Transcription Factor Binding Specificity and Occupancy

Elucidation, modelling and evaluation

Caleb Kipkurui Kibet

Computer Science Department

Rhodes University

This thesis is submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Bioinformatics

Rhodes University April 2017

Abstract

The major contributions of this thesis are addressing the need for an objective quality evaluation of a transcription factor binding model, demonstrating the value of the tools developed to this end and elucidating how *in vitro* and *in vivo* information can be utilized to improve TF binding specificity models.

Accurate elucidation of TF binding specificity remains an ongoing challenge in gene regulatory research. Several *in vitro* and *in vivo* experimental techniques have been developed followed by a proliferation of algorithms, and ultimately, the binding models. This increase led to a choice problem for the end users: which tools to use, and which is the most accurate model for a given TF? Therefore, the first section of this thesis investigates the motif assessment problem: how scoring functions, choice and processing of benchmark data, and statistics used in evaluation affect motif ranking. This analysis revealed that TF motif quality assessment requires a systematic comparative analysis, and that scoring functions used have a TF-specific effect on motif ranking. These results advised the design of a Motif Assessment and Ranking Suite MARS, supported by PBM and ChIP-seq benchmark data and an extensive collection of PWM motifs. MARS implements consistency, enrichment, and scoring and classification-based motif evaluation algorithms. Transcription factor binding is also influenced and determined by contextual factors: chromatin accessibility, competition or cooperation with other TFs, cell line or condition specificity, binding locality (e.g. proximity to transcription start sites) and the shape of the binding site (DNA-shape). *In vitro* techniques do not capture such context; therefore, this thesis also combines PBM and DNase-seq data using a comparative *k*-mer enrichment approach that compares open chromatin with genome-wide prevalence, achieving a modest performance improvement when benchmarked on ChIP-seq data. Finally, since statistical and probabilistic methods cannot capture all the information that determine binding, a machine learning approach (XGBooost) was implemented to investigate how the features contribute to TF specificity and occupancy. This combinatorial approach improves the predictive ability of TF specificity models with the most predictive feature being chromatin accessibility, while the DNA-shape and conservation information all significantly improve on the baseline model of k-mer and DNase data.

The results and the tools introduced in this thesis are useful for systematic comparative analysis (via MARS) and a combinatorial approach to modelling TF binding specificity, including appropriate feature engineering practices for machine learning modelling.

- - I dedicate this thesis to my loving parents - -

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. It is being submitted for the degree of Doctor of Philosophy at Rhodes University. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Caleb Kibet

Caleb Kipkurui Kibet
Appril 2017

Acknowledgements

Personal Acknowledgements

This work could not have been possible without the financial, professional, spiritual and moral support from many people. Firstly, my heartfelt gratitude goes to my supervisor, Prof. Philip Machanick. You have been with me throughout my postgraduate studies, patiently shaping me into an independent researcher by providing guidance, encouragement and freedom to explore. Thank You!

From my undergraduate studies at Kenyatta University, I have met and interacted with various professionals, some who ended up being my mentors. Each of them shaped my scientific life; special mention to Dr Mugiira, who opened up the world of Biosafety and Dr Masiga, who introduced and mentored me in Bioinformatics. Thank You! To my friends and advisors, Dr George Obiero and Dr Rosaline Macharia, thank you for holding my hand in my first steps in Bioinformatics.

To Prof Ozlem Tastan Bishop for running the excellent MSc program and the Research Unit in Bioinformatics (RUBi). Your leadership, suggestions and resilience are worthy of emulation. Thank You! To all my RUBi colleagues, you have been a great company. To the senior Computer Science yellow lab members, Motebang, Jess and Brent, thank you for your support and friendship.

Special thanks to my dear parents, Wilson and Pauline Koros; sisters, Becky, Ruth, Naomi and Anne; and brothers, Seth and Joshua for their immense love and support through my studies. I acknowledge all my friends for their moral support and encouragement; special mention of Coli, Emmanuel and Clement. My gratitude goes Thommas and Hafeni for their friendship and support, and also helping out with the proof-reading sections of this thesis: you guys are awesome!

To my parents while in Grahamstown, pastor Moses Heshu, and wife Bridget, you provided a home away from home. May God bless you richly. To Grahamstown Believers Fellowship, your prayers and spiritual support ensured I remained focused on Christ throughout my studies. God bless you!

Lastly, but most importantly, I thank the Lord God Almighty. For guiding my steps throughout my research, and opening doors when all seemed stuck. Surely, “The Joy of The Lord has been my Strength”.

Financial Acknowledgement

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed, and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Table of contents

List of figures

List of tables

List of Abbreviations

Abbreviation	**Description**
AMA	Average Motif Affinity
AME	Analysis of Motif Enrichment
ANN	Artificial Neural Networks
auPRC	area under Precision Recall Curve
auROC	area under Receiver Operator Curve
BEEML-PBM	Binding Energy Estimation by Maximum Likelihood PBM
bp	base pairs
CB-MAR	Consistency-Based Motif Assessment and Ranking
CentriMo	Centrality of Motifs
ChIP	Chromatin Immunoprecipitation
ChIP-seq	Chromatin Immunoprecipitation and Sequencing
DBD	DNA Binding Domain
DBN	Deep Belief Networks
DNase-seq	sequencing of DNase I hypersensitive sites sequences
ENCODE	ENCyclopaedia of DNA Elements
GBM	Gradient Boosting Machines
gcPBM	genomic context Protein Binding Microarray
GOMER	Generalizable Occupancy Model of Expression Regulation
HelT	Helix Twist
HMG	High Motility Group
HT-SELEX	High Throughput SELEX
IC	Information Content
MARS	Motif Assessment and Ranking Suite
MEME	Multiple Expectation Maximization for Motif Elicitation
MET	Motif Enrichment Tool

MGW	Minor Groove Width
MLNN	Multilayered Neural Network
MNCP	Mean Normalized Conditional Probability
MTAP	Motif Tool Assessment Platform
PBM	Protein Binding Microarray
PFM	Position Frequency Matrix
ProT	Propeller Twist
PWM	Position Weigh Matrix
Slim	Sparse local inhomogeneous mixture model
SnW	Seed-and-Wobble
SVM	Support Vector Machines
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TFFM	Transcription Factor Flexible Models
TSS	Transcription Start Sites
UniPROBE	Universal PBM Resource for Oligonucleotide Binding Evaluation
uPBM	universal Protein Binding Microarray
WAM	Weight Array Matrix
XGBoost	eXtreme Gradient Boosting
XGB-TFBSContext	eXtreme Gradient Boosting using TFBS context information

Chapter 1

Introduction

> *"In a purely technical sense, each species of higher organism—beetle, moss, and so forth, is richer in information than a Caravaggio painting, Mozart symphony, or any other great work of art. Consider the typical case of the house mouse, Mus musculus. Each of its cells contains four strings of DNA, each of which comprises about a billion nucleotide pairs organised into a hundred thousand structural nucleotide pairs, organised into a hundred thousand structural genes. The full information therein, if translated into ordinary-sized printed letters, would just about fill all 15 editions of the Encyclopaedia Britannica published since 1768."*
>
> –Edward O. Wilson [200]

The cell is rich in complex information. Among the billions of nucleotide pairs, the *transcription factors (TFs)* – proteins that bind specific sites in the genome to facilitate the expression of genes to proteins – can still locate their degenerate binding sites to effect the expression of the right genes to produce the right proteins at the right time in the required levels. How do they achieve this? What additional information helps them locate these sites? This task is comparable to finding a needle in the haystack; that is what it has been to researchers interested in this question. The search for a code that describes how the TFs locate their binding sites remains an enduring challenge. It is clear that TFs use more information than is offered by the binding site's pattern; but, the code remains elusive. This thesis is about how the complex information in a cell, specifically about transcription factors binding, is utilised to ensure a TF binds only to a particular site and not to others. This is the elucidation and modelling of TF binding specificity – our area of focus.

This chapter highlights the research gaps addressed in this thesis, provides the aim and objectives, and finally an overview of the thesis. We provide a detailed introduction to the concepts and a review of previous work in the line of this research in Chapter 2.

1.1 Background

The study of transcription factor binding specificity, affinity and occupancy remain an active area of research. This attention is attributable to the key role played by TFs in the regulation of gene expression. A breakdown in the regulatory systems is associated with diseases, including cancer [17], autoimmunity, neurological disorders e.g. epilepsy, developmental syndromes, diabetes, cardiovascular disease, obesity, haemophilia and many others [190, 191, 105]. Therefore, an elucidation of transcriptional gene regulation and a deciphered TF regulatory code is important to understanding how TF binding or the failure thereof can be linked to diseases [58], and ultimately finding the cure for these diseases.

Despite such a significant role for TFs, and the efforts in gene regulation, a complete understanding and mapping of TF binding sites have been impeded by the complexity of the TF recognition mechanism [172, 42, 1]. Though techniques like Chromatin Immuno-Precipitation followed by deep sequencing (ChIP-seq) [77] provide an accurate picture of *in vivo* binding, widespread use is limited by cost and the lack of antibodies against TFs of interest. On the other hand, comprehensive *in vitro* techniques like Protein Binding Microarrays (PBM) [15] are limited in their ability to predict *in vivo* binding [140, 172], and the confounding effect of technology noise: sticky k-mers [76]. *In vitro*-based binding specificities do not capture the TF binding site environment's factors, which among others include chromatin accessibility, TF binding sites (TFBS) flanking sequences, presence of cooperating factors' motifs and proximity to transcription start sites (TSS). All this information is necessary for modelling TF binding specificity.

The presence of a TF binding site in an open chromatin site is a major determinant for binding, and the accessibility information is captured by techniques like the sequencing of DNase I hypersensitive sites sequences (DNase-seq) [173]. However, DNase-seq does not provide information on what binds to the location, only that it is accessible. Therefore, DNase-seq data is combined with other experimental data. Ability to combine DNase-seq and PBM data, given their strengths and weaknesses, would significantly improve our capacity to model TF binding specificity. This is the subject of Chapter 6, where we seek to combine PBM and

DNase-seq data to model TF binding specificity. Our initial attempts to do this were impeded by the lack of a standardised motif evaluation approach, which prompted us to shift our attention to the motif evaluation problem.

Transcription factor binding specificity is mainly modelled as a Position Weight Matrix (PWM) [178, 180]. PWMs are straightforward and easy to use; therefore, a majority of the algorithms that learn binding specificity models from the experimental techniques like ChIP-seq and PBM represent them as PWMs. We define the PWM motif evaluation problem as the lack of a standardised PWM evaluation benchmark, compounded by the presence of multiple instances of PWM models for a given TF. This problem is directly linked to the complexity of the motif discovery due to the complicated nature of TF binding; consequently, the many algorithms developed generate models deposited in the various PWM databases available. We address the motif evaluation problem through a systematic comparative analysis (Chapter 3) and develop a suite of tools for assessment (Chapter 4 and 5).

The PWM cannot fully explain TF binding specificity. In addition to sequence preference, the binding of a TF is also determined by binding site context information previously mentioned and the shape of the binding site, which unifies the degenerate sites [41]. The TF recognises DNA in its 3D conformation, formed by the binding site and flanking sequences [52]: DNA shape. The DNA shape information have been shown to improve binding prediction *in vivo* [121] and *in vitro* [1]. Also, TF binding sites are believed to be under evolutionary constraint [134] and are localised around TSS [93]. Therefore, to adequately model TF binding specificity, we need to take advantage of this information in its entirety.

However, the PWM cannot capture such information, although there have been efforts to extend [194, 215] or introduce new better models [120, 165, 87], including k-mers. None of these, however, can incorporate all the context information. Therefore, machine learning modelling has gained widespread use [99, 121, 103] due to its ability to utilise multiple heterogeneous data. But, the use of these datasets, either when predicting binding sites or modelling TF binding specificity, remain mixed, generating conflicting results. Therefore, in Chapter 6 we investigate how these datasets explain TF binding specificity and occupancy, and how best the data should be processed and used in a machine learning modelling environment, using eXtreme Gradient Boosting (XGBoost) [28].

This section only provides a quick review of the research gaps being addressed; detailed reviews are provided in Chapter 2 and the introduction sections of the respective chapters.

1.2 Problem statement

The project explores a combinatorial approach (using a variety of data sources and techniques) to elucidate, model and evaluate transcription factor binding specificity and occupancy.

1.3 Research hypothesis

Transcription factor binding is determined by sequence specificity as well as the contextual environmental factors of the binding sites – experimental techniques describing these factors are available. Therefore, elucidating TF binding specificity requires a combinatorial approach.

1.4 Research objectives

The principal objective of this study is to elucidate, model and evaluate transcription factor binding specificity and occupancy using data that describe the TF binding site contextual environment by employing statistical and machine learning techniques.

Specific objectives

1. To systematically review and analyse the motif assessment approaches in use and how they influence motif ranking (Chapter 3).

2. To develop tools for comparative motif assessment and ranking, and make them available as a web server (Chapter 4).

3. To evaluate the accuracy and relevance of the motif evaluation tools developed in 2 (Chapter 5).

4. To design an algorithm that combines *in vitro* and *in vivo*-derived data starting with DNase and PBM data to generate motif models that predict *in vivo* binding better than PBM-derived models (Chapter 6).

5. To elucidate transcription factor binding specificity and occupancy as it relates to chromatin accessibility, binding locality (e.g. proximity to transcription start sites) and the shape of the binding site (DNA-shape), and how they can be leveraged to model TF binding specificity using statistical and machine learning techniques (Chapter 6).

1.5 Technical contributions

The work presented in this thesis aimed at elucidating and modelling transcription factor binding specificity using a combinatorial approach. This is addressed on many levels as summarised by the specific objectives. By addressing the aims of this study, this thesis makes the following contributions:

1. We provide a systematic review, categorisation and analysis of the motif evaluation approaches in use. We categorise motif evaluation techniques into motif assessment by binding site prediction, comparison, and scoring and classification. We demonstrate the variability and lack of standardisation in the assessment algorithms in use; how the choice of scoring functions and processing of benchmark data affect motif ranking in a TF-specific manner.

    - Part of this work is published in *F1000Research* [92]. So far cited by two other papers.
    - A reproducible IPython notebook on this work available from GitHub[1].

2. We adopt criteria for designing a good benchmark and algorithms for motif evaluation. A good benchmark should be: *accessible* for easy use; *evolve* with changing techniques and data; be *independent*, not tailored to specific tools or technologies; *relevant*, generating biologically meaningful ranks; *scalable*, expand to new technologies; and *solvable*, not trivial but have a solution. Guided by these criteria, we collate a variety of experimental benchmark data, including an extensive collection of motifs. These are stored in a MySQL database (Chapter 4).

    Furthermore, we develop MARSTools, a suite of Python modules for motif assessment and ranking. They include a consistency-based motif assessment and ranking approach (CB-MAR) that is data-independent; therefore, it is not affected by the benchmark data used, solving the reference motifs bias. The MARSTools are made publicly available as a web server, MARS[2], a motif assessment and ranking suite. These tools are then evaluated in a motif assessment and discovery problems, demonstrating their benefit in both.

    - MARSTools can be obtained from GitHub[3].

[1] https://github.com/kipkurui/MARS_Evaluation/tree/master/Chapter3
[2] www.bioinf.ict.ru.ac.za
[3] https://github.com/kipkurui/MARSTools

- MARS webserver source code can be downloaded from GitHub[4].
- MARS_Evaluation, a collection of reproducible IPython notebooks for the evaluation of the MARSTools is downloadable from GitHub[5]. This includes test data.

3. We employ machine learning modelling in a combinatorial approach to elucidate and model TF binding specificity and occupancy. We demonstrate the benefit of a combinatorial approach and provide an understanding of how the contextual binding site environmental factors contribute to binding specificity. We also develop XGB-TFBSContext, an XGBoost based algorithm for predicting TF binding occupancy using 13 features.

   We also show that sticky k-mers are differentially enriched in open chromatin sites compared with the whole genome. We use this in a background noise and ranking approach to reweight the PBM intensity scores hence incorporating accessibility information to the TF binding specificity models generated.

   - XGB-TFBSContext is available from GitHub[6]
   - IPython notebooks demonstrating the reweighting and background correction approach from XGB-TFBSContext[7]

1.6 Thesis overview

This thesis comprises of 7 chapters, including this one. Chapter 2 introduces the genes, transcription factors, techniques used to study TFs and machine learning concepts. It also provides a review of the literature to highlight the research gaps addressed in this thesis.

Chapter 3 surveys the motif evaluation techniques. Moreover, it provides a systematic analysis of these techniques and highlights some of the drawbacks in the evaluation techniques. The results from this analysis motivate the need for a standardised motif assessment platform.

Chapter 4 introduces MARSTools, standalone tools for motif evaluation, and MARS, a web server for motif assessment and ranking.

Chapter 5 evaluates MARS to demonstrate the application of MARSTools in motif evaluation, as well as in motif discovery.

[4] https://github.com/kipkurui/MARS

[5] https://github.com/kipkurui/MARS_Evaluation/tree/master/MARS_Evaluation

[6] https://github.com/kipkurui/XGB-TFBSContext

[7] https://github.com/kipkurui/XGB-TFBSContext/blob/master/code/Combining%20PBM%20and%20DNase.ipynb

Chapter 6 introduces the combinatorial approach for elucidating TF binding specificity and occupancy. The first section deals with the use of PBM and DNase-seq data to model TF binding specificity. The second section employs an XGBoost-based machine learning modelling to demonstrate the how the contextual environmental factors describe TF binding occupancy, and finally introduces XGB-TFBSContext for modelling and predicting TF binding occupancy using ChIP-seq, PBM, DNAse-seq, DNA shape, proximity to TSS and evolutionary conservation features.

Finally, Chapter 7 concludes the thesis with a summary of our contributions, research limitations and a discussion of possible further research.

Chapter 2

Background and Literature Review

> *"All living organisms are but leaves on the same tree of life. The various functions of plants and animals and their specialized organs are manifestations of the same living matter. This adapts itself to different jobs and circumstances, but operates on the same basic principles. Muscle contraction is only one of these adaptations. In principle it would not matter whether we studied nerve, kidney or muscle to understand the basic principles of life. In practice, however, it matters a great deal."*
>
> –Albert Szent-Gyorgyi [182]

Indeed it does. The cell has to differentiate into the required type and produce the necessary proteins at the correct time and levels, to ensure the wheel of life keeps turning. What, then, ensures this precision? We answer this question, starting with the most basic component of life – DNA – then work our way from there. We investigate why, although DNA information in every cell is similar, it still matters, in practice, which cell we study or from which we extract this information. The main players that cause this variation are transcription factors (TFs) that bind to DNA at different rates, specificity, and affinity leading to the difference in rate and level of gene expression. The long-standing question in gene regulation is to identify where these TFs bind in the genome. This study focuses on this issue and investigates how a variety of experimental data can be leveraged to improve our ability to predict these binding sites.

This chapter reviews the main concepts and techniques important to this thesis. Also, it demonstrates where this study fits in gene regulatory research. Beginning with an introduction to biological concepts, it introduces TFs and some techniques used to investigate TF binding.

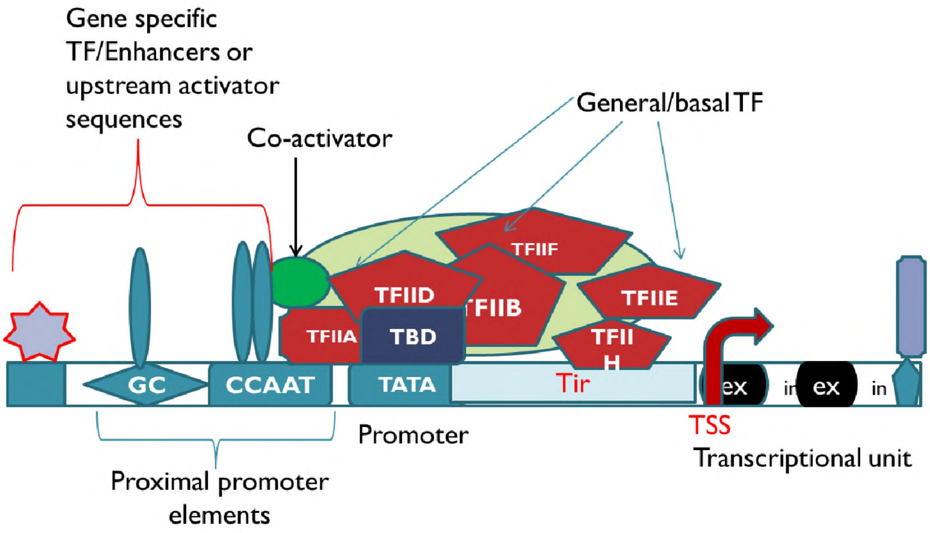


Fig. 2.1 Gene structure and regulatory unit. A simplistic illustration of the gene structure and regulatory unit organization. It shows the number of players involved in gene regulation, demonstrating the complexity of transcription. The sequence-specific TFs recruits the basal TFs, which in turn recruit the RNA polymerase. See Figure 2.2 for further details.

Finally, it provides an introduction to machine learning concepts, one of the techniques utilised in this study.

The genetic information that governs living organisms is contained in the deoxyribonucleic acid (DNA) made up of four bases (nucleotide, when combined with sugar backbone and phosphate group): adenine (A), guanine (G), cytosine (C) and thymine (T). DNA exists as a double helix coiled around a common axis stabilised by hydrogen bonds between specific base pairs, where ‘A’ binds to ‘T’ by two hydrogen bonds, and ‘C’ binds to ‘G’ by three hydrogen bonds. The gene is the basic unit of this genetic information, and it is made up of coding (exons) interrupted by non-coding (introns) regions, and regulatory regions made up of promoters and enhancers as shown in Figure 2.1. The central dogma of molecular biology stipulates that genetic information flows from the DNA to the ribonucleic acid (RNA), via transcription, and then finally to the proteins via translation. The first stage just copies the DNA material into precursor messenger RNA (pre-mRNA) but with ‘T’ replaced with uracil (U), and it takes place in the nucleus. The pre-mRNA then undergoes additional processing: splicing to remove the non-coding regions and adding to the 3’ polyA tail and to 5’ end a methylated cap to form the mRNA; transported to the cytoplasm where the protein is synthesised. The central dogma is illustrated as follows: DNA $\xrightarrow{transcription}$ pre-mRNA $\xrightarrow{processing}$ mRNA $\xrightarrow{translation}$ Protein.

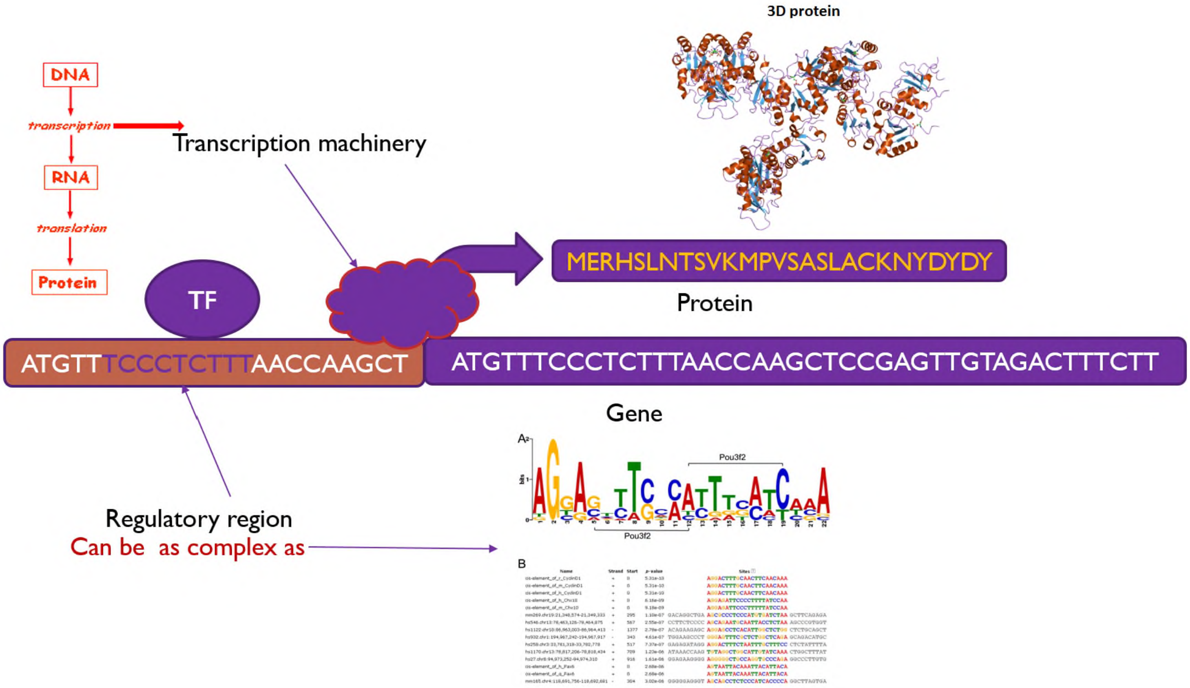


Fig. 2.2 Gene expression and translation. The transcription factors (TFs) play a role in transcription within the central dogma (red arrow). By binding the regulatory region – visualised as a sequence logo (see Figure 2.7 for details) – the TFs recruit the transcription machinery. This facilitates gene transcription opening the way for translation to the protein, visualised in its 3D conformation as shown. The motif logo (A) and sequence alignment (B) portion of the figure is adapted from [205].

The genetic information in every cell is similar, but the cells still achieve specific roles, expressing diverse proteins, at different levels as required at various points of development. Moreover, the proteins expressed by the cell far outnumbers the genes. The cell employs tight regulation at the different stages of the central dogma – gene regulation. For example, the number of genes is increased through alternative splicing, while the rate of transcription initiation, transport to the cytoplasm, post-translational modification, and many others regulates the rate and level of gene expression. Although, as already highlighted, the TFs are central to the regulatory process, gene regulation begins with DNA packaging within the cell.

As illustrated in Figure 2.3, the DNA is wrapped around histone proteins to form the nucleosome, the basic unit of the chromatin, and further packaged into the chromosomes. A collection of all the chromosomes, 46 in human beings, contains all the hereditary information of an organism. The packaging can be considered the first stage of regulation. The gene to be expressed must first undergo changes through the process of chromatin remodelling to allow access of RNA polymerase and TFs to their binding sites to activate gene expression.

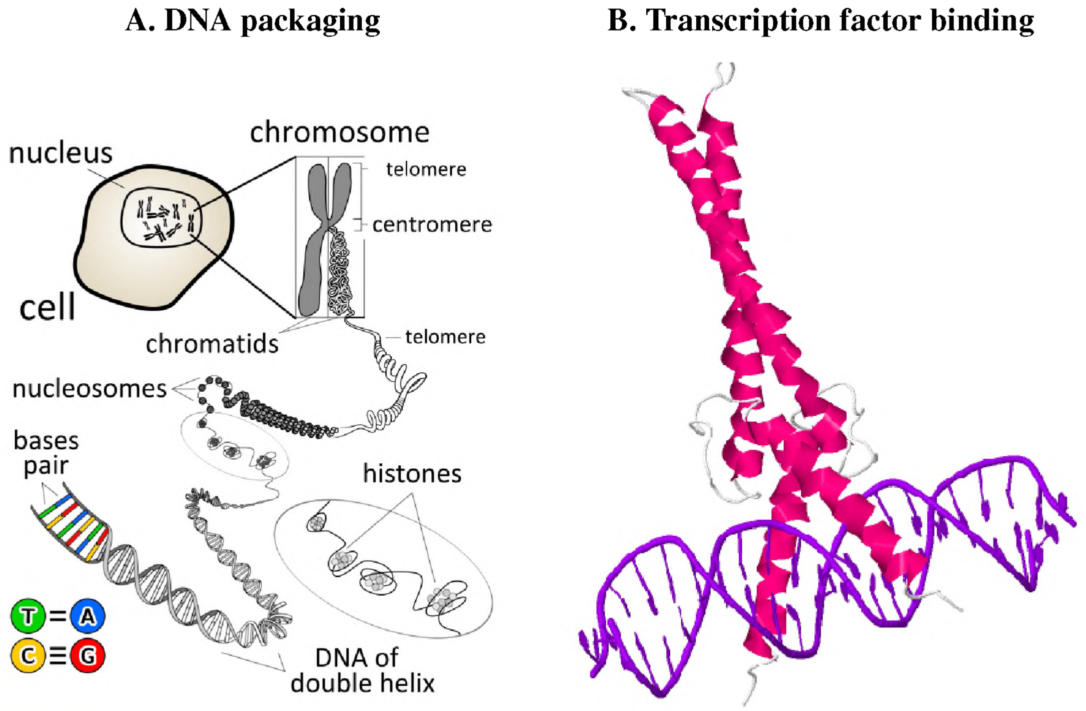


Fig. 2.3 DNA packaging and transcription factor binding. A: Illustration of DNA packaging in the nucleus. **B**: Max TF in pink recognises its DNA (purple) binding site through a dimeric B/HLH/Z domain. The cartoon is extracted using Jsmol [61] from the PDB [16].

Therefore, experimental techniques that provide data on the accessibility of a site are crucial to understanding TF binding and ultimately gene regulation. We introduce these techniques in Section 2.1.1.7.

We have presented the DNA and its packaging, the gene structure, the central dogma and the role of TFs in gene regulation. We now focus our attention to TFs: what they are, where and how they bind, their binding affinity, specificity and occupancy, and finally the experimental techniques used to understand and model TF binding.

2.1 Transcription factors (TFs)

In eukaryotes, regulatory proteins called transcription factors (TFs) initiate transcription by binding to conserved patterns of nucleotide sequence referred to as DNA *motifs* (transcription factor binding sites, TFBS) of the promoter as well as enhancer regions (Figure 2.2). The DNA genetic information is transferred to proteins via RNA by the process of transcription. Transcription requires the binding of RNA polymerase II and basal TFs to the core promoter region as shown in Figure 2.1. This binding modulates (activates or represses) the expression of nearby genes and regulates gene expression. Motifs bound by TFs vary considerably due to the degenerate nature of TFBS. Owing to this variability and the ubiquitous presence of these sites all over the genome, additional information is required to localise the TFs to the gene of interest regulatory sites.

The degeneracy of TF binding sites has been linked to their DNA binding domain (DBD) assuming various conformations during binding [171]. However, although the TFs, as in the ETS family, are similar, they still achieve specificity of binding and function due to minor divergence at DNA-contacting amino acid residues [196, 4]. Transcription factors are broadly classified as basal TFs, which recruit RNA polymerase, and the gene-specific TFs, which activate or repress basal TFs [31]. The gene-specific TFs are further classified into superclass, class, family, subfamily, genus and species, depending on the TFs' DBD and interaction mode with the sequence [201, 176, 202]. The primary superclasses that exist include basic domains, zinc-coordinating domains, helix-turn-helix domains, other all β-helical DBD, α-helices exposed by β-structures, immunoglobulin folds, β-hairpin exposed by an α/β-scaffold, β-sheet binding to DNA and β-barrel DBD. See [202] for complete details on TF classification. Unless otherwise specified, our use of the term *transcription factors* refers to gene-specific TFs.

Transcription factors bind to DNA via protein-DNA hydrogen bonds interacting between the TFs' DNA binding domain (DBD) and the TFBS – this is the *base or direct readout* [159, 172]. Also, the three-dimensional conformation of the DNA sequences within and around the binding site influences TF binding – this is the *shape or indirect readout* [159, 172]. See Figure 2.3A and Section 2.1.1.5. The shape information, therefore, unifies the degenerate sites into a similar shape [41] or provides specificity for TFs that recognise similar motifs. In addition to sequences within the binding site, flanking sequences are also responsible for the formation of a tertiary structure recognised by the TF [41, 52]. The shape readout, however, does not fully explain the specificity of TF binding. Additional players involved include active gene's TFBS exposed by chromatin remodelling [174], combinatorial interaction of nearby motifs [85, 119] and the distance of the putative binding site from the transcription start sites (TSS) of the gene of interest. One of the objectives of this thesis focuses on how to harness this additional information to understand TF binding and finally to model TF binding specificity.

Box 2.1: Key regulatory genomics concepts

Transcription factors (TFs): Proteins that bind to gene promoter and enhancer sites to regulate gene expression.

Sequence motifs: Conserved sequence patterns with biological function.

TF binding site: Sub-sequences or motifs in the genome recognised and bound by TFs.

TF binding affinity: The strength with which a TF binds a given TFBS.

TF binding specificity: The probability that a TF preferentially binds to a given TFBS and not to other putative sites.

TF binding occupancy: The likelihood that a TF will bind a given sequence. *In vivo*, this relates to a measure of TF's actual occupancy in a given genomic site. See Section 2.1.4.2 for more details.

Promoter: A genomic region proximal to a transcription start site bound by regulatory machinery for gene expression.

Enhancer*:* A regulatory region located a distance away from the genes they regulate.

Contextual environmental factors*:* These are the various TFBS environments' factors that determine TF binding specificity. These include chromatin accessibility, TFBS flanking sequences, presence of cooperating factors' motifs and proximity to TSS.

2.1.1 Transcription factor binding affinity, specificity, and occupancy

Transcription factor binding specificity describes the probability that a TF binds to a particular site and not the other putative sites [181, 79]. While *TF binding affinity* is a measure of how strongly a TF interacts with a binding site [7], a *TF binding occupancy* describes the probability that a TF will bind to a sequence [120]. These terms are defined in Box 2.1. Based on DBD, a TF's binding affinity is determined by the sequence and shape of the binding site, whereas a TF's binding specificity is described by a combination the affinities of the TF to all possible sites, in addition to binding site environmental factors. These include chromatin accessibility, the presence of binding sites for cooperating TFs, the distance from TSS, among other determinants to binding that we refer throughout this thesis as *contextual environmental factors*. A *TF's occupancy* is similar to specificity, except that it is measured over a longer sequence or genomic region. The TF binding specificity can be modelled from TF binding affinities to multiple motifs (e.g. in PBM) or learned from TF binding occupancy data, especially *in vivo* (e.g. ChIP-seq).

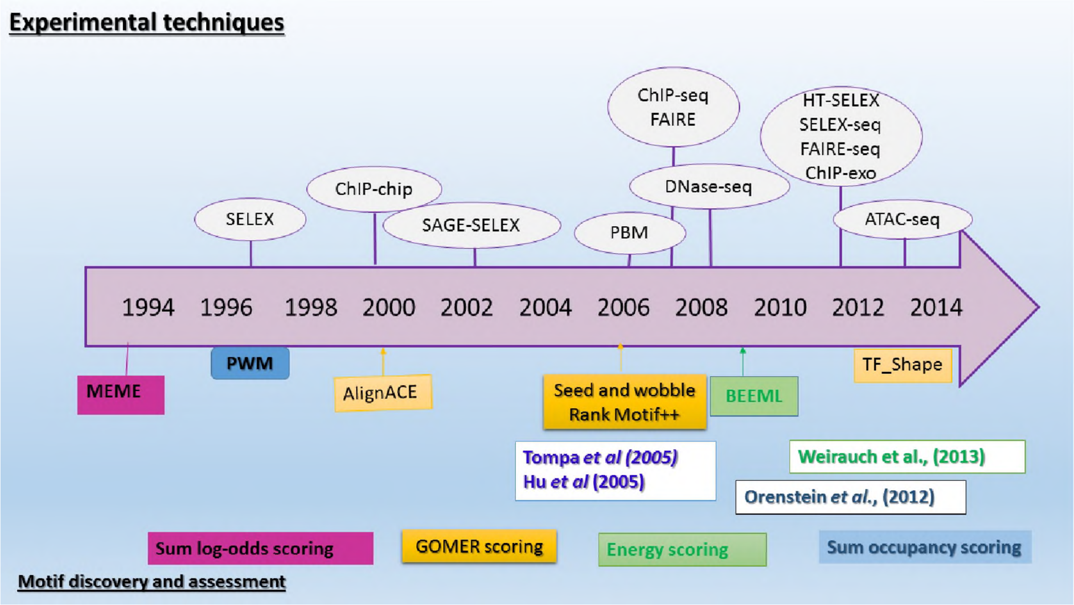


Fig. 2.4 Evolution of motif scoring functions with experimental techniques and algorithms. Tompa et al. [186] and Hu et al. [70] assessed the motifs by binding site prediction while Orenstein et al. [137] and Weirauch et al. [197] used scoring. The scoring techniques are colour coded for the motif discovery or assessment where they were used. Figure used in Kibet and Machanick [92].

Several experimental techniques have been introduced to model TF binding specificity or measure TF occupancy *in vivo*, and these have been changing over the years as our understanding as to what determines a TF's specificity increases. Initially, gene promoter sequences were used, but these led to false positives since overrepresented motifs do not always represent binding sites, and cannot model enhancer sites [20]. Therefore, several experimental techniques that measure TF occupancy *in vivo* have been introduced (for reviews see [20, 140, 181, 78] and Figure 2.4) which have be used to model TF binding specificity. However, for a comprehensive measure of TF binding affinities to all possible sequences, *in vitro* techniques have been used. Each of these come with benefits, but an ability to combine both to model TF binding specificity would be very informative; this is the focus of Chapter 6. In the sections that follow, we review some of these techniques, show their use and demonstrate some of the shortcomings of the current approaches that necessitate a combinatorial approach.

2.1.1.1 *In vitro* techniques

Experimental techniques that measure the binding affinities of TFs to random sequences outside the cell (*in vitro*) provide a comprehensive measure of the binding affinity of a TF to a given sequence. These techniques include microfluidics, surface plasma resonance, protein-binding microarrays (PBM), high-throughput SELEX techniques and many others reviewed in [181, 192, 140]. We only provide a quick introduction to PBM and HT-SELEX, which are relevant to this thesis.

2.1.1.2 Protein Binding Microarray (PBM)

A universal PBM (uPBM) technology [15] is an *in vitro-based* technique that allows a quick measure of TF binding affinity. In a uPBM, over 44,000 double-stranded DNA sequences, 60bp long (24bp linker and 36bp binding sequence), are hybridised to an array. The array will contain all possible 10-mers represented at least once, and nonpalindromic 8-mers 32 times, in an array generated using de Bruijn sequences [146] as demonstrated in Figure 2.5 A and B. In more detail, the arrays are incubated with purified TFs to saturation then extensively washed to eliminate non-specific binding. The TFs bound to the array are visualised using fluorescent-labelled antibodies against the epitope tag (Glutathione S-transferase) expressed with the TF [15]. The fluorescent intensity provides a measure of TF binding affinity to a probe.

A uPBM is a cheap and efficient method widely used to determine TF binding specificities since it is not limited by the availability of highly specific antibodies against a TF, like ChIP-seq, nor the knowledge of the reference genome. It has been used to learn TF binding models deposited in both UniPROBE [133] and JASPAR CORE 2014 databases [122]. However, the sequence length that can be accommodated by the arrays is limited to short binding sites that are less than 12bp long and can, therefore, only be used to investigate short binding sites confidently [78]. Since it is an *in vitro* technique, the learned TF binding specificity models do not include contextual environmental factors relevant to binding; consequently, they may sometimes not generalise to *in vivo* [140, 172]. The primary challenge with uPBM experiments is correctly learning motif models from the 36bp sequences in addition to the effect of technology-specific noise, a phenomenon Jiang et al. [76] referred to as sticky *k*-mers, further limiting our ability to model TF binding specificity confidently.

The genomic-context PBM (gcPBM) [52], a variation of uPBM, uses sequences extracted from the genomic contexts such as ChIP-seq and, therefore, are likely to provide a measure of the binding affinity of sites that are bound *in vivo* [87]. Although limited data for this technique are available, gcPBM is expected to provide an improved measure of TF binding specificity, especially for complex models [172, 87].

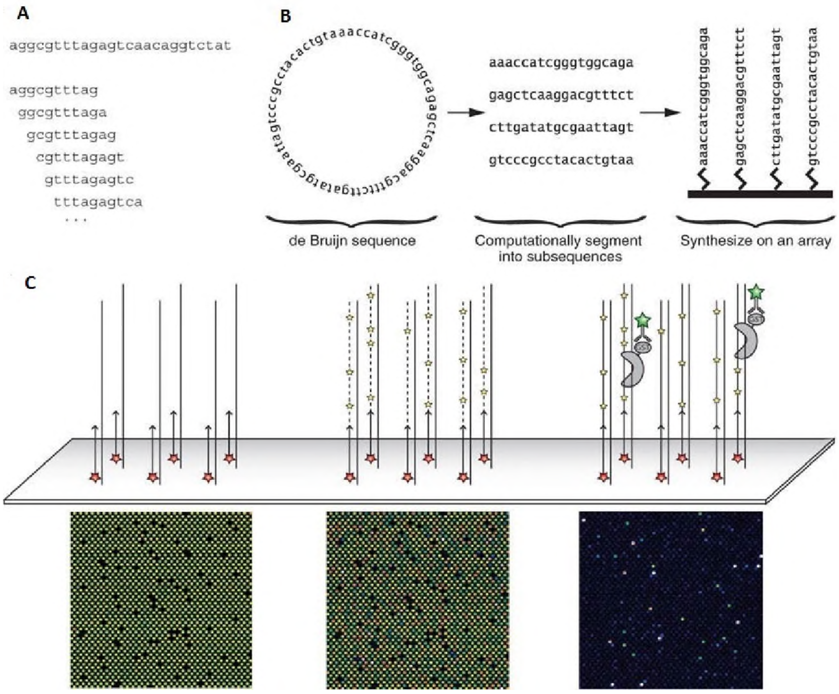


Fig. 2.5 Protein binding microarray experiment. A: A 10-mer with several overlapping 8-mers. **B:** A de Bruijn sequence is used to generate the sequences, which are synthesised and hybridised to an array. **C:** These are then incubated with TFs of interest and visualised by fluorescence. Adapted from [15].

2.1.1.3 HT-SELEX

The other widely used *in vitro* technique is the HT-SELEX [79], which measures TF binding affinity to oligonucleotides of length *n* over multiple enrichment cycles. The method employs random oligonucleotides in the first cycle and amplifies the bound ones by sequencing for use in the next cycles. Some of the motif models utilised in this study are generated using this technique. When compared with uPBM models, HT-SELEX models are better at modelling *in vivo* binding due to the use of longer oligonucleotides (10-40bp) [140, 139] compared with less than 12bp in uPBM [15].

DNA affinity purification sequencing (DAP-seq) [136] is a new disruptive experimental technique [156] that uses *in vitro*-expressed TFs to interrogate genomic DNA and predict TF binding sites. So far, this technique has only been used to investigate TFBS in Arabidopsis.

2.1.1.4 *In silico* techniques

The next two techniques, especially DNA-shape, are generally considered *in vitro*, but since they are not experimental techniques, they should be considered as *in silico* techniques. These are data generated from the genome sequences using computational techniques.

2.1.1.5 DNA-shape

Previously, the focus when modelling TF binding specificity has been on TF-DNA preference to primary sequence; however, the reality is that the TF recognises DNA in its 3D conformation, formed by the binding site and flanking sequences [52]. DNA shape analysis is concerned with the contributions to TF's binding specificity beyond the primary sequence [33]. A PWM (reviewed in Section 2.1.3.1), can only capture sequence specific contribution to TF binding and not nucleotide interdependencies that contribute to DNA shape [159]). Slattery et al. [172] explain the absence of a simple regulatory code for predicting TF binding sites: the failure to use, as well as the difficulty of using, both base and shape readout information in the current techniques modelling TF binding specificity.

The DNA-shape properties are influenced by sequence DNA bending and unwinding, both of which are predicted from the sequence. The static and dynamic properties of DNA structure [159] control the DNA shape, mainly through the enhanced negative electrostatic potential contributed by arginine and histidine residues in the minor groove [172]. To predict the DNA shape of a nucleotide in a sequence, the DNAshape tool [221] is used, which queries a table of pentamers. They generate this table by Monte Carlo simulations on a training set of 2121 different DNA fragments ranging from 12 to 27bp providing about 44 times coverage of all the 512 unique pentamers [221]. The main shape features predicted by DNAshape are the minor groove width (MGW), Roll, propeller twist (ProT) and helix twist (HelT). To facilitate the use of the DNA shape features, in addition to DNAshape tool, other tools in the family of shape tools by the same group are available: a database of binding sites, TFBSshape [204] and a database and browser of multi-organisms genome-wide shape predictions, GBshape [33]. We download the DNA shape features used in Chapter 6 from GBshape.

2.1.1.6 Evolutionary conservation

Functional sequences are expected and have been found to be conserved across genomes [170]; this includes TF binding sites [134, 79], as demonstrated by Schmidt et al. [166] for Cebpa and

Hnf4a TFs using five vertebrate species. Siepel et al. [170] developed a program, phastCons, to predict conserved elements in aligned genomes. PhastCons employs a two-state phylogenetic Hidden Markov Model (phylo-HMM) – conserved and non-conserved – to predict the probability that a site is conserved taking adjacent sites into consideration, or specifically to be under the conserved phylo-HMM model compared with the non-conserved model. Another algorithm, phylogenetic p-values (phyloP) [148], predicts evolutionary conservation and acceleration by measuring non-neutral rates of substitutions in a phylogeny. The conservation scores have been widely applied to model TF binding specificity [58, 132, 69, 217, 147, 193]. We use these data to model TF occupancy in Chapter 6.

2.1.1.7 *In vivo* techniques

In vitro techniques are useful if the interest is to measure TF binding affinities. However, what is required, in the end, is an ability to predict TF binding occupancy *in vivo* in a given context. Therefore, *in vivo* techniques are necessary at least as a reality check. Several experimental techniques are available [20, 172, 78], but the ChIP-seq technique has earned its place as the *de facto* standard for determining TF occupancy *in vivo* [208]. However, this can only be carried out per TF, which is expensive and requires the availability of antibodies of interest. As a result, techniques like DNase-seq, which provide genome information of open chromatin sites, were introduced.

2.1.1.8 ChIP-based techniques

Chromatin immuno-precipitation (ChIP) [154], used to study protein-DNA interactions, is a well-established technique based on enrichment of DNA associated with the protein of interest. ChIP first fixes the cells of interest in chemical cross-linkers like formaldehyde which covalently bind proteins within the cell to each other and their target DNA. Once cross-linked the chromatin is extracted and sheared by sonication to small DNA fragments of 0.2 to 2kb. Protein-DNA complexed fragments are separated by immunoprecipitation using specific antibodies against the cross-linked TF of interest. When antibody of interest is not available, proteins can be tagged using Green fluorescent protein, V5 or Glutathione S-transferase. The next step is to reverse the cross-links, and the purified DNA represents fragments enriched for TF binding sites [24, 149]. The fragments can be analysed using microarray technology (ChIP-chip) or massively parallel sequencing (ChIP-seq [77]), which provides a higher resolution, coverage and sensitivity [25]. ChIP-exo [155] is another variation that uses exonuclease to trim bound sites more precisely to achieve single bp accuracy. Several reviews of the ChIP-seq techniques are available [49, 25, 172]. In this study, we use the ChIP-seq peak information to benchmark, evaluate and rank PWM motifs in Chapters 3 and 4.

2.1.1.9 DNase-seq technique

The chromatin state influences how TFs bind. Open chromatin sites are accessible for binding, expressing genes that are regulated by these transcription factors. Techniques like DNase I footprinting and FAIRE-seq are used to map the open chromatin sites. In classical DNase I footprinting, DNA is incubated with a TF and degraded using DNase I – an enzyme that cleaves phosphodiester linkage adjacent to pyrimidine in DNA. Chromatin regions exposed by nucleosome depletion are potential regulatory sites and are sensitive to DNase I digestion. The fragments can then be analysed using gel electrophoresis to generate a footprint of the TF that was bound to it. The DNA is ^{32}P-end-labeled for visualisation with autoradiography [25]. The distance to the edges of the DNase I footprint from the end label [25] indicates the position of the TF binding site on the DNA fragment. As sequencing technologies developed and cost of sequencing reduced, it became feasible for the DNA fragments from DNaseI digestion of whole cells to be sequenced and mapped to the reference genome for the identification of hypersensitive sites. This technique, referred to as DNase-seq, developed by Boyle et al. [22], is summarised in Figure 2.6.

In most cases, DNase-seq cannot be employed to infer the function of the detected accessible chromatin sites or identify the TF bound to them. Therefore, they are coupled with ChIP-seq or any other TF-specific data to infer with higher confidence regions bound by TFs [174, 217, 113]. More details are in Chapter 6.

2.1.2 Modelling TF binding

The data derived from the above experimental techniques are used on two levels: modelling TF binding specificity or for predicting TFBS. The later can be done in conjunction with available models. These can also be considered as either *de novo* motif discovery or TFBS search problem, respectively. In *de novo* motif discovery, the motif information can be decomposed into a PWM, *k*-mers or one of the other complex models. These models, which capture TF binding specificity, can then be used at a later stage in TFBS search problem. In motif discovery, a variety of algorithms has been developed as new experimental data are generated, which incorporate a variety of information required to model TF binding specificity. For uPBM

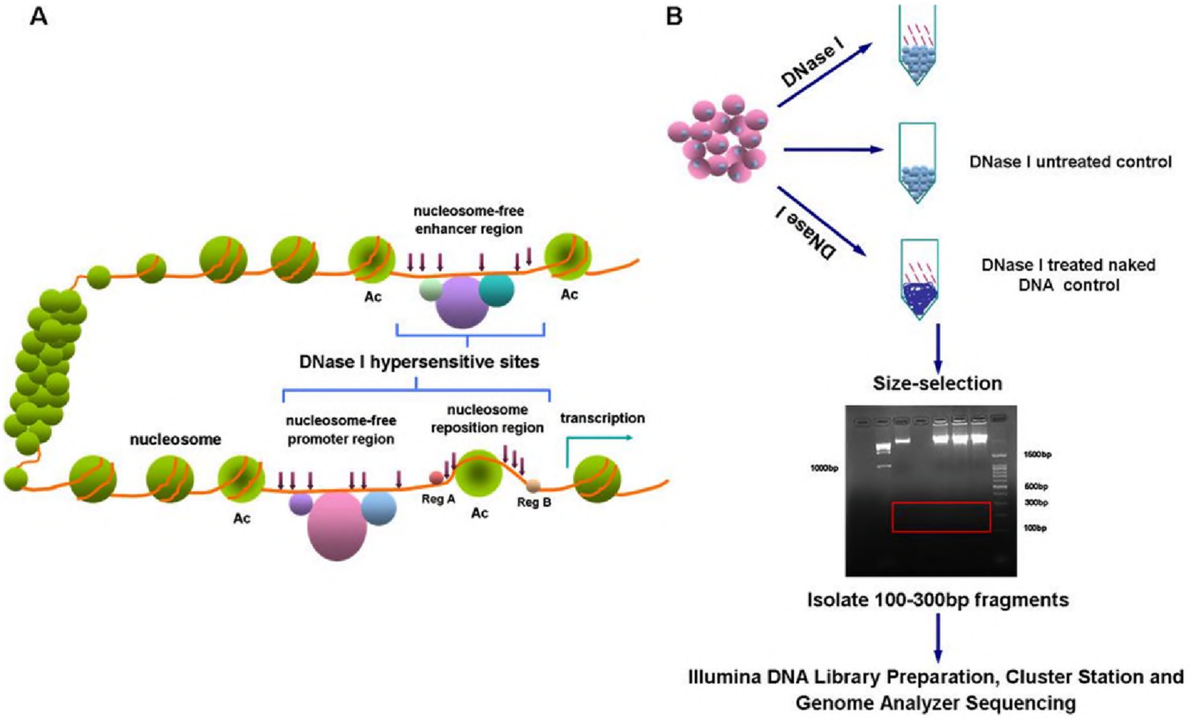


Fig. 2.6 DNase-seq. A: The nucleosome-free regions, hypersensitive to DNase I digestion, are accessible for binding. **B:** The protocol used to generate the DNase-seq data. Figure adapted from [193].

techniques, several algorithms exist [138, 29, 213], reviewed in [197, 137], each tackling different aspect of the problem. On the other hand, ChIP-seq data, considered the *de facto* technique for TF binding occupancy *in vivo*, also has several algorithms specific to it [115, 98]. These have been reviewed [207] and evaluated [187] to provide a guide to the users as to which technique to use. The majority of these reviews demonstrate the lack of a single algorithm that significantly outperforms the rest [186], and some recommend the use of multiple algorithms in motif discovery [62, 70], and to choose more than one model. The problem, however, is how to select a motif from multiple tools, more so given that most of the techniques generate models in different PWM formats. This problem points to the need for ensemble-based motif discovery algorithms [189, 115, 112], which make use of multiple algorithms, after which the models generated are unified and ranked at the end. Lihu et al. [111] reviewed and tested these ensemble algorithms and recommend the use of more than one ensemble approach. One such approach, GimmeMotifs [189], specific for ChIP-seq data, splits the data into training and test sets, then uses nine motif discovery algorithms. The motifs are evaluated and ranked using the held-out test data. We use this technique when evaluating our data-independent motif ranking algorithm introduced in Chapter 4.

The presence of multiple motif discovery algorithms and the need for ensemble methods underpins the difficulty in modelling TF binding specificity and the need for a complex model. The statistical and consensus-based approaches to motif discovery only model low-level representation in the binding site [143] and overlook the complex environmental contribution to TF binding [42], hence the need for sophisticated models [180, 140, 172]. Also, except a few such as Dimont [55], a majority of the techniques cannot be used across or in a combination of experiential techniques to model binding specificity. Several extensions have been introduced, including those that use additional data as a probability prior and those that apply machine learning methods; these are reviewed and tested in [219] to demonstrate a significant benefit to using both statistical and machine-learning approaches. One striking extension is DeepBind [3], which uses deep neural networks to learn binding models visualised as weighted ensembles of PWMs; it can train on uPBM, HT-SELEX and ChIP-seq data outperforming state-of-the-art algorithms developed specifically for these data. However, although Alipanahi et al. [3] make some effort to describe the low-level representation, it remains a black box [143].

We get back to the TFBS search or recognition of binding sites problem (Section 2.1.4) after introducing some of the techniques used to represent TF binding specificity.

2.1.3 Representing TF binding specificity

Representation of TFBS remains an enduring challenge in regulatory research [120]. Initially, consensus sequences [141], which use the most abundant nucleotide at each position to illustrate the binding preference of each TF were used. Later, the use of ambiguity codes to represent the type of bases (see Table 2.1) improved on consensus, but this remained too simplistic to capture the complex nature of TFBS [179]. Therefore, the position weight matrix (PWM) was introduced.

2.1.3.1 PWM

A PWM, reviewed in [180], the most widely used mathematical model for TF binding specificity, was introduced by Stormo and Schneider [178], initially for representing RNA sites as a weight matrix. Stormo defines a PWM as the weight of base b (A,C,G or T) at position i (1 to L) in a L-long binding site given as $W(b,i)$. The score for a given sequence of the same length as the PWM is the weights of each location. Using the sum is the basic approach; other scoring functions and techniques have been used to score sequences. These are reviewed and tested in Chapter 3. PWMs are popular due to their simplicity and ease of visualisation with sequence

Table 2.1 Summary of IUPAC DNA codes

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

logos, first introduced for consensus models. There are many variations to the PWM format including Transfac, MEME [13], PFM [180], JASPAR [122], UniPROBE [133] and many others; in fact, the majority of the tools and experimental techniques developed use or introduce a variation of the available formats. Throughout this thesis, we use PWM motifs represented in MEME format due to familiarity to us. Those we use in this thesis that are in other various formats, we convert to MEME format.

Additionally, a PWM's information content provides a crude measure of a model's specificity; however, there remains disagreement in the field as to what constitutes an accurate model. For example, Orenstein [137] said that high IC motifs are better at modelling *in vivo* binding while Weirauch et al. [197] argues for low IC motifs. In this study, however, we show that IC in itself is not a good measure of motif quality and that this depends on the specificity and variability of a TF's binding site. See Chapter 3 for details.

The PWM model assumes that each position in the TF binding site contributes to the binding affinity of any site independent of other nucleotides. Matrix scores provide the relative preference by TF for the particular base at each position. Therefore, a PWM model only provides an approximate measure of a TF's binding affinity to a given site [181], with shortcomings. These include the inability to model TF interactions and nucleotide dependencies and failure to model a TF with more than one DNA-binding interface or variable-width gaps. Despite its simplicity and shortcomings, a PWM has proven to be a successful model of TF binding specificity [214, 197]. Therefore, the binding specificities of many TFs have been modelled as PWMs stored in motif databases like JASPAR [121], Transfac [123], UniPROBE [133], HOCOMOCO [98], CIS-BP [198] and many others listed in Table 4.1 (Page 72). The motif models stored in these databases are generated from several experimental techniques and motif finding algorithms; therefore redundancy exists. We address this issue in Chapter 3, where we demonstrate the need for a systematic evaluation of PWM models and develop a resource to do this in Chapter 4.

Because of the PWM model's shortcomings, some improvements have been introduced. The main weakness of PWMs is the failure to model nucleotide interdependencies; therefore, Zhou and Liu [218] considered pairwise dependencies in learning the motif model. Additionally, the presence of interacting TFs and their sites influences the binding of TFs to a site. Therefore Wang et al. [194] introduced an interaction-dependent model that considers the PWM of the TF of interest as well as PWMs of other potentially interacting TFs using linear regression. The PWM model was further improved by the inclusion of the independent binding energies at each position (PWM) or considering nucleotide dependencies (dinucleotide) using non-linear regression [215].

Some variations beyond PWMs exist. Weight array matrix (WAM) model [210] incorporates first-order dependencies while the Bayesian trees consider the first-order dependency at any one other position. Finally, the sparse local inhomogeneous mixture model (Slim) [87] considers higher-order dependencies at multiple sites at the same time but can limit to a given number of preceding sites (LSlim) – see [87, 43] for more details.

In addition to nucleotide interdependencies, the length of a TFBS is generally not definitely known prior to modelling. To address this, transcription factor flexible models (TFFM), a technique based on hidden Markov models, was introduced for representing TFBS to model nucleotide interdependencies and flexible length of TFBS [120]. Pairwise interaction model (PIM) [165], which explains pairwise nucleotide interactions based on the principle of maximum entropy, was introduced soon afterwards. Both are claimed by the authors to be superior to previous techniques.

As more reviews and tools become available, that make use of these alternatives to PWMs, we may see wider use and move away from the traditional PWM. For now, the search for a generalised model to represent TF binding specificity continues.

