columns of two motifs (As there has been many work on column comparison: PCC, ALLR, PCS, FIET, KLD, ED, or SW [8]), the question changes to how to use this comparison function to answer motif similarities. Two issues need to be addressed for this question; firstly, as we do not know whether the Q and T lie on one DNA strand or different ones, all various offsets and relative orientations needs to be considered for the motif similarity function, second is column comparison functions return many scores which need to be converted into one final score.

Gupta et al. simplified the questions by assuming a given column comparison function as well as presuming a correct relative offset and orientation for

Q and T , which have equal width (w) and similar orientation with relative offset of zero. They also agreed P-value can be taken by summation of scores from column comparison, as they assumed independency of columns in the motifs. Then they calculated the null distribution of summation of all similarity scores in relation to Q motif using a dynamic programming method. In short, a score function (Q_i, a) returns a positive integer for similarity comparison between i th column of Q and letter a , where $a \in A$. Given we have PDF (probability density function) of first matches in i position of Q , PDF of $A^{(i+1)}$ can be calculated by following:[8]

$$A^{(i+1)}(x) = \sum_{a \in A} A^{(i)}(x - \hat{S}[a, Q_{i+1}])P_a \quad (3.4)$$

x are indices of an array A which correlates with score function results and shows the preferred PDF. P_a refers to null probability of a , $A^{(0)}(0) = 1$ and $A^{(0)}(x) = 0$ are the initial steps of the recursion, and $i = 1 \dots w$ results in PDF of the motif matching to a random sequence. Then under the null hypothesis these can be used to find cumulative probability distribution and the corresponding P-values [8].

3.4.2 Habib et al.

Habib et al. noticed that the commonly used scores for column comparison between motifs cannot differentiate between identical composition in column alignments, for example, ED gives the same perfect score for the figure 3.4 a and b as well as d and c , while it can be easily seen one represents a more informative relation than the other one.[26]



Figure 3.4: Presenting problem of informative and uninformative columns <http://bioinformatics.oxfordjournals.org/content/early/2011/05/04/bioinformatics.btr257.full.pdf>[26]

Habib et al. hence, proposed a new score; Bayesian Likelihood 2-Component (BLiC), which has a Bayesian information criteria that penalizes the similar matches that are close to background distribution. They also assumed motif columns are independent of each other and similarity score can be identified by summation of similarity scores of aligned positions. Then, Bayesian estimator with a Dirichlet mixture prior was used to get each position probability for each DNA nucleotide, which helps to create a model of nucleotide distribution for various binding site positions [9]. BLiC score has two elements, 1) measures the similarity of distribution between two motifs, and 2) shows the difference of yielded similarity distribution to the background model.

$$BLiC_{score} = \log \frac{Pr(m_1, m_2 | common - source)}{Pr(m_1, m_2 | independent - source)} + \frac{Pr(m_1, m_2 | common - source)}{Pr(m_1, m_2 | background)} \quad (3.5)$$

Common, independent, and source refer to distribution of nucleotides at each position. BLiC assumes independency of samples binding sites from the common distribution over nucleotides, hence, likelihood ratio of various source distribution for samples of binding sites can be evaluated [9]. Where m_1 and m_2 are the two motifs. The score can be calculated by summation of individual positions. In short, BLiC evaluates a marginal likelihood score, which is a measure of chance of the nucleotide count in each position of the given motif with a source distribution [9].

BLiC score removes the uninformative columns of false alignments. However that creates another problem: it contains a bias for choosing motifs that have a plentiful instances. Hence BLiC is prone to give a high score to any match irrelevant to the query motif given they have high number of instances. This in turn makes BLiC unsuitable for similarity function between motifs [26]

3.4.3 Tanaka et al.

The method from Tanaka et al. is the one used in this project. The authors eliminate BLiC score problems in two steps; 1) changing popular column similarity scores such as ED to prefer informative columns to uninformative ones, and 2) designing a more sophisticated model rather than independent and identically distributed (iid) model used by Tomtom to better penalize the uninformative columns of alignments. It is important to note that Tanaka et al. also maintained the retrieval accuracy of Tomtom while removing the uninformative alignments [26].

