1-1-2011

Quantitative Analysis Demonstrates Most Transcription Factors Require only Simple Models of Specificity

Yue Zhao
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Program in Computational Biology

Thesis Examination Committee:
Gary D. Stormo, Chair
Michael R. Brent
Barak A. Cohen
Joseph C. Corbo
James J. Havranek
Ting Wang

Quantitative Analysis Demonstrates Most Transcription Factors Require only Simple Models of Specificity

by

Yue Zhao

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

August 2011
Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Quantitative Analysis Demonstrates Most Transcription Factors Require only Simple Models of Specificity

by

Yue Zhao

Doctor of Philosophy in Biology and Biomedical Sciences

Washington University in St. Louis, 2011

Professor Gary D. Stormo, Chair

Organisms must control their gene expression to properly respond to developmental, stress or other environmental cues. A key part of this process is transcriptional regulation, which is largely accomplished by a complex network of transcription factor proteins (TFs) interact with their specific binding sites in the genome. Understanding how TFs select correct binding sites out of the vast number of potential binding sites in the genome is a key challenge in molecular biology. Recently, unprecedented amount of quantitative binding data have become available as results of developments in high-throughput experimental techniques. However, interpretation of high-throughput binding data has proved to be controversial, largely due to the lack of physically principled data analysis methods.

An important question in the analysis of binding data is the complexity of the specificity model needed. This has important implications for both the characterization of specificity and for the prediction of the consequences of mutations. Structurally, TF-DNA interactions are complex with a wide variety of interactions between the protein and DNA making a simple recognition code impossible. Energetically, however, the situation may be much simpler. Detailed studies of a handful of TFs have shown that individual base pairs often contribute independently to the total binding energy. This view of simplicity has been challenged by data from high-throughput binding experiments, although the extent to which the sample model breaks down is uncertain due to lack of rigorous analysis methods.

The goal of this thesis is to assess the complexity of model required to accurately represent TF specificity. To this end, I have developed a new statistical analysis method BEEML (Binding Energy Estimation by Maximum Likelihood) that parameterizes models of TF specificity from high-throughput quantitative binding data, using a realistic biophysical model. Employing the BEEML method, I show that the energetics of most TF-DNA interactions are simple, with bases in the binding site contribute approximately independently to the total binding energy. Further, I show that interactions in the binding site occur mostly between adjacent positions.

ACKNOWLEDGEMENTS

First and foremost I want to thank my advisor Gary Stormo. Not only did he give me an opportunity to work in his group but he also give me many great ideas and discussions regarding both the studies of protein-DNA interaction and quantitative data analysis. I have benefited not only from his deep knowledge of the subjects but also his patience as I find my own way in research.

All Members of the Stormo lab have been supportive and encouraging from day one. I want to thank David Granas, Ryan Christensen, Aaron Spivak and Nnamdi Ihuegbu for advices and interesting discussions over the years. I also want to thank Barak Cohen for being a mentor. I learned much about how think about scientific problems critically from him, especially during my prelim exams.

Finally I want to thank my family for their support. I especially want to thank my wife Hien. What I have accomplished is only possible with her love and support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Proteins, such as many transcription factors, that bind to specific DNA sequences are essential for the regulation of gene expression. Identifying the specific sequences that each factor binds can help us map out transcriptional regulatory networks within cells as well as identify how genetic variation can cause disruption of normal gene expression, which is often associated with disease.

