Winter 12-15-2017

A Generalized Biophysical Model of Transcription Factor Binding Specificity and Its Application on High-Throughput SELEX Data

Shuxiang Ruan
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Gary D. Stormo, Chair
Jeremy D. Buhler
Barak A. Cohen
S. Joshua Swamidass
Ting Wang

A Generalized Biophysical Model of Transcription Factor Binding Specificity and Its Application on High-Throughput SELEX Data

by
Shuxiang Ruan

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2017
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to my advisor, Dr. Gary Stormo, for his guidance on mathematical modeling, data analysis, scientific writing, career planning and other important aspects of personal development. I also would like to thank my thesis committee for their guidance and advice on my research.

I would like to thank the financial support from the Genome Analysis Training Program, which is funded by the grant T32 HG000045 from the National Human Genome Research Institute.

In addition, I would like to thank all the members of the Stormo lab for the discussions we had and the suggestions they made.

Finally, I would like to thank my dear parents and friends. I would not finish my dissertation without their support and encouragement.

Shuxiang Ruan

Washington University in St. Louis

December 2017

ABSTRACT OF THE DISSERTATION

A Generalized Biophysical Model of Transcription Factor Binding Specificity and Its Application on High-Throughput SELEX Data

by

Shuxiang Ruan

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2017

Gary D. Stormo, Chair

The interaction between transcription factors (TFs) and DNA plays an important role in gene expression regulation. In the past, experiments on protein–DNA interactions could only identify a handful of sequences that a TF binds with high affinities. In recent years, several high-throughput experimental techniques, such as high-throughput SELEX (HT-SELEX), protein-binding microarrays (PBMs) and ChIP-seq, have been developed to estimate the relative binding affinities of large numbers of DNA sequences both *in vitro* and *in vivo*. The large volume of data generated by these techniques proved to be a challenge and prompted the development of novel motif discovery algorithms. These algorithms are based on a range of TF binding models, including the widely used probabilistic model that represents binding motifs as position frequency matrices (PFMs). However, the probabilistic model has limitations and the PFMs extracted from some of the high-throughput experiments are known to be suboptimal. In this dissertation, we attempt to address these important questions and develop a generalized biophysical model and an expectation maximization (EM) algorithm for estimating position weight matrices (PWMs) and other parameters using HT-SELEX data. First, we discuss the

inherent limitations of the popular probabilistic model and compare it with a biophysical model that assumes the nucleotides in a binding site contribute independently to its binding energy instead of binding probability. We use simulations to demonstrate that the biophysical model almost always provides better fits to the data and conclude that it should take the place of the probabilistic model in charactering TF binding specificity. Then we describe a generalized biophysical model, which removes the assumption of known binding locations and is particularly suitable for modeling protein–DNA interactions in HT-SELEX experiments, and BEESEM, an EM algorithm capable of estimating the binding model and binding locations simultaneously. BEESEM can also calculate the confidence intervals of the estimated parameters in the binding model, a rare but useful feature among motif discovery algorithms. By comparing BEESEM with 5 other algorithms on HT-SELEX, PBM and ChIP-seq data, we demonstrate that BEESEM provides significantly better fits to *in vitro* data and is similar to the other methods (with one exception) on *in vivo* data under the criterion of the area under the receiver operating characteristic curve (AUROC). We also discuss the limitations of the AUROC criterion, which is purely rank-based and thus misses quantitative binding information. Finally, we investigate whether adding DNA shape features can significantly improve the accuracy of binding models. We evaluate the ability of the gradient boosting classifiers generated by DNAshapedTFBS, an algorithm that takes account of DNA shape features, to differentiate ChIP-seq peaks from random background sequences, and compare them with various matrix-based binding models. The results indicate that, compared with optimized PWMs, adding DNA shape features does not produce significantly better binding models and may increase the risk of overfitting on training datasets.

Chapter 1: Introduction

The interaction between proteins and genomic DNA plays a crucial role in many important cellular processes. For instance, the RNA polymerase interacts with DNA during transcription and uses it as a template for RNA synthesis [1, 2]. Another example is the formation of nucleosomes in which histones and DNA bind together to form a well-defined three-dimensional structure [3]. Finally, some epigenetic modifications such as DNA methylation, which alter DNA accessibility and chromatin structures, are carried out by the DNA methyltransferase and other proteins that mainly target CpG dinucleotides [4]. The transcription factors (TFs), the focus of this dissertation, are a special class of DNA-binding proteins that recognize specific DNA sequences and primarily regulate gene expression [5, 6]. In most species they constitute between 5% and 10% of all genes [7-9]. Most TFs have DNA-binding domains (DBDs), such as zinc fingers or homeodomains, that fold into well-defined three-dimensional structures and interact with DNA mainly through hydrogen bonds [10-12]. Unlike restriction enzymes, which only recognize well-defined restriction sites, a TF may bind a range of similar sequences with varying affinities, a feature crucial to the function of the TF [13, 14]. When occupying a binding site, the TF may recruit other proteins to facilitate the transcription of a nearby gene if it acts as an activator [15], or may simply block the transcription if it is a repressor [16]. Variations in either the TFs or their binding sites are associated with changes in gene expression, often with deleterious phenotypes, but are also associated with evolutionary divergence between species [17, 18]. Although some prominent TFs, including Sox [19], AP-1 [20, 21] and Sp1 [22], have been studied extensively due to their involvement in cell growth, sex determination, apoptosis or cancer development, the binding specificities of a large number of TFs are poorly documented

even in many well-studied species [23]. In recent years, several high-throughput experimental techniques, such as high-throughput SELEX (HT-SELEX), protein-binding microarrays (PBMs) and ChIP-seq, have been developed to estimate the relative binding affinities of large numbers of DNA sequences both *in vitro* and *in vivo* [24-26]. These techniques have greatly accelerated the study of TF binding specificity [5], but the analysis of their results proves to be a challenge and requires the development of novel TF binding models and motif discovery algorithms.

The mathematical models of TF binding specificity are constantly evolving with the experimental methods for measuring the binding affinities of DNA sequences [27]. At the time when the first few TFs such as the *lac* repressor were discovered, only a handful of high-affinity binding sites could be determined for each TF, and their binding specificities were simply represented by consensus sequences, which might include mixed or degenerate base notations [28]. As the number of identified binding sites grew, the consensus sequence was no longer a suitable representation, because all the mismatches to the consensus are treated equally [29]. This prompted the development of perceptron-based scoring matrices for discriminating positive sequences from negative sequences and then position frequency matrices (PFMs) [27]. The PFM assumes that the probability of observing a specific base at a specific position of a binding site is independent of the other positions [27], and is often associated with the information logo of a binding motif, which displays not only the base preference but also the information content at each position [30]. With the development of next-generation sequencing (NGS) and high-throughput methods based on NGS, including bacterial one-hybrid (B1H), HT-SELEX and ChIP-seq, it is possible to estimate the affinities of a TF to large numbers of sequences in one experiment [5]. Other methods, such as the mechanically induced trapping of molecular interactions (MITOMI) [31] and Spec-seq [32], can even directly measure the absolute or

relative association constants between a TF and different DNA sequences in a relatively high-throughput fashion. The large volume of data generated by these methods proved to be a challenge and prompted the development of novel binding models and motif discovery algorithms. A notable example is a biophysical model that is built on the thermodynamic equilibrium of protein–DNA association and contains parameters such as the binding energy and the chemical potential of the TF [33, 34]. It assumes that each base in a binding site contributes independently to the total binding energy of the sequence and thus any binding motif can be represented as a position weight matrix (PWM) of energy contributions [27]. It should be noted that we use the term 'PWM' to denote only the energy matrix in this dissertation, though it may refer to a general matrix-based scoring model in other publications. Other types of TF binding models have also been formulated, including *k*-mer models, hidden Markov models (HMMs), transcription factor flexible models (TFFMs), and deep neural networks [35-38]. Some of these models were compared with the biophysical model on PBM data, and the results indicated that the biophysical model generally provides the best trade-off between prediction accuracy and model complexity [39]. Finally, some studies have examined whether incorporating factors such as flanking sequences and DNA shape features can increase the accuracy of TF binding models [40]. These factors have been shown to have varying impacts on TF binding specificity [41], and their contributions are generally secondary to the DNA sequence of the binding site.

1.1 Matrix-Based Models of Transcription Factor Binding Specificity[1]

Probabilistic models (PMs) for DNA binding proteins were initially introduced by Harr *et al.* for *E. coli* promoters that even treated variable length binding sites [42]. Soon after, Staden converted to the use of log-probability to put the model into a weight matrix (additive) model, also including parameters for variable spacing [43]. Schneider *et al.* drew connections between the probabilistic models and information theory and introduced the log-odds model that accounts for the background distribution of bases [44] and later introduced the popular logo graphical representation of specificity [30]. The probabilistic model was also the basis of the earliest motif discovery algorithms [45-47]. Since then there have been many different algorithms for motif modeling and discovery using probabilistic models (reviewed in [27, 48-51]).

Even earlier von Hippel introduced an energy-based model of protein–DNA interactions [52]. At the time, there were almost no data on actual binding sites so the paper used first principles to describe the informational specificity required for functional regulatory sites. The paper made simplifying assumptions such as the independence between positions and that every mismatch from the preferred sequence had the same energy difference. The first assumption, of independent contributions, has proven to be a reasonably good approximation for most transcription factors, whereas differences in contributions of alternative bases at each position are now well known and form the basis of most specificity modeling approaches. Berg and von Hippel derived an energy model that was identical to the probabilistic one under some simplifying assumptions and connections between the energy approach and the information

[1] Parts of this section are taken from Ruan, S. & Stormo, G. D. Intrinsic limitations of probabilistic models for protein-DNA interactions. *PLOS Computational Biology* (2017).

theory models of specificity became clear [53-55]. Hwa and colleagues put the energy modeling approach into a more general biophysical model that accounts for the effects of protein concentration on binding probabilities [56, 57]. Djordjevic *et al.* pointed out the importance of the biophysical approach in modeling specificity [33]. They further provided an algorithm that is guaranteed, for any collection of known binding sites, to predict the minimum number of additional sites in a genome, thereby minimizing the number of false positive predictions, although the method is not guaranteed to provide a more accurate model of the true specificity [33, 58]. Regression methods have been used to find optimal energy parameters and Foat *et al.* provided the first regression algorithm for motif discovery of optimal energy models [59, 60]. Since then several related methods have been developed to determine biophysical (energy) models of protein specificity from various types of high-throughput experimental data [34, 39, 61-70].

Despite the development of several high-throughput experimental methods for measuring the specificity of protein–DNA interactions [5, 71] and the algorithms described above for modeling them with the biophysical approach, probabilistic models remain the most popular. The purpose of Chapter 2 is to point out that when good energy models are available there is no advantage to using the probabilistic models. In fact, due to inherent limitations the probabilistic models can be misleading and are highly sensitive to the samples used for inference of the parameters. Energy models can be readily obtained and can easily accommodate non-independent contributions between positions [27, 60, 72]. We conclude that energy modeling should become the approach generally used for modeling specificity and predicting protein–DNA interactions.

1.2 High-Throughput Methods for Measuring Binding Specificity[2]

Several high-throughput experimental methods have been developed for determining the specificity of TFs (reviewed in [5]). Some methods measure the fluorescence from TFs over thousands of DNA probes that contain millions of different possible DNA binding sites. Others take advantage of high-throughput sequencing technologies to determine selective enrichments of binding sites from millions of short sequence reads. Regardless of the technology used, computational analysis of the data is required to extract the desired specificity information for the TF and different programs vary widely in their ability to make accurate predictions of binding sites over a wide range of affinities [39].

Protein binding microarrays (PBMs), and related methods [25, 73, 74], utilize arrays of double-stranded DNA to which TF binding can be assayed with fluorescent antibodies to the proteins. Several large-scale PBM experiments have been published in which the specificities of several hundred [75-80], and even over 1000 [23], TFs have been determined. It was shown that the quality of the motif, the fit to the data and the ability to predict TF binding in an independent experiment varied considerably depending on the algorithm used [67]. In a detailed comparison of 26 different algorithms for analyzing PBM data a wide range of accuracies were found [39]. For the vast majority of TFs (~90%), simple PWMs fit the data as well as more complex models when the best algorithms were used. The best methods employed a biophysical model of protein–DNA interactions [33, 34, 59]. Recent enhancements to the FeatureREDUCE algorithm provide further improvements to the accuracy of motif inference from PBM data [65]. The

[2] Parts of this section are taken from Ruan, S., Swamidass, S. J. & Stormo, G. D. BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics* (2017).

BindSter algorithm [63] also provides improved motif inference for both PBM and MITOMI [31] data.

Systematic evolution of ligands by exponential enrichment (SELEX) [81] has also been adapted to utilize high-throughput sequencing [24, 34, 66, 82-85], most commonly called high-throughput SELEX (HT-SELEX) or alternatively SELEX-seq. After one or more rounds of selection, the bound fraction, as well as the input DNA, are sequenced to high depth (see Figure 1.1). Those sequences are used to infer a model of specificity, typically a PWM, for the TF. Some methods using biophysical models have been developed for the HT-SELEX problem [61, 64, 66] but only HTS-IBIS has been widely tested. In addition, the recent DeepBind algorithm [38], which is based on deep convolutional neural networks, reports improved predictions of *in*

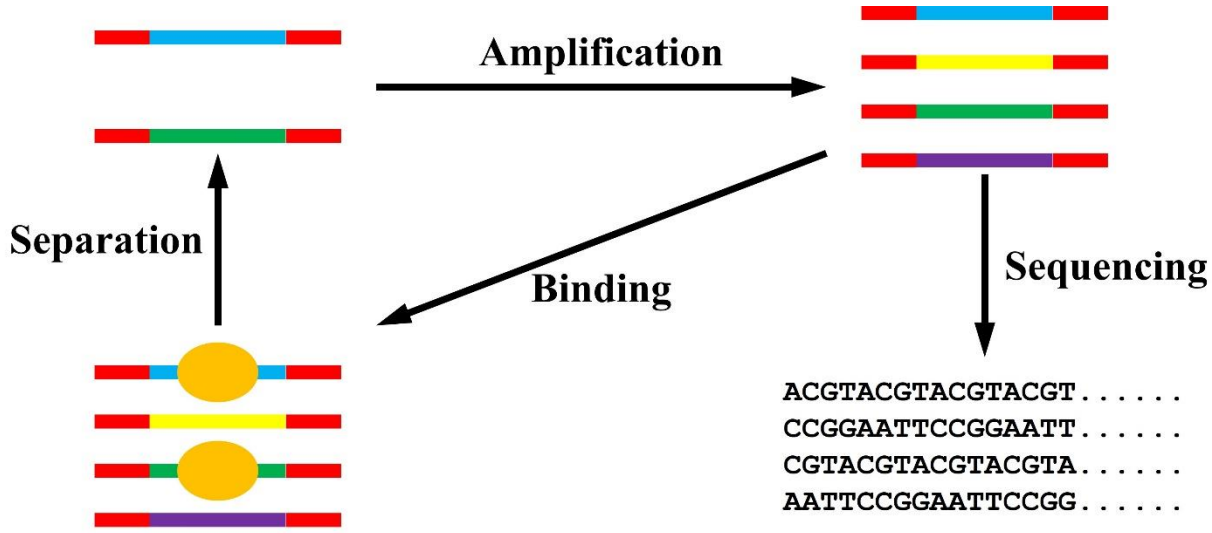


Figure 1.1 The workflow of high-throughput SELEX. Initial random DNA probes or sequences from the previous SELEX cycle are mixed with immobilized TF molecules, allowing for DNA binding by the protein. The unbound DNA sequences are washed away, and then the bound probes are separated from the protein. These bound sequences, which serve as the input of the next SELEX cycle, are amplified and then sequenced using high-throughput sequencing technologies. In this figure, red segments represent flanking regions. Other colored segments represent randomized regions. Orange ovals represent TF molecules.

vivo data, even when it is trained on *in vitro* HT-SELEX data. The most commonly used method for inferring motifs from HT-SELEX data uses the most highly selected sites from later rounds and builds a PFM by comparing the frequency of the preferred site with single-nucleotide variants [24, 86, 87]. This approach does not take advantage of the entire range of binding affinities and has the risk of producing 'over-specified' motifs, with higher information content than the true binding specificity. In a comparison of methods for the zinc finger protein Bcl6 we found that motifs inferred from PBM and bacterial one-hybrid data were very similar to each other and also to a motif inferred using MEME on ChIP-seq data [88]. The HT-SELEX motif for Bcl6 had much higher information content and was less effective in identifying binding sites in the ChIP-seq dataset (at equivalent p-value cutoffs), suggesting the over-specification phenomenon. In another comparison between PBM and HT-SELEX motifs for the same TFs, Orenstein *et al.* also found that the PBM motifs fit the quantitative binding data better, but they found that the HT-SELEX motifs performed better on ChIP-seq data, using the criterion of the area under the receiver operating characteristic curve (AUROC) [89].

Zhao *et al.* introduced the BEEML (short for Binding Energy Estimates using Maximum Likelihood) method that finds the best fit to the data over the parameters of the PWM and the TF concentration [34]. The biophysical model underlying BEEML is suitable for the analysis of HT-SELEX data, but BEEML assumes the location of the binding site on each sequence is known (only the orientation has to be inferred). In a general HT-SELEX experiment, however, the randomized region is much longer than the binding site; typical randomized regions are 20 bp or more while most binding motifs are 10 bp or less. Thus BEEML is limited to HT-SELEX experiments with short randomized regions. Another challenge of HT-SELEX data is that a library of random 20-mers contains over 10^{12} different sequences, well beyond the capacity of

current sequencing approaches. In fact, essentially all of the sequences in the initial pool are unique and the vast majority of possible sequences are missing from the pool. Since BEEML assumes that the read count of a sequence increases with its affinity, we need a new algorithm to tackle the challenge of very small read counts for each sequence.

Chapter 3 has two main purposes. First, we introduce BEESEM (short for Binding Energy Estimation on SELEX with Expectation Maximization), which extends BEEML to work on HT-SELEX data with long randomized regions. We extend the biophysical model used in BEEML to the case with long sequences containing shorter binding sites so that the sites must be inferred in addition to the motif. We do this using an expectation maximization (EM) approach similar to a previously introduced method for motif discovery on unaligned co-regulated gene sets [46], but including non-linear regression to fit the quantitative enrichment data as part of the maximization step. The method also allows us to calculate confidence intervals on the estimated parameters. Second, we assess BEESEM against other modeling approaches using HT-SELEX data, as well as PBM and ChIP-seq data from independent experiments. The results demonstrate that the BEESEM motifs achieve significantly better fits to the quantitative HT-SELEX data and we also show that they perform much better than other HT-SELEX binding models on PBM data and equally well or better on ChIP-seq data.

1.3 Contribution of DNA Shape Features to Binding Specificity

With the rapid advancements in X-ray crystallography [90], Cryo-electron microscopy [91] and nuclear magnetic resonance (NMR) spectroscopy [92], an increasing number of three-dimensional structures of protein–DNA complexes and DNA fragments have been solved. These discoveries not only advanced our understanding of protein–DNA interactions [93] but also

facilitated the design of customizable protein molecules that recognize specific DNA sequences. A notable example is the structure of the zinc finger DBD from Zif268, which was first solved in 1991 [10] and catalyzed the development of zinc-finger nucleases (ZFNs) [94]. As of 2017, the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), one of the largest repositories for three-dimensional structures of large biological molecules, contains more than 130,000 entries, including thousands of protein–DNA complexes [95]. The wealth of structural information enabled the development of computational methods for predicting DNA shape features, such as the helix twist (HelT), the minor groove width (MGW), the propeller twist (ProT), the roll (Roll), and their corresponding second-order shape features. One of these methods, DNAshape, predicts DNA structural features in a high-throughput manner based on Monte Carlo simulations of DNA fragments [96]. In addition, the Genome Browser for DNA shape annotations (GBshape), a database based on DNAshape and related computational tools, provides DNA shape feature predictions for a range of organisms [97]. These computational tools and databases made it possible to study the influence of DNA shape features on TF binding specificity. According to one study, the combination of chromatin endogenous cleavage (ChEC) and high-throughput sequencing reveals two classes (strong and weak) of binding sites for yeast TFs, and similar DNA shape patterns are observed in both classes regardless of binding strength [98]. However, there are doubts about some aspects of these findings [41]. In another study, researchers trained gradient boosting classifiers to differentiate ChIP-seq peaks from random background sequences and claimed that adding DNA shape features can significantly improve the accuracy of the classifiers [40].