2.1.3.2 *k*-mer models

The failure of PWM models to represent positional nucleotide interdependencies has led to the development of other models that do not use a PWM. One such approach, *k*-mer based models [7], has gained popularity for representing DNA-protein interaction, in particular for the uPBM technique. Here, each *k*-mer (a subsequence of length k) is assigned a score that represents the affinity of a TF to that *k*-mer [82]. Markov-based approaches [168] have also been used to model TF binding *in vitro* and *in vivo* where they are used to score every possible sequence of length k to calculate enrichment scores using the seed-and-wobble algorithm, thus overcoming PWM shortcomings. However, *k*-mer models are unwieldy, difficult to visualise and have a propensity to over-fit technique bias [140]. A comparison of performance using *in vitro* uPBM data and *in vivo* ChIP-seq data revealed that *k*-mer models significantly outperformed PWM models *in vitro* but not *in vivo* (p=0.001 and 0.718, respectively; Wilcoxon rank-sum paired test). Like with PWMs, how sequences are scored using *k*-mer models has never been systematically investigated. Therefore, in Chapter 6 we investigate *k*-mer scoring and the functions that have been used.

2.1.3.3 Visualising TFBS

The continued extensive usage of PWM models is due to simplicity and the use of an intuitive visualisation with logos. Sequence logos were introduced by Schneider and Stephens [167] to display the consensus of the sequences by displaying the predominance and frequency of each residue at each site, and information content at each site as shown in Figure 2.7C. The height of the logo at each location represents the information content and conservation of that particular base, while each base at the top of the stack constitutes the consensus sequence. As intuitive as sequence logos are, they do not provide information on interdependencies between the sites. Some advanced models of TF binding specificity introduce visualisation algorithms like dependency logos [87] (Figure 2.7B) for Slim models or TFFM logos (Figure 2.7A) [120] for TFFM. The value of these extensions needs a demonstration of ease of use and added benefits. For *k*-mer models, visualisation is a major problem, but there have been some ideas [140] to use specificity landscapes [26], but none are implemented so far.

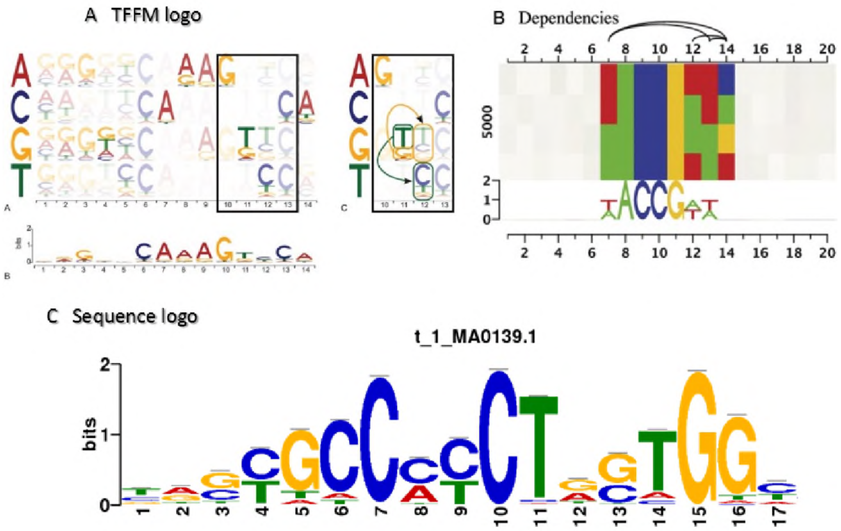


Fig. 2.7 Visualising transcription factor binding specificity. Some forms of visualising TF binding specificity models. **A:** TFFM models, TFFM logos [120]; **B:** Slim models, Dependency logos modified from [87]. **C:** PWM sequence logo predicted from ChIP-seq data using MEME [13].

2.1.3.4 Comparison of models

Weirauch et al. [197] compared the performance of PWM, *k*-mer, dinucleotide and Markov-based algorithms using *in vitro* (uPBM data) and *in vivo* (ChIP-seq) data sets on nine TFs. They found that *k*-mer based models performed best overall in predicting *in vitro* binding. However, PWM models trained on *in vitro* data performed better than *k*-mer or dinucleotide models derived from *in vitro* data and performed comparably to PWMs derived from *in vivo* data in predicting binding in ChIP-seq data. PWM-based models that train their models on an energy-based framework like FeatureREDUCE_PWM [157] and BEEML-PBM [215] performed the best among the PWM based algorithms. In summary, they found that well-trained and well-implemented PWM performs as effectively as more complicated models [197]. These observations were corroborated by a recent review [140]. Although an independent comparison is still lacking for the advanced models – TFFM [120], Slim models [87] and PIM [165] – author comparisons demonstrated their better performance.

2.1.4 Use of TF binding specificity models

As reviewed before, there are two forms of the TFBS problem: *de novo* motif discovery and TF binding site search also called prediction or recognition. *De novo* motif discovery algorithms learn binding models from the experimental techniques while TFBS search uses the model and additional information to scan or identify binding sites in the genome or a set of sequences. The TF binding specificity models can also be used to compute the probability of a TF's occupancy in a sequence or genomic site; we discuss these two in turn below.

2.1.4.1 Predicting transcription factor binding sites

The learned TF binding specificity model can be used to score sequences to identify the binding sites of a TF and calculate the binding affinity to each possible site, and those above a given threshold are considered to be the putative binding sites. However, since a majority of the currently available models of TF binding specificity cannot capture the complex nature of the TFBS, they generate high false positives when used for motif scanning.

For example, when using PWM models to predict a binding site, one can employ one of the many scanning algorithms reviewed in Jayaram et al. [75], where they found FIMO [54] to be the best among the five tools tested. Another comparison [37], which surprisingly did not include FIMO, compared commercial (Biobase [123], MatBase) and public (Match [88] and MatInspector [27]) motif scanners and found public ones to be better. Especially for genome-wide scanning, several hits above a given threshold are found, but most are false positives, which can be reduced by techniques to optimise the threshold [20, 21, 37], but the how is still debated. Another solution is to use additional information to localise the search and increase prediction confidence. Chromatin accessibility data is widely used to narrow down the hits to those within open chromatin sites. Recently, Dror et al. [42] reviewed binding site environments' factors that could help "find the needle (binding site) in the haystack": flanking nucleotides, distal conserved site, and DNA shape. This information is required to learn binding specificity models that can accurately predict TF occupancy in a given genomic site. In addition to this, knowledge of evolutionary conservation of TF binding sites [134, 79] and proximity to TSS are used [58].

Ideally, for a given TFBS search task, the majority of the data described could be used to narrow the search, but simple scanners cannot incorporate such information. Therefore, to fully take advantage of the diverse additional data sets and reduce false positives, machine learning models have been trained, where the scanning algorithms are used to predict putative binding sites, and then machine learning models are trained to discriminate bound from the unbound sites [68]. Others have directly trained machine learning models to predict binding sites without using a PWM. Bauer et al. [14] used sequence-specific chemical and structural

DNA properties in a linear SVM classifier to discriminate true and false TF binding sites in *E-coli*; later improved by using additional physicochemical features [117].

2.1.4.2 Measuring TF binding occupancy

To calculate the occupancy of a TF *in silico*, TF *binding specificity* models are used to compute the *binding affinity* at each possible site in the sequence. Then the probability of occupancy is determined using a variety of ways including multiplying, summing, or finding the maximum affinity in the sequence depending on the scoring function used. Binding affinity is not the only determinant of occupancy; some scoring functions have also modelled the TF concentration levels [32, 215], but this is usually unknown, except in the uPBM technique. *In vivo*, the TF occupancy is also influenced by chromatin accessibility of the site, presence of cooperating or interacting TF sites and many other factors mentioned before. Therefore, computing the probability of occupancy using the TF binding specificity model does not provide the full picture. Indeed, Cheng et al. [32] demonstrated this using ChIP-seq peak score to represent *in vivo* occupancy, by testing the ability to reproduce these ranks using PWM models, DNase accessibility data, competition and cooperative binding with other TFs [84, 32].

The influence of contextual environmental factors to binding occupancy, however, cannot be captured using simple probabilistic modelling or scoring. The role of these factors can be determined using machine and statistical learning, where a combination of features are used to predict TF occupancy [58, 99]. Indeed, machine learning modelling has been pivotal to investigating the role of TF binding site shape to binding affinity and ultimately the occupancy using uPBM [220, 1] and ChIP-seq [121] data. In Chapter 6, we investigate the use of these additional features using *k*-mer scoring models to modelling TF occupancy. However, to confidently do that, we first review and test the scoring functions used to compute binding occupancy, among other factors in Chapter 3 and develop a motif assessment and ranking suite in Chapter 4.

In the following section, we introduce some machine learning concepts and algorithms that are relevant to this thesis.

2.2 Introduction to machine learning

Our understanding as to what contributes to a TF's occupancy *in vivo* using TF binding specificity remains incomplete and under active investigation. As motivated in Section 2.1.4.2,

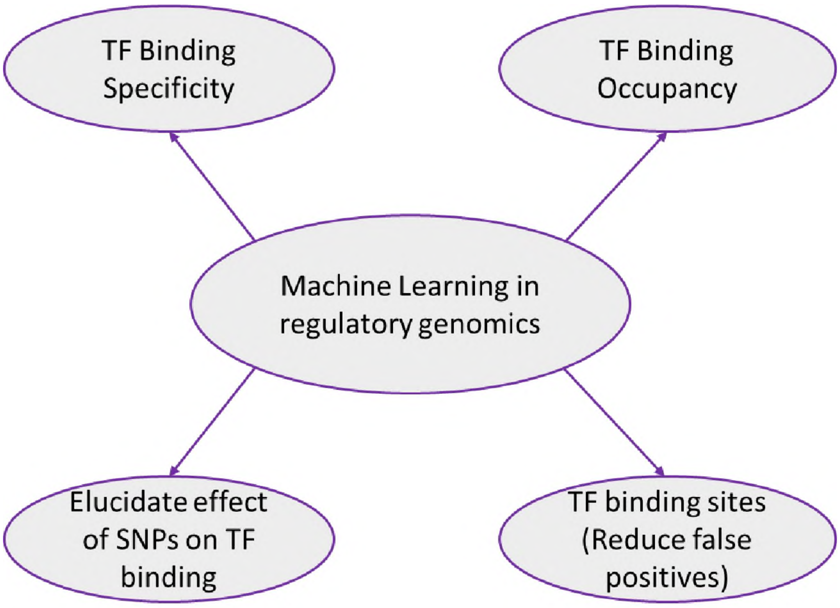

TF Binding Specificity
TF Binding Occupancy
Machine Learning in regulatory genomics
Elucidate effect of SNPs on TF binding
TF binding sites (Reduce false positives)

Fig. 2.8 Some applications of machine learning techniques in regulatory genomics. Machine learning has been applied in the various branches of regulatory genomics, especially with regards to transcription factor (TF) binding.

machine learning and advanced probabilistic modelling are used to measure the contribution of contextual environmental factors to TF occupancy. Furthermore, there is a need for complex binding specificity models that can capture the input of contextual information, and machine learning is a promising approach to that problem. Finally, even when using advanced TF specificity models to predict binding sites, there will be false positives due to the complicated nature of TF binding *in vivo*. Therefore, to reduce false positives, machine learning models play a pivotal role as well. In summary, some of the areas in gene regulation in which machine learning is expected to make contributions are shown in Figure 2.8. The focus of this study is on elucidating TF binding specificity and occupancy.

2.2.1 So, what is machine learning?

Tom Mitchell defines machine learning as “the study of computer algorithms that improve automatically through experience” [128]. Machine learning modelling has found widespread application in the genomic era to make sense of the massive amount of data generated [110, 206]. We use the application of machine learning in gene regulation to demonstrate concepts of interest in this thesis – see Box 2.2 for definitions. A machine learning model could be trained on known (labelled) transcription factor binding sites and used to predict unknown locations in the genome, *supervised learning,* or used to identify functional elements in the genome given a broad range of integrated data [67], *unsupervised learning*. However, since a majority of TF binding sites are unknown, a machine learning model could be trained on data containing a mixture of known and unknown sites, where the model iteratively learns and self-labels the data to use in further iterations, *semi-supervised learning* [44].

To demonstrate some concepts, we use uPBM data. A model can be trained using probe intensity for a given TF in one array and used to predict intensity in a second array with a different spot design, and hence different probe sequences. Two options are to predict correct ranks, *regression models* [2], or to discriminate probes bound by the TF from those that are not, *classification models*. To solve either the classification or regression problem given the probe intensity, the machine model may take as input k-mer frequency or the PWM score in the probe sequence: *features*. In the classification model, for example, a model can be trained to just separate and label probes as either bound vs. unbound, *discriminative*, or train a model that can be used to predict the probability that a TF binds to the probe, *generative model* – as an example, PWM can be learned generatively by maximum likelihood [110], but may also be learned discriminatively by maximum conditional probability. The XGBoost gradient boosting algorithm used in this study is a discriminative model, together with Random Forests, Linear regression, and SVM. These have only become viable options as more training data became available [110]. Although discriminative models are less interpretable, they bring the benefit of higher accuracy [219].

Box 2.2: Key machine learning concepts

Supervised learning: Model is trained with labelled data and used to predict unlabelled data.

Unsupervised learning: Model is used to cluster data into categories without pre-specified labels.

Semi-supervised learning: A model is trained using a combination of labelled and unlabelled data.

Regression: The data label in the model is a real number (binding intensity, expression level).

Classification: The data label in the model is a discrete variable (true vs. false, bound vs. unbound).

***Features**:* A set of properties extracted from the input data that describe the labels.

***Discriminative models**:* The emphasis is on describing the labels given the features – conditional probability, $P(y|x)$.

***Generative models**:* A model that describes the distribution of the features about the classes – joint probability, $P(y,x)$.

2.2.2 Machine learning models

As already reviewed, a variety of machine learning models has been applied to one gene regulation problem or the other. For brevity, we only focus on the most commonly used models: Support Vector Machines (SVMs), Gradient Boosting Machines (GBM), Artificial Neural Networks (ANN), Deep Learning and Random Forest (RF).

Support Vector Machines (SVMs) are arguably the most used in modelling TF binding specificity [104, 103, 2, 8, 68, 58]. An SVM, using a TF binding site classification problem as an example, seeks to find an hyperplane boundary that separates the bound from the unbound sites. An SVM seeks the decision boundary that maximises the separation from the *support vectors* (the sites located nearest to the boundary – see stars with a red outline in Figure 2.9B). Since data points are not always linearly separable, we use a soft margin, which allows some level of misclassification. However, the reality is, especially for TF binding sites classification, the data sites are never linearly separable; hence non-linear kernels are used. Indeed, for the TF binding specificity modelling, several specialised kernels have been used [51, 2].

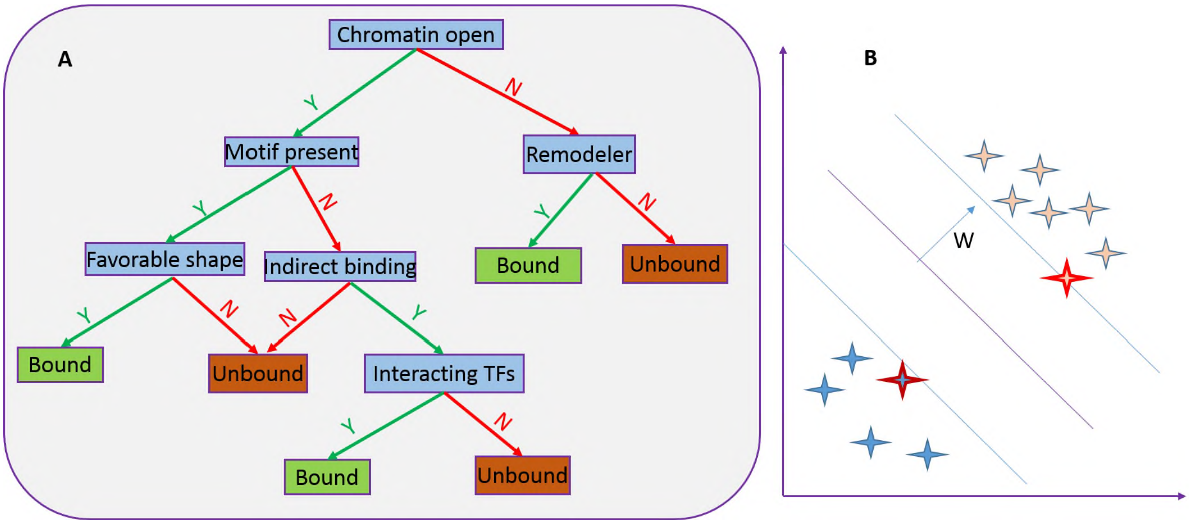


Fig. 2.9 Decision trees and support vector machines. **A:** Simplified decision tree to determine whether a site is bound by a transcription factor (TF) or not. The first decision is whether the site is open for binding, has a TF binding site and if it forms a favourable shape for binding. Other options for binding: binds indirectly or assists in chromatin remodelling. **B:** A graphical representation of a support vector machine classifier. The stars with a red outline, nearest to the decision boundary, are the *support vectors.*

Boosting machines, which belong to the class of ensemble algorithms, use weak learners to boost prediction accuracy of the base learner. The main benefit of the boosting algorithms is the ability to combine evidence without over-fitting the data [100]. This includes Gradient Boosting Machine (GBM) and eXtreme Gradient Boosting (XGBoost [28]). The most common weak learners used are decision trees. For example, to determine whether a site is bound or not, the decision may be made as shown in Figure 2.9A. See [124] for an introduction to boosted algorithms and our application of XGBoost in Chapter 6.

Random Forest (RF) [23], like boosted machines, belongs to the family of ensemble models. The RF trains a combination of decision trees and aggregates the output by majority voting in a classification problem. This has been applied to model TF occupancy *in vivo* [160, 188].

Artificial Neural Networks (ANN) and their derivations have received lots of attention in regulatory genomics recently. ANNs, inspired by the neurons in the brain, are made up of interconnected "neurons" working in concert to solve a given problem. The ANN is extended by adding hidden layers, Multi-Layered Neural Networks (MLNN), or to Convolution Neural Networks (CNN) inspired by the cat's visual cortex which was found to have simple and complex neurons. CNN adds an initial layer of filtering and transformation (convolution), which may be followed by sub-sampling and finally a MLNN – see Zeng et al. [209] for an introduction and application in protein-DNA binding. As the depth increases, we get deep neural networks. Alipanahi et al. [3] recently employed a convolution deep neural network to regulatory genomics, where the first stage of convolution involved motif extraction as k-mers which are used to build a PWM. These are then fed to the deep neural network in different forms to learn higher-level structure and motif interactions responsible for TF binding specificity. The learning in the deep networks is used to determine motif importance to update the thresholds and network weights in the convolution stage by back-propagation. DeepSEA [220] used a similar implementation to detect effects of non-coding variants. Deep learning is gaining traction in regulatory genomics, as evidenced by the number of reviews [143, 5] and applications [3, 220, 89]. The main drawback of neural nets is that they are black boxes; however, they have been combined with generative models based on Bayesian statistics to increase their interpretability [143] – deep belief networks (DBN) [66].

The use of advanced machine learning techniques in regulatory genomics underpins the need for complex models to capture the multi-dimensional interaction that influences the binding of a TF. However, the discriminative nature of most of these techniques can hinder in-depth understanding of the regulatory process due to lack of interpretability. Nevertheless, the quick progress and innovations aimed at opening the black boxes [143] and techniques like belief networks will significantly improve our understanding, driving advances in the field. Before then, our focus is on feature importance analysis to get a glimpse of how they contribute to TF binding affinity, specificity or occupancy. We do this extensively in Chapter 6, using XGBoost, to determine how the various features used contribute to model accuracy.

2.3 Data and algorithms utilised in this thesis

Transcription factor binding is complex, and a variety of data sets – of types reviewed in Section 2.1.1 – have been generated to understand various determinants to TF binding, as shown in the simplified decision tree in Figure 2.9A. Exploring all paths requires a range of data sets for us to understand or predict TF occupancy *in vivo*.

In Figure 2.10, the tools used to process the above data are represented in the figure with **brown** rectangles. MARS is shown in a **brown** oval and encapsulates a number of additional tools. The data formats are in **green** and the data types in **blue**. We provide a quick description of these tools including some encapsulated in MARS below:

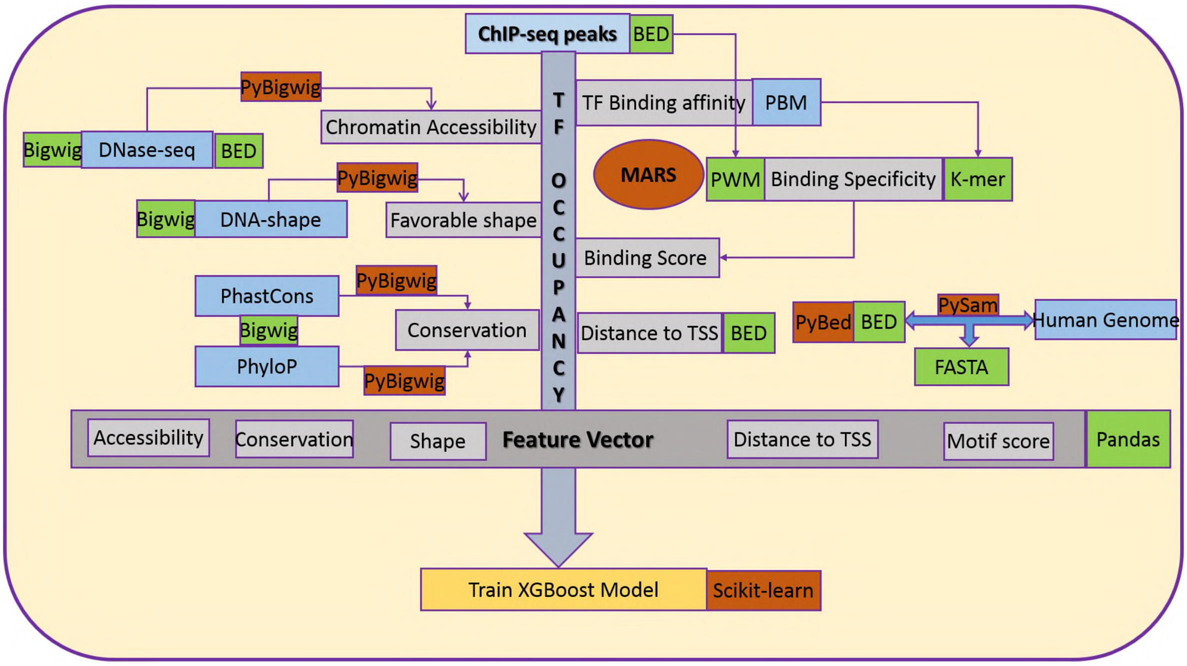


Fig. 2.10 An overview of the data (blue), formats (green), tools (brown), and algorithms used in the thesis. Centred around understanding TF occupancy, the main data used is ChIP-seq, which may be used to obtain TF binding specificity models. In grey background are features that describe binding occupancy *in silico*.

- **Pybedtools** [38] is a Python implementation of the BEDTools [152]. These are mainly used for processing BED files. The BED format stores genomic coordinates (chromosome name, start and end coordinate) and additional information, which could include orientation, enrichment and intensity, depending on the purpose.

- **PyBigWig tools** [161] are used to extract information stored in bigWig files [90] – data format used to store continuous data indexed in a binary compressed format for quick access.

- **Pysam** [64] is a Python wrapper to the SAMtools [109] used to obtain FASTA sequences in the genome on the fly to reduce storage space, especially when processing big BED files. Pysam also allows parallel processing of the BED file when coupled with the Pandas DataFrame.

- **GimmeMotifs** [189] is an ensemble motif discovery algorithm employed in our study.

- **MEME Suite tools** [115] hosts a suite of tools for motif discovery, enrichment analysis, comparison, scanning and ontology analysis. We make use of a large selection of these tools in our study.

- **Scikit-learn** [144] is a machine learning framework for Python, used for feature engineering, parameter optimisation, and Cross-validation.

- **Pandas** [125] is a data analysis and visualisation tool we make use of for a variety of data processing. Pandas DataFrame (DF) is the main format we use in our data analysis. Pandas DF is a two-dimensional data structure with indexed rows and columns; can store heterogeneous data and allows arithmetic calculations on both rows and columns. We use Pandas DF to summarise the data and to directly plot figures at the exploratory stage. We use Pandas DF for direct plotting, with Seaborn [195] for better plots or with Matplotlib [71] for advanced plotting features.

- **Jupyter Notebooks** [95] are used for easy reproducibility of the research carried out. We provide a ready to use Jupyter notebooks that demonstrate step by step the analysis done in each chapter using the IPython framework [145].

2.4 Summary

This literature review chapter has introduced gene regulation and machine learning and put the research carried out in this thesis into perspective. The review has highlighted the need for a combinatorial approach to modelling TF binding specificity and occupancy. It also revealed the lack of and the need for a standardised motif evaluation platform.

The next chapter presents a systematic review and analysis of available motif evaluation techniques.

Chapter 3

Comparative Analysis of Motif Assessment Approaches

"If you can't measure it, you can't improve it."

–Peter Drucker

Transcription factor (TF) binding site prediction remains a challenge in gene regulatory research; this is due to degeneracy and potential variability in binding sites in the genome. Dozens of algorithms designed to learn binding models (motifs) have generated many TF specificity models available in research papers with a subset making it to databases like JASPAR, UniPROBE and Transfac. The presence of many motifs versions from the various databases for a single TF and the lack of a standardised assessment technique makes it difficult for biologists to make an appropriate choice of binding model and for algorithm developers to benchmark, test and improve on their models. In this study, we review and evaluate the approaches in use, highlight differences and demonstrate the difficulty of defining a standardised motif assessment approach. We examine scoring functions, motif length, test data and the type of performance metrics used in prior studies as some of the factors that influence motif ranking. We show that the scoring functions and statistics used in motif assessment influence motif ranking in a TF-specific manner. Furthermore, we show that TF binding specificity can vary by source of genomic binding data. Finally, we demonstrate that motif information content is not in isolation a measure of motif quality but is influenced by TF binding behaviour. We conclude that there is a need for an easy-to-use tool that presents all available evidence for comparative analysis.

3.1 Background

To correctly predict or classify transcription factor binding sites, there is a need for an accurate model, be it k-mer, PWM or machine learning-based. Every algorithm designed and motif discovery run requires model evaluation. Several independent algorithm assessments and benchmarking studies have been performed [186, 197, 137]. Therefore, one would expect this to be a routine task, but it is not; it is an open question [172]. This is because of the diversity of techniques used, which are determined by the experimental data and algorithms used. This research started from the question of how to combine *in vivo* and *in vitro* data to learn TF binding specificity models in PWM (see Chapter 6). However, it quickly became apparent that no standardised protocol for PWM evaluations exists. A literature survey further revealed the disparity in the techniques used, making an informed decision difficult (Figure 3.1). This chapter presents a survey of the various techniques used in motif evaluation, compares their performance, and highlights the areas to be improved.

The initial hindrance to the quality of TF binding specificity models learned was the low resolution of experimental techniques [20]. However, next generation sequencing techniques like ChIP-seq [77] that measure TF *in vivo* occupancy (described in Chapter 2) have improved the resolution. Large scale data are available at ENCODE [45]. Also, data from universal protein binding microarrays (uPBM), which provide comprehensive TF binding affinity, is available in the UniPROBE database [133]. These data sets provide high-resolution data for motif discovery but can also be used as benchmarks to evaluate motifs already available in various databases.

The PWM (described in Section 2.1.3.1) is known to be simplistic but remains widely used due to ease of use, simplicity [179], and the sunk cost effect, especially since many tools are already using PWM in their implementation [140]. The PWM models' ability to describe TF binding may be getting saturated, but the lack of standardised and robust techniques to compare and rank available motifs continues to derail motif quality improvement. In fact, it has been highlighted that new motif discovery algorithms improve in speed and specificity but not substantially in the quality of the generated motifs [187]; 'the first generation' techniques like MEME achieve similar performance to recent ones [207].

The PWM motif evaluation problem can be linked to data choice, assessment approach and statistics used, each with further subcategories (Figure 3.1). In the section that follows, we review and categorise some techniques that have been used in motif evaluation, with a specific focus on the motif scoring approaches.

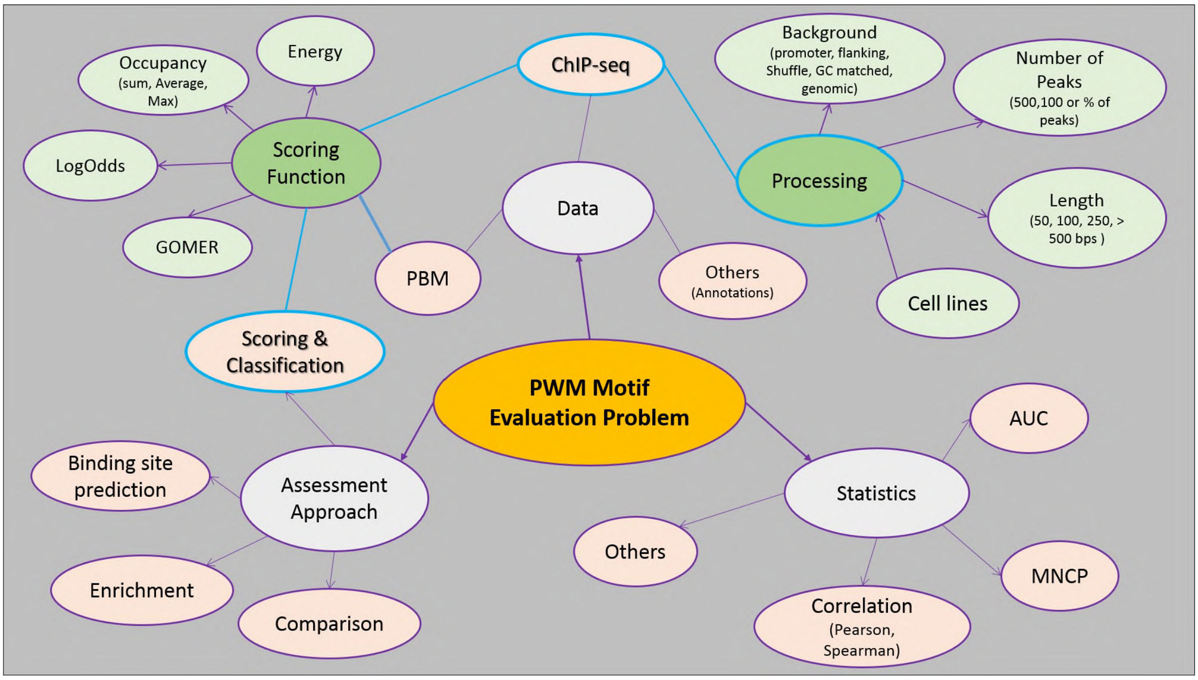

Occupancy (sum, Average, Max)
Energy
ChIP-seq
Background (promoter, flanking, Shuffle, GC matched, genomic)
Number of Peaks (500,100 or % of peaks)
LogOdds
Scoring Function
Processing
Data
Length (50, 100, 250, > 500 bps)
GOMER
PBM
Others (Annotations)
Cell lines
Scoring & Classification
PWM Motif Evaluation Problem
AUC
Binding site prediction
Assessment Approach
Statistics
Others
MNCP
Enrichment
Comparison
Correlation (Pearson, Spearman)

Fig. 3.1 The PWM motif evaluation problem. A schematic of the various factors to be determined in a motif assessment. Each can affect the results obtained in motif ranking.

3.2 Categories of motif evaluation approaches

In this section, we categorise the evaluation approaches available, test their performance, identify bias, and provide a status report on the current approaches and what still needs improvement. From literature, we can categorise motif assessment techniques into assess-by-binding site prediction, motif comparison and by sequence scoring with either classification or enrichment. In the sections that follow, we review these categories in more detail.

3.2.1 Assess by binding site prediction

This approach tests a tools' ability to predict TF binding sites, known or inserted into the sequence. Simply put, the motif prediction algorithms are used to find binding sites in a sequence, and then the predicted sites are compared with known (annotated) sites. The algorithms are then evaluated on their ability to correctly predict known sites (true positives) while minimising incorrect predictions (false positives). Next, statistical measures of accuracy can be used, including sensitivity and specificity [186]. This is the overall idea of this approach. However, there are many variations to it, each trying to alleviate one or more of the technique's known weaknesses.

Tompa et al. [186] performed the most comprehensive assessment of motif finding algorithms based on this approach. They tested the ability of motif discovery algorithms to predict sites of inserted motifs using statistical measures for site sensitivity and correlation coefficient. This study revealed the complexity of a motif assessment problem – artificially inserted motifs do not capture the biological complexity of TF binding. To improve on this, Hu et al. [70] used RegulonDB binding data annotations to compare five motif-finding algorithms. Although these were real binding sites, the annotations were of poor quality, which translated to the tested tool's poor performance [127].

As more algorithms are developed, each with different variables and data, the approaches used above cannot scale well. To solve this, Quest and colleagues [151] developed the Motif Tool Assessment Platform (MTAP) as an automated test of motif discovery tools, complete with genome-wide benchmark data for motif assessment. Nevertheless, it has not achieved wide usage because it is computationally expensive and complicated to set up; in fact, the source code is not readily available from the links provided in the publication. Zhang et al. introduced another variation when evaluating their motif discovery algorithm, MOST+ [211], where they used a technique they called 'site level accuracy'. The ability of the predicted motif models to identify binding sites, as defined by ChIP-seq peaks, when scanned using CisFinder [169] acts as an evaluation of the motif discovery algorithms.

Several other assess by binding site prediction implementations exist [94, 164, 163]. However, TF binding sites the annotations in the human genome is far from complete [10]. Therefore, techniques that depend on the ability to correctly predict these sites to assess the quality of an algorithm provides an incomplete picture; it penalises a tool that predicts unknown but actual binding sites.

3.2.2 Assess by motif comparison

In this approach, the motif discovery algorithm tests its accuracy based on the ability to recover known motifs based on similarity of the discovered motifs with those in 'reference' databases. The metrics used to compute similarity include the sum of square deviation, Euclidean distance, Pearson correlation coefficient, and other statistics that measure divergence between two PWMs [62, 212]. An algorithm is considered accurate or working, based on an arbitrary similarity cut-off. This approach is still widely used [216, 137, 131, 55].

Thomas-Chollier et al. proposed a motif comparison approach for their RSAT algorithm where they combine multiple metrics, including Pearson's correlation, width normalised correlation, logo dot product, correlation of IC, normalised Sandelin-Wasserman, the sum of squared distances and normalised Euclidean similarity for each matrix pair [184]. They then unified these scores to ranks whereby the mean of the ranks is considered the overall score.

Assessing motifs by comparison, as currently implemented, only tests similarity to the available motifs with little information on quality and ranks motifs ranks. It assumes 'reference motifs' accuracy, with no way of assessing novel ones. Besides, 'reference motifs' definition remains largely subjective [207].

3.2.3 Assess by motif enrichment

In this approach, which is very similar to assessment by scoring, the motifs are ranked based on over-representation level in foreground compared to a given background sequences. Analysis using this includes motif enrichment used by Kheradpour and Kellis [91] to rank motifs generated from the ENCODE sequences using five motif finding algorithms. It has also been previously [193] applied with the FIMO algorithm [54] to scan sequences and compute hypergeometric enrichment. Although motif evaluation by enrichment is similar to a scoring approach (described in the next section) but with statistical measure of enrichment, it has never received widespread usage. Some motif enrichment techniques available, which can potentially be used include CentriMo [12], PscanChIP [208] AME [126], MET [18] and many others. However, to be useful in motif evaluation, an enrichment tool should be flexible enough to allow test and background sequences as input.

3.2.4 Assess by scoring and classification

Motif assessment by scoring evaluates TF binding specificity models (mostly in PWM format, although other formats have been used) as opposed to predicting binding sites. Known TF binding specificity models are used to score positive sequences, known to contain binding sites, and negative background sequences, without binding sites; the score is the TF's binding occupancy. The models are then evaluated by their ability to discriminate the two sets of sequences based on their occupancies. A slight variation is an approach called *fragment-based classification*; where models are assessed on how the two sets of sequences are classified by the predicted sites within a *de novo* motif discovery framework [43].

This approach's wide usage is driven by high-throughput sequencing and microarray techniques [137, 139, 197, 217], which have provided high-quality data for motif assessment.

Although widely used, the implementation differs in how the data is prepared (the choice of sequences to use as positive and negative, and the length of the sequences in both sets), the scoring function, and the statistic used to quantify the tools' performance. In the next section, we provide a review and categorise the scoring function and statistics used in motif evaluations.

Processing differences

In this approach, ChIP-seq data is mainly used to define sequences with known binding sites. How these sequences are identified and processed, differs greatly from one evaluation to the other, as summarised in Table 3.1.

Table 3.1 Variations on evaluation benchmark processing. STAP occupancy, calculated from a STAP model (a thermodynamic model), is a variation of sum occupancy which calculates occupancy of a TF (based on the number of bound sites), on bound sites by all other TFs. Weirauch et al. tested three different backgrounds with 100bp.

Data	Length (bp)	Background	Scoring function	Reference
ChIP-seq	250bp	300bp downstream	sum occupancy	[137]
	600bp	300bp downstream	maximum occupancy	[217]
	60bp	300bp downstream	maximum log-odds	[2]
	100bp	random genomic random promoter di-nucleotide shuffled	energy	[197]
	500bp	random non-coding	STAP occupancy	[32]

All these differences, in addition to the scoring functions and statistics used, lead to incomparable evaluation output. Users and algorithm developers, therefore, have to reinvent the wheel continually when testing their tools.

PWM Scoring functions

The main differences in motif assessments stem from the difference in the scoring function used. In the following, we describe the major scoring functions in use and review their usage.

3.2.4.1 GOMER Scoring

The GOMER (generalizable occupancy model of expression regulation) scoring framework was introduced by Granek and Clarke [53] but adapted for PBM sequence scoring [29, 9]. It computes the binding affinity $g(S,\Theta) = exp(f(S,\Theta))$ that is used to model the probability a

TF, given PWM Θ, will bind to at least one of the sub-sequences of S. This assumes that each site can be bound independently:

$$g(S,\Theta) = 1 - \prod_{t=1}^{L-k} 1 - P(S_{t:t+k}|\Theta) \tag{3.1}$$

where L is the sequence length S, and $S_{t:t+k}$ is the sub-sequence of S from position t to $t+k$ inclusive and $P(S_{t:t+k}|\Theta)$ is the probability of that subsequence binding, given Θ. See Chen et al. [29] for more details.

3.2.4.2 Occupancy score

The occupancy score calculates the occupancy of a PWM (Θ) for sub-sequence (S^i) of length k as the product of the probabilities of each base in S^i using Equation 3.2:

$$f(S^i,\Theta) = \prod_{j=1}^{k} \Theta_j[S^i_j]. \tag{3.2}$$

For a sequence S of length L, the sum of the occupancies of all sub-sequences S^i (sum occupancy) [137, 46], the maximum score (maximum occupancy), [217] or the average occupancy (average motif affinity – AMA) have been used.

Sum occupancy is defined in Equation 3.3:

$$f_{sum}(S,\Theta) = \sum_{t=0}^{L-k} \prod_{j=1}^{k} \Theta_j[S^i_{t+j}], \tag{3.3}$$

while average occupancy is defined as:

$$f_{AMA}(S,\Theta) = \frac{f_{sum}(S,\Theta)}{L-k}, \tag{3.4}$$

and finally, maximum occupancy is defined as:

$$f_{max}(S,\Theta) = max\big(f(S^{i=1},\Theta),\ldots,f(S^{i=L-k},\Theta)\big) \tag{3.5}$$

3.2.4.3 BEEML-PBM energy scoring

The energy scoring framework of binding energy estimation by maximum likelihood for protein binding microarrays (BEEML-PBM) [213] computes the logarithm of base frequencies with the idea that this is proportional to the bases' energy contributions. Next, it calculates the binding energy at each location; the lower the binding energy, the higher the binding affinity. For each sequence, the sub-sequence with the lowest binding energy represents the sequence score. It has mainly been used to score PBM data [197, 217].

The probability that sub-sequence S^i is bound is given by Equation 3.6:

$$P(S^i\ is\ bound) = \frac{1}{1+e^{E(S^i)-\mu}}, \tag{3.6}$$

where, for a sub-sequence S^i and assuming a very low TF concentration ($\mu \to -\infty$), $E(S^i)$ is given by Equation 3.7,

$$E(S^i) = \sum_{b=A}^{T}\sum_{t=1}^{L} \varepsilon(b,t)S^i(b,t), \tag{3.7}$$

for a binding site of length L, $\varepsilon(b,t)$ is the energy contribution of base b while $S^i(b,t)$ is an indicator function of site t within S^i(1 with base b, 0 otherwise). The BEEML score for a given sequence, therefore, can be given as:

$$E(S) = min\big(E(S^{i=1}),...,E(S^{i=L})\big) \tag{3.8}$$

3.2.4.4 Log-odds scoring

In log-odds scoring, used by a majority of the MEME Suite tools [11], the score for a given site is the sum of a PWM's log-odds ratios at the match site. For a sub-sequence S^i of length L scored using PWM Θ, the log-odds score is given by Equation 3.9:

$$LogOdds(S^i,\Theta,p) = \sum_{t=1}^{L}\ \sum_{b\in\{A,C,G,T\}} S^i(b,t)log\frac{\Theta_{t,b}}{P_b}. \tag{3.9}$$

where p is $P_b|b \in \{A,C,G,T\}$, which is the background probability (uniform background probability of 0.25 is used) and $S^i(b,t)$ is an indicator function of site t.

For a given sequence, the score can be derived by summing (sum log-odds scoring) individual scores of the sub-sequences:

$$LogOdds_{sum}(S,\Theta,p) = \sum_{i=0}^{L-k} LogOdds(S^i,\Theta,p) \tag{3.10}$$

or by finding the maximum score (maximum log-odds scoring) as follows:

$$LogOdds_{max}(S,\Theta,p) = max\big(LogOdds(S^{i=1},\Theta,p),...,LogOdds(S^{i=L-k},\Theta,p)\big) \tag{3.11}$$

The MEME Suite tools use the sum log-odds scoring to score sequences. Maximum log-odds scoring has been used by Zhong et al. [217], where they compared motifs represented as PWM, k-mer and SVM models [2].

3.2.5 Ranking statistics

After scoring the sequences with the PWM models, the ranks of the models can be determined using one of the following statistical measures.

3.2.5.1 Receiver operating characteristic (ROC) curve

The ROC curve is a plot of true positive rate (TPR or sensitivity, Equation 3.12) against a false positive rate (FPR or 1-specificity, Equation 3.13) at different possible thresholds. The auROC is the area under the ROC curve [59], which provides a measure of a model's ability to correctly classify two sets. The closer to one the score is, the more accurate it is, while a score at 0.5 represents a model that cannot perform better than random.

$$sensitivity = \frac{TP}{TP+TN} \tag{3.12}$$

$$1 - specificity = \frac{FP}{N} \tag{3.13}$$

where TP=True Positives, P=Positives, FP=False Positives, N=Negatives.

In motif assessment, it measures how a PWM can correctly classify test and background sequences [34]. The auROC (commonly named AUC) has been widely used, in particular with the advent of the PBM experimental approach [29, 137, 197].

3.2.5.2 Precision-recall curve

Another variation, used mostly for imbalanced data, is the precision recall curve [40], which is a plot of precision (Equation 3.14) versus recall (Equation 3.15); the area under the PRC curve is auPRC. Precision, or positive predictive value (PPV), emphasises on the relevance of the predictions, while recall emphasises on the correctness of predictions. Since we use balanced positive and negative sequences, we use auROC in our study. PRC has gained widespread use for evaluating genome-wide TFBS predictions like the DREAM challenge.

$$precision = \frac{TP}{TP+FP} \tag{3.14}$$

$$recall = \frac{TP}{TP+FN} \tag{3.15}$$

3.2.5.3 Mean normalised conditional probability (MNCP)

MNCP is a rank-based statistic, introduced by Clarke and Granek [34], that determines if a motif's occupancy in test sequences is higher than the occupancy in a random set. Each sequence set is ranked based on the occupancy, and the MNCP calculated by finding the normalised ratio mean of the two ranks sets. Given a sequence S^i, the rank in the positive set is $R_p(S^i)$ and the rank in the combined positive and negative set is $R_{pn}(S^i)$, its normalised ratio $(NCP(S^i))$ is the slope of the plot $R_p(S^i)$ against $R_{pn}(S^i)$.

$$NCP(S^i) = \frac{(N_p - R_p(S^i))/N_p}{((N_p+N_n) - R_{pn}(S^i))/(N_p+N_n)} \tag{3.16}$$

where N_p and N_n represent the number of sequences in the positive and negative set respectively. The MNCP can then be calculated from the mean the normalised ratios of all the positive sequences.

This statistic has been applied for motif assessment in GIMME motifs [189] and is said to be less affected by the presence of false positives compared with AUC since it places emphasis on true positives [112]. With the advantages mentioned earlier, we use MNCP to test how it contributes to the better prediction to encourage its use.

Pearson's and Spearman's rank correlation are still widely used to measure motif performance. Orenstein et al. [137] used Spearman's rank correlation for PBM and ChIP-seq sequences [137] while Weirauch et al. [197] used Pearson's correlation, but cautioned on the use of Spearman's correlation for PBM data citing its inability to exclude low-intensity probes. We check the usefulness of correlation statistics in motif assessment as part of this work.

3.3 Chapter aim and objectives

As reviewed, several approaches have been used in motif evaluation; however, the technique with the highest variation in implementation is assess by scoring and classification (SC). We hypothesise that differences in application and use of the SC approach influence the outcome, leading to inconsistent ranking. Therefore, this chapter aims to investigate which factors influence motif assessment analysis, and how they do. We emphasise on how SC is affected by choice and length of benchmark sequences, scoring functions, and the statistics as summarised in Figure 3.1. Specifically, we address the following objectives:

1. Identify and classify the motif assessment techniques currently being used
2. Determine how motif scoring functions influence the motif ranks and recommend the least biased combinations
3. Determine how the choice and processing of background sequence influences motif ranking
4. Check how useful the correlation statistics are in motif evaluation and ranking, especially with ChIP-seq data
5. Investigate how choice of benchmark data (PBM and ChIP-seq) influence motifs rankings

3.4 Methods

The main reasons that motif assessment remains an open question are the lack of standardised benchmark data, the dynamic nature of the field, our incomplete understanding of TF-DNA binding and the disparity in the approaches used by various groups to assess their motifs. Mostly, the approaches are assumed to be comparable, but that is not always the case. Progress in the field requires standardised approaches for comparability and reproducibility. Here, we test the methods previously used and determine how each influences motif assessment.

3.4.1 Data

Many experimental techniques, reviewed in Chapter 2, have been used to learn TF binding site models in various forms. For this study, we use PBM and ChIP-seq data as benchmark test data and TF binding specificity models generated from different experimental techniques in MEME format. We used motifs from several databases and publications listed in Table 4.1 (Page 72).

For ChIP-seq data, we downloaded uniformly processed peak sequences from the ENCODE consortium [45][1] is from several cell lines listed in Table A.1. The list of ENCODE data used in this chapter can be accessed from GitHub[2]. The peaks are then converted to BED format using custom scripts, which processed the peak files using the BEDTools v2.17.0 [152] – implemented via pybedtools [38] – to extract the 5% of the highest scored sequences from repeat-masked human genome version hg19. We provide details on the specific data processing in the sections that follow. For PBM data, we adopted the definition of positive and negative sets described by Chen et al. [29]. In short, we select probes whose binding intensity is $4 \times MAD$ (MAD=0.675 for a normal distribution) above the median binding intensity as positive hits with the remaining probes being used as negative hits. We select 500 probes when less than 500 fit the criteria.

The motifs are collected from a variety of databases and publications as summarised in Table 4.1 (Page 72). We select the 20 TFs used in this analysis based on ChIP-seq data availability and at least 10 PWM motifs.

3.4.2 Sequence scoring

For a given TF, each available motif is used to score each sequence in the positive and negative set by directly applying the scoring functions in Section 3.2.4. Next, the performance of the PWM is determined by its ability to correctly rank (correlation) the positive sequences or classify (AUC, MNCP) the two sets. The whole analysis is summarised in Figure 3.2. A completely reproducible IPython notebook for the experiments conducted for this chapter is available from GitHub[3].

3.4.3 Which factors influence motif ranking, and how?

This section describes how we tested for variations in motif evaluation approaches in motif ranking. We start by computing motifs summary statistics, including the average and full-length information content, the length and number of motifs available for each TF. Next, for ChIP-seq

[1] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/

[2] https://github.com/kipkurui/Kibet-F1000Research/blob/master/Supplementary_Tables/Table_S3_list_of_ENCODE_data_used.txt

[3] https://github.com/kipkurui/MARS_Evaluation/tree/master/Chapter3

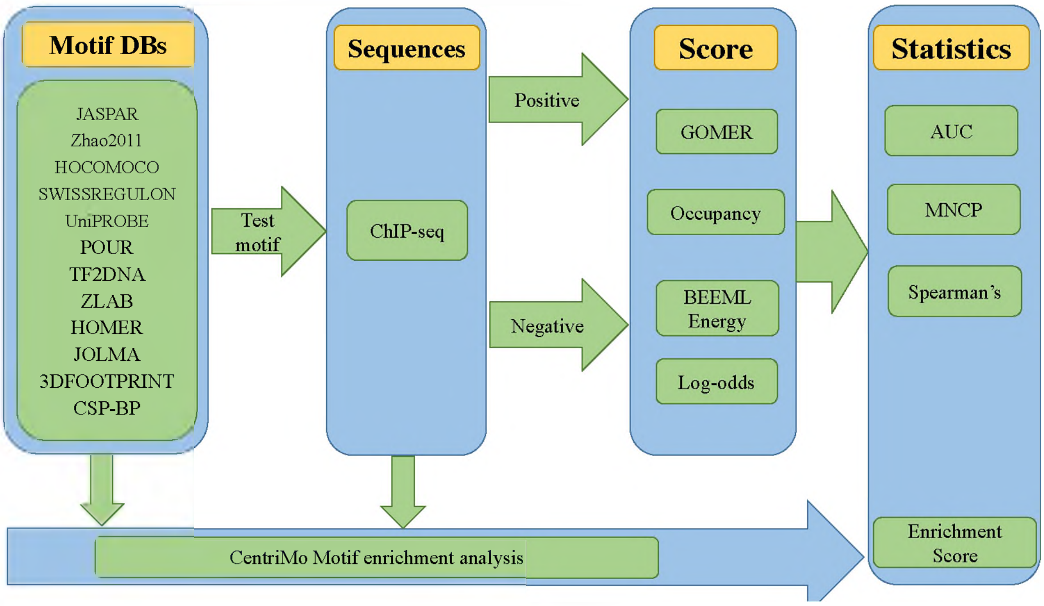


Fig. 3.2 Methodology flow diagram. For a given transcription factor, a majority of motifs available in various databases are extracted and used to score the given test sequences. The motifs are then ranked based on a given statistic.

data, we test the effect of sequence length (50, 100 and 250) centred on the ChIP-seq peaks in motif assessment. In these preliminary analyses, based on our initial results, we use GOMER and Energy scoring functions.

Evaluating choice of background sequences: The selection of background sequences is known to influence motif discovery; therefore, we test how the choices of negative sequences would affect motif ranking. We evaluate the following backgrounds:

- **Flanking sequences**: extracted 500bp from the highest coordinate in the positive set, irrespective of strand
- **Matched genomic**: genomic sequence that correspond to the positive sequences extracted using *gimme background* command from the GimmeMotifs tools [189]
- **GC matched**: the negative sequences are extracted from the genome to match the GC content of the positive set
- **Promoter sequences**: random promoter sequences
- **Dinucleotide shuffle**: dinucleotides in the positive set are shuffled to generate the background sequences

Evaluating statistics used: After calculating the scores of each motif for the sequences acquired, binding prediction can be evaluated by various statistics. We compute AUC using the *scikit-learn metrics* implementation, while for Pearson's and Spearman's correlation, we use *Scipy.stats* package. MNCP is implemented and used as in GimmeMotifs.

Evaluating Scoring functions: Having established the optimal length and the best choice of background sequence in ChIP-seq, we use that to evaluate the scoring functions. First, we determine the level of agreement of motif rankings based on the various scoring functions. Finally, we evaluate the statistics (MNCP, AUC, Pearson's and Spearman's correlation) that provides the best discrimination for each function.

Evaluating benchmark data: For this analysis, we only found nine TFs that had comparable data in ChIP-seq from ENCODE and PBM. These are Egr1, Esrra, Gata3, Hnf4a, Mafk, Max, Myb, Pou2f2 and Tcf3. The data from Badis et al. [9] were downloaded from UniPROBE database [133].

3.5 Results

3.5.1 Length of sequences has a little effect on motif ranking

For a successful ChIP-seq experiment, the TF's putative binding site is located in a narrow region around the peak centre. However, it is not always clear how the sequence length centred on the peak is chosen; different lengths have been previously used in motif discovery. It is for this reason that we sought to investigate if sequence length has an effect on motif ranking, and if so, what the optimum length should be. We tested for 50, 100 and 250bp centred around the peak. For the purpose of this analysis, we use sequences of 100bp as a reference point to compare with the other length variations. We find that ChIP-seq sequence length has no significant effect on motif ranks: p=0.113, for 50 and 100; p=0.0545, 50 and 250; p=0.678, 100 and 250bp – Wilcoxon rank-sum test (Figure A.2). However, some TFs exhibit a significant difference between 50 and 100bp when using GOMER scoring (Elf1 and Sp1; Figure 3.3) and Energy scoring (Elf1, Gata3 and Sp1; Figure A.4).

How well the ranks between 100 and the rest agree seems to indicate how the TFs bind. Transcription factors like Egr1, Ctcf, Cebpb, Srf, Mafk, Me2fa that prefer shorter sequences – higher scores and strong agreement rank correlation at 50 or 100bp – are enriched at the

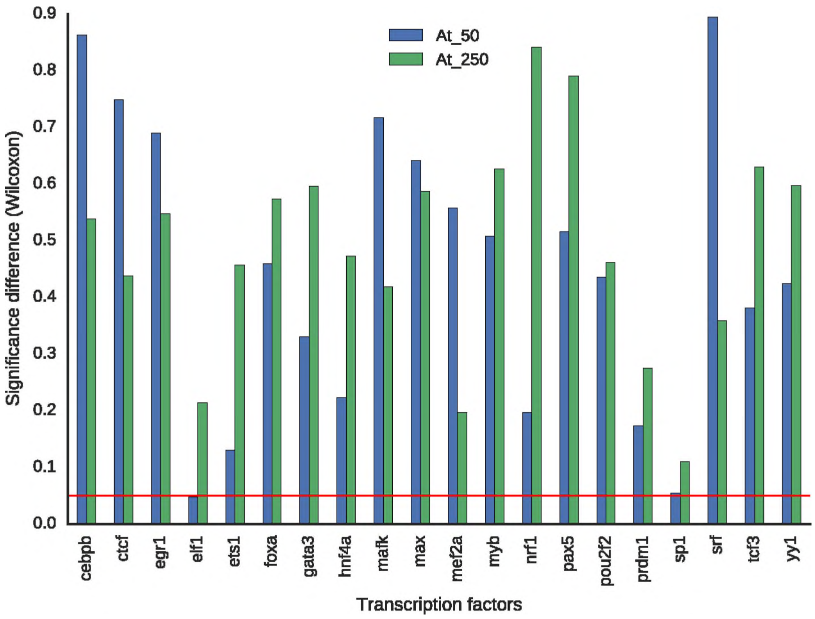


Fig. 3.3 Sequence length has no significant effect on motif ranking, except for some TFs. Using all the motifs for each of the 20 TFs, we tested the effect of sequence length (50bp, 100bp, and 250bp) using GOMER scoring on ChIP-seq data. For each TF, we tested for significant difference in motif rankings between "At 100" and the rest using AUC for all the 20 TFs. The horizontal red line represents the 0.05 significance threshold.

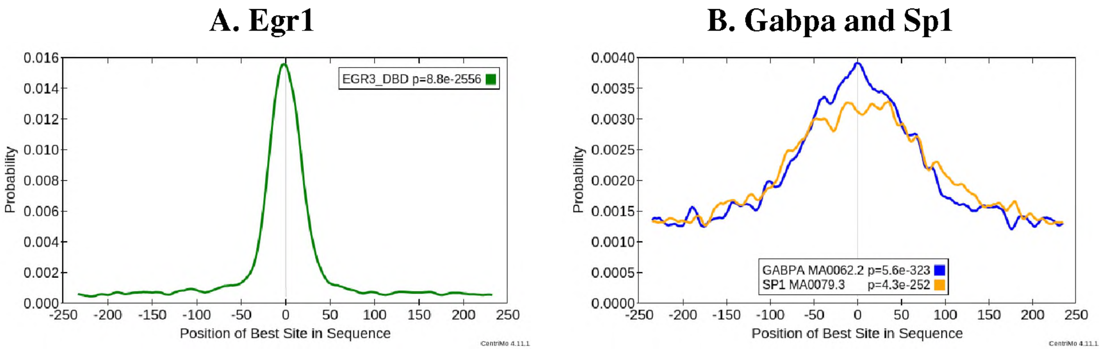


Fig. 3.4 Inferring binding from motif enrichment peaks. A: Egr1 motifs are enriched at the ChIP-seq peak centre in GM12878 demonstrating direct binding. **B:** Motif enrichment of Gabpa and Sp1 with broad peaks demonstrates cooperative binding as well as a preference for longer sequences. Motif enrichment performed using CentriMo 4.11.1, with all eukaryotic motifs hosted in the MEME database.