For many years scientists have been measuring the binding affinity of TFs to specific DNA sequences, but these experiments are low throughput, each experiment determining the affinity of the TF to a single DNA sequence. Moreover, *in vivo* the affinity of the TF is not as crucial as its specificity. Inside a bacterial cell or a eukaryotic nucleus, the concentration of potential binding sites is so high (typically in the millimolar range) that TFs will essentially always be bound to DNA, even if there are no high-affinity sites. Binding to their regulatory sites requires the TFs to pick out the target sites from the vast number of potential binding sites in the genome. The information required for understanding and modeling the regulatory network is not the affinity to the preferred binding sites but the differences in binding affinity for all of the potential binding sites, which is referred to as specificity of the TF. Because the length of a typical binding site is usually about 6–10 base pairs (and can be much longer for some TFs that bind as dimers), it is usually not possible to directly measure the affinity to all potential binding sites, of which there are 4^L for an L-long site. When direct measurement is not possible, recently developed high-throughput

experimental methods (Bulyk et al., 2001; Meng et al. 2005; Berger et al., 2006; Maerkl & Quake, 2007; Zhao et al., 2009; Zykovich et al., 2009) can provide enough data for models of specificity to be construtured.

Models are desirable even when the binding site is short enough that direct affinity measurements of all possible binding sites can be made. First, complexity of the model required for a good fit to the data can provide insight into the recognition mechanism. Second, the model parameters are averaged over many independent measurements and can reduce the uncertainty for any particular sequence compared to the raw data, which tend to be noisier. Third, simple models lend themselves to predicting the effects of variants, both in the binding sites and also in the protein itself, and can facilitate the design of proteins with novel specificity. Fourth, a simple model provides an easy method for scanning a genome and predicting the most likely binding sites as well as the effects of genetic variations.

Structurally, TF-DNA interactions are complex, with a wide variety of interactions between the protein and DNA. As many have pointed out, a simple TF-DNA recognition code does not exist (Matthews, 1988; Mandel-Gutfreund et al., 1995; Pabo & Nekludova, 2000; Luscombe & Thornton, 2002). However, the energetics of the situation appears to be simple, with individual base pairs often contributing approximately independently to the total binding energy. Although deviations from strict independence are common, the non-independent contributions tend to be of smaller magnitude compared to the independent contributions. This allows for simple models of interactions, such as position weight matrices (PWM, Stormo, 2000), to be good approximations to the true binding energies. The physical intuition is that TF-DNA recognition is primarily based on complementarity

between the sequence dependent positioning of hydrogen bond donors and acceptors in the grooves of the double helix and those on surface of the amino acid side chains of the TF. Since most mutations change the shape of this network of hydrogen bond donors and acceptors locally, their effects are also local.

Historically, this view of simplicity was supported by detailed studies of a handful of TFs (Betz et al., 1986; Sarai & Takeda, 1989; Takeda et al., 1989; Fields et al., 1997). Recent advances in high-through analysis of protein-DNA interactions have greatly expanded the knowledge of the specificity of individual TFs (Stormo & Zhao, 2010). The Protein Binding Microarrays (PBM, Bulyk et al., 2001; Berger et al., 2006) approach, in particular, has been used to generate binding data for hundreds of TFs.

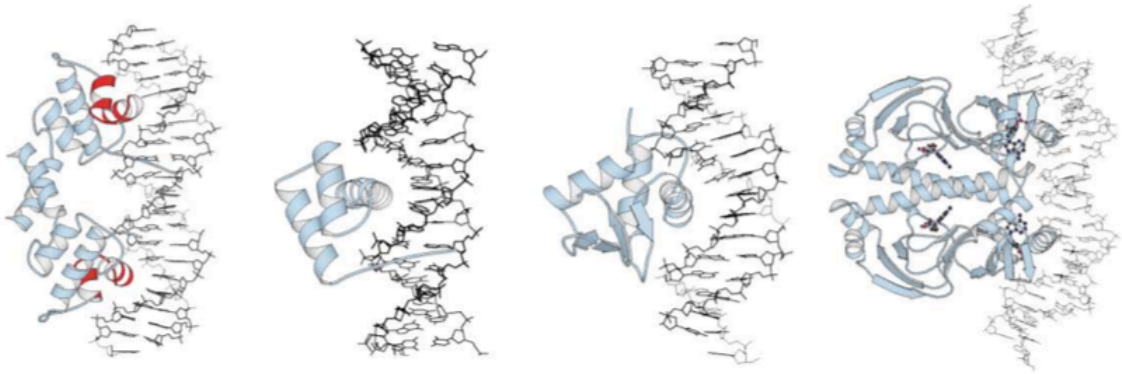
Recently, a large scale PBM study of mouse TFs (Badis et al., 2009) concluded that TF-DNA recognition is highly diverse and complex: 41 out of the 104 TFs studied had clear secondary binding preferences not captured by a single PWM and 89 out of 104 TFs can be better represented by a linear combination of multiple PWMs. However, the authors did not take into account the expected improvement in fit from the additional parameters required for more complex models. Further more, they used three different methods to obtain PWMs and each method was superior to the others on some datasets, indicating that none of the methods can be optimal at determining the PWM parameters. It is possible that the insufficiency of PWMs observed is not due to the complexity of TF-DNA recognition, but rather the algorithms used for parameter estimation.