In Chapter 4 of this dissertation, we seek to replicate the results of the second study mentioned above. In addition, we will compare the performance of the gradient boosting classifiers with

both the probabilistic models and the PWMs generated by DiMO_PWM, a perceptron-based optimization method that finds the optimal PWM with the highest AUROC [99]. Our preliminary results indicate that adding DNA shape features does not significantly improve the performance of the gradient boosting classifiers, and the optimal PWMs generated by DiMO_PWM can achieve similar or even better classification accuracy. More importantly, the matrix-based binding models are less likely to overfit, thus giving them an advantage over the complex gradient boosting classifiers.

Chapter 2: Inherent Limitations of Probabilistic Models[1]

The specificities of transcription factors are most commonly represented with probabilistic models. These models provide a probability for each base occurring at each position within the binding site and the positions are assumed to contribute independently. The model is simple and intuitive and is the basis for many motif discovery algorithms. However, the model also has inherent limitations that prevent it from accurately representing true binding probabilities, especially for the highest affinity sites under conditions of high protein concentration. The limitations are not due to the assumption of independence between positions but rather are caused by the non-linear relationship between binding affinity and binding probability and the fact that independent normalization at each position skews the site probabilities. Generally probabilistic models are reasonably good approximations, but new high-throughput methods allow for biophysical models with increased accuracy that should be used whenever possible.

2.1 Models for Transcription Factor Binding Specificity

2.1.1 Probabilistic Model

The probabilistic model is based on a probability matrix $PM(b, j)$ for each base $b \in \{A, C, G, T\}$ at each position $j = 1, 2, \cdots, m$ for an *m* bp binding site. Any *m* bp DNA sequence S_i can be encoded as a similar matrix $S_i(b, j)$, of 1s and 0s, where a 1 represents the base that occurs at position *j* and all the other elements are 0 [27]. The model assumes that the probability of sequence S_i being among the bound sites is

[1] Parts of this chapter are taken from Ruan, S. & Stormo, G. D. Intrinsic limitations of probabilistic models for protein-DNA interactions. *PLOS Computational Biology* (2017).

$$\tilde{P}(S_i \mid B) = \prod_{j=1}^{m} \prod_{b=A}^{T} PM(b, j)^{S_i(b,j)}, \tag{2.1}$$

where *B* represents the collection of bound DNA molecules. Often $PM(b, j)$ is converted to a log-odds weight matrix $WM(b, j) = \log[PM(b, j) / P(b)]$, where $P(b)$ is the background, or prior, probability of base *b* [27, 44]. For simplicity, we assume the prior probability is a constant, 0.25 for each base, and therefore the two matrix forms give equivalent results.

2.1.2 Biophysical Model

The biophysical model is based on the thermodynamics of the interaction between two molecules, the protein *T* and a binding site S_i. The association constant, which we refer to as the affinity, can be determined by measuring the concentrations of free reactants (protein and DNA) and the complex $T \cdot S_i$:

$$K_i = \frac{[T \cdot S_i]}{[T][S_i]}. \tag{2.2}$$

It is common to assume that the positions contribute independently to the binding affinity, just as the probabilistic model assumes the positions contribute independently to the site probability. This is represented as a matrix of affinity contributions $K(b, j)$ such that

$$K_i = \prod_{j=1}^{m} \prod_{b=A}^{T} K(b, j)^{S_i(b,j)}, \tag{2.3}$$

where $S_i(b, j)$ is the encoding matrix of S_i. From that one can determine the probability of a sequence S_i being bound based on the protein concentration and the association constant:

$$P(B \mid S_i) = \frac{[T \cdot S_i]}{[T \cdot S_i] + [S_i]} = \frac{K_i[T]}{K_i[T] + 1} = \frac{1}{1 + e^{E_i - \mu}}, \tag{2.4}$$

where $E_i = -\ln K_i$ is the free energy of binding to sequence S_i and $\mu = \ln[T]$ is the chemical potential of the protein. Based on the definition of the binding energy, Equation (2.3) can be rewritten as

$$E_i = \sum_{j=1}^{m} \sum_{b=A}^{T} E(b, j) S_i(b, j), \tag{2.5}$$

where $E(b, j) = -\ln K(b, j)$. Equation (2.5) means that the total binding energy can be represented by the inner product of the energy matrix $E(b, j)$ and the encoding of a DNA sequence. The probability of sequence S_i in the bound sequences can be obtained by Bayes' rule:

$$P(S_i \mid B) = P(B \mid S_i) \frac{P(S_i)}{P(B)} \propto \frac{K_i[T]}{K_i[T]+1} = \frac{1}{1+e^{E_i - \mu}}, \tag{2.6}$$

if the background probability $P(S_i)$ is the same for all S_i. In the biophysical model, $P(S_i \mid B)$ is dependent on the chemical potential, which differs from the probabilistic model, where $\tilde{P}(S_i \mid B)$ is entirely determined by the sequence. More importantly $P(S_i \mid B)$ has a non-linear relationship with the binding affinity K_i. This becomes pronounced at high protein concentrations where the energy can be additive across the positions of the binding site and yet the probabilities of the bases at each position are not independent.

2.2 Methods

2.2.1 Simulation Procedure

We developed a program, BEnDS (Binding Energy Distribution Simulations), for generating a probability distribution of bound sequences $P(S_i \mid B)$ based on user-specified values of the motif length *m* and chemical potential μ. It first generates a random energy matrix $E(b, j)$ of length *m*.

One base is randomly chosen as the preferred base at each position and assigned an energy of 0. Energies for the other bases are drawn randomly from a normal distribution with a user-specified mean and standard deviation ($N(2.5,1.0^2)$ by default). The generated energy matrix $E(b, j)$ and the chemical potential μ will serve as the true parameters of the underlying biophysical model. Given the true parameters, probabilities of binding to all possible sites, $P(B \mid S_i)$, are obtained using Equation (2.4) and (2.5). Finally according to Equation (2.6), if we assume an equiprobable prior distribution for all the sequences, we can compute $P(S_i \mid B)$, the probability of sequence S_i among the bound sequences, by simply normalizing $P(B \mid S_i)$. The clean probability $P(S_i \mid B)$, free from measurement errors, will be used to evaluate the predictions of the estimated probabilistic models and the fitted biophysical models (described in following subsections).

As an option, BEnDS can simulate errors on binding energy measurement. It first randomly generates an error ε_i from a normal distribution with a user-specified standard deviation ($N(0,0.5^2)$ by default). The error is then added to the true binding energy of sequence S_i to produce the perturbed energy $E_i^* = E_i + \varepsilon_i$. Then we use E_i^* to compute the probability of a sequence S_i being bound under noise, $P^*(B \mid S_i)$, based on Equation (2.4). Finally, we normalize $P^*(B \mid S_i)$ to obtain the noisy probability distribution $P^*(S_i \mid B)$. When there are measurement errors, only $P^*(S_i \mid B)$ can be observed from experiments. Thus, we use $P^*(S_i \mid B)$ instead of the clean probability $P(S_i \mid B)$ to estimate the parameters in the probabilistic models and the fitted biophysical models.

2.2.2 The Estimated Probabilistic Models

From the observed probability distribution $P^*(S_i \mid B)$, probabilistic models (PMs) were determined by counting the frequencies of the bases at each position. This was done both for the entire distribution and from a subset of high affinity sites (ranked by the observed probability $P^*(S_i \mid B)$), such as the top 1% (as might be expected to be functional sites). When only the top 1% sites are used, the PMs could be obtained either weighted by the site probabilities, or just from the list of sites unweighted, as one might expect from a collection of known regulatory sites or from ChIP-seq type of experiment with a limited sample of observed binding sites. With the estimated PM, we can compute its predicted $\hat{P}(S_i \mid B)$ and compare it with the clean probability $P(S_i \mid B)$ generated by the underlying biophysical model.

2.2.3 The Fitted Biophysical Model

To fit a biophysical model to the noisy observation $P^*(S_i \mid B)$, we solve the following optimization problem:

$$\min_{\theta} \sum_i \left[\ln \frac{A}{1+e^{E_i-\mu}} - \ln P^*(S_i \mid B) \right]^2, \tag{2.7}$$

where θ is the vector of unknown parameters, including the scale factor A, the chemical potential μ, and the energy matrix $E(b, j)$, which is used to compute E_i based on Equation (2.5). The optimization problem is solved using the L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm [100]. The fitted biophysical models will be estimated using either all the sequences or only the top 1% sequences ranked by the observed probability $P^*(S_i \mid B)$. With the estimated parameters, we can compute the $\hat{P}(S_i \mid B)$ predicted by the fitted biophysical model and compare it with the clean probability $P(S_i \mid B)$ generated from the true parameters.

2.3 Results

We compared the predicted probability distribution $\hat{P}(S_i \mid B)$ of the estimated PMs and the fitted biophysical models with the true probability distribution $P(S_i \mid B)$. Of particular interest is how well the rank order of binding site probabilities is preserved.

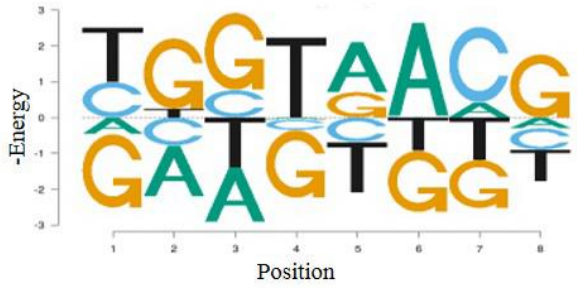
2.3.1 Results in the Absence of Measurement Noise

In the absence of measurement noise, we mainly focus on the estimated PMs as the fitted biophysical model is the same as the underlying biophysical model.

At different protein concentrations, the PMs derived from binding probabilities are usually different. As shown in Equation (2.4), the binding probability of a sequence S_i depends on both its binding affinity K_i (or energy E_i) and the protein concentration (or chemical potential μ). If $K_i[T] \ll 1$, there is a linear relationship between affinity and probability, but that occurs only when $P(B \mid S_i) \ll 0.5$, which is unlikely to be the case *in vivo* for true regulatory sites. At high protein concentrations, where $K_i[T] > 1$ and the preferred binding site is highly occupied, the non-linear relationship between binding probability and affinity has several consequences. One is that the PM itself depends on the protein concentration, whereas the binding energy does not. Figure 2.1a and 2.1b show one example of a simulated energy matrix and its associated energy logo [27, 59]. In the matrix the lowest energy base (preferred base) at each position is assigned energy 0 (using the convention of Berg and von Hippel [53]), while in the logo the average energy for each position is set to 0, with the lower energy (higher affinity) bases on top. Figure 2.1c and 2.1d show the information logo [30] and the PM obtained at very low protein concentration, $\mu = -3$. At low protein concentration the PM corresponds very closely to the independent contributions of each base to the binding affinity (equation (2.6) converges to

equation (2.1)). But at high protein concentration, such as $\mu = 3$, the logo and PM are different (Figure 2.1e and 2.1f). The second logo shows that the information is 'compressed' at $\mu = 3$, with the mean column information content (MCIC) decreasing from 0.9 bit to 0.7 bit. The MCIC is defined as the average information content of all the columns in a matrix. Comparing the two

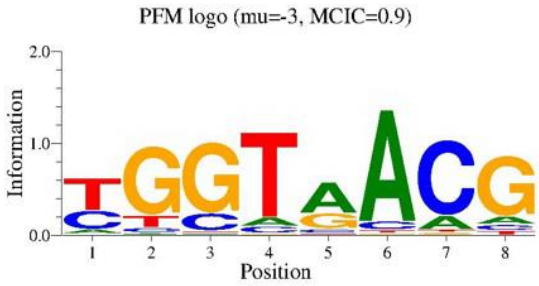
a



b

Base	Position							
	1	2	3	4	5	6	7	8
A	1.98	3.37	3.66	2.25	0	0	1.69	2.04
C	0.55	2.73	1.39	2.50	2.07	2.56	0	2.35
G	3.56	0	0	4.09	0.68	4.28	3.43	0
T	0	1.73	3.53	0	2.78	3.51	3.27	2.60

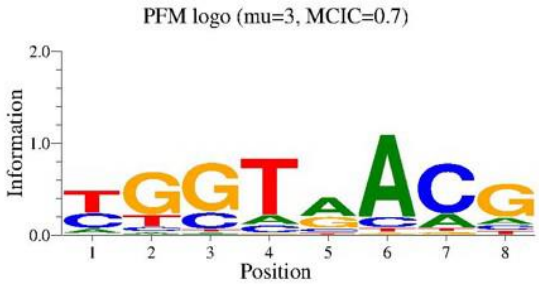
c



d

Base	Position							
	1	2	3	4	5	6	7	8
A	8	3	2	9	59	89	15	10
C	33	5	19	7	7	7	79	7
G	2	78	77	1	30	1	3	77
T	57	14	2	83	4	3	3	6

e



f

Base	Position							
	1	2	3	4	5	6	7	8
A	11	5	3	13	51	82	20	14
C	34	8	24	10	11	11	71	11
G	3	68	69	2	32	2	4	66
T	52	19	4	75	6	5	5	9

Figure 2.1 The energy matrix, derived probabilistic models and corresponding logos of a typical simulation. (a) The energy logo, with the average energy for each position set to 0. (b) The energy matrix generated from simulation. (c) The information logo when $\mu = -3$. (d) The probabilistic model derived from all binding sites when $\mu = -3$. Matrix elements are base frequencies. Each column sums up to 100. (e) The information logo when $\mu = 3$. (f) The probabilistic model derived from all binding sites when $\mu = 3$. Each column sums up to 100.

PMs (Figure 2.1d and 2.1f), at high protein concentration the base probabilities tend to move toward 0.25; the high probability bases decrease in probability and the low probability bases increase. More importantly the magnitude of the change in probability varies from position to position because each column is normalized independently. As a result, the rank order of the

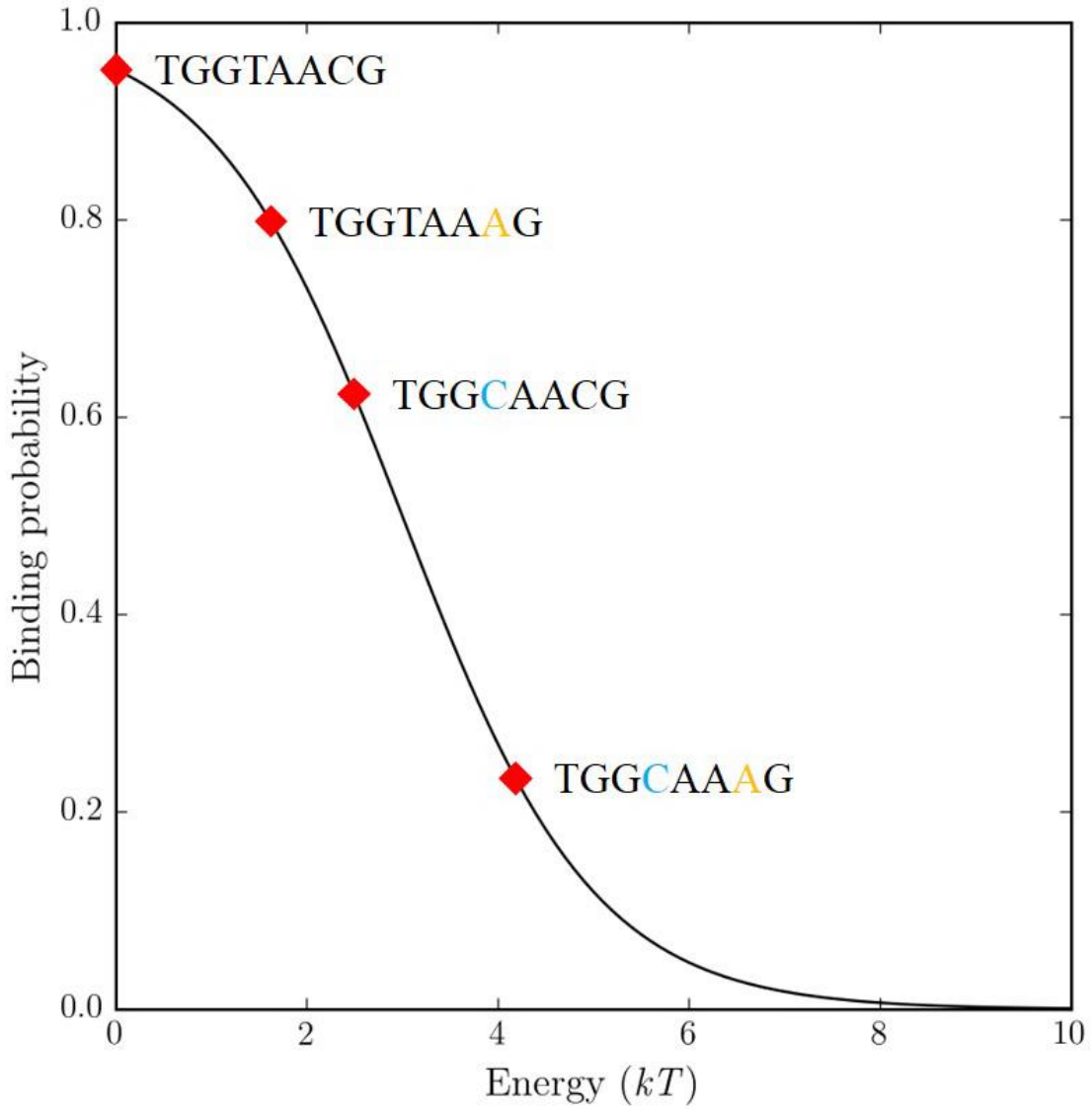


Figure 2.2 The non-linear relationship between binding energy and probability. The C to A mutation that occurs at the same position but in two different sequence contexts causes dramatically different changes in the binding probability of the whole sequence.

probabilities of different binding sites may change, even though the rank order of the probabilities for the four bases remains unchanged at each position. Another consequence of the non-linear relationship between binding affinity and probability is that pairs (and higher order combinations) of positions have non-independent effects on binding probability, even though their contributions to affinity are completely independent. We show this with one example in Figure 2.2 based on the protein with energy matrix shown in Figure 2.1 at $\mu = 3$. When the preferred binding site, TGGTAACG with a true binding probability of 0.95, is mutated to TGGTAAAG, the true binding probability decreases to 0.79, about a 17% decrease in binding probability. If the same C to A mutation occurs in another sequence, TGG*C*AACG to TGG*C*AAAG, the true binding probability decreases from 0.62 to 0.23, a 63% decrease. This apparent non-independence, where the effect of the mutation on site probability varies depending its context, cannot be captured by the PMs, even though the change in binding affinity (1.69 *kT*, Figure 2.1b) is completely independent of context.