Tanaka et al. consider the unaligned section of a motif (these parts are not

often used by other tools, such as Tomtom.) They define a score for these parts to cause uninformative unaligned columns to contribute less than informative unaligned parts, which in turn, would prefer informative alignments better than uninformative ones. The below formula shows the similarity score of α which is the un-gapped alignment between Query (Q) and Target (T) motifs, and α alignment is evaluated by the offset between Q_1 and T_1 [26].

$$\Sigma_s(Q, T, \alpha) = \sum_{i=1}^{|Q|} S(Q_i, T_{i+\alpha}) \quad (3.6)$$

Previously the unaligned columns would not get a score and hence $S(Q_i, \emptyset) := 0$, but Tanaka et al. proposed a 'complete' version, which uses $Sc(Q_i, \emptyset) := m_i$ for unaligned columns, Sc represents similarity score of complete version, and m_i refers to median score when aligning randomly Q_i to a target column ($S(Q_i, T) : T \in T$). Hence, average null score is assigned to random unaligned columns, which scales all alignments scores in the same level, because it considers all query columns for each alignment.[26]

Furthermore, as it was mentioned, most column comparison scores give the same score to any column with identical composition without considering their similarity to background sequence. Consider A and U as informative and uninformative columns respectively; $\max(S(U, T) : T \in T) = \max(S(A, T) : T \in T)$. However, these scores also satisfy $S(A, X) < S(U, X)$ (X being only X columns for any nucleotide base; A,C,G,T), and $\min(S(U, T) : T \in T) > \min(S(A, T) : T \in T)$. Therefore, m_i results in higher score for $Q_i = U$ versus $Q_i = A$ and this in turn causes a lower score for informative unaligned columns in comparison to uninformative unaligned ones, which finally yields more preference to informative aligned columns in Tanaka et al. algorithm. This new score is similar to Tomtom standard score and hence is used in Tomtom as 'complete score'. It is important to note for an aligned query column, the median null score is zero, which is consistent with the original scores. For example, old ED score is now referred to as 'complete-ED'. [26]

$$\Sigma_{sc}(Q, T, \alpha) = \sum_{i=1}^{|Q|} [S_c(Q_i, T_{i+\alpha}) - m_i] \quad (3.7)$$

By looking at different algorithms in this matter, Tanaka et al. algorithm proved to be the most appropriate solution for our work.

3.5 Hierarchical clustering

Our aim is to provide a meaningful method to analyze similarity scores taken from Tanaka's algorithm. Clustering offers the ability to cluster the motifs into groups to gain a better understanding of their relationship and the hierarchical clustering eliminates the necessity to decide on a number of clusters beforehand.

Firstly, hierarchical clustering searches for the most similar pairs in the data set which have the lowest rate of dissimilarity and then joins those two pairs into one in the dendrogram or clustering tree. This process is repeated until all data have been paired. The main challenge is to decide on the similarity and the dissimilarity of pairs. Some choices one could choose are: minimum(3.9), complete(3.10), or average(3.11) distance.

$$d(A, B) = \min(d(a, b)), a \in A, b \in B \quad (3.8)$$

$$d(A, B) = \max(d(a, b)), a \in A, b \in B \quad (3.9)$$

$$d(A, B) = \frac{1}{|A||B|} \sum d(a, b), a \in A, b \in B \quad (3.10)$$

In the complete linkage also known as the maximum method, samples are clustered based on their farthest elements in each group hence samples below a specific level have lower inter-dissimilarity than that particular level. For example, samples below 0.5 threshold would have a 50% similarity with each other which means they have more than a 50% co-presence of species. The minimum method, in turn, finds the closest elements in pairs of samples and hence only one element of the dissimilar pairs is less than the specific level. This method can yield heterogeneous clusters. The average method as its name suggests, takes the average of the dissimilarities in each step[7]. Other methods are e.g. ward, median, centroid, k-means.

Chapter 4

Implementation

This thesis project involved many data processing and programming work, here we discuss the overall steps needed for this project.

1. A total of 1844 vertebrate PWMs were collected from JASPAR (205), TRANSFAC (277), UniPROBE(519), and SELEX (236 used for representation on UCSC genome and 843 used for creating an integrated database) databases.
2. For each PWM a detection threshold was found using 24 million random sequences from 24 human chromosome
3. Cut off P-value= 0.0001 was chosen to find the single threshold for each PWM.
4. Motif files were created for each PWM.
5. Homer script was used to create BED files containing the likely locations of binding sites on human genome. Which further were manipulated to create BigBED files, that are smaller in size.
6. From BigBED files, a hub was created on UCSC genome browser, giving access to visualize all four databases simultaneously.

For the integrated database, following steps were implemented.

1. In this project, we used the TF classification discussed in previous chapters, to classify the proteins from all four databases into one, based on their family and sub family. We were not able to categorize 186 out of 1844 proteins into the families.

