Computational Methods for Dissecting Transcription Regulatory Networks

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Saeed Omidi Klishami

aus Tehran, Iran

Basel, 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Referent: Prof. Erik van Nimwegen

Korreferent: Prof. Sven Bergmann

Basel. den 22, April 2014

Declaration of Authorship

I, Saeed Omidi Klishami, declare that this thesis titled, 'Computational Methods for Dissecting Transcription Regulatory Networks' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________________

Date: ______________________________________

“Scientific knowledge is a body of statements of varying degrees of certainty – some most unsure, some nearly sure, none absolutely certain. ”

Richard Feynman, *The Pleasure of Finding Things Out*

UNIVERSITY OF BASEL

Abstract

Erik van Nimwegen
Biozentrum

Doctor of Philosophy

Computational Methods for Dissecting Transcription Regulatory Networks

by Saeed Omidi Klishami

Gene expression is regulated on many levels of which transcription regulation is the most studied. Transcription factors (TF) arguably are the most integral part of transcriptional regulation; by recognizing and specifically binding to short, but degenerate, DNA sequences, TF control the transcription initiation of genes. In the mid 70s, the first mechanism of TF binding recognition was introduced which is based on simple hydrogen bonding rules. Since then numerous observations of complex TF binding site recognition has substantially altered our viewpoint of the TF binding specificity. In particular, recent studies have uncovered subtle pairwise dependencies (PD) between positions within binding sites, which disapproves the common assumption in many current computational models– that binding positions contribute independently toward the binding affinity of a sequence. Several works already tried to incorporate PD within a framework of binding site recognition, but due to the complexity of this problem, they failed to provide a consistent and rigorous methodology. In my PhD, we have addressed PD from a computational perspective by introducing dinucleotide weight tensors (DWT), which incorporates the entire information on PD into a robust mathematical model. Among several advantages, the DWT model does not have any tunable parameters which makes it highly applicable. Finally, recent boom of high-throughput data has provided a unique window to investigate various questions regarding to binding specificity. Here, we have systematically tested the DWT model against the classical non-dependent model over a large number of human TF ChIP-seq data. The *in vivo* data has clearly demonstrated the role of PD in binding specificity. Remarkably, we found resulting dependency models from ChIP-seq data, outperform non-dependent models on separate *in vitro* data. In fact, testing over the HT-SELEX data, a high-throughput variant of the SELEX, has further corroborated the importance of PD.

Acknowledgements

It is with a great pleasure to present my doctorate dissertation for which I have worked constantly over a time period of four years. My warmest appreciation must go to Professor Erik van Nimwegen, a true scientist with a great passion for perfection in his scientific works, for letting me to work on this project in his lab. Doing this project, would have not been possible without the helpfulness and often ardent supervision of Erik. With a great certainty, what I have learned in his lab will be staying as a precious provision in my future endeavors.

In addition, I had a unique privilege for having two great and reputable scientists, namely Professor Sven Bergmann and Professor Mihaela Zavolan, to whom must go a considerable credit for many useful advises and comments.

I have benefited considerably from working along side with great scientist from whom I learned a lot in these years. To name a few, Dr. Florian Geier, Dr Peter Pemberton-Ross, Dr. Piotr Balwierz, Dr Philipp Berninger, Dr. Mikhail Pachkov and Dr. Evgeniy Ozonov. I am deeply grateful for helpful discussions and invaluable comments on my work. In particular, I am deeply grateful to Florian Geier whose invaluable expertise on high-throughput sequencing data and genome-wide binding site prediction I have benefited so much on the Enhancer project in Chapter 4. I owe a special dept to a great friend Dr. Lukas Burger who has helped me for the Dependency algorithm and numerical computation. It is also important to name Severin Berger for his invaluable contribution to the large-scale ChIP-seq analysis, that has been used greatly for model comparison in Chapter 2.

Last but not least, I wish to use this opportunity to express my deep gratitude toward my beloved family, specially my mother whose support has been an integral part over these years of education. I am thankful to my friends, specially friends from *el tigre* club, namely Arnau, Dominik, Aaron, Hadi and Alex for sharing all the memorable moments in these years.

Contents

List of Figures

Abbreviations

TF	**T**ranscription **F**actors
DBD	**D**NA **B**inding **D**omain
PSWM	**P**osition **S**pecific **W**eight **M**atrix
DWT	**D**inucleotide **W**eight **T**ensor
ChIP-seq	**Ch**romatin **I**mmuno**P**recipitation sequencing
ETS	**E**26 **T**ransformation-**s**pecific
TFBS	**T**ranscription **F**actor **B**inding **S**ite
PD	**P**airwise **D**ependency
BIC	**B**ayesian **I**nformation **C**riterion
ML	**M**aximum **L**ikelihood
MAP	**M**aximum **A** **P**osteriori
DiLogo	**DI**nucleotide Logo
ENCODE	**ENC**yclpedia **O**f **D**NA **E**lements
SELEX	**S**ystematic **E**volution of **L**igands by **Ex**ponential Enrichment
HT-SELEX	**H**igh **T**hroughput SELEX
PBM	**P**rotein **B**inding **M**icroarray
DAG	**D**irected **A**cyclic **G**raph
KL	**K**ullback-**L**eibler divergence
miRNA	**mi**cro RNA
eRNA	**E**nhancer RNA
DHS	**D**NAse I **H**ypersensitive **S**ites
FANTOM	**F**unctional **AN**noTation **O**f the **M**ammalian genome
3C	**C**hromosome **C**onformation **C**apture
5C	**C**arbon **C**opy 3C
Hi-C	**Hi**gh throughput 3C

MARA	**M**otif **A**ctivity **R**esponse **A**nalysis
HMM	**H**idden **M**arkov **C**hain
CAGE	**C**ap **A**nalysis of **G**ene **E**xpression
MDL	**M**inimum **D**escription **L**ength
LS	**L**ocal **S**earch

To my parents

Chapter 1

The complexity of TF-DNA interaction

1.1 The classical Protein-DNA recognition code

In almost every biological process, various proteins recognize and interact with other molecules. Some of these interactions form tightly bound complexes, such as histone interaction with DNA. Other types of interactions, that is the main goal of this work, concerns with dynamically and reversibly binding of proteins to DNA, such as transcription factors (TF) binding to their DNA binding sites [1]. There are different ways that polypeptides interact specifically with other molecules such as DNA. These include hydrophobic interaction, which plays an important role in protein folding. Another type of interactions that are of more specific character owing to the directional character are electrostatic interactions of which the most important form of them is hydrogen bonding. Motivated by the stability of the hydrogen bond interaction between the proteins and double helical DNA, Seeman *et al.* proposed the first protein recognition code in 1976 [2]. The proposed model defines the specific binding by the number of hydrogen bonds, where as it is argued, the specific amino-acid side chain interactions involving two hydrogen bonds are the components of the specific recognition system. In the major groove an argenine binds to guanine, and asparagine or guanine interacts specifically with adenine base pair. Using pairs of hydrogen bonds is sufficient to unambiguously distinguish the A-T or G-C pairs in the major groove of the double helix DNA. However, this oversimplified view to the TF-DNA interaction is rejected after 30 years of subsequent analysis. A multitude of studies have demonstrated myriad complications with such simple recognition code [3]. This chapter briefly reviews various evidences regarding to the complexity of TF binding site recognitions.

1.2 Direct versus Indirect readout mechanism

A direct recognition of DNA binding sequences by TF through direct interactions between amino-acid side chain and DNA base pairs is referred to as direct readout (also known as base readout). Indirect readout is yet another type of TF-DNA interaction that depends upon DNA conformation and its elastic properties [4]. In contrast to the direct readout, in indirect readout the shape and structure of DNA contributes to the binding affinity of DNA [5]. While the elements of direct readout contribute to the binding site recognition by TF, it is clear that in many situation, where the form of DNA deviates from the standard B-form double helix, the indirect readout influences the binding affinity to some degree [6]. Formally, indirect readout, or equivalently shape readout, is defined as TF-DNA interactions which is depend on base pairs that are not directly touching the protein surface .

These observations are in contradiction with the current information-theoretic based methods for binding site prediction, e.g. position specific weight matrix (PSWM) model. For a detailed discussion on the PSWM model refer to the Chapter 2. The DNA consisting of four basic nucleotides form a three-dimensional double-stranded structure in which every base has a different conformation and chemical properties. Studies have demonstrated that the conformation of the DNA is sequence-dependent, and so the structure variation is actually utilized by proteins in order to recognize their binding sites [7] [8]. It is however more challenging to resolve the three-dimensional structure of DNA due to its higher flexibility nature compare with the complexes. Nonetheless, Kardar and colleagues presented the first *structure-based* model for PurR, an *Escherichia coli* TF, recognition binding site [9]. Limited number, and low accuracy, of current structure data has served as hindrance for building mechanistic models, therefore the statistical-based methods that are learned form cognate sites are at the moment more pragmatic.

The miss-conformation of DNA from the standard B-form double helix is categorized into two general groups: local and global shape miss-conformation [6]. In the local shape readout, the deviation from the B-form double helix is localized in rather small regions of the DNA and parameters such as kinks in the DNA can change the affinity of binding to the sequence. On the other hand, the global readout refers to a larger deformation or bent in DNA molecule. One example of the local shape readout is demonstrated by Rohs *et al.* where a mechanism named as minor groove narrowing, often seen in AT-rich sequences enhances the negative electrostatic potential of DNA [10]. It is suggested that the recognition of the local variation in DNA by protein offers an opportunity for forming specific TF-DNA complex. SRF-like and p53-like TFs are two TF classes, according to SCOP superfamilies [11], that contact narrow minor grooves. The global shape readout is observed in situation where the larger deformation of the DNA is involved. In a study on the binding of papillomavirus E2 TF, Rohs and colleagues by using an all-atom Monte Carlo simulation found a greater deviation of DNA from the B-form double helix [12]. This study also suggests that the DNA conformation is highly sequence-dependent.

1.3 Spatial dependency between protein and DNA

Unlike the last section that was concerned with the role of DNA conformation in binding specificity, this section explains an inherent flexibility of the folded protein in recognizing DNA binding sites. As it has been discussed in (section 1.1), the favorable TF-DNA interaction is brought about by hydrogen bonding between these two molecules. This requires an exact spatial localization of TF versus its DNA binding site. The positioning of protein against DNA is also called protein docking geometry. Thus, the correct docking geometry is of an utmost importance in establishing a specific binding interaction [13]. The sheer number of possible geometrical alignments of the protein backbone to the DNA molecules makes it very difficult to specify a simple model of the binding preference of a protein. Very interestingly, it is shown that even a single TF can employ various docking geometry when binding to different sequences. For instance, RelA homodimer, a member of the larger NF-κB family, binds to the sequence 5'-GGAA(A)TTTC-3', where one subunit forms the hydrogen bond interaction with the half-site 5'-GGAA, and the other subunit exhibits a major rotation, causing it not to touch the TTTC-3' half-site [14]. In contrast, both subunits specifically bind to the two half-sites of a symmetrical sequence 5'-GGAA(A)TTCC-3' [15]. In an earlier study, it was illustrated that different types of Zn finger motifs show distinct docking geometry [16]. The two Zn finger motifs in GLI show flexible binding geometry modes: with one mode contacting one strand of DNA (Watson strand), and another motif facilities binding to another DNA strand. It has also been shown that protein with multiple zinc fingers show a large range of docking geometries [13].

1.4 Low-affinity and non-specific binding

Apart from sequences that bind with high specificity, there are roughly two classes of other sequences that exhibit binding. These include low affinity DNA sites and non-specific binding events. In low affinity binding, the sequence content of the site contributes less to the binding affinity of the sequence compared to high affinity sequences. However, several studies have pointed out their abundance in binding. On the other hand, non-specific binding events are also observed specially in experiments such as ChIP-seq. These non-specific binding events are not function of the sequence itself, but different parameters such as the openness of binding site inside the active genomic regions.

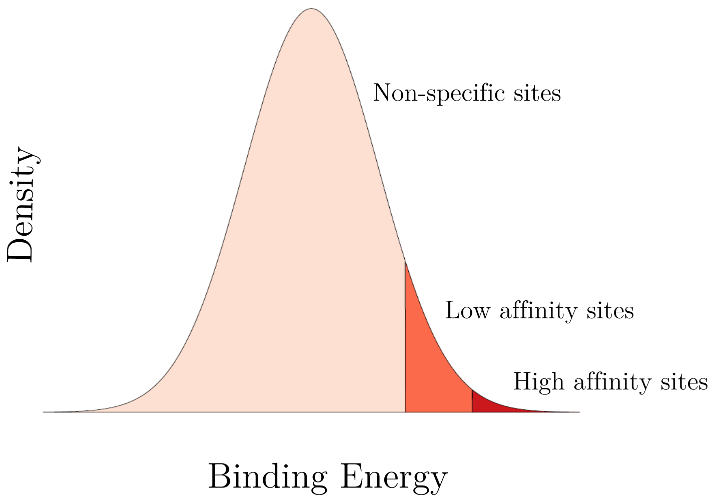


Figure 1.1: Learning the parameters in computational models is often done using only high affinity binding sequences, which represent only a very small fraction of all potential binding sequences. A second class of binding sequences corresponds to 'low affinity' sequences that have sigificantly lower affinity than the high affinity sites, but that still bind the protein more strongly than average sequences. Finally, non-specific binding events can also occur at the bulk of sequences that have no sequence=specific binding affinity.

Figure 1.1 shows a simple representation of the binding sequences in terms of their binding free energy. High affinity sequences are representing a very small fraction of the binding sequences that often perfectly match to the consensus sequences. Only this small class of the sequences is usually the main focus of the published works. It is also true that many of the computational models, e.g. PSWM matrix, and motif databases such as JASPAR and Transfact are solely based on the high-affinity sequences. The other important class of the sequences is low-affinity sequences. Despite their lower chance for binding compared to high-affinity sequences, the low-affinity sequences exhibit a modest number of binding events. Unlike what has been traditionally assumed, these low-affinity sequences represent an important class of the binding sequences. Several studies highlighted their importance in different context, such as *in vivo* experiments.

In a study Jiang and Levin [17] demonstrated the role of low-affinity binding of dorsal (dl) binding in patterning the early embryo. In line with the results from Jiang and Levin, other works showed the low-affinity binding tendency of different TF gradients, which arise specially at the early stages of development in a spatio-temporal manner [18] [19] [20] [21]. In a study on ETS family TFs, it was shown that binding events on genomic regions with low affinity sites were more common than binding events to ETS consensus sequences [22]. It is speculated that the low-affinity binding is pronounced due to the cooperative binding of a different number of proteins to different genomic loci. Moreover, it was shown that the proteins from the same family share the same high-affinity sequences, but they exhibit quite different medium-affinity and low-affinity binding preferences. For example, mouse TFs Irf5 and Irf4, although sharing the same high-affinity site, interestingly show a different lower-affinity sequences 5'-CGAGAC-3' and 5'-TGAAAG-3', respectively. Therefore, to disentangle this subtlety in binding, the computational models should account for lower-affinity sequences. Even though, most of the computational models today are only based on the higher-affinity sequences. The PSWM model can be parametrized in a way to address low-affinity sequences [23], but this might result in a a very flexible model with a high level of false positive rate.

Another class of the sequences that represent the majority of the possible binding sequences is non-specific binding sites (1.1). The non-specificity of binding has been observed in many different studies [24] [25] [26]. The main picture for TF binding site recognition is that the TF searches its functional, and specific, binding sequences in a greater pool of non-specific sequences. During this scanning of the genome, the TF establishes contacts to different genomic sequences, in a way that is not specific to the local sequence. After a short contact, it then releases the DNA and continue searching for the specific site. Once the specific site is visited, it stays there for relatively longer time in order to perform a function: expression, or repression of a gene. How the protein recognizes its binding site and switches from non-specific interaction to a specific binding mode [27].

1.5 Flanking DNA sequences

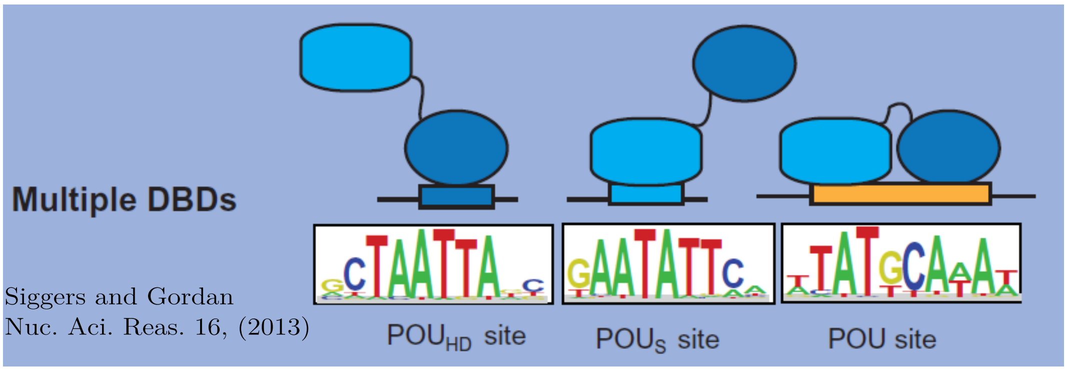
Surprisingly, it is observed that not only the core (or seed) binding sequences are essential for binding specificity, but also the flanking sequences, often non-informative, are exerting a significant effect on binding preference. In our study of the ENCODE ChIP-seq data [28], we have repeatedly observed that PSWM models that include low-information flanking positions, outperform shorter trimmed versions of the same motif. It is also reported by other studies that flanking sequences, when they are taken into account, greatly improve predictions. Interestingly, in a high-throughput study on Gcn4, a master regulator in Yeast, an important role of flanking sequences is illustrated [29].

1.6 Multiplicity in DNA binding modes

Another complicating factor that has been revealed in a number of works is the multiple modes of DNA binding. Over the last few years, high-throughput studies have been increasingly reporting a degeneracy and binding diversity, which leaves the modeling approaches with a great challenge to address the observed binding variability [30] [31] [32] [33] [34]. Some of these intricacies are outlined in this section.

1.6.1 Multiple DNA binding domains

TFs can in general possess a number of distinct DNA binding domains (DBD) [35]. This can result in a flexibility for the DNA binding site recognition. For instance, Evi1, a zinc finger protein, has 10 zinc finger domains which are functioning separately in two autonomous DBD [36]. As a result, it recognizes separate classes of motifs, under different conditions [37]. Yet a more complicated situation has been reported for the mouse TF Oct-1, which contains two DBD. Remarkably, this allows Oct-1 to bind to three classes of motifs, depending on the combinations of its two DBD. In addition to the multiple DBD, each DBD can take on different conformations. For example, for p53, an important regulator for cellular integrity and stress response, the DBD forms two extended and recessed conformations, which can lead to different modes of binding [38].



Siggers and Gordan
Nuc. Aci. Reas. 16, (2013)

FIGURE 1.2: Oct-1 can bind to three distinct classes of motifs, depending on which combination of its two DBD, POU_{HD} and POU_S, is being used [3].

1.6.2 Flexible spacing

One interesting phenomenon regarding TF binding is that some TFs bind to two consecutive *half sites* in the DNA, that are separated by a distance of a few nucleotides, and that this spacing may vary across conditions. This phenomena is known as variable or flexible *spacing*, because the space between two half-sites is not always fixed. It was first reported for the case of leucine zipper (bZIP) proteins [39].

Variable spacing has also been observed for nuclear receptors. It is, for instance, observed that there is a different spacing between the half-site 5'-AGGTCA-3' for binding of peroxisome proliferator activated receptors [40]. This variable spacing also applies to RAR:RXR dimers, which bind to two half sites separated by a varying number of nucleotides [41].

1.6.3 Multimeric binding

Some TFs can form complexes with other proteins, including other TF or co-factors. A recent high-throughput study using HT-SELEX has uncovered the abundance of dimeric binding, for proteins that are previously supposed to act as monomer, such as ELK1 [25]. The observed dimeric sites, are actually observed also in vitro.

1.7 Summary and outline

This chapter has briefly enlisted a number of complications that have been reported, over the last three decades, for TF binding specificity. Starting from a simple model for protein-DNA recognition code, which was a deterministic view to binding recognition, the further researches in this field have revealed more and more discrepancy from this simple view. Today, it has certainly been established that TFs and DNA exhibit a complex landscape of interactions. TFs bind to short, but degenerate sequences under different conditions. In recent years, following a revolution in sequencing technologies, we have been able to collect larger sets of binding sites. This resulted in a remarkable divergence from the simplistic view for binding specificity.

One of the elements that may contribute to the complexity of binding specificity is the pairwise dependency between positions of binding. Unlike, what it has been assumed previously that the binding positions within the TF binding sites are not interacting together, the new observations have repeatedly violated this simplification. Despite the fact that the positional dependency has been proved to impact the binding affinity, these dependencies have not yet been satisfactorily modeled. As we will discuss in Chapter 2, the modeling of dependency is computationally challenging, and current models need to resort to approximations and various simplifications. As a main part of my PhD, we have addressed this difficulty by proposing a novel Bayesian approach toward this problem. We have systematically tested the new model against the classical approach, with no positional dependency, over a large collection of human TFs. We found that our model outperforms the classical model, owing to the incorporation of positional dependencies. In chapter 3, the same model that is used for TF binding site prediction, has been generalized to address other classification problems. Finally, in recent years there has been a great emphasis on enhancer elements, a class of distal regulatory sequences, that play a crucial role in cell-type specific regulation of gene expression. In chapter 4 we propose a new method for incorporating enhancers into computational models of transcription regulation.

Chapter 2

Dinucleotide weighted tensor model

2.1 Introduction

The activity of transcription factors (TF) is highly crucial for the proper cellular response to the external stimuli and stresses. TF function in conjunction as transcription regulators through a myriad of interactions with DNA and other proteins such as cofactors and chromatin modifiers. An integral part of transcription regulation is the *specific* interactions of TF with the short DNA segments, also referred to as TF binding sites (TFBS). In order to promote cooperatively and synergistic effect between TF, it is conjectured that most of TFBS are clustered together within wider regulatory sequences, like promoters and enhancers [42]. Therefore, complete characterization of TF binding specificity is central in further understanding the transcription regulation process. TF often bind to a degenerate set of sequences that makes the task of modeling their binding specificity very challenging [43]. In general, TF-DNA binding free energy is dependent upon numerous interactions at the molecular level [44]; instead of taking into account all these complications, however, many models today–including this work–try to infer TF binding specificity from a set of observed binding sites. The classical approach for inferring TF specificity assumes an independent contribution of binding positions toward the binding affinity. Under this model, the binding specificity of a TF is expressed by a matrix whose entry (n, i) quantifies the binding tendency of a TF towards nucleotide n at position i, independent from any other position. This model is referred to as position-specific weight matrix (PSWM), or weight matrix in short.