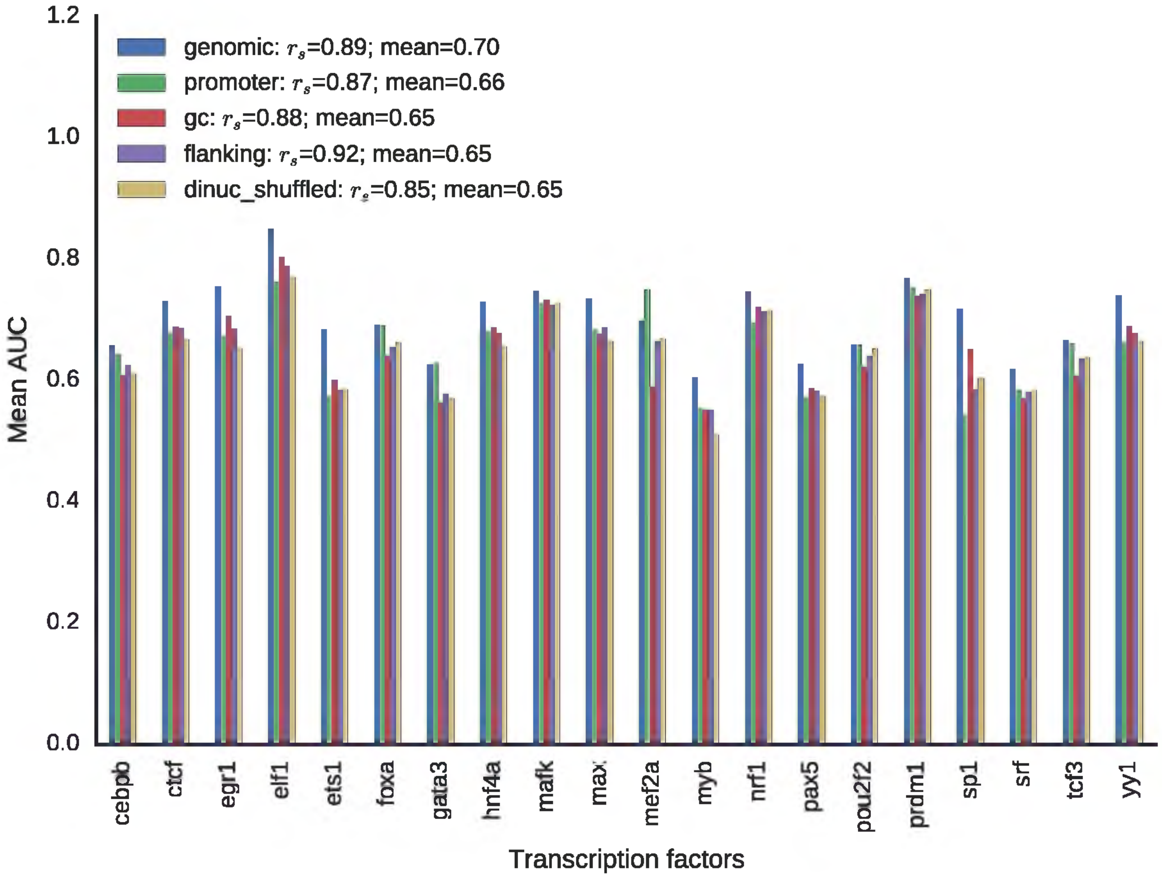


Fig. 3.5 Influence of negative sequences on motif ranking: GOMER. This is a plot of mean AUC values of all the motifs available for each TF for each background set. The figure legend provides the mean Spearman's rank correlation (r_s) and AUC. For each TF, the available motifs are used to score positive and two sets of negative sequence (see text for details).

ChIP-seq peak, which is a reliable indicator of direct binding [12]. Motif enrichment analysis of Egr1 using GM12878 ChIP-seq data confirmed this, where we observed sharp enrichment around the peak centre (Figure 3.4A). Others with significantly better AUC values at 250bp sequence length like Elf1 (p=0.017, Wilcoxon rank-sum test) and Sp1 (p=0.013, Wilcoxon rank-sum test), are known to bind cooperatively with Gabpa [183]. This is confirmed by motif enrichment analysis in GM12878 as shown in Figure 3.4B.

3.5.2 Choice of negative (background) sequences affects motif ranking

Having established the least biased sequence length (100 bp), we use this to evaluate the choice of negative sequence on motif ranks based on GOMER and energy scoring. We evaluate the results based on the motifs mean AUC scores and the Spearman's rank correlation for each TF. We use the consistency of the ranks with the other background choices to evaluate the choice of background sequence type. Using GOMER scoring as an example, although genomic background sequences were the most discriminative (0.7, mean AUC), the ranks based on flanking sequences are more consistent with the rest (0.92, r_s) – Figure 3.5. However, based on energy scoring, the genomic background is the least discriminative (0.58, mean AUC), while flanking background ranks motifs more consistently with the rest (Figure A.3). On average, for energy scoring, the ranks based on genomic background significantly differ from the rest (p=0.03, median Wilcoxon rank sum test), while the rest do not significantly differ from each other, except in a few TFs (Figure 3.6A). Surprisingly, no significant difference in the ranks is observed in GOMER (Figure 3.6B). Similar observations can be gleaned using MNCP statistics – see Appendix A.8 and A.7.

In addition to evaluating the choice of negative sequences, this analysis also reveals how the scoring function affects motif ranking. More on this in Section 3.5.5.

3.5.3 Tissue or cell line of the data could affect enrichment

Transcription factors bind to their possible sites in a sequence-specific manner [58]; some have alternative binding motifs depending on the tissue or cell line. For example, in Figure 3.7, the rank correlation of the motif scores in different cell lines can be as low as 0.83 for GOMER scoring (or as low as 0.65 using energy scoring). The correlation is expected to be close to one if the cell line had no effect. Also, FOXA1_2.GUERTIN motif is differentially enriched only in the A549 cell line (although this could be an outlier). Further scrutiny reveals that this resembles a nuclear receptor motif (Figure A.1), which may be a FOXA1 co-factor identified in motif discovery. Indeed, nuclear receptor motifs are known to be expressed in A549 cell lines [135].

To investigate this further, we test for significant difference in ranks from different cell lines within a TF using GOMER and energy scoring functions. With GOMER, the motif ranks did not differ significantly among the cell lines. However, with energy scoring, there is a significant difference in motif ranks for Cebpb, Srf, Gata3 and Sp1 between the cell lines listed in Table 3.2. With this possible effect in mind, the results displayed throughout this chapter are based on the mean score of all the available ChIP-seq or PBM data sets to avoid a bias towards cell line or experiment-specific motifs.

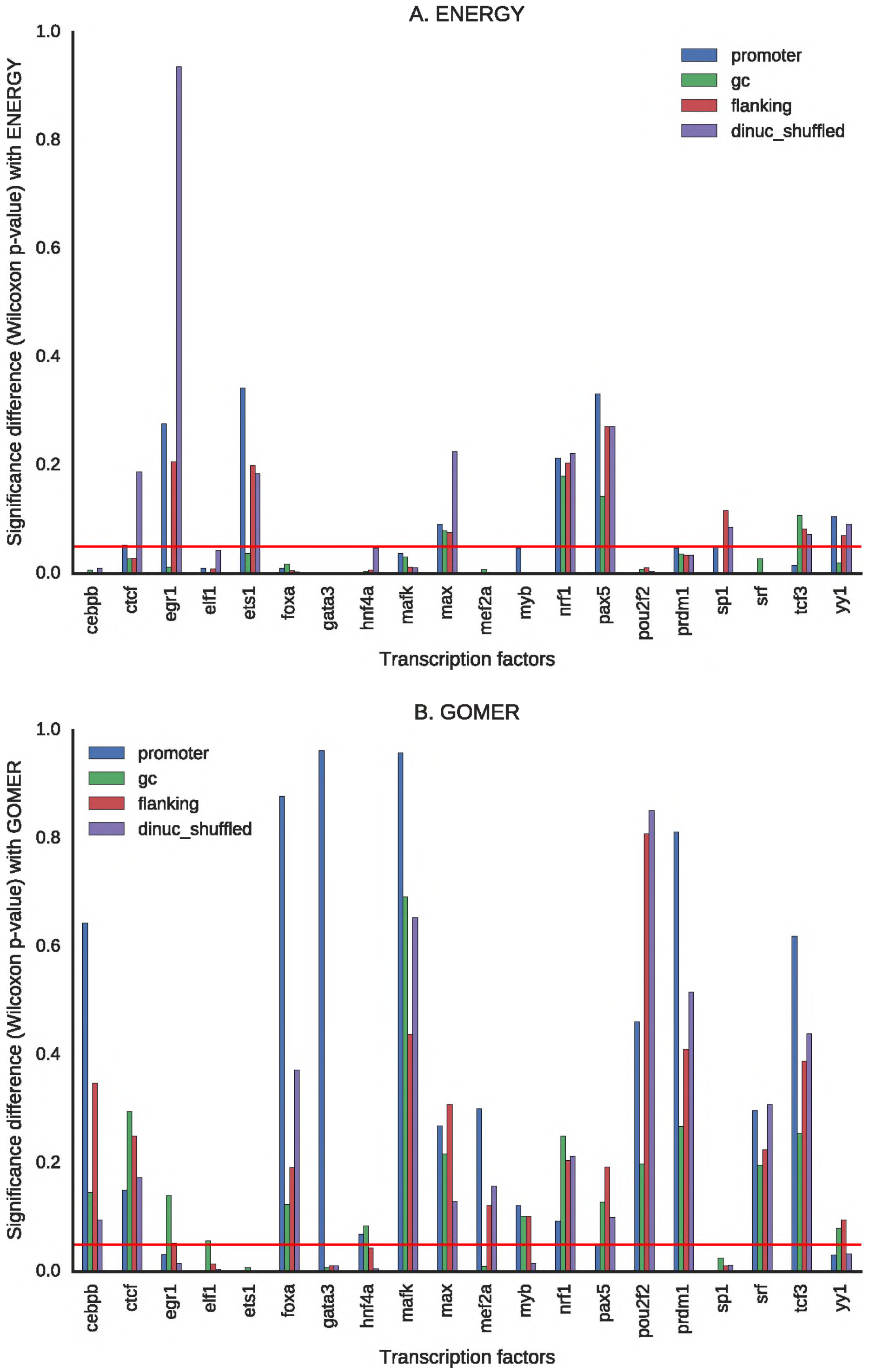


Fig. 3.6 Effect of choice of background sequence on motif ranks A: Based on energy scoring. **B:** Based on GOMER. We evaluate how motif ranks are affected when using the different background sequences with the genomic background. To do this, we tested for a significant difference in the scores assigned to the motifs using Wilcoxon rank-sum test. Significant difference can be linked to the effect of background sequence used. The horizontal red line represents the 0.05 significance threshold.

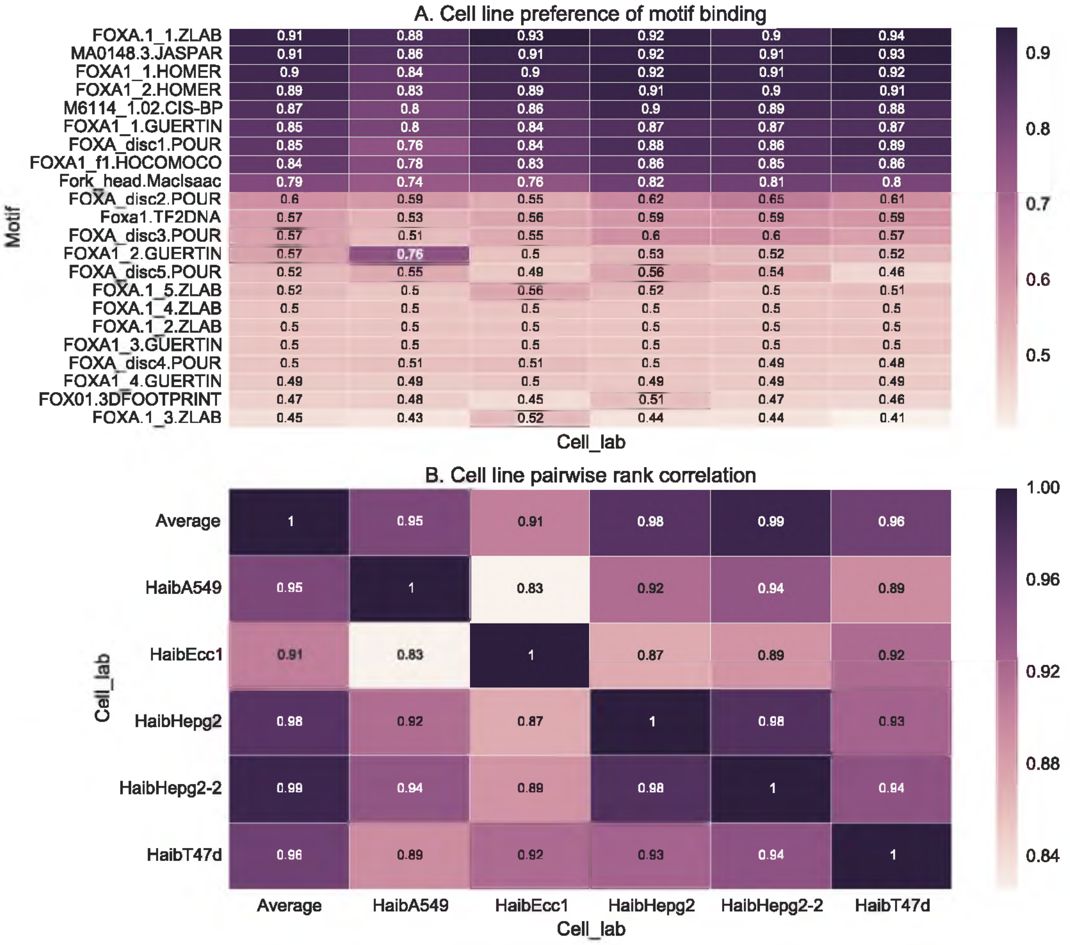


Fig. 3.7 Cell line-specific binding. A: How motif ranks can be influenced by the cell line used in the analysis. Foxa motifs are used to score each of the five cell lines using GOMER scoring and quantified with AUC values. Similar results are obtained with other scoring functions. **B:** How the ranks assigned to the motifs are correlated among the cell lines. Column headings are the same in A and B.

Table 3.2 Effect of cell line on motif ranking. The table displays the cell lines whose motifs significantly differ from each other using Wilcoxon rank-sum test.

TF	Cell line 1	Cell line 2	Wilcoxon p-value
Cebpb	HaibGm12878Cebpb	Average	0.0027
		HaibHepg2Cebpb	0.0004
		HaibK562Cebpb	0.0016
		SydhA549CebpbIggrab	0.0032
		SydhH1hescCebpbIggrab	0.0024
		SydhHelas3CebpbIggrab	0.0022
		SydhHepg2CebpbForskln	0.0029
		SydhHepg2CebpbIggrab	0.0036
		SydhImr90CebpbIggrab	0.0025
		SydhK562CebpbIggrab	0.0063
Gata3	HaibT47dGata3	SydhShsy5yGata3sc	0.0202
	SydhShsy5yGata3sc	SydhMcf7Gata3Ucd	0.0168
sp1	HaibHepg2Sp1Pcr1x	Average	0.0087
		HaibGm12878Sp1Pcr1x	0.0080
		HaibH1hescSp1Pcr1x	0.0027
		HaibK562Sp1Pcr1x	0.0068
Srf	HaibHepg2SrfV0416101	HaibK562SrfV0416101	0.0301

3.5.4 Effect of statistic on motif ranking

The statistic used, whether it tests motif's scores correlation or ability to classify the two sets of sequences, will have an effect on how we interpret the results of the analysis. From our comparisons, the motif ranks when quantified using AUC and MNCP statistics do not differ significantly (p=0.52, Wilcoxon rank-sum test), but the ranks based on Pearson's and Spearman's vary considerably from those based on MNCP or AUC statistics (p=0.006 and 0.002 respectively, Wilcoxon rank-sum test). To test this further, we determined how the scores vary from each other to check consistency and, ultimately, the suitability of statistics using standard deviation (STD) of the maximum normalised scores. The high STD of the correlation statistics' scores, as shown by the error bars in Figure 3.8, shows how unreliable the use of correlation statistics to rank the motifs can be. The correlation scores are also very low. Similar observations are made on PBM data (Figure A.15). When using MNCP, there is a higher rank correlation among the scores assigned by the different scoring functions except log-odds scoring (Figure A.16B). When using the AUC or MNCP statistic, Ctcf, Egr1 and Hnf4a score

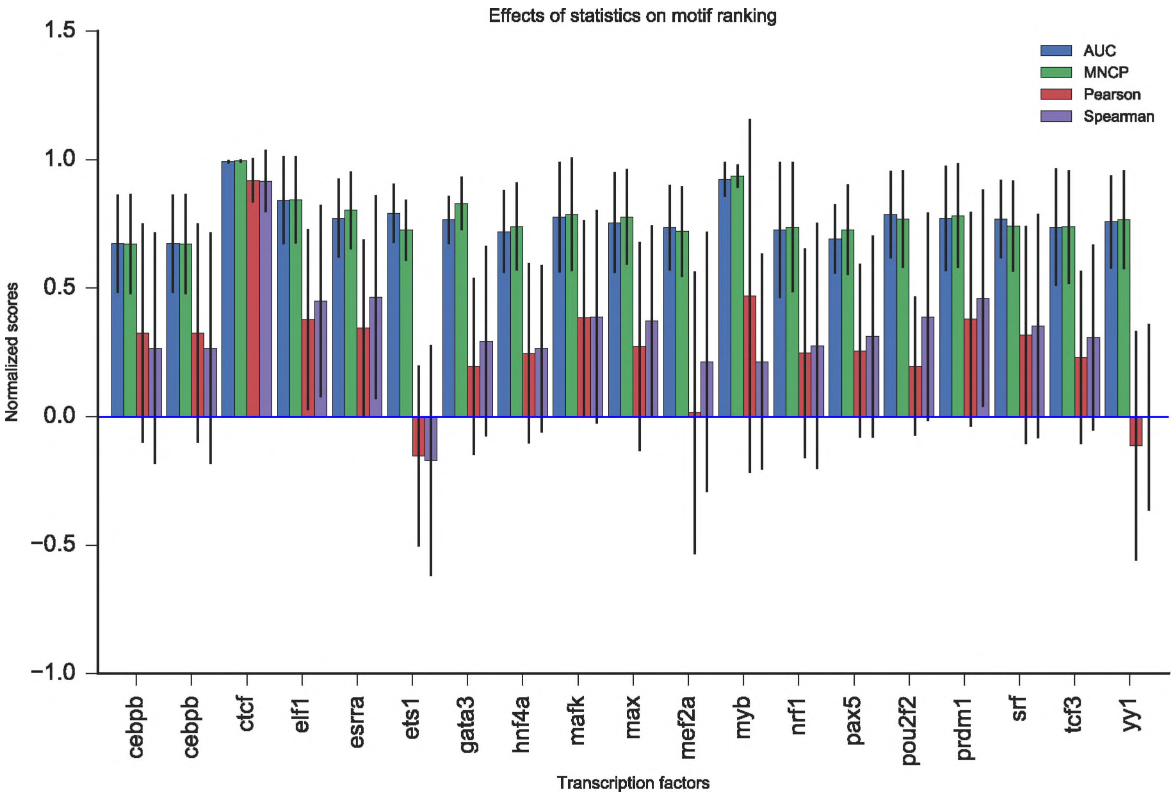


Fig. 3.8 Statistics used influence motif ranks. For each TF, the motifs are used to score sequences using the GOMER scoring function and ranks determined by MNCP, AUC, Pearson's and Spearman's rank correlation. In this figure, we compute the mean normalised scores and compute the standard deviation for each TF, which is displayed as error bars.

significantly higher using energy while for other TFs like Pou2f2 and Esrra, the preference is reversed (Figure A.16A).

3.5.5 The scoring function used has a transcription factor-specific effect

To determine how the scoring function used affects motif ranking during quality assessment, we tested how the PWM models can discriminate positive and negative sequence sets using five scoring functions. For most motifs, the maximum and sum log-odds scoring had low discriminative power when AUC (Figure 3.9) and MNCP (Figure A.16) statistical measures are used. However, sum log-odds scoring performed reasonably (over 0.55 AUC scores) for some TF motifs like Max, Nrf1, Tcf3, and Pax5. The reason behind this performance, however, is not clear. With MNCP, there is a higher rank correlation among the scores assigned by the different scoring functions except for two log-odds scores (Figure A.16B). When using AUC or

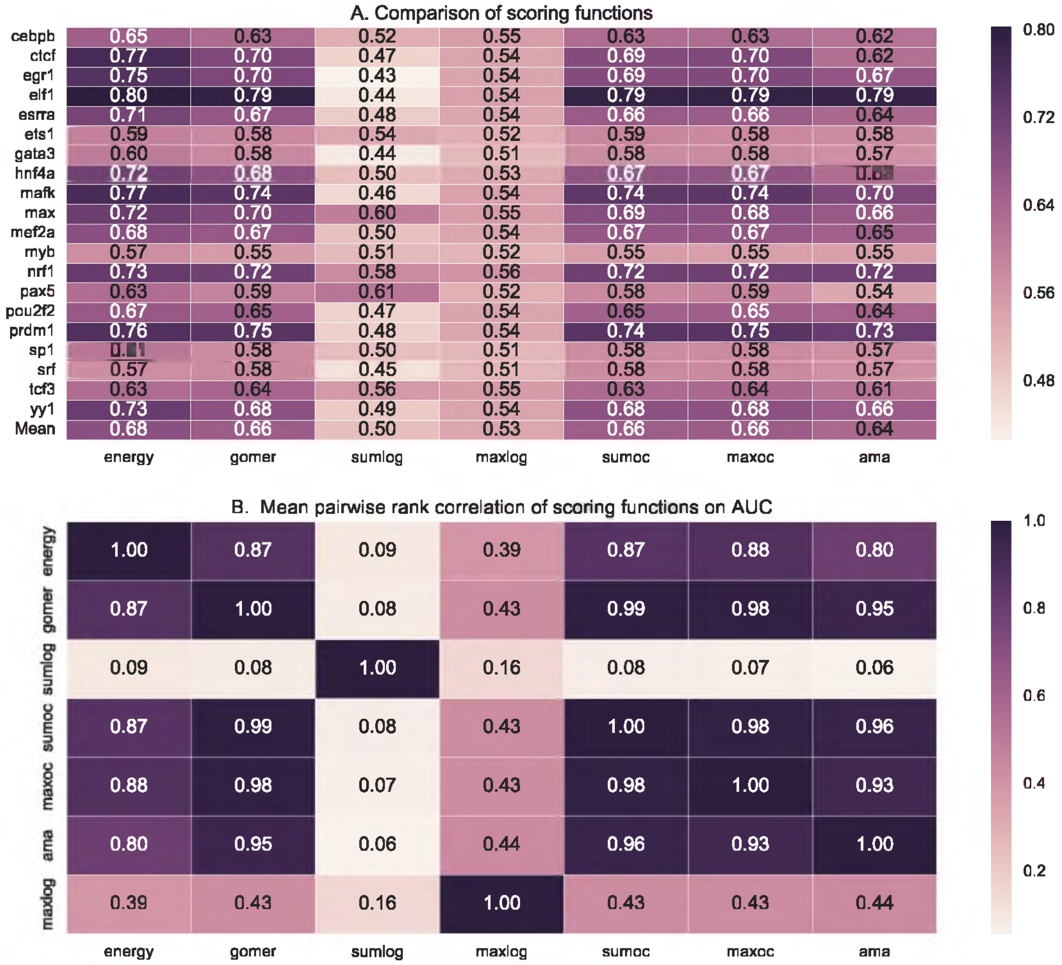


Fig. 3.9 Effect of scoring function on motif ranking using AUC statistic. A: For each TF, the mean AUC score is computed for each of the scoring functions used. **B:** Shows how the ranks assigned to various motifs for a given TF by each scoring function are correlated. It displays the pairwise rank correlation for all TFs in **A**. *Sumlog*: Sum log-odds, *Maxlog*: maximum log-odds, *Sumoc*: sum occupancy score and *Maxoc*: maximum occupancy.

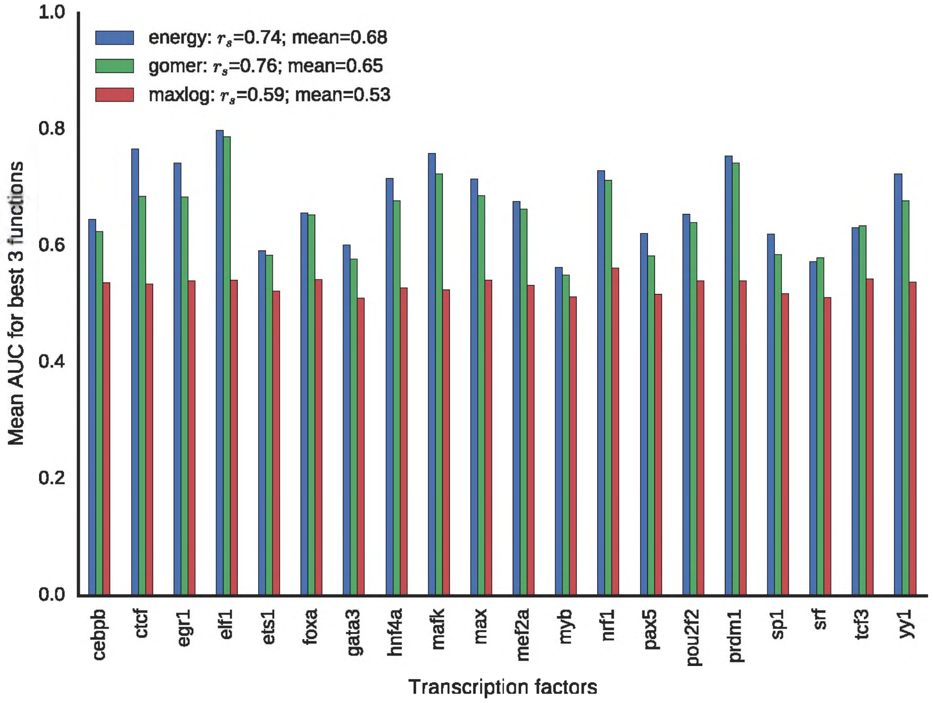


Fig. 3.10 Effect of scoring functions among the best three non-redundant approaches using AUC. The mean Spearman's correlation (r_s) provides a measure of how motif ranks for a function compares with the rest.

MNCP statistic Ctcf, Egr1 and Hnf4a score significantly higher in energy while for other TFs like Pou2f2 and Esrra, the preference is reversed.

To further determine how the scoring functions compare, we use GOMER with AUC as the reference to determine how the motif ranks compare with the rest using Spearman's correlation. The ranks only differ significantly for log-odds scores (sum, 0.04 and max, 0.014) but not for energy, 0.58; Sum occupancy 0.87; max occupancy, 0.79; AMA, 0.87 (Figure 3.9B). The consistency in motif ranks based on mean AUC scores makes this a useful measure of performance. Starting with the occupancy functions (GOMER, AMA, sum, maximum), we find that they perform equally: any can be used with confidence (Figure A.10). Having established that, we select GOMER scoring, energy and Max log score to compare their performance. We find that energy scoring is preferred with MNCP (r_s=0.76; mean=1.35) compared with GOMER r_s=0.75; mean=1.35) while GOMER is preferred with AUC (r_s=0.76; mean=0.65), though energy still has a higher mean AUC (rs=0.74; mean=0.68). See Figures 3.10 and A.9 for AUC and MNCP, respectively.

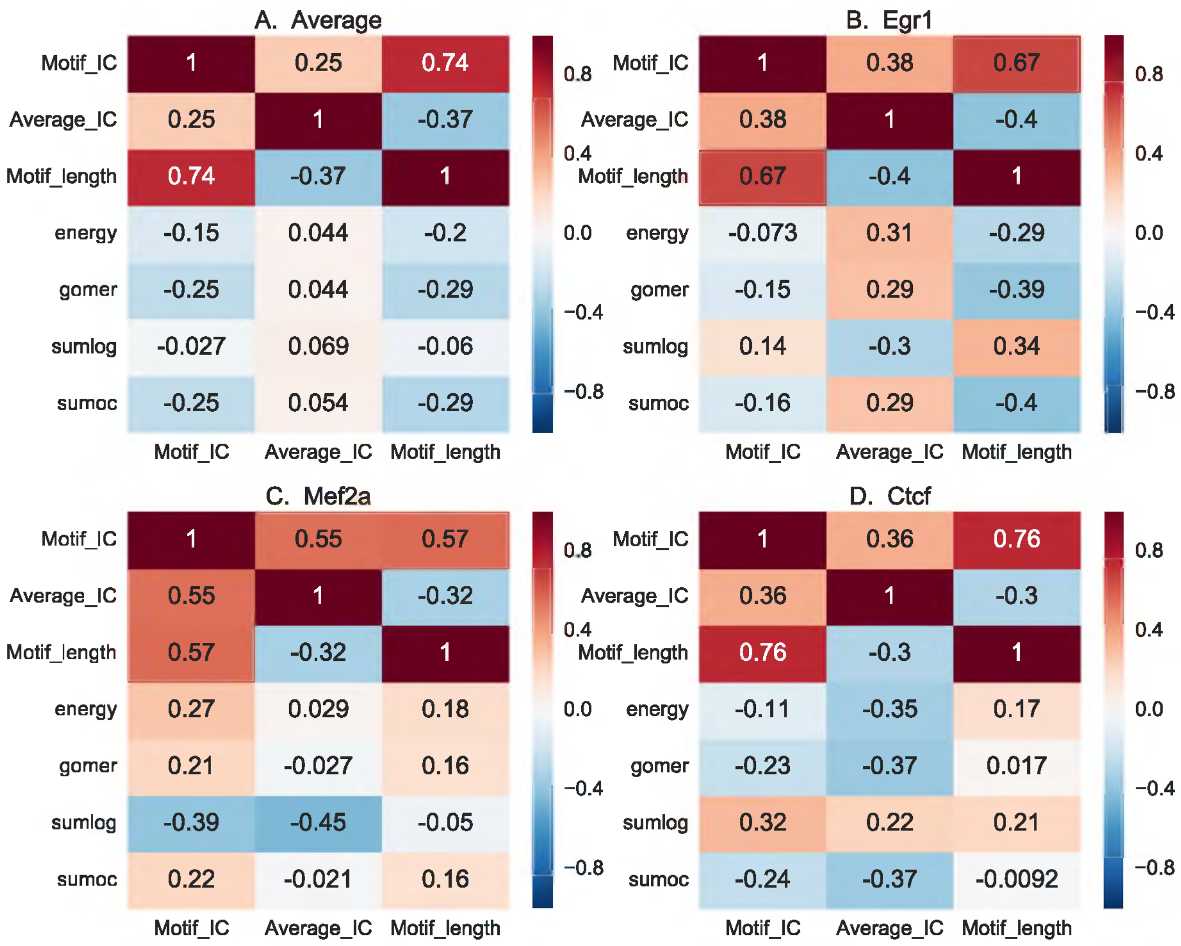
3.5.6 Motif information content is not a good measure of motif quality

Having established a TF-specific effect of scoring functions on motif ranking, we test how their performance can be explained by motif features (information content and length). To check how useful information content (IC) is a measure of motif quality, we determined the level of correlation between the various motif features (motif length, full-length IC, and average IC) and the scores attained by the motifs. We use four scoring functions: GOMER, energy, sum occupancy and sum log-odds score. See Figure 3.11. There is no consistent correlation between the average IC and the scores (Figure 3.11 A). However, on average, there is a negative correlation between both the total IC and motif length, and the scores except for sum log-odds scoring, which has no significant correlation (p=0.34, correlation p-value). This shows that longer motifs are not as discriminative, and also have low sensitivity for most TFs caused by high specificity. This is not a general rule. Some TFs depict a different scenario. For example, Egr1 (Figure 3.11B) has a positive correlation between average IC and scores and a negative correlation with motif length (except for sum log-odds scoring), showing that it has a highly specific binding site [97]. See Figure 3.4A for more evidence. Mef2a, on the other hand, has a positive correlation between motif length and scores showing a preference for longer high IC motifs (Figure 3.11C). This could also reflect variability in binding sites. Ctcf has the highest negative correlation for average IC, with a neutral to positive correlation for motif length (Figure 3.11D), which may indicate a preference for longer low IC motifs.

3.5.7 Effect of benchmark data on motif ranking

The effect of sequence data in motif performance has been previously investigated [140], with the finding that motifs generated from PBM data perform well when tested on the same type data but do not generalise well *in vivo*. Here, we tested how the ranks of the motifs are affected when evaluated against PBM and ChIP-seq data. As shown in Figure 3.12, the data used in motif evaluation affects motif ranks significantly in some TFs but has little or no effect on others. In addition, this also depends on the scoring functions and statistics used (Figure 3.12 and 3.13). Note that TFs with a significant difference with MNCP statistic have a positive effect size value in favour of ChIP-seq data, showing the mean in ChIP-seq is always greater than PBM data.

Also, we make the following observations, directly linked to the data used:



A. Average

	Motif_IC	Average_IC	Motif_length
Motif_IC	1	0.25	0.74
Average_IC	0.25	1	-0.37
Motif_length	0.74	-0.37	1
energy	-0.15	0.044	-0.2
gomer	-0.25	0.044	-0.29
sumlog	-0.027	0.069	-0.06
sumoc	-0.25	0.054	-0.29

B. Egr1

	Motif_IC	Average_IC	Motif_length
Motif_IC	1	0.38	0.67
Average_IC	0.38	1	-0.4
Motif_length	0.67	-0.4	1
energy	-0.073	0.31	-0.29
gomer	-0.15	0.29	-0.39
sumlog	0.14	-0.3	0.34
sumoc	-0.16	0.29	-0.4

C. Mef2a

	Motif_IC	Average_IC	Motif_length
Motif_IC	1	0.55	0.57
Average_IC	0.55	1	-0.32
Motif_length	0.57	-0.32	1
energy	0.27	0.029	0.18
gomer	0.21	-0.027	0.16
sumlog	-0.39	-0.45	-0.05
sumoc	0.22	-0.021	0.16

D. Ctcf

	Motif_IC	Average_IC	Motif_length
Motif_IC	1	0.36	0.76
Average_IC	0.36	1	-0.3
Motif_length	0.76	-0.3	1
energy	-0.11	-0.35	0.17
gomer	-0.23	-0.37	0.017
sumlog	0.32	0.22	0.21
sumoc	-0.24	-0.37	-0.0092

Fig. 3.11 Effect of motif length and IC on scoring functions. In this figure, we show the correlation of motif length, full-length information content (IC) and the assessment scores, to determine how motif characteristics influence the performance of scoring functions. For each motif, the information content is calculated based on information theory for the whole length and also normalised for length. The results for average motif affinity (AMA) and maximum occupancy are similar to sum occupancy, and are not included.

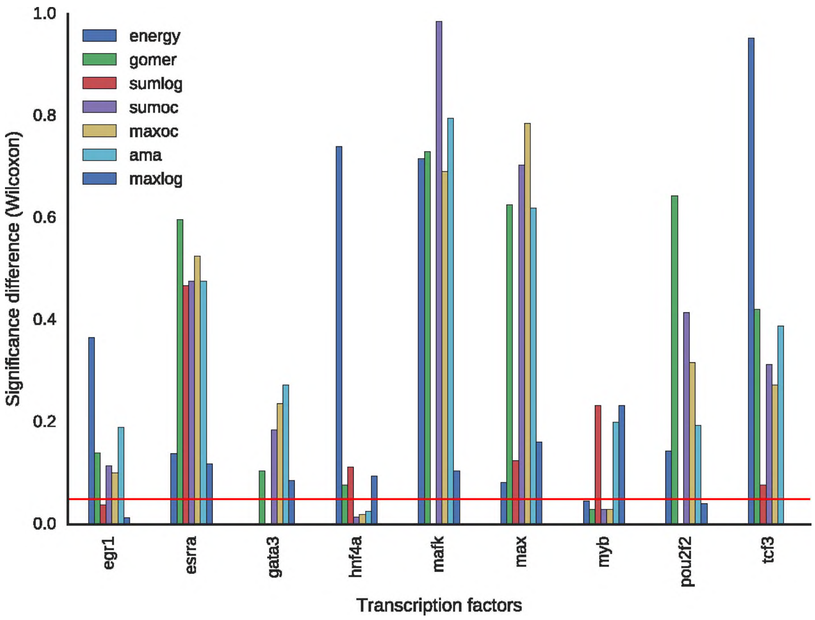


Fig. 3.12 Effect of benchmark data: AUC. We evaluate how the ranks of motifs are affected when using PBM and ChIP-seq data. To do this, we tested for significant difference (Wilcoxon p-values) in the scores assigned to the motifs. Significant difference can be linked to the effect of benchmark data used. The horizontal red line represents the 0.05 significance threshold.

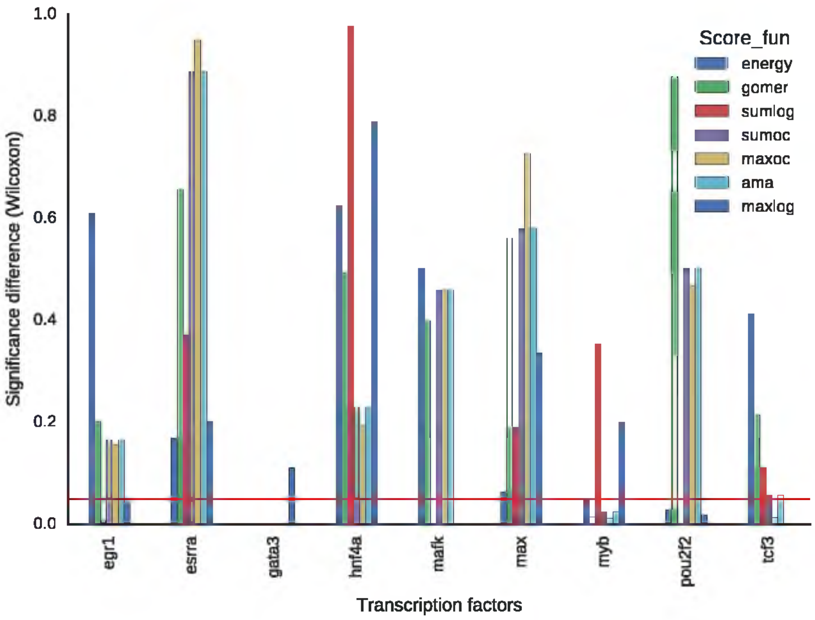


Fig. 3.13 Effect of benchmark data: MNCP. Details same as in Figure 3.12.

1. A much higher energy score in PBM (Figures A.11 and A.12) compared with ChIP-seq (Figures 3.9 and A.16), and a lower correlation between the energy and the occupancy scores.

2. A stronger negative correlation between the GOMER occupancy scores and motif IC -0.47 compared with -0.28 of energy scoring (Figure A.13), a lesser effect when using ChIP-seq data (Figure 3.11). This may explain observation 1.

3. Motifs generated using the PBM technique perform best when using occupancy scores with MNCP or energy scores with AUC or MNCP, except when occupancy scoring and AUC are used (Figure A.14). Poor ranking of UniPROBE PBM-derived motifs by GOMER-AUC may be linked to the fact that they penalise long motifs – UniPROBE motifs are long (mostly over 14bp).

4. Energy scoring with any of MNCP, AUC or occupancy scoring with MNCP display similar behaviour: a preference for specific motifs, which may be longer or have a higher IC. This is supported by the high negative correlation between motif length and occupancy scores with AUC (Figure A.13).

All these support the view that the data used in motif evaluation does have an effect on motif ranking.

3.6 Discussion

In our review, we classified motif assessment approaches into assess by binding site prediction, comparison, and scoring and classification while showing how the techniques in use have changed over time. The review revealed the complexity of the problem, which remains without an appropriate solution, and that the available evaluation approaches are developed for algorithm developers rather than the end users of the generated models. This affirms the need for a solution drawn up with end users in mind, for the purpose of selecting appropriate motifs from the multiple options presented. This study revealed a TF-specific effect of assessment approaches, data, scoring function and statistics used. The results provide a foundation for a user-centred platform.

The analysis on the effect of ChIP-seq sequence length revealed a TF-specific influence on motif ranking determined by the binding behaviour of the TF; whether direct (Figure 3.4A), indirect or cooperative (Figure 3.4B). This supports the observation that binding behaviour can be inferred from the shape of the distribution of ChIP-seq sites [12]. We found 100bp sequence length to provide the necessary discrimination for most of the TFs tested (Figure A.2) in line with the observation that in a ChIP-seq experiment, there is a high probability of a TF binding site being localised within 30bp of a ChIP-seq site [199].

Depending on the data used to generate the motifs, the choice of data to assess the quality of the model does affect motif ranking. As observed in Cebpb, Gata3, Sp1 and Srf, researchers should be aware of bias caused by the use of a single one cell line in an assessment. This could be linked to TFs that are not expressed in some cell lines or poor quality of ChIP-seq data. Therefore, if data from more than one cell line or experiment are available, results averaged over the data should be used to reduce the bias towards cell-line or experiment-specific motifs (Figure 3.7). Further, from the cell line effect, we also confirm a TF-specific influence of negative (background) sequences in motif ranking, an effect well known in motif discovery. Based on the consistency of motif ranks, we recommend flanking sequences followed by GC-matched sequences, consistent with Worsley Hunt et al. [203], where they found GC matched background sequences to provide the least skewed motif over-representation but did not test for flanking sequences.

The statistic used in motif evaluation affects the ranks. Of the statistics examined, we find MNCP and AUC provide ranks that are in agreement but exhibit TF-specific variations. The MNCP and the AUC statistics' results do not differ from each other because they are both rank-order metrics [34]. However, MNCP is more sensitive to differences in the motifs, as observed in this study. It is similar to AUC but derived by plotting ranks of true positive hits against the ranks of all hits. This places emphasis on true positives and therefore is less affected by false positives [34, 112]. Most of the observations from the PBM-based analysis support the conclusion that energy scoring prefers specific motifs (long or with a high IC). We also observe an agreement when we use energy scoring with AUC and MNCP, or occupancy scoring. In MNCP, the preference for specific motifs is expected and has been previously confirmed [112], because the MNCP score provides a rank-based measure of the ratio of motif mean occurrence in test sequences and a random set. These observations are not conclusive, and further research may be required. Finally, correlation statistics are not a reliable measure of motif quality, especially in ChIP-seq data, since the peak scores do not represent TF binding affinity, but the enrichment of bound sequences.

Although there is no clear winner among the scoring functions, occupancy based (GOMER, AMA, sum, and max) and energy scoring functions are preferred. We recommend, based on the presented evidence, using occupancy scoring with the MNCP statistic or energy scoring with AUC or the MNCP statistic. Energy scoring and MNCP are less affected by false positives providing more reliable rankings. We also confirm an observation by Orenstein et al. [137] that sum occupancy was better than maximum occupancy scoring, though not significantly so (p=0.85, Wilcoxon rank-sum test).

The debate as to whether IC is a measure of motif quality or not continues. In our analysis, we do not find any significant correlation (p=0.513, correlation p-value) between the IC and the motif scores (Figure 3.11); an observation contrary to the ones that best-quality motifs have low IC motifs [197], or high IC motifs [138]. Weirauch et al. did not normalise for motif length, which results in high IC motifs that are longer but not necessarily more specific [197]. In this case, a shorter motif with higher IC per nucleotide will be more specific but have lower total IC. The relationship between IC and motif quality, therefore, depends on the TF binding behaviour. TFs with short and specific binding sites will favour high IC while those with long and variable binding sites are modelled more accurately with low IC. Furthermore, it has been shown the low IC flanking motif sites contribute to the specificity of binding *in vivo* [138, 52], and are known to influence TF binding site shape [42], which in turn determines specificity. Taken together, these results show that motif IC should not be used as a measure of motif quality or to optimise an algorithm to generate better motifs, unless the TF is known to have a preference for high IC motifs. Previous studies have shown that TFs with variable binding sites are longer to maintain specificity [177]. This can be confirmed by the negative correlation between the average IC and motif length (Figure 3.11A).

How the data used in evaluation affects motif ranking depends on the data used in motif discovery. As expected, motifs generated from *in vitro* based experimental data perform better when tested on *in vitro* data. The UniPROBE motifs performed better when tested on PBM data, an observation which is expected and has been previously established [137]. TF2DNA motifs also perform better on PBM data, possible because they were generated *in vitro*; however, this could also be explained by the short length of the motifs (7bp). We note a higher negative correlation between length and motif scores in PBM data (Figure A.13) compared with ChIP-seq data (Figure 3.11A), further supporting short length to be the main cause of TF2DNA's better performance. Nonetheless, the results show that approaches used in motif assessment should, therefore, be determined by the intended use of the motifs.

We have confirmed that motif assessment has transcription-specific variability, an observation previously made [211]. Therefore, assessments should on how different motifs for a particular TF rank, rather than on how a particular motif database or algorithm performs. As more information becomes available on how each TF binds, motif discovery, TF binding site prediction and TF occupancy analysis can be tailored for each TF. This will facilitate accurate and specific TF binding specificity models. For the end user, no single database can provide the sole measure of the quality of new data, raising the need for collation of the different motifs tested using a variety of motif assessments to provide information to the end user on their ranks. We address this need in Chapter 4.

3.7 Chapter conclusions

Motif assessment remains an open question. However, the analysis presented in this chapter provides a starting point for a better understanding of the factors that negatively influence motif ranking. From this review and analysis, we can make the following conclusions:

1. The length of ChIP-seq peaks used in motif evaluation has little effect on motif ranking, but with TF-specific variation. Sequences of 100bp long are acceptable for the majority of the TFs, hence this length is recommended.

2. The scoring functions affect motif ranking in a TF-specific manner, but energy scoring is least biased, hence recommended. GOMER scoring achieves similar specificity when used with MNCP statistics.

3. The data, algorithms and the statistics used are confirmed to affect motif ranking.

4. The choice of background sequences in evaluation influence motif ranking, but flanking and GC-matched sequences provide the most consistent ranking.

5. In summary, differences in motif evaluation approaches have a TF-specific effect on motif ranking.

Furthermore, data processing, motif assessment and optimisation during motif discovery do affect the quality of the derived motifs. Therefore, algorithm developers should choose scoring functions that best optimise for generalised models. When it comes to motif discovery pipelines, especially those using more than one algorithm, the motif assessment approach to use when selecting the best motifs is imperative. Our analysis provides information useful when making that decision. We have also shown that IC affects motif quality as influenced by TF binding behaviour; it is not necessarily a measure of motif quality. IC should be used together with other data.

We have demonstrated a need for tools that an end user can use to rank the motifs for a particular TF and make a decision based on the intended end use. This tool should be flexible on the data it accepts, and the scoring functions and statistics to use to facilitate systematic motif quality evaluation. Lessons learned from the analysis in this chapter have been useful in many ways. Firstly, it advised the design of the web-based application that can allow users to compare motifs available in different databases for a particular TF presented in Chapter 4. Secondly, it prompted the extension of motif comparison approach to avoid 'reference motif' bias (Section 4.4.3).

Chapter 4

MARS: Motif Assessment and Ranking Suite

> *"Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand it, you can't control it. If you can't control it, you can't improve it."*
>
> –H. James Harrington

We describe MARS (Motif Assessment and Ranking Suite), a web-based suite of tools used to evaluate and rank PWM-based motifs. The increased number of learned motif models that are spread across databases and in different PWM formats, leading to a choice dilemma among the users, is our motivation. This increase has been driven by the difficulty of modelling transcription factor binding sites and the advance in high-throughput sequencing technologies at a continually reducing cost. Several experimental techniques have been developed resulting in several motif-finding algorithms and databases. We collate a wide variety of available motifs into a benchmark database, including the corresponding experimental ChIP-seq and PBM data obtained from ENCODE and UniPROBE databases, respectively. The implemented tools include a data-independent consistency-based motif assessment and ranking (CB-MAR) and scoring and classification algorithms (SC-MAR). CB-MAR is based on the idea that 'correct motifs' are more similar to each other while incorrect motifs will differ from each other; SC-MAR ranks binding models by their ability to discriminate sequences known to contain binding sites from those without. The CB-MAR and SC-MAR techniques have a 0.86 and 0.73 median rank correlation using ChIP-seq and PBM respectively. In addition, motifs selected by CB-MAR achieve a mean AUC comparable to those ranked by held out data at 0.75 versus 0.76

– we performed ChIP-seq motif discovery using five algorithms in 110 transcription factors. We have demonstrated the benefit of this web server in motif choice and ranking as well as in motif discovery. It can be accessed at www.bioinf.ict.ru.ac.za.

4.1 Background

In Chapter 3, we reviewed and evaluated techniques to assess motifs, with a particular interest in scoring functions used. We demonstrated the lack of a standardised approach and the need for an easily accessible platform that can assist users to choose motifs, and algorithm developers to benchmark their techniques. This chapter addresses that need and briefly reviews what is already available, including gaps and weaknesses, and our attempt at filling that gap. To put this chapter into perspective, we take a step back and talk about the evaluation problem in Bioinformatics, identify where it has been well addressed, and seek to emulate the best practices in our effort.

High throughput sequencing at a continually reducing cost has generated a considerable amount of data, which in turn led to a proliferation of algorithms and statistical tools to deal with the data. The choice dilemma introduced by this increase necessitated independent evaluation of the tools to advise the end users and developers on what works. In fact, previous independent evaluations have helped spark progress and high-quality tools [137]. As expected, this cycle continued to the need to design a biologically relevant and standardised benchmark [6, 158] – a question of "who watches the watchmen" [72]. To this end, we adapt Aniba's criteria [6] of a good benchmark to evaluate the available tools and, ultimately, in the design of our benchmark. These are:

- **Relevant** – biologically meaningful results should be derived from the evaluation
- **Solvable** – tests on the benchmark should not be trivial but must be possible to use with reasonable effort
- **Scalable** – the benchmark should be expandable to cover new techniques and algorithms as they develop
- **Accessible** – data and statistical tools should be easy to source and use to evaluate other algorithms or protocols

- **Independent** – methods should not be tailored or biased to a particular algorithm or experimental techniques

- **Evolvable** – the benchmark should change as new data are made available, as well as to reflect the current problems and challenges in the field

This evaluation problem is widely addressed in multiple sequence alignment [6, 72, 101], 3D structural modelling by CASP challenges [130], Protein–Protein docking by CAPRI [73] and Protein-protein interaction [129]. However, it is still an active challenge in gene regulatory research, especially in predicting TF binding sites and the accuracy of prediction models [207]. This difficulty is directly linked to the motif discovery problem, which is yet to be resolved and is attributed to the degeneracy and ubiquity of TF binding sites in the genome [79, 60]. Nevertheless, there have been some attempts at the evaluation problem. What follows is a short review of the motif assessment tools and techniques against the above criteria. We specifically focus on the web and standalone tools for assess-by-binding site prediction, motif comparison or by sequence scoring and classification [92] approaches – see Chapter 3 for a detailed description of these categories.

An assess-by-binding-site prediction approach evaluates algorithms by their ability to identify known or inserted binding sites in the sequence. A few tools implementing assess-by-binding site prediction exist. Tompa et al. [186] provide a web server[1] accompanied by experimental benchmark data. In this web server, the user downloads the data, predicts binding sites in the data and submits the results back to the web server for evaluation and ranking against the tools in the database. Later, Sandve et al. [163] introduced a web server[2] with both synthetic and experimental benchmark data. The web server uses a machine leaning-based discriminatory algorithm that ranks motif models based on their ability to discriminate between positive (fragments with binding sites, known or inserted) from the negative set (fragments without known sites). However, these tools neither *evolved* nor *scaled* with advances in motif discovery algorithms; this reduced their *relevance* and thus failed to meet major requirements we set for an evaluation benchmark.

Assess-by-scoring and classification tests binding models by their ability to discriminate sequences known to contain binding sites from those without. For PBM data, a web server[3] that was used in the DREAM challenge to evaluate the ability to predict TF binding intensity

[1] http://bio.cs.washington.edu/assessment/
[2] http://tare.medisin.ntnu.no/
[3] http://www.ebi.ac.uk/saezrodriguez-srv/d5c2/cgi-bin/TF_web.pl

[197] exists; evaluates models based on ability to predict PBM binding intensities in held out data. For ChIP-seq data, Swiss Bioinformatics hosts a simple web server, PWMTools[4], that tests motifs on ENCODE data using sum occupancy scoring; it does not allow for comparative evaluation of motifs or sequence data (not *relevant*).

Assess-by-motif-comparison is used to determine if the discovered motifs are similar to those in 'reference databases' using motif comparison algorithms [216, 137, 131, 55]. An algorithm is considered successful if it can predict a motif similar to those in the database. However, this assumes the accuracy of previous predictions (not *relevant* or *scalable*), a weakness we address in this study.

There are few available standalone tools since this is not a regular task performed by the users. MTAP stands out in this category. It was developed specifically to compare the performance of motif finding algorithms [151] rather than the motif models. However, the complexity of the tools and the dynamic nature of the motif finding problem rendered the tool quickly outdated (not *accessible* or *relevant*). Therefore, recently, DynaMIT was developed, offering a dynamic motif integration toolkit that allows the users to add and test new tools continually [39]. Although it is not a motif assessment platform, it offers functionality for running multiple motif-finding tools and ranks the generated models using its *motif-integration* module. GimmeMotifs [189], like DynaMIT, is an ensemble motif discovery algorithm tailored for ChIP-seq data; it also offers functionality for motif assessment, which is not only used internally but available for independent use. Since these are not (in a strict sense) motif assessment tools, they do not ship with benchmark data.

The spread of motif models across DBs and in different PWM formats makes it difficult to create a benchmark that ranks multiple motifs for a given TF, and the growth in available data further compounds this problem [92]. There is a lack of an easily *accessible* and *independent* motif evaluation platform that can allow users to rank PWM models for a given TF. To fill this gap, we introduce a web server that hosts a suite of motif assessment tools used to evaluate and rank motifs. For wider applicability, we collect ChIP-seq and PBM data generated from different labs and use an average score to represent a given motif, with the assumption that this would capture the most general binding behaviour. We also apply a wide variety of scoring functions and statistics to reduce technique bias. Finally, we introduce a novel consistency-based motif evaluation approach that we call Consistency-Based Motif Assessment

[4] http://ccg.vital-it.ch/pwmtools/pwmeval.php

and Ranking (CB-MAR), which can be considered *independent*, hence less biased compared with the scoring-based techniques.

4.2 Chapter aims and objectives

In Chapter 3, we established how the choice of scoring functions and data processing influences motif ranking. In this chapter, we use the lessons learned to create a benchmark and motif evaluation platform that meets the set criteria to ensure biologically significant rankings. Based on the identified gaps and questions that remain unanswered in the motif evaluation problem, we aim, in this chapter, to develop and make accessible tools for motif assessment and ranking. Specifically, we address the following objectives:

1. Generate comprehensive benchmark data sets from a variety of data for evaluating and ranking TF binding specificity models in PWM format

2. Create a database containing the majority of the available published motifs for each TF and their corresponding benchmark data where available

3. Create standalone tools for motif evaluation; addressing weaknesses like the 'reference motifs' bias in the current use of motif assessment by comparison approach

4. Provide a flexible and easy to use interface to the tools to allow users with little or no skill in programming to rank motif models for a given TF against those available in the database

5. Provide a means of visualising the motif ranks based on a given metric and assessment technique

4.3 Benchmark data

For motif assessment and ranking, we require two levels of benchmark data: experimental test data and motifs. In what follows, we describe how the data are sourced and processed.

4.3.1 Experimental benchmark data: ChIP-seq and PBM

We downloaded ChIP-seq data from ENCODE [45][5], PBM data from UniPROBE [133][6], prepared as described Section 3.4.1 and stored in a MySQL database (Table 4.1). For ChIP-seq, there have been many different ways of processing the data to evaluate algorithms and TFBS models as reviewed and tested in Chapter 3. Based on the results of that analysis, we settled on the following protocol for creating positive and negative test data:

1. The ChIP-seq peaks are converted to BED format and widened to 100bp using custom bash scripts.

2. Using BEDTools v2.17.0 [152], we extract the 5% highest-scored sequences or 500 (whichever is greater), from repeat-masked human genome version hg19, as the positive sequences.

3. For the negative set, we extracted 500bp downstream from the highest coordinate (highest coordinate + 500) of the positive sequences.

4.3.2 Benchmark motifs

Transcription factor binding motifs in PWM format were collated from a variety of databases (See Table 4.1 for details). These motifs, in various formats, are converted into MEME format using custom and MEME Suite conversion scripts and finally to SQL tables using custom Python scripts for storage in the database.

4.3.3 Data storage

The benchmark data described are stored in a MySQL relational database. All information about the source of the data, publications and links are stored; this ensures that source is cited for complete transparency and recognition. The basic design of MySQL database hosting the data is as shown in Figure 4.1. Inconsistency in TF names is resolved by linking alternative names to a TF-ID. The TF-ID at the Genus level derived from the TFClass [202] (Figure 4.2) uniquely identifies a TF. Alternative TF names searched from GeneCards [162], are used to link motifs, ChIP-seq or PBM data to TF-ID. For dimers, a combination of two class IDs, separated

[5] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/

[6] http://thebrain.bwh.harvard.edu/uniprobe/downloads.php

Table 4.1 Source of motifs in the benchmark databases. The ‘size’ refers to the number of motifs, and ‘Av. IC’ is the average information content: Total IC/Motif Length.