1 Superclass: Basic domains		TFC
1.1 Class: Basic leucine zipper factors (bZIP)		TFC
1.1.1	Family: Jun-related factors	TGAGTCA
1.1.1.1	Subfamily: Jun factors	TGAGTCA
1.1.1.1.1	JUN::FOS	U MA0099.2
1.1.1.1.2	BATF::JUN	U MA0462.1
1.1.1.1.3	JUN	U MA0488.1
1.1.1.1.4	JUN	U MA0489.1
1.1.1.1.5	JUNB	U MA0490.1
1.1.1.1.6	JUND	U MA0491.1
1.1.1.1.7	JUND	U MA0492.1
1.1.1.1.8	CRE-BP1:c-Jun	U CRE-BP1:c-Jun
1.1.1.1.9	Jun_Fos	U Jun_Fos
1.1.1.1.10	Jundm2_primary	U Jundm2_primary
1.1.1.1.11	Jundm2_secondary	U Jundm2_secondary
1.1.1.2	Subfamily: NF-E2-like factors	GCTGAGTCA
1.1.1.2.1	NF-E2	U NF-E2
1.1.1.2.2	NFE2L1::MafG	U MA0089.1
1.1.1.2.3	NFE2::MAF	U MA0501.1
1.1.1.2.4	Nfe2l2	U MA0150.2
1.1.1.2.5	NFE2_DBD	U NFE2_DBD
1.1.1.2.6	NRF1	U MA0506.1
1.1.1.2.7	NRF1_full	U NRF1_full
1.1.1.2.8	NRF-2	U NRF-2
1.1.1.2.9	Bach1::Mafk	U MA0591.1
1.1.1.2.10	Bach2	U Bach2
1.1.1.2.11	Bach1	U Bach1
1.1.1.3	Subfamily: ATF-2-like factors	TGACGTCA
1.1.1.3.1	ATF7_DBD	U ATF7_DBD
1.1.1.3.2	Creb5_DBD	U Creb5_DBD
1.1.1.3.3	CRE-BP1_1	U CRE-BP1_1
1.1.1.3.4	CRE-BP1:c-Jun	U CRE-BP1:c-Jun
1.1.1.3.5	CRE-BP1_2	U CRE-BP1_2
1.1.2	Family: Fos-related factors	(TGAGTCA)
1.1.2.1	Subfamily: Fos factors	(TGAGTCA)
1.1.2.1.1	JUN::FOS	U MA0099.2
1.1.2.1.2	FOS	U MA0476.1
1.1.2.1.3	FOSL1	U MA0477.1
1.1.2.1.4	FOSL2	U MA0478.1
1.1.2.1.5	Jun_Fos	U Jun_Fos
1.1.2.2	Subfamily: ATF-3-like factors	TGACGTCA
1.1.2.2.1	JDP2_DBD_1	U JDP2_DBD_1
1.1.2.2.2	JDP2_DBD_2	U JDP2_DBD_2
1.1.2.2.3	JDP2_full_1	U JDP2_full_1

Figure 4.1: TF classification for our proteins taken from the four databases

- Then, we analyzed each subfamily or family using Tanaka et al. new algorithm, which allowed us to create a distance matrix using q-values (minimal false discovery rate) taken from the similarity algorithm.
- Hierarchical clustering with complete linkage was then used to cluster our subfamilies. Figures 4.2 to 4.4 give some examples.
- The q-value was equal to 0.01 and from each cluster one PWM was chosen to represent the whole cluster.
- All chosen PWMs went to the previous steps mentioned for databases. and BED files for each family were put together. Hence, in UCSC genome browser, one can see the list of family names, instead of individual proteins. By visualizing the family, one can observe each protein representing different clusters of the family.

Here, some of the clustering figures are presented. Figure 4.2 and 4.3 show two zinc-finger GATA factors and Myc / Max factors subfamilies, respectively. These subfamilies have many clusters using our $q = 0.01$ value, but we were still able to eliminate many of proteins from each of the subfamilies. On the other hand, figure 4.4, shows HOX8 subfamily where no cluster group were present with 0.01 cut off.