Started from mid 70s that Seeman *et al.* [2] proposed the first, and perhaps most idealistic, model for the protein binding site recognition. For more than three decades researches in protein-DNA binding and binding site recognition have revealed numerous complications beyond the simple hydrogen bonding recognition code model. In a recent paper, Siggers and Gordan [3] reviewed the complicating factors in the protein-DNA interaction. Over the past decade, the advancement in experimental techniques, specially high-throughput technologies, has allowed us to pin down a larger number of binding sequences. This provides a unique opportunity to investigate a variety of questions related to the TF binding specificity and build up statistical models for discovering new TFBS. One of the long-standing questions is the extent to which pairwise dependencies (PDs) within binding positions contribute to the total binding affinity of a sequence [45] [46] [32] [47]. In a large-scale study, Bulyk and colleagues assayed 104 distinct mouse TF by using protein binding microarray (PBM) technology [32]. It was shown that the binding energy landscape of human TF is more complex beyond what it had been previously postulated. Notably, a number of assayed TF exhibit strong support for PD within their binding sites. Other studies, such as [45] [46], has earlier pointed out the incompetence of the PSWM model, which partly comes from the pairwise independence assumption. As a final example, Nutiu *et al.* demonstrated that Gcn4p, a yeast amino acid starvation master regulator, shows several strong PD within its binding sites [47]. It is shown that a model that includes PD outperforms the PSWM model to explain the observed binding sequences.

In summary, over the last decade, several studies have identified the role of the PD in TF binding specificity. The impact of PD on binding specificity, at least for a number of TF, is arguably clear; nonetheless the extend in which that PD contributes to the total binding free energy is not yet well understood. Rather, the crux of debate is that how much including dependencies would increase TFBS prediction accuracy. To move forward on this matter more biological evidence is needed, but even more importantly, the complexity of modeling PD and incorporating them into computational models has hindered further developments in this field. As we will discuss below, current computational models have not yet been able to provide a promising replacement over the PSWM model. We believe that before understanding the role of PD in binding specificity, we require a robust and unbiased model that characterizes the PD in TF binding specificity. We are proposing, a novel Bayesian model that rigorously incorporates PD into a computational model. To the best of our knowledge, none of the state-of-the-art models exercise such a proficiency and robustness. Our results show that the proposed model offers a reliable substitution over the classical PSWM model.

2.2 Toward a model for the pairwise dependency

Shortly after the importance of PD became evident, several studies tried to incorporate PD into computational models. The main problem however is that including PD in general is excessively difficult from a computational perspective, which will be explained below. For this reason, current models resort into a variety of approximations and simplifying rules. PD models can be divided into two main categories: models that only allow dependency between certain positions such as neighboring ones [48] [49]. And, models that allow any arbitrary position to be dependent on any other position. In former it is obvious that this approach discards a subset of possible PD. But in the latter, although more realistic, it is computationally expensive to calculate all the PD relations between any pairs of positions. Hence, the models use a range of simplifications and regularization techniques to reduce the computational complexity. In this section, we review some of the models which consider arbitrary PD.

First, a model by Barash *et al.* which uses Bayesian networks to represent arbitrary PD within TFBS [50]. In this model, the maximum likelihood network structure (topology) is learned and employed to characterize the binding specificity. The proposed model entails $3 \times 4 \times N$ free parameters, where N is the length of binding site. Moreover, it is implicitly assumed that the maximum likelihood tree is sufficient to explain the binding variety. This nonetheless can be approximately true if the best tree structure is invariant under different conditions.

Another model that has been proposed by Sharon *et al.* aims to implement a model that includes important PD [51]. The TF binding specificity is modeled by using a *feature-based* approach, called FMM. The idea was to represent binding energy in terms of a weighted sum of features. A feature is a binary statement that can be the appearance of a specific nucleotide at a given position, or it can show a pair of nucleotides in two positions. The authors implemented an iterative approach for learning the best model: starting from an empty model, it tries to incorporate a new feature at a time to improve the model's fit. Using a significant test, only those features which significantly improve the model's fit would be considered to be added to the final model. Obviously, the iterative learning algorithm may result in an over complicated model. To control this undesired behavior, a L_1-regularization is employed that penalizes adding too many parameters into the model. As a result of the L_1-regularization, the final model tends to be relatively sparse. Seemingly, even this regularizing mechanism is not sufficient for controlling the model's complexity; so the authors placed another extra check-point at the end of model learning. Those features that have become *irrelevant* during the last stages of the learning will be removed from the final model. While the term irrelevant is ambiguous in this context, the main reason for doing so, as the authors claimed, is that it is reasonable for features that were added at an early stage to become irrelevant later on.

Finally, a very recent work proposed an alternative model to the PD [52]. This model is very similar in many aspects to the model by Sharon *et al.*. The binding specificity is translated into a statistical mechanics model, an inhomogeneous Potts model, which incorporates information from single and interdependent binding positions. This model is based on interaction parameters, resembling features in FMM approach. The PSWM model is a special treatment of the inhomogeneous Potts model where the associated interaction parameters to the PD were ruled out from the model. Setting interaction parameters to the correct weights plays a crucial role in this model (the same holds for weights of features in the above algorithm). The proposed learning algorithm iteratively constructs the final model; by selecting one single pair of positions and pair of nucleotides at a time (any possible $4^2 \times N(N-1)/2$), and adjusts its associated weight by maximizing the likelihood of the data. The selection of a new interaction term is based on a binomial test. Interactions with a significant deviation from the theoretical binomial distribution are added to the model. Adding more interaction terms, similar to the work from Sharon *et al.*, might lead to overfitting owing to a complex model. To prevent this, the complexity of the model is tracked by the Bayesian information criterion (BIC).

In summary, modeling PD is considered by several works. Using Markovian models, some modeled dependencies by allowing only the interactions between neighboring nucleotides. The problem becomes more complicated, and computationally arduous, if one wishes to allow dependency between any arbitrary positions. As a consequence, all the models so far try to avoid these problems by seeking a range of heuristics and various rules. The authors of this work believe that for at least a pedagogical purpose it is more helpful to provide a formal description of the problem. We discuss the main challenge with PD modeling in the Model section and provide a solution to it. In this paper, we present a novel Bayesian model, dinucleotide weight tensor (DWT) model, that does not have any of the pathologies of the previous models. First, from the modeling point of view, we completely eschew *ad hoc* rules while still managing to computationally tackle taking into account all possible PD. By exploiting from a combinatorial theorem, the matrix-tree theorem [53], our model searches the whole parameter space in a polynomial computational complexity. Most importantly, as it will be shown below, there is no tunable parameter in the DWT model implying that the model is simply applicable in practice. The complexity of the model is automatically adjusted only by the data itself, meaning that there is no need for extra complexity metrics or regularization techniques. As a result, the PD contribute to the scoring if they are supported by data, otherwise the DWT model automatically reduces to the PSWM model. This prevents the model from overfitting. In order to compare the DWT and PSWM models, we developed a testing framework aiming to remove the chance of systematic biases toward any of the approaches. We examined a comprehensive set of human ChIP-seq, consisting of the *in vivo* binding data for 78 human TF. The results show that the DWT model improves over the PSWM model for about one third of the TF. For the rest of TF, the DWT model, as it is expected, reduces to the simpler non-dependent model and performs similar to the PSWM model. In addition, a new graphical representation of the binding specificity is proposed that includes information about the important PD to the traditional sequence logos.

2.3 The dinucleotide weight tensor model

In the current section, we present a Bayesian probabilistic model that quantitatively characterizes TF binding specificity. As a generalization to the PSWM model, we name our model dinucleotide weight tensor (DWT) model. Following the same approach which was introduced originally by Burger and van Nimwegen [54], to predict protein-protein interactions from the amino acid sequences, here we adapt the model for the TF binding site prediction.

Let S indicate a set of nucleotide sequences of the fixed length l. If the sequences in S are aligned together then the likelihood of column i, denoted by S_i, is defined as

$$P(S_i) = \int P(S_i|\omega_i)P(\omega_i)\mathrm{d}\omega_i \tag{2.1}$$

where ω_i denotes the probabilities of different nucleotide at position i, it is equivalent to the i^{th} column of the weight matrix, The reason that we integrate over ω in the above equation is because the matrix ω is unknown. We have chosen Dirichlet's priors for the prior probability $P(\omega)$. The likelihood of a column S_i, given ω is given by

$$P(S_i|\omega) = \prod_{x\in\{A,C,G,T\}} \omega_{i,x}^{n_i^x} \tag{2.2}$$

where n_i^x is the number of the appearance of letter x at position i. Similarly we can define the likelihood function of a pair of columns $S_{i,j}$. We however need to integrate over a tensor ω' that each of its entries encode the probability of having a pair of letters at a pair of columns. Having defined the basic likelihoods for the columns and pair of the columns, the total likelihood of sample S, under i.i.d condition, is represented by the following equation – that models individual positions being *dependent* on another binding position.

$$P(S|\pi) = \prod_{i=1}^{l} P(S_i|S_{\pi(i)}) \tag{2.3}$$

In the above equation, π is a function that maps position i to another position $\pi(i)$ which there is a PD between them. In general, we do not have a prior information about the positional dependencies (PD) and therefore the dependencies, which is encoded by π, is unknown. For this reason, the correct practice of the probability theory instructs us to marginalize over the unknown parameter π.

$$P(S) = \frac{1}{l^{l-2}} \sum_{\pi} \prod_{i=1}^{l} P(S_i|S_{\pi(i)}) \tag{2.4}$$

In the above equation, the pre-factor $1/l^{l-2}$ is associated to a uniform prior over the all possible trees of size l, because the number of nonidentical trees are l^{l-2}. Performing the sum in equation 2.4 is computationally expensive, however. Since the number of all possible PD sets grows exponentially, previous methods have simplified the task by taking into account the maximum likelihood (ML) estimate of the dependencies. In the DWT model, on the other hand, we calculate the sum in equation 2.4 exactly. The equation 2.4 is mathematically equivalent to the following form, which will serve an important role in our model.

$$P(S) = \frac{1}{l^{l-2}} \sum_{\pi} \sum_{\pi} \prod_{i=1}^{l} R_{i,\pi(j)} P(S_i) \tag{2.5}$$

where $R_{i,\pi(j)}$ is an entry of a symmetrical matrix $\mathbf{R}$ that its entries were defined by

$$R_{i,j} = \frac{P(S_i, S_j)}{P(S_i)P(S_j)} \tag{2.6}$$

Matrix $\mathbf{R}$ can be regarded as the matrix of PD scores between any pairs of positions. In fact, for large sample size, the $\log_2(R_{i,j})$ converges to the mutual information between positions i and j. The detailed derivations of $P(S_i, S_j)$ and $P(S_i)$ are given in chapter 3. In order to solve the aforementioned computational difficulty in performing the sum over all possible π, the matrix form of equation 2.5 provides us a way to exploit a well-known theorem in combinatorics known as the matrix-tree theorem, or Kirchhoff's theorem [53]. Here, we refer to this theorem as the matrix-tree theorem. To establish the connection between our problem and combinatorial graph theory, we represent π as a *spanning tree* where nodes are the binding positions and the weights of connections (edges) are provided by the matrix $\mathbf{R}$. Note that the choice of the tree topology in the above model is to avoid circular reasoning in our inference.

The number of nonidentical spanning trees over n nodes follows an exponential function n^{n-2} [55]. Therefore, a brute-force approach for calculating the sum in 2.5 for arbitrary large graph size is computationally intractable. This computational intractability is alleviated by utilizing the matrix-tree theorem which states that the total number of the spanning trees inside a graph G is equal to any *minor* of the graph's *Laplacian* matrix [56]. The Laplacian matrix L of a graph is defined as the graph's adjacency matrix reduced from its degree matrix. The minor $M_{i,j}$ of a matrix A is the determinant of a smaller matrix A', which is obtained by removing a row i and column j from the original matrix A. To this end, as a result of the matrix-tree theorem application, we can rewrite the likelihood function 2.5.

$$P(S) \propto M_{1,1}(\mathbf{L}) \times \prod_{i=1}^{l} P(S_i) \tag{2.7}$$

In the above equation $\mathbf{L}$ is defined as the Laplacian matrix of $\mathbf{R}$, and $M_{1,1}$ is equal to the determinant of the matrix $\mathbf{L}'$ where the first row and column of the Laplacian matrix $\mathbf{L}$ is removed. Note that the above equation is invariant under $M_{i,j}$ for any choice of $1 \leq i, j \leq n$. The first term at the right hand side of the equation 2.5 is equal to the weighted sum of any possible spanning trees weighted by the product of the weights of edges. Although the likelihood function 2.5 and 2.7 are mathematically equivalent, the latter has a useful practical implication. Instead of computing the sum of n^{n-2} trees, we need to compute the matrix determinant which can be numerically performed in $O(n^3)$. One practical issue in calculating equation 2.7 is the wide range of values in the matrix $\mathbf{R}$ that may potentially destabilize the determinant calculation. In order to make sure of the numerical stability we have used a rescaling method on the dependency matrix (section 2.3.1). To summarize, the main strength of the DWT model is that, instead of assuming one fixed set of PD, it rigorously enumerates any possible arrangements of the PD.

The model above, by incorporating PDs, allows a higher flexibility for the binding specificity. If , for instance, a nucleotide x is disfavored by the weight matrix at a certain position i, the mismatch from a favorite nucleotide y costs $\log(p_{i,y}/p_{i,x})$, where $p_{i,x}$ is the probability of nucleotide x at position i and $p_{i,y} > p_{i,x}$, from the binding free energy. Recent high-throughput studies, however, have revealed that it is often not as dramatic the change of the binding free energy for a mismatching sequence as it is modeled by the PSWM model. Under the DWT model, on the other hand, the mismatches from the consensus motif will not be punished as harshly as the PSWM model. It can be a

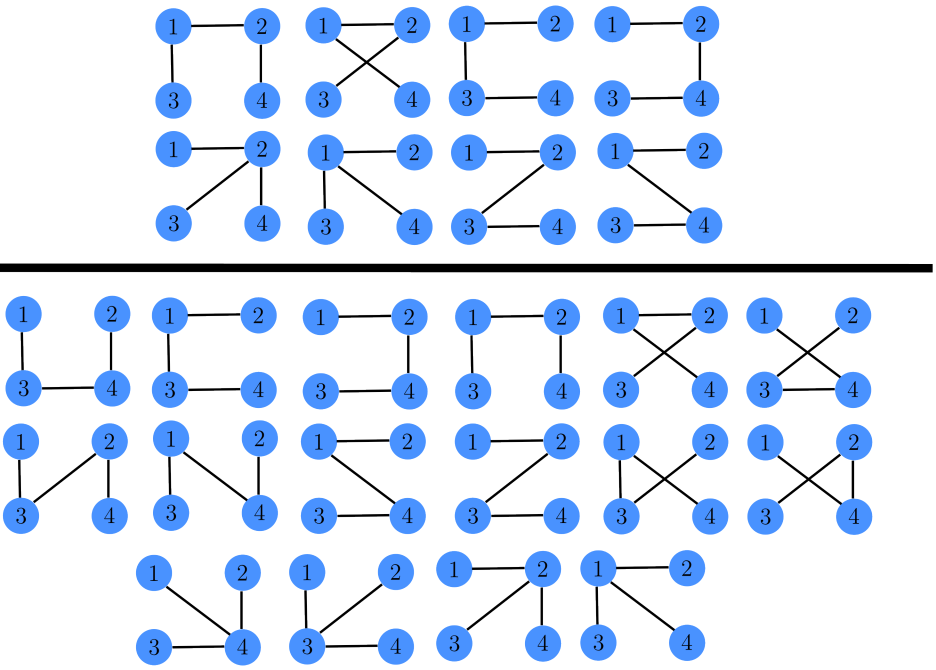


Figure 2.1: The probability of the nodes 1 and 2 being connected. In the numerator, all the trees that have a link between nodes 1 and 2 were depicted. The denominator shows all possible trees that can be made by four nodes.

sequence with an incompatible nucleotide at given position shows a high binding affinity, when the mismatched nucleotide is paired with another nucleotide at a different position,

Of further practical interest is to find the dependent positions with the help of the DWT model. We address this by calculating the posterior probability that a pair of the binding positions are mutually interdependent. The probability of positions i and j being dependent on each other is equal to the fraction of the spanning trees with i and j being connected to the total trees. This idea is shown in figure 2.1, where we are interested to find the fraction of the spanning trees with a connection between nodes 1 and 2, to the total number of nonidentical spanning trees which is possible to build with four nodes. In our model, this is carried out by making matrix $\mathbf{R}^{i-j}$ that adds the rows i and j into a single row and the same operation with the two columns. This intuitively means that the two nodes are shrunk into a single super node. Therefore, the posterior of the positional dependency between positions i and j, given the dependency matrix $\mathbf{R}$, is defined by

$$P((i,j)\,|\,\mathbf{R}) \;=\; R_{i,j} \times \frac{M_{k,l}(\mathbf{L^{i-j}})}{M_{1,1}(\mathbf{L})}\ ,\ \text{for } k,l \neq i,j \tag{2.8}$$

where the numerator is the sum over all the trees with the edge (i,j). Using the above equation we can find out for any pairs of the binding positions if there is a PD relationship between them. Our tests (Result section) show that the nearer positions, such as adjacent positions, are more likely to be dependent.

Finally, and perhaps most importantly, the task of the TF binding site prediction using the DWT model. By calculating $P(s|S)$, under the likelihood model in equation 2.7, we can evaluate any sequence s of being a binding site. By definition, $P(s|S)$ is equal to the ratio of the join probability distribution $P(s,S)$ to the distribution $P(S)$. The join distribution $P(s,S)$ is similar to the $P(S)$, where a new sequence s is added to the initial sample S.

$$P(s|S) \;=\; \frac{M_{1,1}(\mathbf{L}_{new})}{M_{1,1}(\mathbf{L})} \times \prod_{i=1}^{l} f_i^{s_i} \tag{2.9}$$

The $\mathbf{L}_{new}$ is the Laplacian of the new dependency matrix which is obtained by adding s to the sample S. The $f_i^{s_i}$ is the frequency of the letter s_i at the position i in the initial sample S. For the detailed derivation of the above equation refer to chapter 3.

A main property of the equation 2.9 is that it decomposes two aspects of binding specificity: independent positional tendency toward different nucleotides, and PD between the binding positions. The first term characterizes the total PD. The second term in equation 2.9 designates the frequency of the nucleotide s_i at position i independent from any other positions. This term is in fact exactly the same as the PSWM model. Therefore, equation 2.9 can be regarded as a generalization of the PSWM model. Once the evidence of the PD is low, the PD ratio converges to one and hence does not contribute to the final score of the sequence. On the contrary, when there is a high evidence for the PD, the PD ratio, combined with the PSWM score, determines the probability of binding. As a result, whenever the evidence for the PD is negligible, instead of suffering from overfitting, the DWT model reduces in an unsupervised manner to a simpler model – that is the PSWM model. In brief, the DWT model above, unlike its counterparts, relies on no arbitrary assumption or *ad hoc* rules, and it considers all possibilities for the PD in a fast computational time. The complexity of the final model is only controlled by the support of the PD in the data. Therefore, there is no need to monitor and control the model's complexity. These properties make our model easily applicable in practice with no need to account for any unnecessary technicalities.

2.3.1 Re-scaling of the dependency matrix

As it is discussed so far, our model relies upon the concept of the matrix-tree theorem. Accordingly, our model needs to calculate the determinant of dependency matrix, numerically. From the practical point of view, calculating the determinant may lead to numerical instability. For example, when the numbers inside the matrix span a very wide range of values, from very tiny numbers to huge values, their multiplication can potentially lead to the overflow. To eliminate the possibility for the numerical errors, we have used the following equation to scale the matrix entries.

$$H_{ij} = \exp\big(\frac{k}{\log R_{max}} R_{ij}\big) \tag{2.10}$$

where R_{max} is the biggest entry in the $\mathbf{R}$ matrix, and k is a predefined factor to adjust the differences of the values in the new matrix $\mathbf{H}$. The appropriate value of k depends on the machine's constraints. In this study, we have set $k = 15$.

2.4 Results

2.4.1 Graphical representation of the pairwise dependency

We propose a graphical representation of the binding specificity. This new method is a generalized form of the classical sequence logos. The proposed representation method has two main graphical components; the first component shows the nucleotide tendency of the TF at each position, which is exactly similar to the classical sequence logos. The second component provides information about the dependency between positions. Since this representation method has the PD in addition to the sequence logos, it is named dinucleotide logo (DiLogo). For example, figure 2.9a shows the DiLogo for the human TF CEBPB. Each DiLogo has four layers that the top layer is the sequence logo of the TF, as it is in the basic sequence logo. The bottom layer shows the matrix of the PD posterior probability (Supplementary text) for dependency between any given pairs of positions. Since the PD is symmetrical, showing only a half of the posterior matrix is sufficient. The column of the dependency matrix is associated to the labels of the nodes in the linear graph on top of the matrix. The linear graph shows the strong PD where the two positions are connected if the posterior of the PD is higher than a predefined cutoff.

In figure 2.9a we set 0.9 as the posterior cutoff. Note that links are directed, which we call the nodes (positions) i and j as the parent and child nodes, respectively, if there is a direct link from i to j. In the next layer, the nucleotide tendency at the children positions conditioned on the state of the parent positions are shown. The table rows represent the state of the parents. For instance, in figure 2.9, if there is an A at position 2 (the parent of the position 3) there is a high chance to have a T at position 3. On the contrary, if there is T at position 2, the chance of G and A at position 3 is higher. The height of the letters, similar to the classical sequence logos, in the table corresponds to the amount of the information for that letter.

2.4.2 The analysis setup

To pin down the role of PD in defining binding specificity we have developed a comprehensive testing framework to compare the two modeling approaches: the DWT model and PSWM model. To test the models we need to have a set of sequences that are experimentally validated for having binding sites for a specific TF. As a result, we are interested to quantify the performance of the models in recognizing the true binding sequences. Most importantly, care must be taken for any potential biases toward any particular of the approaches. Put it in other terms, the test should not favor a model over the another model due to a systematic bias.

The proposed test setup uses an iterative algorithm for model refinement; each of the models separately learn the best model on the training dataset and the resulted models' performance will be inquired on the test dataset Here, we are introducing an information-theoretic measure that quantifies the performance of each model in distinguishing the true sequences from decoy sequences. In fact, we are measuring that how much of the initial entropy is resolved due to the assigned scores by the models (refer to section 2.4.6). Note that by decoy sequences we are referring to sequences that although similar to the true sequences, they devoid cognate binding site (see section 2.4.5).