The rank correlation (the square of the Spearman's rank correlation coefficient) between the predicted and true all sequence distributions depends on the protein concentration and how the PM is computed. Table 2.1 shows the mean values and standard deviations of r^2 for 100 simulations of 8 bp binding sites with μ of −3, 0 and 3 (which correspond to the preferred sequence being bound at 0.05, 0.5 and 0.95 probability, respectively). The rank correlation is based on PMs generated from the full distribution of binding data and from just the top 1% of sites, either weighted or unweighted. At $\mu = -3$ there is a nearly perfect fit to the true ranking when the PM is derived from the entire distribution. However, when it is derived from the weighted top 1% of sites, the ranking is slightly less accurate (0.994). In both of those cases the PM provides a very good approximation to the true ranking of binding sites. If the unweighted

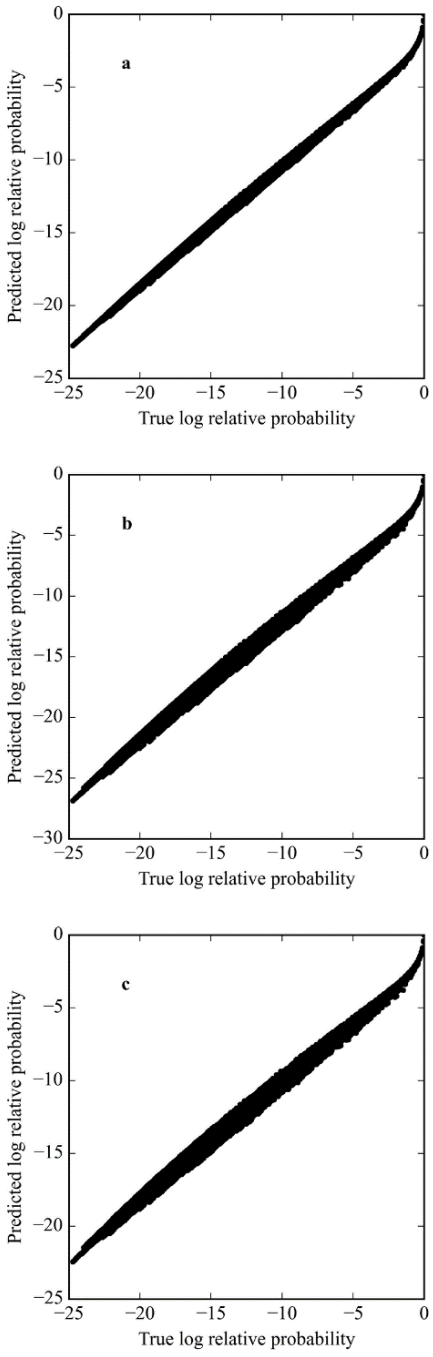


Figure 2.3 The correlation between the predicted and true all sequence distributions. (a) The correlation between the true distribution (in logarithm with highest affinity site set to 0) and that predicted by the PM generated from the weighted all binding sites. (b) The correlation between the true distribution (in logarithm) and that predicted by the PM generated from the weighted top 1% binding sites. (c) The correlation between the true distribution (in logarithm) and that predicted by the PM generated from the unweighted top 1% binding sites.

top 1% sites are used to make the PM, the fit to the true ranking is 0.984 and that is true regardless of the value of μ because the top 1% of sites is the same for different μ and their probabilities are ignored. When $\mu = 0$, the results are very similar. When $\mu = 3$, the rank correlation drops to 0.988 when the weighted top 1% of sites are used to obtain the PM. Figure 2.3 plots the logarithms of the predicted and true relative binding probabilities for the simulated protein in Figure 2.1 and $\mu = 3$. In each case the overall fit is quite good but the width of the plots indicates some degree of mis-ranking of the binding sites.

Table 2.1 The rank correlation between the predicted and true all sequence distributions for probabilistic models in the absence of noise.

	Mean correlations and standard deviations		
Probabilistic model generation method	$\mu = -3$	$\mu = 0$	$\mu = 3$
All binding sites, weighted	1.000 (0.000)	1.000 (0.000)	0.998 (0.001)
Top 1% binding sites, weighted	0.994 (0.010)	0.993 (0.011)	0.988 (0.012)
Top 1% binding sites, unweighted	0.984 (0.014)	0.984 (0.014)	0.984 (0.014)

While the overall rankings are quite good for the estimated PMs, it is the highest affinity sites that are of primary interest. In fact, all DNA-binding proteins exhibit a non-specific binding affinity [57] such that there is a minimum binding affinity below which the sequence no longer matters. In addition, functional regulatory sites must have sufficient occupancy to fulfill their roles, so only sites within some range of the optimum are likely to be functional. Figure 2.4 shows a subset of the data points in Figure 2.3, including only the top 1% of sites. The plots all show substantial mis-ranking of sites. Table 2.2 shows the rank correlations based on the same PMs in Table 2.1, but now focusing on the probabilities of the top 1% binding sites. The values

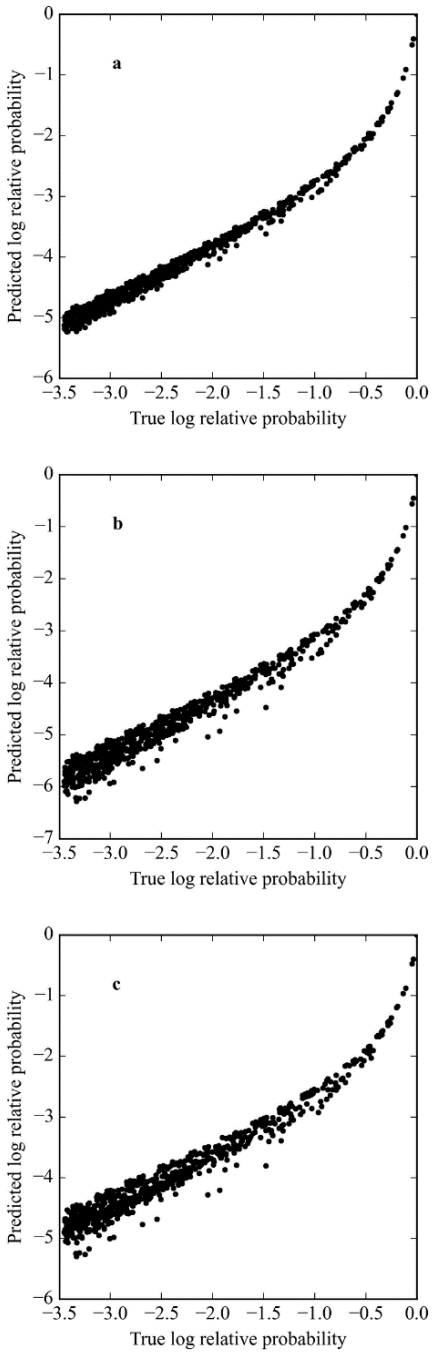


Figure 2.4 The correlation between the predicted and true top 1% sequence distributions. (a) The correlation between the true distribution (in logarithm with highest affinity site set to 0) and that predicted by the PM generated from the weighted all binding sites. (b) The correlation between the true distribution (in logarithm) and that predicted by the PM generated from the weighted top 1% binding sites. (c) The correlation between the true distribution (in logarithm) and that predicted by the PM generated from the unweighted top 1% binding sites.

in Table 2.2 are all lower than those in Table 2.1, indicating that the accuracy decreases when the PM is used to predict the highest affinity sites. In fact, the true top 1% does not contain precisely the same set of sequences as the predicted top 1% of sites. For $\mu = 3$ the rank correlation drops to 0.970 even when the entire distribution is used to generate the PM. When the top 1% of sites are used to generate the PM, weighted by their probabilities, the rank correlations drop substantially for all values of μ, but especially for $\mu = 3$ where it is only 0.876. If the unweighted top 1% are used, the rank correlation drops to 0.840 for all values of μ. The results in Tables 2.1 and 2.2 show that the quality of PMs, their ability to correctly rank binding sites, varies widely depending on both the protein concentration and the set of binding sites used to derive the PM. The effect of the protein concentration is most evident at high values of μ where the non-linearity of equation (2.6) is largest and non-independence of the position probabilities is most pronounced. The effect of site sampling is due to the sensitivity of the PM to the exact set of example sites used.

Table 2.2 The rank correlation between the predicted and true top 1% sequence distributions for probabilistic models in the absence of noise.

	Mean correlations and standard deviations		
Probabilistic model generation method	$\mu = -3$	$\mu = 0$	$\mu = 3$
All binding sites, weighted	1.000 (0.000)	0.995 (0.002)	0.970 (0.010)
Top 1% binding sites, weighted	0.956 (0.032)	0.930 (0.092)	0.876 (0.063)
Top 1% binding sites, unweighted	0.840 (0.075)	0.840 (0.075)	0.840 (0.075)

2.3.2 Results in the Presence of Measurement Noise

When we add measurement noise ($\sigma = 0.5$) to the simulation, the fitted biophysical model derived from the noisy observation $P^*(S_i \mid B)$ may not reproduce the underlying true biophysical

model, thus requiring evaluation of their predictions in addition to the estimated PMs. That amount of noise is larger than what can be obtained with methods such as Spec-seq, where we typically get standard deviations of about 0.2 *kT*, at least for the high affinity sequences [68, 101]. We used the same criterion described in the previous subsection to evaluate the estimated models under noise, which is the rank correlation between the predicted distribution $\hat{P}(S_i \mid B)$ and the true distribution (clean probability) $P(S_i \mid B)$, focusing on either all the sequences or only the top 1%. The results are summarized in the tables below, from which three main conclusions can be drawn. First, the fitted biophysical models almost always achieve higher scores than the corresponding PMs, especially when the predicted top 1% sequence distributions are checked against the true distributions. Also, they can generally recover the true parameters of the underlying model, as indicated by the nearly perfect scores in Table 2.5 and 2.6, especially when they were trained on all the sequences. Second, comparison of Table 2.1 and 2.3 shows that when the all sequence distributions are compared, the accuracy of the PMs derived from the noisy observation $P^*(S_i \mid B)$ is only slightly lower than those derived from the clean observation $P(S_i \mid B)$, regardless of how the PMs are generated. Third, comparison of Table 2.2 and 2.4 shows that when the top 1% sequence distributions are compared, the accuracy of the PMs derived from the noisy observation $P^*(S_i \mid B)$ is lower than those derived from the clean observation $P(S_i \mid B)$. This is especially true for the case $\mu = -3$ or $\mu = 0$, when the PM should be a good approximation to the underlying biophysical model. This may be because when noise level $\sigma = 0.5$, the measurement error on binding energy, rather than the protein concentration, becomes the dominant factor in determining the accuracy of the models.

Table 2.3 The rank correlation between the predicted and true all sequence distributions for probabilistic models in the presence of noise.

	Mean correlations and standard deviations		
Probabilistic model generation method	$\mu = -3$	$\mu = 0$	$\mu = 3$
All binding sites, weighted	0.997 (0.002)	0.997 (0.001)	0.997 (0.001)
Top 1% binding sites, weighted	0.987 (0.030)	0.988 (0.019)	0.987 (0.022)
Top 1% binding sites, unweighted	0.983 (0.014)	0.984 (0.015)	0.986 (0.008)

Table 2.4 The rank correlation between the predicted and true top 1% sequence distributions for probabilistic models in the presence of noise.

	Mean correlations and standard deviations		
Probabilistic model generation method	$\mu = -3$	$\mu = 0$	$\mu = 3$
All binding sites, weighted	0.952 (0.022)	0.959 (0.016)	0.959 (0.013)
Top 1% binding sites, weighted	0.886 (0.090)	0.883 (0.035)	0.872 (0.043)
Top 1% binding sites, unweighted	0.839 (0.069)	0.842 (0.066)	0.846 (0.052)

Table 2.5 The rank correlation between the predicted and true all sequence distributions for biophysical models in the presence of noise.

	Mean correlations and standard deviations		
Biophysical model generation method	$\mu = -3$	$\mu = 0$	$\mu = 3$
All binding sites, weighted	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Top 1% binding sites, weighted	0.992 (0.034)	0.995 (0.022)	0.995 (0.023)

Table 2.6 The rank correlation between the predicted and true top 1% sequence distributions for biophysical models in the presence of noise.

	Mean correlations and standard deviations		
Biophysical model generation method	$\mu = -3$	$\mu = 0$	$\mu = 3$
All binding sites, weighted	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Top 1% binding sites, weighted	0.964 (0.092)	0.972 (0.009)	0.968 (0.012)

2.4 Discussion

Probabilistic models of protein–DNA interactions are commonly used because they are easy to obtain and they provide an intuitive representation of specificity. However, they do not provide the information usually desired, the probability that a specific sequence is bound, $P(B \mid S_i)$, but rather an approximation to the probability of observing a specific sequence given a binding site, $\tilde{P}(S_i \mid B)$. From that one can obtain a predicted rank order of all possible binding sites and, if one assumes a specific probability, or occupancy, for the preferred sequence, the predicted probabilities for all other sequences. To obtain binding probabilities from the biophysical model one needs to know the chemical potential, but just as with the probabilistic model if one assumes the probability, or occupancy, of the preferred sequence, then the probabilities of all other sequences can be obtained from the model. Since both models really return the same information, a predicted ranked list of binding sites and relative binding probabilities, they should be judged on the accuracy of those predictions and the ease of obtaining the model parameters.

The accuracy of PMs is limited by availability of binding site affinity data. When a PM is based on the entire probability distribution of binding sites it is a good approximation overall, even at high μ. However, it does have discrepancies that include mis-ordering of the ranks of binding sites as well as the appearance of non-independence between positions that are in fact independent. These effects are due to the intrinsic lack of proportionality between binding probability and binding affinity that is most problematic at high protein concentrations. More severe defects occur due to incomplete information about the binding probability distribution. Obtaining the full distribution of binding probabilities requires *in vitro* experiments, such as protein binding microarrays, HT-SELEX or other high-throughput methods [5, 24, 25, 34, 62,

66, 73, 82, 89, 102, 103], but many algorithms utilize only the highest affinity binding sites. PMs can be derived from *in vivo* data of binding site locations, and have the advantage of being easily derived from such data using many motif discovery algorithms [27, 45-47, 49, 104, 105]. But in those cases, the PM is derived from only a fraction of the binding sites. Functional regulatory sites will be among the high affinity sequences and in ChIP-seq experiments the peaks will also tend to contain the highest affinity sites. And if the sample size is small, those sites are not even weighted by their binding probabilities. In addition, confounding factors occurring *in vivo*, such as competition and cooperativity with other proteins, lead to incomplete information about the probability distribution and that causes further inaccuracies in the PMs.

Good binding models are still important after the advent of high-throughput methods and their parameters can be readily determined by using appropriate algorithms. Binding affinities to small numbers of sequences can be obtained with arbitrarily high accuracy using a variety of experimental techniques. If the additivity (positional independence) assumption is valid, the relative affinities, compared to the preferred sequence, of only the $3m$ single nucleotide variants are needed for the full energy model. Of course, additivity is unlikely to be completely accurate, but there are still only $3m+9(m-1)$ single variants plus double variants at adjacent positions, where the non-additivity is likely to be most prevalent. But multiple high-throughput methods are now available that provide quantitative binding data from which accurate energy models can be obtained by using appropriate algorithms [34, 39, 59, 61-69, 102]. From sufficiently abundant and accurate quantitative binding data one can even skip the modeling and just use the list of relative binding energies to all possible sites (or at least the highest affinity sites that are likely to function as regulatory sites), avoiding approximations entirely (to the degree allowed by the measurement accuracy). However, models are still useful because they provide a compact

representation of specificity, usefully visualized with logos [27, 30, 59]. It is also important to have good specificity models obtained from *in vitro* binding experiments to compare to data obtained *in vivo*. This allows one to identify cases where interacting TFs alter the specificity of individual TFs, which one can only infer by having good models for each TF alone [85, 106, 107].

We conclude by pointing out that when accurate energy models are available for DNA binding specificity there is no advantage to using probabilistic models, and in fact they can be misleading and provide inaccurate predictions. There are now good high-throughput methods for measuring relative binding affinities to very large collections of sites and good algorithms for determining accurate energy models. We propose that such models become the standard approach for representing specificity and predicting binding sites *in vivo*.

Chapter 3: Estimation of Binding Motifs Using HT-SELEX Data[1]

Characterizing the binding specificities of transcription factors (TFs) is crucial to the study of gene expression regulation. Recently developed high-throughput experimental methods, including protein binding microarrays (PBM) and high-throughput SELEX (HT-SELEX), have enabled rapid measurements of the specificities for hundreds of TFs. However, few studies have developed efficient algorithms for estimating binding motifs based on HT-SELEX data. Also the simple method of constructing a position frequency matrix (PFM) by comparing the frequency of the preferred sequence with single-nucleotide variants has the risk of generating motifs with higher information content than the true binding specificity. We developed an algorithm called BEESEM that builds on a comprehensive biophysical model of protein–DNA interactions, which is trained using the expectation maximization method. BEESEM is capable of selecting the optimal motif length and calculating the confidence intervals of estimated parameters. By comparing BEESEM with the published motifs estimated using the same HT-SELEX data, we demonstrate that BEESEM provides significant improvements. We also evaluate several motif discovery algorithms on independent PBM and ChIP-seq data. BEESEM provides significantly better fits to *in vitro* data, but its performance is similar to some other methods on *in vivo* data under the criterion of the area under the receiver operating characteristic curve (AUROC). This highlights the limitations of the purely rank-based AUROC criterion. Using quantitative binding data to assess models, however, demonstrates that BEESEM improves on prior models.

[1] Parts of this chapter are taken from Ruan, S., Swamidass, S. J. & Stormo, G. D. BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics* (2017).

3.1 Materials

The 2726 HT-SELEX sequencing datasets used in this study were generated by Jolma *et al.* and retrieved from the European Nucleotide Archive (www.ebi.ac.uk/ena) [87]. These datasets were grouped into 547 HT-SELEX experiments, each of which is composed of 4 to 7 SELEX cycles

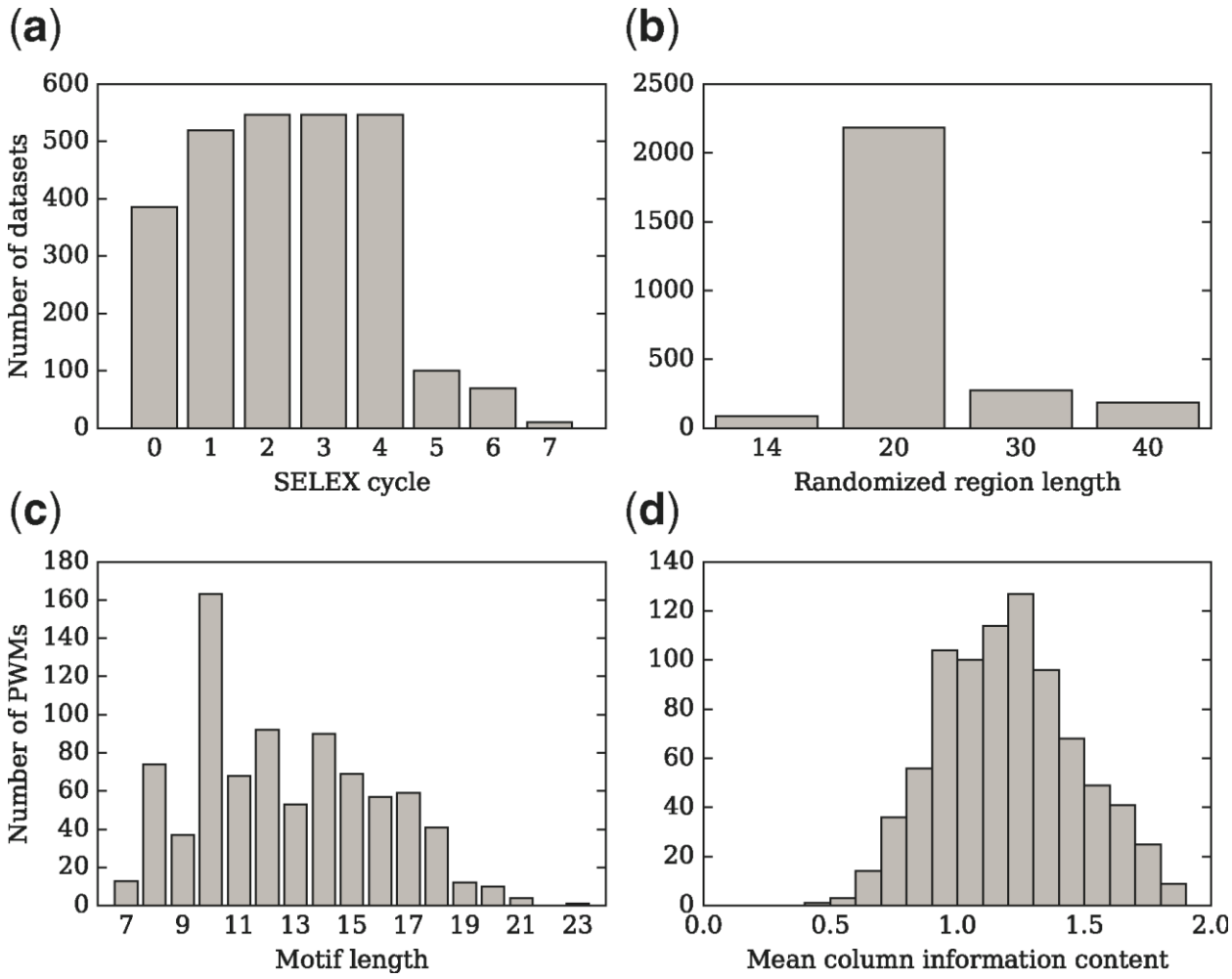


Figure 3.1 The HT-SELEX experiments and the J2013 PFMs. (a) Most of the HT-SELEX experiments have 4 cycles. By convention, the 0th SELEX cycle denotes the initial library of randomly generated DNA probes. Multiple HT-SELEX experiments share the same initial library. 27 sequencing datasets corresponding to the 1st cycle are missing from the database. (b) In 80% of the datasets, the randomized region is 20 bp long. (c) The length of the J2013 PFMs ranges from 7 to 23; the mean length is 12.7 bp. (d) The average mean column information content of the J2013 PFMs is 1.20 bit. The information content is computed based on a uniform background distribution of the four nucleotides.