Database	Source	Size	Av. IC	Av. length	Reference
JASPAR	Mixed	127	1.24	10.56	[122]
UniPROBE	PBM	386	0.72	16.31	[133]
Jolma	HT-SELEX	843	1.22	12.66	[79]
Zhao	PBM-BEEML	419	0.49	10.00	[215]
POUR	ChIP-seq	292	1.24	11.30	[91]
HOCOMOCO	Mixed	426	1.05	12.28	[98]
SwissRegulon	Mixed	297	1.20	12.40	[142]
TF2DNA	3D Structures	1314	1.10	8.44	[150]
HOMER	ChIP-seq	264	1.10	12.01	[65]
Chen2008	ChIP-seq	12	1.26	13.25	[30]
3DFOOTPRINT	3D Structures	297	1.24	10.57	[35]
GUERTIN	ChIP-seq	609	1.16	15.70	[118]
CIS-BP	Mixed	734	0.60	20.00	[198]
ZLAB	ChIP-seq	409	1.18	16.69	[193]
Wei-Human	Microwell-based assay	27	0.99	10.00	[196]
Wei-Mouse	PBM	48	1.24	10.00	[196]
Hallikas	Microwell-based assay	6	1.28	10.33	[57]
MacIsaac	ChIP	56	1.20	9.29	[116]

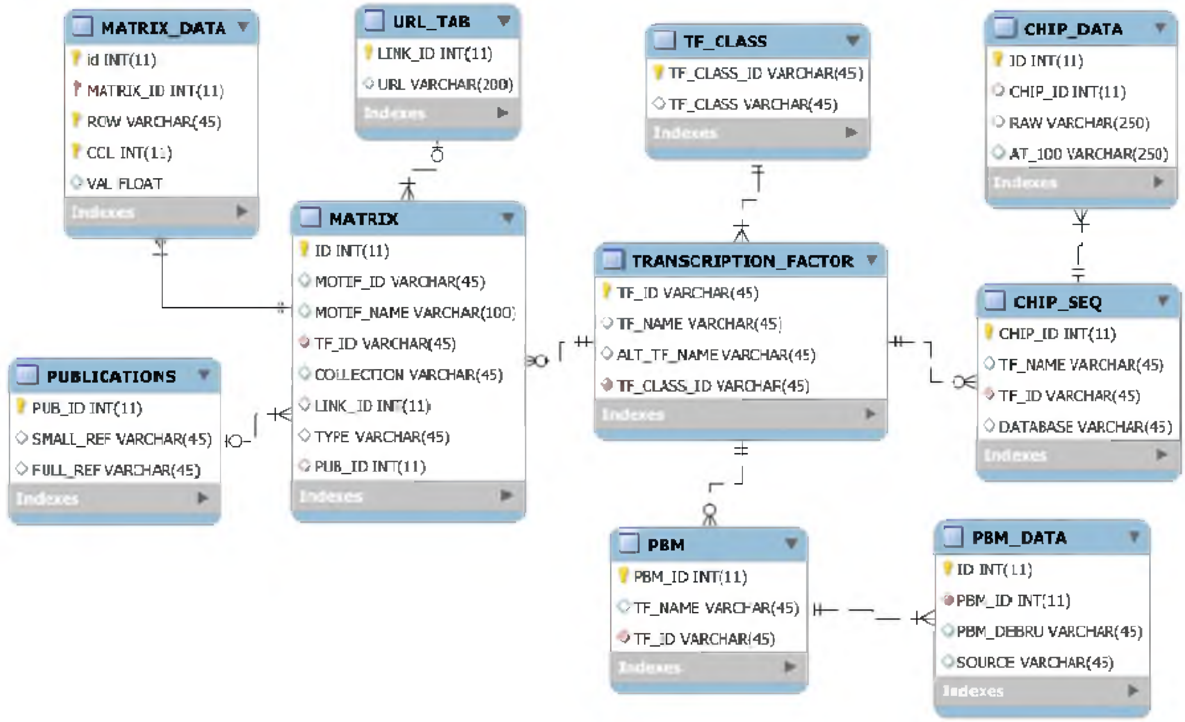
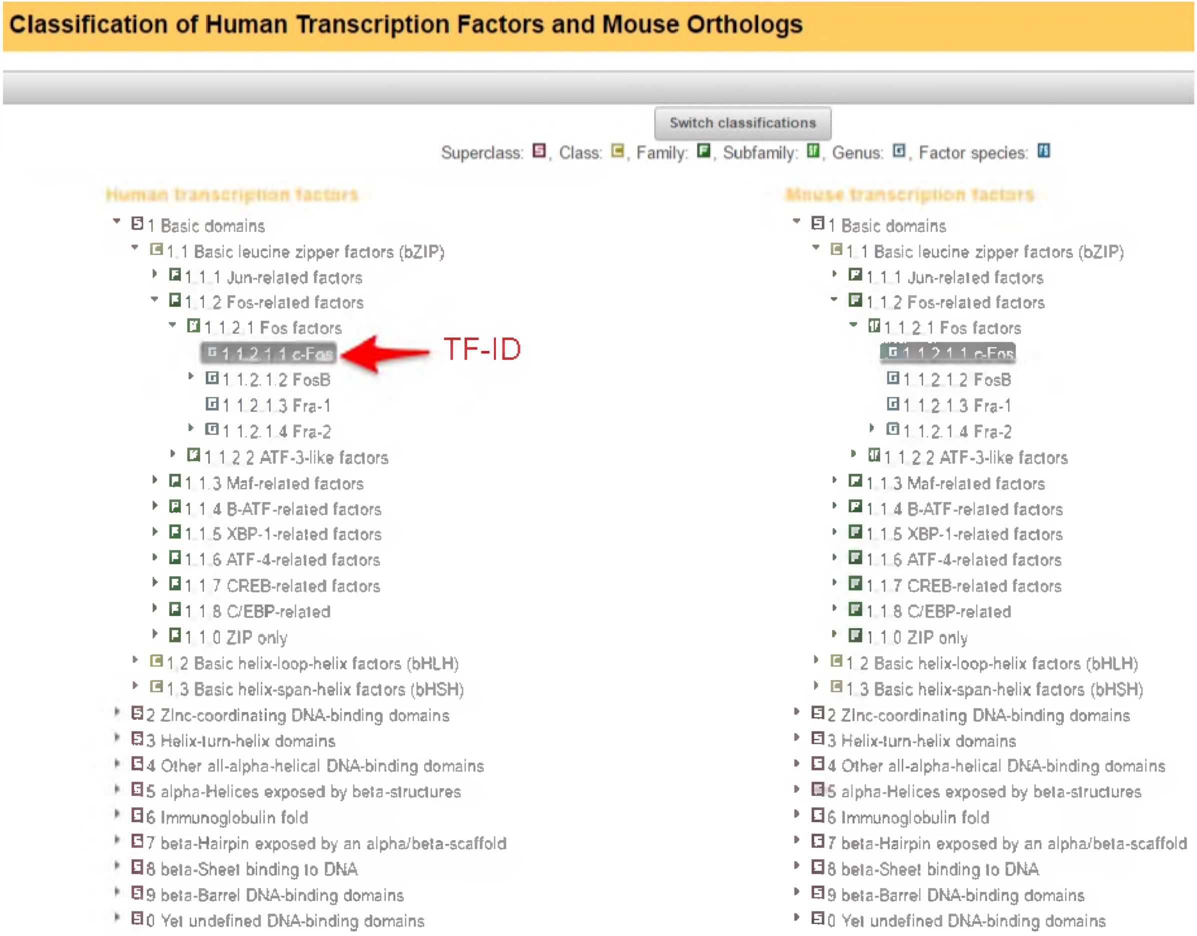


Fig. 4.1 ER schematic representation of the MARS database structure and relationships. Unified by the TRANSCRIPTION_FACTOR table, the PBM, CHIP_SEQ and MATRIX (PWM) tables are liked by the TF_ID sourced from TF-Class [202]. In MATRIX table, the COLLECTION column stores information on the source of the motifs, TYPE column stores the experimental techniques used to generate the motifs and the MATRIX_DATA table stores the actual data – support information like the references and URLs to the motifs are stored in separate tables. The PBM table stores details of the PBM-sourced benchmark data with the actual data stored in the PBM_DATA table. Finally, the CHIP_SEQ table stores the ChIP-seq sourced benchmark information with the actual data stored in the CHIP_DATA table.



Classification of Human Transcription Factors and Mouse Orthologs

Superclass: ■, Class: ■, Family: ■, Subfamily: ■, Genus: ■, Factor species: ■

Human transcription factors

- 1 Basic domains
  - 1.1 Basic leucine zipper factors (bZIP)
    - 1.1.1 Jun-related factors
    - 1.1.2 Fos-related factors
      - 1.1.2.1 Fos factors
        - 1.1.2.1.1 c-Fos
        - 1.1.2.1.2 FosB
        - 1.1.2.1.3 Fra-1
        - 1.1.2.1.4 Fra-2
      - 1.1.2.2 ATF-3-like factors
    - 1.1.3 Maf-related factors
    - 1.1.4 B-ATF-related factors
    - 1.1.5 XBP-1-related factors
    - 1.1.6 ATF-4-related factors
    - 1.1.7 CREB-related factors
    - 1.1.8 C/EBP-related
    - 1.1.0 ZIP only
  - 1.2 Basic helix-loop-helix factors (bHLH)
  - 1.3 Basic helix-span-helix factors (bHSH)
- 2 Zinc-coordinating DNA-binding domains
- 3 Helix-turn-helix domains
- 4 Other all-alpha-helical DNA-binding domains
- 5 alpha-Helices exposed by beta-structures
- 6 Immunoglobulin fold
- 7 beta-Hairpin exposed by an alpha/beta-scaffold
- 8 beta-Sheet binding to DNA
- 9 beta-Barrel DNA-binding domains
- 0 Yet undefined DNA-binding domains

Mouse transcription factors

- 1 Basic domains
  - 1.1 Basic leucine zipper factors (bZIP)
    - 1.1.1 Jun-related factors
    - 1.1.2 Fos-related factors
      - 1.1.2.1 Fos factors
        - 1.1.2.1.1 c-Fos
        - 1.1.2.1.2 FosB
        - 1.1.2.1.3 Fra-1
        - 1.1.2.1.4 Fra-2
      - 1.1.2.2 ATF-3-like factors
    - 1.1.3 Maf-related factors
    - 1.1.4 B-ATF-related factors
    - 1.1.5 XBP-1-related factors
    - 1.1.6 ATF-4-related factors
    - 1.1.7 CREB-related factors
    - 1.1.8 C/EBP-related
    - 1.1.0 ZIP only
  - 1.2 Basic helix-loop-helix factors (bHLH)
  - 1.3 Basic helix-span-helix factors (bHSH)
- 2 Zinc-coordinating DNA-binding domains
- 3 Helix-turn-helix domains
- 4 Other all-alpha-helical DNA-binding domains
- 5 alpha-Helices exposed by beta-structures
- 6 Immunoglobulin fold
- 7 beta-Hairpin exposed by an alpha/beta-scaffold
- 8 beta-Sheet binding to DNA
- 9 beta-Barrel DNA-binding domains
- 0 Yet undefined DNA-binding domains

Fig. 4.2 Transcription factor classification scheme. A TF can be uniquely identified by Genus ID (TF-ID), but factor species ID further classifies TFs with isoforms. Using c-Fos as an example, the TF-ID (1.1.2.1.1) provides information on its Superclass (1), Class (1.1), Family (1.1.2) and Subfamily (1.1.2.1) [202] (http://tfclass.bioinf.med.uni-goettingen.de/tfclass).

by double columns (::) were used. However, not all TFs have been assigned TF class IDs. These were left blank with the intention of resolving these names with the TF-Class curators.

4.4 MARS Tools

The Motif Assessment and Ranking Suite of Tools (MARSTools) consist of standalone tools for motif evaluation and ranking. They include Scoring and classification Based Motif Assessment and Ranking (SC-MAR), Enrichment-Based Motif Assessment and Ranking (EB-MAR), and Consistency-Based Motif Assessment and Ranking (CB-MAR). This section describes the implementation of these tools. MARSTools are hosted in Github[7].

4.4.1 Scoring and classification motif assessment and ranking (SC-MAR)

SC-MAR implements two separate algorithms: assess by score (hereinafter, SCORE) and *Gimme roc*. Gimme roc is a motif evaluation function which is part of GimmeMotifs, an ensemble motif discovery pipeline.

Table 4.2 Scoring functions implemented and the recommended statistics. The table lists the implemented scoring function and the corresponding recommended statistics for the best results. Log-odds functions, as implemented, are not discriminative; they are only available in MARSTools for comparative purposes, and are not recommended for use.

Scoring functions	Preferred stat.	Recommended?	Reference
Energy	AUC or MNCP	High (Default)	[213]
GOMER	MNCP	High	[29, 53]
Sum Occupancy	MNCP	Average	[137]
Max Occupancy	MNCP	Average	[217]
Max Log-odds		Not Recommended	[217]
Sum Log-odds		Not Recommended	[11]

The implementation of SCORE is described in Chapter 3. In summary, SCORE uses PWM motifs to score PBM (36 bp) or ChIP-seq (100 bp sequences partitioned into positive (foreground) and the negative (background) using one of the implemented scoring functions. It then evaluates the ability to classify the two sets using area under receiver operator characteristic curve (AUC) or the mean normalized conditional probability (MNCP) statistics. This is repeated for all the available motifs for a given TF, which can then be ranked based on one of the statistics. This

[7] https://github.com/kipkurui/MARSTools

is implemented in the *Assess_by_score.py* Python module. This module implements all the scoring functions and statistics described in Chapter 3 and summarised in Table 4.2. SCORE takes as input motifs in MEME format, scoring function and one or more test sequences for use in evaluation. The benchmark data can be a tab-delimited file with positive and negative sequences of equal sizes or BED file from which the positive and negative sequences are extracted as described Section 4.3.1. With these data entered, motif evaluation by SCORE proceeds as follows:

1. Convert each motif in the MEME input to a dictionary of lists. Each base is a key, and the value is a list of probabilities of the base appearing at each position in the PWM.

2. For each benchmark data, score each sequence with the PWM using the selected function.

3. Evaluate the motif using the AUC and MNCP. Pearson's and Spearman's statistics are also used but do not provide the best results, especially on ChIP-seq data.

4. Repeat the above for each motif in the input file.

5. Finally, SCORE returns the following data:

    - Summary scores for each motif (in a file named like *TF.scoring-function*, e.g. *Ctcf.gomer*)
    - The raw data from each test data used (*TF_raw.scoring-function*, e.g. *Ctcf_raw.gomer*)
    - Ranked PWM motifs in MEME format with AUC score appended to the motif names
    - A ranked histogram plot for each statistic used (Figure 4.12)
    - A clustergram of raw data (Figure 4.13)

On the other hand, *Gimme roc* takes as input motifs in PFM format, a foreground and background FASTA files, and prints to the console the motif scores based on MNCP, AUC 'maximum f-measure' and 'sensitivity at maximum f-measure' scores. We implement this as a *run_gimme.py* Python module that converts MEME motifs to PFM format used by *Gimme roc*, parses the output from different test data (cell lines) and plots the summary data of motif ranks as combined bar-plot for each statistic used.

4.4.2 Enrichment-based motif assessment and ranking (EB-MAR)

The differential enrichment of motifs in test as compared with the background sequences provides an indirect measure of motif quality. In this study, EB-MAR adapts the CentriMo motif enrichment analysis tool to the motif evaluation problem. It uses CentriMo version 4.10.0 in differential mode [106] – an option that tests differences in motif enrichment between two sequence sets – in a novel way for motif assessment. The differential mode parameters are set to search for local (anywhere in the full length of the sequences) rather than the central enrichment of all the input motifs in the positive (primary) and negative (control) set, derived as described in Section 4.3.1, by using a very high E-value threshold (100000). This threshold ensures CentriMo returns the enrichment of all the motifs instead of those above a threshold. The negative log of the E-value is used as the measure of a motif quality and for ranking. CentriMo scores sequences using log-odds scoring, declaring a maximally scoring site above a given threshold. It then counts the number of sites declared in each window. Finally, it determines the significance of central enrichment of the sites using a binomial p-value, which is normalised to E-value by the number of PWMs in the database.

4.4.3 Consistency-based motif assessment and ranking (CB-MAR)

> *"All happy families are alike; each unhappy family is unhappy in its own way."*
>
> – Leo Tolstoy, *Anna Karenina*

CB-MAR is based on the idea that 'correct motifs' are more similar to each other while incorrect motifs will differ from each other. The logic for this view is that differing methods are unlikely to reproduce each other's errors. This idea is used in evaluating sequence alignments: correct ones are assumed to compare with each other in a consistent manner, while inaccurate ones will differ from each other in various ways, generating inconsistent alignments [101, 72].

CB-MAR is based on Tomtom [56] and FISim [50] motif comparison algorithms, but any motif comparison algorithm could be used. CB-MAR first calculates a Pairwise Similarity Score (*PSS*) between all motif pairs and finally a Mean Similarity Score (*MSS*), which it uses as a measure of motif quality. For best results, the benchmark motif set should be: (*a*) generated from a variety of data and motif finding algorithms – with (*b*) completely similar motifs eliminated (especially in a small set) – and (*c*) large enough to capture variation in binding behaviours of the TF. The optimum number depends on the TF: one with uniform behaviour can be characterised by a smaller set of motifs than one with variable binding affinity, for example.

In more detail, CB-MAR is implemented as follows. Given a TF with a collection of motifs M of size n and using Tomtom Euclidean distance (ED) for motif comparison, we define (PSS) based on Tomtom P-value P for motif M_i and M_j, computed as:

$$PSS(M_i,M_j) = -log(P_{M_i,M_j}), \tag{4.1}$$

and then normalized by the maximum score of all PSS scores of M_i. The MSS, which we use as the measure of quality, of motif M_i, is then computed as:

$$MSS(M_i) = \frac{\sum_n^j PSS_{(M_i,M_j)}}{n} \tag{4.2}$$

To obtain a comparable metric, we modified FISim to take MEME files as input and return a matrix of PSS and the MSS for each motif.

4.5 MARS web server

The MARS web server is implemented in Django, a Python web framework, and hosted on an Apache web server with PWM motifs and sequence benchmark data stored in a MySQL database. MARS follows the traditional three-tier architecture with the data storage layer, the server layer, and the presentation layer (Figure 4.3). In addition to these layers, we refer to the MARSTools as the application layer. What follows is a brief description of how these layers are connected and the design principles we followed. We use the SCORE tools to describe the flow of information from the presentation layer, through the server layer which queries the data stored in the database layer to be used by the application layer.

4.5.1 Presentation layer

The presentation layer utilises open source Bootstrap[8], jQuery[9] and django-crispy-forms[10] frameworks for fluidity and interactivity. The forms (Figure 4.8), designed using django-crispy-forms, receive the user information, with interactivity enabled by the use of jQuery. The form

[8] http://getbootstrap.com/

[9] http://jquery.com/

[10] http://django-crispy-forms.readthedocs.io/en/latest/

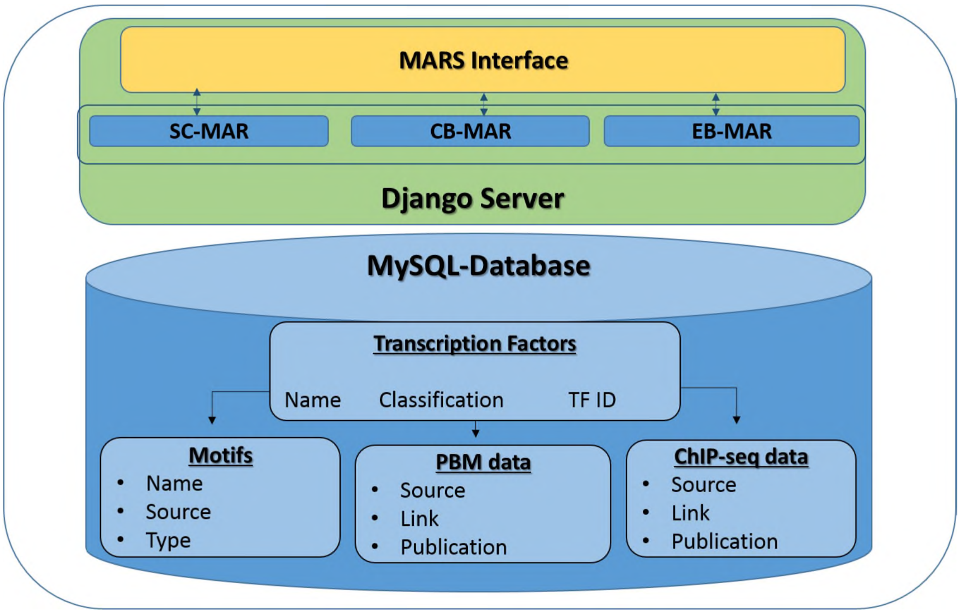
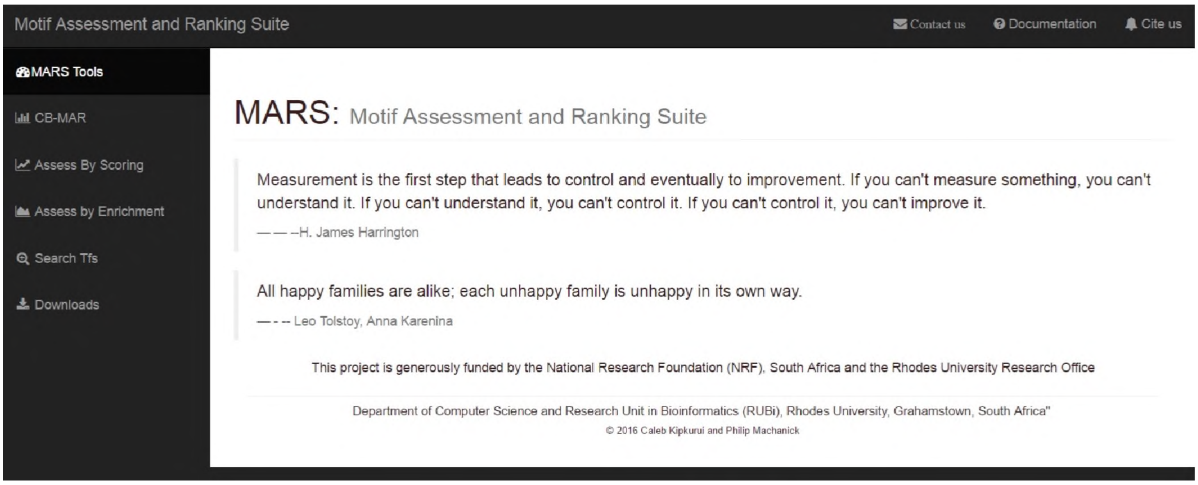


Fig. 4.3 MARS design diagram. MARS web server is designed using the Django framework with a MySQL back-end. The database currently contains 6,530 human and mouse motif models derived from available databases and publications. SC-MAR: Scoring and Classification Motif Assessment and Ranking, CB-MAR: Consistency-Based Motif Assessment and Ranking and EB-MAR: Enrichment-Based Motif Assessment and Ranking. See Figure 4.1 for details on the MySQL database design.

is responsive, in that it adapts to the user selections, thus only displaying the relevant fields. The user data is then submitted via AJAX calls to the server, where views dispatch jobs based on user requests, and the results are returned and presented interactively to the user.

4.5.2 Server layer

The server layer is designed using the Django web framework and is responsible for processing user data from the presentation layer. It consists of two Django modules: the *MARS* and the *search* module. *MARS* handles all the evaluation views and is responsible for distributing the jobs to particular algorithms based on user request. The search module (as the name suggests), allows the user to query the database for available motifs and benchmark data. The server layer is the driver of the whole system: it receives user information or data and queries the database for data, which it passes on to the application layer for processing.



MARS: Motif Assessment and Ranking Suite

Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand it, you can't control it. If you can't control it, you can't improve it.

— — —H. James Harrington

All happy families are alike; each unhappy family is unhappy in its own way.

— - – Leo Tolstoy, Anna Karenina

This project is generously funded by the National Research Foundation (NRF), South Africa and the Rhodes University Research Office

Department of Computer Science and Research Unit in Bioinformatics (RUBi), Rhodes University, Grahamstown, South Africa"

© 2016 Caleb Kipkurui and Philip Machanick

Fig. 4.4 MARS homepage and layout. The sidebar contains the MARSTools while the menu bar contains links to resources of help to the user: documentation, citation information and how to contact MARS administrator.

4.5.3 Data storage layer

Within the Django framework, the *models.py* Python module provides a connection to the MySQL database that stores the benchmark data.

4.5.4 Application layer

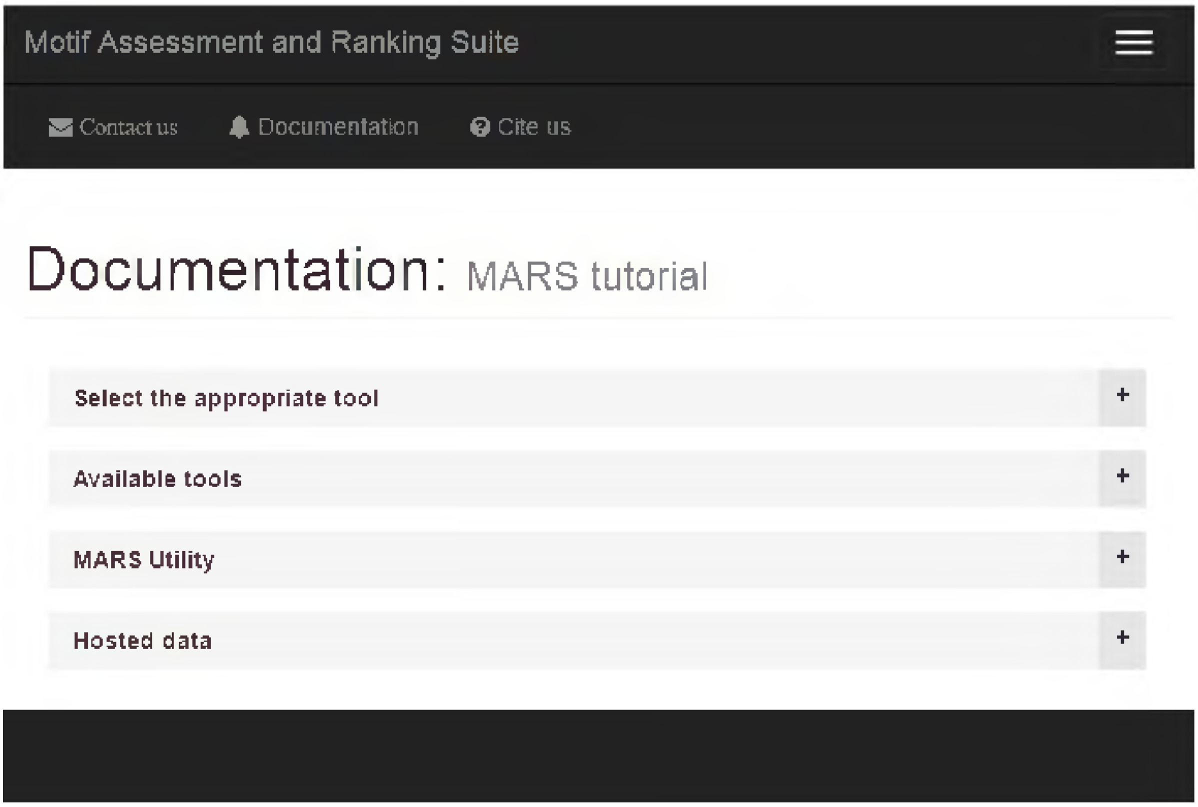
The MARSTools described in Section 4.4 form the application layer. These are standalone tools that perform motif assessment and provide the ability to visualise the results. The visualization plots are generated using Seaborn [195], MatplotLib [71] and Pandas [125] data analysis and visualization frameworks. Details of requirements and installation instructions are provided on the downloads page[11].

4.6 Results and discussion

4.6.1 MARS interface

The main interface to MARS is as shown in Figure 4.4. The MARSTools can be accessed from the sidebar, while additional pages of benefit to the user – documentation, contacts, and citation information – can be found on the menu-bar.

[11] www.bioinf.ict.ru.ac.za/downloads



Motif Assessment and Ranking Suite

Contact us Documentation Cite us

Documentation: MARS tutorial

Select the appropriate tool +

Available tools +

MARS Utility +

Hosted data +

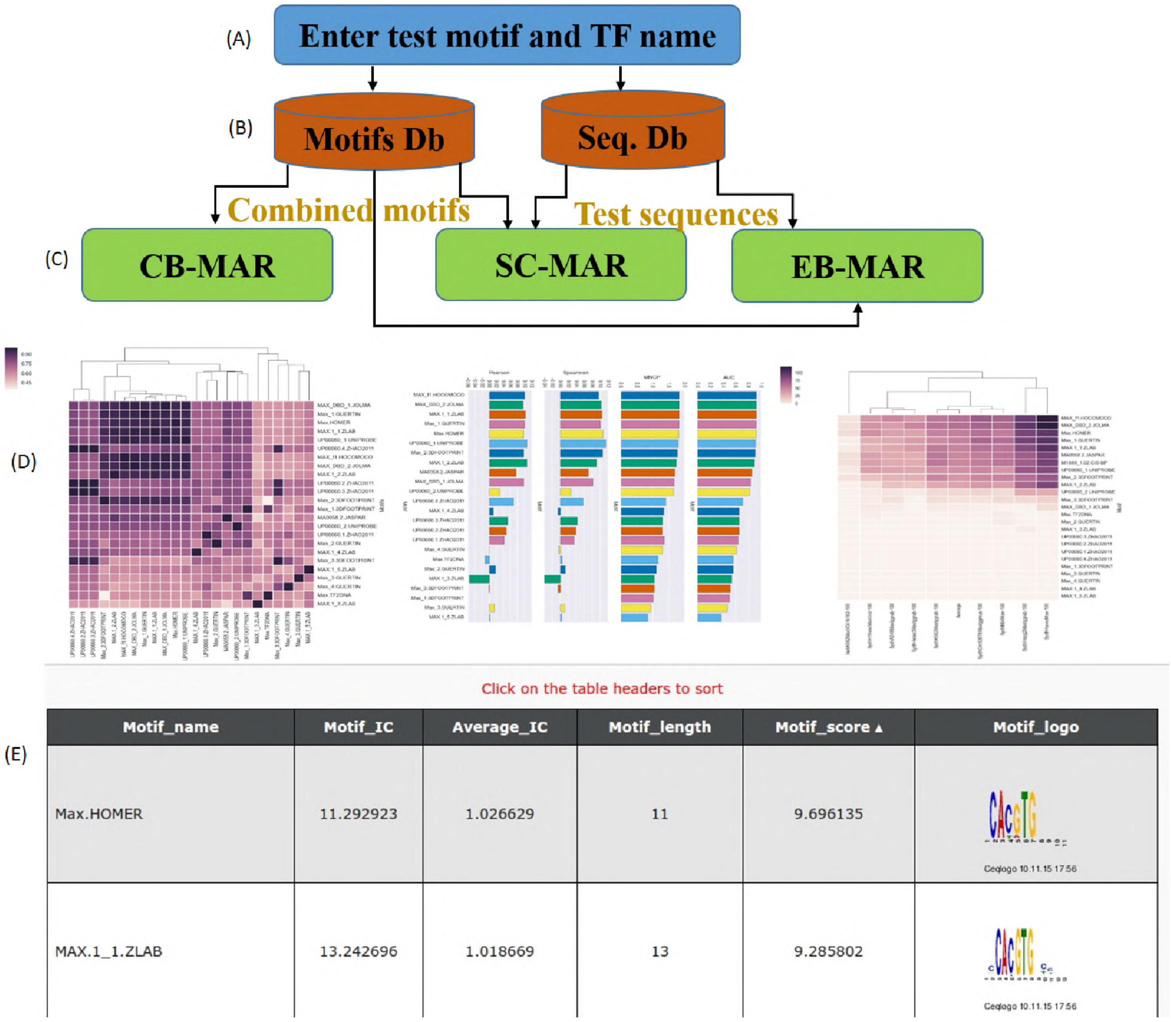
Fig. 4.5 MARS documentation. Provides details of accepted data formats, implemented tools and a short tutorial on how to use the tools in MARS.

Documentation: MARS documentation (Figure 4.5) provides quick information to guide the users on how to use tools, as well as providing a summary of benchmark data available. It includes a decision diagram (Figure 4.7) to guide the user on the best tool to use for a particular task, based on the available data.

Search: In addition to MARSTools, MARS also provides a search functionality that allows the users to explore and even downloaded available data (only motifs, for now) for a given TF.

4.6.2 MARS utility

MARS is designed to allow the users to either retrieve ranked motifs for a given TF or rank their own, as long as the required test data is available or uploaded (Figure 4.6). For any analysis, the TF name is the only required input; used to retrieve the available motifs from the database, and benchmark data in the case of SC-MAR and EB-MAR methods. Where the data is not available, the user is prompted to upload a motif. MARS currently accepts motifs in MEME format and



Click on the table headers to sort

Motif_name	Motif_IC	Average_IC	Motif_length	Motif_score ▴	Motif_logo
Max.HOMER	11.292923	1.026629	11	9.696135	Ceqlogo 10.11.15 17:56
MAX.1_1.ZLAB	13.242696	1.018669	13	9.285802	Ceqlogo 10.11.15 17:56

Fig. 4.6 MARS usage flow diagram. **(A)** User uploads a motif in MEME format and enters its TF name. **(B)** Motifs and test sequences linked to the TF are extracted from the database. **(C)** Motifs can be ranked, by comparison, used to score test sequences (rank by score) or their enrichment determined using CentriMo (rank by enrichment). **(D)** The results are visualised interactively, **(E)** with additional information like motif length, information content, and logo. The clustergram (combined heatmap and cladogram) offers additional information on the motif or test data clustering. In the end, the user can download ranked motifs in MEME format, as well as raw data for further analysis. SC-MAR: Scoring and Classification Motif Assessment and Ranking; CB-MAR: Consistency-Based Motif Assessment and Ranking; EB-MAR: Enrichment-Based Motif Assessment and Ranking.

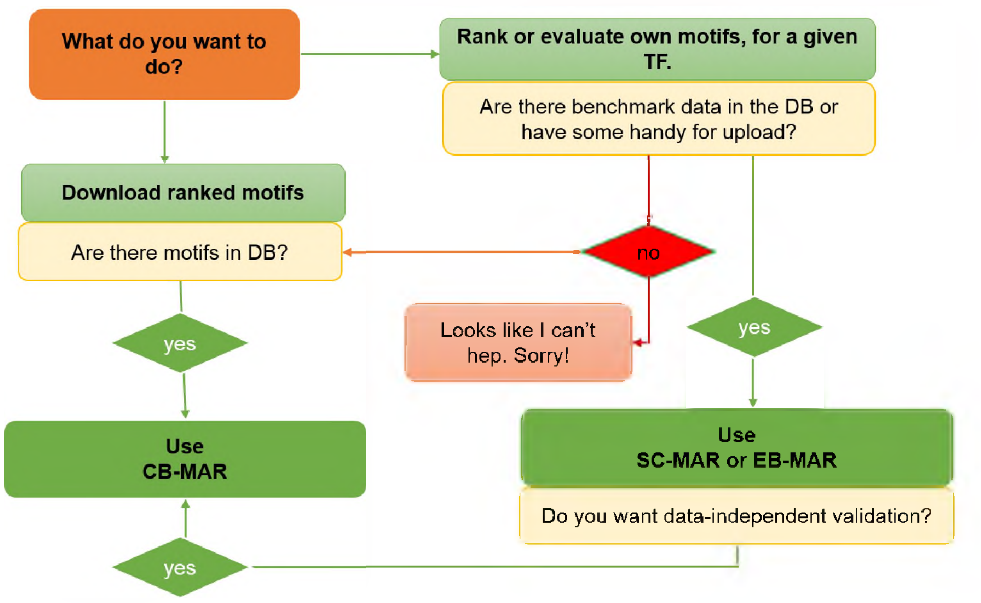

What do you want to do?
Rank or evaluate own motifs, for a given TF.
Are there benchmark data in the DB or have some handy for upload?
Download ranked motifs
Are there motifs in DB?
no
yes
Looks like I can't hep. Sorry!
yes
Use CB-MAR
Use SC-MAR or EB-MAR
Do you want data-independent validation?
yes

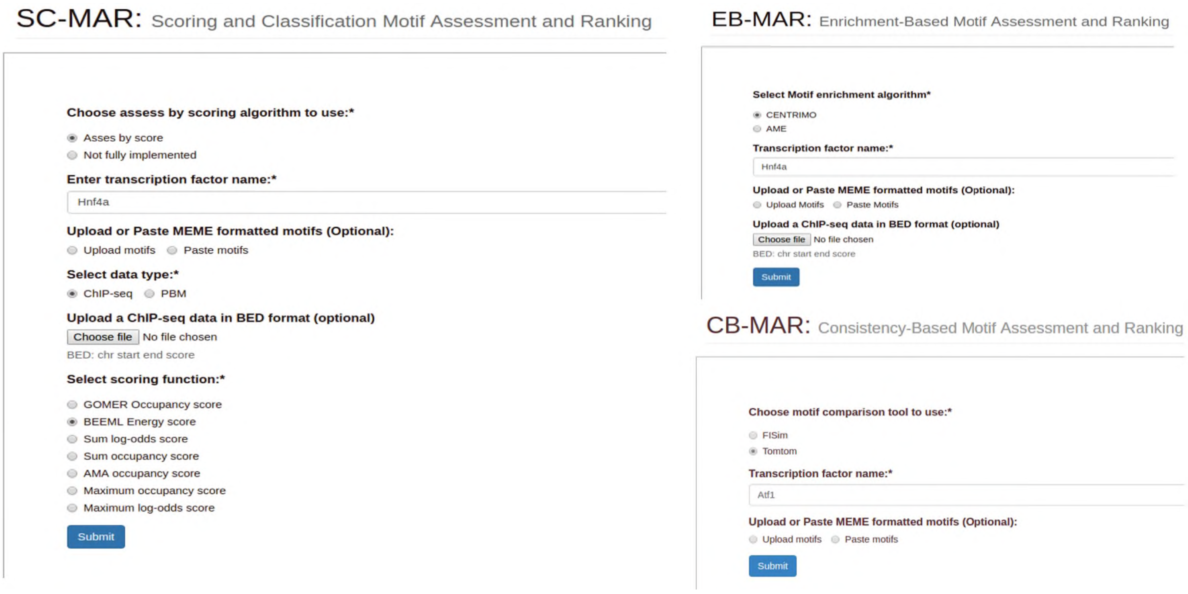
Fig. 4.7 Decision flow diagram on tools to use in MARS. Provides a quick guide on the appropriate tool for a given task. DB refers to MARS database while SC-MAR, CB-MAR, EB-MAR are defined in Figure 4.6 caption.

ChIP-seq data in BED format, only for version hg19 of the human genome. A guided search function for the available motif and benchmark data assists the users when choosing the tools to use, based on the decision flow diagram in Figure 4.7. The MARS documentation[12] also provides a guide on the accepted data formats and best practices when using the various tools in the web-server.

4.6.2.1 The motif assessment process: Form submission

MARS is designed for the end users of the generated motif models, providing them with a platform to compare and test the quality of the available motifs or those they generate from motif discovery. Given a TF name, complete evaluation and ranking of motifs in the database can be performed using available benchmark data and default parameters – chosen based on results from Chapter 3. The submission forms are as shown in Figure 4.8. Alternatively, the user can upload their own motifs in MEME format, which are ranked against those in the MARS database, or on their own when none is available in the database. For SC-MAR and EB-MAR, the user can also upload test sequences in BED format from which MARS retrieves the negative set from the hg19 genome, as described in Section 3.4.1.

[12] www.bioinf.ict.ru.ac.za/documentation



SC-MAR: Scoring and Classification Motif Assessment and Ranking

Choose assess by scoring algorithm to use:*

● Asses by score
○ Not fully implemented

Enter transcription factor name:*

Hnf4a

Upload or Paste MEME formatted motifs (Optional):

○ Upload motifs ○ Paste motifs

Select data type:*

● ChIP-seq ○ PBM

Upload a ChIP-seq data in BED format (optional)

Choose file No file chosen

BED: chr start end score

Select scoring function:*

○ GOMER Occupancy score
● BEEML Energy score
○ Sum log-odds score
○ Sum occupancy score
○ AMA occupancy score
○ Maximum occupancy score
○ Maximum log-odds score

Submit

EB-MAR: Enrichment-Based Motif Assessment and Ranking

Select Motif enrichment algorithm*

● CENTRIMO
○ AME

Transcription factor name:*

Hnf4a

Upload or Paste MEME formatted motifs (Optional):

○ Upload Motifs ○ Paste Motifs

Upload a ChIP-seq data in BED format (optional)

Choose file No file chosen

BED: chr start end score

Submit

CB-MAR: Consistency-Based Motif Assessment and Ranking

Choose motif comparison tool to use:*

○ FISim
● Tomtom

Transcription factor name:*

Atf1

Upload or Paste MEME formatted motifs (Optional):

○ Upload motifs ○ Paste motifs

Submit

Fig. 4.8 Submission forms for each of the three MARSTools. Each tool requires the user to select the algorithm (first choice), enter TF name and upload the test motifs. For SC-MAR and EB-MAR, the user has an option of uploading experimental data in BED format. SC-MAR allows the user to evaluate motifs in ChIP-seq or PBM data and choose the scoring function. For each tool, the default options are designed to provide the best results.

AJAX calls to the server submit the completed data in the forms. Form validation is carried out to check for data availability, completeness and correct MEME or BED formats, returning appropriate error messages to the user for correction.

4.6.2.2 The motif assessment process: Results visualization

On submission, the data is queued and executed by the specific algorithm. The results, returned via AJAX calls, are in three formats: a clustergram or clustered heatmap, bar plots and tables. See the captions in Figures 4.11, 4.12, and 4.9, respectively for details.

4.6.3 Other uses of MARS

The main aim for developing MARS is to provide researchers with a platform for motif assessment and ranking. In the course of testing and optimisation, we found that MARS can also be used to answer the following research questions:

- Perform discriminatory motif enrichment in multiple cell lines. As an addition, especially using the visualisation platform provided (Figure 4.11), a user can obtain information on

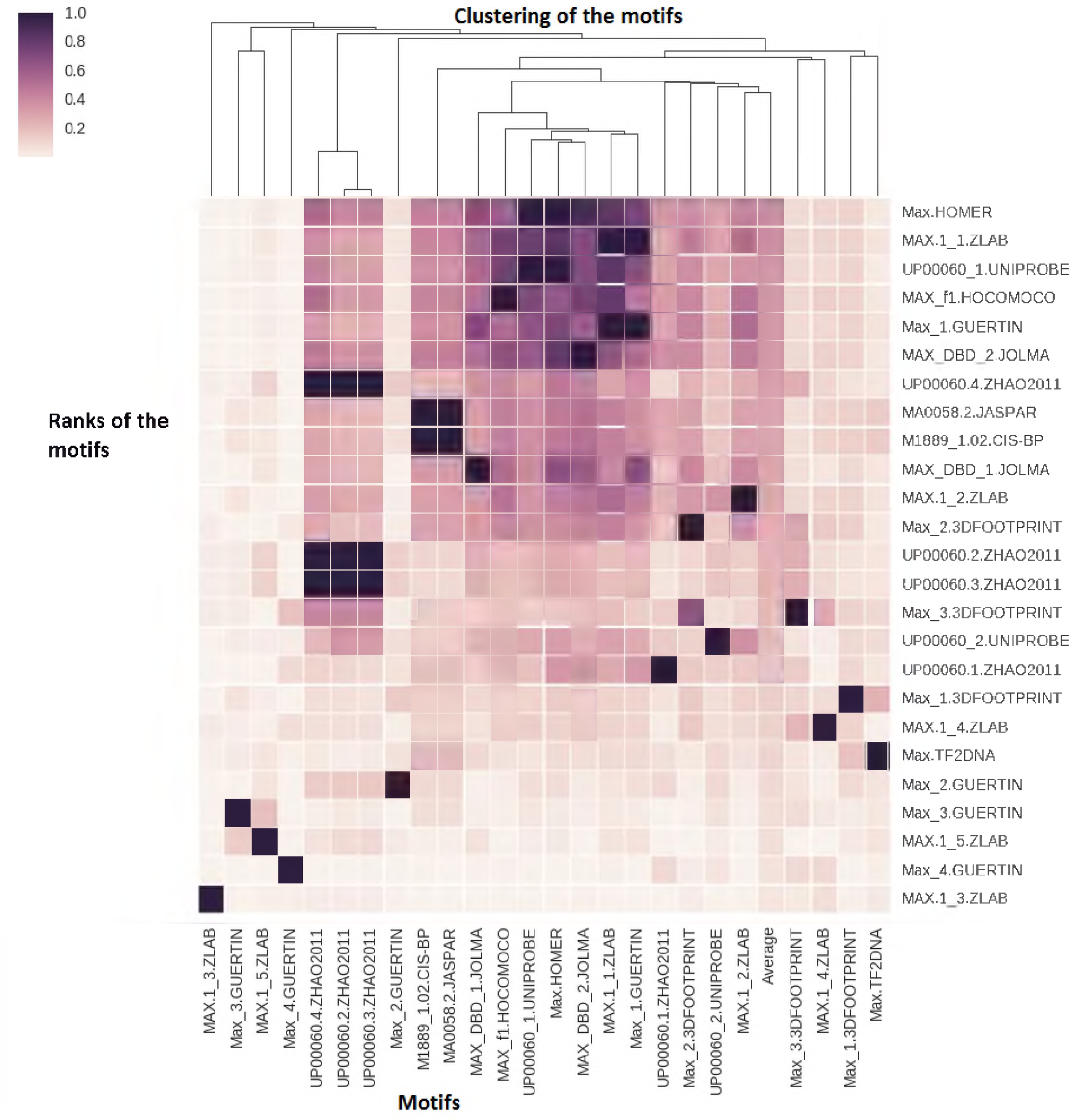
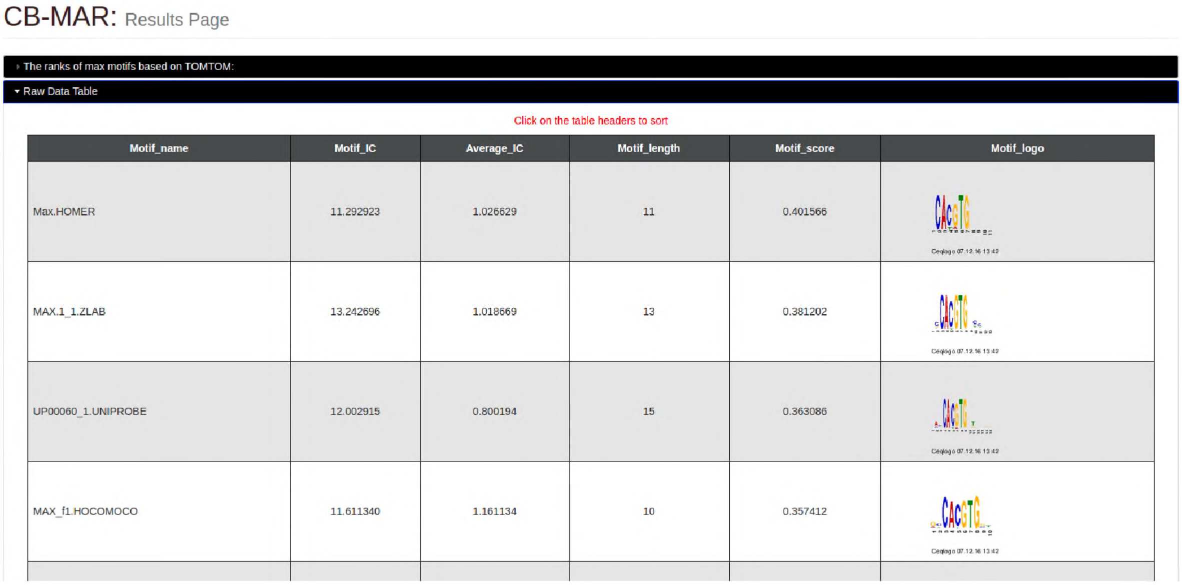


Fig. 4.9 CB-MAR results: motif ranking and clustering using Tomtom. The figure displays the ranks of the motifs in the vertical axis (based on average similarity score) and shows how they cluster in the horizontal axis (based on Pairwise similarity score) – see Section 4.4.3 for details. This provides information on motif's similarity as well as help detect duplicates.



CB-MAR: Results Page

▸ The ranks of max motifs based on TOMTOM:

▾ Raw Data Table

Click on the table headers to sort

Motif_name	Motif_IC	Average_IC	Motif_length	Motif_score	Motif_logo
Max.HOMER	11.292923	1.026629	11	0.401566	Ceqlogo 07.12.16 13:42
MAX.1_1.ZLAB	13.242696	1.018669	13	0.381202	Ceqlogo 07.12.16 13:42
UP00060_1.UNIPROBE	12.002915	0.800194	15	0.363086	Ceqlogo 07.12.16 13:42
MAX_f1.HOCOMOCO	11.611340	1.161134	10	0.357412	Ceqlogo 07.12.16 13:42

Fig. 4.10 CB-MAR results: raw data table. CB-MAR results are also displayed as a sortable table that displays: Information content (full length and Average), the length of the motif, the average similarity score (Motif score) and the motif logo. This information assist the user in gauging the quality of the motif and understanding why a motif ranks as it does.

motif enrichment in various cell lines or sequences. This, however, should only be used in research exploration stage since MARS does not provide statistical significance of the enrichment.

- Download ranked motifs for a given TF from a wide variety of databases hosted in MARS. No database currently offers this functionality and users have to scout for motifs for various databases and publications and deal with format conversion.
- Using the search functionality, a user can obtain summary information like IC, logos, length and source of the motifs hosted in the MARS database.

4.7 Chapter conclusions

We have developed MARS, a web server hosting a suite of tools for comparative analysis. It offers choice and flexibility to users through the ability to upload additional test data, and motif; also, we do not impose an assessment approach on the users. A major contribution to motif evaluation in this study is the data-independent consistency-based approach (CB-MAR), which offers an excellent alternative in the absence of benchmark sequence data. We believe that

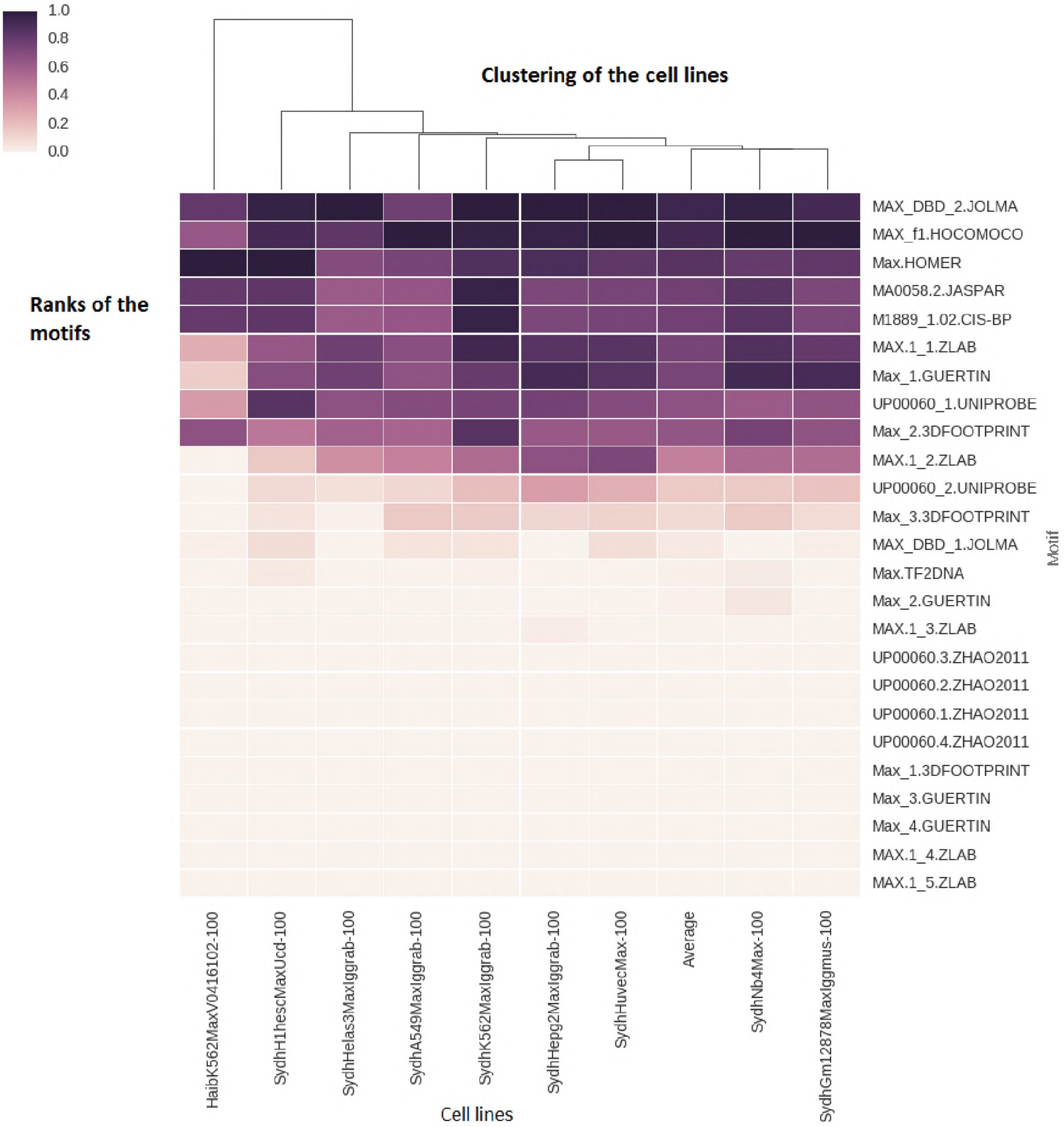
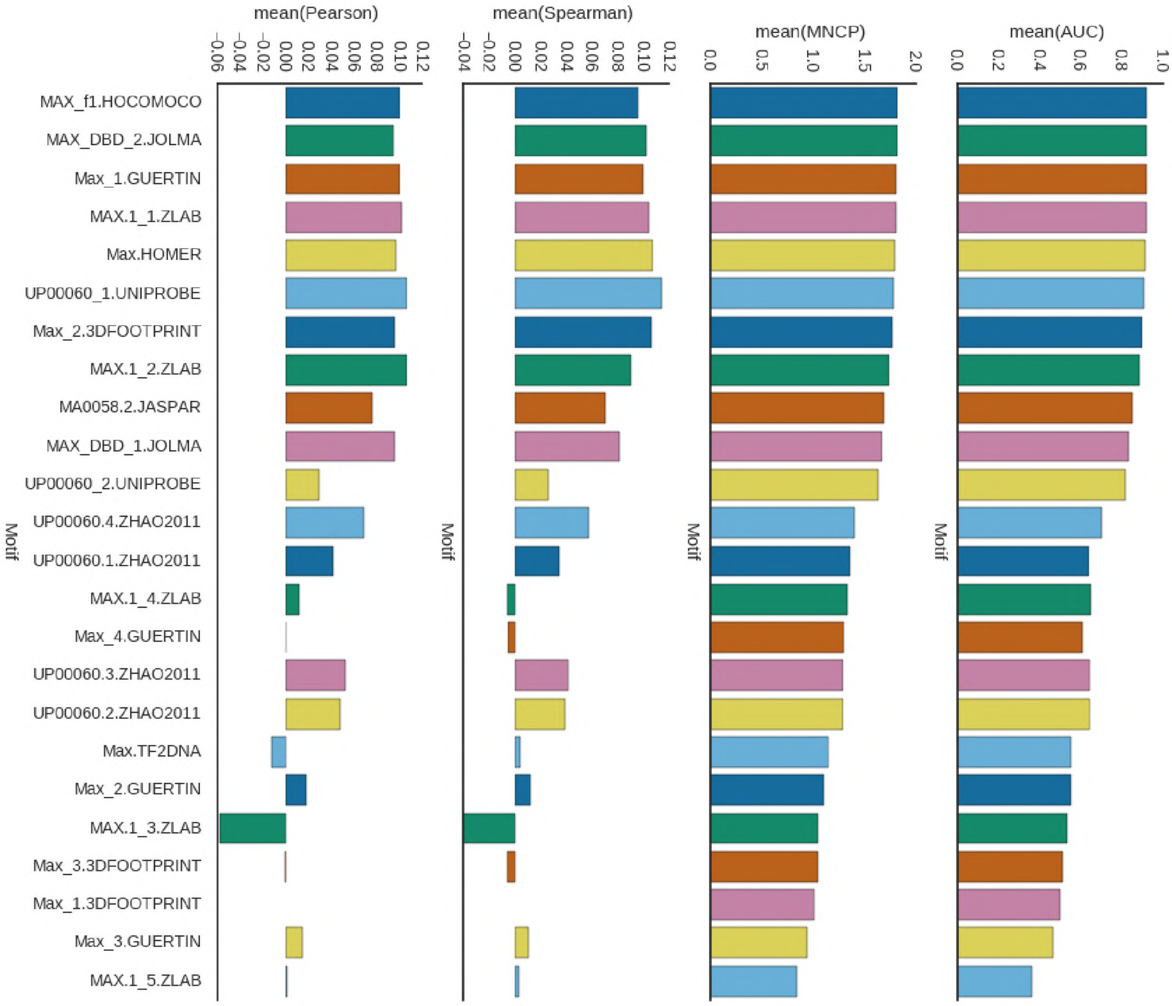
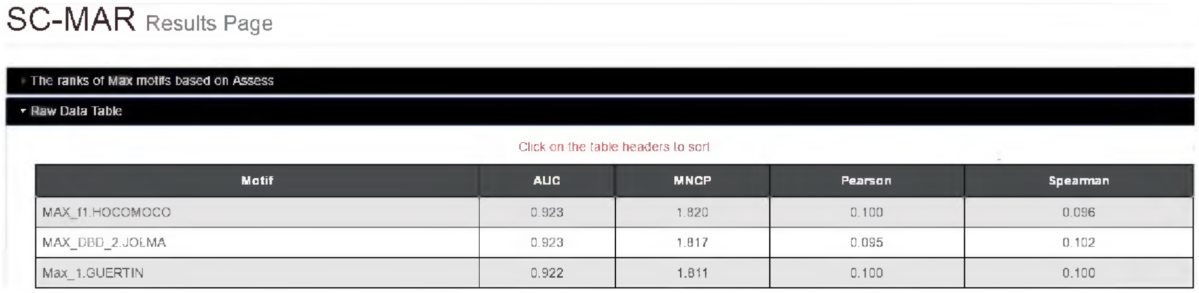


Fig. 4.11 EB-MAR results: ranks of Max motifs using CentriMo. The figure combines a sorted heat map and a dendrogram to display ranks of the motifs and the clustering of the benchmark data used, respectively. This provides information on how the motifs are enriched in different cell lines.