The goal of this thesis is to assess the complexity of model required to accurately represent TF specificity. To this end, I have developed a new statistical analysis method

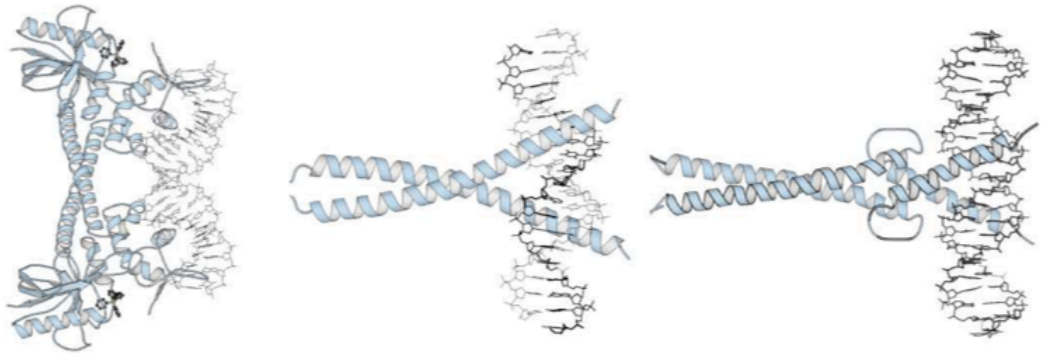
BEEML (Binding Energy Estimation by Maximum Likelihood) that parameterizes models of TF specificity from high-throughput quantitative binding data using on a realistic biophysical model. Employing the BEEML method, I show that the energetics of most TF-DNA interactions are simple, with each base in the binding site contributes approximately independently to the total binding energy. Further, I show that pairwise interactions not captured by the PWM occur mostly between adjacent positions in the binding site. This simplicity has important implications for our understanding of the molecular basis of TF specificity and demonstrates the importance of the analysis method in the interpretation of high-throughput data.

1.1 Biochemistry of TF Binding Specificity

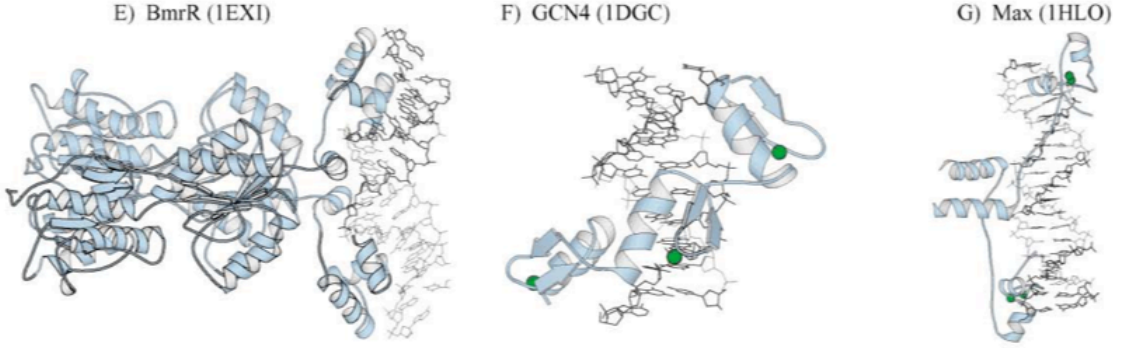
Our understanding of the detailed mechanisms of TFs specificity has mostly come from the study of crystal structures. As of April 2011, there are more than 1600 high-resolution (< 3Å) structures of protein-DNA complexes in the Protein Data Bank (Rose et al., 2011). These structures have shown us that TFs use a variety of folds to recognize specific DNA sequences. A number of classification schemes have been used to categorize the different folds (Harrison, 1991; Luscombe et al., 2000), the most recent of which identified more than 30 different families of DNA binding folds in TFs. Analysis of genomes have shown that some folds, such as C2H2 zinc finger or homeodomain, are very common while other folds occur very rarely or only in a restricted sets of organisms (Garvie & Wolberger, 2001). Figure 1.1 shows the structures of some representative DNA binding domains.