(Figure 3.1a). A total of 463 distinct TFs were surveyed in these experiments. In a typical experiment, the TF of interest binds to DNA probes that consist of a randomized region and two constant flanking regions. The length of the randomized region varies between experiments (Figure 3.1b). We focus on the experiments in which the randomized region is 20 bp because they form the largest group. A total of 843 PFMs were also published by Jolma *et al.* [87], which we refer to as the J2013 PFMs. To assess their specificity, we calculated the mean column information content (MCIC), which is defined as the average information content of all the columns in a matrix. Figure 3.1d shows that the average MCIC of the J2013 PFMs is 1.20 bit.

3.2 Generalized Biophysical Model

The biophysical model underlying BEESEM generalizes the biophysical model described in the preceding chapter and allows simultaneous estimation of the binding motif and the locations of binding sites. In the original biophysical model, Equation (2.6) states that the probability of finding a sequence S_i among all the DNA molecules bound by the TF is

$$P(S_i \mid B) = \frac{P(S_i)}{P(B)} \frac{1}{1+e^{E_i-\mu}}. \tag{3.1}$$

$P(S_i)$ is the proportion of sequence S_i before TF binding and $P(B)$ is the overall probability of a DNA molecule being bound by the TF. Equation (3.1) assumes that the DNA sequence length *l* is the same as the motif length *m*. However, *l* is generally larger than *m* and the binding site may be located anywhere on the sequence. As a result, each protein–DNA complex has $2(l-m+1)$ possible configurations, if we account for both orientations. To keep track of the different complexes and their configurations, we use $T \cdot S_i^k$ to denote the *k*th configuration of the protein–DNA complex of S_i, where *T* represents the TF, $k = 1, 2, \cdots, 2(l-m+1)$. Under this notation, S_i^k

can also be interpreted as the binding sequence in $T \cdot S_i^k$ and is therefore associated with a binding energy (denoted by $E(S_i^k)$). $E(S_i^k)$ is entirely determined by the DNA sequence of S_i^k, regardless of its location on sequence S_i. Thus $E(S_i^k)$ can also be written as $E(s_j)$, if $S_i^k = s_j$, where s_j represents the DNA sequence of a specific *m*-mer. To generalize Equation (3.1), we rearrange it and replace the symbol S_i with s_j, namely

$$\frac{P(s_j \mid B)}{\bar{P}(s_j)} = \frac{1}{P(B)} \frac{1}{1+e^{E(s_j)-\mu}}. \tag{3.2}$$

$\bar{P}(s_j)$ represents the effective proportion of s_j prior to TF binding, where 'effective' means the sequence count of s_j is discounted because we assume that two or more proteins cannot bind to the same sequence at the same time. For example, let us consider the scenario that the same *m*-mer s_1 occurs twice on a DNA sequence. Since at most one of them can be bound by the TF due to physical hindrance, the effective count of s_1 on this DNA sequence is only one. To simplify the LHS of Equation (3.2), we use R_j to denote the ratio $P(s_j \mid B) / \bar{P}(s_j)$. In the following section, we will show how to use the EM algorithm to compute R_j from data. By fitting Equation (3.2) to the computed R_j, we can estimate the unknown parameters in our model (collectively represented by a vector θ), which include the PWM (which contains the energy of each base relative to the preferred base in units of *kT*), the chemical potential and some auxiliary parameters (see Appendix A for a detailed description of the model and its parameters).

3.3 BEESEM Algorithm

3.3.1 Parameter Estimation

BEESEM uses the EM algorithm to iteratively find both the optimal PWM and the most likely binding position on each sequence read. The EM algorithm consists of multiple rounds and each round has two steps: an expectation step (E step) and a maximization step (M step) [46]. In the E step, we use the current estimate of the PWM (or an initial guess) to calculate the probability distribution of binding sites on each sequence. Specifically, we assume the probability of an *m*-mer subsequence being the binding site is proportional to its affinity score predicted by the current PWM. Also we require that the probabilities of all the *m*-mers on sequence S_i sum to $P(1|S_i)$. $P(1|S_i)$ represents the probability that S_i is bound by the TF and therefore contains a binding site. By computing the $P(1|S_i)$ for each sequence in the after-binding library we can exclude those free-rider sequences which are carried to the next cycle through non-specific binding [89], and it can be computed using the Bayes' theorem:

$$P(1|S_i) = \frac{P(S_i|1)P(1)}{P(S_i|1)P(1) + P(S_i|0)P(0)}. \tag{3.3}$$

In the above equation, $P(S_i|1)$ is the proportion of S_i among bound sequence reads, $P(S_i|0)$ is the proportion of S_i among unbound sequence reads, $P(1)$ is the *ex ante* probability that a sequence read is bound by the TF, and $P(0)$ is the probability of the complementary event, thus $P(0) = 1 - P(1)$. $P(1)$ can be approximated by the mean of the $P(1|S_i)$ values from the previous round. After computing the probability distribution of binding sites on each sequence, we can easily calculate the probability $P(s_j|B)$ in Equation (3.2) and finally the ratio R_j. In the M step, we search for a parameter vector θ that maximizes the probability of observing the R_j

values computed in the E step. R_j is a ratio of two positive numbers, so it likely follows a log-normal distribution. However, for convenience, we assume it can be approximated by a normal distribution in the regime of the data, and define the maximum likelihood estimate of θ by fitting Equation (3.2) with least squares minimization. In other words, the objective function of the *n*th round is

$$\hat{\theta}_n = \arg\min_{\theta} \sum_j \left[R_j^{(n)} - R_j(\theta) \right]^2 , \tag{3.4}$$

where $R_j^{(n)}$ denotes the computed R_j value in the *n*th round and $R_j(\theta)$ represents the RHS of Equation (3.2). Appendix B contains a proof that Equation (3.4) is the correct objective function and therefore the procedure described above is indeed an EM algorithm. The minimization is performed using the L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm [100], which is a quasi-Newton hill-climbing method for solving non-linear optimization problems. A major challenge for any hill-climbing method is that the search for a global minimum of the objective function may be attracted to a local minimum. Thus it is important to select a proper starting point (denoted by θ_0), since it largely determines the endpoint of a search. Conventionally θ_0 is randomly generated and multiple searches are attempted. However, our preliminary results indicate that $\hat{\theta}$ may change significantly between successive rounds, if random starting points are used in each round. In order to facilitate the convergence of $\hat{\theta}$, we always use $\hat{\theta}_{n-1}$ to initialize a search in the *n*th round. This modification greatly improves the stability of the $\hat{\theta}$ series. The iterative EM algorithm keeps refining $\hat{\theta}$ until the maximum number of rounds (10 by default) is reached.

3.3.2 Confidence Intervals

BEESEM can calculate the confidence interval for each parameter in the $\hat{\theta}$. It can be proved that $\hat{\theta}$ is both asymptotically unbiased and asymptotically normal [108]. Thus we can assume that $\hat{\theta}$ obeys a multivariate normal distribution $N(\lambda, \mathbf{\Sigma})$, where λ is the mean vector and $\mathbf{\Sigma}$ represents the covariance matrix. Since Equation (3.4) gives an (asymptotically) unbiased estimate of λ, the only unknown parameter of the multivariate normal distribution is $\mathbf{\Sigma}$. The covariance matrix is related to the Hessian matrix of the objective function in Equation (3.4). To proceed, we assume that the measurement errors of R_j are independent and obey the same normal distribution $N(0, \sigma^2)$. σ^2 can be estimated using the sum of squared residuals, which equals to the value of the objective function at $\hat{\theta}$. If $R_j(\theta)$, namely the RHS of Equation (3.4), is linear with respect to θ, it can be shown that [108]

$$\mathbf{\Sigma} = \sigma^2 \mathbf{H}^{-1}. \tag{3.5}$$

In this equation, **H** represents the Hessian matrix of the objective function. Although $R_j(\theta)$ is actually a non-linear function with respect to θ, we can assume that the objective function is well approximated by a parabola in the vicinity of a minimum (namely the Laplacian approximation). This approximation enables us to compute the confidence intervals using second order derivatives, which means Equation (3.5) still holds.

3.3.3 Motif Length Selection

BEESEM can infer the optimal motif length if a user-defined value is not supplied. To optimize the motif length, BEESEM first generates a series of energy PWMs, ranging from 7 to 10 bp. Then the candidate PWM that contains the *k*-mer submatrix (*k* defaults to 5) with the highest MCIC is identified and it is called the core PWM. Next, BEESEM collects any candidate PWM

that meets the following three criteria: 1) it is longer than the core PWM, 2) one of its *k*-mer submatrices is highly correlated with any submatrix in the core PWM, and 3) it does not contain a non-informative position at either end. A non-informative position is a PWM column with low information content (less than 0.25 bit by default). Finally, BEESEM chooses the longest one from the collected PWMs (including the core PWM).

3.3.4 Initialization of BEESEM with Seed Sequences

We generated a total of 660 PWMs by applying BEESEM to the aforementioned HT-SELEX datasets. Each BEESEM PWM comes in two types: seeded and unseeded. BEESEM takes an optional seed sequence of length *m* as input. The seed sequence is used to generate a seed matrix, which serves as the initial guess of the PWM in the first BEESEM round. To generate the seed matrix, BEESEM counts the number of occurrences of the seed sequence, as well as all the sequences with one mismatch, in the sequencing dataset of the highest SELEX cycle. The highest SELEX cycle is used because we expect high-affinity sequences to be enriched after multiple SELEX cycles. Next, BEESEM uses the computed sequence counts to build a position frequency matrix (PFM), with a default pseudocount of 1. Finally, the PFM is rescaled so that its mean column information content equals 0.5 bit. The rescaling generally reduces the information content of the seed matrix, making it a starting point of weak bias. Ideally the seed sequence should be close to the true consensus sequence, in which case the convergence of BEESEM will be greatly accelerated. For example, when generating the seeded BEESEM PWMs, we truncated the consensus sequences of the J2013 PFMs and used them as seed sequences. If a seed sequence is not specified by the user, however, BEESEM will automatically generate one using the input HT-SELEX sequencing datasets. For example, the seed sequences of the unseeded BEESEM PWMs were all automatically generated, in contrast to the seeded PWMs. Specifically, BEESEM

chooses the most abundant *m*-mer in the highest SELEX cycle that contains all 4 nucleotides. We require the seed sequence to contain all 4 bases in order to filter out low-complexity sequences that may have large sequence counts.

3.3.5 Evaluation of Binding Models

We evaluated the BEESEM PWMs on HT-SELEX data, as well as PBM and ChIP-seq data from independent experiments. In addition, we compared their performance in the evaluation tests with 5 other binding models, including the J2013 PFMs, HTS-IBIS, DiMO, BEEML and DeepBind. HTS-IBIS is derived from the RAP algorithm [109] and optimized for HT-SELEX data. The HTS-IBIS PWMs were generated by applying the HTS-IBIS program to the HT-SELEX datasets used in this study. DiMO is based on perceptron learning [99] and aims to maximize the AUROC score of a PFM based on ChIP-seq data. To generate a DiMO PFM, we initialize DiMO with a J2013 PFM and then train it on the ChIP-seq data of the corresponding TF. BEEML is a motif finding algorithm built on a biophysical model of protein–DNA interactions, and the corresponding PWMs were trained on PBM data [67], using a specialized version of BEEML for PBM data (BEEML-PBM). DeepBind uses deep learning to train binding models on *in vitro* and *in vivo* data, including HT-SELEX data [38]. The DeepBind binding models were trained on HT-SELEX data. The BEEML PWMs are not evaluated on ChIP-seq data, because few TFs have both BEEML PWMs and ChIP-seq data. For similar reasons, the DiMO PFMs are not evaluated on PBM data.

In the HT-SELEX evaluation tests, we tested the internal consistency of the HT-SELEX based methods and their abilities to accurately model the results of the HT-SELEX experiments. We developed two methods for evaluating binding models on HT-SELEX data, which are called the ‘consistency’ test and the ‘goodness-of-fit’ test respectively. In both methods, the assessment is

based on the difference between two computed sequence distributions, which should be the same under an accurate specificity model. The difference can be measured in two ways: the square of the Pearson's r (denoted by r^2) and the symmetric divergence (denoted by D_{sym}). The symmetric divergence, with a minimum value of 0 that indicates a perfect fit, is a symmetrized version of the Kullback–Leibler divergence [110]. In the consistency test, we assess the difference between two predicted binding site distributions, namely two predictions of the $P(s_j \mid B)$ in Equation (3.2). The two distributions are computed with different approaches: one is calculated using the PWM and the sequencing dataset before TF binding, and the other is calculated using the PWM and the dataset after TF binding. Since both approaches require a presumptive PWM, the resulting distributions are both theoretical predictions. Notwithstanding, we expect the two distributions to be similar, if the binding model is consistent. In the goodness-of-fit test, we assess the difference between two after-binding (overall) *m*-mer distributions. One distribution is calculated directly using the sequencing dataset after TF binding, thus representing the empirical distribution. The other is theoretical and can be calculated using the PWM and the dataset before TF binding. If the binding model is accurate, the predicted distribution should agree with the empirical one.

In the PBM evaluation tests, we tested the ability of the HT-SELEX based methods to predict PBM data. PBMs are an independent *in vitro* experiment, so this is an important external validation. In the PBM test, we evaluate the ability of a binding model to predict the *in vitro* affinities of PBM probes, as measured by their fluorescence intensities. The PBM datasets generated by multiple studies were retrieved from the CIS-BP database (cisbp.ccbr.utoronto.ca) [23]. These datasets include both mouse and human TFs. To evaluate a binding model, we use it to calculate the average affinity score of all the *m*-mers on each PBM probe (including the

reverse complement), where *m* represents the motif length and depends on the binding model. The correlation between the predicted scores and the probe intensities, as measured by the square of the Pearson's r, is used for evaluating the binding models.

In the ChIP-seq evaluation tests, we assessed how well the BEESEM PWMs, which are trained on HT-SELEX data, could identify *in vivo* TF binding sites. Here, we use the ChIP-seq data from the ENCODE project database (www.encodeproject.org) [111] and the ROC curve to quantify the ability of binding models to identify ChIP-seq peaks among random DNA sequences [89]. To perform the evaluation test, we first rank all the peaks in a ChIP-seq dataset based on their enrichment scores. Then we retrieve the center sequences of the top 500 peaks, which are all 250 bp long, and treat them as the positive sequence set (as in [89]). To generate the negative set, we collect the DNA sequence 200 bp downstream from each positive sequence. Next, we assign an affinity score to each sequence. The affinity score of a sequence is defined as the highest affinity of its constituent *m*-mers (including the reverse complement), where *m* is the motif length and depends on the binding model. Finally, we rank all the sequences based on their affinity scores, plot the ROC curve, and compute the AUROC score.

3.4 Results

3.4.1 Characterization of BEESEM PWMs

Reproducibility

The reproducibility of an BEESEM PWM can be measured by its standard deviation or the confidence intervals of its elements. In this study, the standard deviation of a PWM is defined as the mean of the standard deviations of its elements, which can be individually estimated using the covariance matrix of $\hat{\theta}$. Although the confidence interval is a more natural way to quantify

the uncertainty in an estimate, we focus on the standard deviation for two reasons. First, the two measures are equivalent if the estimate obeys a normal distribution, as in our case. Second, it is difficult to summarize the confidence intervals of multiple estimates, while it is straightforward to average their standard deviations. The average standard deviations of seeded and unseeded BEESEM PWMs are 0.014 and 0.015 respectively. The standard deviations are very small compared with the absolute values of PWM elements, averaging 1.73 for seeded and 1.76 for unseeded PWMs. Thus we consider the estimated binding energies to be highly reproducible.

Information Content

The average MCIC of seeded and unseeded BEESEM PWMs is 0.58 bit and 0.60 bit respectively. They are both smaller than the average MCIC of the corresponding J2013 PFMs, which is 1.19 bit. Thus the BEESEM PWMs are generally less specified than the J2013 PFMs. It was reported that motifs with lower MCIC often fit quantitative binding data better [39]

3.4.2 Evaluation Results

Evaluation on HT-SELEX Data

Table 3.1 shows the evaluation results for BEESEM, BEEML, HTS-IBIS, the J2013 PFMs and DiMO on HT-SELEX data. The BEESEM PWMs on average achieve better scores than the other algorithms in both the consistency and the goodness-of-fit tests, by either the r^2 or the D_{sym} criterion. Based on the two-sided T-test, the difference in the consistency r^2 between the unseeded BEESEM PWMs and BEEML ($\text{p-value} = 1.8\times10^{-5}$), HTS-IBIS ($\text{p-value} = 1.7\times10^{-47}$), J2013 ($\text{p-value} = 6.6\times10^{-48}$), DiMO ($\text{p-value} = 3.0\times10^{-29}$) is significant. The difference in the goodness-of-fit r^2 between the unseeded BEESEM PWMs and HTS-IBIS ($\text{p-value} = 1.3\times10^{-15}$

), J2013 (p-value $= 2.0 \times 10^{-19}$), DiMO (p-value $= 4.7 \times 10^{-12}$) is significant, but not for BEEML (p-value $= 0.089$). It is noteworthy that there is only a small difference in performance between the two BEESEM PWM types (the p-value for the consistency and goodness-of-fit r^2 is 0.88 and 0.77 respectively). The DiMO PFMs significantly lag behind the other algorithms presumably because DiMO overfits the characteristics of the ChIP-seq data that it was trained on. Column-wise comparisons show that the r^2 of a goodness-of-fit test is generally lower than the corresponding consistency test, which is consistent with the goodness-of-fit test being more stringent, although the D_{sym} is better on the goodness-of-fit test.