In the test setup, we use the top 1000 enriched sequences from the chromatin immunoprecipitation by sequencing (ChIP-seq) method as the *true* binding sequences. ChIP-seq method is a common technique that measures the binding affinity of the genomic regions. A large number of studies have published ChIP-seq data for many TF in a vast range of organisms and cell-lines. ENCODE project, for instance, published ChIP-seq data for many human TF on a different cell types [28]. This technique, however, does not pinpoint the exact location of the binding site; instead, it retrieves larger genomic regions, with a few hundred base-pairs resolution, that the TF was bound, directly or indirectly. Despite the fact that all peaks in ChIP-seq may not be resulted from the direct interaction of the TF to the associated DNA region, provided that the TF can be presented because of a multi-meric interaction with another TF, we believe that the top

1000 peaks contains a significant fraction of the sequences with the desired TF's binding sites. Moreover, 1000 sequences provide us with enough material for a robust statistical analysis.

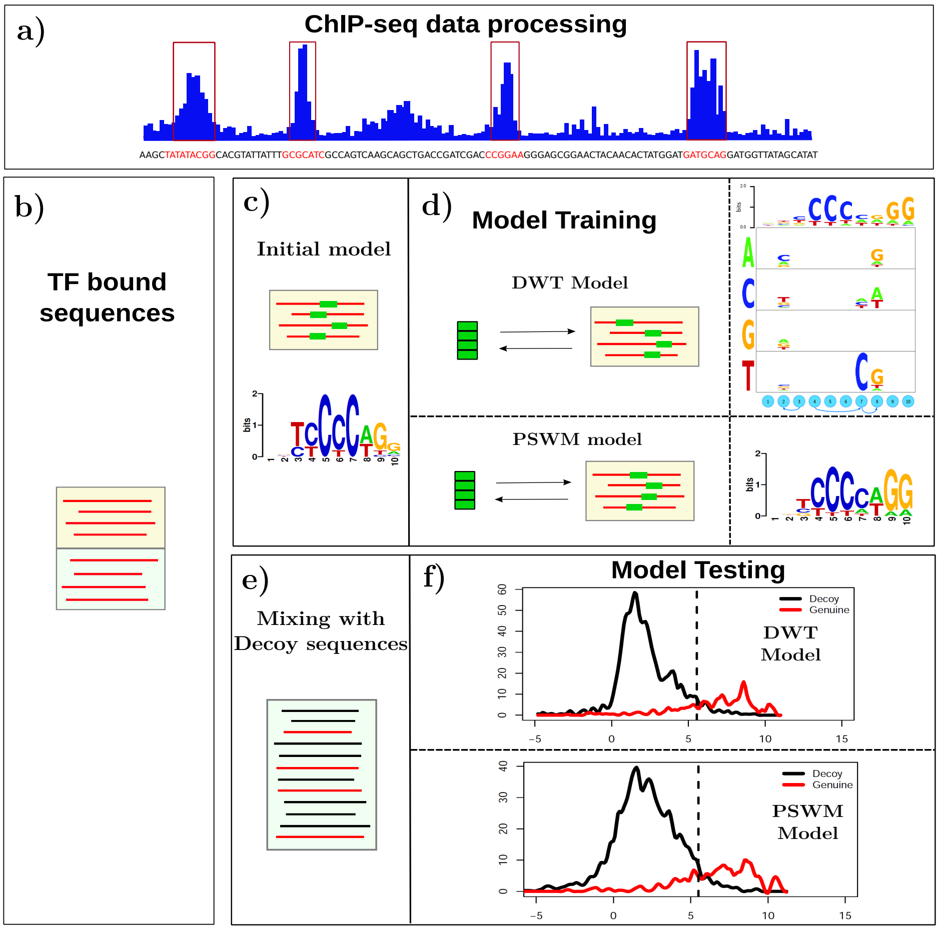


FIGURE 2.2: **The Testing Framework for PD.** 1) We start with extracting the sequences from the genomic regions which are highly enriched under the ChIP-seq assay. In this work, we selected top 1000 ranking sequences. 2) The selected sequences then divided into two separate sets of equal size: training and test sets. 3) An initial motif is then used to predict the binding sites that will be used to build the initial models. 4) Both models separately refine the motif in an iterative manner until convergence. 5) The resulted models from the iterations are tested over the test set that is mixed with the decoy sequences. The decoy sequences, although similar to the test sequences from different aspects, are supposedly devoid of the TF binding sites. 6) The DWT and PSWM models, by using the learned models from the iteration, assign a score to every sequence. The distribution of the scores is shown here where the black line indicates the scores of the decoy sequences and the red line for the genuine sequences' scores. By setting a cutoff over the scores, we call a sequence as genuine if its score is above the cutoff and decoy with the score below the cutoff. Therefore, we measure how well such score cutoff performs in order to distinguish the real genuine sequences from the decoy ones.

Figure 2.2 summarizes the test setup. This framework is also implemented as a website that by uploading the binding sequences (in FASTA format) and an initial motifs performs the PD analysis. Here we explain the different stages as it appears in the figure, but more details will be provided in following sections.

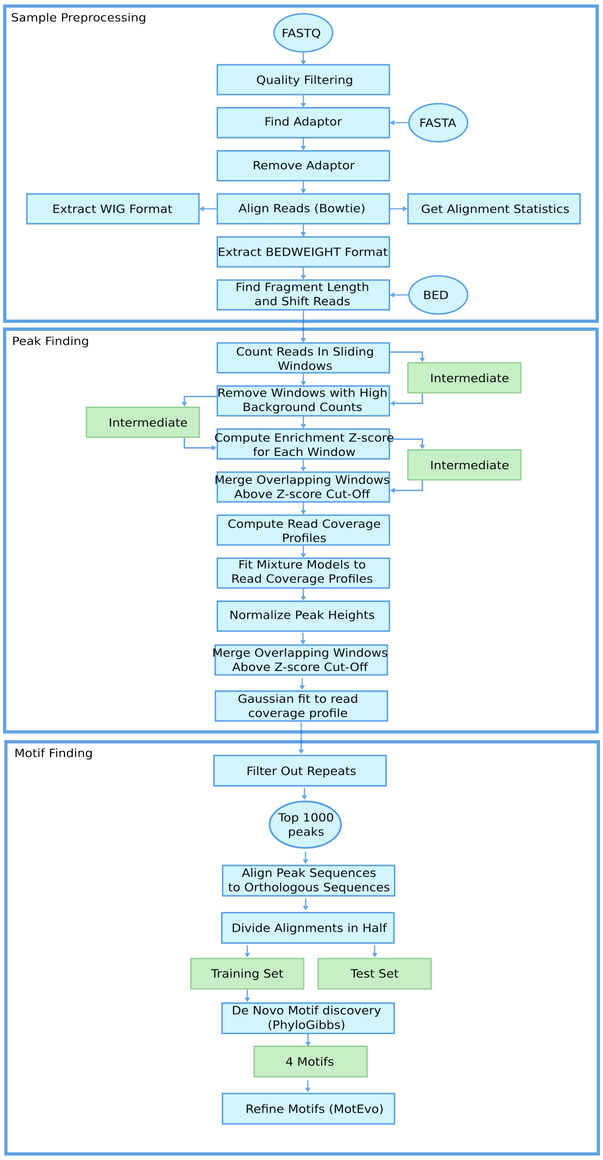

Sample Preprocessing
FASTQ
Quality Filtering
Find Adaptor
FASTA
Remove Adaptor
Extract WIG Format
Align Reads (Bowtie)
Get Alignment Statistics
Extract BEDWEIGHT Format
Find Fragment Length and Shift Reads
BED
Peak Finding
Count Reads In Sliding Windows
Intermediate
Remove Windows with High Background Counts
Intermediate
Compute Enrichment Z-score for Each Window
Intermediate
Merge Overlapping Windows Above Z-score Cut-Off
Compute Read Coverage Profiles
Fit Mixture Models to Read Coverage Profiles
Normalize Peak Heights
Merge Overlapping Windows Above Z-score Cut-Off
Gaussian fit to read coverage profile
Motif Finding
Filter Out Repeats
Top 1000 peaks
Align Peak Sequences to Orthologous Sequences
Divide Alignments in Half
Training Set
Test Set
De Novo Motif discovery (PhyloGibbs)
4 Motifs
Refine Motifs (MotEvo)

FIGURE 2.3: The flowchart of the ChIP-seq data processing pipeline. Starting from the quality filtering and removing the sequence adapters, the sequenced reads were mapped to the reference genome, e.g. hg19. The problem with multi-mappers here is dealt with defining a weight to each read that maps a multiple region. If a read maps to n genomic loci, it will be counted $1/n$ at each of the regions. This is all encoded in an invented format, called BEDWEIGHT format. After this phase, using a statistical model we assign a Z-score to each of the genomic region and select the ones with a score bigger than a cutoff, to be set from the dataset.

The ChIP-seq data was processed using CRUNCH, an in-house ChIP-seq data analysis pipeline (to be published). Figure 2.3 summarizes the ChIP-seq data processing pipeline. We have used Bowtie [57], an ultra-fast and memory-efficient algorithm for short read

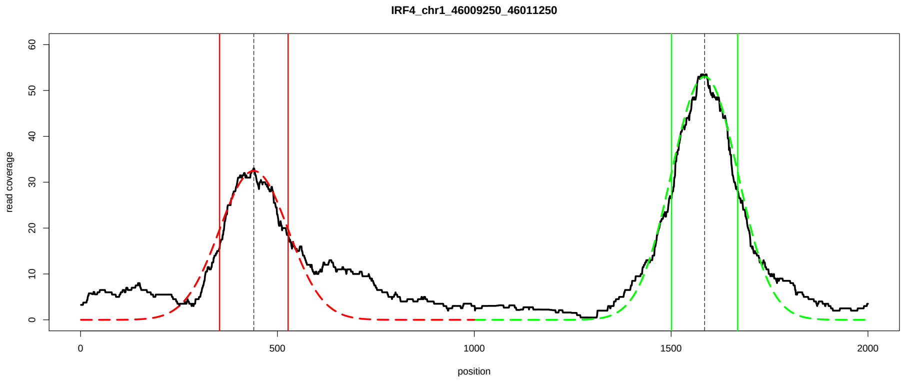


FIGURE 2.4: Read coverage profile of IRF4 in a 2 kbp long region. The actual read coverage is shown by the solid black line, and the two Gaussian fits by green and red dashed lines.

alignment, to map the sequenced reads to the reference genome. The Bowtie options are given below.

```
bowtie -f -v 3 -a -B 1 --quiet --best --strata
```

The issue with the multimappers is treated by assigning a weight to each read. Reads that are mapped to n genomic regions receive a weight of $1/n$ at each of the associated regions. After this, we have a statistical model, originally developed for FANTOM4 project [58], to calculate the significance of the regions. In this study, we have selected the top 1000 regions according to the whole-genome Z-scores. After peak calling and calculating their Z-scores, we have selected the top 1000. peaks. The selected peaks were further processed by extracting sub-regions that contain high density of sequenced reads (2.4).

The resulted sequences were then divided into two disjoint sets: training and test sets. An initial set of predicted binding sites is used to start the iterative model refinement. The initialization is performed by running a de novo motif discovery algorithm, called PhyloGibbs, which uses phylogenetic information [59]. Multiple alignments of the sequences were obtained by T-Coffee program [60]. The resulted binding sites from PhyloGibbs on the training set were then used to build the first DWT and PSWM models. For those TF that we already have a motif in databases such as SwissRegulon [61] or JASPAR [62], but under several tests we have found that the pipeline motifs, although quite similar to the database motifs, are better models in the context of the processed ChIP-seq dataset. Hence, the pipeline motifs were used in the consequent analysis.

After the initial models are made, by sliding a window over the training sequences each model separately assigns score to each window. The background frequency model is set according to the nucleotides frequencies in the sequences. Then the windows with the relatively high score are selected, and the selected windows serve to build a new model. This iterative process continues until convergence. For the DWT model, the convergence is tested by calculating the distance between the two dinucleotide frequency matrices in the last two rounds. Whereas for the PSWM model, the convergence is checked by the distance of the single nucleotide frequency matrices. At the end of this iteration every model proposes the best refined model for predicting the binding sites.

After the final models are learned, each model is examined over the test sequences. The test sequences are mixed with decoy sequences that almost likely do not possess binding sites. The decoy sequences, however, are similar to the test sequences from different aspects such as nucleotide frequency and sequence length. Here, we chose to preserved dinucleotide frequencies that might even make it more difficult for the dependency model whenever there is dependency between adjacent positions. In the current setting, the number of decoy sequences is four times bigger than the number of true sequences. The two models assign a score to each sequence in the test set including the decoy sequences. We calculate the binding affinity of each sequence as its score. The binding affinity $E(S)$ of a sequence S is defined as:

$$E(S) = \log\big(\sum_{\forall i} \exp(score(S_i))\big) \tag{2.11}$$

where s_i denotes a site inside the larger sequence S. Note that we consider both positive and negative strands of the sequence when calculating the binding affinity. A similar approach is also used to compare experimental methods for measuring TF binding specificity [63].

Finally, given the scores of all sequences we measure the classification accuracy of the true binding sequences from the decoy ones. The score distribution for both classes of sequences were represented by histograms in the figure 2.2. We then ask how much the two classes of the sequences are separable by setting a cutoff over the sequence scores, shown by the dashed line in the figure 2.2. In other words, if a score cutoff indicates the class of sequence, how well this cutoff can separate the decoy and true sequences. After having the cutoff, that provides the maximum separation between the genuine from the decoy sequences, we propose a robust measure that in an information-theoretic sense quantifies the performance of each model. The proposed measure calculates the amount of the explained or resolved entropy inherent in the given decision-theory problem (see section 2.4.6). Equivalently, the amount of information gain due to the usage of a model, if one wants to distinguish the genuine from the decoy sequences. Hence, higher the value is, better the performance of a model is in stratification of the two classes of the sequences. The best score is one which implies the there is absolutely no uncertainty in distinguishing the decoy and true sequences. A totally random predictor would have a score of zero. As it is shown in section 2.4.7, the DWT model is always performs at least as good as the PSWM model, and for several TF, including PD information, resulted in a more accurate model on the binding specificity. An important implication of this result is that the DWT model does not suffer from overfitting, or mistaking noise for signal, which is owing to the rigorous integration of all the dependency trees. Once the evidence of the PD is low, the DWT model automatically reduces to the simpler PSWM model.

2.4.3 Iterative model refinement

The test framework is summarized in figure 2.2. Note that, the two models are learned separately via an iteratively approach. The iteration algorithm, except the convergence test, is similar for both models. The iteration is represented in 1, which requires a start weight matrix as well as a set of sequences, which in our setting is the training sequences. By running MotEvo [64] over the sequences, we select the sites with a posterior at least 0.5. This selected set of sites constitutes the initial model. In line 9 of 1, a model is created using the predicted binding sites. A model can be either the PSWM model or the dependency model. Next, by maximizing equation 2.15 with respect to a free parameter c a score cutoff is fitted. This cutoff is then used to convert the log-likelihood score $L(s)$ of the sequence s, to a probability space by applying the following equation:

$$P(s_i, c) = \frac{e^{s_i - c}}{1 + e^{s_i - c}} \tag{2.12}$$

As it is shown at line 14 of 1, using (2.12), we select the sites with a minimum probability of 0.5. The predicted sequences at run i is compared with the predictions from the last run $i-1$, and if the difference (distance) is less than a predefined value ϵ the iteration will be terminated. The comparison, however, is performed differently for each of the models. For the PSWM model the distance of the simple frequency matrices is calculated. Each column k of S_i is associated to a column k in the frequency matrix and four rows that are representing the normalized frequency of the nucleotides at each column. The distance of two frequency matrices W_i and W_j is measured by,

$$dist(W_i, W_j) = \sum_{m=1}^{l} \sum_{\alpha \in \{A,C,G,T\}} \frac{2|W_i[\alpha][m] - W_j[\alpha][m]|}{W_i[\alpha][m] + W_j[\alpha][m]} \tag{2.13}$$

Testing for the convergence in the dependency model is done on four-dimensional matrices that represent the normalized pair-nucleotides frequency on every pair of the positions. Each entry (m, n, α, β) of this matrix determines the frequency of the pair (α, β) at positions (m, n). Similar to (2.13), the distance between two four-dimensional matrices Z_i and Z_j is defined according to the following.

$$dist(Z_i, Z_j) = \sum_{m=1}^{l} \sum_{n=1}^{l} \sum_{\alpha \in \{A,C,G,T\}} \sum_{\beta \in \{A,C,G,T\}} \frac{2|Z_i[\alpha][\beta][m][n] - Z_j[\alpha][\beta][m][n]|}{Z_i[\alpha][\beta][m][n] + Z_j[\alpha][\beta][m][n]} \tag{2.14}$$

The iteration will be terminated if the distance between the two last matrices is less than ϵ. In our tests, we set $\epsilon = 0.05$. Finally, the iteration algorithm 1 returns the last set of predictions as well as the last fitted value for the score cutoff c.

Algorithm 1 Iteration method

```
Require: $D$ and $wm$ #$D$ is indicating the input sequences and $wm$ the initial weight matrix
$i \leftarrow 1$    #number of iterations
$S_i \leftarrow [\,]$
for $s$ in $[$ MotEvo($D$, $wm$) $]$ do
    if $P(s) \geq 0.5$ then
        $S_i \leftarrow s$
    end if
end for
repeat
    $M \leftarrow Model(S_i)$
    $c \leftarrow argmax_c \left[F(S_i, c)\right]$    #fit the cutoff
    $i \leftarrow i + 1$
    $S_i \leftarrow [\,]$
    for $s$ in $D$ do
        if $\left[\exp(L(s) - c)/(1 + \exp(L(s) - c))\right] \geq 0.5$ then
            $S_i \leftarrow s$
        end if
    end for
until distance($S_i$, $S_{i-1}$) $\leq \epsilon$
return $S_i$ and $c$
```

2.4.4 Fitting the score cutoff

As it has been mentioned above, at the end of each iteration step, we select a number of binding sites. This selection is based on setting a score cutoff over the sequence scores that are calculated by (2.15). If the likelihood of the total scores is written as,

$$F = \sum_{\forall i} log\Big[\frac{P(s_i|B)}{(1+e^{-c})} + P(si|M) \times \frac{e^{-c}}{1+e^{-c}}\Big] \tag{2.15}$$

where $P(si|M)$ is equal to the score of sequence i according to the model and $P(s_i|B)$ is the sequence score under the background model, which here is a Markov process of order zero. The best score cutoff would be the value of c, which maximizes the above equation. Therefore, by determining the value of c at each round of the iteration, we select a set of binding sites by choosing those sites with a minimum probability of 0.5 (equation (2.12)).

2.4.5 Test set and decoy sequences

Another important point in the explained motif refinement procedure in figure 2.2 is the way we make the test set. A fraction of the test set consists of the true regions, the ones with the real binding site. The other part of the test set is made of noise sequences. This noise or false regions are generated by a Markov model that is trained over the true regions. Therefore, pair frequency of letters is conserved. The other scenario that we have also considered is to make the false regions by selecting some genomic regions that were not enriched in the ChIP-seq experiment. We have found that this method, although works acceptable for some TF, runs into problems for many TF. This maybe due to the fact that the motif refinement instead of capturing the real motifs, goes to a wrong direction by learning some sequence features that are not relevant to the actual protein sequence specificity of interest. This then makes the models to mistake noise for signal, by just learning the irrelevant sequence features. This behavior apparently is not desirable in our purpose.

We have found that the sequences, which are generated by the Markov model, work stable in the different circumstances. One choice that we made is the order of Markov process, we have tested different Markov orders and we found the change of order from one to higher ones does not change the final results considerably. Therefore, we kept the Markov order one for all the tests that we have performed.

2.4.6 Performance assessment

At the end of the iteration, which is explained above, each of the models predict a set of final binding sites in the training set. This set of predicted binding sites serve to create the final model for testing. We have designed an experiment, where the experimentally validated sequences (genuine) were mixed with a number of random sequences (decoy).

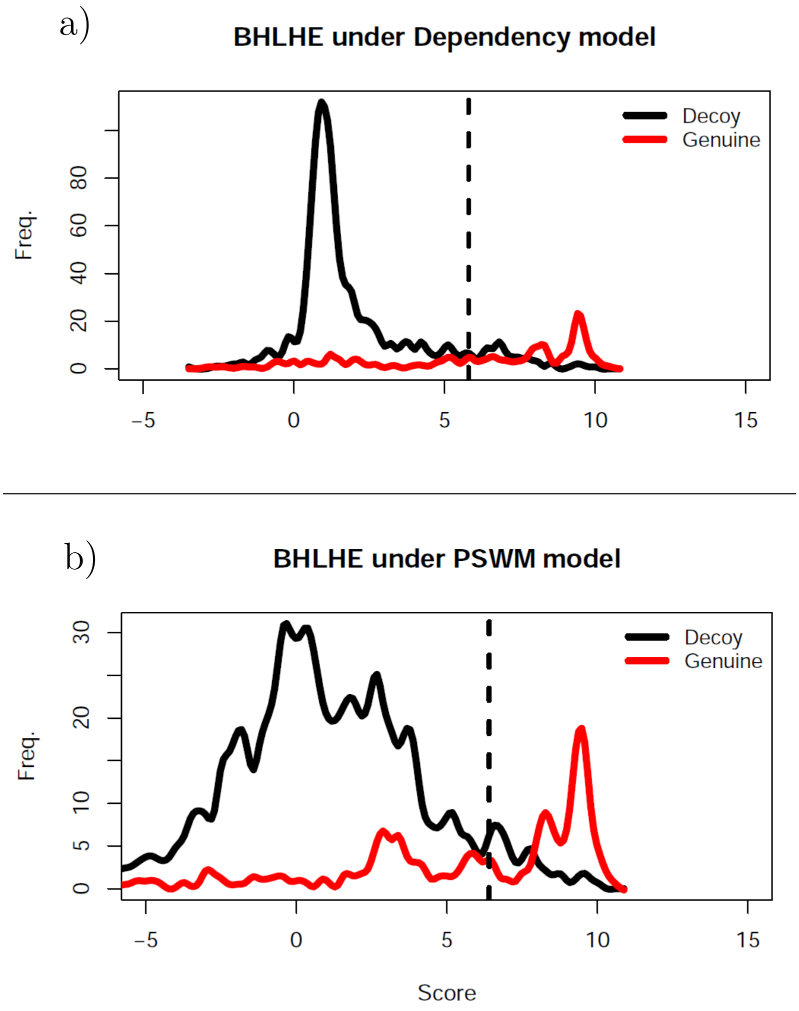


FIGURE 2.5: Histogram of scores. a) Using the dependency model to score the sequences and (b) where the PSWM was used

For each genuine sequence we have four decoy sequences in the mixture. We assume that the chance of having a binding site in the decoy sequences is very low. The decoys were generated by a Markov model of order one which is trained over the genuine sequences. Hence, the decoys have the same dinucleotide frequency as the genuine sequences, but they are expected to be mostly devoid of the true binding sites. Next we challenged the models to predict the binding sites in all the sequences. Note that we have used the same background frequency that had been fitted over the training set. Having the log-likelihood ratio for each binding site in the sequences in both forward and reverse strand calculated by the two models, we assigned a score to each sequence by the following equation.

$$Score(S_i) = \log\left(\sum_{k=1}^{|S_i - l|} \exp\left(L(s^{+,-}_{k,k+l})\right)\right) \tag{2.16}$$

The $s^{+,-}_{k,k+l}$ indicates a sequence, in forward and reverse strand, starting at position k with length l, which is the length of the binding site. The models are expected to assign smaller scores to the decoy sequences comparing to the assigned scores to the genuine sequences. For this reason, we ought to get a bimodal distribution of the scores from (2.16), where the two peaks of the distribution represent two classes of sequences: the bound and not-bound sequences. The shape of the final distribution however depends on different factors, such as the quality of the experiment, the sequence specificity of the TF and so on. Therefore, the score distribution will not necessary follow a bimodal form. This, for instance, is demonstrated in figure 2.5 where the score histogram of the two models were drawn. The two classes of sequences are represented by different colors, red line indicates the score of the genuine sequences and black line shows the score distribution of the decoy sequences. As it is clear, the genuine sequences have relatively higher scores than the decoy sequences, but there is still some genuine sequences with low scores that resemble the decoy sequence scores. It is probably because the TF was not directly bound to these sequences, and instead it formed a dimer with other TF(s) which was directly binding to the sequences.