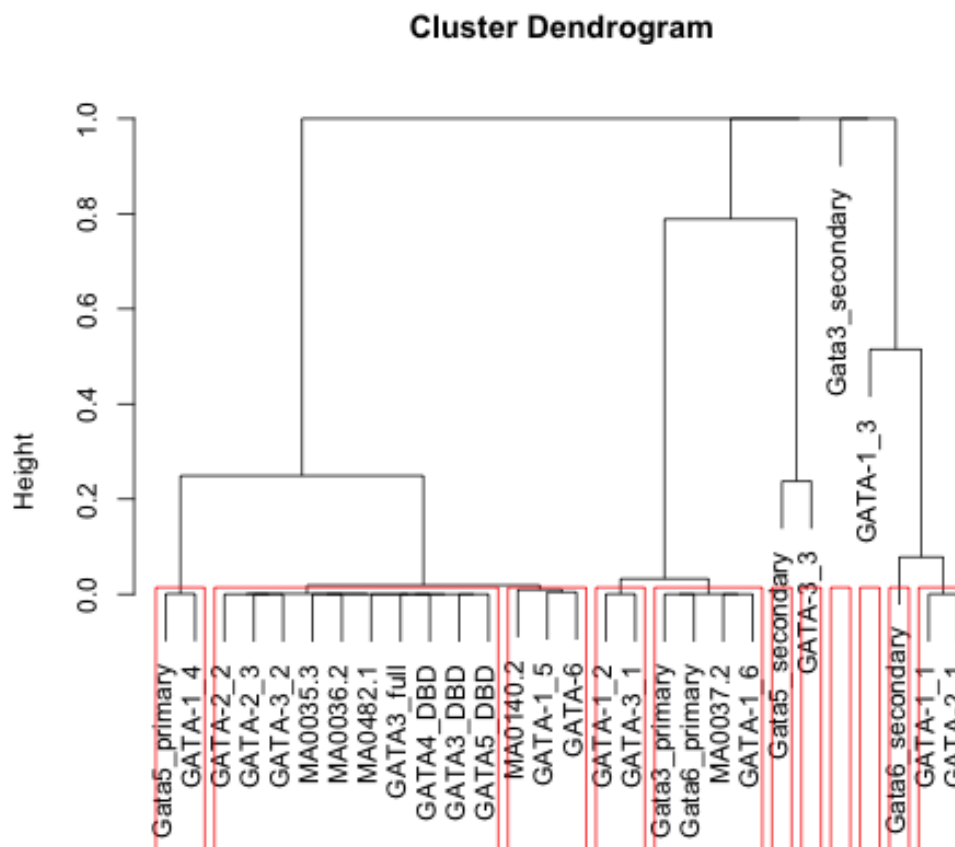


Figure 4.2: Subfamily of two zinc-finger GATA factors of GATA-type zinc fingers family from class 'other C4 zinc finger-type factors', superclass 2

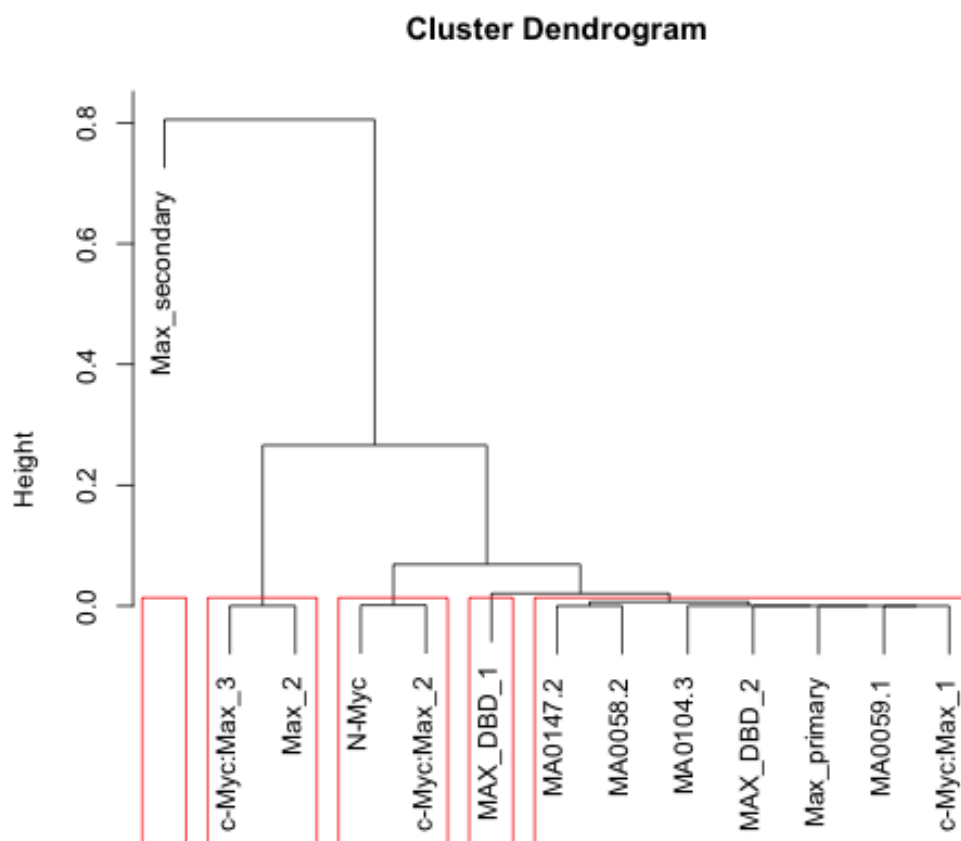


Figure 4.3: Myc / Max factors of bHLH-ZIP factors family and class of 'basic helix-loop-helix factors' (bHLH), superclass 1