Table 3.1 The results of the HT-SELEX evaluation tests. There are fewer DiMO or BEEML PWMs because only a subset of the TFs have ChIP-seq or PBM data. All the tests are performed using the 2nd SELEX cycle as the prior and the 3rd cycle as the posterior. All the PWMs are trimmed to 8 bp in order for direct comparison. The best score in each column is boldfaced. Scores that are not significantly different from the best scores are italicized ($\text{p-value} \geq 0.05$). DeepBind is excluded from the HT-SELEX test because the output of its models cannot be interpreted as simple binding probabilities.

Algorithm	Sample size	Consistency r^2 (s.d.)	Consistency D_{sym} (s.d.)	Goodness-of-fit r^2 (s.d.)	Goodness-of-fit D_{sym} (s.d.)
BEESEM seeded	660	**0.79 (0.26)**	**0.46 (0.48)**	**0.47 (0.26)**	**0.40 (0.31)**
BEESEM unseeded	660	*0.79 (0.27)*	*0.49 (0.50)*	*0.46 (0.26)*	*0.43 (0.35)*
BEEML	76	0.65 (0.28)	0.90 (0.80)	*0.41 (0.29)*	0.50 (0.32)
HTS-IBIS	660	0.55 (0.30)	2.60 (1.26)	0.35 (0.22)	1.23 (0.52)
J2013	660	0.53 (0.35)	3.86 (2.49)	0.33 (0.26)	1.63 (0.93)
DiMO	73	0.39 (0.33)	28.3 (23.1)	0.24 (0.22)	5.52 (3.14)

Evaluation on PBM Data

The mean r^2 achieved by each motif finding algorithm in the PBM evaluation tests is shown in Figure 3.2a. In this comparison, the BEEML PWMs are a positive control because they were trained on PBM data. The real assessment is how the HT-SELEX based methods (BEESEM, J2013, DeepBind and HTS-IBIS) can predict the PBM data. Because the PBM test is an independent validation for the HT-SELEX based methods, their r^2 scores in Figure 3.2a are generally lower than the corresponding goodness-of-fit r^2 in Table 3.1. Compared with the other methods, the seeded and unseeded BEESEM PWMs are ranked first and second respectively, with a mean r^2 of 0.27 and 0.24, and approach the performance of the positive control ($r^2 = 0.44$). Based on the two-sided T-test, the difference between the seeded and unseeded BEESEM PWMs is not significant ($\text{p-value} = 0.42$). The remaining three algorithms (J2013: 0.14, HTS-IBIS: 0.08 and DeepBind: 0.08) achieve much lower r^2, compared with BEEML and BEESEM. The difference between the unseeded BEESEM PWMs and J2013 ($\text{p-value} = 0.0014$), DeepBind ($\text{p-value} = 1.5\times10^{-8}$), HTS-IBIS ($\text{p-value} = 1.3\times10^{-8}$) is significant.

Evaluation on ChIP-seq Data

The performance of each motif finding algorithm in the ChIP-seq evaluation tests, as measured by the mean AUROC score, is shown in Figure 3.2b. The results show that the DiMO PFMs achieve the highest average AUROC score (0.84). This is mainly because these PFMs were trained on the same ChIP-seq data used for evaluation, while all the other motifs were trained on HT-SELEX data. In addition DiMO is designed to maximize the AUROC score of a PFM, the same as our evaluation criterion. The seeded BEESEM PWMs, the J2013 PFMs and the

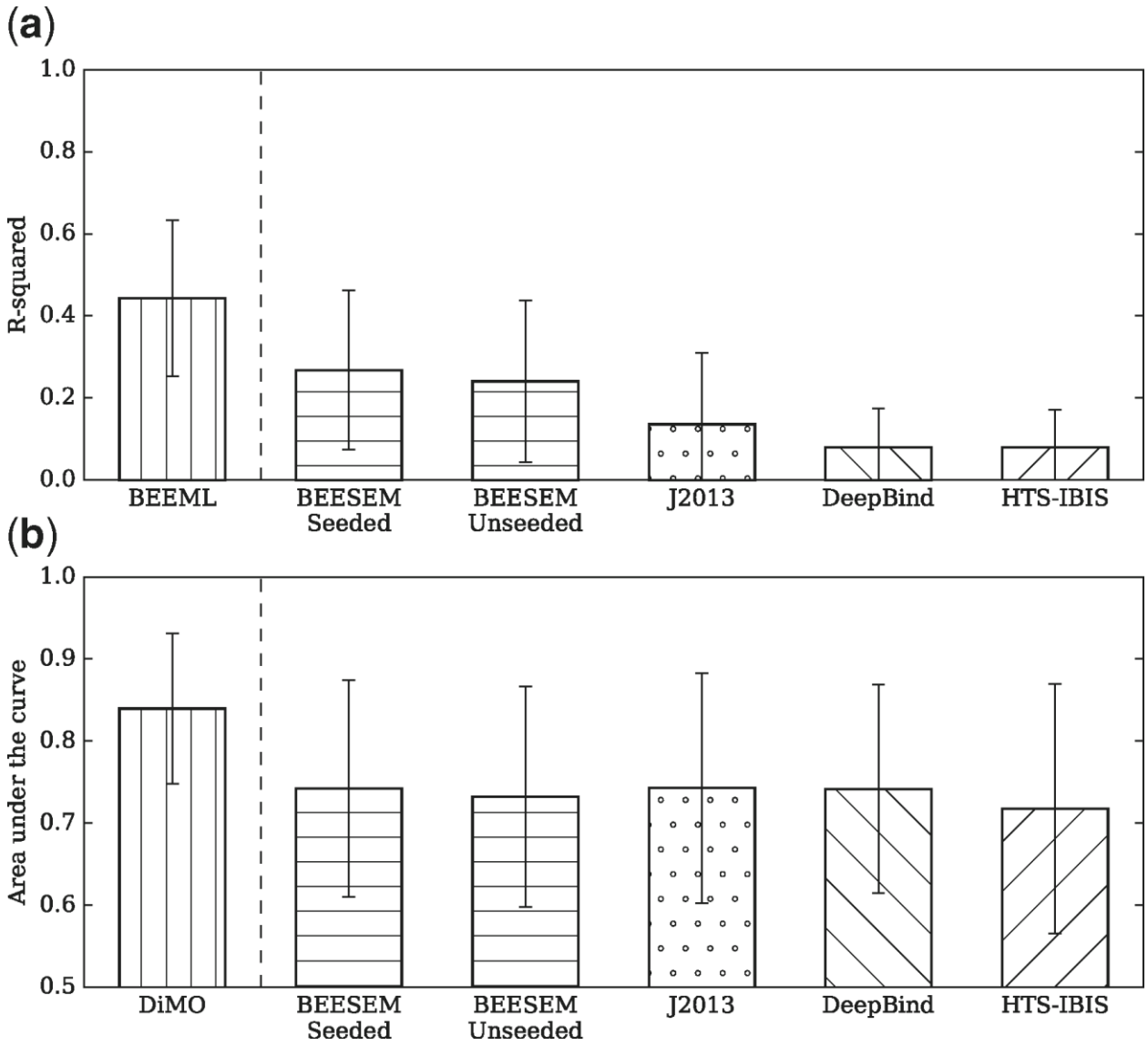


Figure 3.2 The results of the PBM and ChIP-seq evaluation tests. (a) In the PBM tests, the number of binding models tested is 67 for each algorithm. The error bars mark the standard deviation of the scores. The BEEML bar is singled out because the corresponding PWMs were trained on PBM data. For the other binding models trained on HT-SELEX data, the PBM test is an external validation on *in vitro* data. (b) In the ChIP-seq tests, the number of binding models tested is 72 for each algorithm. The error bars mark the standard deviation of the scores. The y axis starts from 0.5, the expected score of a random classifier. The DiMO bar is singled out because the corresponding PWMs were trained on ChIP-seq data. For the other binding models trained on HT-SELEX data, the ChIP-seq test is an external validation on *in vivo* data.

DeepBind models all achieve a mean AUROC score of 0.74, whereas the unseeded BEESEM achieves 0.73 and HTS-IBIS achieves 0.72 (similar to the result reported in [64]). Based on the two-sided T-test, there is no significant difference among the algorithms trained on HT-SELEX data.

3.5 Discussion

BEESEM infers the specificity of TFs based on HT-SELEX data by extending our previous development of BEEML [5]. BEESEM allows the sequences to be much longer than the binding sites, which requires the simultaneous estimation of the binding site locations and the specificity model. This general problem was addressed using the EM algorithm [46], but in that case the data consisted simply of a collection of sequences containing sites without quantitative binding information. Now we include the enrichment of sites by comparing the posterior distribution to the prior (the distribution before the binding site selection). This requires non-linear parameter estimation as part of the EM maximization step. We previously showed that standard motif finding algorithms, that don't account for the ratios of the posterior to the prior distributions and that don't apply non-linear parameter estimation, perform much less well even on relatively simple datasets [5]. We also use the EM algorithm to filter out low-affinity sequence reads that are carried over from the previous SELEX cycle and thus do not contain any high-affinity binding site. In fact, both the seeded and unseeded BEESEM PWMs predict that on average only 60% of the sequence reads actually contain a binding site. We are still employing some approximations, such as the PWM model of specificity [39, 67] and assuming each sequence is bound by at most one protein, but we expect these do not have large effects on the models.

The estimation of the binding site location is important because of the very large libraries from which the selections are made. A library of random 20-mers contains over 10^{12} different

sequences, well beyond the capacity of current sequencing approaches. In fact essentially all of the sequences in the initial pool are unique. Even after four rounds of selection the most abundant 20-mer usually occurs between 10–100 times. In a 20-mer library, every 10-mer occurs in more than 10^7 different contexts so that even if selection was very stringent, most selected and sequenced sites would be unique. By focusing on typical motif lengths, 7–10 bp, and summing over their occurrences in both the prior and posterior distributions, we can obtain models with good fits to the selection data.

Despite the clear improvements in modeling the *in vitro* quantitative binding data (both HT-SELEX and PBM), the BEESEM motifs achieve essentially equal AUROC scores in the ChIP-seq test, which is a valuable independent assessment but also has limitations. Certainly an important use of motifs is in predicting binding events *in vivo*, and also in accounting for changes in expression that are correlated with genetic variants [17]. The AUROC criterion has been cited previously as a method for evaluating the quality of motifs determined from *in vitro* binding data [39, 89]. But ROC curves, and the AUROC measures based on them, have important limitations. First, there are biological considerations. Binding of TFs *in vivo* is confounded by a myriad of other proteins, including nucleosomes, that compete or cooperate in binding. In addition, defining an appropriate negative sequence set is challenging. Ideally these would be genomic regions that are accessible for the TF to bind, but to which it does not under conditions in which it does bind to the positive dataset, but that criterion is seldom used. Second, ROC curves are intrinsically rank-based. The scores assigned to each peak, predicted binding energies in our case or log-probabilities when using probabilistic models, can be multiplied by any constant without altering the ROC curve, and the AUROC measurement. In fact, the J2013 motifs, which perform well in the ChIP-seq test, have very high information content. We had

previously suggested, based on comparisons with different high-throughput methods, that the J2013 algorithm for HT-SELEX was producing over-specified motifs [88]. It was observed for PBM data that algorithms that generated higher information content motifs tended to fit the quantitative binding data less well [39]. Finally, our results using the DiMO method of motif optimization are enlightening. We reasoned that if the AUROC is a good criterion for evaluating a motif on ChIP-seq data, one could use it as the objective function for motif optimization. Using a simple perceptron algorithm we showed that nearly any motif, obtained by a variety of motif discovery algorithms on ChIP-seq data, could be modified to increase its AUROC score [99]. When applying DiMO to the ChIP-seq datasets analyzed in this work, and using the J2013 motifs as starting points, we could in every case obtain a new motif with a higher AUROC score. This was accompanied by an increase in MCIC. However, those DiMO motifs perform significantly worse on the quantitative HT-SELEX data and are further over-specified, highlighting the limitation of using the AUROC as the sole criterion for evaluating motif quality. To gain more insights into the ChIP-seq test, we assess each binding model using the median relative affinity (MRA) of each ChIP-seq dataset, which is defined as the median affinity score of the top 500 peaks divided by their highest affinity score. Figure 3.3 shows that the median MRA for the BEESEM PWMs is about 0.2 while the other binding models predict very low MRAs (their median MRAs are all less than 0.005). It means J2013, DiMO and HTS-IBIS generally predict a >200-fold affinity difference between the peak of the highest affinity and the median peak. Although the true differences in the relative affinity for the different peaks are unknown, it seems unlikely that half of the binding sites in these top scoring peaks would have such low binding affinities. In fact changes in binding affinity of 10-fold are often considered deleterious when inferring causal variants responsible for changes in gene expression [112, 113].

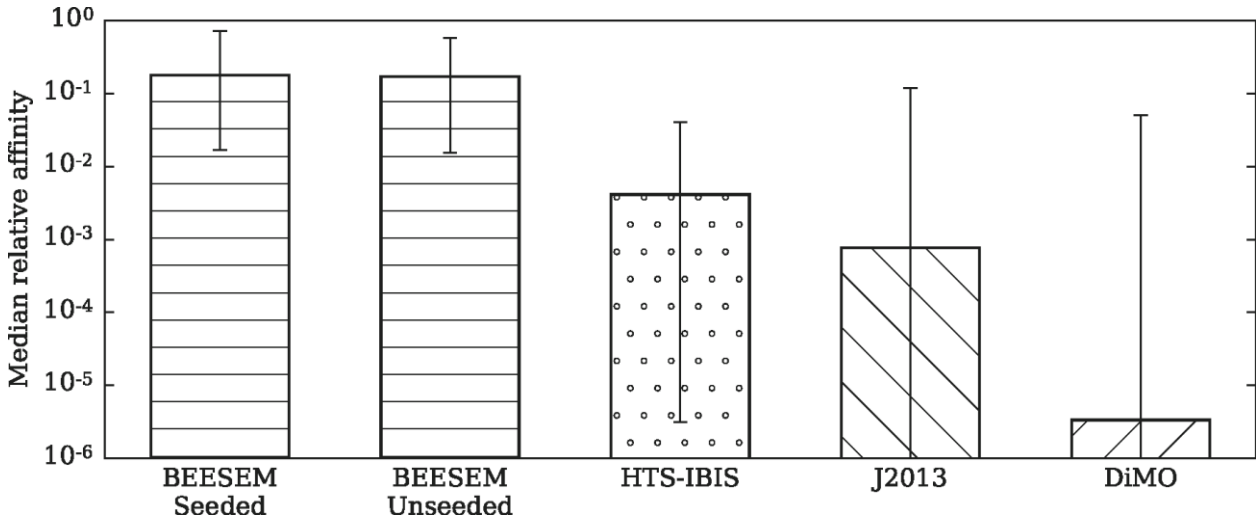


Figure 3.3 The median relative affinities (MRAs) predicted by different binding models. The number of PWMs tested is 73 for each algorithm. The rectangular bars mark the 50th percentile (the median) of the 73 MRAs for each algorithm, and the error bars mark the 5th and 95th percentiles. DeepBind is excluded from the HT-SELEX test because the output of its models cannot be interpreted as simple binding probabilities.

We consider the fit to *in vitro* binding data to be the best criterion for judging the quality of TF binding motifs. The *in vitro* data measure intrinsic binding specificity without the confounding effects that occur *in vivo*. When obtained over a wide range of affinities, either HT-SELEX or PBM data can provide good quantitative models of specificity. The ROC curves are still of value, but should not be the primary means of evaluation. Low AUROC values can point to important information that is missing from the models of intrinsic specificity alone. For example, a highly enriched ChIP-seq peak without a high-affinity binding site may indicate indirect binding, although a low-affinity site that is bound cooperatively with another factor seems more likely. In general, reliable specificity models, which are most easily obtained *in vitro*, are the most useful information for understanding regulatory sites *in vivo* and the alterations in gene regulation that occur in genetic variants. In particular, accurate relative affinity estimates for genetic variants are useful for distinguishing likely deleterious variants from fairly benign ones.

Chapter 4: Contribution of DNA Shape Features to Binding Specificity[1]

With the improvements in X-ray crystallography [90], Cryo-electron microscopy [91] and nuclear magnetic resonance (NMR) spectroscopy [92], an increasing number of three-dimensional structures of protein–DNA complexes and DNA fragments become available. These structural data catalyzed the development of computational methods for predicting DNA shape features [96] and algorithms that use them for motif discovery. In a recent study, Mathelier *et al.* described a motif finding algorithm called DNAshapedTFBS, which trains gradient boosting classifiers to differentiate ChIP-seq peaks from random background sequences, and claimed that taking account of DNA shape features improves TF binding site predictions [40]. In this chapter, we attempt to replicate the results of the study and compare the gradient boosting classifiers with both the probabilistic models [27] and the PWMs generated by the DiMO_PWM program [99]. Our preliminary results indicate that the gradient boosting classifiers, even when they incorporate DNA shape features, are not significantly better than the optimal PWMs generated by DiMO_PWM.

4.1 Materials

4.1.1 JASPAR PFMs

Mathelier *et al.* identified 76 JASPAR PFMs [114] that can be associated with ChIP-seq datasets generated by the ENCODE project [115]. These PFMs were used to locate the best binding site within each positive and negative sequence for training and testing classifiers. In this study, we used the same set of JASPAR PFMs, except for one PFM profile (ID: MA0133.1) that is no

[1] This work is being prepared for publication and will be submitted once the final results are completed.

longer available in the March 2017 JASPAR CORE collection (jaspar.genereg.net/html/DOWNLOAD/JASPAR_CORE/).

4.1.2 ChIP-seq Datasets

We used the same ChIP-seq datasets analyzed by Mathelier *et al.* [40]. 396 uniformly processed human ENCODE ChIP-seq datasets associated with the aforementioned 75 JASPAR PFMs were downloaded from the UCSC Genome Browser (hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/). For each ChIP-seq peak, we retrieved the 100 bp sequence centered on the point-source of the peak from the human genome assembly hg19, which serves as a positive sequence for training and testing TF binding models. For each positive sequence, we also constructed a negative sequence, which is the 100 bp sequence 100 bp downstream from the positive sequence in the human genome. For each ChIP-seq dataset, we constructed 10 training and 10 testing sets, where each training set is 9 times the size of a testing set.

4.1.3 DNA Shape Features

We retrieved the same DNA shape features used by Mathelier *et al.* from GBshape (ftp://rohslab.usc.edu/hg19/) [97]. The features include the helix twist (HelT), the minor groove width (MGW), the propeller twist (ProT), the roll (Roll), and the corresponding second-order shape features. These features were only used for training and testing DNAshapedTFBS models.

4.2 Methods

4.2.1 Motif Discovery Algorithms Evaluated

Table 4.1 lists the motif discovery algorithms evaluated in this study. Out of the 9 algorithms, 7 are based on the DNAshapedTFBS program, which trains a binary classifier using the gradient boosting algorithm [40]. These 7 algorithms differ in two aspects: 1) how the feature vector is

encoded, and 2) whether the feature vector includes DNA shape features. All the DNAshapedTFBS-based algorithms incorporate sequence encodings in feature vectors except for JASPAR + shape. We tested three ways of encoding DNA sequences: 4-bit, 3-bit, and 3-bit plus adjacent dinucleotide (3-bit dinuc) encoding. In the 4-bit encoding, which was used by Mathelier *et al.*, A is encoded as 1000, T as 0100, G as 0010, and C as 0001. In the more compact 3-bit encoding, A is encoded as 100, C as 010, G as 001, and T as 000. In the 3-bit dinuc encoding, the feature vector contains the 3-bit encoding plus 9 additional bits that represent each adjacent dinucleotide in a sequence [116]. The remaining two algorithms are the probabilistic model and the DiMO_PWM program, detailed in Chapter 2 and Chapter 3 respectively. DiMO is based on perceptron learning and finds the optimal PWM by maximizing its AUROC score [99]. The original DiMO program outputs a normalized PFM derived from the optimal PWM, even though it uses a PWM internally for optimization. Because energy PWMs do not have the limitations of probabilistic PFMs (see Chapter 2), we used a modified DiMO program (DiMO_PWM) in this study that outputs the optimal PWM directly.