The above experimental setting is equivalent to a decision-theory problem that is how well we can categorize observations X into two classes c_1 and c_2. Here we would like to collect the genuine sequences from the decoys based on the observed scores from equation (2.16). Therefore, Different models are evaluated according to how well they perform in separating the score distributions of the genuine sequences $P_1(X)$ and decoy sequences $P_2(Y)$. Intuitively, apreferred model is the one that brings the lowest overlap between the two distributions $P_1(X)$ and $P_2(Y)$. As a measure of the separation, here we propose the likelihood ratio of the scores being picked up from a joint distribution to the two independent distributions.

$$I = \frac{P_1(X)P_2(Y)}{P(X,Y)} \tag{2.17}$$

Finding the distributions in (2.17) requires to bin the scores, but due to the finite sample size this may affect the final measure. Therefore, we suggest a robust way of calculating the distributions by removing the possible biases due to the binning effects.

$$P_1(X) = \int_\rho (\prod_i^{|\rho|} \rho_i^{n_i}) P(\rho) d\rho \tag{2.18}$$

The ρ is the distribution function resulted from the binning, and n_i is the number of scores falling in the i^{th} bin. The product term in (2.18) is the likelihood. Assuming a Dirichlet prior on the ρ, and using the same integral identity (**??**), results in:

$$P_1(X) = \frac{\prod_i^{|\rho|} \Gamma(n_i + \zeta)}{\Gamma(N + |\rho|\zeta)} \tag{2.19}$$

N is the total number of genuine sequences, which is equal to 500 here. The quantity ζ serves as a pseudo-count for the bins, which we set it to 0.5 here. The same applies for the decoy sequences distribution $P_2(Y)$ as well as the joint distribution $P(X, Y)$. As a result, the closed-form of (2.17) is:

$$I \propto \frac{1}{N + M} \prod_i \left[\frac{\Gamma(n_i + \zeta)\Gamma(m_i + \zeta)}{\Gamma(n_i + m_i + \zeta)\Gamma(\zeta)} \right] \tag{2.20}$$

The M and N are the number of the decoy and genuine sequences, respectively. The above equation in fact is a measure of the divergence between two distributions of counts (or scores). I is equal to one, if the two distributions are completely independent. On the other hand, the value of I becomes smaller as the two distributions diverge. For large number of counts it can be shown that $\log(I)$ converges to Jensen-Shannon divergence $JS_{\{\pi,1-\pi\}}(P_1(X), P_2(Y))$.

$$JS_{\{\pi,1-\pi\}}(P_1(X), P_2(Y)) = H(\pi P_1(X) + (1-\pi)P_2(Y)) - \pi H(P_1(X)) - (1-\pi)H(P_2(Y)) \tag{2.21}$$

Equation (2.4.6) measures the distance between two distributions that are weighted by π. In our tests, $\pi = 0.2$, because it indicates the sample mixture ratio. Notice that the equation is an approximation to the , and the approximation becomes more precise if we have more sequences. As a result, we are in effect calculating the distance between the two score distributions. We prefer the method with a higher divergence between the genuine and decoy score distributions. As the final score, we have chosen the following scoring schema.

$$S = 1 + \frac{I}{H(\{\pi, 1-\pi\})} \tag{2.22}$$

For large sample size, the above equation is equal to *equivocation* $E_x[H(C|x)]$, as a measure of the information gain, which sets an upper-bound for the *probability of error* [**?**]. The entropy function $H(C|x)$ measures the complexity, or the degree of uncertainty,

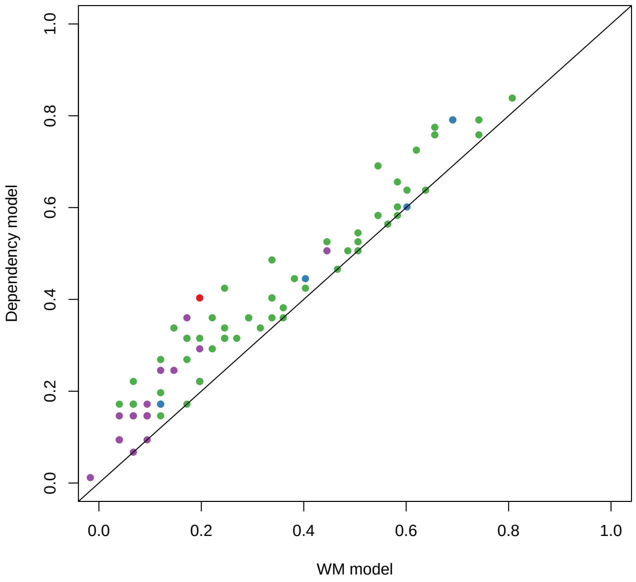


FIGURE 2.6: Comparison of the PSWM and DWT models scores.

in assigning the observation x to the different classes. For our purpose, it can be shown that

$$H(C|x) \;=\; H(\{\pi, 1-\pi\}) - JS_{\{\pi,1-\pi\}}(P_1(x), P_2(x)) \tag{2.23}$$

where $H(\{\pi, 1-\pi\})$ is a measure of the initial uncertainty, for example here we have $H(0.2, 0.8)$, which it says how much the *prior* uncertainty we have in assigning a sequence, without knowing its score, to any of the two classes. If the two distributions $p_1(x)$ and $P_2(x)$ is exactly equivalent the measured distance between them is zero, and therefore the information gained measured by equation (2.23) is equal to the initial uncertainty $H(C)$; meaning that there is no information gained by the scores X. As a conclusion, we are measuring the information gain in a robust way according to the scores we observed. A given model is better than the other model if it has a bigger score S. Figure 2.6 shows the results for 78 TF, and as it is clearly shown the DWT model performs always at least as good as the PSWM model.

2.4.7 Comparison of the DWT and PSWM models over the ChIP-seq data

We have processed published ChIP-seq data from the ENCODE project [28]. For consistency, we have chosen all the TF from a single human cell-line (GM12878) and a few additional TF from other line. We have tested the DWT and PSWM models within the testing framework that is explained in the last section. For each TF in the end we report the measured performances of the two models. Figure 2.7 shows a summary of the test for all 78 TF. Each point represents one TF and the X and Y axis indicate the PSWM and the DWT models scores, respectively. If a point lies on the straight $y = x$ line means that the both models perform exactly the same. When however a point is above the $y = x$ line, it means that for the corresponding TF, the DWT model outperforms the PSWM model. As it is clearly shown in figure 2.7 the DWT model performs always at least as good as the PSWM model. For a number of TF the DWT helped to improve the binding sequences recognition. Within the TF that are examined, we have found 26 TF ($\tilde{3}$0% of the total TF) with the PD that their inclusion improves the model's performance (supplementary text). We have also looked at the different protein structural families and found there is no correlation between the DWT model's performance and the protein families. While this classification may not be indicative for the exhibiting PD, the detailed analysis of the protein structures is beyond the scope of the current study. We observed that many TF exhibiting PD, but only a fraction of them led to an increase in the model's accuracy. For instance, we found two strong PD in GABP binding sites (supplementary text). Including these dependencies, however, did not make any appreciable improvement over the PSWM model. Even though, it is still possible that the PD for GABP can improve the performance on other dataset.

2.4.8 The DWT models explain the HT-SELEX data better than the PSWM models

Systematic evolution of ligands by exponential enrichment (SELEX) is a well-established *in vitro* method for studying protein-DNA binding specificity [65]. This technique relies on mechanisms commonly referred to evolution, such as selection and replication. Starting from a random pool of double-stranded DNA, the sequences are subjected to selection for binding. The selected sequences are then amplified to make the next sequence pool. After a number of cycles the last pool of sequences contains DNA sequences with a very high affinity of binding.

A high-throughput variant of this method (HT-SELEX) is introduced by Jolma *et al.* [26] that the amplified sequences from each round of selection are sequenced using massively parallel sequencing. In a separate study, Jolma *et al.* [25] published a large number of HT-SELEX data for human TF. This dataset has provided a good opportunity to investigate different questions related to the TF binding specificity. In this work, we have used the published HT-SELEX data as an independent dataset to test the DWT and PSWM models that are resulted from the ChIP-seq study. We have used several

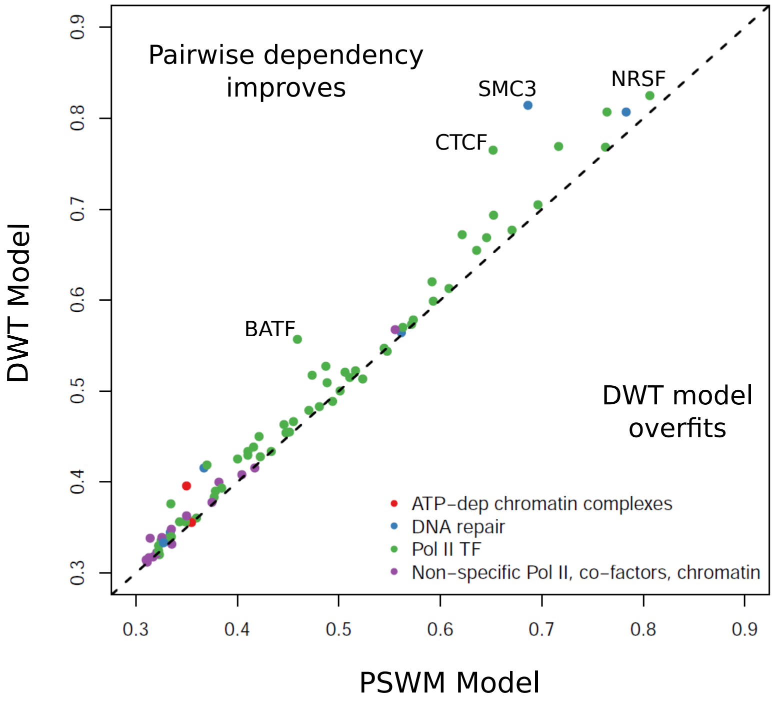


FIGURE 2.7: **The Comparison between DWT and PSWM Models.** The PSWM and the DWT model scores on the x-axis and y-axis, respectively. The units in the x and y axis is the fraction of the explained entropy or uncertainty regarding to the detecting of the real binding sequences from the decoy ones. The dashed $y = x$ line shows the equal performance. The dots are color coded according to a rough protein family assignment. As expected, the PolII TF which are the specific factors show a high level of specificity. This is in contrast with the Non-specific TF and co-factors such as EP300.

criteria to compare the performance of the DWT and PSWM models for those TF that there is a HT-SELEX data available.

First complication for model testing in the HT-SELEX data is that the read lengths are generally larger than the length of the binding site. Instead of taking the score of each window in the sequenced reads, we calculate the binding affinity of reads by the similar method as we used in ChIP-seq data. Hence, the binding affinity $E(S)$ of a read S is calculated by equation 2.11. Using the DWT and PSWM models that are already learned from ChIP-seq data we calculated two binding affinity scores. Therefore, the predicted frequency of a sequence S at selection cycle $(t + 1)$ is defined as

$$\hat{p}_{t+1}(S) = \frac{f_t(S)\exp\{E(S) + E_{ns}\}}{\sum_{S'} f_t(S')\exp\{E(S') + E_{ns}\}} \tag{2.24}$$

where $f_t(S)$ is the observed frequency of the sequence prob S at the round t, and E_{ns} denotes the *non-specific* binding to the sequence S. For simplicity of the analysis here we assume $E_{ns} = 0$.

In the subsequent analysis we have compared the predicted frequencies $p_{t+1}(S)$, for every consequent SELEX selection cycle, from the DWT and PSWM model to the observed frequencies $f_{t+1}(S)$.

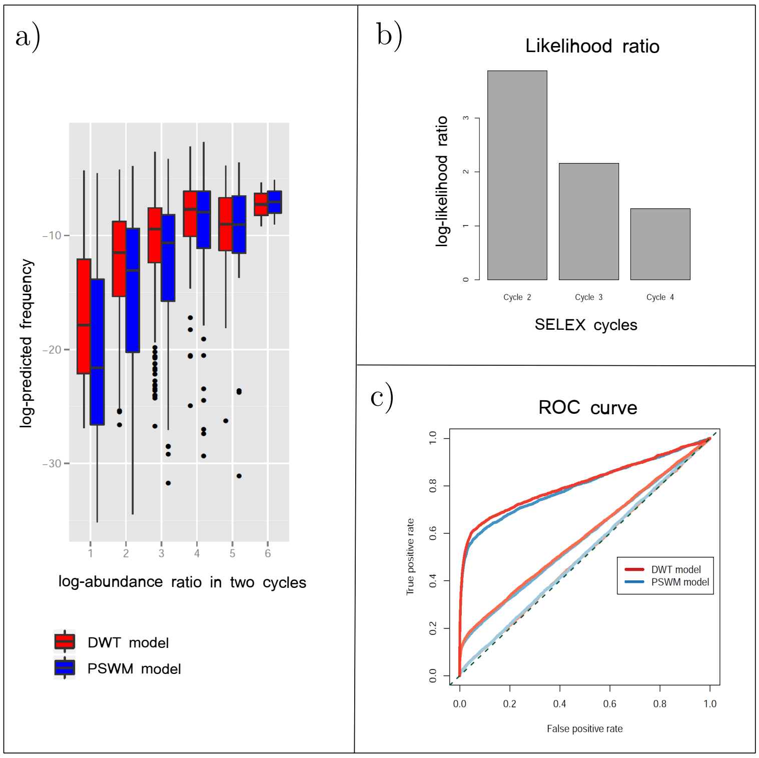


FIGURE 2.8: **The Comparison between the DWT and PSWM Models over the HT-SELEX data.** The PSWM and DWT model predict the abundance of the sequences at the next selection cycle in SELEX, given the frequency of the sequences at the current sequence pool, by using equation 2.24. **a)** The predicted frequency of the top frequent sequences binned by the log transformed abundance ratio is compared. The DWT model's predictions are clearly higher in compare with the PSWM model and matches the observed high abundance sequences. **b)** The likelihood ratio of the two models is calculated for the three adjacent SELEX selection cycle. **c)** In this test we simply asked whether the probability of the selection, or the predicted frequency, provides us with a good information about the fact that a sequence would be selected in the next round or not. The color's darkness denote the SELEX cycle. At the first cycle, where we are calculating the selection probability in a randomized pool of sequences, the both model performs near a random classifier. However, as the more specific sequences are aggregating in the sequence library the performance of the models also improve.

Figure 2.8 represents the results for three separate tests on the HT-SELEX data for human TF MAFK. At first, we have compared the predicted frequency of the sequence to the observed binding affinity of the sequence between any two adjacent SELEX cycle. The observed affinity $\Omega(S)$ of each prob sequence S is basically calculated by $\Omega(S) =$

$\log(\frac{f_{t+1}(S)}{f_t(S)})$. In figure 2.8a we have shown the distribution of the predicted frequencies for the top 1000 high affinity sequences according to the values of $\Omega(S)$. Notably, The DWT model assigns relatively higher prediction scores to these highly abundance sequences. In another test, we have calculated the log-likelihood ratio of the two models. For the log-likelihood ratio bigger than zero means a higher likelihood of the observed data under the DWT model. The log-likelihood of the observed sequence abundances is defined as below.

$$\log(\mathcal{L}) = \sum_S \big(n_{t+1}(S) \times \log p_{t+1}(S)\big) \tag{2.25}$$

where $n_{t+1}(s)$ denotes the number of sequence S observed in the $(t+1)^{th}$ sequence library. It is shown in figure 2.8b that the log-likelihood ratio is always positive, indicating the the dependency model has a higher likelihood for the observed data. This is the case for all of the tested TF.

At last, we have considered a simple scenario; how much the predicted frequency $p_{t+1}(S)$ of the sequences are indicative of the fact that the sequence is being selected in the next round. It can be represented as a bi-classification problem to decide whether or not a sequence is being selected based on its predicted abundance. The three adjacent cycles were considered. As it is shown in figure 2.8c at the cycles 1 (totally random sequences) to 2, the both models perform poorly near to a random classifier. However, as the library become more and more specific with the real binding sequences the predictions become also more accurate. It is also clear here that the DWT model provides a better classifier in compare with the PSWM model. We have calculated the area under the curve (AUC) for all the tested TF and it is always the case that the DWT model outperforms the PSWM model.

2.4.9 Determining pairwise dependencies

One question that naturally arises with respect to the PD is which pairs of positions are dependent. We address this here by calculating the equation 2.8 to determine the posterior probabilities of the PD between any pairs of the binding positions. Figure 2.9 represents the binding specificity of CEBPB factor. Our representation is similar to the weight matrix sequence logo, which is augmented with the positional dependencies information. The graph in the bottom of figure 2.9a shows the dependent positions. The dependency relationship is not a directed, but for the sake of representation we made the edges directed. Therefore, we can read the graph as to say position 5 is dependent on position 4, because there is a directed edge from position 4 to 5. The middle part of the figure 2.9a characterizes the nucleotide frequencies for the dependent positions, conditioned over the parent node. For example, whenever there is A, G or T at position 4, it is highly likely to see G at position 5. Having C at position 4, however, renders not much information about the status of position 5. We have created similar plots for all tested TF, which can be found in the supplementary materials.

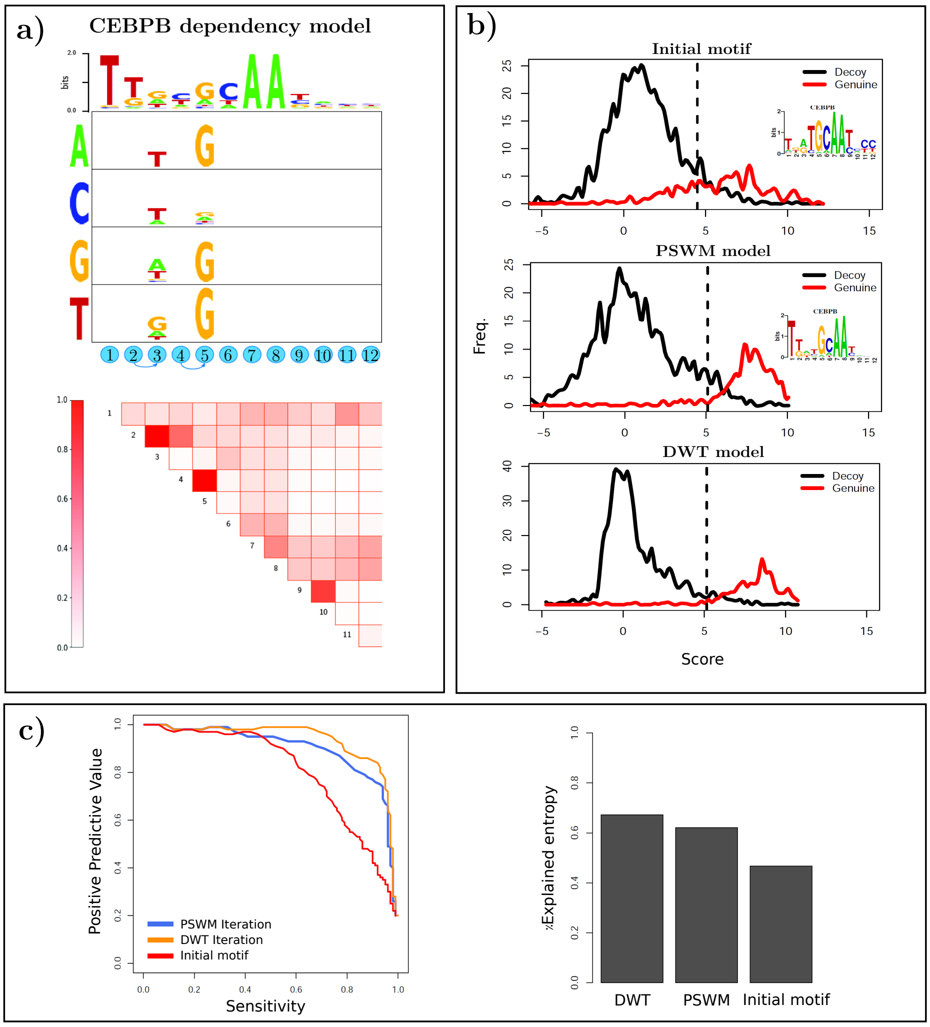


FIGURE 2.9: **PD Analysis for CEBPB.** a) The DiLogo representation for CEBPB. Dependent positions are selected with at least 0.90 posterior of dependency, the posteriors are shown with the heat map. Position 2 and 3 are dependent and as it is shown in this figure whenever there are A and C at position 2, there is a high chance of T at position 3. If there is G at position 2, the chance of having A at 3 is higher than T. This is even completely different if we had T at position 2, which then it makes G very likely at position 3. The interpretation of the PD (4,5) is more straightforward. If there is either of A, G or T at position 4, it is almost surely expected to have G at the fifth position. But if 4 is C, every nucleotide at position 5 is equally likely. b) The score distribution of the genuine (foreground) and decoy (background) sequences for three models: the initial motif, learned PSWM model through iteration, and the DWT model. c) Model comparison, a precision recall curve on the left side and the explained entropy score on the right side based on the score distributions in (b).