SC-MAR Results Page

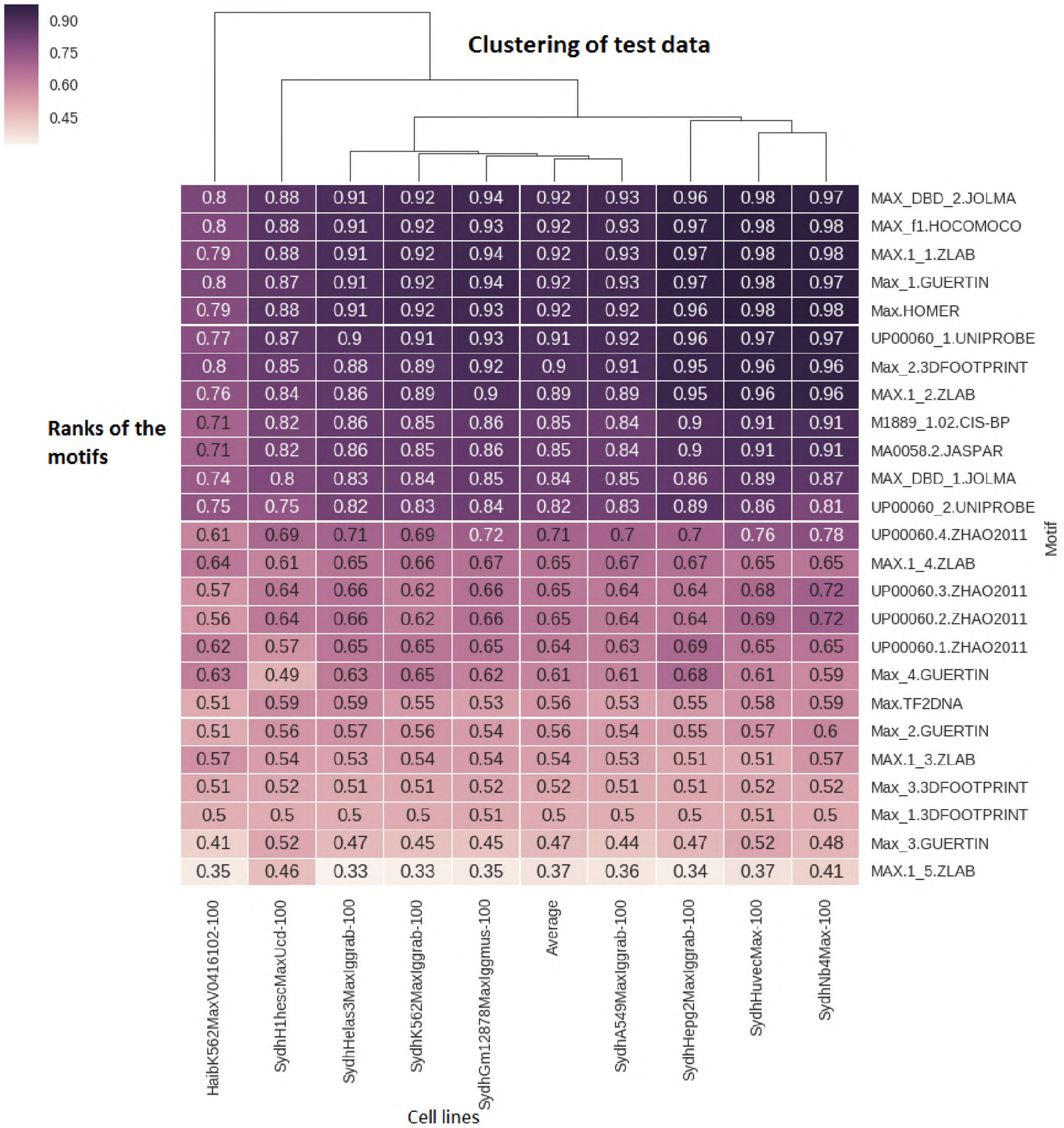
▸ The ranks of Max motifs based on Assess

▾ Raw Data Table

Click on the table headers to sort

Motif	AUC	MNCP	Pearson	Spearman
MAX_f1.HOCOMOCO	0.923	1.820	0.100	0.096
MAX_DBD_2.JOLMA	0.923	1.817	0.095	0.102
Max_1.GUERTIN	0.922	1.811	0.100	0.100

Fig. 4.12 SC-MAR results: motif ranks in SC-MAR sorted by the AUC scores. The figure provides information on how the motifs rank when using the various statistics; the raw scores are displayed as a sortable table (below bar plot), which allows the user to rank motifs based on the different statistics. The user can then check the information presented in Figure 4.13 to determine how the motifs rank in different test data (cell lines in this case).



Motif	HaibK562MaxV0416102-100	SydhH1hescMaxUcd-100	SydhHelas3MaxIggrab-100	SydhK562MaxIggrab-100	SydhGm12878MaxIggmus-100	Average	SydhA549MaxIggrab-100	SydhHepg2MaxIggrab-100	SydhHuvecMax-100	SydhNb4Max-100
MAX_DBD_2.JOLMA	0.8	0.88	0.91	0.92	0.94	0.92	0.93	0.96	0.98	0.97
MAX_f1.HOCOMOCO	0.8	0.88	0.91	0.92	0.93	0.92	0.93	0.97	0.98	0.98
MAX.1_1.ZLAB	0.79	0.88	0.91	0.92	0.94	0.92	0.93	0.97	0.98	0.98
Max_1.GUERTIN	0.8	0.87	0.91	0.92	0.94	0.92	0.93	0.97	0.98	0.97
Max.HOMER	0.79	0.88	0.91	0.92	0.93	0.92	0.92	0.96	0.98	0.98
UP00060_1.UNIPROBE	0.77	0.87	0.9	0.91	0.93	0.91	0.92	0.96	0.97	0.97
Max_2.3DFOOTPRINT	0.8	0.85	0.88	0.89	0.92	0.9	0.91	0.95	0.96	0.96
MAX.1_2.ZLAB	0.76	0.84	0.86	0.89	0.9	0.89	0.89	0.95	0.96	0.96
M1889_1.02.CIS-BP	0.71	0.82	0.86	0.85	0.86	0.85	0.84	0.9	0.91	0.91
MA0058.2.JASPAR	0.71	0.82	0.86	0.85	0.86	0.85	0.84	0.9	0.91	0.91
MAX_DBD_1.JOLMA	0.74	0.8	0.83	0.84	0.85	0.84	0.85	0.86	0.89	0.87
UP00060_2.UNIPROBE	0.75	0.75	0.82	0.83	0.84	0.82	0.83	0.89	0.86	0.81
UP00060.4.ZHAO2011	0.61	0.69	0.71	0.69	0.72	0.71	0.7	0.7	0.76	0.78
MAX.1_4.ZLAB	0.64	0.61	0.65	0.66	0.67	0.65	0.67	0.67	0.65	0.65
UP00060.3.ZHAO2011	0.57	0.64	0.66	0.62	0.66	0.65	0.64	0.64	0.68	0.72
UP00060.2.ZHAO2011	0.56	0.64	0.66	0.62	0.66	0.65	0.64	0.64	0.69	0.72
UP00060.1.ZHAO2011	0.62	0.57	0.65	0.65	0.65	0.64	0.63	0.69	0.65	0.65
Max_4.GUERTIN	0.63	0.49	0.63	0.65	0.62	0.61	0.61	0.68	0.61	0.59
Max.TF2DNA	0.51	0.59	0.59	0.55	0.53	0.56	0.53	0.55	0.58	0.59
Max_2.GUERTIN	0.51	0.56	0.57	0.56	0.54	0.56	0.54	0.55	0.57	0.6
MAX.1_3.ZLAB	0.57	0.54	0.53	0.54	0.54	0.54	0.53	0.51	0.51	0.57
Max_3.3DFOOTPRINT	0.51	0.52	0.51	0.51	0.52	0.52	0.51	0.51	0.52	0.52
Max_1.3DFOOTPRINT	0.5	0.5	0.5	0.5	0.51	0.5	0.5	0.5	0.51	0.5
Max_3.GUERTIN	0.41	0.52	0.47	0.45	0.45	0.47	0.44	0.47	0.52	0.48
MAX.1_5.ZLAB	0.35	0.46	0.33	0.33	0.35	0.37	0.36	0.34	0.37	0.41

Fig. 4.13 SC-MAR results: ranks of Max motifs in cell lines used. Provides information on how the motifs perform in each of the test data used (cell lines in this case). This figure displays how Max TF motifs rank based on AUC statistic on the vertical axis and how the cell lines cluster based on AUC scores.

this web server and the algorithms implemented will help reduce motif redundancy and the continued dependence of ‘reference motifs’ of unknown quality, due to lack of evaluation. Our suite also acts as a hub for motifs, currently scattered in various databases and publications, though this was not the main purpose. To ensure continued relevance, we hope in future to expand MARS to:

1. Allow motif uploads in a variety of formats; this will ensure users do not have to convert their motifs to MEME PWM format.

2. Expand the evaluation to other types of models for representing TF binding specificity – *k*-mer, Slim and TFFM – in addition to PWMs. Functionality to evaluate *k*-mer-based models is currently an experimental feature in standalone MARSTools.

3. Increase the variety of benchmark data accepted, especially as newer experimental techniques are introduced.

4. Expand CB-MAR to utilise other motif comparison algorithms.

5. Improve speed, flexibility and efficiency of MARS.

6. Allow users to contribute benchmark and motif data to the MARS for further comparison.

In this chapter, we have introduced a web server hosting a suite of tools. In Chapter 5, we evaluate MARS following Aniba’s criteria as adapted in Section 4.1.

Chapter 5

Evaluation of MARS

> *"Who watches the watchmen"*
>
> –Alan Moore *"Watchmen"* and Iantorno et al. [72]

Motif evaluation studies in effect watched motif discovery algorithms to advise the end users on algorithm's performance. However, they end up being in need to be watched themselves, hence the question "who watches the watchmen?" To help "watch" or rather keep the watchmen in check, Aniba and colleagues offered some characteristics of a good benchmark as scalability, relevance, evolution, independence, solvability, and accessibility. The number of motifs available for a single TF continues to increase. This increase offers variability by expanding the binding spectra captured, since TFs are known to bind to degenerate sites spread around the genome. However, this is also a challenge. Choosing a binding model is now a daunting task, given that we can already have up to 47 PWM models in our database (for Hnf4a, as an example) generated from a variety of data and algorithms. How can these models be ranked to obtain generalised or specialised models for a given task? It is harder to select a generalised model because of bias towards the data used during evaluation. A data-free approach of assessment is required to attain evaluation independence. In Chapter 4, we introduced MARS, a web server hosting PWM motif evaluation and ranking techniques based on sequence scoring and classification (SC-MAR), motif enrichment (EB-MAR), and a data-*independent* consistency-based motif assessment and ranking (CB-MAR), making these tools *accessible* to the end users. The web server is supported by a database of benchmark data and motifs against which a user can test and rank their motifs. MARS also allows the users to upload their benchmark data which, in combination with the modular design of the algorithms, ensures MARS can *scale* and *evolve* with new data and algorithms.

5.1 Chapter aim and objectives

The main goal of this chapter is to evaluate MARS: to demonstrate how MARS meets the *relevance*, *independence* and *solvability* criteria. Since no other tool is available that provides functionalities similar to MARS, independently evaluating MARS can be a challenge. Notwithstanding, we evaluate MARS by addressing the following objectives:

1. Evaluate MARS against Aniba's criteria by testing for the usefulness of these tools and how the motif ranks by various tools correlate

2. Compare the specific tools in SC-MAR (SCORE and gimme roc) and CB-MAR (Tomtom and FISim) to provide guidance on the defaults

3. Demonstrate the *relevance* of CB-MAR in motif discovery and how MARS facilitates systematic comparative analysis in motif evaluation

We do not evaluate design goals such as usability, since such an evaluation would be more suited to a software engineering project rather than a Bioinformatics thesis.

5.2 Methods

This section describes the approaches and data used to evaluate MARS. Detailed reproducible IPython notebook for the analysis carried out in this chapter is available at from GitHub[1].

5.2.1 Comparison of MARS tools

How well tools implementing different algorithms and data reproduce each other can act as a crude evaluation. For our evaluation, we select a total of 127 TFs with a TF-ID, have more than ten motifs and have benchmark data sourced from either PBM (60 TFs) or ChIP-seq (83 TFs) to rank all the available motifs for each TF using the different tools available. For simplicity of analysis and comparison, we use SCORE-Energy and AUC statistics throughout these evaluations – a combination we found in Chapter 3 to produce consistent rankings and is least biased by motif length and information content (IC).

[1] https://github.com/kipkurui/MARS_Evaluation/blob/master/MARS_Evaluation/Complete_Analysis-MARS_Evaluation.ipynb

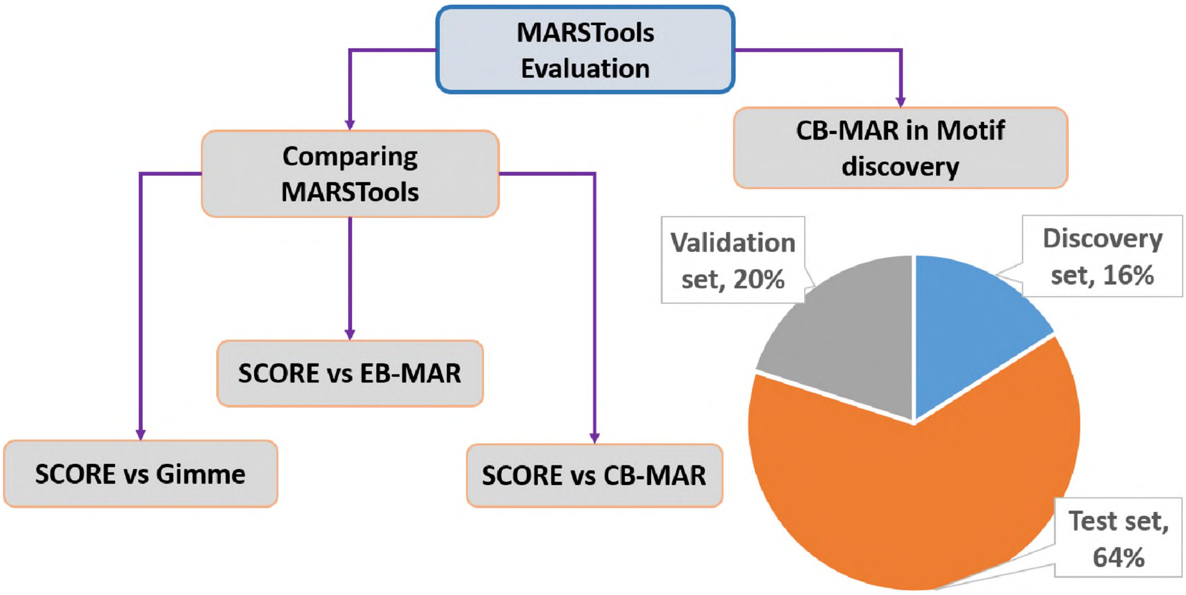


Fig. 5.1 MARSTools Evaluation flow diagram. Demonstrates how the MARSTools are evaluated by comparing against each other and applying CB-MAR in motif discovery. SCORE: Scoring and Classification Motif Assessment and Ranking using Energy scoring; CB-MAR: Consistency-Based Motif Assessment and Ranking; EB-MAR: Enrichment-Based Motif Assessment and Ranking. We set aside 80% of data fro discovery but GimmeMotifs use 20% of that for discovery and the rest for evaluation (test).

5.2.2 CB-MAR in *ab initio* motif discovery

We apply CB-MAR to choose the top motifs in *ab initio* discovery, a task commonly accomplished using held out data. We take advantage of an ensemble motif finding tool, GimmeMotifs [189], which performs *ab initio* motif discovery from ChIP-seq data using nine algorithms. We chose a total of 110 TFs, which had ENCODE ChIP-seq data and a corresponding TF-class ID. For all the data available from different cell lines for a given TF, we extracted the top 500 peaks, widened them to 100 bp around the peak centre, then combined and shuffled the data. Next, we randomly extracted 80% of the sequences for *ab initio* motif discovery and kept the rest for validation. For CTCF, we randomly sampled 5000 sequences out of the combined and shuffled data as it had a large data set, which would take too long to run. After motif discovery, we used the CB-MAR (using Tomtom) approach in combination with motif clustering (using *gimme cluster* from GimmeMotifs at 95% similarity and *kcmeans.py* clustering from FISim [50]) to rank and narrow down the motif predictions to the best three non-redundant motifs. Finally, we compared the top motifs identified by CB-MAR with those from GimmeMotifs (it uses 20% of the provided data for discovery and the rest for validation) on the validation set using the *gimme roc* command. This process can be summarised as follows:

1. Extract top 500 peaks from each cell line and merge them to a single Pandas DataFrame
2. Shuffle the DataFrame using Numpy random permutations
3. Randomly extract 80% of the peaks for motif discovery and reserve the rest for validation
4. Run motif GimmeMotifs with MDmodule, MEME, MotifSampler, trawler, Improbizer, BioProspector, Posmo, ChIPMunk, AMD, HMS and Homer algorithms– *gimme motifs* uses 20% of the input sequences for discovery and the rest for evaluation
5. Keep the intermediate raw motif from GimmeMotifs then use CB-MAR to rank the motifs
6. Use *gimme cluster* and FISim *kcmeans.py* clustering and extract top three non-redundant motifs
7. Evaluate the top motifs by *gimme roc* and CB-MAR on the held out validation data

NB: kcmeans.py computes pairwise similarity matrix using FISim, eliminates negative eigenvalues to produce a kernel, computes a distance matrix and finally clusters the motifs using c-means [50].

5.3 Results

To demonstrate the relevance and usability of the MARSTools, we employ them in a systematic analysis of motif quality as well as in motif discovery. What follows are the results from optimisation of the various algorithms and options, comparison to similar tools and application of MARS in a motif discovery case study.

5.3.1 Scoring and classification motif assessment and ranking (SC-MAR)

SC-MAR comprises two tools: SCORE and *gimme roc*. This section compares the motif ranks based on the two techniques and recaps of our review of SCORE performed in Chapter 3.

Effect of the scoring function, motif IC and length: For the SCORE approach, we performed thorough comparison and testing in Chapter 3: how the scoring functions and motif characteristics affect ranks of motifs an evaluation. In summary, we found that the scoring function used influences motif ranking in a TF-specific manner and that motif IC is not a predictor of motif quality. Although most of the scoring functions do not differ much in performance, energy function (hereafter SCORE-Energy) generates the most 'biologically relevant' rankings. Therefore, we use SCORE-Energy for all the comparisons in this chapter, unless otherwise stated.

SCORE-Energy reproduces *gimme roc* rankings: GimmeMotifs [189], an ensemble motif discovery pipeline for ChIP-seq data, also includes *gimme roc* for motif quality analysis and ranking. We use this to benchmark our approach on ChIP-seq data, and found that *gimme roc* produces motif rankings that are significantly correlated with the SCORE-Energy ranks (R=0.999 Pearson's, p-value=1.9×10^{-105}) and (R=0.995 Spearman's, correlation p-value=1.7×10^{-108}) – Figure 5.2. Also, no significant difference between the two sets of scores (p-value=0.825, Wilcoxon rank-sum test) is observed showing that the scoring function used by GimmeMotifs tools is similar to SCORE-Energy. Furthermore, the reproducibility of results validates our implementation of energy scoring function.

5.3.2 Comparison of EB-MAR and SCORE

For most TFs, SCORE and EB-MAR motif ranks agree (0.829, Median correlation) as shown in Figure 5.3A. However, for some TFs, these scores do not agree, like in SP4, where the scores have a negative correlation, while some have a correlation close to zero. This poor performance may be attributed to the lack of statistical enrichment of the motifs in the sequences; compare the motif enrichment plots of Mafk and Sp4 in Figure 5.3B and C, respectively. Mafk and Sp4 plots are representative of the binding behaviours – see distribution plots in Figure 5.4 for other typical examples. Sequence-specific TFs (direct binding) are easy to evaluate and rank, but indirect binding TFs can not be evaluated reproducibly by two techniques – sequence scoring or enrichment in this case. This clearly shows one of the additional benefits of using statistical enrichment as a measure of motif quality: ability to infer TF binding behaviour.

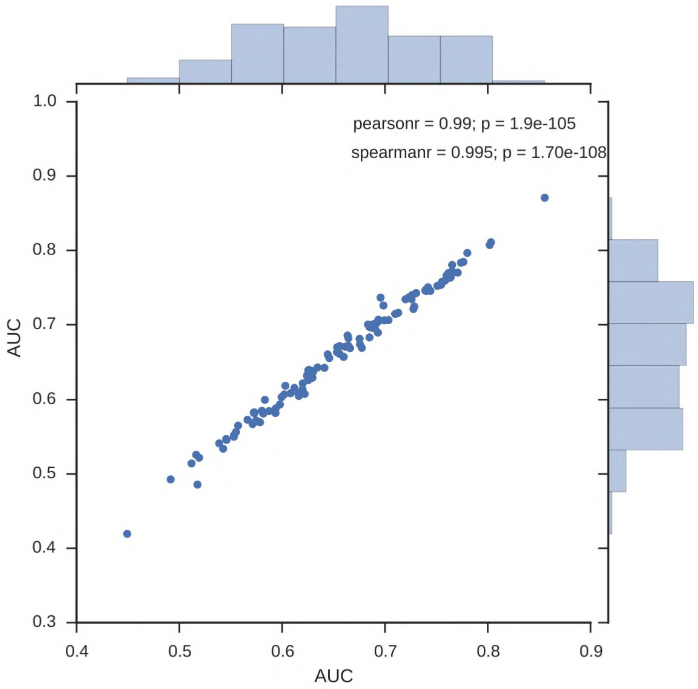
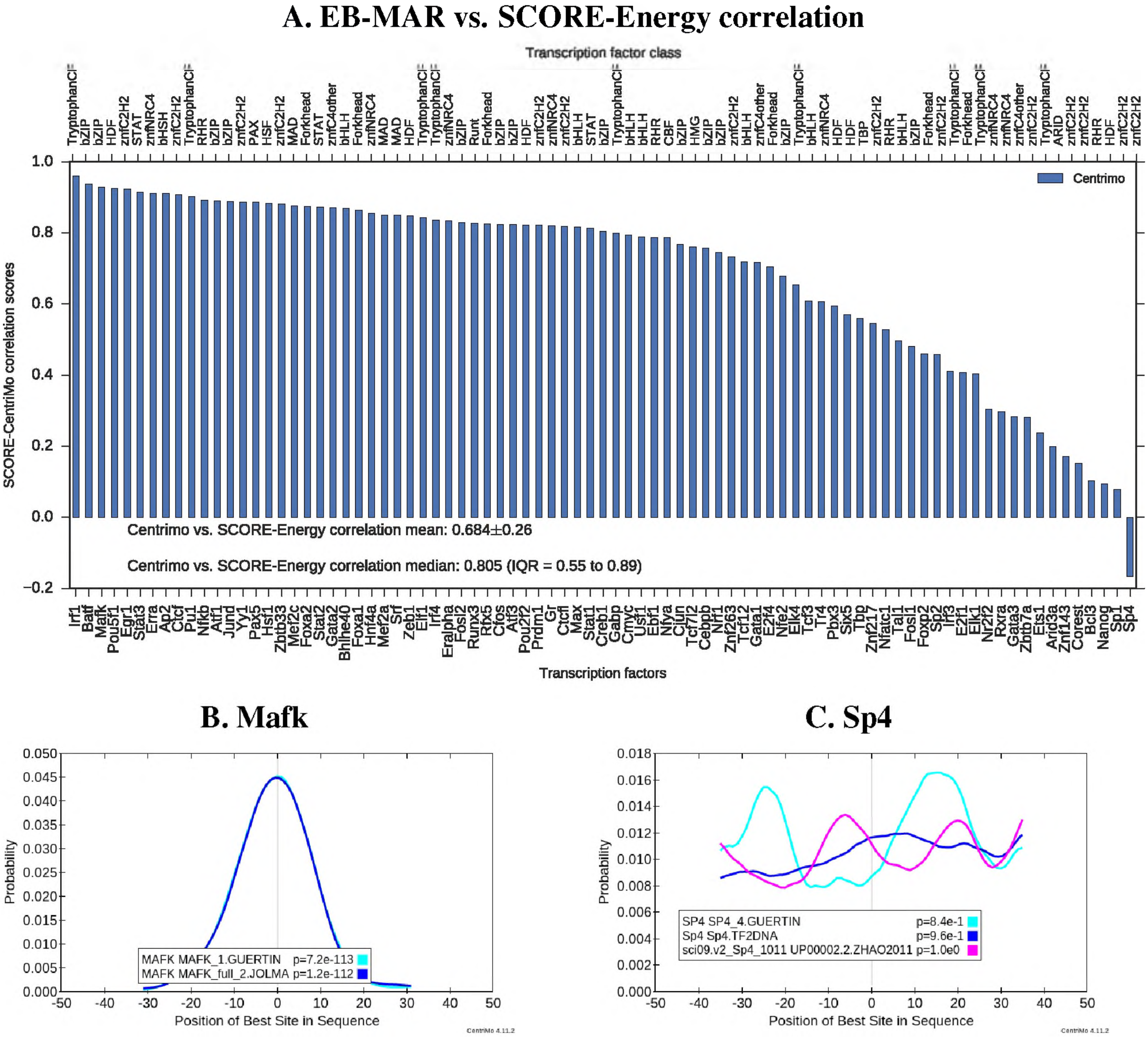


Fig. 5.2 Joint scatter plot and histogram for *gimme roc* and SCORE-Energy scoring. Shows the correlation of AUC scores in the two approaches are in agreement and the data are normally distributed.

5.3.3 Comparison of CB-MAR and SCORE

Tomtom comparison produces 'biologically relevant' rankings: For consistency-based motif ranking (CB-MAR), we decide on the best motif comparison algorithms that generate biologically relevant rankings – as defined by how well the motif ranks reproduce those based on *in vivo* data – by correlating with ranks based on energy scoring. From Figure 5.4A, we observe that the scores and ranks based on Tomtom (median=0.88; Interquartile range (IQR)=0.63-0.93) can better reproduce AUC scores based on energy scoring compared with FISim (median=0.78; IQR=0.52-0.86). We use the median to summarise the performance of the two techniques since the correlation scores were skewed (Figure 5.5). The level of correlation between CB-MAR and SCORE-Energy AUC rankings also seem to predict the binding behaviour of the TFs. For Tomtom, we find highly correlated motifs also have centrally enriched distributions in CentriMo (Figure 5.4 C) while less or negatively correlated TFs have broad distributions (Figure 5.4 D and E), a known predictor of indirect or cooperative binding [12]. The most common poorly correlated TF family, znfC2H2, is also known to bind in a sequence-independent manner [79], revealing this to be a generalised observation.


A. EB-MAR vs. SCORE-Energy correlation

B. Mafk

C. Sp4

Fig. 5.3 EB-MAR and SCORE correlation. A: Bar graph shows how rankings based on SCORE-Energy correlate with EB-MAR (CentriMo). The mean±STD and median with interquartile range statistics are annotated onto the figure. The bottom labels are the TF names while the top labels are the TF class. **B:** The ranks between the two techniques are in agreement when there is direct binding as illustrated in this CentriMo distribution. **C:** Ranks do not agree in those that are not statistically enriched in the test sequence. The motif names and the *p*-value of central enrichment of the ChIPed motifs are provided in the figure legends.

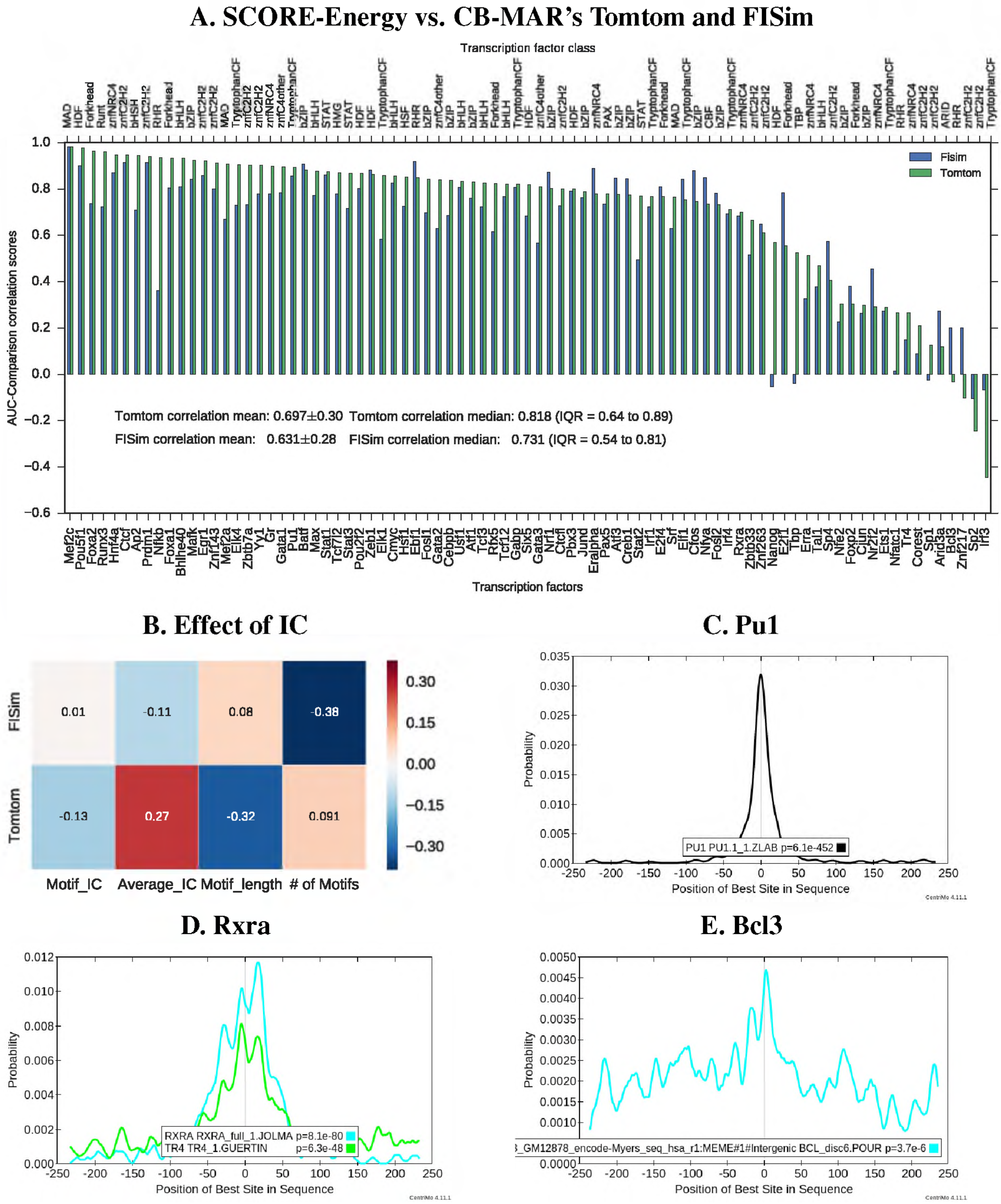


Fig. 5.4 Correlating CB-MAR (Tomtom and FISim) with energy AUC ranking in ChIP-seq data. **A:** Bar graph shows how rankings based on energy scoring correlate with consistency-based techniques. The bottom labels are the TF names while the top labels are the TF class. The mean±STD and median with interquartile range statistics are annotated onto the figure. **B:** The effect of motif information content (total and averaged by length), the number of motifs (size) and length on motif ranking. The CentriMo plots predict the possible direct binding behaviours (based on the sharp, centred distribution) of Pu1 motifs (**C**) in the distribution of best binding sites in ChIP-seq peaks, and indirect or cooperative binding of Tr4 (**D**) and Bcl3 (**E**) motifs. The motif names and the *p*-value of central enrichment of the ChIPed motifs is provided in the figure legends. For Rxra, the other centrally enriched motif (Tr4) could bind cooperatively with it.

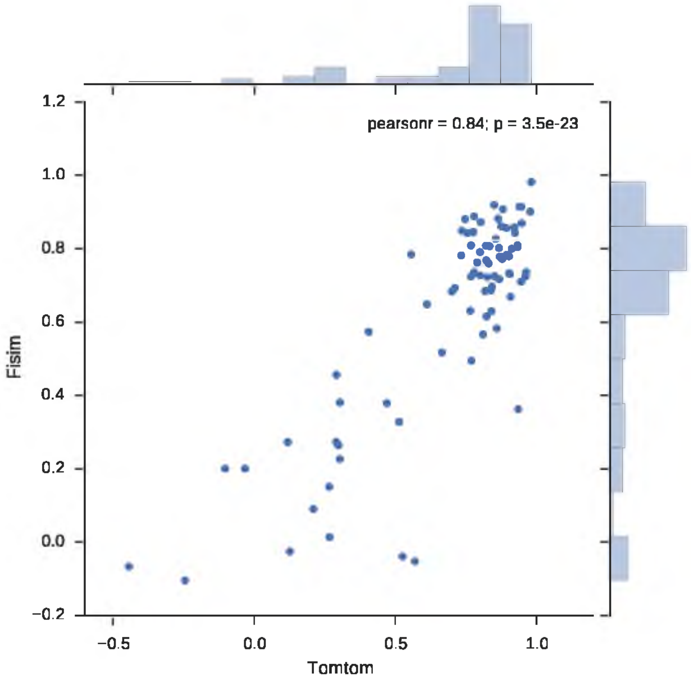


Fig. 5.5 Skewed joint histogram and scatter plot of Tomtom and FISim correlated with SCORE-Energy. Plot shows the skewed distribution of motif ranks based on Tomtom (x-axis) and FISim (y-axis) correlated with motif ranks based on SCORE-Energy.

Effect of benchmark data: When we determined the correlation between CB-MAR and energy scoring on PBM data, we do not find a clear performance difference between Tomtom and FISim average scores (Median of 0.7 and 0.72, respectively) – Figure 5.6. However, FISim has a higher mean (0.56 vs 0.52), hinting that FISim could better model *in vitro* while Tomtom models *in vivo* binding better. We also note that the TFs with low correlation between scores are known to bind indirectly or cooperatively. Specifically, the TFs from high mobility group (HMG) which have a negative correlation, are known to bind both directly and cooperatively, but they may have a different binding behaviour *in vivo* and *in vitro*; CB-MAR captures *in vivo* binding better than *in vitro* binding.

5.3.4 Consistency-Based Motif Assessment and Ranking (CB-MAR)

CB-MAR is implemented using two motif comparison algorithms: Tomtom and FISim. The ranks obtained by the two differ, an effect we believe can be explained by the inherent difference in the algorithms. What follows is an attempt to understand these differences by looking into how the algorithms deal with motif length, information content and size on motif ranking.

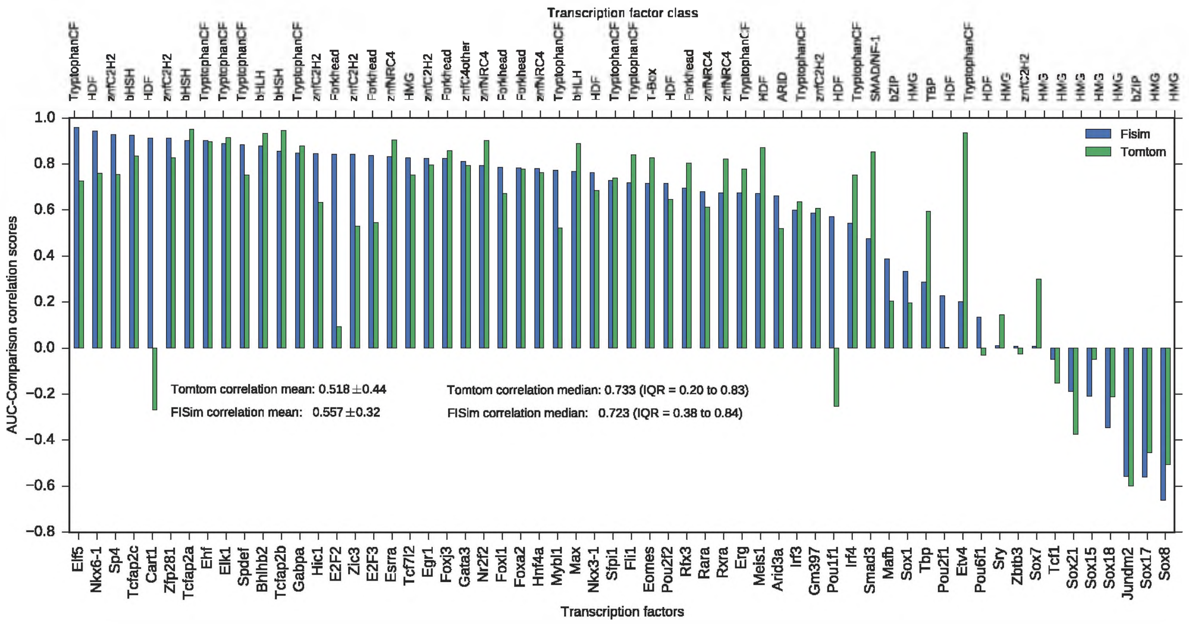


Fig. 5.6 Correlating CB-MAR (Tomtom and FISim) with energy AUC ranking in PBM data. How FISim and Tomtom correlate energy ranking in PBM data. The TF family class is given on the top axis and the TF names on the bottom. The mean±STD and median with interquartile range statistics are annotated onto the figure.

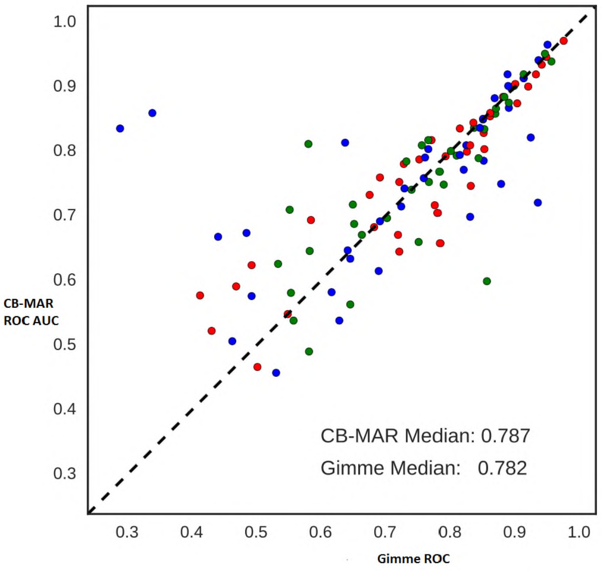


Fig. 5.7 CB-MAR in motif discovery. Demonstrates the use of CB-MAR as part of an ensemble motif discovery pipeline, where ranks based on both approaches performed similarly. Colours have no specific meaning.

Effect of motif length, size and information content: We find that Tomtom scores have a positive correlation with average IC normalised over motif length (R=0.24, favouring higher IC motifs) while FISim scores have a negative correlation (-0.11, penalise higher IC motifs). FISim is not influenced by length (R=0.078) while Tomtom penalises longer motifs (R=-0.25) since average IC is lower for longer motifs. Surprisingly, the number of motifs for the TF seem to negatively affect motif scores in FISim (R=-0.38) – due to higher IC as the number of motifs is increased – but has no effect for Tomtom (R=0.011), possibly explaining their difference in performance (Figure 5.4B).

CB-MAR generates relevant ranks in motif discovery: The first level of application of motif evaluation is in *ab initio* motif discovery, where an algorithm has to narrow down the identified motifs to a few that reflect the binding behaviour of a TF. The advent of ensemble motif discovery pipelines makes proper motif assessment and ranking even more desirable since the motifs from a variety of algorithms have to be ranked and unified. By purely using CB-MAR and motif clustering with *kcmeans.py* (we found *kcmeans.py* to be faster and better than *gimme cluster*), we were able to correctly identify better or similar motifs in a majority of the cases. Overall, the best motifs identified by GimmeMotifs and CB-MAR are similar (0.97; Wilcoxon Rank-sum test) with median AUC scores of 0.782 and 0.787 respectively (Figure 5.7). It is good practice for a few of the top motifs to be considered for further analysis [186]. Consequently, for the TFs whose performance differed by over 0.1, we tested the quality of the top three motifs: six using GimmeMotifs and 11 using CB-MAR (Figure 5.8). For CB-MAR, we find that choosing the second motif improves the quality in Maz, Yy1, and Atf1, while the third motif is always of a lower quality except for Irf3 (Figure 5.8A); for E2f4, second or third motifs have a lower performance and for Tcfap2 (Ap2) the top 3 motifs do not differ in quality significantly. For GimmeMotifs, we observed an improvement for 7 TFs but found no effect in E2f1, Ikzf1, Znf263 and Atf3 TFs (Figure 5.8B).

5.4 Discussion

A comparative approach to motif evaluation, using a variety of data and techniques, is necessary for to understand and capture the different binding behaviours and to make an informed decision on motif quality. For example, we can observe differences in binding behaviour of Zbtb3 TF *in vivo* and *in vitro* by correlating motif performance in PBM and ChIP-seq

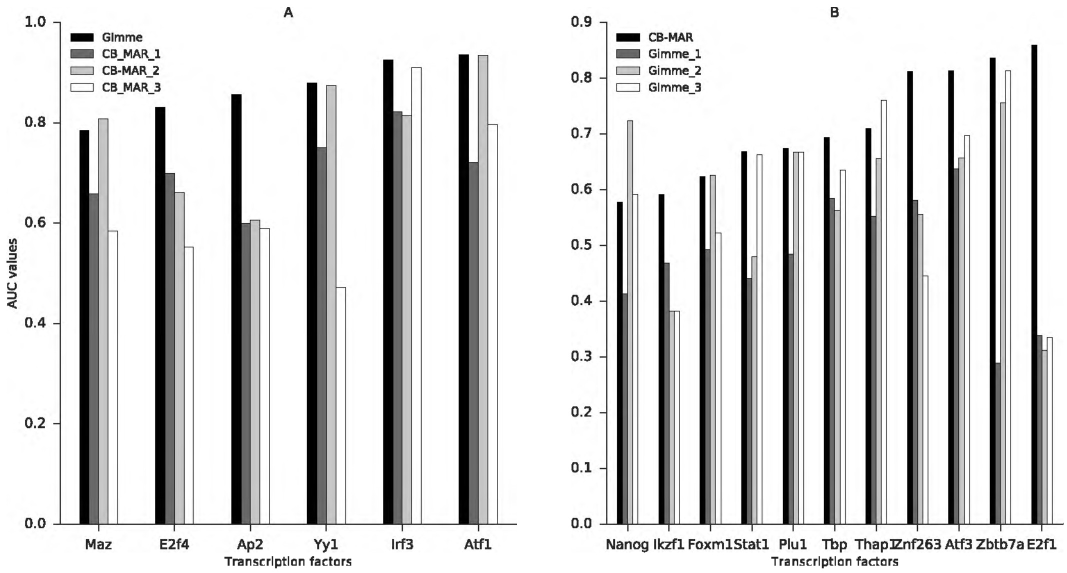


Fig. 5.8 How predicting top three motifs influences the performance of Gimme and CB-MAR. A: Compares the performance of top GimmeMotifs identified motif against the top three CB-MAR motifs. **B:** Compares the performance of top CB-MAR motif against the top three GimmeMotifs motifs. The ranks of the top three motifs as determined by GimmeMotifs and CB-MAR are represented by the suffixes (Gimme_1,2,3 and CB_MAR_1,2,3; respectively). The Y-axis is the mean AUC scores of the motifs on the validation sequences.

data (Figure 5.9); Zbtb3 is known to recognise unmethylated motifs *in vivo* and methylated ones *in vitro* [19]. Furthermore, we can discover that HMG TFs may also bind indirectly, cooperatively or a variation of binding behaviour, as captured in PBM data by a low or negative correlation between CB-MAR and SCORE-Energy derived ranks (Figure 5.6). Indeed HMG TFs, specifically the SOX-related factors, are known to bind cooperatively with partner TFs [83, 96]. They are believed to form complexes with partner proteins before recognising the binding site [83]. Therefore, we expect that predicted models based on PBM data would differ from how one of these TFs binds *in vivo*.

The data-independent approach, CB-MAR, produces biologically relevant motif ranking. The evaluation of CB-MAR reveals that it better reproduces ranks based on ChIP-seq (R=0.88) than PBM (R=0.73) data, showing that it captures motif binding behaviour *in vivo* better – see Figure 5.4A and 5.6. We further support this argument by using it to identify best models in motif discovery successfully; more details later in the section. CB-MAR reduces the 'reference motifs' bias: an approach in which users consider an algorithm successful if it can predict motifs similar to those in a 'reference database' at a given (usually arbitrary) similarity threshold. The current collections of ChIP-seq and PBM experimental data in our database can only facilitate

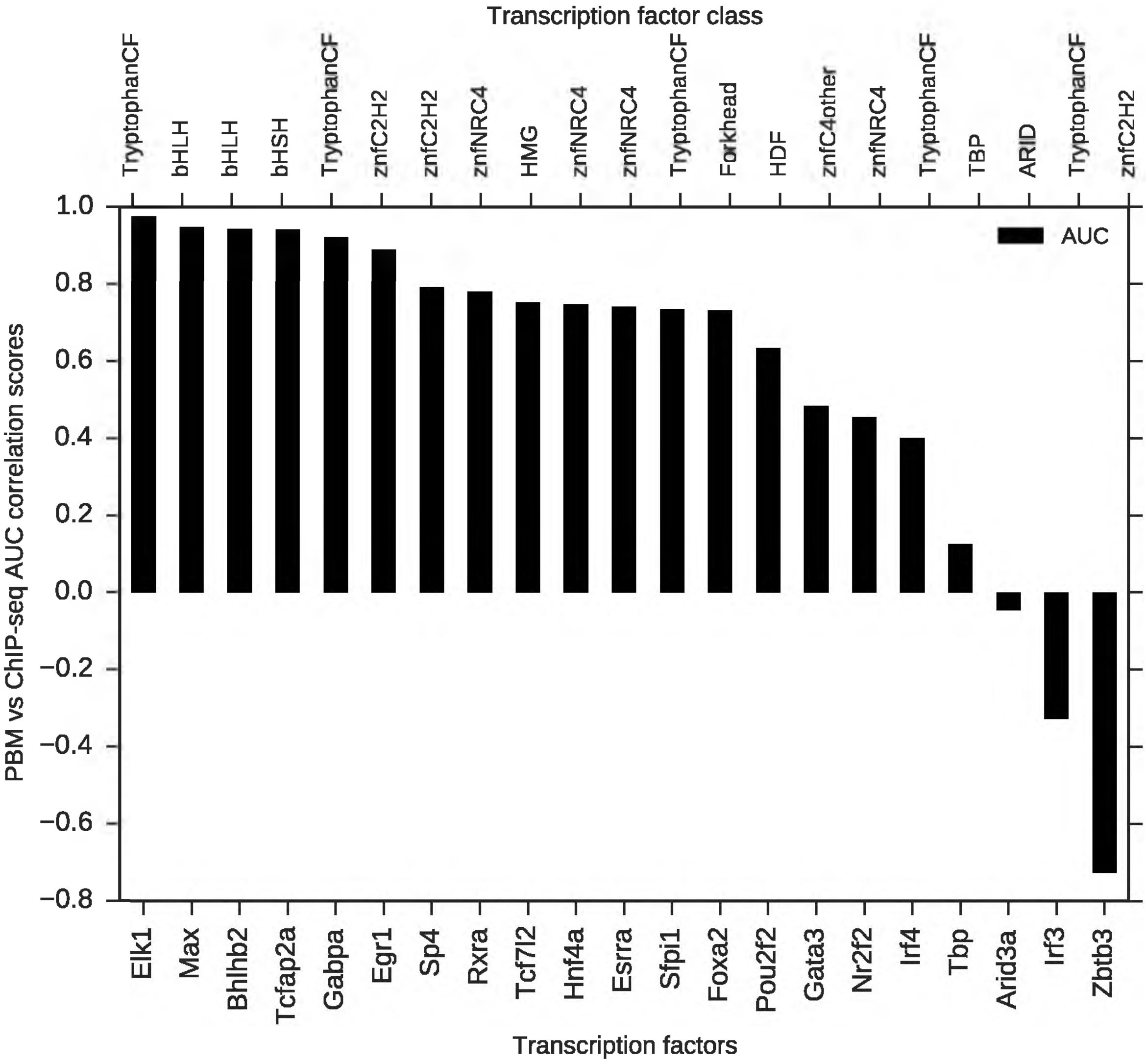


Fig. 5.9 Correlating SCORE-Energy ranking in ChIP-seq and PBM data. For the 21 transcription factors which had data in PBM and ChIP-seq, we tested how well the motif ranks *in vivo* and *in vitro* compare.

quality evaluation for less the 300 TFs out of 1352 that have motifs in our database. This demonstrates that CB-MAR is even more desirable when evaluating motifs for the majority of the TFs.

Between the two motif comparison algorithms tested for CB-MAR (Tomtom and FISim), we show that Tomtom better captures *in vivo* motif rankings in ChIP-seq data compared with FISim. FISim is designed to favour similarity of high information or conserved sites [50], showing that scoring high IC sites higher may not match biologically similar motifs. Additionally, low information flanking sites have been reported to increases binding specificity in some TFs [79, 108, 42].

CB-MAR is valuable in motif discovery. The first step after motif discovery is to filter and narrow down to significant motifs. Usually, a partition of the data is held out for testing, but with limited data, this may not be feasible. Besides, this is only available to the algorithm developers and to motifs generated using sequencing or microarray data (promoter sequences, ChIP-seq, PBM, etc.), and not to those from TF tertiary structures like 3DFootprint [35]. We have demonstrated that the top performing motifs can be identified using CB-MAR in combination with motif clustering to avoid duplicates. We do not average similarly clustered motifs, as sometimes done by GimmeMotifs. Rather, we choose the motif that consistently compares with the rest within the cluster and in the whole set. Motif averaging may produce a motif that does not fit biology or reflect TF binding behaviours as demonstrated by the cases where GimmeMotifs identified motifs that performed significantly worse than CB-MAR (Figure 5.8B).

In addition to providing the ranks of the motifs, the EB-MAR evaluation approach provides additional information on the relevance of the motifs. Are they statistically enriched in the sequences? Do they bind directly or indirectly? Motifs that are not statistically enriched in the sequences could point to indirect binding, mostly, though it could also mean the motifs are not enriched in the sequences (cell lines) used. In this case, comparing the performance of the motifs in different cell lines through the cluster heatmap and table provides the user with this information. MARS is, therefore, a useful resource for systematic comparative analysis.

5.5 Chapter conclusions

We evaluated MARS on two levels: comparing how the tools replicate each other and by applying CB-MAR in motif discovery. The usefulness of MARS in motif discovery demonstrates the *relevance*, *solvability* and *independence* of MARSTools in motif evaluation. We have also demonstrated the importance of systematic comparative analysis, which MARSTools facilitates, further supporting the *relevance* of MARSTools.

Chapter 6

Elucidating Transcription Factor Binding Occupancy using *in vivo* and *in vitro* Data

> *"Linked together as a team with one goal, we soon realised we were only as strong as our weakest link. However, did we condemn the weaker member? That would not serve any purpose. Instead, the stronger guys responded by carrying more weight than the weaker teammate. Encouragement was key in reaching the top of the stadium, standing as one."*
>
> –Jake Byrne, *First and Goal: What Football Taught Me About Never Giving Up*

So that is how it is when predicting TF binding. Teamwork in the sense that no data set is sufficient on its own to attain a strong predictive ability. A combined approach is necessary: the potential for each data set is not the same, but they are all necessary for a complete elucidation of the gene regulatory code.

The holy grail of determining transcription factor (TF) binding specificity is a method with the relative ease and low cost of *in vitro* methods [140], with the accuracy of *in vivo* methods. There have been many attempts at combining *in vivo* data with *in vitro* binding models, with limited success [147, 113, 132, 36, 217]. Ideally, it should be possible to use general *in vivo* information to improve a specific *in vitro* model, combining the benefits of *in vivo* (biologically accurate) and *in vitro* (relatively inexpensive and easier to perform) methods. We aim here to elucidate the problem of arriving at the holy grail rather than solve it.

6.1 Background

Accurate elucidation of TF binding specificity remains an active challenge in gene regulatory research. Several *in vitro* and *in vivo* experimental techniques have been developed to elucidate this, as reviewed in Chapter 2. One *in vitro* technique, a protein binding microarray (PBM) [15], generates high throughput binding data that covers all possible 10-mers and has been used to predict TF binding sites (TFBS). PBMs are less expensive, easier to generate than *in vivo* models, and provide comprehensive TF-DNA binding affinity without the confounding contextual binding site environmental factors [107]. These contexts include chromatin accessibility [147], competition or cooperation with other TFs [80, 171], cell line or condition specificity [8] among other factors. In spite of these benefits, these factors are also responsible for cases where *in vitro* techniques have poor performance when predicting *in vivo* binding [140, 42].

In vivo techniques capture the contextual binding site environmental information. These include chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) [77], which can accurately provide TF's genome-wide binding information. However, ChIP-seq is limited to specific cell lines or cellular conditions, is experimentally expensive, and can only be used to study a single TF at a time. On the other hand, DNase hypersensitivity site information, captured by DNase I digestion and high-throughput sequencing (DNase-seq) [173], can provide chromatin accessibility data for multiple cell types, but cannot be used independently to predict binding since they are not specific to any TF [174]. Therefore, they are combined with other data sets to predict TFBS though they have also been used to model TF binding specificity [217].

A combinatorial approach can alleviate the weaknesses of the *in vitro* and *in vivo* techniques: multiple data sets, *in vivo* and *in vitro*, are combined to improve TF binding site prediction.

Several studies have used various combinations of *in vivo* and *in vitro* information to predict TF binding sites (TFBS), with varying results. One approach of relevance to this study is the probabilistic integration of PBM, epigenetic and sequence data (PIPES) [217], a technique that combines PBM, DNase and evolutionary data (sequence conservation) to predict tissue-specific TFBS using a probabilistic graphical model. This method uses the combined *in vitro* occupancy learnt from PBM data and the probability that the site (36bp overlapping sequences) is located in an open chromatin site. PIPES infers this probability from DNase tag densities, considering a site open if it contained a high tag density over 15 and closed if it has a low tag density. PIPES does not use a PWM to model binding, rather, it extends the *k*-mer based method using a biophysically-motivated model. However, this technique is tissue specific. Also, Zhong et al.

reported conservation data did not improve on their model, possibly due to use of conservation over a longer sequence – only TF binding sites are expected to be conserved and not the entire sequence. Other techniques using DNase-seq data worth noting include protein interaction quantification (PIQ), which combines DNase and PWM models for different TFs to predict binding sites; CENTIPEDE [147] and MILLIPEDE [113], which additionally use conservation data; Romulus [74], which uses PWM and DNase-seq data to classify bound from unbound binding sites; and others like intersect [132] and epigenetic priors [36].

The TF binding specificity information from the experimental techniques is used to predict TF binding sites, decomposed into PWM, *k*-mer and machine learning models. PWMs are straightforward and easy to visualise with sequence logos. However, they do not capture complex binding environment, and can only describe the base and not the shape readout [172] (see Section 2.1 for an explanation of these TF-DNA interaction modes). On the other hand, *k*-mer models are better but are unwieldy hence difficult to apply when predicting TFBS. Besides, the currently available tools are tailored for PWM models [140], explaining the low adoption of other approaches by the research community. The models learned from *in vitro* based data do not generalise well when predicting *in vivo* binding [139, 140], which is attributable to overfitting and background noise bias. Also, in PBM and HT-SELEX data, artefactual over-representation of certain *k*-mers has been reported: the 'sticky *k*-mers' phenomenon [76, 139]. Because of the preceding, a single model to describe TF binding specificity remains elusive [172, 1, 140], with the best option being the use of statistical and machine learning models to combine a variety of prediction information. These, however, also run the risk of over-fitting the data, and are difficult to explain and visualise.

Nevertheless, a machine learning model's ability to integrate complex information including the DNA-shape of the binding site partly explains their recent increase in usage [99, 121, 103]. The use of the shape readout information of binding sites has long been known to play a role in TF binding [81, 102], but has only recently been generated [159] and demonstrated to improve prediction of TF binding *in vivo* [121] and *in vitro* [1]. Moreover, TFBS are believed to be under evolutionary constraint [134] and are localised around transcription start sites (TSS) [93]. All these can be modelled using statistical and machine learning modelling, but the use of these data sets, either when predicting TFS or modelling TF binding specificity, remain varied generating conflicting results – see Section 6.3.4.

In this study, we start our attempt to combine *in vivo* (PBM) and *in vitro* (DNase-seq) to model TF binding specificity using *k*-mer and PWM models, with modest success. Next, we apply machine learning techniques to combine PBM-derived *k*-mer models and DNase information clustered from different cell lines, with some success. Motivated by this, we move on to include the above-reviewed data as features (DNA-shape, TSS and conservation) in an XGBoost [28] gradient boosting classification model, using ChIP-seq data for training. We also employ XGBoost to investigate how TF binding specificity can be explained by these features and investigate the best way (scoring function) to score sequences using *k*-mer models.

6.2 Chapter aim and objectives

The main aim of this chapter is to investigate how TF binding site environment influences or contributes to its binding and how we can utilise this information to improve prediction by combining with *in vitro* data. We are dealing with thesis objectives four and five as re-listed and expanded below:

1. Investigate the use of chromatin accessibility, evolutionary conservation, proximity to transcription start sites, DNA-shape among other factors influencing TF binding to improve *in vivo* prediction

2. Design an algorithm that incorporates the factors in objective one starting with PBM and DNase data to generate motif models that predict *in vivo* binding better than models derived from PBM only

3. Investigate the phenomenon of sticky *k*-mers

4. Investigate the use of machine learning to tackle the above objectives

6.3 Methodology

In this chapter, we first deal with the attempt to combine PBM and DNase-seq data using frequency *k*-mer counts difference in open chromatin sites and the human genome as noise or preferred *k*-mer prior information. Next, we use the PBM-based *k*-mer models to score ChIP-seq data in addition to information on chromatin accessibility of the hits, conservation of the binding site and proximity of the transcription factor start sites information to train an XGBoost model to classify bound versus unbound sites. Finally, we explain how the various features contribute and explain TF binding specificity using feature-importance studies.

6.3.1 Data

Transcription factor binding site prediction has been investigated using a variety of experimental techniques described in Chapter 2. In this chapter, we show the need for and benefit of augmenting *in vitro* with *in vivo*-based techniques to improve model's predictive ability. The main data used in this study are:

- **PBM:** the de Bruijn sequences and the all 8-mer contigmers derived from the UniPROBE database
- **ChIP-seq:** uniformly-processed peak data from the ENCODE database
- **DNase:** fold enrichment information clustered over 125 cell types generated by ENCODE [185][1]

6.3.2 Method: A background correction approach using PBM and *k*-mer counts in DNase and human genome

In vitro techniques like PBM have been used to quantify TF binding intensities, and in most cases reflect *in vivo* binding. However, the binding site environmental factors like accessibility, the binding site flanking sequences, cooperative binding with other TFs, and many others influence TF binding. The chromatin accessibility of the binding, measured by DNase-seq data, cannot be captured by PBM technique. We hypothesise that the *k*-mer frequency in the open chromatin sites when compared with the genome-wide frequency, will provide data useful to calibrate the TF preference for a given *k*-mer. This section describes how we address this hypothesis.