Table 4.1 Descriptions of the motif discovery algorithms evaluated

Algorithm	Output	Description
DiMO_PWM	Position weight matrix	Modified DiMO program that outputs PWMs instead of PFMs
JASPAR	Position frequency matrix	PFMs from the JASPAR database
DNAshapedTFBS_4bit	Gradient boosting classifier	DNAshapedTFBS with 4-bit encoding
DNAshapedTFBS_3bit	Gradient boosting classifier	DNAshapedTFBS with 3-bit encoding
DNAshapedTFBS_3bit_dinuc	Gradient boosting classifier	DNAshapedTFBS with 3-bit dinuc encoding
JASPAR + shape	Gradient boosting classifier	JASPAR PFM score plus DNA shape features
DNAshapedTFBS_4bit + shape	Gradient boosting classifier	DNAshapedTFBS_4bit plus DNA shape features
DNAshapedTFBS_3bit + shape	Gradient boosting classifier	DNAshapedTFBS_3bit plus DNA shape features
DNAshapedTFBS_3bit_dinuc + shape	Gradient boosting classifier	DNAshapedTFBS_3bit_dinuc plus DNA shape features

4.2.2 Training and Testing Binding Models

The training and testing procedures are based on the methods described by Mathelier *et al.* [40].

Preprocessing

We first use the JASPAR PFM to scan all the positive and negative sequences in both the training and testing set, and identify the best binding site, which has the same length as the PFM, within each sequence. Then we use the DNAshapedTFBS program to extract the DNA sequence of each best site and the corresponding DNA shape features. The sequences of the best sites, instead of the full-length positive and negative sequences, are used in the following steps for training and testing TF binding models.

Training

The training procedure depends on the motif discovery algorithm. For the DNAshapedTFBS-based methods, we first construct, for each best site in the training set, a feature vector containing its JASPAR PFM score or encoded DNA sequence. If the method takes account of the DNA shape features, the feature vector also contains the normalized values of the 8 DNA shape features at each position. Then a gradient boosting classifier is trained on the positive and negative feature vectors. For DiMO_PWM, the sequences of the positive and negative sites are directly fed into the program along with the JASPAR PFM, which serves as a seed matrix. DiMO_PWM then finds the optimal PWM that maximizes the AUROC on the training set and outputs it without any transformation or normalization.

Testing

The testing procedure is the same for all the algorithms. We first use the trained gradient boosting classifiers, the JASPAR PFMs, and the DiMO PWMs to score all the positive and negative sites in the testing set. We then use the scores predicted by each model to rank the sites and compute the area under the precision recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) based on the true labels of the ranked sites.

4.3 Results

For each algorithm, the mean and standard deviation of the AUPRC and AUROC scores for the 396 samples are summarized in Table 4.2. The table shows that the average scores on the testing sets are similar for all the algorithms except for the JASPAR PFMs, while the average scores on the training sets vary greatly. The gap between the training and testing scores is a measure of overfitting for a model, which is generally in line with the complexity of the model according to our results. Specifically, the highly non-linear DNAshapedTFBS-based models, which use an ensemble of decision trees for classification, have larger gaps than the JASPAR PFMs and the DiMO PWMs, which are both linear models. In fact, the training and testing scores of the two types of linear models are generally comparable. The largest gaps (0.125 for AUPRC and 0.142 for AUROC) are associated with the DNAshapedTFBS-based models with DNA shape features, presumably because their feature vectors are most complex. If we focus on the scores on the testing sets, several conclusions can be drawn from Table 4.2. First, the algorithms with the highest AUPRC are DiMO_PWM, DNAshapedTFBS_4bit and JASPAR + shape, while the JASPAR PFMs have the lowest scores, due in part to the inherent limitations of probabilistic models described in Chapter 2. Second, taking account of adjacent dinucleotides or DNA shape features does not improve the performance of the DNAshapedTFBS-based models. It should be noted that JASPAR is not a DNAshapedTFBS-based model even though JASPAR + shape

outperforms JASPAR. In Figure 4.1, we compare the AUPRC scores generated by the DNAshapedTFBS-based methods with and without DNA shape features. In all the subplots except for Figure 4.1a, the data points cluster evenly around the diagonal, indicating that the difference between the two scores of the same sample is likely due to random noise in model training. Figure 4.1a shows that JASPAR + shape largely outperforms JASPAR, a result also reported by Mathelier *et al*. Finally, the 3-bit sequence encoding is on a par with the 4-bit encoding, indicating that the latter is redundant and unnecessary. Figure 4.2 compares the AUPRC scores generated by DNAshapedTFBS_4bit + shape, one of the best algorithms reported by Mathelier *et al.*, with several other methods. It shows that DiMO_PWM slightly outperforms DNAshapedTFBS_4bit + shape, while the JASPAR PFMs generally underperform.

Table 4.2 Evaluation scores for the motif discovery algorithms on ChIP-seq data

Algorithm	Training AUPRC	Training AUROC	Testing AUPRC	Testing AUROC
DiMO_PWM	0.842 (0.112)	0.828 (0.112)	0.829 (0.123)	0.812 (0.126)
JASPAR	0.812 (0.132)	0.788 (0.139)	0.813 (0.131)	0.788 (0.140)
DNAshapedTFBS_4bit	0.896 (0.083)	0.892 (0.083)	0.829 (0.126)	0.813 (0.128)
DNAshapedTFBS_3bit	0.891 (0.084)	0.887 (0.084)	0.826 (0.127)	0.811 (0.128)
DNAshapedTFBS_3bit_dinuc	0.903 (0.079)	0.898 (0.080)	0.826 (0.127)	0.810 (0.128)
JASPAR + shape	0.951 (0.051)	0.951 (0.050)	0.829 (0.124)	0.811 (0.128)
DNAshapedTFBS_4bit + shape	0.948 (0.051)	0.949 (0.050)	0.824 (0.128)	0.807 (0.130)
DNAshapedTFBS_3bit + shape	0.948 (0.051)	0.949 (0.050)	0.823 (0.128)	0.807 (0.130)
DNAshapedTFBS_3bit_dinuc + shape	0.948 (0.051)	0.949 (0.050)	0.825 (0.127)	0.808 (0.129)

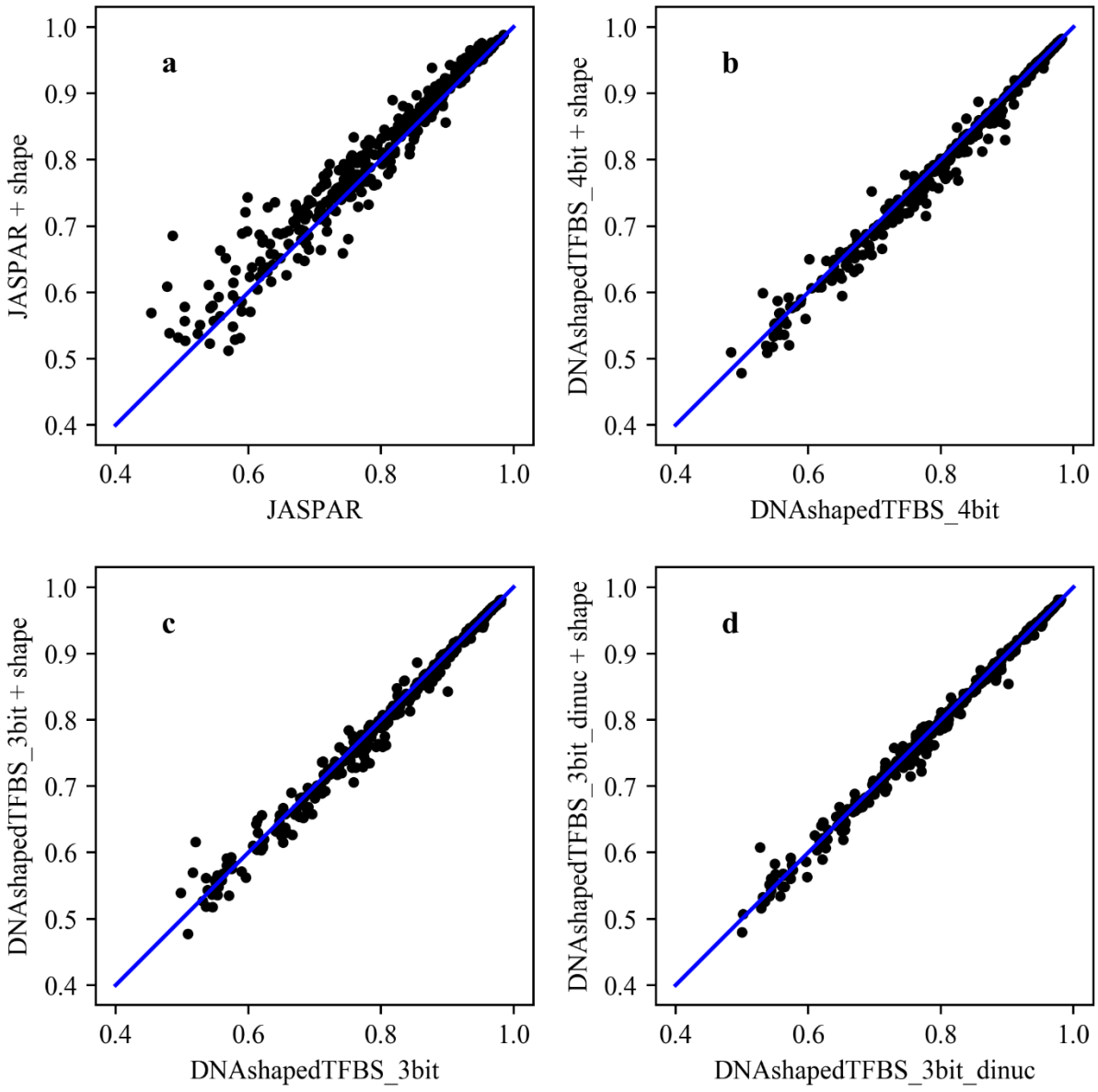


Figure 4.1 Comparison of the AUPRC scores generated by the DNAshapedTFBS-based methods with and without DNA shape features.

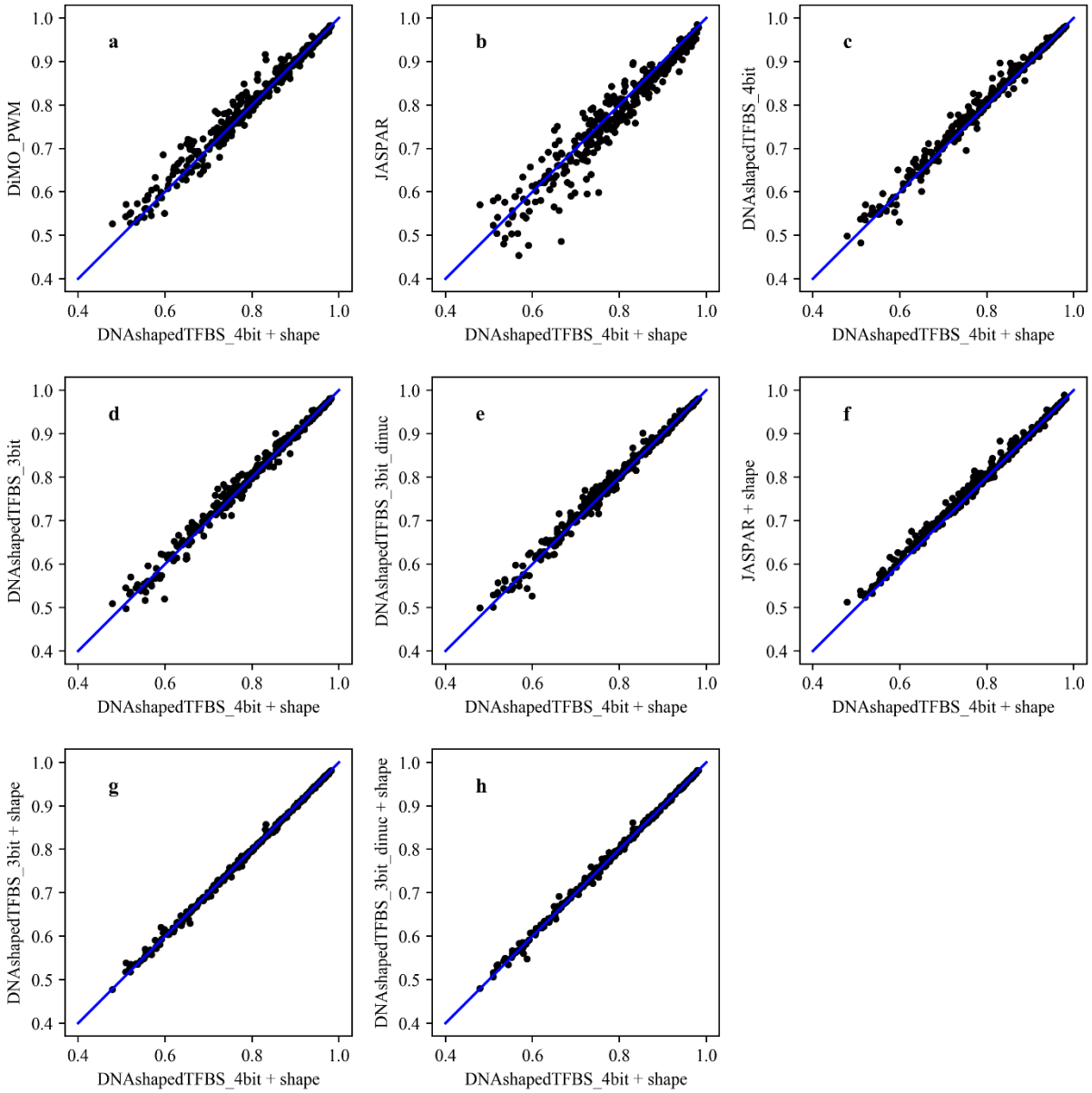


Figure 4.2 Comparison of the AUPRC scores generated by DNAshapedTFBS_4bit + shape with other methods.

4.4 Discussion

Our preliminary results indicate that adding DNA shape features does not significantly improve the performance of the DNAshapedTFBS-based models, which seems to be different from the results in the Mathelier study. However, after analyzing the raw evaluation scores reported by Mathelier *et al*. (github.com/amathelier/DNAshapedTFBS_notebooks), we found that the mean AUPRC scores for DNAshapedTFBS_4bit and DNAshapedTFBS_4bit + shape are 0.861 and 0.867 respectively, with a difference of only 0.006. In addition to the small improvements, the incorporation of DNA shape features into motif discovery algorithms may be challenging. First, the GBshape database only provides predictions instead of measurements of DNA shape features [97]. These predictions are based on Monte Carlo simulations of DNA pentamers and may deviate from actual DNA structures [96], which can only be measured using X-ray crystallography. Moreover, X-rays can only shed light on the structures of protein–DNA complexes and DNA fragments in crystals, but these macromolecules may constantly transition from one configuration to another under normal conditions. The flexibility of protein domains and DNA may play a role in binding site recognition [117], and it cannot be captured by static shape parameters. Second, taking account of DNA shape features significantly increases the number of parameters in a binding model, which will increase the cost of training and may result in overfitting. For example, it only takes $3k$ parameters to encode a k bp sequence using 3-bit encoding. However, k additional parameters are needed for incorporating each type of DNA shape feature. In the Mathelier study, 8 types of DNA shape features were added, thus tripling the number of total features in some models. This further increases the complexity of the intrinsically non-linear gradient boosting classifiers, which use an ensemble of decision trees for classification.

The fact that the performance of DiMO_PWM is better than or equal to the non-linear gradient boosting classifiers underscores the validity of the assumption that each position in a binding site contributes independently to its total binding energy. However, the success of the PWMs does not mean that the structure of DNA plays no role in binding site recognition. In fact, there are good examples showing that it does [118]. The JASPAR PFMs, DiMO PWMs and DNAshapedTFBS_4bit models are all agnostic with respect to the mechanisms of specificity. They only describe mathematically how much each base at each position contributes to binding affinity (or preference) and cannot infer these parameters from more basic principles of chemistry or quantum mechanics. But for the purpose of predicting binding sites, the actual mechanism is irrelevant. In this case, it may be advantageous to choose the most efficient model, one that optimizes the fit to the data and has minimum complexity, since it is easier to interpret, takes less time to train, and is less susceptible to overfitting.

Future Plans

A number of improvements can be made to the BEESEM program. For example, BEESEM currently uses all the *m* bp sequences (*m* is the motif length) to construct the objective function and then searches for the optimal parameters. This is very time-consuming and can be replaced with mini-batch gradient descent, in which the true gradient of the overall objective function is approximated by the gradient of a series of partial objective functions constructed from mini-batches of random *m* bp sequences [119]. The mini-batch method has been shown to improve the optimization of deep neural networks. Since the structure of the BEESEM objective function is very similar to a neural network, the mini-batch method may also make its optimization procedure more efficient. In addition to gains in efficiency, new features can extend the applicability of BEESEM. For example, BEESEM currently cannot model gaps within binding motifs, although its PWMs may happen to contain a string of low-information positions. If BEESEM could model gaps, the user can explicitly specify the location and length of the gap based on prior knowledge and BEESEM can use this information to generate more accurate motifs in a more efficient fashion. Another feature that can be added to BEESEM is the support for ambiguous bases within input sequence reads. Currently any sequence with ambiguous bases is filtered out before the sequence reads are fed into BEESEM. Finally, although BEESEM is designed to extract motifs from HT-SELEX data, its underlying biophysical model can be extended to other types of high-throughput experiments such as PBMs [67]. A generalized BEESEM model may also be used to infer the cooperativity between two TFs from CAP-SELEX experiments [106], which are similar to HT-SELEX but use two types of TFs in each experiment.

References

1. D.G. Vassylyev, S. Sekine, O. Laptenko, et al., *Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution.* Nature, 2002. **417**(6890): p. 712-9.

2. S. Buratowski, *The basics of basal transcription by RNA polymerase II.* Cell, 1994. **77**(1): p. 1-3.

3. J.D. McGhee and G. Felsenfeld, *Nucleosome structure.* Annu Rev Biochem, 1980. **49**: p. 1115-56.

4. P.A. Jones, *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.

5. G.D. Stormo and Y. Zhao, *Determining the specificity of protein-DNA interactions.* Nat Rev Genet, 2010. **11**(11): p. 751-60.

6. C.O. Pabo and R.T. Sauer, *Transcription factors: structural families and principles of DNA recognition.* Annu Rev Biochem, 1992. **61**: p. 1053-95.

7. C.G. de Boer and T.R. Hughes, *YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities.* Nucleic Acids Res, 2012. **40**(Database issue): p. D169-79.

8. D.Y. Rhee, D.Y. Cho, B. Zhai, et al., *Transcription factor networks in Drosophila melanogaster.* Cell Rep, 2014. **8**(6): p. 2031-43.

9. J.M. Vaquerizas, S.K. Kummerfeld, S.A. Teichmann, et al., *A census of human transcription factors: function, expression and evolution.* Nat Rev Genet, 2009. **10**(4): p. 252-63.

10. N.P. Pavletich and C.O. Pabo, *Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A.* Science, 1991. **252**(5007): p. 809-17.

11. L. Fairall, J.W. Schwabe, L. Chapman, et al., *The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition.* Nature, 1993. **366**(6454): p. 483-7.