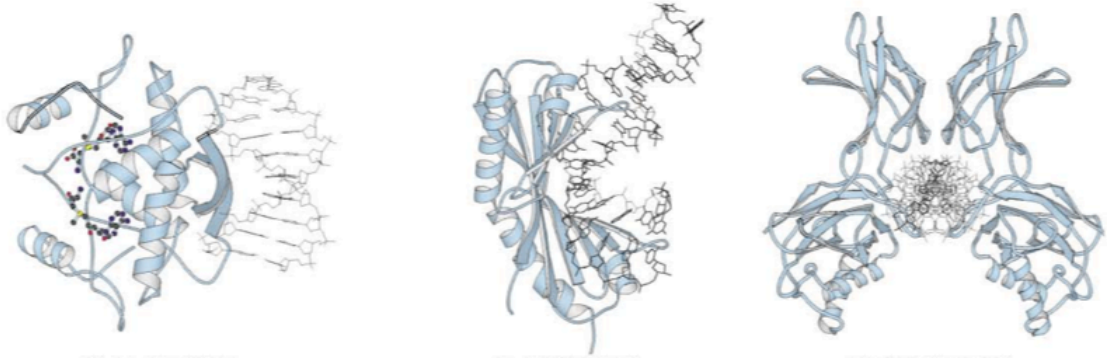
A) λ repressor (1LMB) B) engrailed (1HDD) C) PU.1 (1PUE) D) CAP (1CGP)



E) BmrR (1EXI) F) GCN4 (1DGC) G) Max (1HLO)



H) PurR (1PNR) I) Zif268 (1ZAA) J) GAL4 (1D66)



K) MetJ (1CMA) L) TBP (1YTB) M) NFκB (1SVC)

Figure 1.1 Structures of DNA binding folds demonstrating the mechanisms of TF specificity, obtained from (Garvie & Wolberger, 2001). Structural family of the TFs are: A) bacterial helix-turn-helix B) homeodomain C) winged helix-turn-helix, ETS domain D) helix-turn-helix E) unclassified F) basic leucine zipper G) basic helix-loop-helix H) LacI I) zinc finger J) zinc domain, GAL4 type K) β sheet recognition L) TATA binding protein M) Rel homology domain

The α helix is the most common protein structural element used for base recognition, typically through contacts in the major groove. Early analysis based on model building (Church et al., 1977) showed that proportions of the α helix were ideal for presenting amino acid side chains for interaction with bases in the major groove of B-DNA. However, many TFs contain flexible N- or C-terminal tails that are unstructured in the absence of the DNA but bind in one of the DNA grooves. For example, λ repressor has an N-terminal arm that contacts bases in the major groove (Jordan & Pabo, 1988) and homeodomain proteins have N-terminal arms that dock in the minor groove (Gehring et al., 1994). Examples of TFs using β sheets or loops also exist, although their use is not nearly as prevalent as the use of α helixes.