2.4.10 Pairwise dependency often occurs between neighboring positions

We have also investigated the distance between positions that we have found to be dependent on each other. We found that the chance of PD decreases as we consider distal positions rather than the adjacent ones. In other words, PD between two immediately adjacent positions is more likely than between the distant positions. Figure 2.10 shows the PD as a function of the distance between positions. As it is clearly demonstrated, the neighboring positions have the highest probability of being in PD. For distal positions, the trend can be divided into different regimes. For positions with the distance two and three there is about the same chance of PD. For more distant positions, between four until seven, there is no clear preference over the distance.

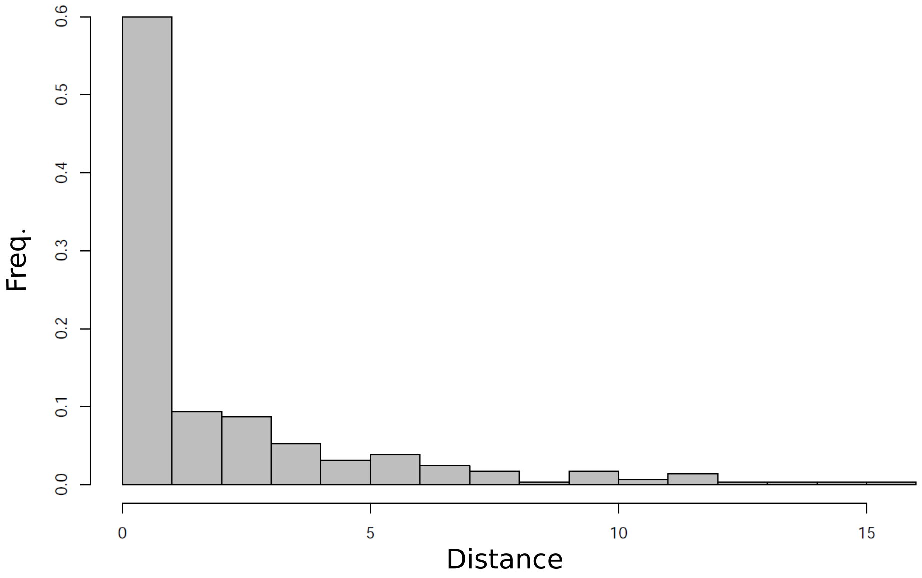


FIGURE 2.10: **PD Occurs Mostly between Neighboring Positions.** The distribution of PD as a function of the distance between the positions. Here we show the dependent positions with at least 0.9 posterior of PD. As it is shown here the adjacent positions are more likely to be dependent on each other than the more distal ones.

2.4.11 The case of GABP

One of the TF that shows a strong dependencies is GA-binding protein (GABP) that exhibits dependencies between positions 9 and 10 and another one between positions 2 and 3 (??). These dependencies, even though are strongly supported by the binding data, their usage did not lead to an appreciable improvement over the PSWM model. This might be due to the fact that these dependencies may improve the predictions in other datasets. This also means that not all the dependencies that are supported by the data are going to necessarily improved the predictions.

Summary

The PD within the TF binding sites is addressed. The presence of the PD and the effect they have on the TF binding specificity have been the matter of discussions. From a computational perspective, modeling the binding specificity with the PD is computationally intractable. For that reason the current methods simplify the problem by taking into account only a small subset of the total PD. We proposed here a Bayesian model, the DWT model, that rigorously brings a comprehensive view to the PD within the TF binding sites. The DWT model, taking advantage of a generalization of the matrix-tree theorem, explores the entire parameter space of PD in a computationally feasible manner. Unlike its peers the DWT model is without any tunable parameter and relies upon no extra regularization techniques. We show that the DWT model is in fact a generalization to the PSWM model, which it automatically reduces to the PSWM model whenever the support of PD is rare. In order to understand if the PD are helpful in a context of TF binding sites prediction we have compared the two approaches: PSWM and DWT models. A comprehensive testing framework is developed which removes the possibility for systematic biases toward any of the two approaches. We have tested 78 human TF ChIP-seq data and found for about a third of them that the DWT model outperforms the PSWM approach. For those TF without the PD, as it is predicted, the DWT model performs the same as the PSWM model. The improvements that we have been gained by the DWT model, although modest, proves the role of PD in the binding specificity. In addition, we propose a new graphical representation for the TF binding specificity that combines the classical logo with the significant PD. The results for all the 78 tested TF is accessible via [HERE]. We have also developed a website that by uploading the binding sequences (in FASTA format) as well as an initial weight matrix, performs the PD analysis. The source code of DWT model (in C++) and other useful scripts (often in Python) is available by request.

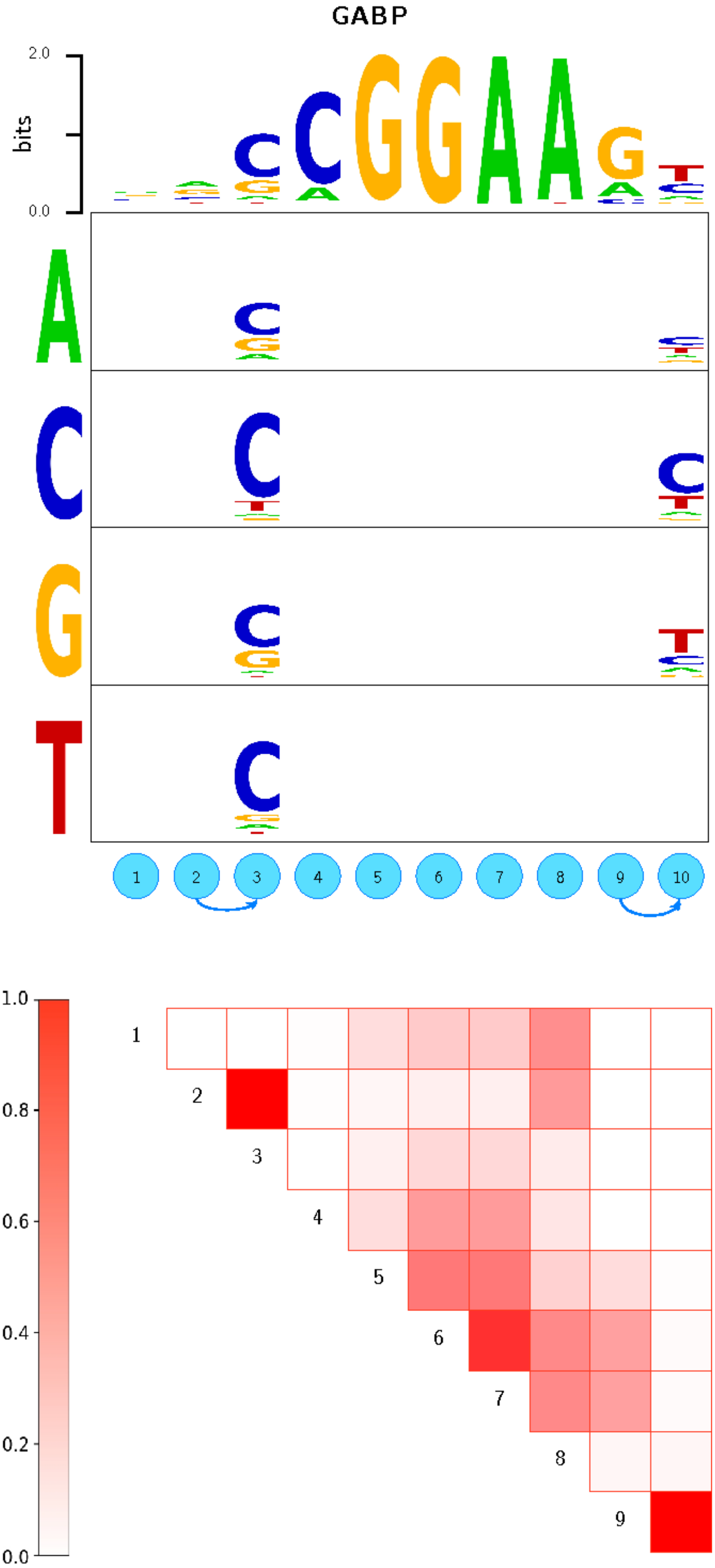


FIGURE 2.11: GABP diLogo: As it is shown there are dependencies between positions 9-10 and 2-3.

Chapter 3

Feature-coupling based classification

3.1 Introduction

Often in science and engineering we are interested to partition a group of entities into a number of categories. In biology, for example, species are categorized into different groups: eukaryotes for any organism whose cells contain a nucleus like man and yeast, prokaryotes for organisms without nucleus like Bacteria and Archaea. At first sight, it may sound queer having human and yeast grouped together, but this classification is merely based on common cellular organelles and structure. Other classifications, however, may distinguish between man and yeast using other features. For instance, human is classified as a member of mammalian class along with mouse, cow and other mammalians. In software engineering we might be interested to label an incoming email as spam or normal message. Therefore, researchers search for the best set of features that with a high accuracy indicates the nature of the message.

Formally, the problem of classification is stated in terms of the probability $p(C_i|\{f\}, \boldsymbol{\theta})$, for $i = 1, 2, 3...N$. The $\{f\}$ is the observed data consisting of the desired features, note that the set notation is because that the order is not relevant. The symbol $\boldsymbol{\theta}$ stands for model parameters. Using Bayes' theorem, we can write

$$p(c_i|\{f\}, \boldsymbol{\theta}) = p(c_i)\frac{p(\{f\}|c_i, \boldsymbol{\theta})}{p(p(\{f\}|\boldsymbol{\theta})} \tag{3.1}$$

where $p(c_i)$ is the *prior* probability to classify as c_i, before seeing the input data $\{f\}$. The function $p(\{f\}|c_i, \boldsymbol{\theta})$ is referred to as the *likelihood* function. Note that the likelihood is not a probability distribution, it quantifies the likelihood of the parameters given the data. In the denominator $p(p(\{f\}|\boldsymbol{\theta})$ is for normalization purpose and is usually called *evidence* of the data, where the class label c_i is marginalized.

Since we are interested in the relative probability of any of the N classes, we can for now just calculate

$$p(c_i|\{f\}, \boldsymbol{\theta}) \propto p(c_i)p(\{f\}|c_i, \boldsymbol{\theta}) \quad \text{for } i = 1, 2, ..., N \tag{3.2}$$

In this chapter different ways of treating the equation (3.2) is discussed.

3.2 Naive Bayes classifier

Calculating the likelihood function in (3.2), in a general sense, can be tractable. Features can be dependent with each others in a pairwise manner or even higher orders of dependency. Therefore, before proceeding with calculating the likelihood function (3.2) we might need to make some assumptions. These assumptions can be very strong, for example by assuming that the features are completely independent from each other. In some situations, this may be a good approximation and sometimes it turns out to be a very naive assumption which leads to a poor classifier. Hence, this method is ironically named as Naive Bayes method [1]. Naive Bayes classifier has been used in different domains, such as automatic text categorization and disease diagnostic systems.

Despite the strong independent assumption in this model, because of its simplicity and the small number of parameters it has been used in many classification problems. However, one of the pathologies of this method is that its underlying non-dependency assumption sometimes lead to *overcounting*. For example, two variables hypertension and obesity are independently strong indicators of heart disease, but they are also highly correlated. Therefore, the Naive Bayes approach leads to overcounting the importance of these two variables [66].

Naive Bayes method is based on a strong independent assumption, that is

$$p(\{f\}|c_i, \boldsymbol{\theta}) = \prod_{n=1}^{M} p(f_n|c_i, \boldsymbol{\theta}) \tag{3.3}$$

where M is the cardinality of the feature set $\{f\}$. In equation (3.3) the likelihood function is equal to the product of each of the features separately.

[1] In some literature it is referred to as *Idiot* Bayes!

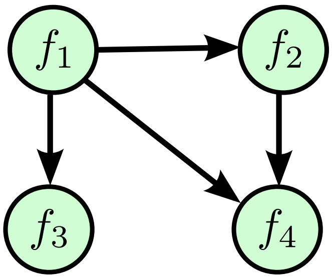


FIGURE 3.1: Example of a DAG that represent a Bayesian network consisting of four features.

One of the application of the Naive Bayes classifier is in transcription factors binding site prediction. In this problem, every feature f_i is the appearance of a nucleotide at i^{th} position in the binding site. If it appeared that each binding position contributes to the total binding affinity of the sequence independent from the other positions, then using Naive Bayes classifier is acceptable. This method in the context of binding site prediction is commonly referred to as the position specific weight matrix (PSWM) model. However, several lines of evidence have revealed that the positional non-interdependence might not be a good approximation of the binding affinity. This will be discussed in more detailed in the upcoming chapter.

3.3 Methods for incorporating feature dependency

In order to improve the performance of Naive Bayes method, several algorithms have been suggested that consider the dependency between features. Solving this problem in general is computationally arduous. But in there are several methods for taking the dependent features into account that are usually named as graphical models.

3.3.1 Bayesian networks

This section briefly discusses Bayesian networks (or also referred to as belief network) as a model for incorporating feature dependencies. Bayesian networks are widespread in different fields that are trying to build up more realistic models by allowing dependencies between input random variables. It is common to depict the Bayesian networks in terms of directed acyclic graphs (DAG), where each node represents a random variable (technically it is better to referred to nodes as features instead of random variables, because the definition of random variables in Bayesian context is rather obscure. It is more a term that is borrowed from frequentist statistics). And each directed edge stands for a dependency from one feature to another feature.

Figure 3.1 shows an example of a Bayesian network with four features f_1, f_2, f_3 and f_4. As it is represented in figure 3.1 pair of features f_1 and f_2 are connected by a directed edge from f_1 to f_2; this is equivalent of the conditional probability $p(f_2|f_1)$. In mathematical form, this example Bayesian network is written as

$$p(f_1, f_2, f_3, f_4|I) = p(f_1|I)p(f_2|f_1, I)p(f_3|f_1, I)p(f_4|f_1, f_2, I) \tag{3.4}$$

where notation I indicates any prior information regarding to the model, such as the topology of the Bayesian network.

In equation (3.4), feature f_4 is dependent on two features simultaneously. This shows that the dependency structure in Bayesian network is not only of first order, but higher-orders of dependency can be encapsulated in the Bayesian network.

It is easy to see that the Naive Bayes model is in fact a specialized form of the general Bayesian networks. By definition, when pairs of variables are conditionally independent from each other, the resulted model is Naive Bayes model. Naive Bayes is just the most compact representation of the Bayesian networks.

We have assumed above that the topology (or structure) of the Bayesian network is previously known. This however may not be the case in some situation. Moreover, even when the structure of the network is yielded, we might think of the parameters in the model; the strength of connections. Below we discuss some of these issues.

3.3.2 Model learning in Bayesian networks

Learning algorithms for learning the parameters of Bayesian networks can be divided into two categories [67]. First, several algorithms try to infer the conditional dependency from the data empirically. By different test methods for the interdependency between the variables of data, these algorithms find the best Bayesian network to explain the observed data. An example of this algorithms is Bayesian Network Power Constructor (BNPC) which uses independent tests and mutual information scores for characterizing the variable pairwise dependency [68].

Another class of algorithms are based on a scoring function, or *metric*. The scoring functions are based on different principles, such as maximum entropy scores, or minimum description length. For instance, in an algorithm that is introduced by Heckerman *et al.* [69] it searches the space of DAG for the best graph to explain the data. This algorithm is based on local search (LS) method, that by starting from an initial DAG, by local changes in the structure of the DAG improves the model's fit.

3.4 Ensemble tree model

So far we considered models that by incorporating variable dependency avoid overcounting problem that may arise using Naive Bayes method. These models are based on only one structure, where the dependency relationship between variables are fixed. There are several learning algorithms for the network structure as well as the model's parameters. In this section, we are presenting a novel model that does not a priori assume that there is one structure for the variable dependency. This model, that is called the ensemble tree (ET) model, marginalize the tree structure. Therefore, there is no need for learning the best dependency network topology.

The number of all possible (spanning) trees for n nodes is equal to $n^{(n-2)}$. Therefore, it is computationally very expensive, in fact, impossible to calculate the sum of all trees of any arbitrary size n. This problem formally is defined by

$$p(\{f\}|\boldsymbol{\theta}) = \sum_{\pi} \Big[p(\pi) \prod_{i=1}^{M} p(f_i|f_{\pi(i)}, \boldsymbol{\theta}) \Big] \tag{3.5}$$

where π denotes the tree structure that encodes the dependency relationships between pair of variables. As it is discussed above, the number of times that the sum in (3.5) should be performed scaled with M^{M-2}. For an arbitrary large feature set $\{f\}$, computing the equation (3.5) is tractable. Here we suggest another approach for computing the above equation that only requires polynomial calculation time. For that, first we need to rewrite the equation (3.5) in the following form:

$$p(\{f\}|\boldsymbol{\theta}) = \sum_{\pi} \Big[p(\pi) \prod_{i=1}^{M} \Big(\frac{p(f_i, f_{\pi(i)}|\boldsymbol{\theta})}{p(f_{\pi(i)}|\boldsymbol{\theta})} \times \frac{p(f_i|\boldsymbol{\theta})}{p(f_i|\boldsymbol{\theta})} \Big) \Big] \tag{3.6}$$

In equation (3.6) by using the product rule in probability the conditional probability $p(f_i|f_{\pi(i)}, \boldsymbol{\theta})$ is replaced by $p(f_i, f_{\pi(i)}|\boldsymbol{\theta})/p(f_{\pi(i)}|\boldsymbol{\theta})$.

Defining a symmetrical square matrix $\mathbf{R}$ that whose elements are

$$R_{ij} = R_{j_i} = \frac{p(f_i, f_j|\boldsymbol{\theta})}{p(f_i|\boldsymbol{\theta})p(f_j|\boldsymbol{\theta})} \tag{3.7}$$

we can rewrite equation (3.6) as,

$$p(\{f\}|\boldsymbol{\theta}) = \sum_{\pi} \Big[p(\pi) \prod_{i=1}^{M} \big(R_{i\pi(i)}.p(f_i|\boldsymbol{\theta}) \big) \Big] \tag{3.8}$$

This alternative matrix form of the equation allows us to exploit a well-known combinatorial theorem, known as matrix-tree theorem[2]. According to the matrix-tree theorem sum of all nonidentical spanning trees for a connected graph that is represented by adjacency matrix **G** is equal to any cofactor of the Laplacian form of matrix **G** [70] [56]. The Laplacian of a matrix **G** is defined as the difference of the degree matrix **D** and the adjacency matrix **G**.

Therefore, using a uniform distribution over the spanning trees for prior probability $p(\pi)$, and applying the matrix-tree theorem, we have:

$$p(\{f\}|\boldsymbol{\theta}) = \frac{1}{M^{(M-2)}} \times \det(\tilde{\mathbf{R}}) \prod_{i=1}^{M} p(f_i|\boldsymbol{\theta}) \tag{3.9}$$

In the above equation $\tilde{\mathbf{R}}$ is obtained by transforming the original **R** matrix into the Laplacian matrix and then removing one arbitrary row and column from the matrix; this is often called the *minor* of the graph. Equation (3.9) mathematically is equivalent to the first equation (3.5), but it has a significant implication from a computational perspective. Calculating the determinant of matrix $\tilde{\mathbf{R}}$ is central in (3.9), which its computational complexity is polynomial. In fact, calculating the determinant of a matrix of size n by standard numerical methods, such as the Cholesky decomposition or the QR decomposition, takes $O(n^3)$ computational time [71]. Hence, using the matrix-tree theorem we could replace the original computational intractable problem with conventional numerical algorithms for calculating matrix determinant.

3.5 Likelihood of single features and feature pairs

So far we were concerning with the high level structure of the model, such as interdependency between pairs of features or non-dependency like in Naive Bayes models. However, we have not talked about the basic probability functions: $p(f_i|\theta)$ and $p(f_i, f_j|\theta)$.

Here we first assume that features are discrete with a finite feature space, meaning that they accept countable and finite discrete values. For instance, binary features like the pass or fail result for a course. Alternatively, it can be larger feature set like the IQ test result that takes integer values from a minimum IQ to the maximum possible score. In biology, example is the status of a position on DNA that can have either of four possible nucleotides: A, C, G, or T.

[2] In some text, it is alternatively referred to as Kirchhoff's theorem.

Let x be a discrete variable taking on values $1, 2, ..., S$; given the probabilities of taking any of the possible values as $\mathbf{p}$ the likelihood of a dataset d consisting of N observation is defined as:

$$p(d|\mathbf{p}) = \prod_{k=1}^{S} p_k^{N_k} \tag{3.10}$$

where N_k is the number of time that value k is observed in d and by definition $N = N_1 + N_2 + ... + N_S$. The evidence of the data d is obtained by application of product rule as below.

$$p(d) = \int_{\mathbf{p}} \mathrm{d}\mathbf{p}.p(d|\mathbf{p})p(\mathbf{p}) \tag{3.11}$$

In the equation (3.11) we are needed to decide on the prior probability over the probabilities $\mathbf{p}$. For that, it is common to use Dirichlet distribution. The Dirichlet distribution is defined as:

$$p(\mathbf{p}|\lambda) = \frac{1}{Z(\lambda)} \prod_{k} p_k^{\lambda_k - 1} \tag{3.12}$$

where λ_k is the hyperparameter of the Dirichlet distribution. The values of λ_k can be treated as pseudo-counts, or virtual counts, that states a priori how much we expect to see the count associated to the k-th state [72]. It can be shown that the Beta distribution is a special case of the Dirichlet distribution when $S = 2$.

Note that the normalization constant $Z(\lambda)$ is obtained by solving the following integral.

$$\int_{\mathbf{p}} \prod_{k} p_k^{\lambda_k - 1} = 1 \tag{3.13}$$

Solving the above integral yields:

$$Z(\lambda) = \frac{\prod_k \Gamma(\lambda_k)}{\Gamma(\sum_k \lambda_k)} \tag{3.14}$$

Hence, the evidence of the observed sample X is obtained by plugging equation (3.10) and (3.12) in equation (3.11).

$$p(X|\lambda) = \int_{\mathbf{p}} \mathrm{d}\mathbf{p} \prod_{k} p_k^{N_k + \lambda_k - 1} \tag{3.15}$$

The above integral is analogous to the integral in equation (3.13) that results in (3.14); as a result, the evidence of the data is also Dirichlet distribution, where the hyperparameters are $N_k + \lambda_k$. For this reason, in some literature the usage of Dirichlet distributions for prior is referred to as *conjugate* priors.

$$p(X|\lambda) = \frac{\Gamma(\sum_k \lambda_k)}{\Gamma(N + \sum_k \lambda_k)} \prod_k \frac{\Gamma(N_k + \lambda_k)}{\Gamma(\lambda_k)} \tag{3.16}$$

which is the probability that the sample data X comes from one multinomial distribution.