12. W.J. Gehring, M. Muller, M. Affolter, et al., *The structure of the homeodomain and its functional implications.* Trends Genet, 1990. **6**(10): p. 323-9.

13. A. Tanay, *Extensive low-affinity transcriptional interactions in the yeast genome.* Genome Res, 2006. **16**(8): p. 962-72.

14. A.I. Ramos and S. Barolo, *Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution.* Philos Trans R Soc Lond B Biol Sci, 2013. **368**(1632): p. 20130018.

15. H.L. Pahl, *Activators and target genes of Rel/NF-kappaB transcription factors.* Oncogene, 1999. **18**(49): p. 6853-66.

16. M. Lewis, G. Chang, N.C. Horton, et al., *Crystal structure of the lactose operon repressor and its complexes with DNA and inducer.* Science, 1996. **271**(5253): p. 1247-54.

17. W. Zheng, T.A. Gianoulis, K.J. Karczewski, et al., *Regulatory variation within and between species.* Annu Rev Genomics Hum Genet, 2011. **12**: p. 327-46.

18. S.B. Carroll, *Evolution at two levels: on genes and form.* PLoS Biol, 2005. **3**(7): p. e245.

19. S. Mertin, S.G. McDowall, and V.R. Harley, *The DNA-binding specificity of SOX9 and other SOX proteins.* Nucleic Acids Res, 1999. **27**(5): p. 1359-64.

20. T. Kouzarides and E. Ziff, *Leucine zippers of fos, jun and GCN4 dictate dimerization specificity and thereby control DNA binding.* Nature, 1989. **340**(6234): p. 568-71.

21. T. Hai and T. Curran, *Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity.* Proc Natl Acad Sci U S A, 1991. **88**(9): p. 3720-4.

22. A. Al-Sarraj, R.M. Day, and G. Thiel, *Specificity of transcriptional regulation by the zinc finger transcription factors Sp1, Sp3, and Egr-1.* J Cell Biochem, 2005. **94**(1): p. 153-67.

23. M.T. Weirauch, A. Yang, M. Albu, et al., *Determination and inference of eukaryotic transcription factor sequence specificity.* Cell, 2014. **158**(6): p. 1431-43.

24. A. Jolma, T. Kivioja, J. Toivonen, et al., *Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities.* Genome Res, 2010. **20**(6): p. 861-73.

25. M.F. Berger, A.A. Philippakis, A.M. Qureshi, et al., *Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities.* Nat Biotechnol, 2006. **24**(11): p. 1429-35.

26. A. Valouev, D.S. Johnson, A. Sundquist, et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nat Methods, 2008. **5**(9): p. 829-34.

27. G.D. Stormo, *Modeling the specificity of protein-DNA interactions.* Quant Biol, 2013. **1**(2): p. 115-30.

28. R.H. Ebright, Y.W. Ebright, and A. Gunasekera, *Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site.* Nucleic Acids Res, 1989. **17**(24): p. 10295-305.

29. T.D. Schneider, *Consensus sequence Zen.* Appl Bioinformatics, 2002. **1**(3): p. 111-9.

30. T.D. Schneider and R.M. Stephens, *Sequence logos: a new way to display consensus sequences.* Nucleic Acids Res, 1990. **18**(20): p. 6097-100.

31. S. Rockel, M. Geertz, and S.J. Maerkl, *MITOMI: a microfluidic platform for in vitro characterization of transcription factor-DNA interaction.* Methods Mol Biol, 2012. **786**: p. 97-114.

32. G.D. Stormo, Z. Zuo, and Y.K. Chang, *Spec-seq: determining protein-DNA-binding specificity by sequencing.* Brief Funct Genomics, 2015. **14**(1): p. 30-8.

33. M. Djordjevic, A.M. Sengupta, and B.I. Shraiman, *A biophysical approach to transcription factor binding site discovery.* Genome Res, 2003. **13**(11): p. 2381-90.

34. Y. Zhao, D. Granas, and G.D. Stormo, *Inferring binding energies from selected binding sites.* PLoS Comput Biol, 2009. **5**(12): p. e1000590.

35. M. Annala, K. Laurila, H. Lahdesmaki, et al., *A linear model for transcription factor binding affinity prediction in protein binding microarrays.* PLoS One, 2011. **6**(5): p. e20059.

36. J. Keilwagen, J. Grau, I.A. Paponov, et al., *De-novo discovery of differentially abundant transcription factor binding sites including their positional preference.* PLoS Comput Biol, 2011. **7**(2): p. e1001070.

37. A. Mathelier and W.W. Wasserman, *The next generation of transcription factor binding site prediction.* PLoS Comput Biol, 2013. **9**(9): p. e1003214.

38. B. Alipanahi, A. Delong, M.T. Weirauch, et al., *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.* Nat Biotechnol, 2015. **33**(8): p. 831-8.

39. M.T. Weirauch, A. Cote, R. Norel, et al., *Evaluation of methods for modeling transcription factor sequence specificity.* Nat Biotechnol, 2013. **31**(2): p. 126-34.

40. A. Mathelier, B. Xin, T.P. Chiu, et al., *DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo.* Cell Syst, 2016. **3**(3): p. 278-86 e4.

41. M.J. Rossi, W.K.M. Lai, and B.F. Pugh, *Correspondence: DNA shape is insufficient to explain binding.* Nat Commun, 2017. **8**: p. 15643.

42. R. Harr, M. Haggstrom, and P. Gustafsson, *Search algorithm for pattern match analysis of nucleic acid sequences.* Nucleic Acids Res, 1983. **11**(9): p. 2943-57.

43. R. Staden, *Computer methods to locate signals in nucleic acid sequences.* Nucleic Acids Res, 1984. **12**(1 Pt 2): p. 505-19.

44. T.D. Schneider, G.D. Stormo, L. Gold, et al., *Information content of binding sites on nucleotide sequences.* J Mol Biol, 1986. **188**(3): p. 415-31.

45. C.E. Lawrence, S.F. Altschul, M.S. Boguski, et al., *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.* Science, 1993. **262**(5131): p. 208-14.

46. C.E. Lawrence and A.A. Reilly, *An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.* Proteins, 1990. **7**(1): p. 41-51.

47. G.D. Stormo and G.W. Hartzell, 3rd, *Identifying protein-binding sites from unaligned DNA fragments.* Proc Natl Acad Sci U S A, 1989. **86**(4): p. 1183-7.

48. G.D. Stormo, *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.

49. E. van Nimwegen, *Finding regulatory elements and regulatory motifs: a general probabilistic framework.* BMC Bioinformatics, 2007. **8 Suppl 6**: p. S4.

50. P. D'Haeseleer, *How does DNA sequence motif discovery work?* Nat Biotechnol, 2006. **24**(8): p. 959-61.

51. H.J. Bussemaker, B.C. Foat, and L.D. Ward, *Predictive modeling of genome-wide mRNA expression: from modules to molecules.* Annu Rev Biophys Biomol Struct, 2007. **36**: p. 329-47.

52. P.H. Von Hippel, *On the Molecular Bases of the Specificity of Interaction of Transcriptional Proteins with Genome DNA.* Biological Regulation and Development. Vol. 1. 1979, New York, NY: Plenum Publishing Corp. 279-347.

53. O.G. Berg and P.H. von Hippel, *Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.* J Mol Biol, 1987. **193**(4): p. 723-50.

54. G.D. Stormo, *Computer methods for analyzing sequence recognition of nucleic acids.* Annu Rev Biophys Biophys Chem, 1988. **17**: p. 241-63.

55. G.D. Stormo and D.S. Fields, *Specificity, free energy and information content in protein-DNA interactions.* Trends Biochem Sci, 1998. **23**(3): p. 109-13.

56. L. Bintu, N.E. Buchler, H.G. Garcia, et al., *Transcriptional regulation by the numbers: models.* Curr Opin Genet Dev, 2005. **15**(2): p. 116-24.

57. U. Gerland, J.D. Moroz, and T. Hwa, *Physical constraints and functional characteristics of transcription factor-DNA interaction.* Proc Natl Acad Sci U S A, 2002. **99**(19): p. 12015-20.

58. D.S. Homsi, V. Gupta, and G.D. Stormo, *Modeling the quantitative specificity of DNA-binding proteins from example binding sites.* PLoS One, 2009. **4**(8): p. e6736.

59. B.C. Foat, A.V. Morozov, and H.J. Bussemaker, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.* Bioinformatics, 2006. **22**(14): p. e141-9.

60. G.D. Stormo, T.D. Schneider, and L. Gold, *Quantitative analysis of the relationship between nucleotide sequence and functional activity.* Nucleic Acids Res, 1986. **14**(16): p. 6661-79.

61. J. Atherton, N. Boley, B. Brown, et al., *A Model for Sequential Evolution of Ligands by Exponential Enrichment (Selex) Data.* Annals of Applied Statistics, 2012. **6**(3): p. 928-49.

62. P.M. Fordyce, D. Gerber, D. Tran, et al., *De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis.* Nat Biotechnol, 2010. **28**(9): p. 970-5.

63. G. Locke and A.V. Morozov, *A Biophysical Approach to Predicting Protein-DNA Binding Energetics.* Genetics, 2015. **200**(4): p. 1349-61.

64. Y. Orenstein and R. Shamir, *HTS-IBIS: fast and accurate inference of binding site motifs from HT-SELEX data.* bioRxiv, 2015.

65. T.R. Riley, A. Lazarovici, R.S. Mann, et al., *Building accurate sequence-to-affinity models from high-throughput in vitro protein-DNA binding data using FeatureREDUCE.* Elife, 2015. **4**.

66. T.R. Riley, M. Slattery, N. Abe, et al., *SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes.* Methods Mol Biol, 2014. **1196**: p. 255-78.

67. Y. Zhao and G.D. Stormo, *Quantitative analysis demonstrates most transcription factors require only simple models of specificity.* Nat Biotechnol, 2011. **29**(6): p. 480-3.

68. Z. Zuo and G.D. Stormo, *High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding.* Genetics, 2014. **198**(3): p. 1329-43.

69. S. Ruan, S. Joshua Swamidass, and G.D. Stormo, *BEESEM: Estimation of Binding Energy Models Using HT-SELEX Data.* Bioinformatics, 2017.

70. J. Liu and G.D. Stormo, *Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions.* Nucleic Acids Res, 2005. **33**(17): p. e141.

71. H.J. Bussemaker, *Recent progress in understanding transcription factor binding specificity.* Brief Funct Genomics, 2015. **14**(1): p. 1-2.

72. Y. Zhao, S. Ruan, M. Pandey, et al., *Improved models for transcription factor binding site identification using nonindependent interactions.* Genetics, 2012. **191**(3): p. 781-90.

73. R. Nutiu, R.C. Friedman, S. Luo, et al., *Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument.* Nat Biotechnol, 2011. **29**(7): p. 659-64.

74. J.W. Puckett, K.A. Muzikar, J. Tietjen, et al., *Quantitative microarray profiling of DNA-binding molecules.* J Am Chem Soc, 2007. **129**(40): p. 12310-9.

75. G. Badis, M.F. Berger, A.A. Philippakis, et al., *Diversity and complexity in DNA recognition by transcription factors.* Science, 2009. **324**(5935): p. 1720-3.

76. R. Gordan, K.F. Murphy, R.P. McCord, et al., *Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights.* Genome Biol, 2011. **12**(12): p. R125.

77. G. Badis, E.T. Chan, H. van Bakel, et al., *A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters.* Mol Cell, 2008. **32**(6): p. 878-87.

78. M.F. Berger, G. Badis, A.R. Gehrke, et al., *Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences.* Cell, 2008. **133**(7): p. 1266-76.

79. K. Narasimhan, S.A. Lambert, A.W. Yang, et al., *Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities.* Elife, 2015. **4**.

80. H.S. Najafabadi, S. Mnaimneh, F.W. Schmitges, et al., *C2H2 zinc finger proteins greatly expand the human regulatory lexicon.* Nat Biotechnol, 2015. **33**(5): p. 555-62.

81. C. Tuerk and L. Gold, *Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.* Science, 1990. **249**(4968): p. 505-10.

82. A. Zykovich, I. Korf, and D.J. Segal, *Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing.* Nucleic Acids Res, 2009. **37**(22): p. e151.

83. D. Wong, A. Teixeira, S. Oikonomopoulos, et al., *Extensive characterization of NF-kappaB binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits.* Genome Biol, 2011. **12**(7): p. R70.

84. N. Ogawa and M.D. Biggin, *High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro.* Methods Mol Biol, 2012. **786**: p. 51-63.

85. M. Slattery, T. Riley, P. Liu, et al., *Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.* Cell, 2011. **147**(6): p. 1270-82.

86. K.R. Nitta, A. Jolma, Y. Yin, et al., *Conservation of transcription factor binding specificities across 600 million years of bilateria evolution.* Elife, 2015. **4**.

87. A. Jolma, J. Yan, T. Whitington, et al., *DNA-binding specificities of human transcription factors.* Cell, 2013. **152**(1-2): p. 327-39.

88. A. Gupta, R.G. Christensen, H.A. Bell, et al., *An improved predictive recognition model for Cys(2)-His(2) zinc finger proteins.* Nucleic Acids Res, 2014. **42**(8): p. 4800-12.

89. Y. Orenstein and R. Shamir, *A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data.* Nucleic Acids Res, 2014. **42**(8): p. e63.

90. Y. Shi, *A glimpse of structural biology through X-ray crystallography.* Cell, 2014. **159**(5): p. 995-1014.

91. R. Fernandez-Leiro and S.H. Scheres, *Unravelling biological macromolecules with cryo-electron microscopy.* Nature, 2016. **537**(7620): p. 339-46.

92. P.R. Markwick, T. Malliavin, and M. Nilges, *Structural biology by NMR: structure, dynamics, and interactions.* PLoS Comput Biol, 2008. **4**(9): p. e1000168.

93. S. Jones, P. van Heyningen, H.M. Berman, et al., *Protein-DNA interactions: A structural analysis.* J Mol Biol, 1999. **287**(5): p. 877-96.

94. F.D. Urnov, E.J. Rebar, M.C. Holmes, et al., *Genome editing with engineered zinc finger nucleases.* Nat Rev Genet, 2010. **11**(9): p. 636-46.

95. P.W. Rose, A. Prlic, A. Altunkaya, et al., *The RCSB protein data bank: integrative view of protein, gene and 3D structural information.* Nucleic Acids Res, 2017. **45**(D1): p. D271-D81.

96. T. Zhou, L. Yang, Y. Lu, et al., *DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W56-62.

97. T.P. Chiu, L. Yang, T. Zhou, et al., *GBshape: a genome browser database for DNA shape annotations.* Nucleic Acids Res, 2015. **43**(Database issue): p. D103-9.

98. G.E. Zentner, S. Kasinathan, B. Xin, et al., *ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo.* Nat Commun, 2015. **6**: p. 8733.

99. R.Y. Patel and G.D. Stormo, *Discriminative motif optimization based on perceptron training.* Bioinformatics, 2014. **30**(7): p. 941-8.

100. C.Y. Zhu, R.H. Byrd, P.H. Lu, et al., *Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization.* Acm Transactions on Mathematical Software, 1997. **23**(4): p. 550-60.

101. B. Roy, Z. Zuo, and G.D. Stormo, *Quantitative specificity of STAT1 and several variants.* Nucleic Acids Res, 2017.

102. A. Isakova, R. Groux, M. Imbeault, et al., *SMiLE-seq identifies binding motifs of single and dimeric transcription factors.* Nat Methods, 2017. **14**(3): p. 316-22.

103. S.J. Maerkl and S.R. Quake, *A systems approach to measuring the binding energy landscapes of transcription factors.* Science, 2007. **315**(5809): p. 233-7.

104. T.L. Bailey, J. Johnson, C.E. Grant, et al., *The MEME Suite.* Nucleic Acids Res, 2015. **43**(W1): p. W39-49.

105. G.Z. Hertz and G.D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.* Bioinformatics, 1999. **15**(7-8): p. 563-77.

106. A. Jolma, Y. Yin, K.R. Nitta, et al., *DNA-dependent formation of transcription factor pairs alters their binding specificity.* Nature, 2015. **527**(7578): p. 384-8.

107. M. Slattery, T. Zhou, L. Yang, et al., *Absence of a simple code: how transcription factors read the genome.* Trends Biochem Sci, 2014. **39**(9): p. 381-99.

108. J.A. Rice, *Mathematical statistics and data analysis*. 3rd ed. Duxbury advanced series. 2007, Belmont, CA: Thomson/Brooks/Cole.

109. Y. Orenstein, E. Mick, and R. Shamir, *RAP: accurate and fast motif finding based on protein-binding microarray data.* J Comput Biol, 2013. **20**(5): p. 375-82.

110. S. Kullback and R.A. Leibler, *On Information and Sufficiency.* The Annals of Mathematical Statistics, 1951. **22**(1): p. 79-86.

111. S.G. Landt, G.K. Marinov, A. Kundaje, et al., *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.* Genome Res, 2012. **22**(9): p. 1813-31.

112. M. Kasowski, F. Grubert, C. Heffelfinger, et al., *Variation in transcription factor binding among humans.* Science, 2010. **328**(5975): p. 232-5.

113. T.E. Reddy, J. Gertz, F. Pauli, et al., *Effects of sequence variation on differential allelic transcription factor occupancy and gene expression.* Genome Res, 2012. **22**(5): p. 860-9.

114. A. Mathelier, O. Fornes, D.J. Arenillas, et al., *JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2016. **44**(D1): p. D110-5.

115. E.P. Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

116. G.D. Stormo, *Maximally efficient modeling of DNA sequence motifs at all levels of complexity.* Genetics, 2011. **187**(4): p. 1219-24.

117. L. Mathiasen, E. Valentini, S. Boivin, et al., *The flexibility of a homeodomain transcription factor heterodimer and its allosteric regulation by DNA binding.* FEBS J, 2016. **283**(16): p. 3134-54.

118. C. Spiro, D.P. Bazett-Jones, X. Wu, et al., *DNA structure determines protein binding and transcriptional efficiency of the proenkephalin cAMP-responsive enhancer.* J Biol Chem, 1995. **270**(46): p. 27702-10.

119. O. Dekel, R. Gilad-Bachrach, O. Shamir, et al., *Optimal Distributed Online Prediction Using Mini-Batches.* Journal of Machine Learning Research, 2012. **13**: p. 165-202.

120. A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data Via Em Algorithm.* Journal of the Royal Statistical Society Series B-Methodological, 1977. **39**(1): p. 1-38.

Appendix A: Biophysical Model of BEESEM

In this appendix, we give a detailed description of the biophysical model of BEESEM in four steps.