In addition to sequence-specific interactions mediated by hydrogen bonding between DNA bases and protein side chains, electrostatics driven non-specific interactions between phosphate backbone of DNA and positively charged amino acids of the protein are also crucial for the proper functioning of the transcriptional regulatory system (von Hippel, 2007). For example, non-specific binding to the genome decreases the amount of TF that is free in solution and thus is directly available to support the specific binding equilibrium.

Non-specific binding is also important for the kinetics of specific binding. In 1970, Riggs et al. (Riggs et al., 1970) reported that lac repressor is able to find its target site at a rate much faster than predicted by 3D-diffusion. This implies non-specific binding to the genome increases the rate of target site location by the repressor. In a series of papers, Berg, Winter and von Hippel (Berg et al., 1981; Winter et al., 1981; Winter & von Hippel, 1981) proposed three mechanisms for how lac repressor locates its operator that all depended on the "facilitated transfer" of TF to target site by one dimensional diffusion while in the non-specifically complexed state: sliding, intersegment transfer, and hopping. All three mechanisms were later observed to contribute to the kinetics of specific target site location. Sliding mechanism is dominant at low salt concentration on naked DNA. However, hopping and intersegmental transfer becomes more important as DNA is covered by proteins, as is likely the case *in vivo*.

While structural studies have provided us with many insights into the mechanism of TF-DNA interaction, it is currently not possible to accurately predict TF binding specificity based on this knowledge. For practical tasks such as predicting the regulatory targets of TFs or predicting the consequence of mutations in the binding site, statistical models based on thermodynamics properties of specific TF-DNA recognition must be used.

The bimolecular interaction between TF and a particular DNA binding sequence, S_i, is governed by two rate constants, k_{on} for the formation of the complex, and k_{off} for the dissociation:

$$assosciation: TF + S_i \xrightarrow{K_{on}} TF \cdot S_i$$

$$dissociation: TF \cdot S_i \xrightarrow[K_{off}]{} TF + S_i$$

The equilibrium binding constant K_i of the TF to the site S_i is:

$$K_i = \frac{K_{on}}{K_{off}} = \frac{[TF \cdot S]}{[TF][S]}$$

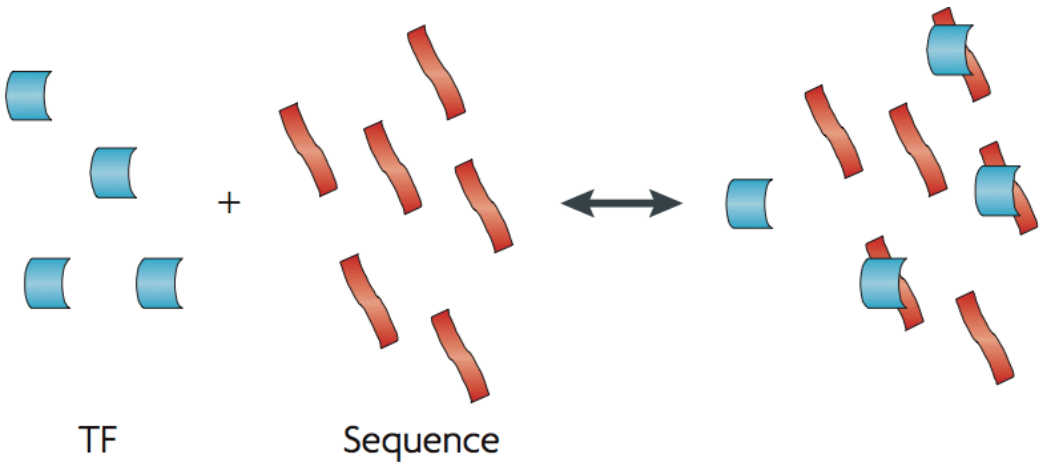
where square brackets indicate concentration. At a specific instant, S_i can be in two possible states, bound or free (denoted by s = 1 or s = 0). The probability of TF binding to sequence S_i is:

$$P(s = 1 \mid S_i) = \frac{[TF \cdot S_i]}{[TF \cdot S_i] + [S_i]} = \frac{1}{1 + \frac{1}{K_i[TF]}} = \frac{1}{1 + e^{E_i - \mu}} \tag{1.1}$$

where $E_i \equiv -\ln K_i$ is the Gibbs free energy of binding (often referred to as ΔG_i), in units of RT (R is the gas constant and T the temperature in Kelvin) and $\mu = \ln[TF]$ is the chemical potential.