The choices for the pseudo-count values λ_k is dependent on the data and the system that we are modeling. But there are also some standard values for λ values. For instance, having for any $\lambda = 1/2$ is known as the Jeffreys' prior, or $\lambda = 1$ is equivalent to the uniform prior [73].

According to the Dirichlet distribution, the posterior predictive distribution is equal to

$$p(x = k|X, \lambda) = \frac{N_k + \lambda_k}{N + \sum_k \lambda_k} \tag{3.17}$$

Interestingly, if we use a uniform prior $\lambda = 1$, the above posterior is equal to $\frac{N_k+1}{N_k+K}$. For the case $K = 2$, this posterior is equivalent to the Laplace's succession rule. In machine learning community the usage of Dirichlet conjugate priors is sometimes referred to as Laplace smoothing, or additive smoothing.

The connection of the evidence of data as in equation (3.16) to the concept of entropy, when a uniform prior $\lambda_k = 1$ is employed, is given by the following equation.

$$\log p(X|\lambda_k = 1) \approx -N \log N + N + \sum_k \big(N_k \log N_k - N_k\big) \tag{3.18}$$

$$= \sum_k N_k \log \frac{N_k}{N} \tag{3.19}$$

$$= -N\, H(N_k/N) \tag{3.20}$$

This result shows that less entropic samples are more probable [73]. The approximation in (3.18) is by the usage of Stirling's approximation [3]. In equation (3.20) function

[3] Stirling's approximation: $\log \Gamma(n + 1) = n \log n - n$ for any positive integer n

$H(p_k)$ is the entropy function that is defined by $H(p_k) = -\sum_k p_k \log p_k$. Note that the base of logarithm, when we are using the natural logarithm the uncertainty measure of the entropy is in *nats* units; whereas when we using logarithm base 2, the measured information is in *bits*.

In the same way that we have calculated the evidence $p(X|\lambda)$ for a single observed counts, we can define $p(X, Y|\lambda)$ for a pair of data counts. The joint distribution $p(X, Y|\lambda)$ using the same conjugate Dirichlet distribution with the hyperparameters $\lambda_{k,l}$ is obtained by solving the following integral.

$$p(X, Y|\lambda) = \int_{\mathbf{p}} \mathrm{d}\mathbf{p} \prod_{k,l} p_{k,l}^{N_{k,l}+\lambda_{k,l}-1} \tag{3.21}$$

Note that the resulted distribution is again a Dirichlet distribution.

$$p(X, Y|\lambda) = \frac{\Gamma(\sum_{k,l} \lambda_{k,l})}{\Gamma(N + \sum_{k,l} \lambda_{k,l})} \prod_{k,l} \frac{\Gamma(N_{k,l} + \lambda_{k,l})}{\Gamma(\lambda_{k,l})} \tag{3.22}$$

The value of the hyperparameters in equation (3.22) is obtained by the following relations.

$$\lambda_{k,.} = \sum_{l=1}^{R} \lambda_{k,l} \tag{3.23}$$

$$\lambda_{.,l} = \sum_{k=1}^{Q} \lambda_{k,l} \tag{3.24}$$

where R and Q are the size of feature space for X and Y, respectively. The reason for the form of hyperparameters $\lambda_{k,l}$ is because of the additive property of the Dirichlet distribution.

By replacing equations (3.22) and (3.16) into the equation (3.7) the values of the dependency matrix $\mathbf{R}$ can be calculated.

$$R_{ij} = \frac{\Gamma(N + \sum_{k,l} \lambda_{k,l})}{\Gamma(\sum_{k,l} \lambda_{k,l})} \prod_{k,l} \frac{\Gamma(N_{k,l} + \lambda_{k,l})}{\Gamma(\lambda_{k,l})} \prod_{k=1}^{R} \frac{\Gamma(\lambda_{k,.})}{\Gamma(N_{k.} + \lambda_{k,.})} \prod_{l=1}^{Q} \frac{\Gamma(\lambda_{.,l})}{\Gamma(N_{.l} + \lambda_{.,l})} \tag{3.25}$$

We can use the the entropy approximation in equation (3.20) for the logarithm of R_{ij} values [73].

$$\log R_{ij} \approx N\big[H(N_{k,.}/N) + H(N_{.,l}/N) - H(N_{k,l}/N)\big] \tag{3.26}$$

$$= -N \times D_{KL}(\frac{N_{k,.}}{N} \times \frac{N_{.,l}}{N}||\frac{N_{k,l}}{N}) \tag{3.27}$$

$$= -N \times I(\frac{N_{k,.}}{N} \times \frac{N_{.,l}}{N}; \frac{N_{k,l}}{N}) \tag{3.28}$$

where the function $D_{KL}(x||y)$ is the Kullback-Leibler divergence function (or relative entropy), and $I(x;y)$ is the mutual information between two variables [74]. Note that the Kullback-Leibler divergence is not a distance measure; meaning that $D_{KL}(x||y)$ is not essentially equal to $D_{KL}(y||x)$.

According the the Gibbs inequality [72], the Kullback-Leibler divergence of two variables is always at least zero. In other words, $D_{KL}(x||y) \leq 0$, with the equality when x and y are identical distribution. Hence, this measure of dependency (correlation) between two variables is biased toward the hypothesis that the two variables are dependent [73]. It can be shown that the value of R_{ij} is resulted from the following hypothesis testing.

$$R_{ij} = \frac{p(D|dep)}{p(D|indep)} \tag{3.29}$$

where $p(D|dep)$ is the probability that the data is drawn from a joint distribution, and $p(D|indep)$ is the opposite, where the data is the result of two separate multinomial distributions.

Chapter 4

Modeling gene expression regulation with enhancers

4.1 Introduction

The proper functioning of any living organism relies on the flawless gene expression regulation. The precise regulation of genes in space and time is highly crucial for any biological process, including proliferation, development and differentiation in the multicellular organisms. In unicellular organisms it is critical to express proper genes in a fluctuating environment in order to cope with various stress factors, such as the lack of vital nutrients or the presence of lethal substances. A myriad number of interactions between proteins, such as transcription factors (TF) and chromatin re-modelers, to DNA sequences constitute the core of gene expression regulation. I think I can safely say that this mechanism is yet too far from our total grasp.

Today we know that the gene expression regulation consists of a number of layers. At first, it must be decided what part of the genome to be transcribed, or in other words to control the production of RNA. This work, however, mainly concerns with the transcription regulation. At the next layer, the transcribed RNA are processed, such as splicing. Different splicing might lead to completely different function for the final protein product. It is also shown, in the past decade, that a regulatory mechanism is imposed on the RNA level. A class of RNA, generally referred to as microRNA (miRNA), can block the translation of RNA to protein.

In this chapter, a new computational model for the transcription regulation is discussed. It is widely known that TF through binding to promoters, DNA regions that are proximal to the transcription start site, control the transcription of genes. In an abstract term, the RNA production of a gene is a function of its promoter sequence as well as the presence or absence of regulatory proteins. Binding of different TF to the promoter can activate or deactivate the transcription from the gene. We have discussed in the last chapters that it is in fact very complicated to calculate the binding affinity of TF to the DNA sequences. What makes it more difficult is the fact that a combination of TF can behave differently under different conditions. This is referred to as combinatorial binding, or combinatorial regulation.

In the last decades, another class of regulatory sequences are recognized that act somehow mysteriously. It was recognized that a DNA sequence, although sitting far away from the transcription start site, plays an important role in the transcription regulation of the gene [75]. These distal elements are generally, and somehow meaninglessly, referred to as *enhancers*. Same as promoters, enhancers contain a number of binding sites for some TF, and TF through binding to these sequences control the transcription of a gene at a far distance. This somehow bizarre observation becomes maybe easier to understand if we imagine the genome as a long polymer chain that is crumbled inside a small nuclei compartment. Consequently, different parts of the polymer get closer together in a three-dimensional space. Put in other terms, the DNA bend to get closer two linearly distal regions. Recently another mechanism for the function of enhancers has been suggested. It is shown that the enhancers are also transcribed into a non-coding RNA, referred to as eRNA (for enhancer RNA) [76], and the resulted RNA has a regulatory effect on far reaching genomic loci.

It is conjectured that the activity of enhancers is highly cell-type specific [77] [78]. Enhancers originally defined as regulatory elements that act at distance, and independent from orientation, can mediate the gene expression of a remote genomic loci [75].

4.2 Current methods for identifying enhancers

In the last few years, a number of methods has been proposed to identify potential enhancer elements. This section provides an overview on these methods. As it is discussed in the last section, enhancers posses binding sites for TF and via binding to enhancers, TF mediate the gene expression regulation at the distant genes. Remarkable, more than 90% of binding events occur outside proximal promoters, at regions that resembling enhancers [79] [80]. The presence of TF at active enhancers is associated to the low nucleosome occupancy [81]. Put in other terms, the active enhancers are accessible to the TF owing to the lack of nucleosome. It has been demonstrated that DNAse I hypersensitive sites (DHS) are very useful for detecting nucleosome free regions [82].

For instance, the results of ENCODE project shows that more than 97% of the non-promoter elements, that includes enhancers and other regulatory units, to be located at nucleosome free regions [82]. Even though useful to detect active regions of the genome, it is not completely known that if all the open regions are fully functions. Furthermore, enhancers constitute a subclass of the regulatory regions which include silencers and insulators; therefore not all the open elements are associated to the enhancer function.

Several studies have recognized the role of co-factor EP300 (or sometimes referred to as P300) in gene expression regulation, specifically in marking enhancer regions [83] [84] [85] [86]. In a large-scale study, Visel *et al.* [87] explored the role of enhancers in different embryo tissues of mouse by mapping the regions that are enriched by EP300 binding. Along with the binding of EP300, several studies have shown a correlation between enhancers and some chromatin marks, namely the presence of acetylation of histone H3 at lysine 27 (H3K27ac) and H3 lysine 4 monomethylation (H3K4me1) and the absence of promoter-associated mark H3 lysine 27 trimethylation (H3K27me3) and H3 lysine 4 trimethylation (H3K4me3) [88]. This study has found 5,118 genomic regions with chromatin features resembling enhancers. In addition, it is shown that the active enhancers lack H3K27me3, a modification associated with polycomb silencing.

It is just recently that it is discovered that enhancers are also transcribed. The resulted transcribed often referred to as enhancer RNA (eRNA) [76]. Understanding the possible role of eRNA in the activity of enhancers has become an active topic of research [89] [90]. In a very recent study by FANTOM consortium [91], 43,011 potential enhancers, defined as the to be expressed and directional-independent, for many human cell lines, encompassing 432 primary cell, 135 tissue and 241 cell line samples, have been localized. Remarkably, this study has revealed that many disease-associated SNP occur more often at regulatory regions, specially those that are related to enhancers, rather than exonic regions.

In summary, there are several approaches for localizing enhancers, mainly active enhancers, within the genome. These approaches ranging from chromatin marks, comparative genomic and accessibility of the genomic elements to the expressed transcripts from enhancers. Within the last years, using variety of techniques many collections of enhancers have been published; but it is still not quite clear that which enhancer regulates which gene, or group of genes. One experimental technique that tries to answer this is the family of chromosome conformation capture (3C) methods [92]. In particular a variant of 3C method, namely chromosome conformation capture carbon copy (5C), has been of interest in the ENCODE project [93]. Using this technique, they were able to query a small subset of human region, Beta-glubin, to detect possible long-range interactions between the genomic elements. As it was expected, it is shown that the previously recognized enhancers are interacting with the promoter regions. One issue with this technique, apart from the difficulty in interpreting data, is that it is applicable at a low scale for only a small number of regions. A more recent high-throughput variant of this technique introduced by Lieberman-Aiden *et al.* [94] which is called Hi-C method. The authors were able to apply this technique on a genome-wide scale and showed the interaction of distant genomic regions follow a scale free distribution. This technique has been used to investigate 1 Mega bp human genome, which is still a very low resolution for understanding the detailed mechanisms of gene expression regulation by TF.

In this chapter, we are presenting a novel Bayesian approach that characterize enhancer elements. Most importantly, this method provides a computational framework for finding the best promoter-enhancer linkage. We can show that a model with enhancers along with promoters explain the observed gene expression dynamics better than a model that consists only of promoters. Finally, the proposed framework allows incorporating new evidences or the available data for a more accurate enhancer and enhancer-promoter recognition.

4.3 MARA: Transcription regulation as a function of promoter sequence

In a probabilistic model Balwierz et al. [95] characterized the gene expression dynamic across different samples, or time, as the function of predicted binding sites within genes' promoters. This model, called motif activity respond analysis (MARA), infers the transcription regulatory circuit that is behind the observed gene expressions. Formally, the observed expression E_{ps} at a promoter p, in sample s, varies linearly with the number of binding sites N_{pm}, for all possible (available) motif m.

$$E_{ps} = \sum_{m} N_{pm} A_{ms} + \text{noise} \tag{4.1}$$

The values of N_{pm} for every motif on each promoter is calculated by summing the posterior probabilities of binding for each site with promoters. For calculating the site counts N_{pm}, conservation of each sequence as well as its matching to the motif is considered by MotEvo program [64]. What is needed to be inferred from the data are the values of A_{ms}, which denotes *activity* of motif m in sample s. The motif activity determines the contribution of the motif in a particular sample to the observed gene expression profile. Note that the motif m has no activity per se, it is the binding of TF that defines the motif activity. In the MARA model the noise is assumed to be originated due to the model's error; therefore, the noise is modeled as Gaussian distribution with an unknown variance σ^2, which is integrated out from the likelihood function.

One more point is the that the MARA model is in fact a so-called *Ridge regression* model, meaning that it uses L2 regularizing term, also known as Tikhonov regularization, to control the complexity of the final model. From the Bayesian point of view, it is similar to have a Gaussian prior over the activity parameters.

The posterior of motif activities has a form of multi-variate Gaussian distribution

$$p(A|E,N,\sigma) \propto \sigma^{-PS} \exp\left[-\frac{\sum_{p,s}\left((E_{p,s} - sum_m N_{pm} A_{ms})^2 + \lambda^2 \sum_m A_{ms}^2)\right)}{2\sigma^2}\right] \tag{4.2}$$

where P and S are the total number of promoters and samples, respectively; and λ sets the width of prior distribution over A_{ms} relative to the width of the likelihood function. The value of λ is determined through a 80/20 cross-validation scheme. In the next section, we develop a new model that includes potential enhancers to the basic MARA model.

4.4 The probabilistic Model

In this section, we present a statistical approach to characterize the role of distal regulatory elements, also referred to as enhancers, in the gene expression regulation. Instead of detecting the distal elements using different techniques such as epigenetic signals that mark the enhancers or chromatin conformation capture method, here we are concerned with the *functional* activity of enhancers. The enhancers, similar to promoters, harbor a set of binding sites for different transcription factors (TF) that upon binding will potentially activate the regulatory effect of the enhancer. This is physically done by loop formations in the DNA that brings close the proximal promoters and the enhancers (figure 4.1). The detailed mechanism for loop formation is still not fully understood, but it is suggested that the loop stabilization might be mediated by TF and co-factors that are bound, directly or indirectly, to the two regions: promoters and the enhancers. Therefore, the arrangement of TF in both proximal and distal sequences would convey information about their distal interaction and consequently their contribution toward the gene expression regulation.

Under a simple linearity assumption, the observed gene expression of promoters is modeled as it is explained in section 4.3 and equation 4.1. This idea is illustrated in figure

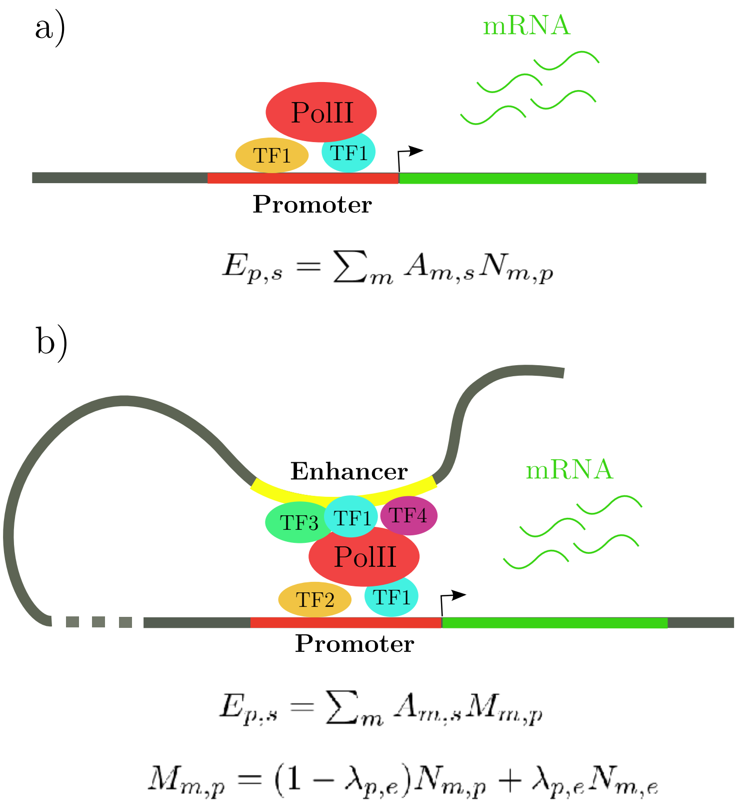

a)
mRNA
PolII
TF1
TF1
Promoter
$E_{p,s} = \sum_m A_{m,s} N_{m,p}$
b)
Enhancer
mRNA
TF3
TF1
TF4
PolII
TF2
TF1
Promoter
$E_{p,s} = \sum_m A_{m,s} M_{m,p}$
$M_{m,p} = (1 - \lambda_{p,e}) N_{m,p} + \lambda_{p,e} N_{m,e}$

FIGURE 4.1: Models of transcription regulation. a) when there is only the proximal promoter plays the regulatory role and b) when the contact between distal enhancer and the promoters control the expression of the gene.

4.1a, where the proximal regulatory unit is responsible in deriving the expression of the gene next to it. Another mode of regulation can be done via interaction of the proximal promoter to the remote regulatory element, facilitated via DNA loop formation. To include the regulatory activity of the enhancer into the above model, as it is shown in figure 4.1b, we model the binding sites composition as the mixture of binding sites from the promoter and the enhancer.

$$E_{p,s} = \sum_m A_{m,s}((1 - \lambda_{p,e}) N_{m,p} + \lambda_{p,e} N_{m,e}) \tag{4.3}$$

The mixture of TF binding sites is determined by the quantity $\lambda_{p,e}$ that can reflect the rate of physical contact between promoter p and the distal enhancer e. Note that any other functional form can be assumed for the promoter-enhancer interaction. Here, we only consider the weighted sum of the motifs in both enhancer and promoter. The site-count $N_{m,e}$ is calculated by the same way as $N_{m,p}$. By assuming a Gaussian noise model and independence between different samples, we can write the likelihood model.

$$\begin{aligned} P(E_p|N,A,\lambda,\sigma) &= \Big(\frac{1}{\sqrt{2\pi\sigma}}\Big)^n \exp\Big(- \frac{\sum_s \big(\epsilon_{p,s} - \lambda \widetilde{E}_{p,s}\big)^2}{2\sigma^2}\Big) \\ \epsilon_{p,s} = E_{p,s} - \sum_m A_{m,s}N_{m,p} \ &, \ \widetilde{E}_{p,s} = \sum_m A_{m,s}(N_{m,p} - N_{m,e}) \ , \end{aligned} \tag{4.4}$$

where n is the number of samples and to avoid the notational clutter we have replaced λ for $\lambda_{p,e}$. Integrating out σ^2 from the above equation, with a scale-invariant Jeffrey's prior over σ^2, yields

$$P(E_p|N,A,\lambda) \propto \int_0^\infty \sigma^{-n-2} \exp\Big(- \frac{\sum_s(\epsilon_{p,s} - \lambda \widetilde{E}_{p,s})^2}{2\sigma^2}\Big) \mathrm{d}\sigma^2 \tag{4.5}$$

By substituting $\sigma^2 = 1/t$ (and so that $\mathrm{d}\sigma^2 = -\mathrm{d}t/t^2$), and using Gamma integral equation, we can work out the integration for σ^2.

$$P(E_p|N,A,\lambda) \propto \Gamma(\frac{n}{2})\Big(\sum_{s=1}^{n} \big(\epsilon_{p,s} - \lambda \widetilde{E}_{p,s}\big)^2\Big)^{-n/2} \tag{4.6}$$

The maximum likelihood (ML) estimation for λ^*_{MLE} (λ^*, in short) can be found by finding the root of the first derivative of equation 4.6 with respect to the λ. We can do the same over the log-transformed $\mathcal{L}_\lambda = \log\big(P(E_p|N,A,\lambda)\big)$ of the equation 4.6.

$$\frac{\partial \mathcal{L}_\lambda}{\partial \lambda} = \frac{n \times \sum_s \widetilde{E}_{p,s}\big(\epsilon_{p,s} - \lambda \widetilde{E}_{p,s}\big)}{\sum_s \big(\epsilon_{p,s} - \lambda \widetilde{E}_{p,s}\big)^2} \tag{4.7}$$

The above equation is equal to zero if the sum term inside the numerator becomes zero. Hence,

$$\lambda^* = \frac{\big\langle \epsilon_p . \widetilde{E}_p \big\rangle}{\big\langle \widetilde{E}_p^2 \big\rangle} \tag{4.8}$$

Negative inverse of the square root of the second derivative of $\mathcal{L}_\lambda$ provides the error-bar of the ML estimation λ^*.

$$\frac{\partial \mathcal{L}_\lambda^2}{\partial^2 \lambda} = \frac{-n \times \big\langle \widetilde{E}_p^2 \big\rangle}{\big\langle (\epsilon_{p,s} - \lambda \widetilde{E}_{p,s})^2 \big\rangle} \tag{4.9}$$

To summarize, the ML estimation of λ as well as the error-bar of the estimate is given by:

$$\lambda = \lambda^* \pm \frac{\sqrt{\left\langle (\epsilon_{p,s} - \lambda^* \widetilde{E}_{p,s})^2 \right\rangle}}{\sqrt{\langle \widetilde{E}_p^2 \rangle}} n^{-1/2} \tag{4.10}$$

The above calculation is a valid guess for the value of λ, as long as we pose an ignorant prior over the value of λ. In section 4.2, we have discussed that currently there are several methods for detecting candidate enhancers. This has led to a large number of available genome-wide data, hence we prefer to have more informative prior over λ rather than a totally ignorant prior. The idea of integrating previous observations into the prior of λ is explained in the following section

4.4.1 Informative prior over λ

Given the pile of the available genomic data that has been collected in the last decade regarding to the functional part of the genome, we might be able to incorporate previous knowledge into the prior of λ. There is a number experiments that are designed to draw a genome-wide picture of the state of the different genomic loci. For example, DNAse I hypersensitivity assay is one of those experiments which reveals the open chromatin, which are possibly active parts of the genome, and closed or inactive parts. Having an experiment such this can help us to disentangle the potentially active regulatory regions from the silence parts of genome. In the last years, ENCODE consortium has released the DNAse data for 98 human cell-lines. As it has been shown there is only about two percent of the genome being in open state. Therefore, the above approach to calculate λ might lead to a lot of undesired false positives.