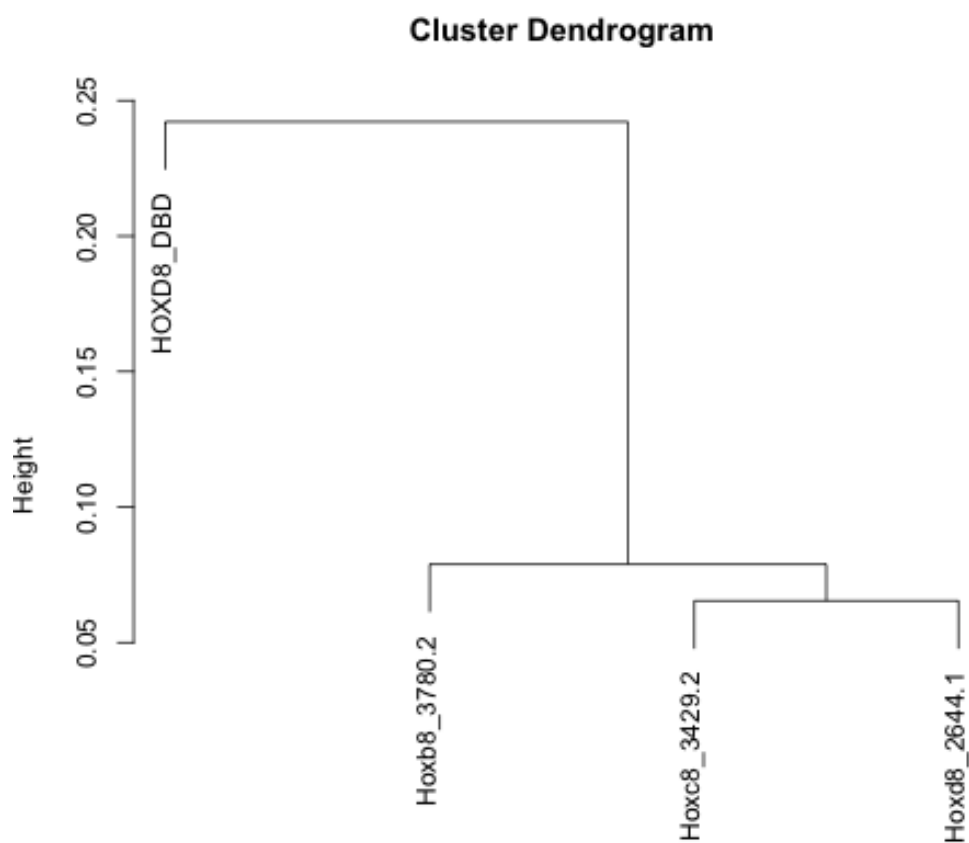


Figure 4.4: HOX8 subfamily from family of hox related factors and class of 'homeo domain factors', superclass 3

Chapter 5

Results

This thesis project tackles challenges biologists face to analyze transcription factor binding sites. To the best of our knowledge, prediction of putative TF binding sites is done individually for a given PWM using various softwares, this makes it cumbersome to analyze PWM datas simultaneously. Here, we presented a pipeline that enables simultaneous study of many PWMs in an easy and time efficient manner.

The figure 5.1 to 5.3 presents the outcome of this work. UCSC genome browser puts the data into tracks, one can hide, pack, squish, or get the full view of the tracks[13]. As it can be seen, datasets are shown by 5 different sections: 1) SELEX with 36 families, 236 motif. 2) JASPAR which has 205 motif tracks. 3) TRANSFAC with 277 motifs. 4) 519 motifs for PWM. And finally, 5) integrated database with 247 subfamilies of TF classification. Also, each databases has one track which contains all the protein datas in that database. This is useful for looking into a section of human genome to find all possible TF binding sites.

It is important to note that showing many tracks concurrently might slow down UCSC genome browser. Hence, using integrated database for analyzing subfamilies, could be a better solution given this issue. This is because the integrated database contains PWMs that best represent a subfamily irrespective of its original dataset. For example, figure 5.1 shows the R-SMAD subfamilies of integrated database and some of the proteins of SMAD family from each database. R-SMAD subfamily does not form any clusters in our clustering step and the integrated database 'R-SMAD' track contains all the PWMs from each of the databases. This can be observed by zooming in the tracks shown in the figure at different locations of the human genome. While, the other two figure (5.2 and 5.3) show the Ahr subfamily of integrated database, which forms 3 clusters in our clustering step, therefore, integrated dataset shows a reduced version of PWMs from the four database.