In the first step, we consider the interaction between the TF (denoted by *T*) and one DNA species (denoted by its sequence *S*). We use $\bar{S}$ to represent the reverse complement of *S*. We assume that the DNA binding site of the TF is asymmetrical and the DNA sequence is non-palindromic. In this case, the protein–DNA complex may have two configurations, which we refer to as the top configuration $T \cdot S$ and the bottom configuration $T \cdot \bar{S}$ respectively. These two configurations need not have the same binding energy, and they correspond to the two possible orientations of a non-palindromic DNA molecule in an asymmetrical binding site. We also assume the system has reached equilibrium, namely

$$T \cdot S \rightleftharpoons T^{\mathrm{U}} + S^{\mathrm{U}} \rightleftharpoons T \cdot \bar{S}, \tag{A.1}$$

where X^{U} denotes unbound molecules of *X*. In equilibrium, binding and unbinding should proceed at the same rate. As a result, we can calculate the binding (or association) constants of the two configurations [34]:

$$K = \frac{[T \cdot S]}{[T^{\mathrm{U}}][S^{\mathrm{U}}]}, \quad \bar{K} = \frac{[T \cdot \bar{S}]}{[T^{\mathrm{U}}][S^{\mathrm{U}}]}, \tag{A.2}$$

where [*X*] denotes the concentration of *X*. We are interested in the proportion of DNA molecules that are 'trapped' in the top configuration, denoted by $P(T \cdot S \mid S)$. In this notation, the second *S* represents any DNA molecule, either bound or unbound. We can prove that

$$P(T \cdot S \mid S) = \frac{K}{K + \bar{K} + 1/[T^{\mathrm{U}}]}. \tag{A.3}$$

Proof of Equation (A.3):

$$\begin{aligned} P(T\cdot S \mid S) &= \frac{[T\cdot S]}{[S]} \\ &= \frac{[T\cdot S]}{[T\cdot S]+[T\cdot \bar{S}]+[S^{\mathrm{U}}]} \\ &= \frac{K[T^{\mathrm{U}}][S^{\mathrm{U}}]}{K[T^{\mathrm{U}}][S^{\mathrm{U}}]+\bar{K}[T^{\mathrm{U}}][S^{\mathrm{U}}]+[S^{\mathrm{U}}]} \\ &= \frac{K}{K+\bar{K}+1/[T^{\mathrm{U}}]}. \end{aligned}$$

In order to proceed, we need to introduce some definitions [34]:

$$E \equiv -\ln K, \quad \bar{E} \equiv -\ln \bar{K}, \quad \mu \equiv \ln[T^{\mathrm{U}}]. \tag{A.4}$$

E and $\bar{E}$ are called the binding energies of the top and bottom configuration, respectively. And μ is called the chemical potential of the TF. According to these definitions, we have

$$P(T\cdot S \mid S) = \frac{e^{-E}}{e^{-E}+e^{-\bar{E}}+e^{-\mu}}. \tag{A.5}$$

Obviously, there is a similar equation for $P(T\cdot \bar{S} \mid S)$.

In the second step, we consider a more complex system. In this system, instead of just one DNA species, there are n distinct DNA species, denoted by S_i, $i = 1, 2, \cdots, n$. After the system reaches equilibrium, some DNA molecules be bound by the TF. We use B to represent any bound DNA molecule, regardless of its actual sequence. It follows from this notation that $P(S_i \mid B)$ represents the proportion of bound DNA molecules that have sequence S_i. If S_i is non-palindromic, we can prove that

$$P(S_i \mid B) = P(T\cdot S_i \mid B) + P(T\cdot \bar{S}_i \mid B). \tag{A.6}$$

In practice, only $P(S_i \mid B)$ can be measured directly, because we cannot distinguish $T \cdot S_i$ from $T \cdot \bar{S}_i$ in a typical experiment. But in terms of developing the model, it is more convenient to work with $P(T \cdot S_i \mid B)$ or $P(T \cdot \bar{S}_i \mid B)$ alone. It can be computed in the following way:

$$P(T \cdot S_i \mid B) = P(T \cdot S_i \mid S_i) \frac{P(S_i)}{P(B)}. \tag{A.7}$$

In the above discussions, we implicitly assume that the DNA molecules all have the same length, denoted by *l*. In addition, we assume $l = m$, where *m* represents the length of actual binding sites. In the third step, we still assume that the DNA molecules all have the same length. But we relax the second assumption and only assume $l \geq m$. Also we assume that the DNA binding site of the TF is asymmetrical and that the DNA molecules are all non-palindromic. Even if some DNA molecules are palindromic, we expect our conclusions to be largely valid. As a result, each protein–DNA complex may have $2(l-m+1)$, instead of only two, configurations. The factor '2' reflects the two possible orientations of a non-palindromic DNA molecule in an asymmetrical binding site. To keep track of the different configurations and complexes, we use $T \cdot S_i^k$ to denote the *k*th configuration of the *i*th protein–DNA complex, $k = 1, 2, \cdots, 2(l-m+1)$ and $i = 1, 2, \cdots, n$. In other words, the superscript *k* indicates where the actual binding site is located within sequence S_i. Thus S_i^k can also represent the sequence of the binding site, which is an *m*-mer. If the system has reached equilibrium, we can prove that

$$P(T \cdot S_i^k \mid S_i) = \frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q} + e^{-\mu}}. \tag{A.8}$$

It is a natural generalization of Equation (A.5). In this equation, E_i^k is simply a shorthand for $E(T \cdot S_i^k)$, the binding energy of configuration $T \cdot S_i^k$. In theory, E_i^k is entirely determined by the

DNA sequence of S_i^k, regardless of its location. Specifically, it is usually assumed that the binding energy can be represented by the inner product of a PWM and the encoding vector of a DNA sequence. Thus E_i^k can also be written as $E(T \cdot s_j)$, if $S_i^k = s_j$, where s_j represents the DNA sequence of a particular *m*-mer. The subscript *j* distinguishes one *m*-mer from another, and it ranges from 1 to 4^m (the total number of single-stranded *m*-mers). It should be noted that S_i has a length of *l*, while S_i^k and s_j both have a length of *m*. In addition, we can prove a result similar to Equation (A.7):

$$P(T \cdot S_i^k \mid B) = P(T \cdot S_i^k \mid S_i) \frac{P(S_i)}{P(B)}. \tag{A.9}$$

To proceed, we need to introduce an indicator function I_{ij}^k, which satisfies the following definition:

$$I_{ij}^k \equiv \begin{cases} 1, & \text{if } S_i^k = s_j, \\ 0, & \text{if } S_i^k \neq s_j. \end{cases} \tag{A.10}$$

We use $P(T \cdot s_j \mid B)$ to represent the proportion of actual binding sites that have sequence s_j, namely

$$P(T \cdot s_j \mid B) \equiv \sum_{i=1}^{n} \sum_{k=1}^{2(l-m+1)} I_{ij}^k P(T \cdot S_i^k \mid B). \tag{A.11}$$

In addition, we define a weighting factor w_{ij}:

$$w_{ij} \equiv \frac{e^{-E(T \cdot s_j)} + e^{-\mu}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q} + e^{-\mu}}. \tag{A.12}$$

Then we can prove that

$$P(T\cdot s_j \mid B)=\frac{1}{P(B)}\frac{e^{-E(T\cdot s_j)}}{e^{-E(T\cdot s_j)}+e^{-\mu}}\sum_{i=1}^{n}w_{ij}P(S_i)\sum_{k=1}^{2(l-m+1)}I_{ij}^{k}. \tag{A.13}$$

Proof of Equation (A.13):

$$\begin{aligned}
P(T\cdot s_j \mid B) &\equiv \sum_{i=1}^{n}\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}P(T\cdot S_i^k \mid B)\\
&=\sum_{i=1}^{n}\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}P(T\cdot S_i^k \mid S_i)\frac{P(S_i)}{P(B)}\\
&=\frac{1}{P(B)}\sum_{i=1}^{n}P(S_i)\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}P(T\cdot S_i^k \mid S_i)\\
&=\frac{1}{P(B)}\sum_{i=1}^{n}P(S_i)\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}\frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)}e^{-E_i^q}+e^{-\mu}}\\
&=\frac{1}{P(B)}\sum_{i=1}^{n}\frac{P(S_i)}{\sum_{q=1}^{2(l-m+1)}e^{-E_i^q}+e^{-\mu}}\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}e^{-E_i^k}\\
&=\frac{1}{P(B)}\sum_{i=1}^{n}\frac{P(S_i)}{\sum_{q=1}^{2(l-m+1)}e^{-E_i^q}+e^{-\mu}}\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}e^{-E(T\cdot s_j)}\\
&=\frac{e^{-E(T\cdot s_j)}}{P(B)}\sum_{i=1}^{n}\frac{P(S_i)}{\sum_{q=1}^{2(l-m+1)}e^{-E_i^q}+e^{-\mu}}\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}\\
&=\frac{1}{P(B)}\frac{e^{-E(T\cdot s_j)}}{e^{-E(T\cdot s_j)}+e^{-\mu}}\sum_{i=1}^{n}\frac{e^{-E(T\cdot s_j)}+e^{-\mu}}{\sum_{q=1}^{2(l-m+1)}e^{-E_i^q}+e^{-\mu}}P(S_i)\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}\\
&=\frac{1}{P(B)}\frac{e^{-E(T\cdot s_j)}}{e^{-E(T\cdot s_j)}+e^{-\mu}}\sum_{i=1}^{n}w_{ij}P(S_i)\sum_{k=1}^{2(l-m+1)} I_{ij}^{k}.
\end{aligned}$$

Then we introduce a second approach to calculate $P(T\cdot s_j \mid B)$, which is based on $P(S_i \mid B)$ instead of $P(S_i)$. $P(S_i \mid B)$ again represents the proportion of bound DNA molecules that have sequence S_i, and it can be measured directly in an experiment. To achieve that goal, we first need to reveal the relationship between $P(T\cdot S_i^k \mid B)$ and $P(S_i \mid B)$. According to the definition of the conditional probability, it can be proved that

$$P(T \cdot S_i^k \mid B) = \frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} P(S_i \mid B). \tag{A.14}$$

Proof of Equation (A.14):

$$\begin{aligned}
\frac{P(T \cdot S_i^k \mid B)}{P(S_i \mid B)} &= \frac{P(T \cdot S_i^k \cap B)}{P(B)} \frac{P(B)}{P(S_i \cap B)} \\
&= \frac{P(T \cdot S_i^k \cap B)}{P(S_i \cap B)} \\
&= \frac{[T \cdot S_i^k]}{[T \cdot S_i]} \\
&= \frac{[T \cdot S_i^k]}{\sum_{q=1}^{2(l-m+1)} [T \cdot S_i^q]} \\
&= \frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}}.
\end{aligned}$$

This equation states the relationship between $P(T \cdot S_i^k \mid B)$ and $P(S_i \mid B)$. And it enables us to derive $P(T \cdot s_j \mid B)$ from $P(S_i \mid B)$:

$$P(T \cdot s_j \mid B) = e^{-E(T \cdot s_j)} \sum_{i=1}^{n} \frac{P(S_i \mid B)}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} \sum_{k=1}^{2(l-m+1)} I_{ij}^k. \tag{A.15}$$

Proof of Equation (A.15):

$$
\begin{aligned}
P(T \cdot s_j \mid B) &\equiv \sum_{i=1}^{n} \sum_{k=1}^{2(l-m+1)} I_{ij}^{k} P(T \cdot S_i^{k} \mid B) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{2(l-m+1)} I_{ij}^{k} \frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} P(S_i \mid B) \\
&= \sum_{i=1}^{n} \frac{P(S_i \mid B)}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} \sum_{k=1}^{2(l-m+1)} I_{ij}^{k} e^{-E_i^k} \\
&= \sum_{i=1}^{n} \frac{P(S_i \mid B)}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} \sum_{k=1}^{2(l-m+1)} I_{ij}^{k} e^{-E(T \cdot s_j)} \\
&= e^{-E(T \cdot s_j)} \sum_{i=1}^{n} \frac{P(S_i \mid B)}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} \sum_{k=1}^{2(l-m+1)} I_{ij}^{k}.
\end{aligned}
$$

In the last step, we derive the objective function for estimating the unknown parameters.

Rearranging Equation (A.13) shows that

$$
\frac{P(T \cdot s_j \mid B)}{\sum_{i=1}^{n} w_{ij} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^{k}} = \frac{1}{P(B)} \frac{e^{-E(T \cdot s_j)}}{e^{-E(T \cdot s_j)} + e^{-\mu}}. \tag{A.16}
$$

For simplicity, we use R_j to denote the LHS of Equation (A.16), namely

$$
R_j \equiv \frac{P(T \cdot s_j \mid B)}{\sum_{i=1}^{n} w_{ij} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^{k}}. \tag{A.17}
$$

Then based on Equation (A.16) it can be proved that

$$
R_j = \frac{1}{P(B)} \frac{1}{1 + e^{E(T \cdot s_j) - \mu}}. \tag{A.18}
$$

Obviously, Equation (A.18) can be rewritten as

$$
R_j(\theta) = \frac{A}{1 + e^{E(T \cdot s_j) - \mu}}. \tag{A.19}
$$

In the equation, *A* is a scaling factor. Equation (A.19) is at the core of estimating unknown parameters, which include the PWM, the chemical potential μ and the scaling factor *A*.

Collectively, they are denoted by θ, the vector of unknown parameters. To estimate the unknown parameters, we seek to minimize the sum of squared differences between $R_j(\text{data})$, computed using its definition Equation (A.17), and $R_j(\theta)$, derived from Equation(A.19). In other words, we assume that

$$R_j(\text{data}) = R_j(\theta) + \varepsilon_j, \tag{A.20}$$

where ε_j is i.i.d. Gaussian noise, and the objective function is defined as

$$\hat{\theta} = \arg\min_{\theta} \sum_j \left[R_j(\text{data}) - R_j(\theta) \right]^2. \tag{A.21}$$

We choose this objective equation because it can be efficiently evaluated and existing optimization tools can be readily applied to it.

Appendix B: Proof of BEESEM Being an Expectation Maximization Algorithm

In this section, we show that BEESEM is indeed an EM algorithm. We first introduce some notations.

- **x** is a vector and $x_i = P(S_i \mid B)$, $i = 1, 2, \cdots, n$. **x** represents the observed data in an experiment.
- **Z** is a matrix with *n* rows and $2(l-m+1)$ columns, where *l* is the randomized region length and *m* is the motif length. $Z_{ik} = 1$ if S_i^k is a binding site, and $Z_{ik} = 0$ otherwise. **Z** represents the collection of hidden parameters, namely the location of each actual binding site.
- θ denotes the vector of unknown parameters.
- $\hat{\theta}^{(n)}$ and $\hat{\theta}^{(n+1)}$ represent the vector of estimated parameters in the *n*th and the $(n+1)$ th round, respectively.
- $P(\mathbf{x}, \mathbf{Z} \mid \theta)$ denotes the (joint) conditional probability of **x** and **Z** given θ. When viewing it as a function of θ, we use $L(\theta; \mathbf{x}, \mathbf{Z})$ instead and call it the likelihood function.
- $l(\theta; \mathbf{x}, \mathbf{Z})$ represents the natural logarithm of the likelihood function $\ln L(\theta; \mathbf{x}, \mathbf{Z})$.
- $\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}}$ is an operator that calculates the expected value of its operand with respect to $P(\mathbf{Z} \mid \mathbf{x}, \hat{\theta}^{(n)})$, the conditional probability of **Z** given **x** and $\hat{\theta}^{(n)}$.

The maximum likelihood estimate (MLE) of the unknown parameters can be found by maximizing the (marginal) conditional probability of **x** given θ, namely

$$\hat{\theta} = \arg\max_{\theta} P(\mathbf{x} \mid \theta) = \arg\max_{\theta} \sum_{\mathbf{Z}} P(\mathbf{x}, \mathbf{Z} \mid \theta). \tag{B.1}$$

However, this optimization is usually very time-consuming, because the number of possible values of **Z** is very large. Instead of solving the optimization directly, the EM algorithm maximizes $P(\mathbf{x} \mid \theta)$ using an iterative two-step approach [120]. First, it calculates the expected value of the log-likelihood with respect to $P(\mathbf{Z} \mid \mathbf{x}, \hat{\theta}^{(n)})$. This expected value is a function of θ and we denote it by $Q(\theta; \hat{\theta}^{(n)})$, namely

$$Q(\theta; \hat{\theta}^{(n)}) \equiv \mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}}\, l(\theta; \mathbf{x}, \mathbf{Z}) = \sum_{\mathbf{Z}} l(\theta; \mathbf{x}, \mathbf{Z}) P(\mathbf{Z} \mid \mathbf{x}, \hat{\theta}^{(n)}). \tag{B.2}$$

Second, it updates the MLE by maximizing $Q(\theta; \hat{\theta}^{(n)})$, namely

$$\hat{\theta}^{(n+1)} = \arg\max_{\theta} Q(\theta; \hat{\theta}^{(n)}). \tag{B.3}$$

This iterative process keeps refining the MLE until it converges.

To prove BEESEM is an EM algorithm, we only need to prove that

$$Q(\theta; \hat{\theta}^{(n)}) \propto -f^{(n)}(\theta), \tag{B.4}$$

where $f^{(n)}(\theta)$ is the objective function of BEESEM (see Equation (A.21)). For simplicity, we assume $\mu \to -\infty$ in the following discussions, although similar arguments can be made in more general cases. The assumption implies that $w_{ij} = 1$, for any *i* or *j*. Thus according to Equation (A.15) and Equation (A.17), we have

$$
\begin{aligned}
R_j &= \frac{P(T \cdot s_j \mid B)}{\sum_{i=1}^{n} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^k} \\
&= \frac{1}{\sum_{i=1}^{n} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^k} \sum_{i=1}^{n} P(S_i \mid B) \sum_{k=1}^{2(l-m+1)} I_{ij}^k \frac{e^{-E(T \cdot s_j)}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}} \\
&= \frac{1}{\sum_{i=1}^{n} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^k} \sum_{i=1}^{n} P(S_i \mid B) \sum_{k=1}^{2(l-m+1)} I_{ij}^k \frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}}.
\end{aligned}
\tag{B.5}
$$

To proceed, we define a new variable K_j that is a function of **x** and **Z**:

$$
K_j \equiv \frac{\sum_{i=1}^{n} P(S_i \mid B) \sum_{k=1}^{2(l-m+1)} I_{ij}^k Z_{ik}}{\sum_{i=1}^{n} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^k}. \tag{B.6}
$$

In the definition, the denominator represents the prior sequence count of s_j, and the numerator represents the number of actual binding sites that have sequence s_j. Then we assume that the conditional probability of **x** and **Z** given θ has the following form:

$$
P(\mathbf{x}, \mathbf{Z} \mid \theta) \propto \prod_j \exp\left\{ -\frac{[K_j - R_j(\theta)]^2}{2\sigma^2} \right\}, \tag{B.7}
$$

where $R_j(\theta)$ is defined in Equation (A.19) and σ is a constant. Or equivalently, we have

$$
l(\theta; \mathbf{x}, \mathbf{Z}) \propto -\sum_j [K_j - R_j(\theta)]^2. \tag{B.8}
$$

The above equation implies that

$$
\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}}\, l(\theta; \mathbf{x}, \mathbf{Z}) \propto -\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} \sum_j [K_j - R_j(\theta)]^2. \tag{B.9}
$$

To proceed, we make the following assumption:

$$
\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} \sum_j [K_j - R_j(\theta)]^2 \approx \sum_j [\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}}\, K_j - R_j(\theta)]^2. \tag{B.10}
$$

To compute $\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}}\, K_j$, we need to make use of the fact that

$$\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} Z_{ik} = P(Z_{ik} = 1 | \mathbf{x}, \hat{\theta}^{(n)}) = \frac{e^{-E_i^k}}{\sum_{q=1}^{2(l-m+1)} e^{-E_i^q}}. \tag{B.11}$$

Thus according to the definition of K_j, we have

$$\begin{aligned} \mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} K_j &= \mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} \left[\frac{P(S_i | B) \sum_{k=1}^{2(l-m+1)} I_{ij}^k Z_{ik}}{\sum_{i=1}^{n} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^k} \right] \\ &= \frac{P(S_i | B) \sum_{k=1}^{2(l-m+1)} I_{ij}^k \mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} Z_{ik}}{\sum_{i=1}^{n} P(S_i) \sum_{k=1}^{2(l-m+1)} I_{ij}^k} \\ &= R_j. \end{aligned} \tag{B.12}$$

The above equation implies that

$$\sum_j [\mathrm{E}_{\mathbf{Z}|\mathbf{x},\hat{\theta}^{(n)}} K_j - R_j(\theta)]^2 = \sum_j [R_j - R_j(\theta)]^2. \tag{B.13}$$

As a result, we can finally prove that our objective function is essentially the same as $Q(\theta; \hat{\theta}^{(n)})$. In fact, their difference is only a negative scaling factor. Thus BEESEM is indeed an EM algorithm.