To encode the open-chromatin data into a prior model, we assume that the closed state of a locus across many different cell-lines indicates a small λ, showing its lack of active regulatory activity. On the other hand, when a locus is open in some circumstances, this may show its regulatory activity, however not necessary toward a particular promoter that we are testing for. For this purpose, we have chosen a truncated (between zero and one) exponential distribution which the scaling-factor of the distribution is controlled by the openness of the chromatin for a given locus.

$$P(\lambda|\alpha) = \frac{1}{\alpha} \exp(-\frac{\lambda}{\alpha}) \tag{4.11}$$

The value of α in equation 4.11 is set according to the fraction of time that the region is being open across all samples. As a result, this value reflects our prior expectation toward the value of λ; if the region is almost never open we do not expect it to have a functional regulatory activity and so the λ might be close to zero. On the other hand, if we find a region to be open in a number of samples, this information is transferred to the prior of λ by allowing relatively higher values of λ.

4.4.2 Identifying novel enhancers and enhancer-promoter linkage

Using the prior probability 4.11 and the already defined likelihood model 4.6, we are able to write down the posterior over λ.

$$P(\lambda|E_p, N, A, \alpha) \propto \frac{1}{\alpha} \exp(-\frac{\lambda}{\alpha}) \times \left(\sum_{s=1}^{n} \left(\epsilon_{p,s} - \lambda \widetilde{E}_{p,s}\right)^2 \right)^{-n/2} \tag{4.12}$$

Using this posterior, we can calculate the maximum a posteriori estimation (MAP) for λ in a brute-force manner. The above model, gives us a way to assess the enhancer activity of any region in regard to the expression of a gene. Care must be taken however when we are considering this model, because from the modeling perspective we have introduced more complexity to the model in compare with the more simpler approach which is only-promoter method (figure 4.1a). This might be true that the explained model gives us a good fit to the data but maybe the promoter-only model also fits the data decently. Therefore we must compare both strategy in the light data and systematically consider the complexity of each model in compare with each other. To this end, one might think of employing different model comparison methods. We use here *Bayes factor* statistic that is defined as:

$$\underbrace{\frac{P(M_1|D)}{P(M_2|D)}}_{\text{Posterior odds}} = \underbrace{\frac{P(M_1)}{P(M_2)}}_{\text{Prior odds}} \times \underbrace{\frac{\int_{\theta_1} P(D|M_1, \theta_1) P(\theta_1) \mathrm{d}\theta_1}{\int_{\theta_2} P(D|M_2, \theta_2) P(\theta_2) \mathrm{d}\theta_2}}_{\text{Bayes factor } K} \tag{4.13}$$

The equality between Posterior odds and Bayes factor is established if the Prior odds is one, or in other words we do not *apriori* prefer any model for the other. There are observations that show the genomic distance between two distal regions determines the background frequency in which two regions initiate a contact due to the loop formation of the DNA polymer. For instance, the recent Hi-C data empirically shows a power-law distribution, with scaling factor 1.08, for the background contact frequency between two regions that are d nucleosome fiber ($\sim$ 1.6 kb) apart. It is also theoretically shown that in self-avoiding polymer model the contact probability of two distal parts are proportional to the reverse of their distance. This information can be encoded in the prior of the enhancer-promoter model and as a result scales the Bayes factor in equation 4.13. But for now we assume that the Prior odds are equal to one, so let us focus on deriving the Bayes factor.

Deriving the enhancer-promoter model part of the Bayes factor K requires to integrate the 4.6 for λ.

$$P(E|N,A,\alpha) = \int_0^1 P(E|N,A,\lambda,\alpha)P(\lambda|\alpha)\mathrm{d}\lambda \tag{4.14}$$

Performing the above integral analytically is quite cumbersome, hence we can either resort numerical computation of the integral or using some approximations. If we are given a big enough sample size (say $n > 20$), the approximation might work good enough to estimate of the integral. Taylor expansion of the log of $P(E|N,A,\lambda,\alpha)$, defined above as $\mathcal{L}_\lambda$, around its maximum with regard to λ, equation 4.8, can provide an estimate for the true value of $\mathcal{L}_\lambda$.

$$\mathcal{L}_\lambda = \mathcal{L}_{\lambda^*} + \frac{\partial \mathcal{L}_\lambda}{\partial \lambda}\bigg|_{\lambda=\lambda^*}(\lambda-\lambda^*) + \frac{1}{2}\frac{\partial \mathcal{L}_\lambda^2}{\partial^2 \lambda}\bigg|_{\lambda=\lambda^*}(\lambda-\lambda^*)^2 + \dots \tag{4.15}$$

Where the second term in the above sum is equal to zero as we defined the first derivative around the maximum λ^*. Ignoring the higher terms of the Taylor expansion 4.15, and replacing $\phi^2 = -(1/\mathcal{L}''_{\lambda^*})$ results in:

$$P(E|N,A,\alpha) = \frac{\exp(\mathcal{L}_{\lambda^*})}{\alpha}\int_0^1 \exp\Big(-\frac{(\lambda-\lambda^*)^2}{2\phi^2} - \frac{\lambda}{\alpha}\Big)\mathrm{d}\lambda \tag{4.16}$$

With a little basic algebra, and defining $z = \lambda^* - (2\phi^2)/\alpha$, the above integral is equal to:

$$P(E|N,A,\alpha) = \frac{\phi\sqrt{2\pi}\exp(\mathcal{L}_{\lambda^*})}{2\alpha} \times \exp\big(\frac{z^2 - {\lambda^*}^2}{2\phi^2}\big)\,\mathrm{erf}(1) \tag{4.17}$$

In this equation $\mathrm{erf}(x)$ is the standard error-function. By using the above integral, and the likelihood of only-promoter model, we can define the Bayes factor K as it follows:

$$K \;=\; c \times \frac{\phi}{\alpha}\exp\big(\frac{z^2 - {\lambda^*}^2}{2\phi^2}\big)\left(\frac{\sum_s\big(\epsilon_{p,s} - \lambda^*\widetilde{E}_{p,s}\big)^2}{\sum_s \epsilon_{p,s}^2}\right)^{-n/2} \tag{4.18}$$

Where c is defined as $(\sqrt{2\pi}/2) \times \text{erf}(1)$, which is equal to 1.05617 . In addition to the Bayes factor K as it defined by 4.18, one can take into account the genomic distance d into this statistics. As it has been explained we would expect that the more closer both regions are, the more frequent would be their interaction. Here we incorporate this information into the Prior odds, in equation 4.13, by using a power-law distribution.

$$P = d^{-\beta} \times K \tag{4.19}$$

where d indicates the genomic distance between two genomic loci. For instance, one can define the d based on how many kilo base-pair two regions are apart, or similarly how many nucleosome fiber (about 1600 bp) as the basic packing unit of DNA. The scaling factor $\beta > 1$ can be set according to the other experiments or theoretical works. Here we set $\beta = 1.08$ which is previously fit to the Hi-C data.

As a result of equation 4.19 we propose a method that for a given promoter tests whether or not any genomic region can potentially be an enhancer. As long as we are in possession of the motif activity matrix $\mathbf{A}$ and the expression matrix $\mathbf{E}$ for a given system, such as stem-cell differentiation or immune respond, we can systematically determine the enhancers and more importantly link the promoters to the discovered enhancers. Therefore this method does not assume that any enhancer interacts specifically the closest promoter, whereas the association is established merely by the function of enhancers. In order to learn the motif activity matrix $\mathbf{A}$ we need however to have the enhancer promoter relationships with their associated interaction λ. Since this is not possible to have enhancers in advance, we simplify the problem by learning $\mathbf{A}$ under the promoter-only model. And later use this matrix in order to detect functional enhancers.

4.5 Results

We have tested this model over the FANTOM time-course CAGE expression data on Adipocyte. Using an iterative algorithm that will be explained below, we have found a number of enhancers that adding their site-counts to the site-counts of associated promoter regions increases the goodness of fit, if it is compared to a simpler promoter-only model.

4.5.1 Fitting motif activities in the enhancer-promoter model

In the last sections a model is presented that can be used to identify candidate enhancers for promoters. The new model, which consists of enhancers and promoters, proposes a new way to investigate on the motif activities across the gene expression dynamics. Instead of a site-count matrix **N** that includes promoters' site-counts for each motif, we have replaced it by new matrix **M** whose elements indicate the weighted sum of the sites in promoter and its associated enhancer.

To learn the model parameters λ for each promoter and its enhancer, and incorporate it into a new site-count matrix, we propose an iterative procedure. By starting from a simple model, that only includes promoters – that is simply the basic MARA model – we fit the activities A_{ms}. After that we sample a small number of promoters, and for each of the sampled promoters, slide a fixed-size window up and down stream of the promoter. The exonic and promoter regions are excluded from the windows. Then the Bayes factor, equation 4.18, is calculated to test whether the window can be a potential enhancer. We then select the window with the highest Bayes factor among the windows with a Bayes factor of at least 10. For the selected window, the maximum a posteriori (MAP) estimation for the λ is found. Using the MAP estimate of λ, we add the site-count of the identified enhancer to the promoter. After that, a new site-count matrix is built over the basic MARA site-count matrix. By the same method as MARA, the new motif activities are calculated, and the fraction of explained variance between the new model and the previous model is found. Once the difference in the fraction of explained variance between two consequent models is low enough, the iteration stops. Otherwise, a new set of promoters are sampled and their best enhancers site-counts are added to the site-count matrix by the same way as it has been explained.

4.5.2 Adipocyte time-course data

Adipose tissue can account for between 5% in athletes to 60% of total body mass, which makes it one of the most plastic organs in the body. It is demonstrated that under stable conditions there is about 10% of the adipocytes turn over annually [96]. As a part of FANTOM5 project, a time-course data of in vitro differentiation of adipocyte from the adipose-derived mesenchymal stem cells (hADSC) is produced that consists of 17 time points and done in triplicates. We have used this data to build a model for explaining the gene expression dynamics across time point. The final model includes enhancers along side with promoters. By the approach that is explained in last section, a set of candidate enhancers are identified and the site-count of the best enhancer is added to the associated promoter's site-count. This result is shown in figure 4.2 that in the course of iteration for learning the model, the goodness of fit, or the fraction of explained variance, improves. This is shown by the red curve. To test whether there is a systematic bias for the fact that a more complex model tend to explain the data better than the simpler model, we have shuffled the enhancer-promoter linkages. This is

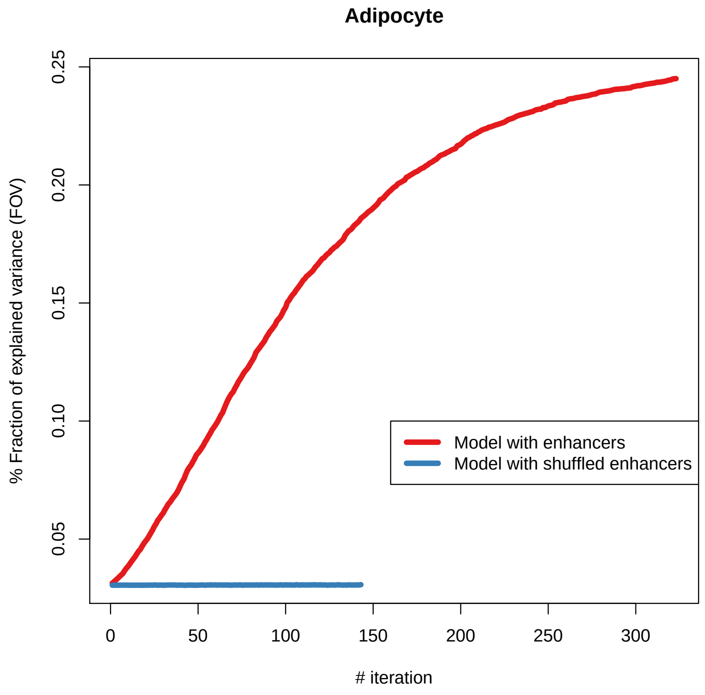


FIGURE 4.2: The improvement of fit for a model with enhancer once it is started from the basic MARA model, which only includes the promoters. The x-axis indicate the iteration time, at each step a small subset of promoters are sampled and for each one the best enhancer within 100 kb up/downstream is identified. The blue curve is a situation that the identified enhancers, not necessary the best one, is assigned to a randomly chosen promoter.

done by picking an arbitrary candidate enhancer, not necessary the best enhancer, and instead of adding its site-count to its promoter, adding it a randomly selected promoter.

We then looked at the activity profile of motifs. It is observed that for many motif, the activity of the motif from a simple no-enhancer model amplifies once the enhancers are added to the model. However, for some motifs the activity profile changes with the enhancer-promoter model. This is shown for the top four motif in this system (figure 4.3), selected according to their Z-scores.

In the figure 4.4, the Chi-square score of each promoter is shown under the two models. For each promoter there are two Chi-square scores which is calculated by the basic MARA model and the final enhancer-promoter model. It is clearly shown that for the promoters that an enhancer is assigned, the Chi-square score improves under the new augmented model. Whereas for the promoter-only model, the difference between the intact promoters – the ones that no enhancer is found – in the two models is negligible.

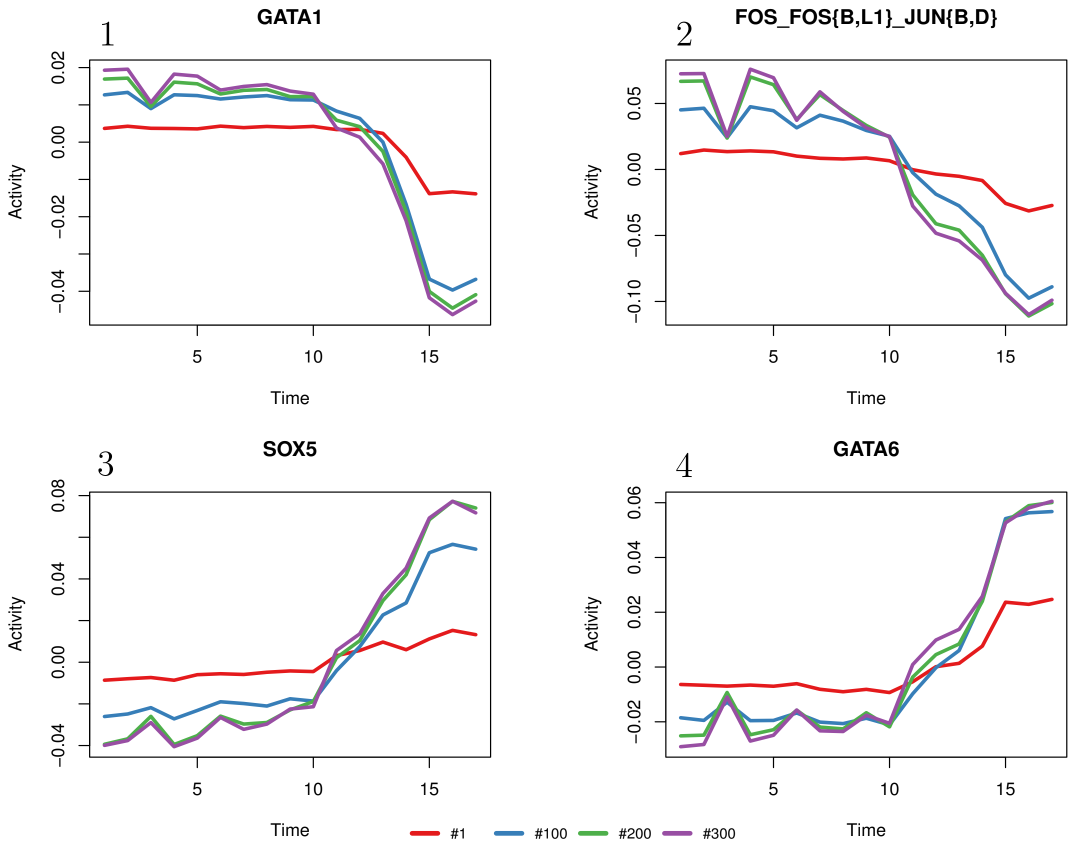


FIGURE 4.3: The activity profile of the top four motifs in adipocyte system. The activities are drawn for four time point of the iteration. Red curve indicates basic MARA model. It is then followed by three other time points, after adding more enhancers to the promoters' site-counts.

4.6 Dynamics of promoter-enhancer interaction

In the previous section we outlined a statistical investigation on the function of enhancers as the potential gene regulators. We, however, did not address the enhancers activity in action. It has been shown and discussed that enhancers, unlike promoters, are of more fluid nature. The DNA polymer is highly dynamics, with the different parts of the chain, under many known and unknown physical constraints, is flopping around dynamically, which this might bring a potential enhancer element in a close proximity to a promoter of a gene. Therefore, enhancer-promoter interaction can be variable under time and space. It is possible that an enhancer controls the transcription of a number of genes, and a gene is under the control of multiple enhancers at different times. In this section, we are suggesting a model or maybe better described as conjecture, for the dynamics of enhancer-promoter interaction. Figure **??** shows the aforementioned problem, multiple enhancers, in combination or solo, can play an activator role at a given time.

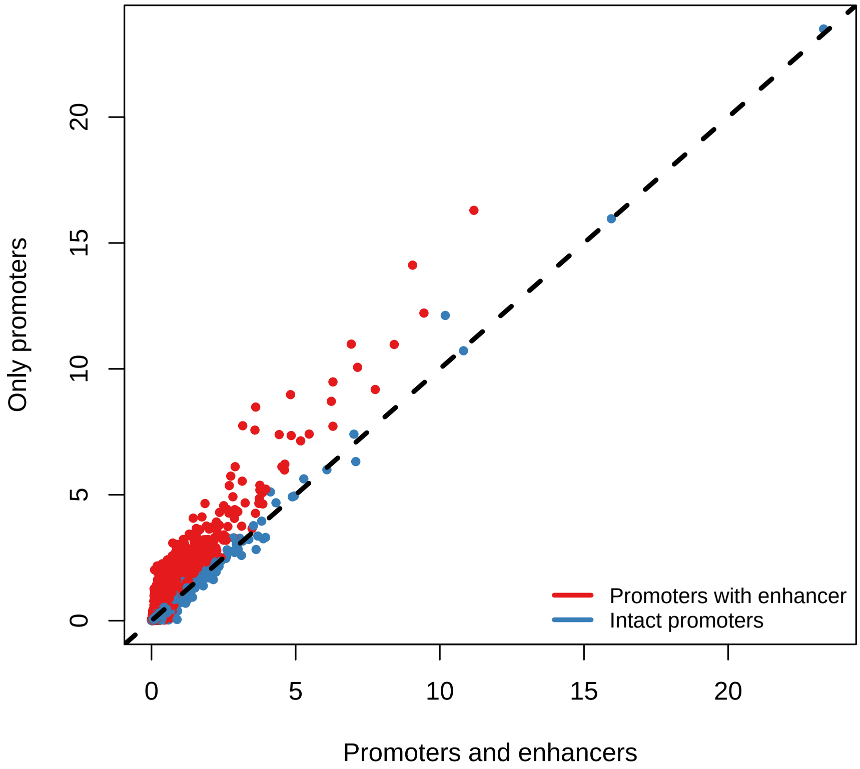


FIGURE 4.4: Each point represents a promoter and for each promoter the Chi-square score is calculated under both models: enhancer-promoter, and only promoter models. The promoters are divided into two classes, promoters with an enhancer assigned to them which is indicated in red; and promoters that no enhancer is found for them, which are represented in blue.

A precise physical model that explains which enhancers are activating the gene expression under a set of initial conditions, is highly complicated. Here, however, we tackle this problem in a less mathematical and computational complexity. By sticking to the same principles as the last section, we would like to model the dynamics of the enhancer-promoter interactions. More specifically, we are interested in knowing that the expression of a gene, in a given cellular condition, is under which set of enhancers' control. Putting differently, which of the potential enhancers are more plausible to regulate the expression of the gene. The gene expressions profiles at a different time points constitute the observables, and the other prerequisite is a list of candidate enhancers for each gene. To infer that which enhancer is playing an activator role at a time or what is the dynamic of promoter-enhancer interactions we use the Hidden Markov Model (HMM).

So far it is just as idea for modeling the dynamic of enhancer-promoter interaction, which can be explored today owing to the great collection of enhancers and other experimental

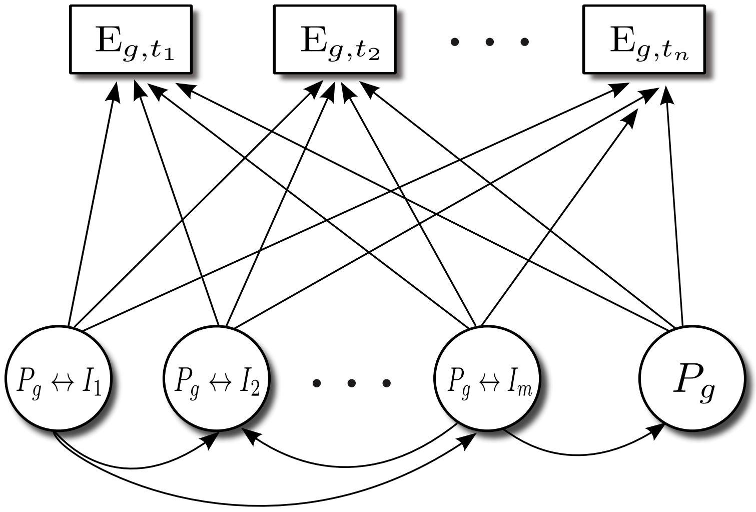

E_{g,t_1}
E_{g,t_2}
• • •
E_{g,t_n}
$P_g \leftrightarrow I_1$
$P_g \leftrightarrow I_2$
• • •
$P_g \leftrightarrow I_m$
P_g

FIGURE 4.5: The HMM model for enhancer-promoter interaction gene expression regulation. At a given time it can be any enhancer interacting with the gene's promoter upon which mediate the expression of the gene, or even the promoter alone does all the regulatory job. In addition, there is a dynamics between switching of enhancers from time to time.

data. One can study what features are contributing to the interaction of a specific enhancer with a particular promoter at a given time and space. It can be due to the activity of a subset of TF, or maybe other physical constraints such as the distance along the genome or the state of DNA inside the nucleus.