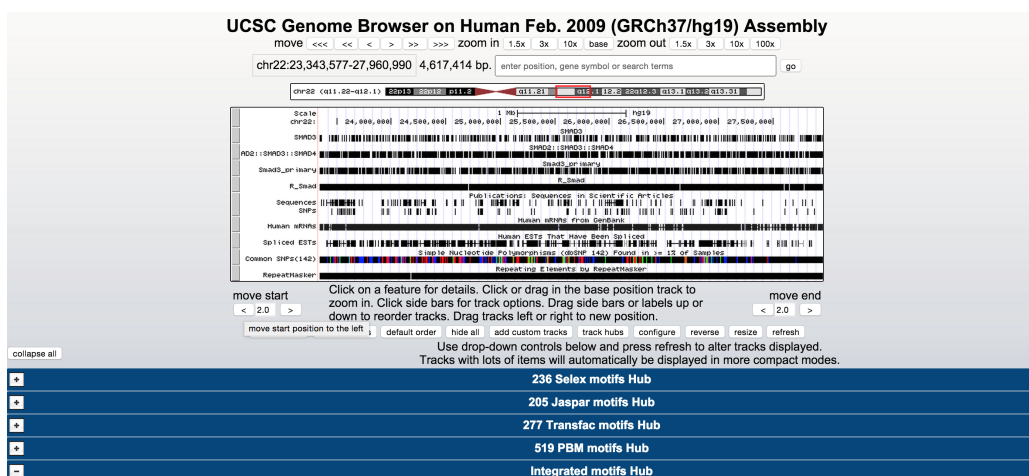


Figure 5.1: SMAD factors shown from JASPAR, PBM, and SELEX and Integrated database

Figure 5.3 shows a zoomed in version of figure 5.2, one can see a small black line at each position where there is a transcription factor binding site, with the name of its protein adjacent to the line. We believe this project has useful results that could help many biologist interested in knowing transcription factor binding sites of various PWMs. It enables anyone without any bioinformatician or programming knowledge to visualize TF binding sites on human genome. The data is hosted on Aalto computer resources and is accessible through links provided in the appendix of this project. There are 5 links; one for each database and one for the integrated database. This makes it quite easy for interested individuals to access and study the results as one only need to click or copy-paste the link on their internet browser to access the results.