6.3.2.1 *k*-mer counts

As shown in Figure 6.1, we extracted all possible 8-mers with up to two gaps from the PBM data using the Seed-and-Wobble (SnW) algorithm. We then obtained the counts of these *k*-mers in the Clustered DNase and the human genome (hg19). We then normalised them by dividing each *k*-mer count by the total count to obtain frequency counts, which is comparable across sequences. Next, we computed frequency difference between the *k*-mer frequency counts in the human genome and in the open chromatin sites to obtain two sets: *k*-mers differentially

[1] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/

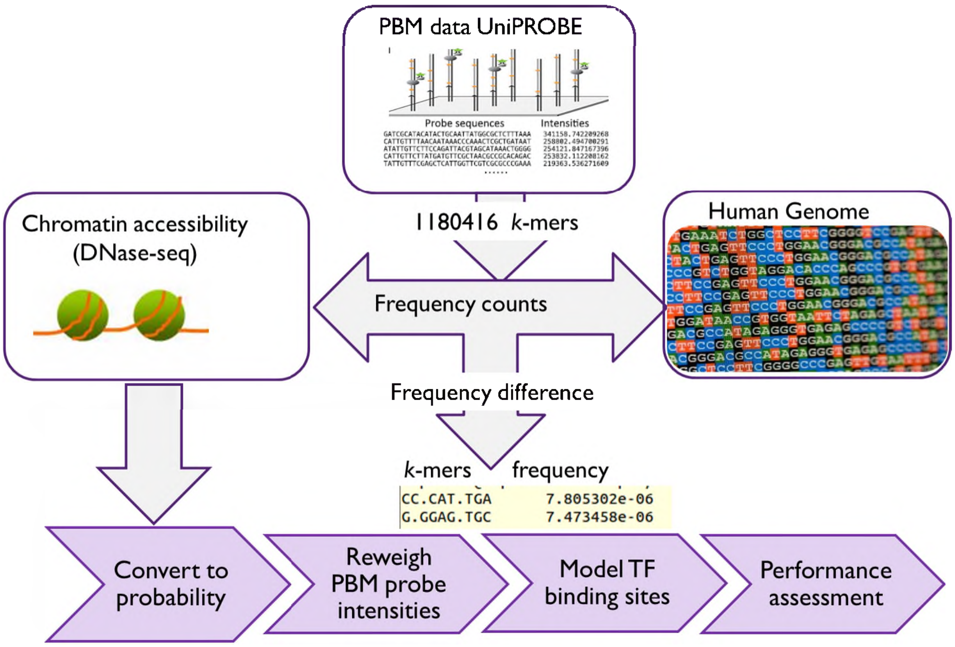

PBM data UniPROBE
Probe sequences
Intensities
GATCGCATACATACTGCAATTATGGCGCTCTTTAAA 341158.742209268
CATTGTTTTAACAATAAACCCAAACTCGCTGATAAT 258802.494700291
ATATTGTTCTTCCAGATTACGTAGCATAAACTGGGG 254121.847167396
CATTGTTCTTATGATGTTCGCTAACGCCGCACAGAC 253832.112208162
TATTGTTTCGAGCTCATTGGTTCGTCGCGCCCGAAA 219363.536271609
......
1180416 k-mers
Chromatin accessibility (DNase-seq)
Human Genome
Frequency counts
Frequency difference
k-mers
frequency
CC.CAT.TGA 7.805302e-06
G.GGAG.TGC 7.473458e-06
Convert to probability
Reweigh PBM probe intensities
Model TF binding sites
Performance assessment

Fig. 6.1 Methodology flow diagram. Flow diagram showing how *in vivo* prediction by PBM data was improved.

enriched in the human genome form background noise ($hg - dn$) while those enriched in the chromatin sites represent preferred k-mers information ($dn - hg$). See Box 6.1 for definition of some of these terminologies and notations.

The process of combining PBM and genomic information involved some iterations, outlined in Section 6.4.1.2. However, two main techniques that use these data directly deserve a detailed description.

6.3.2.2 Reweighting PBM intensity scores

The PBM technique generates intensity scores for each of the 36 bp probes in the array. Our first approach sought to reweight the intensity score based on dn and hg frequency k-mer count difference. We achieved this by:

- **Reranking:** This option directly uses the $dn - hg$ frequency counts to reweight the probe intensity scores. To do this, we score each probe sequence directly using the normalised $dn - hg$ frequency scores, use the score to reweight the intensity score, and finally transform them back to original scale, thereby altering the ranks of the sequences. A normal SnW run is then carried out to learn a PWM model. At this point, one can use any algorithm that learns k-mer or PWM models from PBM data.

- **Secondary motif:** We take advantage of the *rerank.pl* the algorithm that comes with the SnW code [15]. It scores probe sequences using a given primary PWM, generated from the first SnW run, giving any match a lower score. We modify the SnW algorithm to generate the primary PWM based purely on normalised $hg-dn$ frequency counts and use it to rerank PBM intensity scores. After reranking, a normal SnW run is used to generate a PWM, which should ideally eliminate the noise based on k-mers differentially enriched in the genome.

6.3.2.3 Background noise correction

In a PBM, the preference of a given TF for each of the k-mers (Enrichment scores) is computed using an L-statistic [15], a modified form of the Wilcoxon-Mann-Whitney statistic. Certain k-mers have been observed to be spuriously over-represented, leading to noise [76]: a "sticky k-mers" phenomenon. Jiang et al. [76] used ANOVA to detect these k-mers and showed that correcting for the noise using a background correction algorithm improved the quality of the k-mer models. However, they could not explain why they exist. We hypothesise that these k-mers are overrepresented in the whole genome compared with the open chromatin sites. We tested this by simply determining the percentage of the top 50 sticky k-mers identified by Jiang differentially enriched in the genome.

This next approach modifies the k-mer enrichment score to reduce background noise (also called the residual or sticky k-mer effect). Since the k-mers differentially enriched in the genome can be considered to represent noise (just like sticky k-mers), we use a modified background noise correction algorithm developed by Jiang to transform the k-mer E-scores, a ranked-based enrichment score [15]. Using the same notations used by Jiang: for a TF i, k-mer j and background noise τ, we can correct for the bias from the E-scores y_j as:

$$y' = y_j - E(\tau_j|Y)' \tag{6.1}$$

where $E(\tau_j|Y)$ is the posterior mean of τ_j. Finally, y' is transformed to the original scale to obtain a *corrected E-score*.

Box 6.1: Key terminologies

k-**mers**: DNA sub-sequences of length *k* contained in the genome.

Sticky *k*-**mers:** These are *k*-mers that are spuriously over-represented in PBMs, leading to noise.

Frequency counts: The counts of a *k*-mer in the sequences divided by the counts of all the *k*-mers in the sequences. The frequency counts of a *k*-mer are comparable across sequences of different sizes – open chromatin and whole genome in our case.

$\mathbf{dn-hg}$: *k*-mer in the DNase-seq data minus frequency count in the human genome – represent *preferred k-mers*.

$\mathbf{hg-dn}$: *k*-mer frequency counts in human genome minus frequency count in the DNase-seq data – represent *k-mer noise*.

E-score: The enrichment scores of each *k*-mer as determined by TF binding intensity in PBM data.

Corrected E-score: Enrichment score of *k*-mers after correcting for *sticky k-mer* effect.

6.3.3 Method: Evaluating *k*-mer scoring functions

This chapter mainly uses *k*-mer scoring, either to assess the modified *k*-mer E-score models or to score sequences to create the machine learning features (see more on machine learning in Section 6.3.4). Therefore, it was necessary to optimise for the scoring approach that best captures the binding information from the *k*-mer models. The scoring techniques in the literature include the following.

Sum occupancy *k*-**mer scoring:** This approach is similar to Equation 3.3, where the score to a given sequence is the sum of the scores of all *k*-mers in the sequence. In this case, we use a scoring function as in Jiang et al. [76], but since we score genomic sequences, we do not consider the positional effect score – influence of the position of *k*-mer along the probe on intensity score on the E-scores [15].

Maximum E-score *k*-**mer scoring:** Sum occupancy assumes the score over a whole sequence predicts presence of a binding site, but this may not hold true for longer sequences since the presence of sites with a poor match reduce the overall score. In maximum occupancy scoring, the score for the sequence is the score of the best matching *k*-mer in the sequence [15].

Sum at maximum scoring position: Max occupancy scoring may not capture the score of a TF that has a binding site much longer than the *k*-mer length. In this case, a better approach would be to use a sum of all *k*-mers within the highest scoring site or by binning the sequences and using the maximum sum over any bin as the score. The size of the bin is a question, but 36bp is a good start [76].

We evaluate the above functions on their ability to discriminate bound (ChIP-seq peaks) from unbound (background) sequences, prepared as in Section 3.4.1. Furthermore, we test how well a machine learning model based on each of the function scores as features perform, providing additional support for their reliability and performance against each other.

6.3.4 Method: A machine learning approach

Augmenting the PBM intensity scores with *k*-mer frequency count differences in open chromatin sites did not generate convincing improvement in performance – see results in Section 6.4.1.2. Therefore, we sought for a different approach. This section describes how we used the PBM *k*-mer models and clustered DNase fold enrichment scores to training a supervised gradient boosting classifier (XGBoost [28]) on paired ChIP-seq data. The idea being, if we can train a generalised model using one cell line, which can reliably predict binding in a different cell line, then our ability to predict binding sites in cell lines without ChIP or DNase data would be improved. Additionally, machine learning enables us to investigate the contribution of other features to TF binding specificity: evolutionary conservation, DNA-shape and proximity to transcription start sites. The sections that follow motivate our choice of algorithm and describe the data, features, and techniques used for parameter optimisation and feature importance.

6.3.4.1 Data selection

Training a generalised TF binding site prediction model benefits from integrating a variety of data that capture diverse binding site environments' information. Therefore, in addition to PBM and DNase data used in the previous section, we use additional features as detailed below.

ChIP-seq data: ChIP-seq [77] provides high-confidence TF binding information in a given cell line or condition. This is the primary data used for training and evaluation of our models; we only use half the peaks as positive training set to create a high confidence training set. A motif has been found in up to 85% of the top peaks [3, 193]; therefore, using half of the peaks rated higher by peak calling is likely to contain a TFBS. For the negative set, we extract a similar size of sequences located 500bp downstream of the positive set, a region we found to provide appropriate background sequences – see Section 3.5.2.

DNase I hypersensitivity data: The chromatin accessibility of a site for TF binding is arguably the most important feature for discriminating bound versus unbound sites [217, 132]. However, DNase-seq is cell-type specific and does not provide information on the TF that binds to the open site. We use DNase-seq data clustered from over 125 cell types from ENCODE, to generate a more generalised model. We specifically use the normalised scores based on narrowPeak signalValue[2] as features. We convert the DNase signal scores for each BED coordinate into a bigWig file using *BEDtoBigwig* from UCSC.

PBM-derived *k*-mer models: The uPBM data exhaustively captures TF preference without confounding environmental effects [140]. In this study, all 8-mer contigmers are used to score the ChIP-seq sequences to obtain a PBM *k*-mer score used as a training feature.

Sequence conservation: Transcription factor binding sites have been reported to be conserved across species [134, 79], with divergence in some being linked to gene duplication [134]. PhastCons [170] measures the probability of a site, considering adjacent sites, being conserved among a given set of aligned species. PhyloP [148] provides a similar measure but doesn't consider adjacent sites. We obtained genome-wide conservation data in bigWig format, calculated from multiple sequence alignment of 100 vertebrates based on the two techniques: phastCons[3] and phyloP[4]. A list of the vertebrates used, and the processing techniques used can be obtained from UCSC[5].

DNA shape: Transcription factors recognise their binding sites in their 3D conformation. This conformation is described by DNA-shape data generated by the DNAshape tool [221]. We obtain genome-wide shape predictions of minor groove width (MGW), Roll, Propeller twist (ProT) and Helix Turn (HelT) from GBshape genome browser[6] [33]. For each hit site, the shape scores are extracted using pyBigWig [161].

[2] http://genome-euro.ucsc.edu/cgi-bin/hgc?db=hg19&c=chr21&o=33032260&t=33033430&g=wgEncodeRegDnaseClustered&i=58&l=33032260&r=33033430

[3] ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP100way/

[4] ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons100way/

[5] http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=cons100way

[6] ftp://rohslab.usc.edu/hg19/

Transcription start sites: Since TF binding regulates gene expression, the proximity of a binding site to a TSS can provide additional support that a site is active. We downloaded the RefSeq TSS information from the UCSC Table browser [86][7] by selecting RefSeq, hg19, genes and predictions, and finally checking for chromosome name and TSS coordinates on the web page that follows.

6.3.4.2 Feature engineering

Given BED coordinates (chromosome: start, end) of 100bp ChIP-seq sequences, the features from the above data (summarised in Figure 6.2) are derived as follows:

1. Obtain the corresponding sequences using pysam, then score by all the PBM k-mers to get the **k-mer score feature**. Similarly, the frequency count differences converted to noise ($hg-dn$ scores) are also used score each sequence to obtain the **hg-dn score feature**. We retain the hit coordinates based on PBM k-mers for the next steps.

2. From the hit coordinates, we get the maximum accessibility scores using pyBigWig, which we use as the **DNase score feature**. Further, the maximum mean **phastCons** and **phyloP conservation score features** are also obtained in a similar way from the phastCons and phyloP bigWig files respectively.

3. Again, using the hit coordinates, we use Pybedtools' *closest-bed* command with a $-A$ flag to obtain the distance to the nearest TSS irrespective of the strand; this is the **TSS score feature**.

4. Finally, we extract the $4n$ **hit sites DNA-shape features** from the corresponding shape bigWig files using *pyBigWig*, where n is the length of the binding site. The 4 shape features are: Roll, Minor groove width, Propeller Twist and Helix turn Helix score.

5. In addition to the core features, we also test how well the scoring functions for k-mers capture models' binding information. These are k-mer sum occupancy score, the sum at maximum scoring site and k-mer the maximum occupancy score.

The above protocol is repeated for each BED coordinate in the positive and negative set. We then store the results in a Pandas DataFrame that we use directly in scikit-learn as a feature vector.

[7] http://genome.ucsc.edu/cgi-bin/hgTables

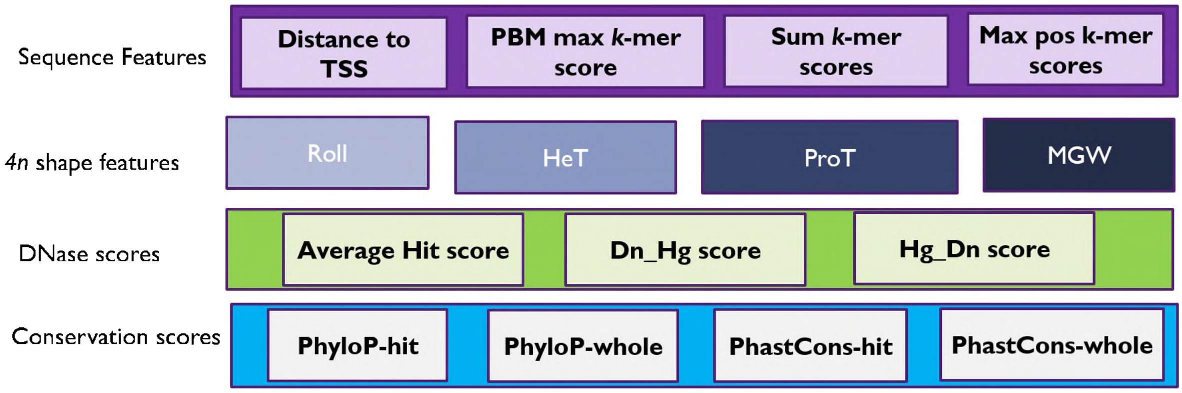


Fig. 6.2 The features selected for training the classification model.

6.3.4.3 The Algorithm

Given the size of the training data, and following scikit-learn [144] recommendations for the machine learning methods to use, we test Gradient Boosting Classifier, Stochastic Gradient Descent, Support Vector Machines classifier and the XGBoost algorithm. XGBoost can directly access scikit-learn's API, though it is not yet directly implemented in scikit-learn. Our initial test uses each tool with default parameters. In this section, we describe the gradient boosting methods with specific emphasis on XGBoost [28].

Gradient Boosting Freund and Schapire introduced the concept of boosting first in 1996 [47]. It starts with a base learner then takes advantage of weak learners (just slightly better than random) to boost the model while minimising misclassification error gradually. The input data is reweighted in each iteration while assigning higher weights to the hard-to-classify data. Finally, all the results of the base learner through the iterations are combined by majority vote. Adaptive Boosting (AdaBoost) implemented boosting algorithm with automatic updating of weights in a sequence of consecutive rounds using decision trees or stumps as base learners. Later, Friedman [48] introduced gradient boosting, where the weak learners are added by steepest gradient descent optimising the loss function; either by cross-entropy for classification or mean square error (MSE) for regression. Therefore, the residual errors of previous trees are corrected by new learners.

Extreme Gradient Boosting (XGBoost) [28], is a recent implementation of gradient boosted decision trees, optimised for speed and performance. Chen and Guestrin implement a sparsity-aware algorithm for sparse data, and to decide on split points during tree learning they use a weighted quantile sketch. XGBoost prevents over-fitting by shrinkage, which reduces the influence of each tree after each boosting round by a numeric learning rate η between 0 and 1, and by column sub-sampling. The ability to prevent overfitting allows us to add and test more features without the risk of over-fitting. The main benefit of XGBoost is its speed, achieved through parallelization of tree construction and cache optimisation of data structures. Also, the ability to perform out-of-core (external memory) and distributed computing ensure scalability of model learning.

6.3.4.4 Parameter optimization

We use scikit-learns' *GridSearchCV* function to optimise the XGBoost parameters using 10-fold cross-validation (CV). We select for the following: logistic loss function; learning rate $\eta = 0.1$, to reduce over-fitting; a maximum tree depth of 8; early stopping after 3000 boosted trees; and row and column sampling rates of 0.8. We leave the rest of the parameters at their defaults.

6.3.4.5 Feature importance

We use feature importance to understand how the features contribute to TF binding specificity. We did this using a variation of Recursive Feature Elimination (RFE) approach akin to leave-one-out CV, where we first eliminated a single feature and created a model using the remaining features, with importance measured by the decrease in model performance. Additionally, we start with a baseline model and sequentially added the rest of the features, creating a model for each group; we use this specifically to optimise individual features. We also used the inbuilt XGBoost feature importance function to extract and plot feature importance information. However, this only reveals their importance based on how often they were used to make decisions during tree building, and not how they contribute to the accuracy of the model. Therefore, the importance assigned by inbuilt *plot_importance()* function differs from our RFE: the plot does not provide a true picture of how the features contribute to a model's accuracy.

6.4 Results

The analysis in this chapter is divided into two sections: a background correction approach of combining PBM and DNase-seq, and a machine learning approach to elucidating and modelling TF binding occupancy. Encompassing both parts is the use of a k-mer scoring approach. In

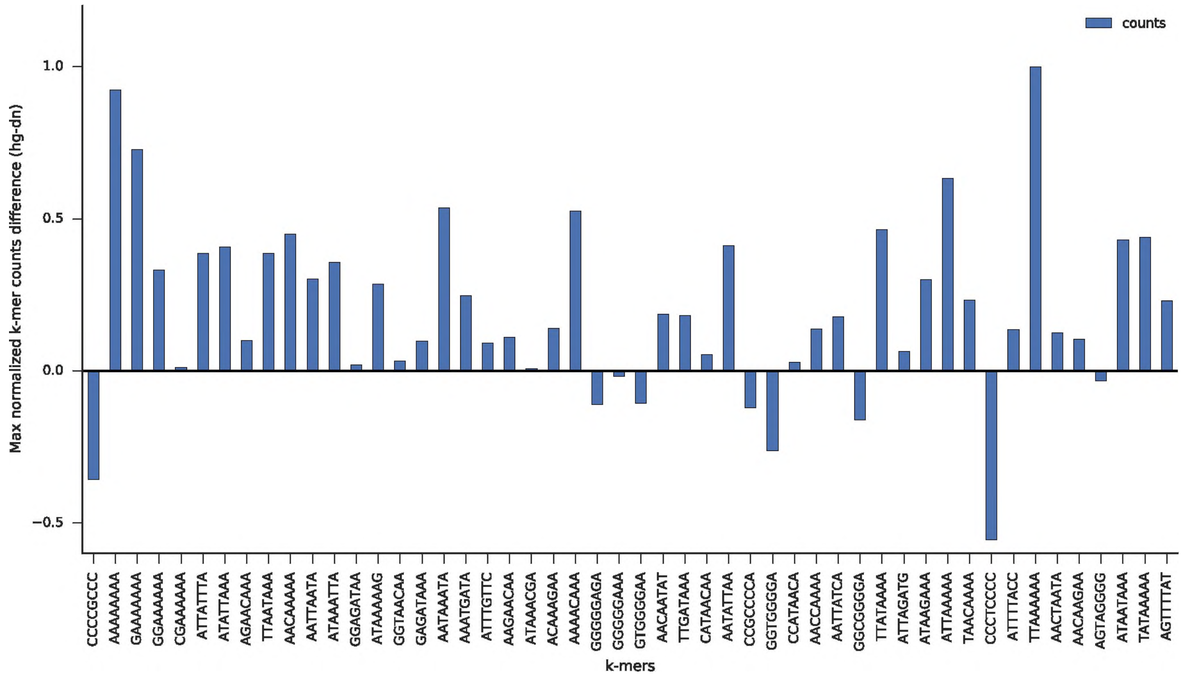


Fig. 6.3 Sticky *k*-mers are differentially enriched genome-wide compared with open chromatin sites. We count all possible 8-mers in a repeat-masked human genome and clustered DNase-seq data – see Section 6.3.2.1 for details.

the first section, we use *k*-mer counts in the open chromatin sites to either represent noise or preferred *k*-mers information as described in Section 6.3.2.1.

6.4.1 Results: A background correction approach

To combine PBM and DNase-seq data to improve *in vivo* prediction, we use the *k*-mer counts in the human genome and open chromatin sites to either rerank probes or eliminate background noise in the uPBM probes. But first, we investigate the sticky *k*-mer effect.

6.4.1.1 Sticky *k*-mers are differentially enriched genome-wide compared with open chromatin sites

Certain *k*-mers are systematically overrepresented in PBM experiments for unknown reasons; described as *k*-mer with background noise with a standard deviation greater than one from two different PBM array experiments [76]. These have also been identified in HT-SELEX. A large fraction (81%) of these *k*-mers are differentially enriched genome-wide compared with open chromatin sites (Figure 6.3) possibly explaining their existence. The similarity in the sticky

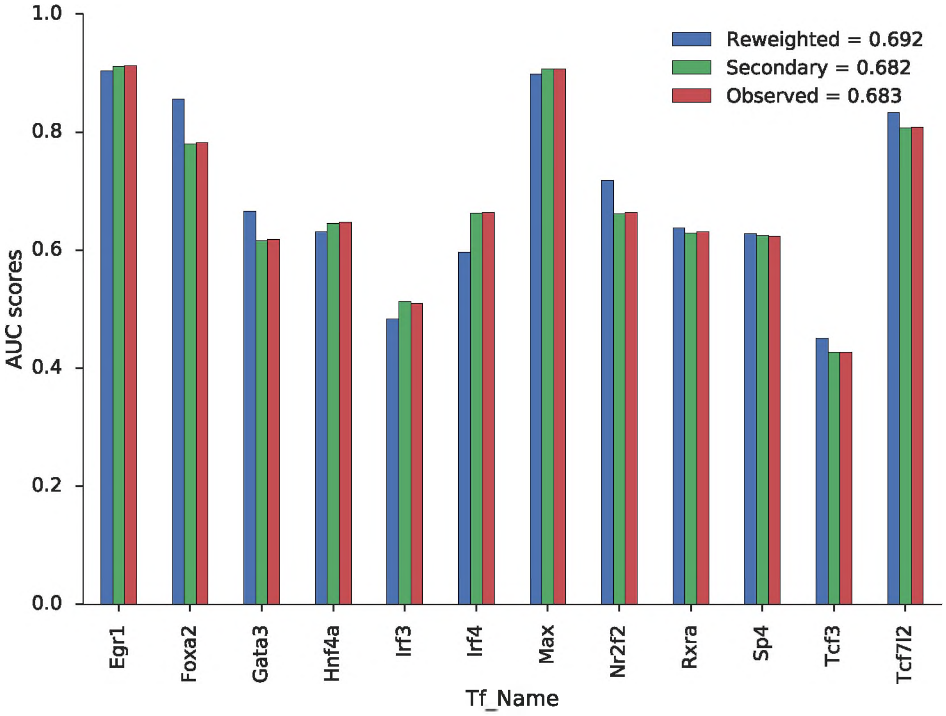


Fig. 6.4 Combining PBM and DNase-seq data by reranking PBM probes. The PBM probes are reranked either using the primary PWM to generate a "Secondary" PWM or directly using preferred *k*-mers frequency counts to generate a "Reweighted" PWM. We compared the evaluation scores of these models against the score of "Observed" PWM, which is the PWM generated from a normal Seed-and-wobble run on original PBM data. The PWMs are evaluated in ChIP-seq data using MARSTools' SCORE-Energy to obtain the AUC scores for each TF. The legend displays the mean AUC scores for each approach.

k-mers and those differentially enriched genomewide as opposed to open sites means that we can employ a similar approach to background correction. The section that follows describe the results of PBM background correction approach.

6.4.1.2 Background noise correction modestly improves *in vivo* prediction

Taking differential enrichment genome-wide as noise, and using the background correction equation 6.1, we achieved a modest prediction improvement *in vivo*. What follows is a description of the iterations undertaken to combine these data sets.

1. From the PBM data, the binding affinity of a TF is measured, using a rank-based statistics. Seed-and-Wobble algorithm chooses a seed *k*-mer then it wobbles around the bases of the seed, calculating statistics for each variant. Therefore, our first approach used the *k*-mer frequency counts in the DNase data and E-score to choose a seed *k*-mer, by modifying the

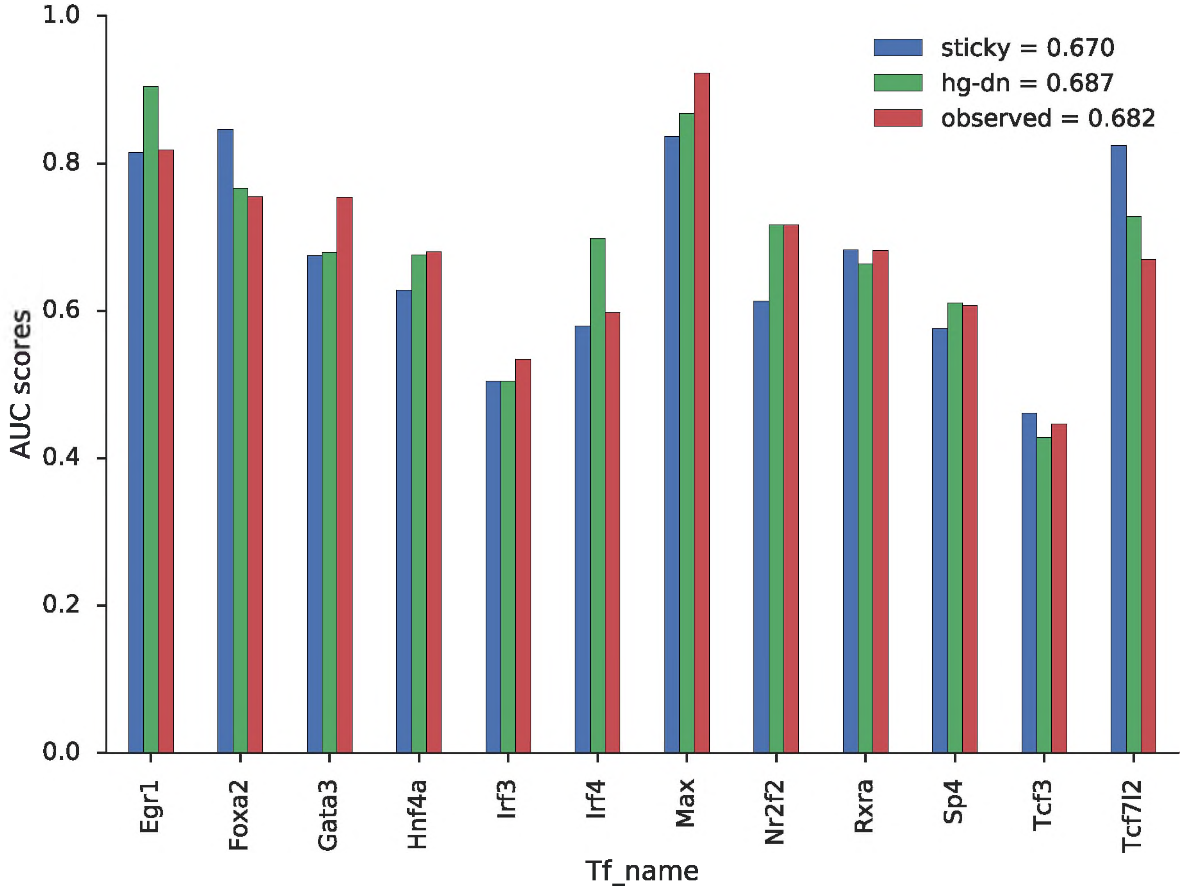


Fig. 6.5 Combining PBM and DNase-seq data by background noise correction. Taking k-mers differentially enriched genome-wide as noise, we use a background noise correction algorithm to model TF binding as PWM. $hg - dn$, represent k-mer frequency counts in human genome minus frequency count in the DNase-seq data. "observed": PWMs generated from a normal Seed-and-wobble run on original PBM data. The PWMs are evaluated in ChIP-seq data using MARSTools' SCORE-Energy to obtain the AUC scores for each TF. The legend displays the mean AUC scores for each approach.

SnW algorithm. Although the seed does influence the *k*-mers used to build the model, this approach did not work because DNase *k*-mer counts are quite noisy and uninformative on their own.

2. Next, we modified the SnW algorithm to take DNase frequency counts as input and used to re-compute the E-scores. The motifs generated by this algorithm performed poorly; the noise problem persisted.

3. The SnW algorithm also provides a function to rerank the intensity scores, by scoring the sequences with a PWM, the sequences that score well are down-ranked. The rerank function is used to capture secondary binding modes. By taking advantage of the rerank function, our next approach was a modification of (2) whereby the PWM models generated in the first run were used to rerank the intensity score of de Bruijn sequences followed by the usual SnW run. This approach uses the *k*-mer frequency counts in DNase data as noise, which are eliminated in the reranking step. This approach demonstrated limited success; of the five chosen motifs, Max motifs performed better than JASPAR and UniPROBE motifs, Gabpa better than UniPROBE while the rest fared poorly. We performed assessment based on motif enrichment in ChIP-seq data using CentriMo.

4. From the above step, we realised that the reranking approach had some potential in our algorithm. Our fourth approach used the modified SnW algorithm, but instead of combing frequency counts with E-scores, the frequency counts alone were used to generate a motif which reflected the probability of DNase *k*-mer counts. This was then used as in attempt (3) above. The models generated by this approach were better than those from UniPROBE in 8 of the 11 TFs tested. The improvement, however, was modest.

5. Next, we used the *k*-mer frequency count differences in the human genome and the DNase data to generate the PWM, which we then use to rerank the intensity data. We argued that, since reranking assigns a lower score to the sequences that match the PWM, *k*-mers over-represented in the genome as compared with the open chromatin sites will be reduced among the top *k*-mers; and possibly edge closer to reflect the *in vivo* binding behaviours in the open sites. This attempt performed in a similar manner to the normal SnW run (Figure 6.4, secondary). The poor performance is attributable to the information loss by the PWM.

6. Therefore, our next approach directly used the k-mer frequency difference of preferred k-mers $(dn-hg)$. The $dn-hg$ scores for a given TF were normalised by standardization to obtain a z-score, which is then scaled to positive (by adding the minimum score to all the values). The scaled score is then used to score the probe sequences to obtain a probe score, which is then normalised to between 0 and 1 (by dividing all scores by the maximum score) and used to reweight the probe intensity scores by simple multiplication. This approach performed better than previous attempts, but this improvement is modest (Figure 6.4, reweighted).

7. Our final approach uses Jiang et al.'s background noise correction algorithm to reduce k-mer frequency counts difference represented as noise $(hg-dn)$ – see Section 3.5.2. This method also only modestly improves the PBM models (Figure 6.5, $hg-dn$), while correcting for Jiang et al. background noise leads to a performance drop (Figure 6.5, sticky).

All evaluations are carried out using MARSTools, energy scoring and ChIP-seq benchmark data. A reproducible IPython notebook describing this work is available[8].

6.4.2 Results: k-mer scoring function affects TFBS prediction

As demonstrated in Chapter 3, the scoring function used does affect the predictive ability of a k-mer model. To test this, we first scored positive and negative sequences using each of the functions and assessed how well they classify them by using the AUC score. We find max scoring to be the most predictive scoring approach (Figure 6.6) followed by the sum of k-mer scores and maximally scoring bin (max_kmer_pos). We further support these observations through feature importance studies for machine learning modelling (Figure 6.8A).

6.4.3 Results: Machine learning approach

Transcription factor binding specificity is influenced by both shape and base readout information. In this study, we investigate how we can improve the predictive power of PBM-derived k-mer binding models by augmenting with *in vivo* DNase information as our baseline model. We use DNase-seq data clustered from a variety of cell lines by the ENCODE Analysis Working Group [185]. Furthermore, we investigate the contribution of the following features to TF

[8] https://github.com/kipkurui/XGB-TFBSContext/blob/master/code/Combining%20PBM%20and%20DNase.ipynb

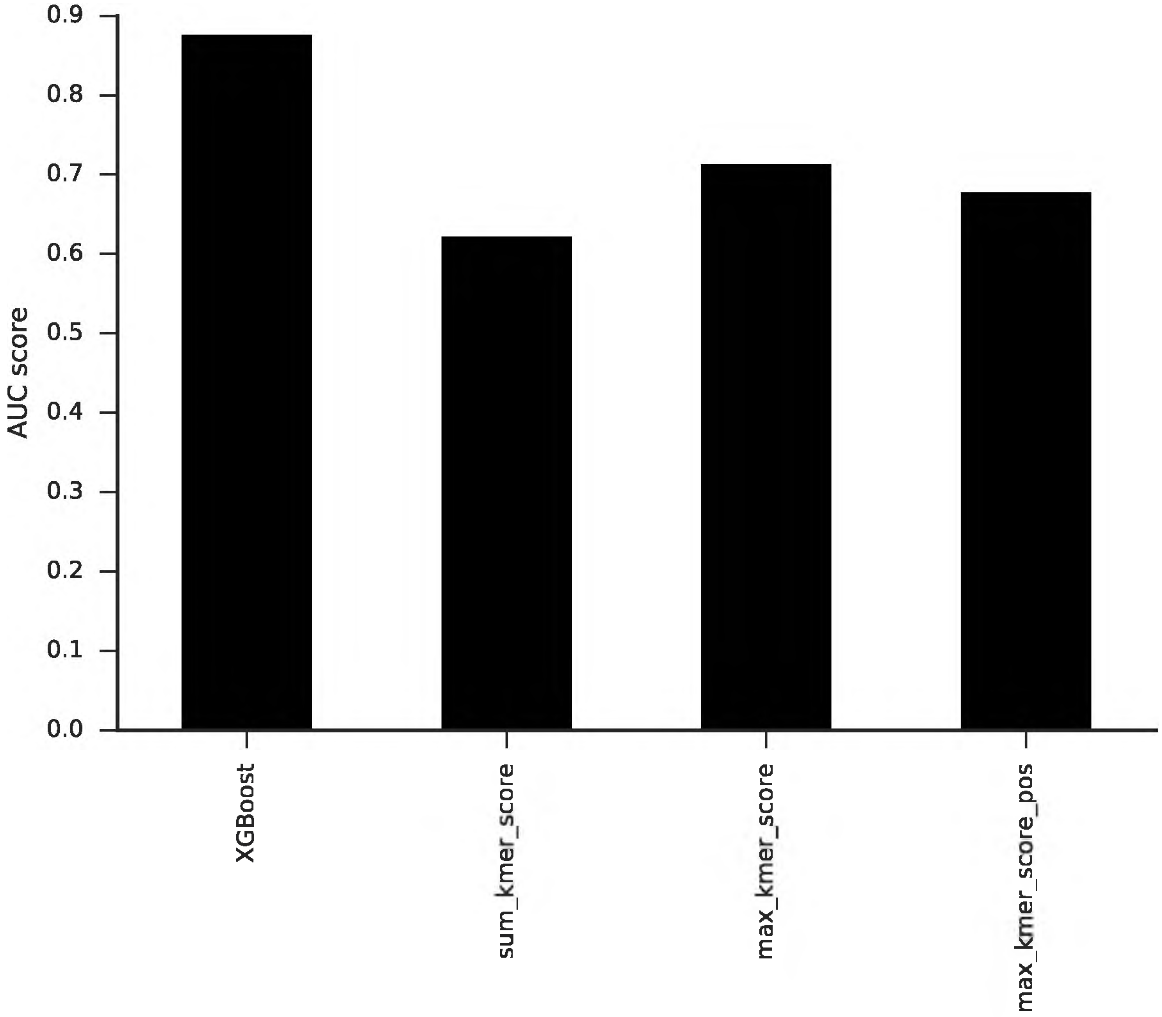


Fig. 6.6 Effect of *k*-mer scoring function. We test the ability of three scoring functions to classify bound vs. unbound sequences. The XGBoost bar shows results from multi-featured machine learning based model–see Section 6.4.3 for more details.

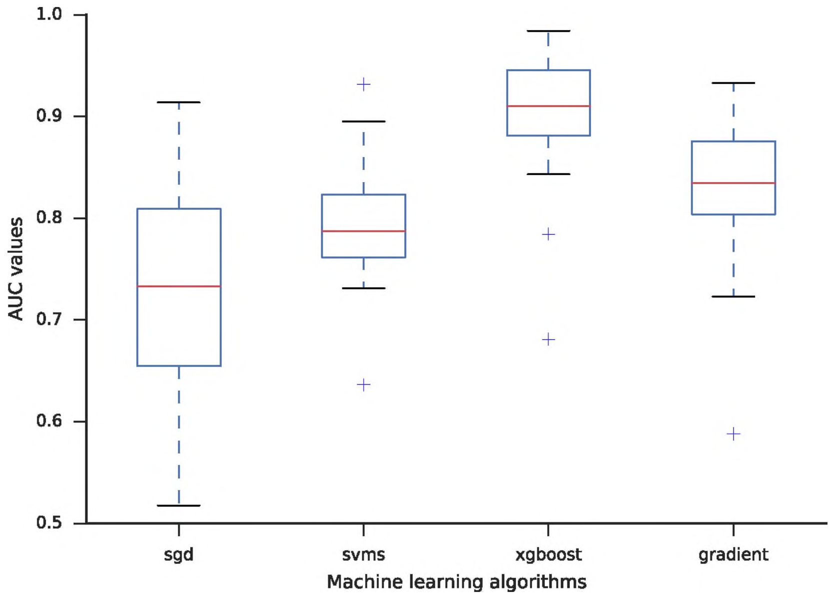


Fig. 6.7 Compare machine learning algorithms. To make a choice of the machine learning model to use, we compared the performance of the models trained with default parameters in SGD (stochastic gradient descent), SVMs (support vector machines) and gradient (gradient boosted machines).

binding specificity: proximity to TSS, evolutionary conservation and DNA-shape information. In addition to investigating the importance of these features, we optimise for feature extraction approaches.

For a given TF and cell line, we use 50% of total ChIP-seq peaks (with high confidence of containing binding sites) and a similar size derived 500bp away (negative sequences without binding sites) to train a supervised classification model with XGBoost. For TFs with data in more than one cell line, we train the model on one cell line and test on another. XGBoost is a gradient boosting based approach, where weak decision tree models are boosted to improve predictive power. Since we are using multiple heterogeneous features, there is a risk for overfitting, but this is avoided by XGBoost using shrinkage and column sub-sampling, a major reason for our choice. In addition, XGBoost, run on default parameters against SGD, SVM and GBM had a significantly better performance (p=9.15×10^{-6}, 3.93×10^{-5} and 1.3×10^{-3} respectively; Wilcoxon rank-sum test, Figure 6.7).

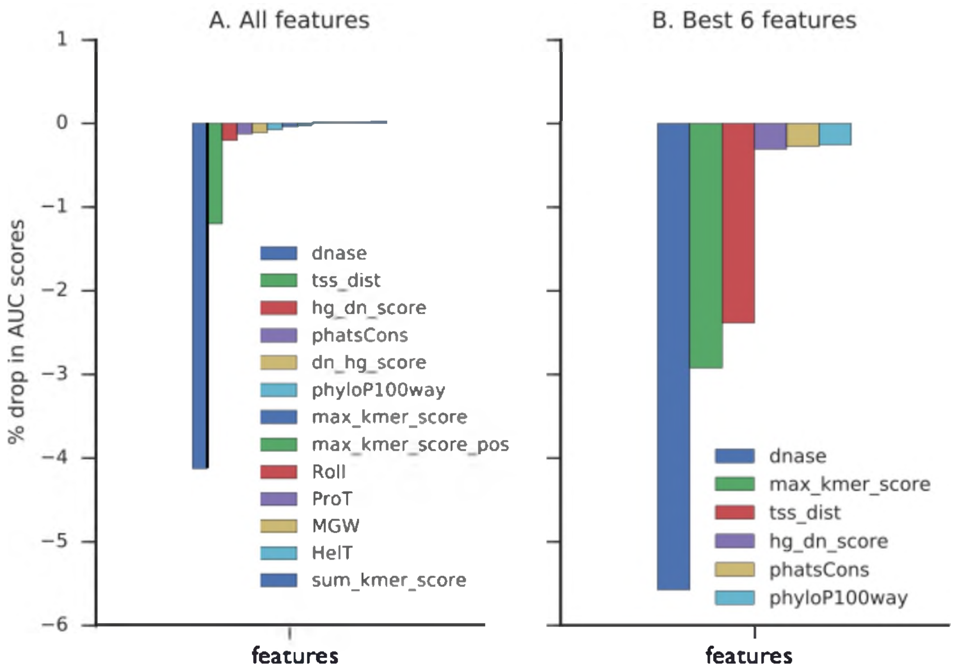


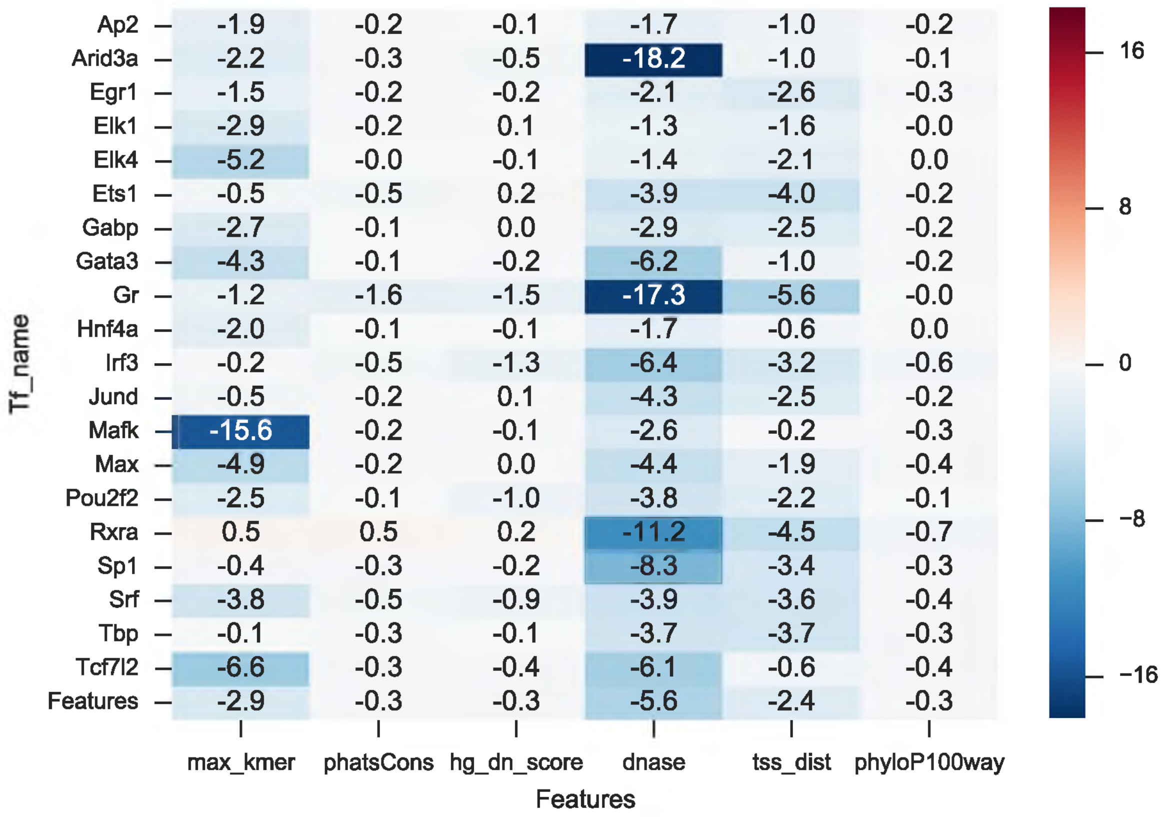
Fig. 6.8 Feature importance. Percentage drop in performance (AUC scores) when a given feature is eliminated. **A:** we use all the features. **B:** using the best performing features from **A**. The feature with the highest drop is the most important for model's predictive ability. The scores are ranked, top to bottom in the key and left to right in the plot.

6.4.3.1 Baseline model: DNase chromatin accessibility information is the most important feature

Using all the features, we first established those with the greatest contribution to the model's predictive ability by feature elimination. From this, we found the DNase information in the hit site to be the most informative feature for the model's predictive ability. From the 13 initial features, eliminating the DNase information leads to the greatest and most significant drop in performance (p=0.028, Wilcoxon rank-sum test; d=0.62, Cohen's effect size test) of over 4% (Figure 6.8A). This, however, is TF-specific with the TFs like Gr, Arid3a and Rxr3, with *k*-mer models of poor predictive value, being the most affected – see Figure 6.9.

Next, we eliminate the features that carry similar information and test for feature importance by recursive elimination with replacement (Figure 6.8B). The results do not differ much, except for an improved importance rank of the *k*-mer score; using *k*-mer binding model features based on different scoring functions, the remaining features complemented the eliminated feature.

Next, we plot the difference between the full model and those lacking one of the features (Figure 6.9) from which we can observe TF-specific feature importance. For example, Arid3a,



Tf_name	max_kmer	phatsCons	hg_dn_score	dnase	tss_dist	phyloP100way
Ap2	-1.9	-0.2	-0.1	-1.7	-1.0	-0.2
Arid3a	-2.2	-0.3	-0.5	-18.2	-1.0	-0.1
Egr1	-1.5	-0.2	-0.2	-2.1	-2.6	-0.3
Elk1	-2.9	-0.2	0.1	-1.3	-1.6	-0.0
Elk4	-5.2	-0.0	-0.1	-1.4	-2.1	0.0
Ets1	-0.5	-0.5	0.2	-3.9	-4.0	-0.2
Gabp	-2.7	-0.1	0.0	-2.9	-2.5	-0.2
Gata3	-4.3	-0.1	-0.2	-6.2	-1.0	-0.2
Gr	-1.2	-1.6	-1.5	-17.3	-5.6	-0.0
Hnf4a	-2.0	-0.1	-0.1	-1.7	-0.6	0.0
Irf3	-0.2	-0.5	-1.3	-6.4	-3.2	-0.6
Jund	-0.5	-0.2	0.1	-4.3	-2.5	-0.2
Mafk	-15.6	-0.2	-0.1	-2.6	-0.2	-0.3
Max	-4.9	-0.2	0.0	-4.4	-1.9	-0.4
Pou2f2	-2.5	-0.1	-1.0	-3.8	-2.2	-0.1
Rxra	0.5	0.5	0.2	-11.2	-4.5	-0.7
Sp1	-0.4	-0.3	-0.2	-8.3	-3.4	-0.3
Srf	-3.8	-0.5	-0.9	-3.9	-3.6	-0.4
Tbp	-0.1	-0.3	-0.1	-3.7	-3.7	-0.3
Tcf7l2	-6.6	-0.3	-0.4	-6.1	-0.6	-0.4
Features	-2.9	-0.3	-0.3	-5.6	-2.4	-0.3

Fig. 6.9 Feature importance: TF specificity. Testing for feature importance by recursive feature elimination. The heatmap displays the percentage change (negative: drop; positive: increase) in performance when a given feature is eliminated for each TF.

Gr and Rxra are greatly dependent on the DNase data (18, 17 and 11% drop, respectively), which could be associated with poor quality of the k-mer models. On the other hand, Mafk is distinct in that it depends more on the k-mer model (15% drop); hinting that base readout captures its binding specificity well.

6.4.3.2 DNA-shape features improve the baseline model's performance in a complementary manner

When using nine feature vectors and $4n$ shape features (ProT, HelT, MGW and Roll), the contribution of each of the shape features could not be established (Figure 6.8A). This failure is attributable to the complementary contribution of the shape features, as established below. In addition, we use early stopping to avoid over-fitting, therefore, the per binding site nucleotide specific shape feature does not easily reveal the importance of these features, since the tree building may stop before they are utilised. Therefore, starting with a baseline model of max

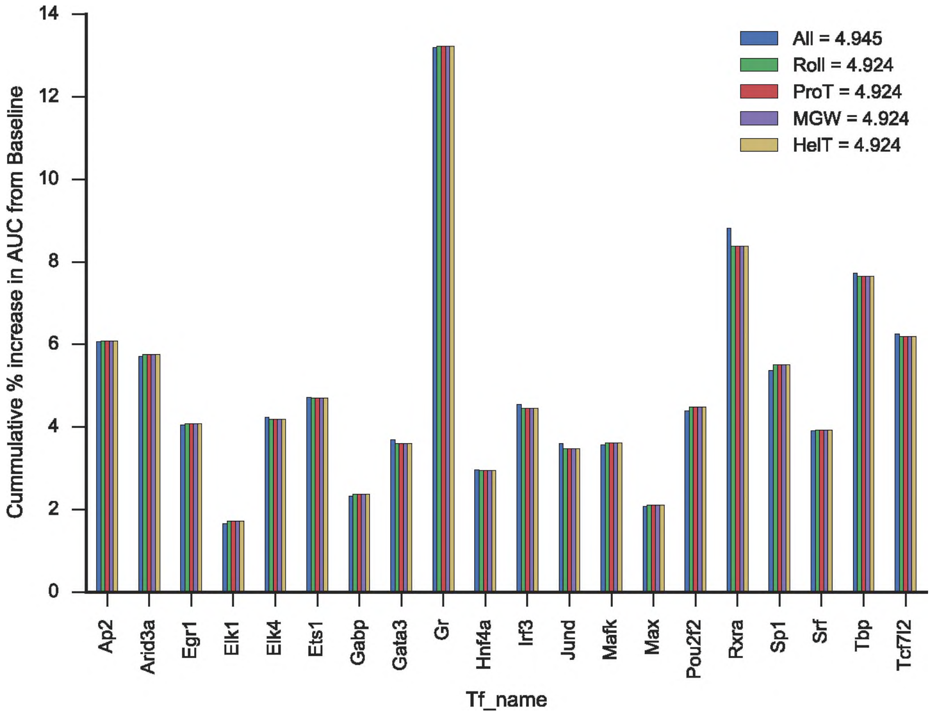


Fig. 6.10 Feature importance: DNA-shape. Testing for DNA-shape importance by feature addition to the baseline model. The figure displays the cumulative percentage gain in performance when a given shape feature is added.

k-mer score and DNase, we add each of the shape features and then determine the contribution to prediction ability of the model. A model with all the shape features performs equally to that with any one of them (Figure 6.10), and is significantly better than the baseline model (p=0.035, Wilcoxon rank-sum test; d=0.58, Cohen's effect size test). DNA-shape feature is most informative for Gr and Rxra, and least for Max and Elk1 TFs.

6.4.3.3 Conservation information of the *k*-mer hit improves model

Having established the importance of chromatin accessibility (DNase feature) and the 3D conformation of the binding site (DNA-shape feature), we sought to determine the predictive ability of the binding site conservation information. The sequences around the TF binding site are evolutionarily conserved for some TFs [134], making conservation scores an important TFBS prediction feature. However, contradictory results have been reported on how to use this data [217, 69, 1], whether at the hit site or whole sequence, or even the overall usefulness of each of the conservation feature (phyloP or phastCons score). To determine usefulness and use of conservation information, we test each of these variations: phastCons-hit, phastCons-whole, phyloP-hit and phyloP-whole.

To begin, we establish that conservation information significantly improves the performance of the baseline model (p=0.023, Wilcoxon rank-sum test; d=0.66, Cohen's effect size test), with an average performance difference of 5.4% from the model with all the conservation features (Figure 6.11). We also observe that the phyloP score for the whole site is the most informative conservation score feature with a significant performance improvement from the baseline model (p=0.035, d=0.62). On the other hand, phastCons is a better predictor for the hit site (Figure 6.11), but the improvement is just barely significant (p=0.048, Wilcoxon rank-sum test; d=0.60, Cohen's effect size test). The rest of the features do not lead to any significant performance improvement, but with a medium effect size of over 0.5. Gr and Rxra benefit most from the conservation data, while other like Gata3, Hnf4a, Max and Mafk gain little.

6.4.3.4 Augmenting baseline model with the noise information has no significant effect on baseline model

Initially, we attempted to correct for background in the PBM models but did not produce significant improvement (Section 6.4.1). Therefore, we tested whether the k-mer noise or preferred k-mers information based on frequency counts could improve the model performance. Again, we start with the baseline model. Although the two ($dn-hg$, preferred k-mers and $hg-dn$, noise) complement each other and have TF-specific performance difference, they do not lead to a significant performance improvement, with k-mer background noise information being a slightly better predictor of *in vivo* binding (Figure 6.12). On average, using all these features leads to 3.6% increase and up to 11% for Rxra. TFs with a less predictive k-mer model are greatly affected, an observation that has remained consistent throughout this study. TFs with weak k-mer models greatly benefit from additional features.

For some TFs like Max, however, the k-mer counts information are not predictive of the binding site, in fact preferred k-mers information leads to a drop in performance.

6.4.3.5 XGBoost Model predict cell type-specific binding but with generalizable predictions

Until this point, we have been training a model on one cell type and testing on another (Section 6.4.3). Therefore, we wanted to know if the models retain cell type-specific binding, and how well they can be generalised to diverse cell types. To test for cell-type generalisation,

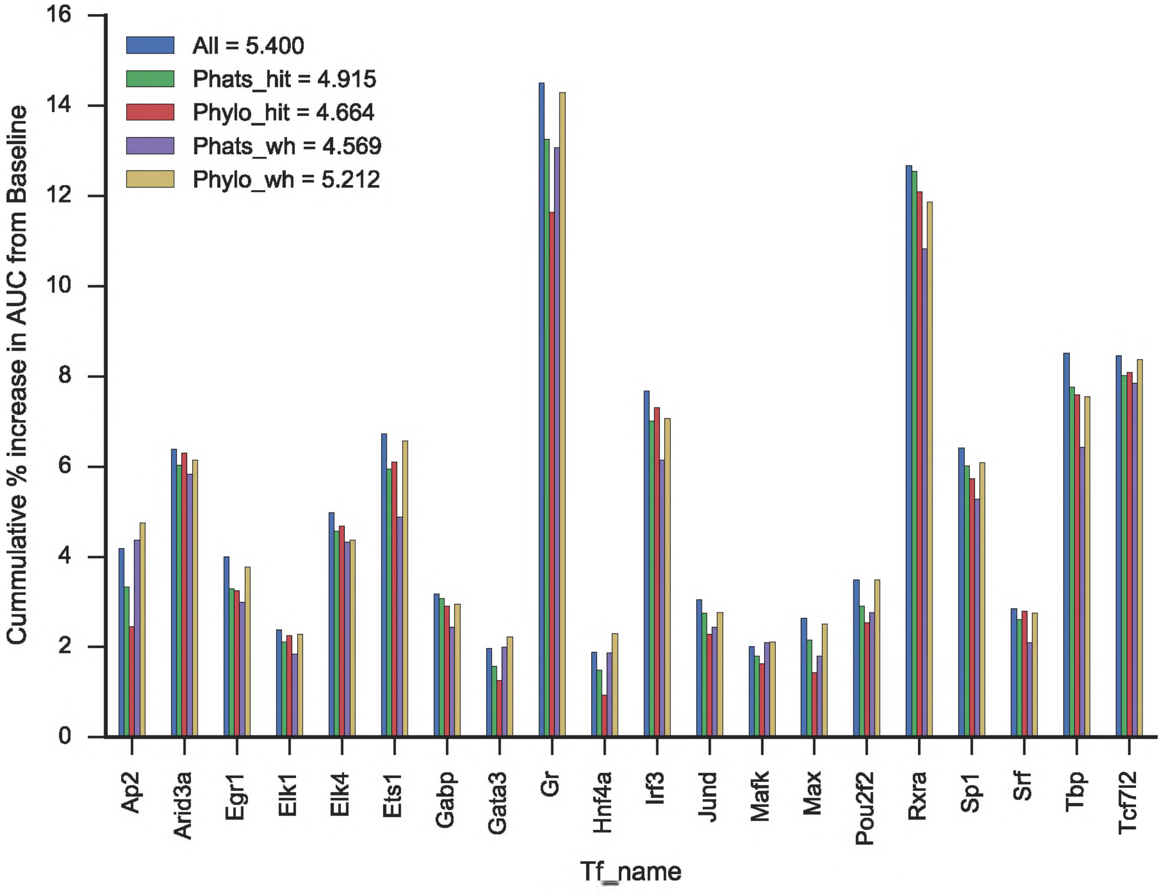


Fig. 6.11 Feature importance: conservation. Testing for feature importance (conservation) by recursive feature addition to the baseline model. The figure displays the percentage gain in performance when a given feature is added.