As a summary, we have investigated the role of enhancers from a transcription regulatory point of view that is mediated by the interaction of TF with the enhancer and promoters. We have extended the classical model for transcription regulation beyond the promoter sequences, by incorporating the possible interaction of distal enhancer regions to the promoters. The new model has shown an improvement in the model's fit, and resulted to a collection of the potential enhancers for each promoters. The dynamic of the enhancer-promoter is briefly explored by introducing a latent variable model for the enhancer usage. The motivation for this came from the fact that a given enhancer can control several promoters at different times, and each promoter can have multiple enhancers under different conditions.

Bibliography

[1] Jiro Kondo and Eric Westhof. Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. *Nucleic Acids Res*, 39(19):8628–37, October 2011. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/21737431.

[2] Nadrian C. Seeman, John M. Rosenberg, and Alexander Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA*, 73 (3):804–808, March 1976.

[3] Trevor Siggers and Raluca Gordan. Protein-dna binding: complexities and multi-protein codes. *Nucleic Acids Res*, 42(4):2099–111, February 2014. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/24243859.

[4] Shandar Ahmad, Hidetoshi Kono, Marcos J Araazo-Bravo, and Akinori Sarai. Readout: structure-based calculation of direct and indirect readout energies and specificities for protein-dna recognition. *Nucleic Acids Res*, 34(Web Server issue):W124–7, July 2006. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/16844974.

[5] Z Otwinowski, R W Schevitz, R G Zhang, C L Lawson, A Joachimiak, R Q Marmorstein, B F Luisi, and P B Sigler. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, 335(6188):321–9, September 1988. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/3419502.

[6] Remo Rohs, Xiangshu Jin, Sean M West, Rohit Joshi, Barry Honig, and Richard S Mann. Origins of specificity in protein-dna recognition. *Annu Rev Biochem*, 79: 233–69, 2010. ISSN 1545-4509. URL http://www.ncbi.nlm.nih.gov/pubmed/20334529.

[7] Remo Rohs, Sean M West, Peng Liu, and Barry Honig. Nuance in the double-helix and its role in protein-dna recognition. *Curr Opin Struct Biol*, 19(2):171–7, April 2009. ISSN 1879-033X. URL http://www.ncbi.nlm.nih.gov/pubmed/19362815.

[8] Z Shakked and D Rabinovich. The effect of the base sequence on the fine structure of the dna double helix. *Prog Biophys Mol Biol*, 47(3):159–95, 1986. ISSN 0079-6107. URL http://www.ncbi.nlm.nih.gov/pubmed/3544051.

[9] Sahand Jamal Rahi, Peter Virnau, Leonid A Mirny, and Mehran Kardar. Predicting transcription factor specificity with all-atom models. *Nucleic Acids Res*, 36(19): 6209–17, November 2008. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/18829719.

[10] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein-dna recognition. *Nature*, 461(7268):1248–53, October 2009. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/19865164.

[11] A G Murzin, S E Brenner, T Hubbard, and C Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, April 1995. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/7723011.

[12] Remo Rohs, Heinz Sklenar, and Zippora Shakked. Structural and energetic origins of sequence-specific dna bending: Monte carlo simulations of papillomavirus e2-dna binding sites. *Structure*, 13(10):1499–509, October 2005. ISSN 0969-2126. URL http://www.ncbi.nlm.nih.gov/pubmed/16216581.

[13] C O Pabo and L Nekludova. Geometric analysis and comparison of protein-dna interfaces: why is there no simple code for recognition? *J Mol Biol*, 301(3):597–624, August 2000. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/10966773.

[14] Y Q Chen, S Ghosh, and G Ghosh. A novel dna recognition mode by the nf-kappa b p65 homodimer. *Nat Struct Biol*, 5(1):67–73, January 1998. ISSN 1072-8368. URL http://www.ncbi.nlm.nih.gov/pubmed/9437432.

[15] Y Q Chen, L L Sengchanthalangsy, A Hackett, and G Ghosh. Nf-kappab p65 (rela) homodimer uses distinct mechanisms to recognize dna targets. *Structure*, 8(4):419–28, April 2000. ISSN 0969-2126. URL http://www.ncbi.nlm.nih.gov/pubmed/10801482.

[16] M Suzuki, M Gerstein, and N Yagi. Stereochemical basis of dna recognition by zn fingers. *Nucleic Acids Res*, 22(16):3397–405, August 1994. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/8078776.

[17] J Jiang and M Levine. Binding affinities and cooperative interactions with bhlh activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72

(5):741–52, March 1993. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/8453668.

[18] Michael A White, Davis S Parker, Scott Barolo, and Barak A Cohen. A model of spatially restricted transcription in opposing gradients of activators and repressors. *Mol Syst Biol*, 8:614, 2012. ISSN 1744-4292. URL http://www.ncbi.nlm.nih.gov/pubmed/23010997.

[19] Raffaella Scardigli, Nicola Baumer, Peter Gruss, FranÃois Guillemot, and Isabelle Le Roux. Direct and concentration-dependent regulation of the proneural gene neurogenin2 by pax6. *Development*, 130(14):3269–81, July 2003. ISSN 0950-1991. URL http://www.ncbi.nlm.nih.gov/pubmed/12783797.

[20] G Struhl, K Struhl, and P M Macdonald. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell*, 57(7):1259–73, June 1989. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/2567637.

[21] Sheldon Rowan, Trevor Siggers, Salil A Lachke, Yingzi Yue, Martha L Bulyk, and Richard L Maas. Precise temporal control of the eye regulatory gene pax6 via enhancer-binding site affinity. *Genes Dev*, 24(10):980–5, May 2010. ISSN 1549-5477. URL http://www.ncbi.nlm.nih.gov/pubmed/20413611.

[22] Peter C Hollenhorst, Atul A Shah, Christopher Hopkins, and Barbara J Graves. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ets gene family. *Genes Dev*, 21(15):1882–94, August 2007. ISSN 0890-9369. URL http://www.ncbi.nlm.nih.gov/pubmed/17652178.

[23] Matthew T Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, DREAM5 Consortium, Harmen J Bussemaker, Quaid D Morris, Martha L Bulyk, Gustavo Stolovitzky, and Timothy R Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol*, 31(2): 126–34, February 2013. ISSN 1546-1696. URL http://www.ncbi.nlm.nih.gov/pubmed/23354101.

[24] EA Feingold, PJ Good, MS Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, FS Collins, TR Gingeras, D. Kampa, EA Sekinger, et al. The encode(encyclopedia of dna elements) project. *Science(Washington)*, 306(5696):636–40, 2004.

[25] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M Vaquerizas, Renaud Vincentelli, Nicholas M Luscombe, Timothy R Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale.

Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39, January 2013. ISSN 1097-4172. URL http://www.ncbi.nlm.nih.gov/pubmed/23332764.

[26] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanpaa, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R Hughes, Nicholas M Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Res*, 20(6):861–73, June 2010. ISSN 1549-5469. URL http://www.ncbi.nlm.nih.gov/pubmed/20378718.

[27] Simone Furini, Paolo Barbini, and Carmen Domene. Dna-recognition process described by md simulations of the lactose repressor protein on a specific and a non-specific dna sequence. *Nucleic Acids Res*, 41(7):3963–72, April 2013. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/23430151.

[28] ENCODE Project Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/22955616.

[29] Raluca Gordan, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L Bulyk. Genomic regions flanking e-box binding sites influence dna binding specificity of bhlh transcription factors through dna shape. *Cell Rep*, 3 (4):1093–104, April 2013. ISSN 2211-1247. URL http://www.ncbi.nlm.nih.gov/pubmed/23562153.

[30] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–35, November 2006. ISSN 1087-0156. URL http://www.ncbi.nlm.nih.gov/pubmed/16998473.

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

[32] Gwenael Badis, Michael F Berger, Anthony A Philippakis, Shaheynoor Talukder, Andrew R Gehrke, Savina A Jaeger, Esther T Chan, Genita Metzler, Anastasia Vedenko, Xiaoyu Chen, Hanna Kuznetsov, Chi-Fong Wang, David Coburn, Daniel E Newburger, Quaid Morris, Timothy R Hughes, and Martha L Bulyk. Diversity and

complexity in dna recognition by transcription factors. *Science*, 324(5935):1720–3, June 2009. ISSN 1095-9203. URL http://www.ncbi.nlm.nih.gov/pubmed/19443739.

[33] Trevor Siggers, Abraham B Chang, Ana Teixeira, Daniel Wong, Kevin J Williams, Bilal Ahmed, Jiannis Ragoussis, Irina A Udalova, Stephen T Smale, and Martha L Bulyk. Principles of dimer-specific gene regulation revealed by a comprehensive characterization of nf-îb family dna binding. *Nat Immunol*, 13(1):95–102, January 2012. ISSN 1529-2916. URL http://www.ncbi.nlm.nih.gov/pubmed/22101729.

[34] Gwenael Badis, Esther T Chan, Harm van Bakel, and Lourdes Pena-Castillo. A library of yeast transcription factor motifs reveals a widespread function for rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell*, 32(6):878–87, December 2008. ISSN 1097-4164. URL http://www.ncbi.nlm.nih.gov/pubmed/19111667.

[35] N M Luscombe, S E Austin, H M Berman, and J M Thornton. An overview of the structures of protein-DNA complexes. *Genome biology*, 1:REVIEWS001, 2000. ISSN 1465-6906. doi: 10.1186/gb-2000-1-1-reviews001.

[36] T Funabiki, B L Kreider, and J N Ihle. The carboxyl domain of zinc fingers of the evi-1 myeloid transforming gene binds a consensus sequence of gaagatgag. *Oncogene*, 9(6):1575–81, June 1994. ISSN 0950-9232. URL http://www.ncbi.nlm.nih.gov/pubmed/8183551.

[37] R Delwel, T Funabiki, B L Kreider, K Morishita, and J N Ihle. Four of the seven zinc fingers of the evi-1 myeloid-transforming gene are required for sequence-specific binding to ga(c/t)aaga(t/c)aagataa. *Mol Cell Biol*, 13(7):4291–300, July 1993. ISSN 0270-7306. URL http://www.ncbi.nlm.nih.gov/pubmed/8321231.

[38] Suryani Lukman, David P Lane, and Chandra S Verma. Mapping the structural and dynamical features of multiple p53 dna binding domains: insights into loop 1 intrinsic dynamics. *PLoS One*, 8(11):e80221, 2013. ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/24324553.

[39] J Kim and K Struhl. Determinants of half-site spacing preferences that distinguish ap-1 and atf/creb bzip domains. *Nucleic Acids Res*, 23(13):2531–7, July 1995. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/7630732.

[40] Sepideh Khorasanizadeh and Fraydoon Rastinejad. Nuclear-receptor interactions on DNA-response elements, 2001. ISSN 09680004.

[41] R Kurokawa, J DiRenzo, M Boehm, J Sugarman, B Gloss, M G Rosenfeld, R A Heyman, and C K Glass. Regulation of retinoid signalling by receptor polarity and

allosteric control of ligand binding. *Nature*, 371(6497):528–31, October 1994. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/7935766.

[42] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, pages 99–104, 2004. doi: 10.1038/nature02886. Published. URL http://www.nature.com/nature/journal/v431/n7004/full/nature02800.html.

[43] Guillaume Paillard and Richard Lavery. Analyzing protein-dna recognition mechanisms. *Structure*, 12(1):113–22, January 2004. ISSN 0969-2126. URL http://www.ncbi.nlm.nih.gov/pubmed/14725771.

[44] Robert G Endres, Thomas C Schulthess, and Ned S Wingreen. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins*, 57(2):262–8, November 2004. ISSN 1097-0134. URL http://www.ncbi.nlm.nih.gov/pubmed/15340913.

[45] T K Man and G D Stormo. Non-independence of mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (qumfra) assay. *Nucleic Acids Res*, 29(12):2471–8, June 2001. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/11410653.

[46] Martha L Bulyk, Philip L F Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids res*, 30(5):1255–61, March 2002. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=101241&tool=pmcentrez&rendertype=abstract.

[47] R. Nutiu, R.C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G.P. Schroth, and C.B. Burge. Direct measurement of dna affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol*, 29(7):659–664, 2011. doi: 10.1038/nbt.1882. URL http://www.nature.com/nbt/journal/v29/n7/full/nbt.1882.html.

[48] Rahul Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS One*, 5(3):e9722, 2010. ISSN 1932-6203. URL http://www.ncbi.nlm.nih.gov/pubmed/20339533.

[49] Anthony Mathelier and Wyeth W Wasserman. The next generation of transcription factor binding site prediction. *PLoS Comput Biol*, 9(9):e1003214, September 2013. ISSN 1553-7358. URL http://www.ncbi.nlm.nih.gov/pubmed/24039567.

[50] Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. Modeling dependencies in protein-DNA binding sites. *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03*, pages 28–37, 2003. doi: 10.1145/640075.640079. URL http://portal.acm.org/citation.cfm?doid=640075.640079.

[51] Eilon Sharon, Shai Lubliner, and Eran Segal. A feature-based approach to modeling protein-dna interactions. *PLoS Comput Biol*, 4(8):e1000154, 2008. ISSN 1553-7358. URL http://www.ncbi.nlm.nih.gov/pubmed/18725950.

[52] Marc Santolini, Thierry Mora, and Vincent Hakim. Beyond position weight matrices: nucleotide correlations in transcription factor binding sites and their description. *arXiv:1302.4424v1*, 2013. URL http://arxiv.org/pdf/1302.4424v1.

[53] Marina Meila and Tommi Jaakkola. Tractable bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92, 2006.

[54] Lukas Burger and Erik van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. *Mol Syst Biol*, 4:165, 2008. ISSN 1744-4292. URL http://www.ncbi.nlm.nih.gov/pubmed/18277381.

[55] M. Townsend. *Discrete Mathematics: Applied Combinatorics and Graph Theory*. Benjamin Cummings, Menlo Park, 1987.

[56] S. S. Skiena. *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. Addison-Wesley, Reading, Massachusetts, 1990.

[57] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/19261174.

[58] Harukazu Suzuki. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics*, 41:553–562, 2009. ISSN 1061-4036. doi: 10.1038/ng.375.

[59] Rahul Siddharthan and Erik van Nimwegen. Detecting regulatory sites using phylogibbs. *Methods Mol Biol*, 395:381–402, 2007. ISSN 1064-3745. URL http://www.ncbi.nlm.nih.gov/pubmed/17993687.

[60] C Notredame, D G Higgins, and J Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302:205–217, 2000. ISSN 0022-2836. doi: 10.1006/jmbi.2000.4042.

[61] Mikhail Pachkov, Piotr J Balwierz, Phil Arnold, Evgeniy Ozonov, and Erik van Nimwegen. Swissregulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Res*, 41(Database issue):D214–20, January 2013. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/23180783.

[62] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–6, January 2008. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/18006571.

[63] Yaron Orenstein and Ron Shamir. A comparative analysis of transcription factor binding models learned from pbm, ht-selex and chip data. *Nucleic Acids Res*, February 2014. ISSN 1362-4962. URL http://www.ncbi.nlm.nih.gov/pubmed/24500199.

[64] P. Arnold, I. Erb, M. Pachkov, N. Molina, and E. van Nimwegen. Motevo: integrated bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of dna sequences. *Bioinformatics*, 28(4):487–494, 2012. URL http://bioinformatics.oxfordjournals.org/content/28/4/487.abstract.

[65] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *Science*, 249(4968):505, 1990.

[66] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[67] Silvia Acid, Luis M. de Campos, Juan M. Fernández-Luna, Susana Rodríguez, José María Rodríguez, and José Luis Salcedo. A comparison of learning algorithms for bayesian networks: A case study based on data from an emergency medical service. *Artificial Intelligence in Medicine*, 30(3):215–232, 2004.

[68] J. Cheng, DA Bell, and W. Liu. An algorithm for Bayesian network construction from data. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

[69] David Heckerman, Dan Geiger, and David M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20: 197, 1995.

[70] Sriram Pemmaraju and Steven Skiena. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. Cambridge University Press, Cambridge, 2003.

[71] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge Univ. Press, 2007.

[72] David MacKay. *Information Theory, Inference, and Learning Algorithms*. September 2003.

[73] Thomas P. Minka. Bayesian inference, entropy, and the multinomial distribution, February 10 2000. URL http://citeseer.ist.psu.edu/171980.html;ftp://vismod.www.media.mit.edu/pub/tpminka/papers/minka-multinomial.ps.gz.

[74] David Wolf. Mutual information as a bayesian measure of independence, November 06 1995. URL http://arxiv.org/abs/comp-gas/9511002.

[75] J Banerji, S Rusconi, and W Schaffner. Expression of a beta-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2 Pt 1):299–308, December 1981. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/6277502.

[76] Michael T Y Lam, Wenbo Li, Michael G Rosenfeld, and Christopher K Glass. Enhancer rnas and regulated transcriptional programs. *Trends Biochem Sci*, March 2014. ISSN 0968-0004. URL http://www.ncbi.nlm.nih.gov/pubmed/24674738.

[77] Wei Xie and Bing Ren. Developmental biology. enhancing pluripotency and lineage specification. *Science*, 341(6143):245–7, July 2013. ISSN 1095-9203. URL http://www.ncbi.nlm.nih.gov/pubmed/23869010.

[78] Michael Bulger and Mark Groudine. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol*, 339(2):250–7, March 2010. ISSN 1095-564X. URL http://www.ncbi.nlm.nih.gov/pubmed/20025863.

[79] Richard A. Young. Control of the embryonic stem cell state, 2011. ISSN 00928674.

[80] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, Yuriy L Orlov, Weiwei Zhang, Jianming Jiang, Yuin-Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing-Kin Sung, Neil D Clarke, Chia-Lin Wei, and Huck-Hui Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–17, June 2008. ISSN 1097-4172. URL http://www.ncbi.nlm.nih.gov/pubmed/18555785.

[81] Housheng Hansen He, Clifford A Meyer, Hyunjin Shin, Shannon T Bailey, Gang Wei, Qianben Wang, Yong Zhang, Kexin Xu, Min Ni, Mathieu Lupien, Piotr Mieczkowski, Jason D Lieb, Keji Zhao, Myles Brown, and X Shirley Liu. Nucleosome dynamics define transcriptional enhancers. *Nat Genet*, 42(4):343–7, April 2010. ISSN 1546-1718. URL http://www.ncbi.nlm.nih.gov/pubmed/20208536.

[82] Robert E Thurman, Eric Rynes, and Richard Humbert. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/22955617.

[83] M Merika, A J Williams, G Chen, T Collins, and D Thanos. Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Molecular cell*, 1:277–287, 1998. ISSN 10972765. doi: 10.1016/S1097-2765(00)80028-3.

[84] Marcelo A Nobrega, Ivan Ovcharenko, Veena Afzal, and Edward M Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, October 2003. ISSN 1095-9203. URL http://www.ncbi.nlm.nih.gov/pubmed/14563999.

[85] Elisa de la Calle-Mustienes, CÃ¡rmen Gloria FeijÃo, Miguel Manzanares, Juan J Tena, Elisa RodrÃguez-Seguel, Annalisa Letizia, Miguel L Allende, and JosÃ Luis GÃmez-Skarmeta. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate iroquois cluster gene deserts. *Genome Res*, 15(8):1061–72, August 2005. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/16024824.

[86] Adam Woolfe, Martin Goodson, Debbie K. Goode, Phil Snell, Gayle K. McEwen, Tanya Vavouri, Sarah F. Smith, Phil North, Heather Callaway, Krys Kelly, Klaudia Walter, Irina Abnizova, Walter Gilks, Yvonne J K Edwards, Julie E. Cooke, and Greg Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, 3, 2005. ISSN 15449173. doi: 10.1371/journal.pbio.0030007.

[87] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M Rubin, and Len A Pennacchio. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–8, February 2009. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/19212405.

[88] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A Brugmann, Ryan A Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–83, February 2011. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/21160473.

[89] Nasun Hah, Charles G Danko, Leighton Core, Joshua J Waterfall, Adam Siepel, John T Lis, and W Lee Kraus. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, 145(4):622–34, May 2011. ISSN 1097-4172. URL http://www.ncbi.nlm.nih.gov/pubmed/21549415.

[90] Dong Wang, Ivan Garcia-Bassets, Chris Benner, Wenbo Li, Xue Su, Yiming Zhou, Jinsong Qiu, Wen Liu, Minna U Kaikkonen, Kenneth A Ohgi, Christopher K Glass, Michael G Rosenfeld, and Xiang-Dong Fu. Reprogramming transcription by distinct classes of enhancers functionally defined by erna. *Nature*, 474(7351):390–4, June 2011. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/21572438.

[91] Robin Andersson. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61, March 2014. ISSN 1476-4687. URL http://www.ncbi.nlm.nih.gov/pubmed/24670763.

[92] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science (New York, N.Y.)*, 295:1306–1311, 2002. ISSN 00368075. doi: 10.1126/science.1067799.

[93] JosÃe Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D Green, and Job Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, 16(10):1299–309, October 2006. ISSN 1088-9051. URL http://www.ncbi.nlm.nih.gov/pubmed/16954542.

[94] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, October 2009. ISSN 1095-9203. URL http://www.ncbi.nlm.nih.gov/pubmed/19815776.

[95] Piotr J Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J Gruber, Mihaela Zavolan, and Erik van Nimwegen. Ismara: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res*, March 2014. ISSN 1549-5469. URL http://www.ncbi.nlm.nih.gov/pubmed/24515121.

[96] Kirsty L Spalding, Erik Arner, Pål O Westermark, Samuel Bernard, Bruce A Buchholz, Olaf Bergmann, Lennart Blomqvist, Johan Hoffstedt, Erik Näslund, Tom Britton, Hernan Concha, Moustapha Hassan, Mikael Rydén, Jonas Frisén, and Peter

Arner. Dynamics of fat cell turnover in humans. *Nature*, 453:783–787, 2008. ISSN 0029-7828. doi: 10.1097/01.ogx.0000325910.81966.ac.