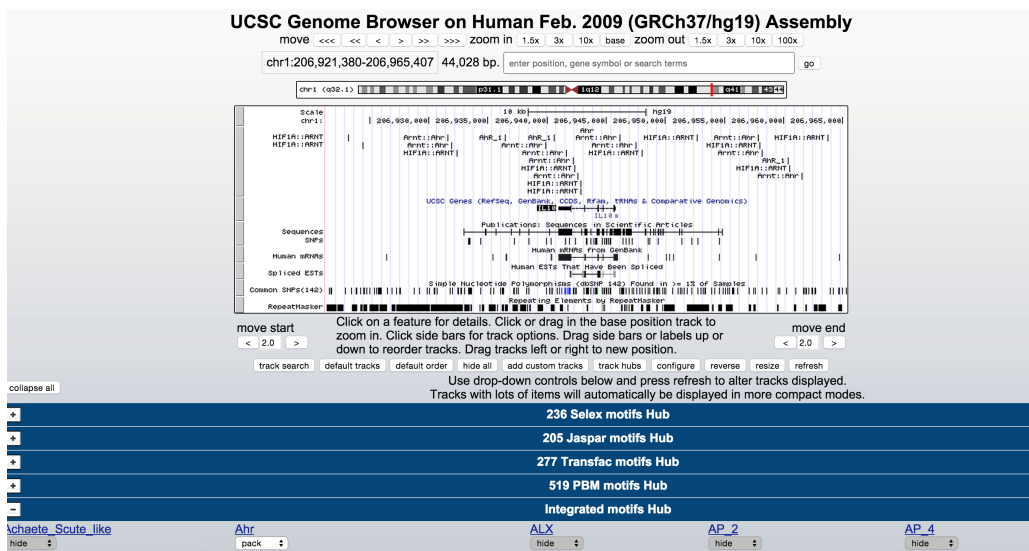


Figure 5.2: Ahr subfamily from integrated database

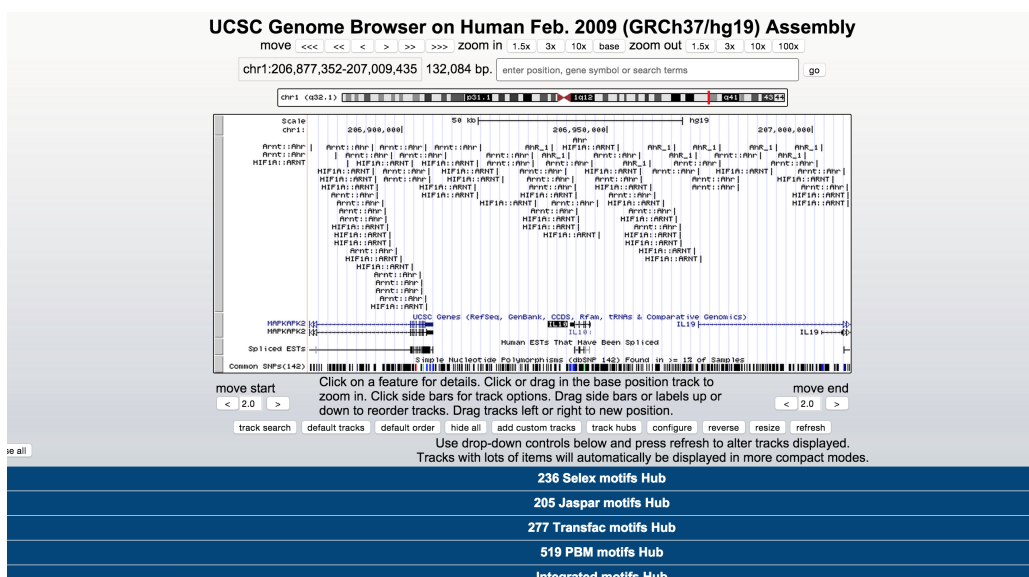


Figure 5.3: Ahr subfamily from integrated database

Chapter 6

Conclusion

In this project we worked with position weight matrices from JASPAR, TRANSFAC, SELEX, and PBM databases. We were able to use all these PWMs and predict transcription factor binding sites on whole human genome. Creating a result that can be used by scientists for effortless analyzation of potential DNA-binding sites. Furthermore, we integrated all four databases into one, using classification of human transcription factors[6]. Then, by performing a clustering on each subfamily, we aquired the PWMs that can represent the whole subfamily and were able to reduce the data.

Future work could look into more TF databases (only vertebrates data was used in this project) using the same pipeline as used in this project. Also, 168 proteins were not fitted in any TF subfamily; hence, it would be good to receive the help of biologist in classifying the remaining uncategorized data or one could use the same similarity pipline to classify them into TF classification subfamilies.

Bibliography

- [1] http://jaspar.genereg.net/cgi-bin/jaspar_db.pl-ID=MA0133.1&rm=present&collection=CORE. Accessed 20 May 2015.
- [2] http://www.scottbot.net/HIAL/wp-content/uploads/2013/04/P-value_Graph1.png. Accessed 20 May 2015.
- [3] BENIZRI, GINOUVÈS, AND BERRA. The magic of the hypoxia-signaling cascade. *Cellular and Molecular Life Sciences* 65, 7-8 (2008), 1133–1149.
- [4] BRIVANLOU, A. H., AND JR., J. E. D. Signal transduction and the control of gene expression. *Science* 295, 5556 (2002), 813–818.
- [5] DÖNITZ, J. Classification of human transcription factors. http://www.edgar-wingender.de/huTF_classification.html. Accessed 20 May 2015.
- [6] EVAN, GERARD, HARRINGTON, ELIZABETH, FANIDI, ABDALLAH, LAND, HARTMUT, AMATI, BRUNO, BENNETT, AND MARTIN. Integrated control of cell proliferation and cell death by the c-myc oncogene. *The Royal Society* 345, 1313 (1994), 269–275.
- [7] GREENACRE, M. Hierarchical cluster analysis. *SORT* 29, 1 (2008), 27–42.
- [8] GUPTA, SHOBHIT, STAMATOYANNOPOULOS, JOHN, BAILEY, TIMOTHY, NOBLE, AND WILLIAM. Quantifying similarity between motifs. *Genome Biology* 8, 2 (2007), R24.
- [9] HABIB, KAPLAN, MARGALIT, AND FRIEDMAN. A novel bayesian dna motif comparison method for clustering and retrieval. *PLoS Computational Biology* 4, 2 (2008).
- [10] HEINZ, BENNER, SPANN, BERTOLINO, AND ET AL. Simple combinations of lineage-determining transcription factors prime cis-regulatory