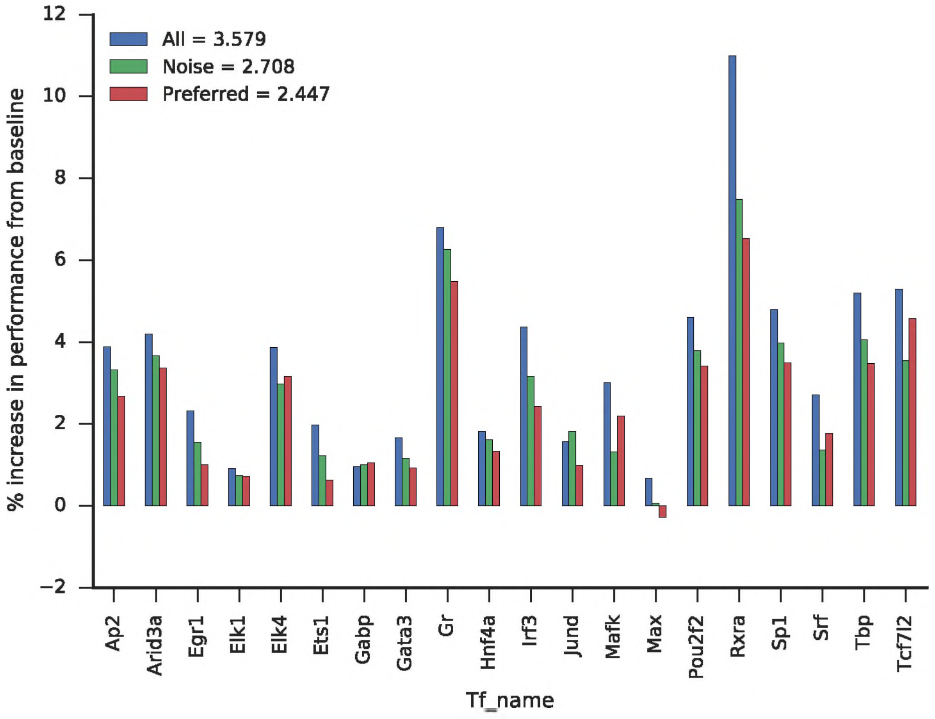


Fig. 6.12 Feature importance: *k*-mer noise. Testing for feature importance (noise) by recursive feature addition. The plot displays the percentage gain in performance when a given noise feature is eliminated.

we recursively trained a model on each of the cell types and tested how well the model can predict binding in the other cell types; we are also interested in determining the cell types whose models can generalise better. To do this, we used TFs with ChIP-seq peaks data in more than three cell lines.

First, we determine how different the performance of the model is when predicting other cell lines. We show this using the standard deviation (STD) of the models' performance in various cell types. We find that the majority of the TFs have an STD of less than 0.05 AUC score with Sp1 having the highest deviation at 0.06 (Figure 6.13). We can, therefore, propose that, except for a few TFs, a model learned from one cell line can generalise to other cell lines.

For *n* different cell-type data sets available for a given TF, we perform *n* leave-one-out CVs, each time leaving out a different cell-type and test how generalizable the models from the cell lines are. For most TFs, the models perform well across cell lines: models from some cell lines are more generalised (perform well across cell types), and some cell lines' TF occupancy can be easily predicted irrespective of model source. Cell-type *model generality* refers to how well a model trained on the cell type can predict TF occupancy in other cell-types; cell-type *occupancy generality* refers to how well TF occupancy in the cell type can be predicted by

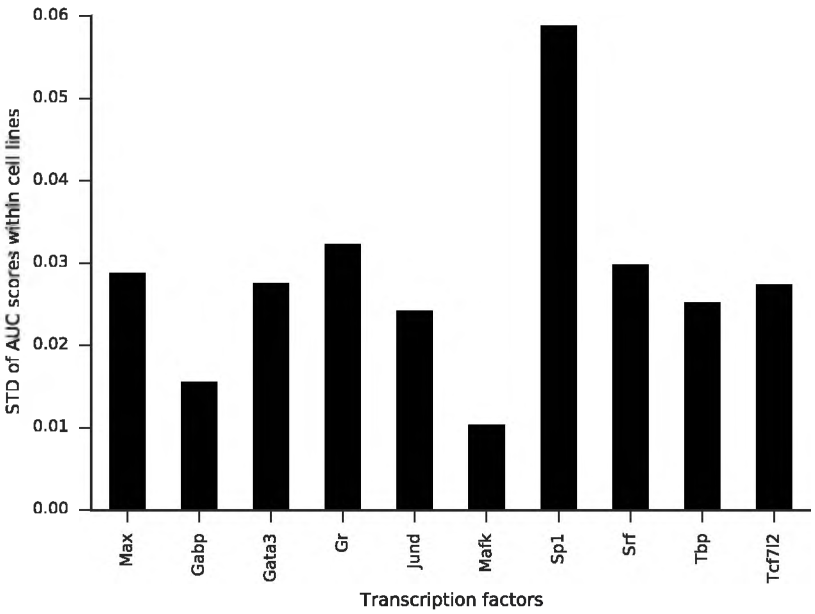


Fig. 6.13 Cell-type specificity of the model. Plot displays the standard deviation of the performance of a model trained in one cell line and tested for all the available cell lines for a given TF.

models trained on other cell types. See Figure 6.14 for analysis carried out on Max TFs and full results in the accompanying IPython notebook[9].

6.5 Discussion

Although *in vitro* techniques provide comprehensive binding intensity data for a TF, they do not generalise well in some cases to predicting *in vivo* binding because they do not capture binding site environments' contextual information. Therefore, it is desirable to combine with *in vivo* data to improve performance; this has been widely investigated with mixed performance [147, 36, 132, 113]. These data can be combined in two levels: to model TF binding specificity or to predict TFBS. Our first attempt in this study combined the data to learn better PWM or k-mer models that incorporate the accessible *in vivo* environment. Accessibility is captured by the preferred k-mer information (k-mers differentially enriched in open chromatin sites) or by eliminating k-mer background noise (k-mers differentially enriched genome-wide). These techniques marginally improved on the native PBM technique, and this could be explained by the weakness of our approach or the failure of the k-mer and PWM models to capture this

[9] https://github.com/kipkurui/XGB-TFBSContext/blob/master/code/Machine_learning_plots_and_explore.ipynb

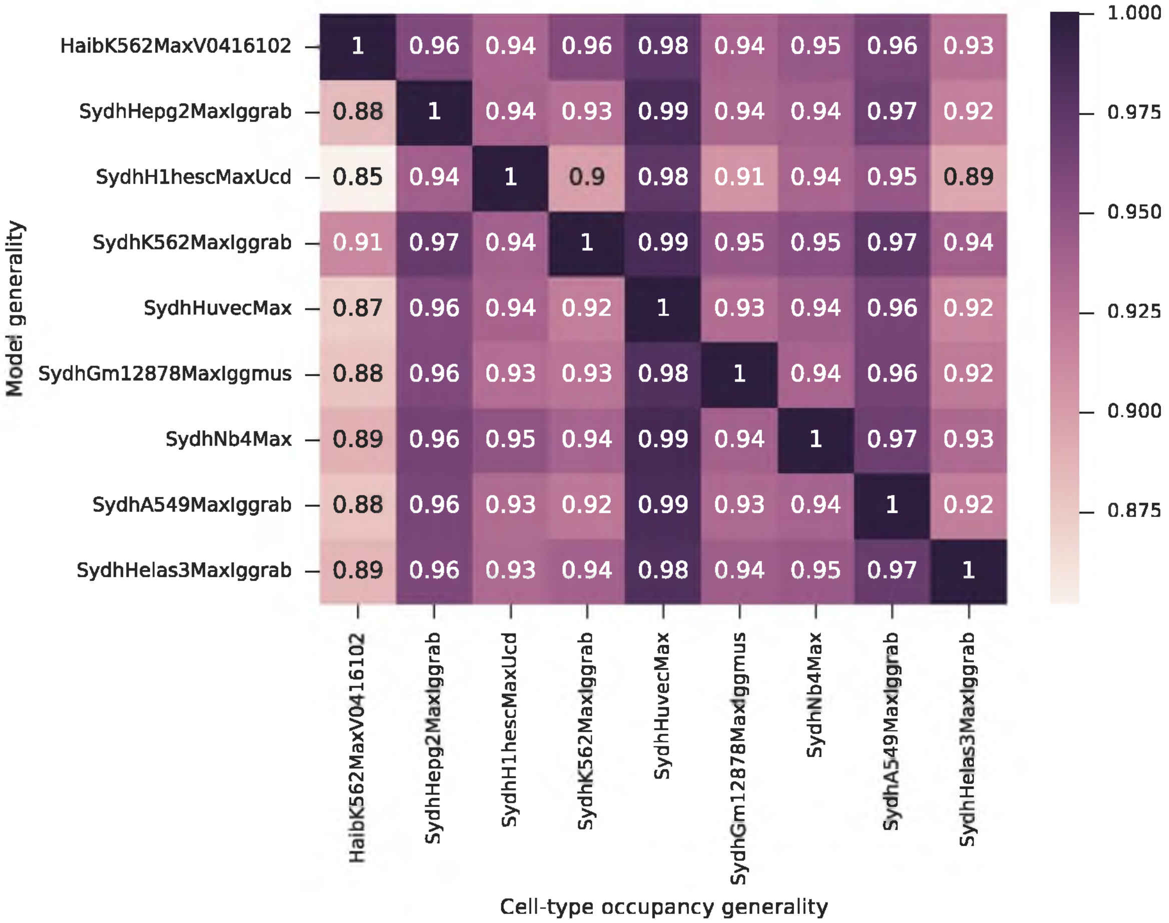


Fig. 6.14 Cell-type model generality and cell-type occupancy generality. Heatmap displays the AUC scores for a model trained on one cell type (Y-axis, model generality) and used to predict TF occupancy on another cell type (X-axis, cell-type occupancy generality) for Max TF. Cell-type model generality refers to how well a model trained on the cell type can predict TF occupancy in other cell-types; cell-type occupancy generality refers to how well TF occupancy in the cell type can be predicted by models trained on other cell types. The diagonal scores are ones because a model is trained and tested on the same cell line.

information. Advanced Bayesian models that use this information as prior may be a better fit. In fact, complex models have been proposed as necessary by some studies [140]. Nonetheless, the background correction approach showed that sticky *k*-mers [76] are differentially enriched in the genome compared with the open chromatin sites; this further demonstrates some of the weaknesses of *in vitro* approaches in addition to partly explaining the discrepancy between their *in vivo* and *in vitro* binding affinity.

The preferred and *k*-mer noise information, when used in a machine learning model do not significantly improve the performance of the model, which could explain why they were not significantly useful for correcting for E-scores. That said, they provide useful discriminatory information, especially for TFs with less predictive models. *K*-mer counts, therefore, may not best capture the *in vivo* information, and alternative usage should be sought. Some techniques have used prior binary information that a given site is located in an open chromatin site when learning *k*-mer models [217].

The *k*-mer scoring functions used influences the sequences scoring. We make a similar observation in Chapter 3, where we demonstrated a TF-specific effect of the scoring functions used. For *k*-mer models, their short length and the comprehensive nature of the scores makes the use of the *k*-mer sum occupancy inappropriate. Indeed, some studies have chosen to use only *k*-mers that pass a certain threshold for scoring, but at the expense of information loss [51]. We argue that, especially when discriminating the bound from unbound sequences, the low scoring *k*-mers (which can be negative for E-scores) are informative. Therefore, using the maximum *k*-mer score for a *k*-mer is more accurate and better captures the TF binding specificity. However, this may fail to capture the binding specificity of TFs with longer binding sites where the 8-mers are not sufficient to describe the binding site. Longer sequences can be scored in bins and the bin with the maximum bin score is the binding score for the TF.

Although we did not achieve a significant improvement in TF binding specificity decomposed into *k*-mer or PWM models, we show that, when combined in machine learning model, augmenting the *k*-mer scores with the DNase scores significantly improves the model's predictive ability. This is consistent with previous observations for PWM models [147]. What sets this study apart, however, is the use of combined data from multiple cell lines to create a model that can generalise to other cell lines.

Motivated by the above, we additionally investigated the use of conservation information to model TF binding specificity with some success. Some previous studies did not find these data useful in their models [217, 69]. We show that the site chosen to obtain conservation information directly influences the results. We show that phastCons scores should be obtained from the hit sites [114, 104] since it is designed to identify conserved blocks [170]. Therefore, predicting phastCons scores over a longer sequence is misleading. On the other hand, the phyloP scores at the whole length sequence (100bp for our case) are the most informative. PhyloP measures per nucleotide conservation information, which may explain why it is not affected by averaging scores over a longer sequence. Additionally, the evolutionary conservation of TF binding sites by structural limitation [134] spans to binding sites' flanking sequences [1]. The contribution of the shape readout (structural information) to the TF binding specificity has been established [1, 159]. This observation is not generalised. Some TFs bind in a sequence-specific manner, and for these, we would expect the conservation for the hit site alone as opposed to the whole environment to be informative.

These results are contrary to previous observation that the phastCons data were not informative when augmented with the DNase information [217] and can be explained by the use of the conservation scores over longer sequences. Another study [69] tested both the phyloP and phastCons scores in 46-way alignment of vertebrate genomes and did not find them informative – worth noting is that they were also using maximum conservation scores for the full-length ChIP-seq peaks. As already demonstrated, this could explain poor performance for phastCons scores, but not for phyloP. The use of maximum conservation scores, rather than the mean, assumes conservation at a single-nucleotide level, which is not true [134]. We speculate phyloP captures the conservation of the shape readout [1], which is within the binding environment (binding site and flanking sequences).

DNA-shape information has been shown to contribute to predictive modelling for PWM and TFFM models *in vitro* and *in vivo* [121]. However, a similar study has not been conducted for *k*-mer models. In this study, we show that shape features (ProT, HelT, Roll and MGW) improve TF binding specificity in a complementary manner. Using all the features may not contribute as much to the predictive ability and could lead to over-fitting, especially if early stopping is not employed for XGBoost models. The significant contribution of shape features to some TFs whose base readout is not well-defined supports previous results [1]. Rxra and Gr benefit greatly from additional features, and this can be attributed to their binding behaviours. Gr TFs are known to bind with multiple behaviours: direct as a heterodimer or homodimer, indirectly recruited by FOX and STAT TFs [175], and cooperative binding with NF1. Therefore obtaining a sequence specific generalised model is difficult. See Ratman et al. [153] and Starick et al. [153] for a review of Gr binding behaviours. However, the binding region, irrespective of the type of binding behaviour of Gr, would be located in open chromatin sites [175], be evolutionarily conserved and located proximally to TSS. Interestingly, Rxra is also known to bind indirectly with Sp1 or as heterodimers with retinoid acid receptor (RAR) proteins [63]. These results support our observation of TF-specific binding behaviours and contribution of the features to predictive ability. These results further show the benefit of systematic comparative analysis to elucidating TF binding specificity.

The distance of a hit site from the TSS is also informative when predicting TFBS. TF binding is responsible for regulating the level of gene expression, and it is expected for these data to be informative.

6.6 Chapter conclusions

In the chapter, we have carried out both a combinatorial approach to elucidating and modelling TF binding specificity and occupancy. We demonstrate the benefit of this approach and provide an understanding on how the contextual binding site environmental factors contribute to binding specificity. We also confirm a TF specific contribution of these features, pointing to the need for a TF-specific modelling of TF binding specificity in the future. This type of modelling will enable the model to take advantage of the features unique to the TF binding behaviour. Currently, the models generated by various algorithms do well for some TFs but not for others.

In summary, we can draw the following specific conclusions from this study:

1. A combinatorial approach improves our ability to model TF binding occupancy and specificity from *in vitro* (specifically PBM) data

2. DNase accessibility is the most predictive feature, with clustering from a variety of cell lines providing a generalizable model that still retains some specificity

3. TFs with indirect or cooperative binding benefit the most from contextual features

4. It matters how the contextual features are used; feature engineering explains some reports of lack of predictive ability

5. A *k*-mer scoring function can affect the outcome of an analysis, but maximum occupancy provides the most predictive features

6. Sticky *k*-mers are overrepresented in the whole genome compared with open chromatin sites; they are therefore spuriously overrepresented in PBM microarrays

7. XGBoost is a powerful tool for elucidating and modelling TF binding occupancy and specificity

Limitations

Our attempt to combine *in vivo* DNase and *in vitro* PBM data using *k*-mer frequency counts difference did not perform well. This failure could either be explained by the lack of strong predictive ability by these features, as confirmed by the XGbooost-based approach, or a shortfall in the approaches we applied. Therefore, how to combine PBM and DNase data remains unresolved.

Also, it is worth pointing out that the random choice of ChIP-seq cell line used in training and testing the XGB-TFBSContext algorithm could influence the results – although we did not observe an effect in our cell lines cross-validation.

Although we do come up with an XGB-based model for predicting TF binding occupancy, this was not the core purpose; therefore, the model is not fully optimised. The work to make this a fully functional optimised algorithm is ongoing.

Chapter 7

Conclusions and Future Work

> *"The conclusion of things is the good. The good is, in other words, the conclusion at which all things arrive. Let's leave doubt for tomorrow," Komatsu said. "That is the point."*
>
> –Haruki Murakami, *1Q84*

Although we are yet to achieve a complete understanding of transcription factor binding, the research community makes steps continually towards this. This thesis has presented our contribution to this quest.

7.1 Conclusions

Three themes encompassed this thesis: elucidation, modelling and evaluation of transcription factor binding specificity and occupancy. We show that despite the continued dominance of the PWMs as the models of choice for representing TF binding specificity, no standardised approach to evaluating and ranking the ever-increasing PWM models in publications and databases exists. We demonstrate the need for a systematic comparative assessment of these motifs to understand why they rank as they do. The functions used for motif scoring influence the ranks of the motifs in a TF-specific manner; we further support this conclusion through XGBoost feature importance studies. This TF-specificity points to the need, in future, for a TF-specific modelling of TF binding specificity. Majority of the algorithms do not generalise well in performance due to over-fitting TFs used for evaluation, especially when developers evaluate their algorithms on a hand-picked set of TFs.

On motif evaluation, we demonstrate the need for a standardised benchmark, and went ahead and adapted Aniba et al's criteria for a good benchmark to the motif evaluation problem. We used these criteria to create a benchmark data and MARSTools, a collection of tools for PWM motif assessment and ranking. This research further wrapped these tools in Motif Assessment and Ranking Suite (MARS), a web server for motif ranking and visualisation for comparative evaluation. We also collate motifs from a variety of databases against which users can benchmark their motifs for a given TF. These data also make a data-independent consistency-based motif evaluation approach possible. Accurate motifs are likely to be similar, but incorrect ones are likely to differ in various ways. Our tool ensures that the motifs are ranked even in the absence of benchmark data. Additionally, the ranks are not biased by the data used.

Modelling of TF binding specificity requires a combinatorial approach to unravel the regulatory code. TF binding specificity cannot be described purely by the sequence preference. However, despite the popularity of PWMs, they fall short when there is a need to model TF binding from a variety of data. We demonstrate the benefit of a combinatorial approach in XGBoost feature importance and engineering studies. These reveal that the DNase-seq data is the most predictive of TF occupancy and ultimately specificity. Furthermore, in the quest to combine the PBM and the DNase-seq data, we show that the simple statistical techniques are inadequate. Solving the same problem with machine learning allowed use to extend the model to include other data sets, improving its accuracy. From feature engineering studies, we also show that transcription factors that bind indirectly or cooperatively benefit the most from additional data to localise binding. Furthermore, this study established that DNA-shape features improve model's predictive ability in a complimentary manner, that is, a single shape feature was enough to improve the model; additional DNA-shape features mostly do not provide an additional performance gain. We show that an XGBoost model that combines multiple data sets performs significantly better than a k-mer or PWM model at predicting *in vivo* occupancy.

7.2 Limitations and future work

This study is not without limitations. First, in its current form, the MARS web-server is not as fast as we would like since we are only using a single server. We plan to scale this to use a cluster as the demand for the service increases. Also, the benchmark data in MARS is not complete; there is plentiful additional data available in various databases and publications.

However, since they are not uniformly processed or generated, they are not included in this study. There is a need to normalise the data from various sources and expand the available benchmarks.

The XGBoost model developed and used in this study to elucidate the contribution of different features to TF binding occupancy is not fully optimised. The current form is designed for the elucidation and not modelling. We intend to optimise this algorithm for modelling, then benchmark it against other approaches with similar functionality.

Although we established the need for advanced models to capture the complex nature of TF binding, this conclusion is limited to *k*-mer and PWM models. We did not test some of the recent models like Slim and TFFM, which consider nucleotide interdependencies and flexible length models, respectively. In future, we intend to include these models in our study and evaluate how they help improve TF binding specificity modelling.

References

[1] Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., Rohs, R., and Mann, R. S. (2015). Deconvolving the recognition of DNA shape from sequence. *Cell*, 161(2):307–318. (Cited on pages 2, 3, 27, 108, 128, and 135.)

[2] Agius, P., Arvey, A., Chang, W., Noble, W. S., and Leslie, C. (2010). High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Computational Biology*, 6(9):12. (Cited on pages 29, 30, 40, and 43.)

[3] Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838. (Cited on pages 21, 32, and 114.)

[4] Andrilenas, K. K., Penvose, A., and Siggers, T. (2015). Using protein-binding microarrays to study transcription factor specificity: homologs, isoforms and complexes. *Briefings in Functional Genomics*, 14(1):17–29. (Cited on page 12.)

[5] Angermueller, C., Pärnamaa, T., Parts, L., and Oliver, S. (2016). Deep Learning for Computational Biology. *Molecular Systems Biology*, 12(7):878. (Cited on page 32.)

[6] Aniba, M. R., Poch, O., and Thompson, J. D. (2010). Issues in bioinformatics benchmarking: The case study of multiple sequence alignment. *Nucleic Acids Research*, 38(21):7353–7363. (Cited on pages 67 and 68.)

[7] Annala, M., Laurila, K., Lähdesmäki, H., and Nykter, M. (2011). A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PLoS ONE*, 6(5):1–13. (Cited on pages 13 and 24.)

[8] Arvey, A., Agius, P., Noble, W. S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, 22(9):1723–34. (Cited on pages 30 and 107.)

[9] Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. a., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723. (Cited on pages 40 and 48.)

[10] Bahrami, S., Ehsani, R., and Drabløs, F. (2015). A property-based analysis of human transcription factors. *BMC Research Notes*, 8(1):82. (Cited on page 38.)

[11] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2):W202–W208. (Cited on pages 42 and 75.)

[12] Bailey, T. L. and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17):1–10. (Cited on pages 39, 50, 62, and 96.)

[13] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34:W369–W373. (Cited on pages 22 and 25.)

[14] Bauer, A. L., Hlavacek, W. S., Unkefer, P. J., and Mu, F. (2010). Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS Computational Biology*, 6(11):e1001007. (Cited on page 26.)

[15] Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–35. (Cited on pages 2, 15, 16, 107, 112, and 113.)

[16] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242. (Cited on page 11.)

[17] Bhagwat, A. S. and Vakoc, C. R. (2015). Targeting Transcription Factors in Cancer. *Trends in Cancer*, 1(1):53–65. (Cited on page 2.)

[18] Blatti, C. and Sinha, S. (2014). Motif enrichment tool. *Nucleic Acids Research*, 42(W1). (Cited on page 39.)

[19] Blattler, A., Yao, L., Wang, Y., Ye, Z., Jin, V. X., and Farnham, P. J. (2013). ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. *Epigenetics & Chromatin*, 6(1):13. (Cited on page 102.)

[20] Boeva, V. (2016). Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Frontiers in Genetics*, 7(February). (Cited on pages 14, 18, 26, and 36.)

[21] Boeva, V., Clément, J., Régnier, M., Roytberg, M. A. A., and Makeev, V. J. J. (2007). Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in computational annotation of cis-regulatory modules. *Algorithms for molecular biology*, 2:13. (Cited on page 26.)

[22] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*, 132(2):311–322. (Cited on page 19.)

[23] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. (Cited on page 31.)

[24] Buck, M. J. and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360. (Cited on page 18.)

[25] Cai, Y. H. and Huang, H. (2012). Advances in the study of protein-DNA interaction. 43(3):1141–1146. (Cited on pages 18 and 19.)

[26] Carlson, C. D., Warren, C. L., Hauschild, K. E., Ozers, M. S., Qadir, N., Bhimsaria, D., Lee, Y., Cerrina, F., and Ansari, A. Z. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proceedings of the National Academy of Sciences of the United States of America*, 107(10):4544–9. (Cited on page 24.)

[27] Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21(13):2933–42. (Cited on page 26.)

[28] Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *22nd SIGKDD Conference on Knowledge Discovery and Data Mining*. (Cited on pages 3, 31, 109, 114, and 117.)

[29] Chen, X., Hughes, T. R., and Morris, Q. (2007). RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics*, 23(13):72–79. (Cited on pages 20, 40, 41, 43, 46, and 75.)

[30] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells (Supplemental Info). *Cell*, 133(6):1106–17. (Cited on page 72.)

[31] Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., Davis, C. A., Carninci, P., Lassman, T., Gingeras, T. R., Guigó, R., Birney, E., Weng, Z., Snyder, M., and Gerstein, M. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, 22(9):1658–1667. (Cited on page 12.)

[32] Cheng, Q., Kazemian, M., Pham, H., Blatti, C., Celniker, S. E., Wolfe, S. A., Brodsky, M. H., and Sinha, S. (2013). Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy. *PLoS Genetics*, 9(8). (Cited on pages 27 and 40.)

[33] Chiu, T. P., Yang, L., Zhou, T., Main, B. J., Parker, S. C. J., Nuzhdin, S. V., Tullius, T. D., and Rohs, R. (2015). GBshape: A genome browser database for DNA shape annotations. *Nucleic Acids Research*, 43(D1):D103–D109. (Cited on pages 17 and 115.)

[34] Clarke, N. D. and Granek, J. A. (2003). Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics*, 19(2):212–218. (Cited on pages 43, 44, and 62.)

[35] Contreras-Moreira, B. (2010). 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Research*, 38(Database):D91–D97. (Cited on pages 72 and 104.)

[36] Cuellar-Partida, G., Buske, F. a., McLeay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, 28(1):56–62. (Cited on pages 106, 108, and 132.)

[37] Dabrowski, M., Dojer, N., Krystkowiak, I., Kaminska, B., and Wilczynski, B. (2015). Optimally choosing PWM motif databases and sequence scanning approaches based on ChIP-seq data. *BMC Bioinformatics*, 16(1). (Cited on page 26.)

[38] Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–4. (Cited on pages 33 and 46.)

[39] Dassi, E. and Quattrone, A. (2015). DynaMIT: the dynamic motif integration toolkit. *Nucleic Acids Research*, page gkv807. (Cited on page 69.)

[40] Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine learning – ICML'06*, pages 233–240. (Cited on page 43.)

[41] Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment on transcription factor binding across diverse protein families. *Genome Research*, pages 1268–1280. (Cited on pages 3 and 12.)

[42] Dror, I., Rohs, R., and Mandel-Gutfreund, Y. (2016). How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays*, pages 1–8. (Cited on pages 2, 21, 26, 63, 104, and 107.)

[43] Eggeling, R., Roos, T., Myllymäki, P., and Grosse, I. (2015). Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, 16(1):375. (Cited on pages 23 and 39.)

[44] Ernst, J., Beg, Q. K., Kay, K. A., Balázsi, G., Oltvai, Z. N., and Bar-Joseph, Z. (2008). A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*. *PLoS Computational Biology*, 4(3). (Cited on page 29.)

[45] Feingold, E., Good, P., Guyer, M., and Kamholz, S. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 9305(301). (Cited on pages 36, 46, and 71.)

[46] Foat, B. C., Morozov, A. V., and Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):141–149. (Cited on page 41.)

[47] Freund, Y. and Schapire, R. R. E. (1996). Experiments with a New Boosting Algorithm. *International Conference on Machine Learning*, pages 148–156. (Cited on page 117.)

[48] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232. (Cited on page 117.)

[49] Furey, T. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12):840–852. (Cited on page 19.)

[50] Garcia, F., Lopez, F. J., Cano, C., and Blanco, A. (2009). FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC Bioinformatics*, 10:224. (Cited on pages 77, 93, 94, and 104.)

[51] Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A. (2014). Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Computational Biology*, 10(7). (Cited on pages 30 and 134.)

[52] Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L. (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports*, 3(4):1093–1104. (Cited on pages 3, 12, 15, 17, and 63.)

[53] Granek, J. A. and Clarke, N. D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome biology*, 6(10):R87. (Cited on pages 40 and 75.)

[54] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–8. (Cited on pages 26 and 39.)

[55] Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197. (Cited on pages 21, 38, and 69.)

[56] Gupta, S. and Stamatoyannopoulos, J. (2007). Quantifying similarity between motifs. *Genome Biology*, 8(24). (Cited on page 77.)

[57] Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59. (Cited on page 72.)

[58] Håndstad, T., Rye, M., Mocnik, R., Drabløs, F., and Sætrom, P. (2012). Cell-type specificity of ChIP-predicted transcription factor binding sites. *BMC Genomics*, 13(1):372. (Cited on pages 2, 18, 26, 27, 30, and 51.)

[59] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36. (Cited on page 43.)

[60] Hannenhalli, S. (2008). Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, 24(11):1325–31. (Cited on page 68.)

[61] Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T., and Sussman, J. L. (2013). JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Israel Journal of Chemistry*, 53(3-4):207–216. (Cited on page 11.)

[62] Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104. (Cited on pages 20 and 38.)

[63] He, Y., Tsuei, J., and Wan, Y.-J. Y. (2014). Biological functional annotation of retinoic acid alpha and beta in mouse liver based on genome-wide binding. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 307(2):G205—-G218. (Cited on page 136.)

[64] Heger, A. (2009). pysam: Python interface for the SAM/BAM sequence alignment and mapping format. https://github.com/pysam-developers/pysam. (Cited on page 33.)

[65] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589. (Cited on page 72.)

[66] Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5):5947. (Cited on page 32.)

[67] Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–6. (Cited on page 29.)

[68] Holloway, D. T., Kon, M., and DeLisi, C. (2007). Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Systems and Synthetic Biology*, 1(1):25–46. (Cited on pages 26 and 30.)

[69] Hombach, D., Schwarz, J. M., Robinson, P. N., Schuelke, M., and Seelow, D. (2016). A systematic, large-scale comparison of transcription factor binding site models. *BMC Genomics*, 17(1):388. (Cited on pages 18, 128, 134, and 135.)

[70] Hu, J., Li, B., and Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33(15):4899–4913. (Cited on pages 14, 20, and 38.)

[71] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95. (Cited on pages 34 and 80.)

[72] Iantorno, S., Gori, K., Goldman, N., Gil, M., and Dessimoz, C. (2014). Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods in Molecular Biology*, 1079:59–73. (Cited on pages 67, 68, 77, and 91.)

[73] Janin, J., Henrick, K., and Moult, J. (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Structure, Function and Genetics*, 9:2–9. (Cited on page 68.)

[74] Jankowski, A., Tiuryn, J., and Prabhakar, S. (2016). Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics*, 32(16):2419–2426. (Cited on page 108.)

[75] Jayaram, N., Usvyat, D., and R. Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, (i):1–12. (Cited on page 26.)

[76] Jiang, B., Liu, J. S., and Bulyk, M. L. (2013). Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, 29(11):1390–8. (Cited on pages 2, 15, 108, 112, 113, 114, 119, and 134.)

[77] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–502. (Cited on pages 2, 18, 36, 107, and 114.)

[78] Jolma, A. and Taipale, J. (2011). Methods for analysis of transcription factor DNA-binding specificity in vitro. *Sub-Cellular Biochemistry*, 52:155–173. (Cited on pages 14, 15, and 18.)

[79] Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339. (Cited on pages 13, 16, 17, 26, 68, 72, 96, 104, and 115.)

[80] Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–8. (Cited on page 107.)

[81] Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. (2007). Functional Specificity of a Hox Protein Mediated by the Recognition of Minor Groove Structure. *Cell*, 131(3):530–543. (Cited on page 108.)

[82] Kähärä, J. and Lähdesmäki, H. (2013). Evaluating a linear k-mer model for protein-DNA interactions using high-throughput SELEX data. *BMC Bioinformatics*, 14 Suppl 1(Suppl 10):S2. (Cited on page 24.)

[83] Kamachi, Y. and Kondoh, H. (2013). Sox proteins: regulators of cell fate specification and differentiation. *Development*, 140(20):4129–44. (Cited on page 102.)

[84] Kaplan, T., Li, X.-Y., Sabo, P. J., Thomas, S., Stamatoyannopoulos, J. A., Biggin, M. D., and Eisen, M. B. (2011). Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *PLoS Genetics*, 7(2):e1001290. (Cited on page 27.)

[85] Karczewski, K. J., Tatonetti, N. P., Landt, S. G., Yang, X., Slifer, T., Altman, R. B., and Snyder, M. (2011). Cooperative transcription factor associations discovered using regulatory variation. *Proceedings of the National Academy of Sciences of the United States of America*, 108(32):13353–13358. (Cited on page 12.)

[86] Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue):D493–6. (Cited on page 116.)

[87] Keilwagen, J. and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, 43(18). (Cited on pages 3, 15, 23, 24, and 25.)

[88] Kel, A., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2003). MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research*, 31(13):3576–3579. (Cited on page 26.)

[89] Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999. (Cited on page 32.)

[90] Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207. (Cited on page 33.)

[91] Kheradpour, P. and Kellis, M. (2013). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–87. (Cited on pages 39 and 72.)

[92] Kibet, C. K. and Machanick, P. (2016). Transcription factor motif quality assessment requires systematic comparative analysis [version 2; referees: 2 approved]. *F1000Research*, 4(ISCB Comm J):1429. (Cited on pages 5, 14, 68, and 69.)

[93] Kim, N.-K., Tharakaraman, K., Mariño-Ramírez, L., and Spouge, J. L. (2008). Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics*, 9:262. (Cited on pages 3 and 108.)

[94] Klepper, K., Sandve, G. K., Abul, O., Johansen, J., and Drabløs, F. (2008). Assessment of composite motif discovery methods. *BMC Bioinformatics*, 9:123. (Cited on page 38.)

[95] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. (2016). Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press. (Cited on page 34.)

[96] Kondoh, H. and Kamachi, Y. (2010). SOX-partner code for cell specification: Regulatory target selection and underlying molecular mechanisms. *International Journal of Biochemistry and Cell Biology*, 42(3):391–399. (Cited on page 102.)

[97] Kubosaki, A., Tomaru, Y., Tagami, M., Arner, E., Miura, H., Suzuki, T., Suzuki, M., Suzuki, H., and Hayashizaki, Y. (2009). Genome-wide investigation of in vivo EGR-1 binding sites in monocytic differentiation. *Genome Biology*, 10(4):R41. (Cited on page 58.)

[98] Kulakovskiy, I., Levitsky, V., Oshchepkov, D., Bryzgalov, L., Vorontsov, I., and Makeev, V. (2013). From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *Journal of Bioinformatics and Computational Biology*, 11(1):1340004. (Cited on pages 20, 23, and 72.)

[99] Kumar, S. and Bucher, P. (2016). Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features. *BMC Bioinformatics*, 17(S1):4. (Cited on pages 3, 27, and 108.)

[100] Kundaje, A., Lianoglou, S., Li, X., Quigley, D., Arias, M., Wiggins, C. H., Zhang, L., and Leslie, C. (2007). Learning regulatory programs that accurately predict differential expression with MEDUSA. *Annals of the New York Academy of Sciences*, 1115:178–202. (Cited on page 31.)

[101] Lassmann, T. and Sonnhammer, E. L. L. (2005). Automatic assessment of alignment quality. *Nucleic Acids Research*, 33(22):7120–7128. (Cited on pages 68 and 77.)

[102] Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., Sabo, P. J., Lu, Y., Rohs, R., Stamatoyannopoulos, J. A., and Bussemaker, H. J. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16):6376–81. (Cited on page 108.)

[103] Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., and Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, 47(8):955–961. (Cited on pages 3, 30, and 108.)

[104] Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Research*, 21(12):2167–2180. (Cited on pages 30 and 135.)

[105] Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251. (Cited on page 2.)

[106] Lesluyes, T., Johnson, J., Machanick, P., and Bailey, T. L. (2014). Differential motif enrichment analysis of paired ChIP-seq experiments. *BMC Genomics*, 15(1):752. (Cited on page 77.)

[107] Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics*, 15(7):453–68. (Cited on page 107.)

[108] Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A. C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R., and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, 25(7):1018–1029. (Cited on page 104.)

[109] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. *Bioinformatics*, 25(16):1–2. (Cited on page 33.)

[110] Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332. (Cited on pages 28 and 29.)

[111] Lihu, A. and Holban, t. (2015). A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings in Bioinformatics*, 16(6):964–73. (Cited on page 20.)

[112] Liseron-Monfils, C., Lewis, T., Ashlock, D., McNicholas, P. D., Fauteux, F., Strömvik, M., and Raizada, M. N. (2013). Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas. *BMC Plant Biology*, 13(1):1–17. (Cited on pages 20, 44, and 62.)

[113] Luo, K. and Hartemink, A. J. (2013). Using DNase digestion data to accurately identify transcription factor binding sites. *Pacific Symposium on Biocomputing*, (1):80–91. (Cited on pages 19, 106, 108, and 132.)

[114] Ma, W. and Wong, W. H. (2011). The analysis of ChIP-seq data. *Methods in Enzymology*, 497:51–73. (Cited on page 135.)

[115] Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12):1696–1697. (Cited on pages 20 and 34.)

[116] Macisaac, K. D., Gordon, D. B., Nekludova, L., Odom, D. T., Schreiber, J., Gifford, D. K., Young, R. a., and Fraenkel, E. (2006). A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics*, 22(4):423–9. (Cited on page 72.)

[117] Maienschein-Cline, M., Dinner, A. R., Hlavacek, W. S., and Mu, F. (2012). Improved predictions of transcription factor binding sites using physicochemical features of DNA. *Nucleic Acids Research*, 40(22):e175. (Cited on page 27.)

[118] Martins, L., Siepel, A., Lis, J. T., Guertin, M. J., and Martins, A. L. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genetics*, 8(3):e1002610. (Cited on page 72.)

[119] Mathelier, A., Shi, W., and Wasserman, W. W. (2015). Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76. (Cited on page 12.)

[120] Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS Computational Biology*, 9(9):e1003214. (Cited on pages 3, 13, 21, 23, 24, and 25.)

[121] Mathelier, A., Xin, B., Chiu, T.-P., Yang, L., Rohs, R., and Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*, pages 1–9. (Cited on pages 3, 23, 27, 108, and 135.)

[122] 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. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(Database issue):D142—-7. (Cited on pages 15, 22, and 72.)

[123] Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378. (Cited on pages 23 and 26.)

[124] Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms: From machine learning to statistical modelling. *Methods of Information in Medicine*, 53(6):419–427. (Cited on page 31.)

[125] McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt, S. and Millman, J., editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56. (Cited on pages 34 and 80.)

[126] McLeay, R. C. and Bailey, T. L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, 11:165. (Cited on page 39.)

[127] Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J., and Van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Research*, 39(3):808–824. (Cited on page 38.)

[128] Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill. (Cited on page 28.)

[129] Moretti, R., Fleishman, S. J., Agius, R., Torchala, M., Bates, P. a., Kastritis, P. L., Rodrigues, J. P. G. L. M., Trellet, M., Bonvin, A. M. J. J., Cui, M., Rooman, M., Gillis, D., Dehouck, Y., Moal, I., Romero-Durana, M., Perez-Cano, L., Pallara, C., Jimenez, B., Fernandez-Recio, J., Flores, S., Pacella, M., Praneeth Kilambi, K., Gray, J. J., Popov, P., Grudinin, S., Esquivel-Rodríguez, J., Kihara, D., Zhao, N., Korkin, D., Zhu, X., Demerdash, O. N. a., Mitchell, J. C., Kanamori, E., Tsuchiya, Y., Nakamura, H., Lee, H., Park, H., Seok, C., Sarmiento, J., Liang, S., Teraguchi, S., Standley, D. M., Shimoyama, H., Terashi, G., Takeda-Shitaka, M., Iwadate, M., Umeyama, H., Beglov, D., Hall, D. R., Kozakov, D., Vajda, S., Pierce, B. G., Hwang, H., Vreven, T., Weng, Z., Huang, Y., Li, H., Yang, X., Ji, X., Liu, S., Xiao, Y., Zacharias, M., Qin, S., Zhou, H. X., Huang, S. Y., Zou, X., Velankar, S., Janin, J., Wodak, S. J., and Baker, D. (2013). Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins: Structure, Function and Bioinformatics*, 81(11):1980–1987. (Cited on page 68.)

[130] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2):1–6. (Cited on page 68.)

[131] Narlikar, L. (2013). MuMoD: a Bayesian approach to detect multiple modes of protein-DNA binding from genome-wide ChIP data. *Nucleic Acids Research*, 41(1):21–32. (Cited on pages 38 and 69.)

[132] Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, 150(6):1274–86. (Cited on pages 18, 106, 108, 115, and 132.)

[133] Newburger, D. E. and Bulyk, M. L. (2009). UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37(SUPPL. 1):77–82. (Cited on pages 15, 22, 23, 36, 48, 71, and 72.)

[134] Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. M., and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4:1–20. (Cited on pages 3, 17, 26, 108, 115, 128, and 135.)

[135] Oda, T., Tian, T., Inoue, M., Ikeda, J., Qiu, Y., Okumura, M., Aozasa, K., and Morii, E. (2009). Tumorigenic role of orphan nuclear receptor NR0B1 in lung adenocarcinoma. *Am J Pathol*, 175(3):1235–1245. (Cited on page 51.)

[136] O'Malley, R., Huang, S.-s., Song, L., Lewsey, M., Bartlett, A., Nery, J., Galli, M., Gallavotti, A., and Ecker, J. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, 165(5):1280–1292. (Cited on page 16.)

[137] Orenstein, Y., Linhart, C., and Shamir, R. (2012). Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PloS ONE*, 7(9):e46145. (Cited on pages 14, 20, 22, 36, 38, 39, 40, 41, 43, 44, 63, 67, 69, and 75.)

[138] Orenstein, Y., Mick, E., and Shamir, R. (2013). RAP: accurate and fast motif finding based on protein-binding microarray data. *Journal of Computational Biology*, 20(5):375–82. (Cited on pages 20 and 63.)

[139] Orenstein, Y. and Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, 42(8):e63. (Cited on pages 16, 39, and 108.)

[140] Orenstein, Y. and Shamir, R. (2016). Modeling protein–DNA binding via high-throughput in vitro technologies. *Briefings in Functional Genomics*, page elw030. (Cited on pages 2, 14, 15, 16, 21, 24, 25, 36, 58, 106, 107, 108, 115, and 134.)

[141] Osada, R., Zaslavsky, E., and Singh, M. (2004). Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–25. (Cited on page 21.)

[142] Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. (2007). SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Research*, 35(Database):D127–D131. (Cited on page 72.)

[143] Park, Y. and Kellis, M. (2015). Deep learning for regulatory genomics. *Nature Biotechnology*, 33(8):825–826. (Cited on pages 21 and 32.)

[144] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. (Cited on pages 34 and 117.)

[145] Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29. (Cited on page 34.)

[146] Philippakis, A. A., Qureshi, A. M., Berger, M. F., and Bulyk, M. L. (2008). Design of compact, universal DNA microarrays for protein binding microarray experiments. *Journal of Computational Biology*, 15(7):655–665. (Cited on page 15.)

[147] Pique-regi, R., Degner, J. F., Pai, A. A., Boyle, A. P., Song, L., Lee, B.-k., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–55. (Cited on pages 18, 106, 107, 108, 132, and 134.)

[148] Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121. (Cited on pages 18 and 115.)

[149] Pugh, B. F. and Gilmour, D. S. (2001). Genome-wide analysis of protein-DNA interactions in living cells. *Genome Biology*, 2(4):REVIEWS1013. (Cited on page 18.)

[150] Pujato, M., Kieken, F., Skiles, A. A., Tapinos, N., and Fiser, A. (2014). Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic Acids Research*, 42(22):13500–12. (Cited on page 72.)

[151] Quest, D., Dempsey, K., Shafiullah, M., Bastola, D., and Ali, H. (2008). A parallel architecture for regulatory motif algorithm assessment. In *IPDPS Miami 2008 - Proceedings of the 22nd IEEE International Parallel and Distributed Processing Symposium, Program and CD-ROM*. (Cited on pages 38 and 69.)

[152] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842. (Cited on pages 33, 46, and 71.)

[153] Ratman, D., Vanden Berghe, W., Dejager, L., Libert, C., Tavernier, J., Beck, I. M., and De Bosscher, K. (2013). How glucocorticoid receptors modulate the activity of other transcription factors: A scope beyond tethering. *Molecular and Cellular Endocrinology*, 380(1-2):41–54. (Cited on page 135.)

[154] Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309. (Cited on page 18.)

[155] Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419. (Cited on page 18.)

[156] Rhee, S. Y., Parker, J. E., and Mockler, T. C. (2016). A glimpse into the future of genome-enabled plant biology from the shores of Cold Spring Harbor. *Genome Biology*, 17(1):3. (Cited on page 16.)

[157] Riley, T. R., Lazarovici, A., Mann, R. S., and Bussemaker, H. J. (2015). Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE. *eLife*, 4:e06397. (Cited on page 25.)

[158] Rodrigues, M. R. and Luck, M. (2006). Evaluating dynamic services in bioinformatics. *Cooperative Information Agents X*, 4149:183–197. Lecture Notes in Artificial Intelligence. (Cited on page 67.)

[159] Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–53. (Cited on pages 12, 17, 108, and 135.)

[160] Rube, H. T., Lee, W., Hejna, M., Chen, H., Yasui, D. H., Hess, J. F., LaSalle, J. M., Song, J. S., and Gong, Q. (2016). Sequence features accurately predict genome-wide MeCP2 binding in vivo. *Nature Communications*, 7:11025. (Cited on page 31.)

[161] Ryan, D., Grüning, B., and Ramirez, F. (2016). pyBigWig 0.2.4. (Cited on pages 33 and 115.)

[162] Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. (2010). Genecards version 3: the human gene integrator. *Database*, 2010. (Cited on page 71.)

[163] Sandve, G. K., Abul, O., Walseng, V., and Drabløs, F. (2007). Improved benchmarks for computational motif discovery. *BMC Bioinformatics*, 8:193. (Cited on pages 38 and 68.)

[164] Sandve, G. K. and Drabløs, F. (2006). A survey of motif discovery methods in an integrated framework. *Biology Direct*, 1:11. (Cited on page 38.)

[165] Santolini, M., Mora, T., and Hakim, V. (2014). A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites. *PLoS ONE*, 9(6):e99015. (Cited on pages 3, 23, and 25.)

[166] Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–40. (Cited on page 17.)

[167] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100. (Cited on page 24.)

[168] Schütz, F. and Delorenzi, M. (2008). MAMOT: hidden Markov modeling tool. *Bioinformatics*, 24(11):1399–400. (Cited on page 24.)

[169] Sharov, A. L. A. L. A. and Ko, M. I. S. H. (2009). Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA research*, 16(5):261–73. (Cited on page 38.)

[170] Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050. (Cited on pages 17, 18, 115, and 135.)

[171] Siggers, T. and Gordân, R. (2013). Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Research*, pages 1–13. (Cited on pages 12 and 107.)

[172] Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9):381–399. (Cited on pages 2, 12, 15, 17, 18, 19, 21, 36, and 108.)

[173] Song, L. and Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb.prot5384. (Cited on pages 2 and 107.)

[174] Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. a., Simon, J. M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N. D., Birney, E., Iyer, V. R., Crawford, G. E., Lieb, J. D., and Furey, T. S. (2011). Open chromatin

defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, 21(10):1757–67. (Cited on pages 12, 19, and 107.)

[175] Starick, S. R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M. I., Chung, H.-R., Vingron, M., Thomas-Chollier, M., and Meijsing, S. H. (2015). ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome Research*, 25(6):825–35. (Cited on pages 135 and 136.)

[176] Stegmaier, P., Kel, A. E., and Wingender, E. (2004). Systematic DNA-binding domain classification of transcription factors. *Genome Informatics*, 15(2):276–86. (Cited on page 12.)

[177] Stewart, A. J. and Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(November):973–985. (Cited on page 63.)

[178] Stormo, G. and Schneider, T. (1982). Use of the 'Perceptron'algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9):2997–3011. (Cited on pages 3 and 21.)

[179] Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23. (Cited on pages 21 and 36.)

[180] Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quantitative Biology*, 1(2):115–130. (Cited on pages 3, 21, and 22.)

[181] Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. *Nature Reviews Genetics*, 11(11):751–60. (Cited on pages 13, 14, 15, and 22.)

[182] Szent-Gyorgyi, A. (1949). Muscle Research. *Scientific American*, 22(180):6. (Cited on page 8.)

[183] Takahashi, K., Hayashi, N., Shimokawa, T., Umehara, N., Kaminogawa, S., and Ra, C. (2008). Cooperative regulation of Fc receptor gamma-chain gene expression by multiple transcription factors, including Sp1, GABP, and Elf-1. *Journal of Biological Chemistry*, 283(22):15134–41. (Cited on page 50.)

[184] Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, 40(4):e31. (Cited on page 39.)

[185] Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82. (Cited on pages 110 and 123.)

[186] Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. a., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–44. (Cited on pages 14, 20, 36, 37, 38, 68, and 101.)

[187] Tran, N. T. L. and Huang, C.-H. (2014). A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biology Direct*, 9(1):4. (Cited on pages 20 and 36.)

[188] Tsai, Z. T. Y., Shiu, S. H., and Tsai, H. K. (2015). Contribution of Sequence Motif, Chromatin State, and DNA Structure Features to Predictive Models of Transcription Factor Binding in Yeast. *PLoS Computational Biology*, 11(8). (Cited on page 31.)

[189] van Heeringen, S. J. and Veenstra, G. J. C. (2011). GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics*, 27(2):270–271. (Cited on pages 20, 33, 44, 47, 69, 93, and 95.)

[190] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263. (Cited on page 2.)

[191] Villard, J. (2004). Transcription regulation and human diseases. *Swiss Medical Weekly*, 134(39-40):571–9. (Cited on page 2.)

[192] Wang, J., Lu, J., Gu, G., and Liu, Y. (2011). In vitro DNA-binding profile of transcription factors: Methods and new insights. *Journal of Endocrinology*, 210(1):15–27. (Cited on page 15.)

[193] Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., and Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812. (Cited on pages 18, 20, 39, 72, and 114.)

[194] Wang, L., Jensen, S., and Hannenhalli, S. (2006). An interaction-dependent model for transcription factor binding. *Systems Biology and Regulatory Genomics*, pages 225–234. (Cited on pages 3 and 23.)

[195] Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., Miles, A., Augspurger, T., Yarkoni, T., Evans, C., Wehner, D., Rocher, L., Megies, T., Coelho, L. P., Ziegler, E., Hoppe, T., Seabold, S., Pascual, S., Cloud, P., Koskinen, M., Hausler, C., kjemmett, Milajevs, D., Qalieh, A., Allan, D., and Meyer, K. (2015). Seaborn: v0.6.0 (June 2015). doi:10.5281/zenodo.19108. (Cited on pages 34 and 80.)

[196] Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H. G., Ukkonen, E., Hughes, T. R., Bulyk, M. L., and Taipale, J. (2010). Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *The EMBO Journal*, 29(13):2147–60. (Cited on pages 12 and 72.)

[197] Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., and Hughes, T. R. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–34. (Cited on pages 14, 20, 22, 23, 25, 36, 39, 40, 42, 43, 44, 63, and 69.)

[198] Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R., and Hughes, T. R. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443. (Cited on pages 23 and 72.)

[199] Wilbanks, E. G. and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PloS ONE*, 5(7):e11471. (Cited on page 62.)

[200] Wilson, E. O. (1985). The biological diversity crisis: A challenge to science. *Issues in Science and Technology*, 2(1):20–29. (Cited on page 1.)

[201] Wingender, E. (1997). Classification scheme of eukaryotic transcription factors. *Molecular Biology*, 31(4):584–600. (Cited on page 12.)

[202] Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, 41(Database issue):D165–70. (Cited on pages 12, 71, 73, and 74.)

[203] Worsley Hunt, R., Mathelier, A., Del Peso, L., and Wasserman, W. W. (2014). Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, 15(1):472. (Cited on page 62.)

[204] Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research*, 42(Database issue):D148–55. (Cited on page 17.)

[205] Ying, L., Haiyan, H., and Li, C. (2011). Prediction of Transcriptional Regulatory Networks for Retinal Development. In Lopes, H., editor, *Computational Biology and Applied Bioinformatics*, pages 357–374. Intech. (Cited on page 10.)

[206] Yip, K. Y., Cheng, C., and Gerstein, M. (2013). Machine learning and genome annotation: a match meant to be? *Genome Biology*, 14(5):205. (Cited on page 28.)

[207] Zambelli, F., Pesole, G., and Pavesi, G. (2013a). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–37. (Cited on pages 20, 36, 39, and 68.)

[208] Zambelli, F., Pesole, G., and Pavesi, G. (2013b). Pscanchip: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from chip-seq experiments. *Nucleic Acids Research*, 41(Web Server issue):W535–W543. (Cited on pages 18 and 39.)

[209] Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, 32(12):i121–i127. (Cited on page 31.)

[210] Zhang, M. O. and Marr, T. G. (1993). A weight array method for splicing signal analysis. *Bioinformatics*, 9(5):499–509. (Cited on page 23.)

[211] Zhang, Y., He, Y., Zheng, G., and Wei, C. (2015). MOST+: A de novo motif finding approach combining genomic sequence and heterogeneous genome-wide signatures. *BMC Genomics*, 16 Suppl 7(Suppl 7):S13. (Cited on pages 38 and 64.)

[212] Zhang, Z., Chang, C. W., Hugo, W., Cheung, E., and Sung, W. K. (2012). Simultaneously learning DNA motif along with Its position and sequence rank preferences through EM algorithm. In *Research in Computational Molecular Biology: 16th Annual International Conference, RECOMB 2012, Barcelona, Spain, April 21-24, 2012. Proceedings*, volume 7262 LNBI, pages 355–370, Berlin, Heidelberg. Springer Berlin Heidelberg. (Cited on page 38.)

[213] Zhao, Y., Granas, D., and Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5(12):e1000590. (Cited on pages 20, 41, and 75.)

[214] Zhao, Y., Ruan, S., Pandey, M., and Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, 191(3):781–90. (Cited on page 23.)

[215] Zhao, Y. and Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, 29(6):480–483. (Cited on pages 3, 23, 25, 27, and 72.)

[216] Zheng, Y., Li, X., and Hu, H. (2014). Comprehensive discovery of DNA motifs in 349 human cells and tissues reveals new features of motifs. *Nucleic Acids Research*, 43:74–83. (Cited on pages 38 and 69.)

[217] Zhong, S., He, X., and Bar-Joseph, Z. (2013). Predicting tissue specific transcription factor binding sites. *BMC Genomics*, 14:796. (Cited on pages 18, 19, 39, 40, 41, 42, 43, 75, 106, 107, 115, 128, 134, and 135.)

[218] Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916. (Cited on page 23.)

[219] Zhou, Q. and Liu, J. S. (2008). Extracting sequence features to predict protein-DNA interactions: A comparative study. *Nucleic Acids Research*, 36(12):4137–4148. (Cited on pages 21 and 29.)

[220] Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 112(15):4654–4659. (Cited on pages 27 and 32.)

[221] Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research*, 41(Web Server issue):56–62. (Cited on pages 17 and 115.)

Appendix A

Additional Figures and Tables

Some of the figures that could not be included in the main thesis are included here for completeness. For each chapter, complete details including additional figures are provided in the respective IPython notebooks.

Additional data and scripts

Supplementary data that accompany this thesis can be accessed from Github[1]. These include:

- Raw motifs from various databases added to MARS database
- Scripts used to convert the motifs from various PWM formats to MEME format
- A list of raw ChIP-seq data used throughout this thesis
- Details of length and information content of all the motifs in the MARS database

[1] https://github.com/kipkurui/MARS_Evaluation/tree/master/Additional_Data

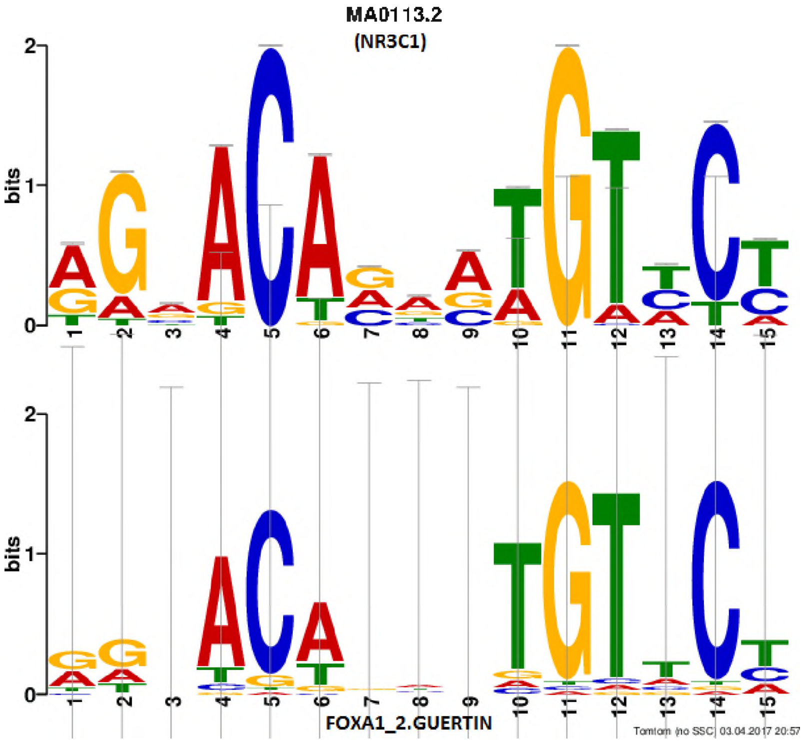


Fig. A.1 FOXA1_2_GUERTIN resembles NR3C1. TomTom motif comparison of FOXA1_2_GUERTIN differentially enriched in A549 cell lines reveals that it is not a Foxa1 motif; it is nuclear receptor-like motif.