- elements required for macrophage and b cell identities. *Mol Cell* 38, 4 (2010), 576–589.
- [11] HUME, M. A., BARRERA, L. A., GISSELBRECHT, S. S., AND BULYK, M. L. Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein-dna interactions. *Nucleic Acids Research* 43 (2015).
- [12] JOLMA, KIVIOJA, TOIVONEN, CHENG, WEI, ENGE, AND TAIPALE. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Research* 20, 6 (2010), 861–873.
- [13] KENT, SUGNET, FUREY, ROSKIN, PRINGLE, ZAHLER, AND HAUSSLE. The human genome browser at ucsc. *Genome Res.* 12, 6 (2002), 996–1006.
- [14] KENT, ZWEIG, BARBER, HINRICHS, AND KAROLCHIK. Bigwig and bigbed: enabling browsing of large distributed data sets. *Bioinformatics* 26, 17 (2010), 2204–2207.
- [15] LATCHMAN. Transcription factors: an overview. *The International Journal of Biochemistry Cell Biology* 29, 12 (1993), 1305–1312.
- [16] LEMONS, D., AND MCGINNIS, W. Genomic evolution of hox gene clusters. *Science* 313 (2006), 1918–1922.
- [17] MATHELIER, A., ZHAO, X., ZHANG, A. W., PARCY, F., WORSLEY-HUNT, R., ARENILLAS, D. J., BUCHMAN, S., CHEN, C.-Y., CHOU, A., IENASESCU, H., LIM, J., SHYR, C., TAN, G., ZHOU, M., LENHARD, B., SANDELIN, A., AND WASSERMAN, W. W. Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research* 42 (2013), D124–D127.
- [18] MATYS, V., FRICKE, E., GEFFERS, R., GOSSLING, E., HAUBROCK, M., HEHL, R., HORNISCHER, K., KARAS, D., KEL, A. E., KEL-MARGOULIS, O. V., KLOOS, D., LAND, S., LEWICKI-POTAPOV, B., MICHAEL, H., MUNCH, R., REUTER, I., ROTERT, S., SAXEL, H., SCHEER, M., THIELE, S., , AND WINGENDER, E. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* 31, 1 (2003), 374–378.

- [19] MÜLLNER, D. The fastcluster package: User's manual. <http://cran.r-project.org/web/packages/fastcluster/vignettes/fastcluster.pdf>. Accessed 20 May 2015.
- [20] RANEY, DRESZER, BARBER, CLAWSON, FUJITA, WANG, NGUYEN, PATEN, ZWEIG, KAROLCHIK, AND KENT. Track data hubs enable visualization of user-defined genome-wide annotations on the ucsc genome browser. *Bioinformatics* 30, 7 (2014), 1003–1005.
- [21] SANDELIN, A., ALKEMA, W., ENGSTRÖM, P., WASSERMAN, W. W., AND LENHARD, B. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32 (2004), D91–D94.
- [22] SHAMOVSKY, AND NUDLER. New insights into the mechanism of heat shock response activation. *Cellular and Molecular Life Sciences* 65, 6 (2008), 855–861.
- [23] STEGMAIER, P., KEL, A. E., AND WINGENDER, E. Article: Systematic dna-binding domain classification of transcription factors. *Genome Inform.* 15, 2 (2004), 276–286.
- [24] STORMO, G. Dna binding sites: representation and discovery. *Bioinformatics*, 16 (2000), 16–23. Accessed 20 May 2015.
- [25] STORMO, G. Modeling the specificity of protein-dna interactions. *Quant. Biol.* (2013), 115–130. Accessed 20 May 2015.
- [26] TANAKA, EMI, BAILEY, TIMOTHY, GRANT, E., C., NOBLE, STAFFORD, W., KEICH, AND URI. Improved similarity scores for comparing motifs. *Bioinformatics* 27, 12 (2011), 1603–1609.
- [27] THOMAS, M. C., AND CHIANG, C.-M. The general transcription machinery and general cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, 41 (2006), 105–178.
- [28] WASSERMAN, W., AND SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* 5 (2004), 276–287. Accessed 20 May 2015.
- [29] WEBER, L. W., BOLL, M., AND STAMPFL, A. Maintaining cholesterol homeostasis: sterol regulatory element-binding proteins. *World J Gastroenterol* 10, 21 (2004), 3081–3087.

- [30] WINGENDER, E., SCHOEPS, T., , AND DONITZ, J. Tfclass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research* 41 (2013).

Appendix A

UCSC genome browser links

- 1) http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr22:25916196-28832390&hubUrl=http://research.ics.aalto.fi/csb/share/Jaspar_Transfac_PBM_ucscgenomebrowser/Jaspar/hub.txt
- 2) http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr22:25916196-28832390&hubUrl=http://research.ics.aalto.fi/csb/share/Jaspar_Transfac_PBM_ucscgenomebrowser/Transfac/hub.txt
- 3) http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr22:25916196-28832390&hubUrl=http://research.ics.aalto.fi/csb/share/Jaspar_Transfac_PBM_ucscgenomebrowser/PBM/hub.txt
- 4) http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr22:25916196-28832390&hubUrl=http://research.ics.aalto.fi/csb/share/Jolma_DNA-BindingSpecificitiesOfHumanTranscriptionFactors_BEDfiles/hub.txt
- 5) http://genome-euro.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr22:25916196-28832390&hubUrl=http://research.ics.aalto.fi/csb/share/Jaspar_Transfac_PBM_ucscgenomebrowser/IntegratedDatabase/hub.txt