Table A.1 Description of cell lines information derived from the ENCODE database

Cell Line ID	Cell Line Description
A549	Epithelial cell line derived from a lung carcinoma tissue
GM12878	B-lymphocyte
H1-hESC	Embryonic stem cells inner cell mass
HEK293	Embryonic kidney, cells contain Adenovirus 5 DNA
HeLa-S3	Cervical carcinoma, ectoderm
HepG2	Hepatocellular carcinoma, endoderm
HCT-116	Colorectal carcinoma, colon cancer, endoderm
HUVEC	Umbilical vein endothelial cells, mesoderm
IMR90	Fetal lung fibroblasts
K562	Established from a patient with chronic myelogenous leukemia
MCF-7	Mammary gland, adenocarcinoma, ectoderm
NB4	Acute promyelocytic leukemia cell line.
PANC-1	Pancreatic carcinoma
SH-SY5Y	Neuroblastoma clonal subline of the neuroepithelioma cell line
T-47D	Epithelial cell line derived from a mammary ductal carcinoma

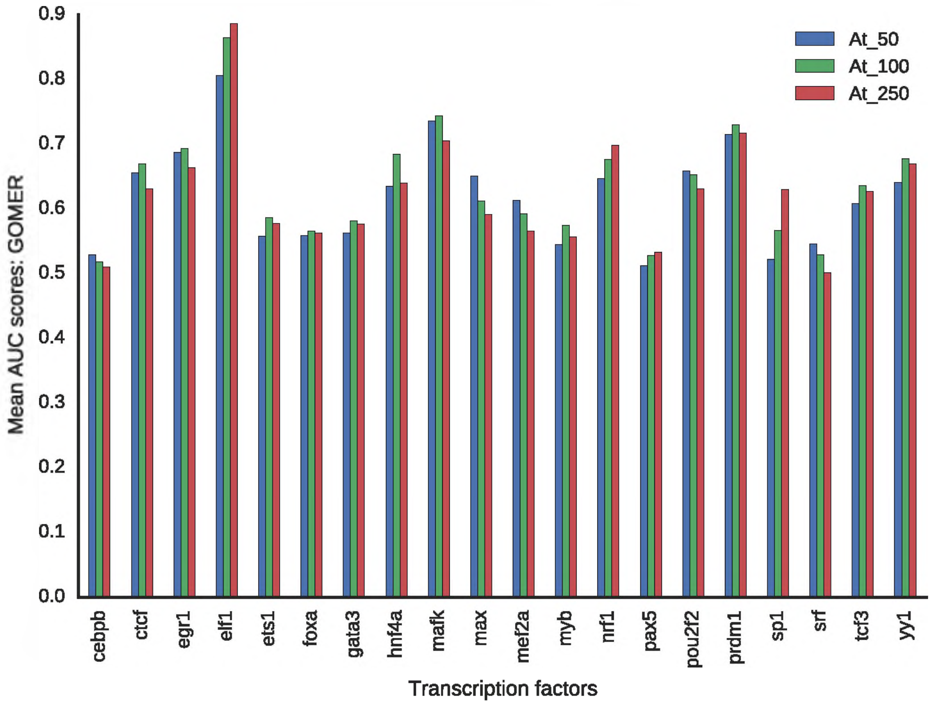


Fig. A.2 Effect of sequence length on motif ranking. Using all the motifs for each of the 20 TFs, we tested the effect of sequence length (50bp, 100bp, and 250bp) using GOMER scoring on ChIP-seq data. For each TF, the mean of the AUC of the motifs is computed.

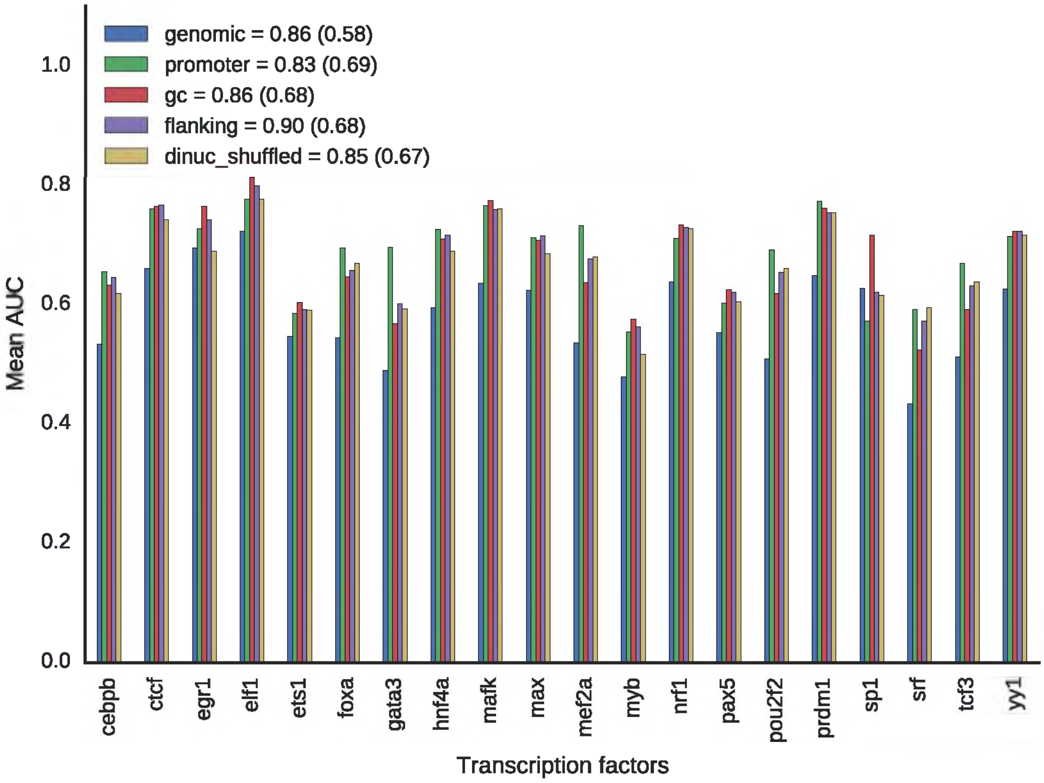


Fig. A.3 Influence of negative sequences on motif ranking: Energy. Caption same as in Figure 3.5, but this time using Energy scoring.

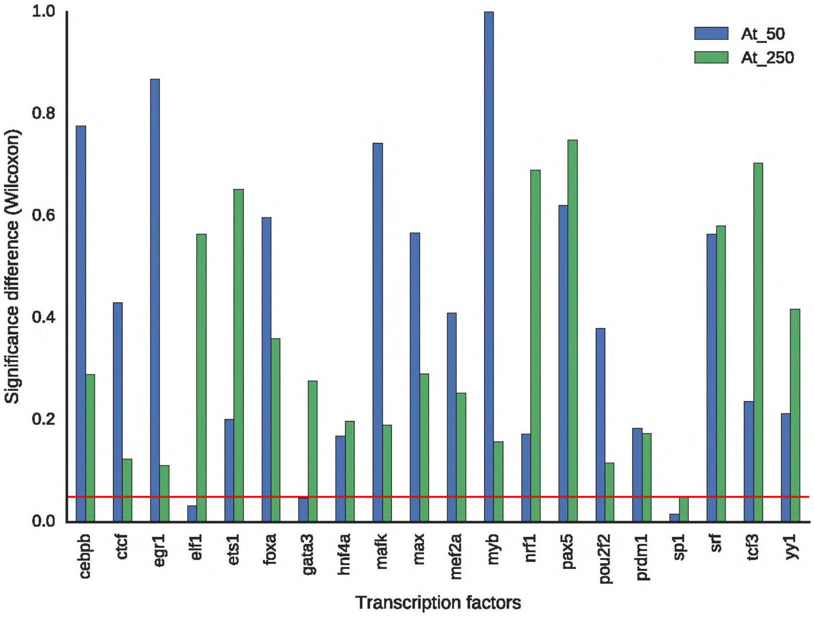


Fig. A.4 Effect of sequence length on motif ranking for Energy scoring (Wilcoxon rank-sum test). The horizontal red line represents the 0.05 significance threshold.

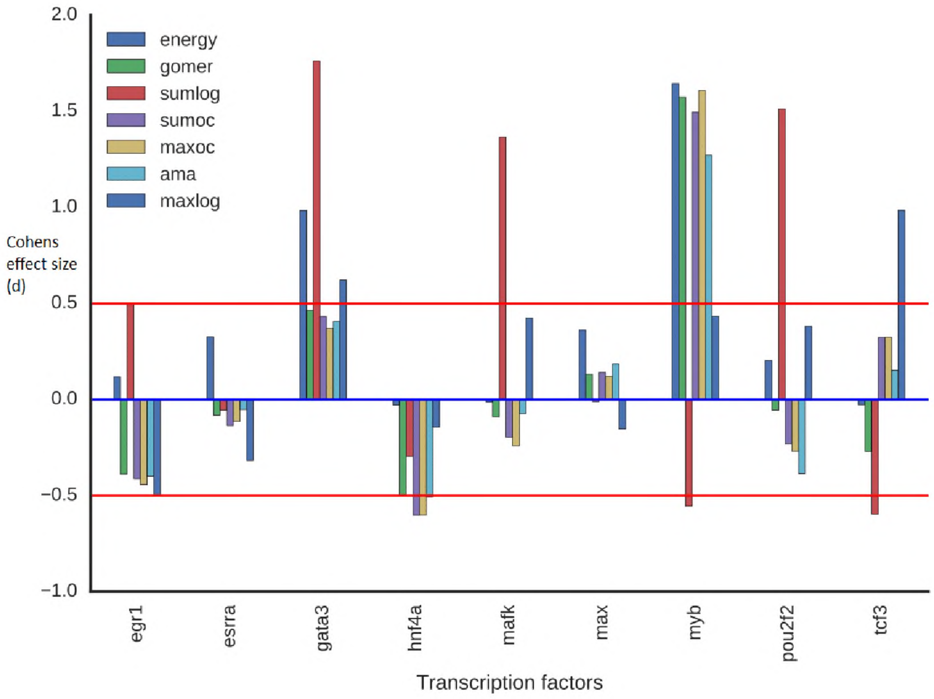


Fig. A.5 Effect of data as measured by effect size (Cohen's *d*) using AUC. The horizontal red line represents the 0.5 medium Cohen's effects size.

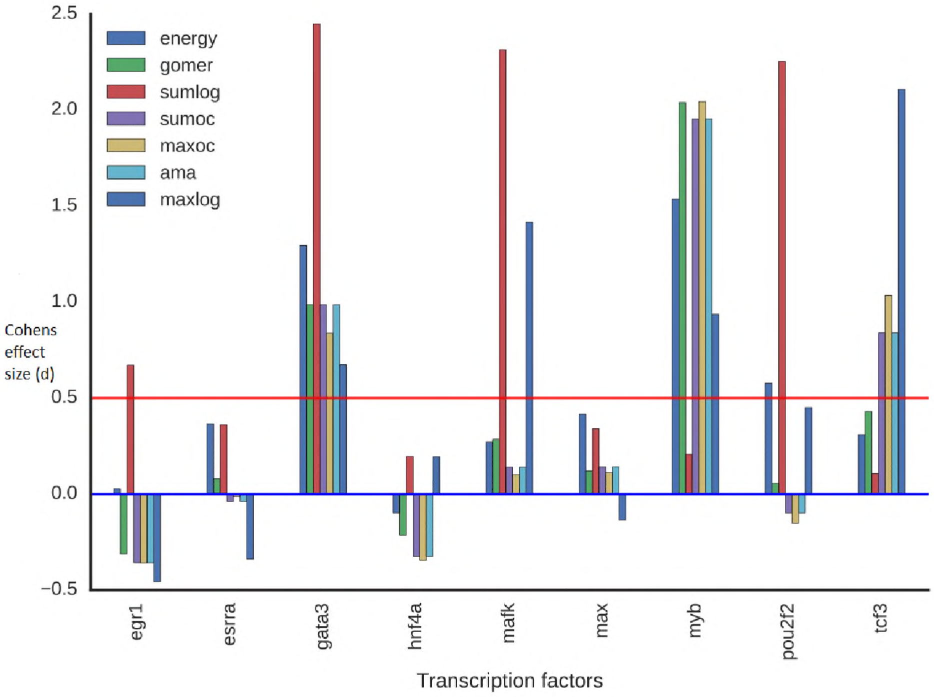


Fig. A.6 Effect of data as measured by effect size (Cohen's *d*) using MNCP.

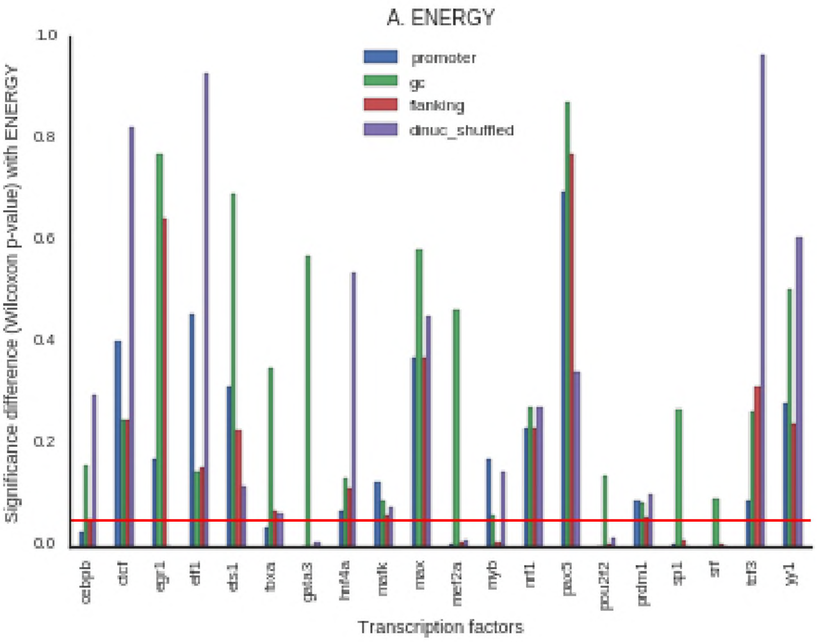


Fig. A.7 Significance difference (Wilcoxon) from choice of background sequence with Energy.

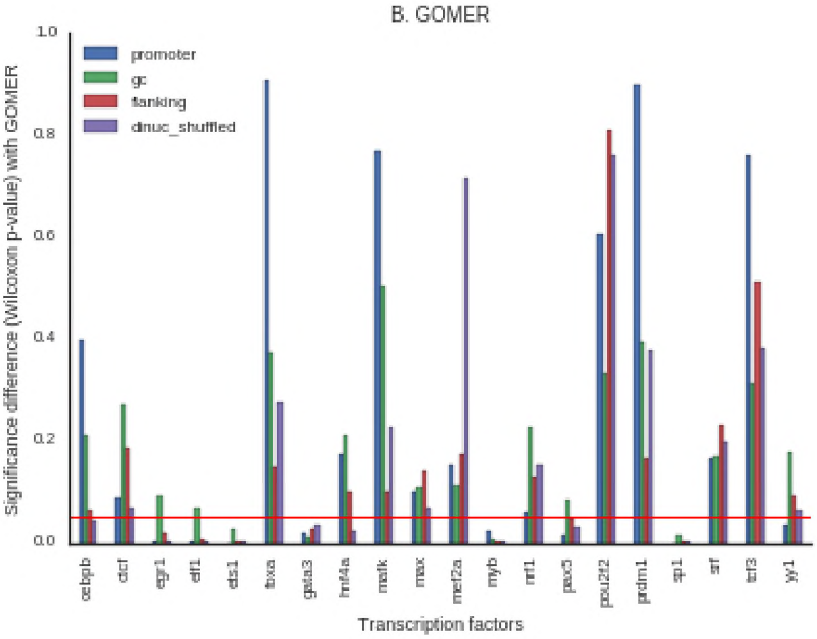


Fig. A.8 Significance difference (Wilcoxon) from choice of background sequence with GOMER.

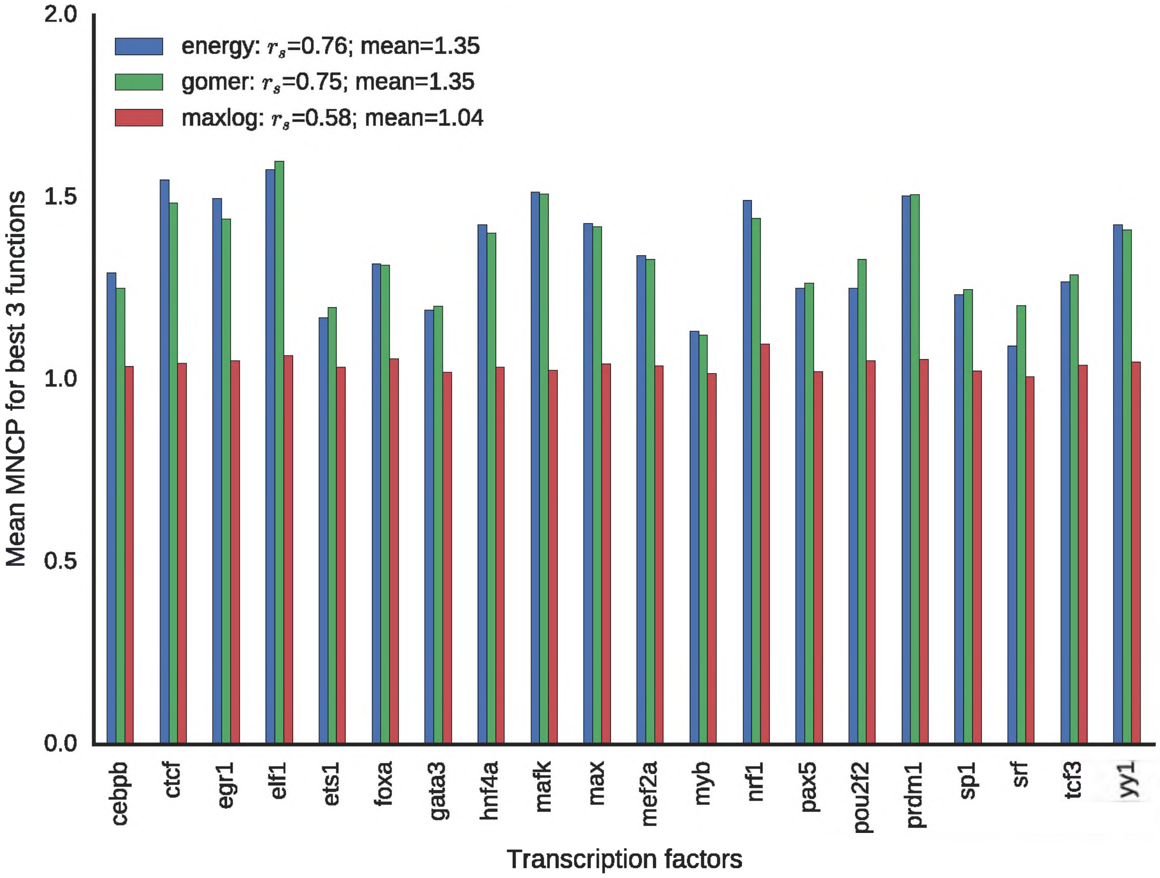


Fig. A.9 Effect of scoring functions among the best three non-redundant approaches using MNCP. The mean Spearmans correlation (r_s) provides a measure of how motif ranks for a function compare with the rest.

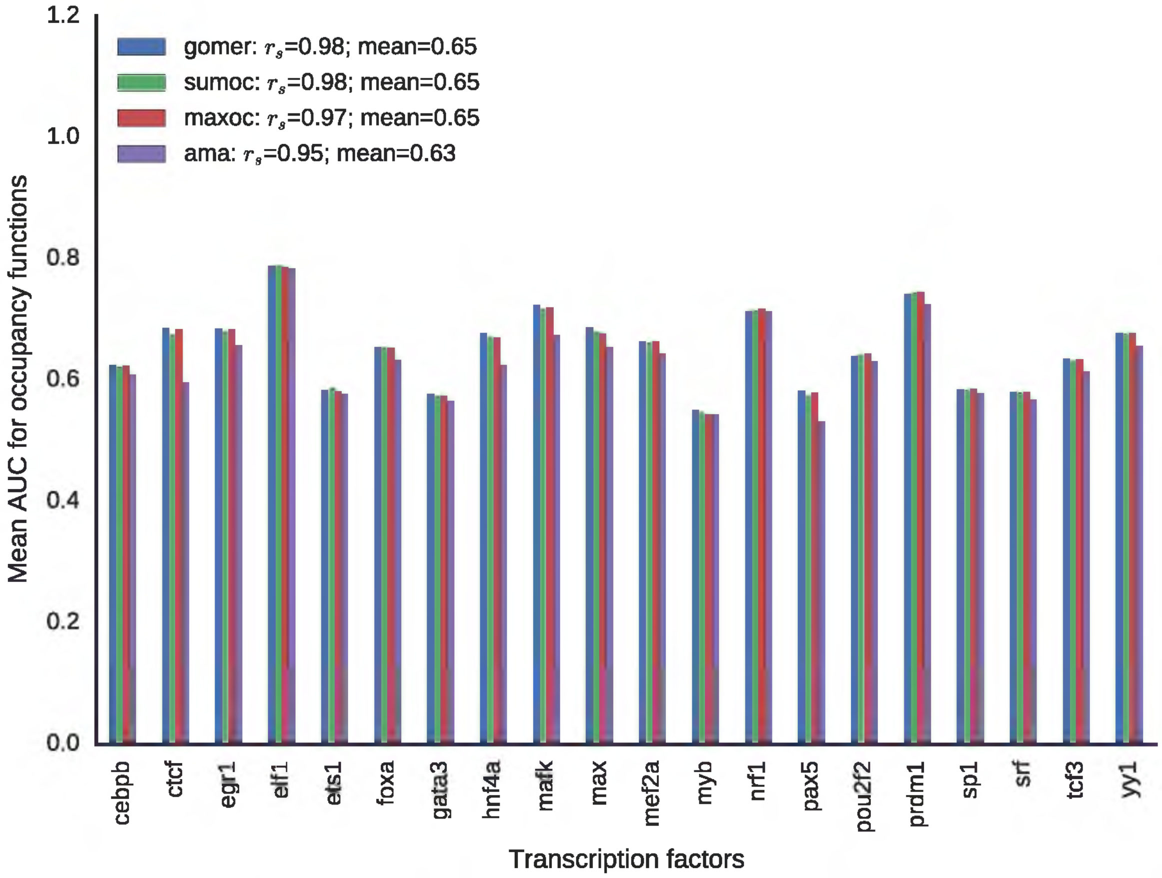


Fig. A.10 Effect of scoring functions using AUC. The mean Spearmans correlation (r_s) provides a measure of how motif ranks for a function compare with the rest.

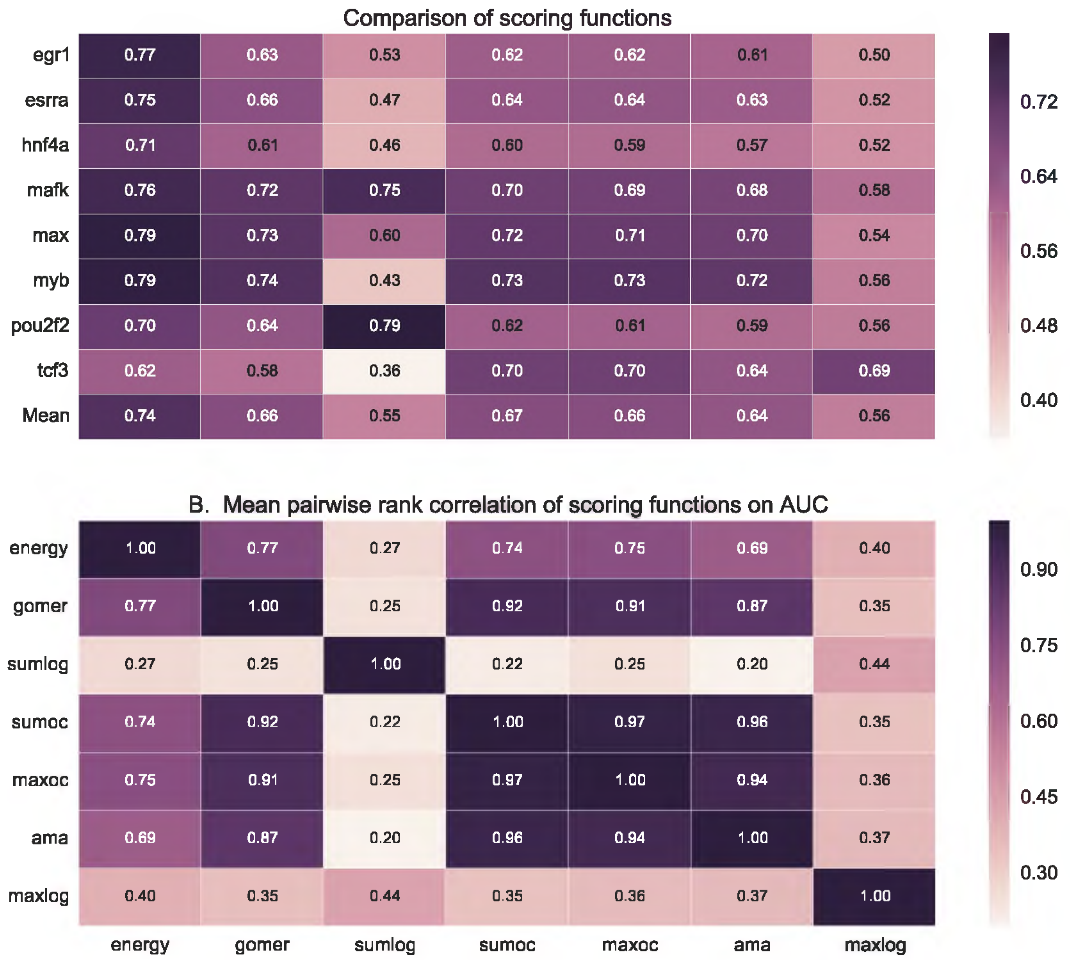
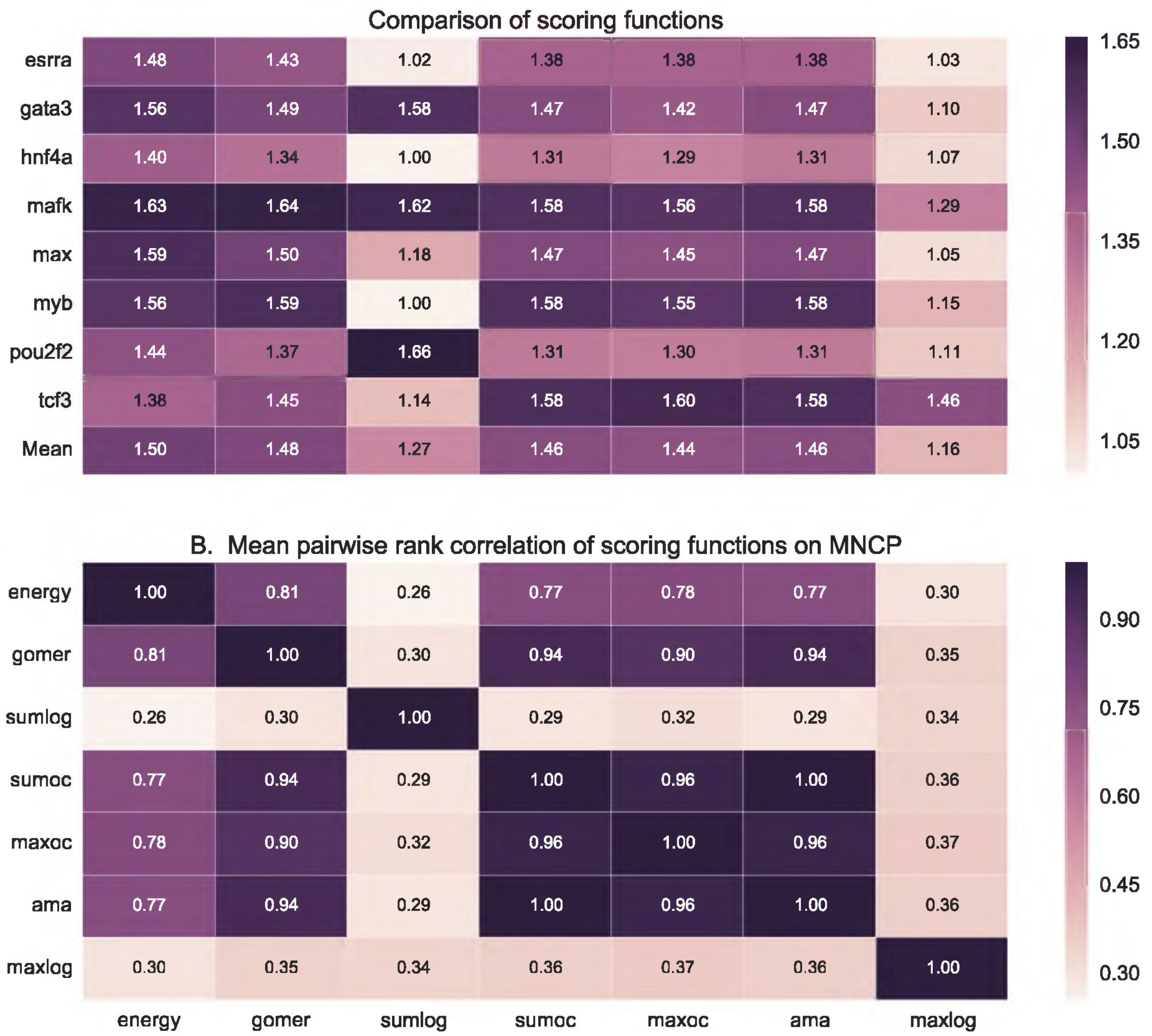


Fig. A.11 Effect of scoring function on motif ranking using AUC statistic for PBM data. A. For each transcription factor (TF), the mean AUC score is used to represent it for each scoring functions used. In **B**, we show how the ranks assigned to various motifs for a given TF by each scoring function are correlated. It displays the pairwise rank correlation for all TFs in **A**. *Sumlog*: Sum log-odds function, *Sumoc*: sum occupancy score and *Maxoc*: maximum occupancy.



Comparison of scoring functions

esrra	1.48	1.43	1.02	1.38	1.38	1.38	1.03
gata3	1.56	1.49	1.58	1.47	1.42	1.47	1.10
hnf4a	1.40	1.34	1.00	1.31	1.29	1.31	1.07
mafk	1.63	1.64	1.62	1.58	1.56	1.58	1.29
max	1.59	1.50	1.18	1.47	1.45	1.47	1.05
myb	1.56	1.59	1.00	1.58	1.55	1.58	1.15
pou2f2	1.44	1.37	1.66	1.31	1.30	1.31	1.11
tcf3	1.38	1.45	1.14	1.58	1.60	1.58	1.46
Mean	1.50	1.48	1.27	1.46	1.44	1.46	1.16

B. Mean pairwise rank correlation of scoring functions on MNCP

	energy	gomer	sumlog	sumoc	maxoc	ama	maxlog
energy	1.00	0.81	0.26	0.77	0.78	0.77	0.30
gomer	0.81	1.00	0.30	0.94	0.90	0.94	0.35
sumlog	0.26	0.30	1.00	0.29	0.32	0.29	0.34
sumoc	0.77	0.94	0.29	1.00	0.96	1.00	0.36
maxoc	0.78	0.90	0.32	0.96	1.00	0.96	0.37
ama	0.77	0.94	0.29	1.00	0.96	1.00	0.36
maxlog	0.30	0.35	0.34	0.36	0.37	0.36	1.00

Fig. A.12 Effect of scoring function on motif ranking based on MNCP statistic in PBM data. See caption in Figure A.11 for details.

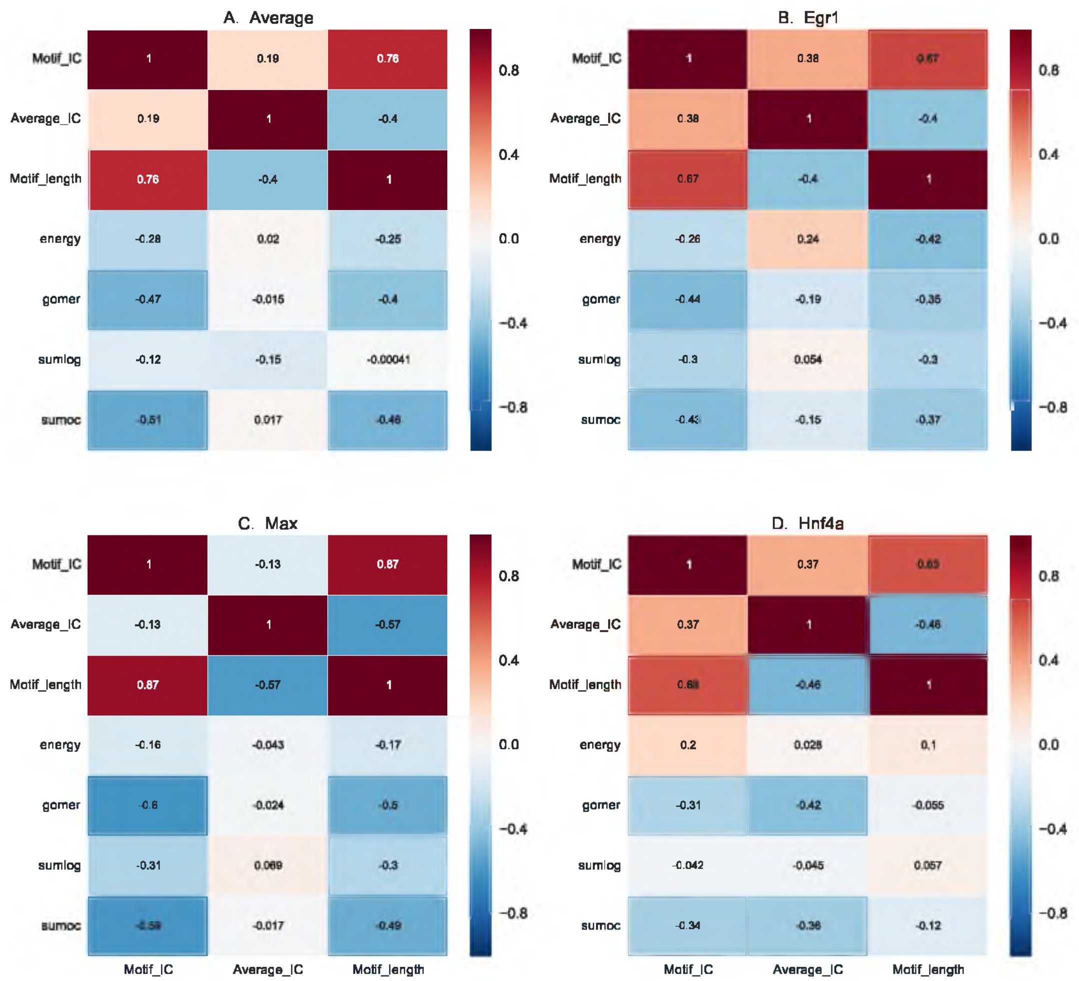
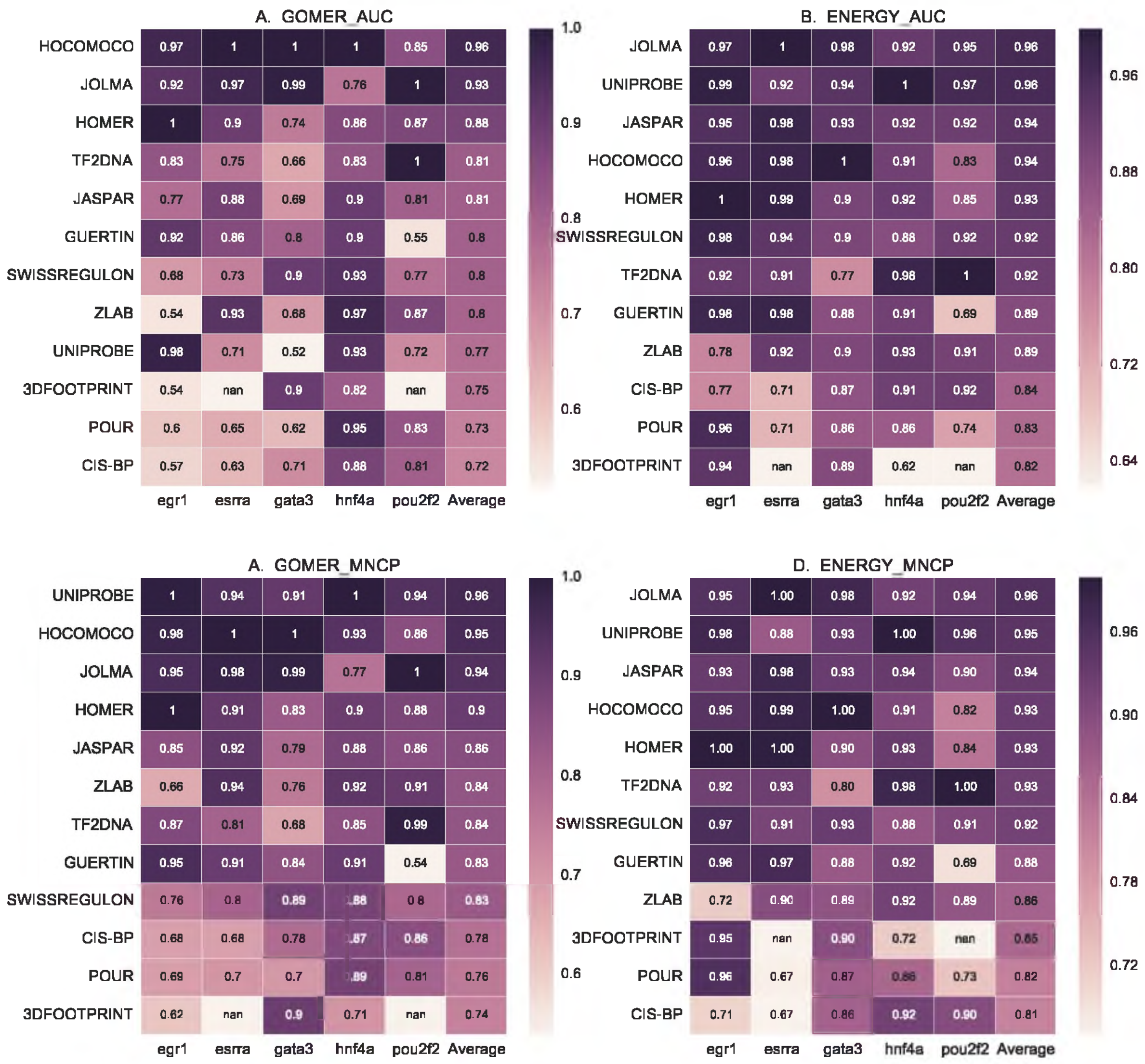


Fig. A.13 Effect of motif length and IC on scoring functions using PBM data. In this figure, we show the correlation of motif length, full-length information content (IC) and the assessment scores, to determine how the performance of scoring functions is influenced by motif characteristics. For each motif, the information content is calculated based on information theory for the whole length and also normalized for length. The results for average motif affinity (AMA) and maximum occupancy are similar to sum occupancy, and are not included.



A. GOMER_AUC

	egr1	esrra	gata3	hnf4a	pou2f2	Average
HOCOMOCO	0.97	1	1	1	0.85	0.96
JOLMA	0.92	0.97	0.99	0.76	1	0.93
HOMER	1	0.9	0.74	0.86	0.87	0.88
TF2DNA	0.83	0.75	0.66	0.83	1	0.81
JASPAR	0.77	0.88	0.69	0.9	0.81	0.81
GUERTIN	0.92	0.86	0.8	0.9	0.55	0.8
SWISSREGULON	0.68	0.73	0.9	0.93	0.77	0.8
ZLAB	0.54	0.93	0.68	0.97	0.87	0.8
UNIPROBE	0.98	0.71	0.52	0.93	0.72	0.77
3DFOOTPRINT	0.54	nan	0.9	0.82	nan	0.75
POUR	0.6	0.65	0.62	0.95	0.83	0.73
CIS-BP	0.57	0.63	0.71	0.88	0.81	0.72

B. ENERGY_AUC

	egr1	esrra	gata3	hnf4a	pou2f2	Average
JOLMA	0.97	1	0.98	0.92	0.95	0.96
UNIPROBE	0.99	0.92	0.94	1	0.97	0.96
JASPAR	0.95	0.98	0.93	0.92	0.92	0.94
HOCOMOCO	0.96	0.98	1	0.91	0.83	0.94
HOMER	1	0.99	0.9	0.92	0.85	0.93
SWISSREGULON	0.98	0.94	0.9	0.88	0.92	0.92
TF2DNA	0.92	0.91	0.77	0.98	1	0.92
GUERTIN	0.98	0.98	0.88	0.91	0.69	0.89
ZLAB	0.78	0.92	0.9	0.93	0.91	0.89
CIS-BP	0.77	0.71	0.87	0.91	0.92	0.84
POUR	0.96	0.71	0.86	0.86	0.74	0.83
3DFOOTPRINT	0.94	nan	0.89	0.62	nan	0.82

A. GOMER_MNCP

	egr1	esrra	gata3	hnf4a	pou2f2	Average
UNIPROBE	1	0.94	0.91	1	0.94	0.96
HOCOMOCO	0.98	1	1	0.93	0.86	0.95
JOLMA	0.95	0.98	0.99	0.77	1	0.94
HOMER	1	0.91	0.83	0.9	0.88	0.9
JASPAR	0.85	0.92	0.79	0.88	0.86	0.86
ZLAB	0.66	0.94	0.76	0.92	0.91	0.84
TF2DNA	0.87	0.81	0.68	0.85	0.99	0.84
GUERTIN	0.95	0.91	0.84	0.91	0.54	0.83
SWISSREGULON	0.76	0.8	0.89	0.88	0.8	0.83
CIS-BP	0.68	0.68	0.78	0.87	0.86	0.78
POUR	0.69	0.7	0.7	0.89	0.81	0.76
3DFOOTPRINT	0.62	nan	0.9	0.71	nan	0.74

D. ENERGY_MNCP

	egr1	esrra	gata3	hnf4a	pou2f2	Average
JOLMA	0.95	1.00	0.98	0.92	0.94	0.96
UNIPROBE	0.98	0.88	0.93	1.00	0.96	0.95
JASPAR	0.93	0.98	0.93	0.94	0.90	0.94
HOCOMOCO	0.95	0.99	1.00	0.91	0.82	0.93
HOMER	1.00	1.00	0.90	0.93	0.84	0.93
TF2DNA	0.92	0.93	0.80	0.98	1.00	0.93
SWISSREGULON	0.97	0.91	0.93	0.88	0.91	0.92
GUERTIN	0.96	0.97	0.88	0.92	0.69	0.88
ZLAB	0.72	0.90	0.89	0.92	0.89	0.86
3DFOOTPRINT	0.95	nan	0.90	0.72	nan	0.85
POUR	0.96	0.67	0.87	0.86	0.73	0.82
CIS-BP	0.71	0.67	0.86	0.92	0.90	0.81

Fig. A.14 Ranking of motif databases when based on PBM data. We compare the motif databases by using the best ranking for each motif using GOMER and energy AUC and MNCP values, and CentriMo enrichment values. For each scoring function, the scores for each TF are normalized by dividing each value with the maximum, which are then averaged to rank the different databases.

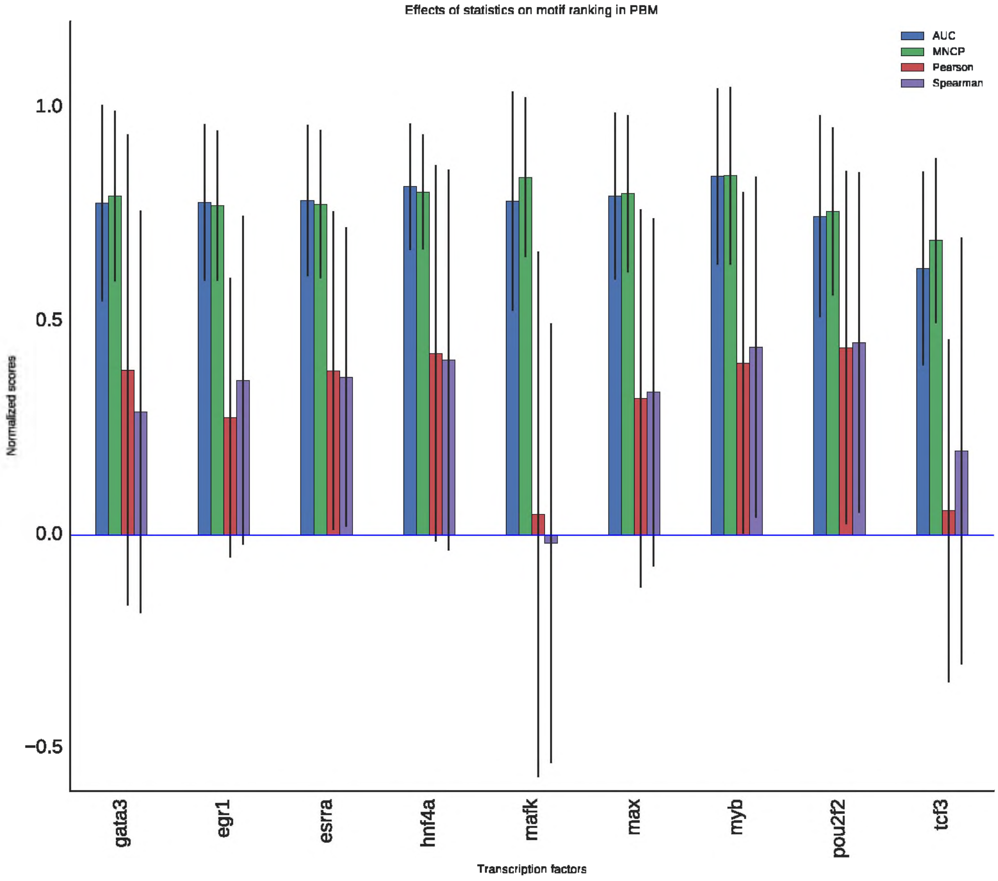
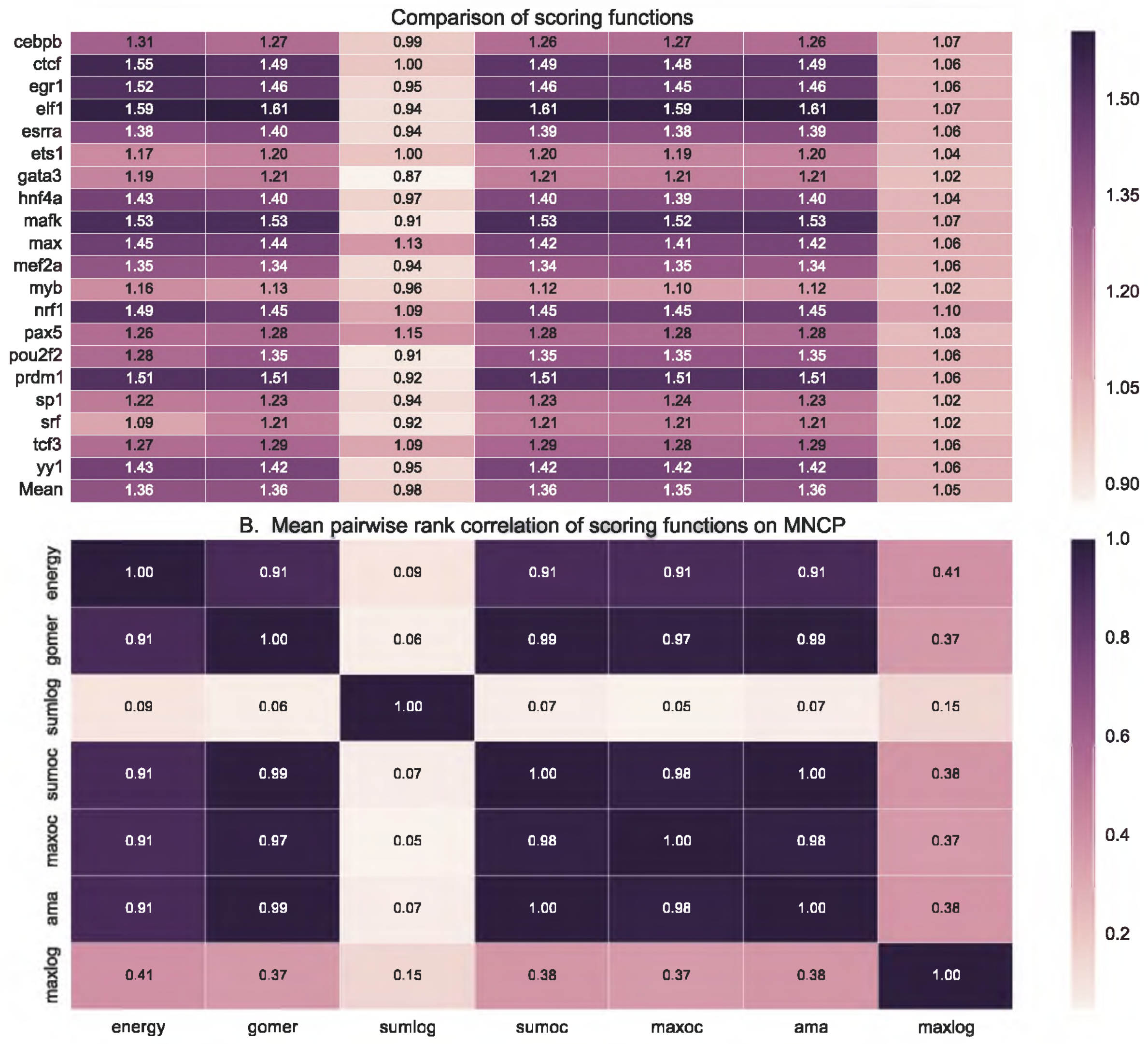


Fig. A.15 Statistics used influence motif ranks. For each TF, the motifs are used to score sequences using GOMER scoring function and ranks determined by MNCP, AUC, Pearson and Spearman's rank correlation. In this figure, we compute the mean normalized scores and compute the standard deviation for each TF, which is displayed as error bars.



Comparison of scoring functions

cebpb	1.31	1.27	0.99	1.26	1.27	1.26	1.07
ctcf	1.55	1.49	1.00	1.49	1.48	1.49	1.06
egr1	1.52	1.46	0.95	1.46	1.45	1.46	1.06
elf1	1.59	1.61	0.94	1.61	1.59	1.61	1.07
esrra	1.38	1.40	0.94	1.39	1.38	1.39	1.06
ets1	1.17	1.20	1.00	1.20	1.19	1.20	1.04
gata3	1.19	1.21	0.87	1.21	1.21	1.21	1.02
hnf4a	1.43	1.40	0.97	1.40	1.39	1.40	1.04
mafk	1.53	1.53	0.91	1.53	1.52	1.53	1.07
max	1.45	1.44	1.13	1.42	1.41	1.42	1.06
mef2a	1.35	1.34	0.94	1.34	1.35	1.34	1.06
myb	1.16	1.13	0.96	1.12	1.10	1.12	1.02
nrf1	1.49	1.45	1.09	1.45	1.45	1.45	1.10
pax5	1.26	1.28	1.15	1.28	1.28	1.28	1.03
pou2f2	1.28	1.35	0.91	1.35	1.35	1.35	1.06
prdm1	1.51	1.51	0.92	1.51	1.51	1.51	1.06
sp1	1.22	1.23	0.94	1.23	1.24	1.23	1.02
srf	1.09	1.21	0.92	1.21	1.21	1.21	1.02
tcf3	1.27	1.29	1.09	1.29	1.28	1.29	1.06
yy1	1.43	1.42	0.95	1.42	1.42	1.42	1.06
Mean	1.36	1.36	0.98	1.36	1.35	1.36	1.05

B. Mean pairwise rank correlation of scoring functions on MNCP

	energy	gomer	sumlog	sumoc	maxoc	ama	maxlog
energy	1.00	0.91	0.09	0.91	0.91	0.91	0.41
gomer	0.91	1.00	0.06	0.99	0.97	0.99	0.37
sumlog	0.09	0.06	1.00	0.07	0.05	0.07	0.15
sumoc	0.91	0.99	0.07	1.00	0.98	1.00	0.38
maxoc	0.91	0.97	0.05	0.98	1.00	0.98	0.37
ama	0.91	0.99	0.07	1.00	0.98	1.00	0.38
maxlog	0.41	0.37	0.15	0.38	0.37	0.38	1.00

Fig. A.16 Effect of scoring function on motif ranking based on MNCP statistic. **A:** For each TF, the mean AUC score is computed for each of the scoring functions used. **B:** How the ranks assigned to various motifs for a given TF by each scoring function are correlated. It displays the pairwise rank correlation for all TFs in **A**. *Sumlog*: Sum log-odds function, *Sumoc*: sum occupancy score and *Maxoc*: maximum occupancy.

Appendix B

ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge

> *"...in this community challenge, we ask participants to develop and apply computational methods that can integrate genomic DNA sequence, in vitro DNA shape parameters, in vivo chromatin accessibility (DNase-seq) profiles and overall gene expression (RNA-seq) data to predict in vivo binding maps of a diverse collection of TFs in a variety of cell lines. We aim to perform systematic comparisons to benchmark and identify methods with high predictive performance."*
>
> –ENCODE-DREAM Organizers

The aims and objectives of the ENCODE-DREAM challenge encompassed two of the themes covered in our research objectives: combining *in vitro* and *in vivo* data, and systematic performance comparisons of the methods. Therefore, this was a timely opportunity to learn and benchmark our approaches community-wide. Also, it was a confirmation that the questions we are tackling are relevant to the research community. The difference, however, was on our approach; our model evaluations were focused on PWM while the challenge on more advanced models and in general binding site prediction. On combining *in vivo* and *in vitro* data, our focus was mainly on PBM and DNase-seq data while the challenge was focused on ChIP-seq and DNase-seq, and DNA-shape. While the challenge was limited to the provided datasets, our objectives went beyond these, including the use of conservation data. This section provides a general idea of the ENCODE-DREAM challenge and our approach and its shortcomings.

Challenge aims and objectives

The main aim is to identify the best performing model for predicting positional *in vivo* TF binding maps within and across cell types and tissues, which can also act as a systematic benchmark to compare and test current and future methods. The objectives are further broken down to:

1. identify TFs and families of TFs that are predictable across cellular contexts
2. assess the influence of training and testing context on performance
3. determine how the *in vivo* context-specific features contribute to prediction
4. determine the extent to which calibrated binding maps guarantee performance

Challenge Data

The challenge tested the participants by their ability to model and predict *in vivo* binding as defined by ChIP-seq data by integrating a variety of datasets described below:

- **ChIP-seq data**: For a given cell line and transcription factor, the genome is binned in a 200bp window sliding every 50bp and labelled: bound (B), high confidence peaks based on 5% FDR; Ambiguous (A) if it doesn't pass; the rest are Unbound (U).
- **Chromatin accessibility (DNase-seq)**: Fold-enrichment signal coverage tracks, peaks and alignment files are provided.
- ***In-vitro* DNA shape**: Shape data, obtainable from GBshape[1] or using the DNAShapeR-tools[2].
- **Gene expression (RNA-seq)**: gene-level expression levels of all human genes as defined in the GENCODEv19 gene annotations.

Our approach: TeamKE

The TeamKE submission was implemented with XGBoost using PWM scores, DNase scores, GC content and shape information. Specifically, for any set of training cell lines, we filter for

[1] http://rohsdb.cmb.usc.edu/GBshape/

[2] http://tsupeichiu.github.io/DNAshapeR/

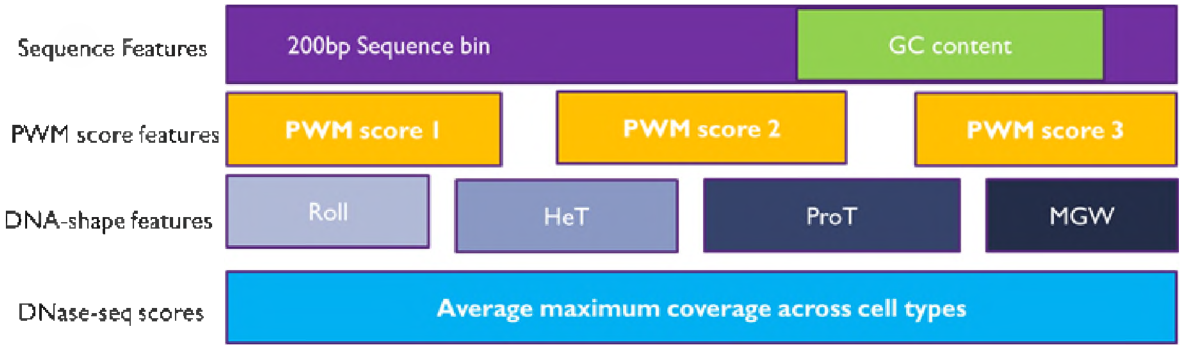


Fig. B.1 Feature engineering for the DREAM challenge.

the sites labelled B (positive) or U (background) in all of them. We do this to ensure the model trained can be applied to different cell lines. For PWM scores, we used a collection of PWM in our database, sourced from a variety of databases, ranked and clustered to obtain the best three motifs located in different clusters, where available. We also use the average maximum fold enrichment over all the cell lines for DNase data and an average DNA shape information in each bin (could have used shape information for the hit site but ran out of time).

Our choice of XGBoost algorithm to train the model is motivated by its speed (parallel tree construction), memory efficiency and ease of handling large data. The parameters were optimised for a few TFs and used in training the rest of the TFs. We also tested SGD, SVM and Gradient boosting but found XGBoost to provide the best predictions within a reasonable time.

From this challenge, we learned that the most predictive feature is the DNase and PWM score. Also, the quality of some of the PWMs in the databases is wanting, for example, ATF2 PWMs were least predictive, attaining auPRC of 0.061. Finally, although using average shape information did have some predictive value, the improvement is not significant. As we later figured out from further feature engineering, the shape features should be extracted from the hit site.

Some design principles

The amount of data we had to process for the DREAM challenge was just massive. Therefore, there was a need to optimise each stage of the modelling, from feature extraction to training. Some of the approaches we employed include:

- Computation speed
  - Parallelization of code and use of clusters

- Large data, requiring large processing memory
  - Chunk-wise data processing
  - Out-of-memory data processing with HDF5
- How to select positive and negative data for training
  - Querying HDF5, just like a database
- Data storage, how to minimise intermediate data
  - On demand processing from BED files using pysam

Limitations

Our approach did not perform very well in the challenge because we did not fully utilise the data provided. The main limitation was time and the fact that most of the skills required to process the data were being acquired on the challenge. The following could have and will be done to improve further on our current model:

- Per TF parameter optimization
- Better predictive PWM models
- Using gene expression data as additional feature
- Further feature engineering

Some of these will be tested for the benchmarking stage of the challenge, and are being put aside as a future direction of research.

Availability

Our code used for the challenge is available from GitHub[3]. Details of further work on this will be made available from this repository.

[3] https://github.com/kipkurui/TeamKE_DreamChallenge