Experimentally, the value of E_i can be determined by measuring the fractional occupancy of S_i at several different TF concentrations, then use curve fitting algorithms to determine to the value of E_i that best fit the data according to equation 1.1. An example of this is shown in figure 1.2. Common approaches used to measure binding probabilities are

Electrophoresis mobility shift assay (EMSA, Fried & Crothers, 1981; Garner & Revzin, 1981) and DNase footprinting assay (Galas & Schmitz, 1978). These types of experiments are slow and laborious since TF affinity for each sequence must be measured separately. Attempts have been made to use in vitro affinity data to predict in vivo TF binding patterns (Liu & Clarke, 2002). However, accuracy of the analysis was limited by the fact that only 44 affinity measurements could be made due to the use of low-throughput EMSA method.



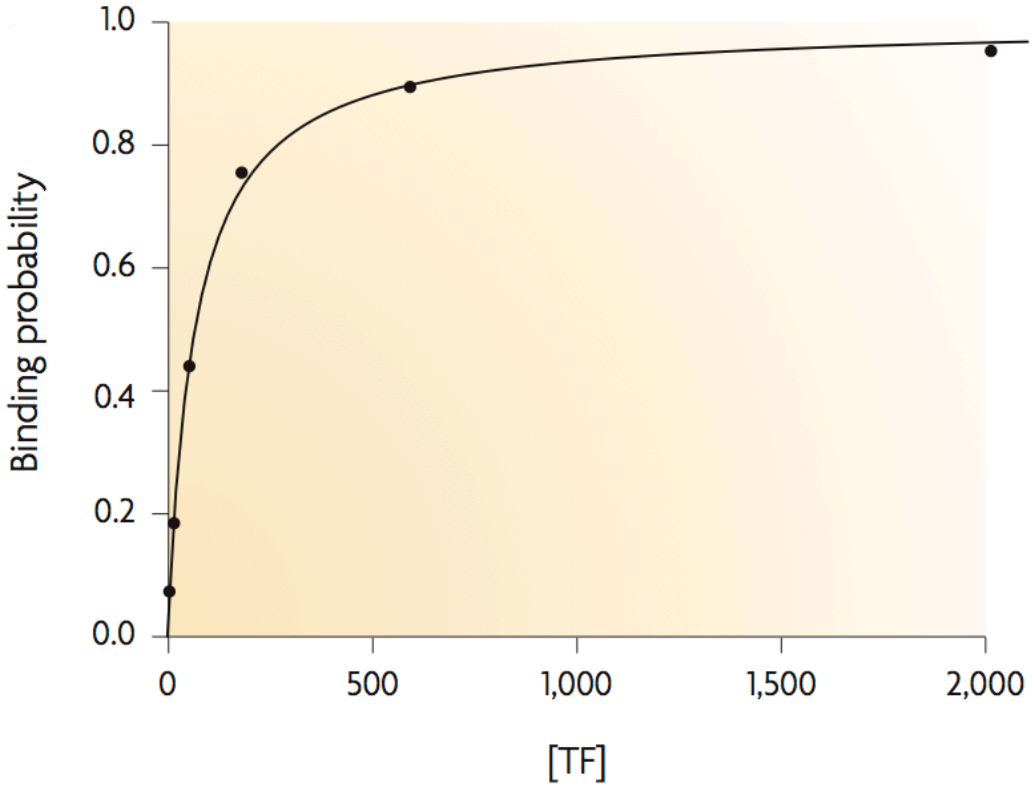
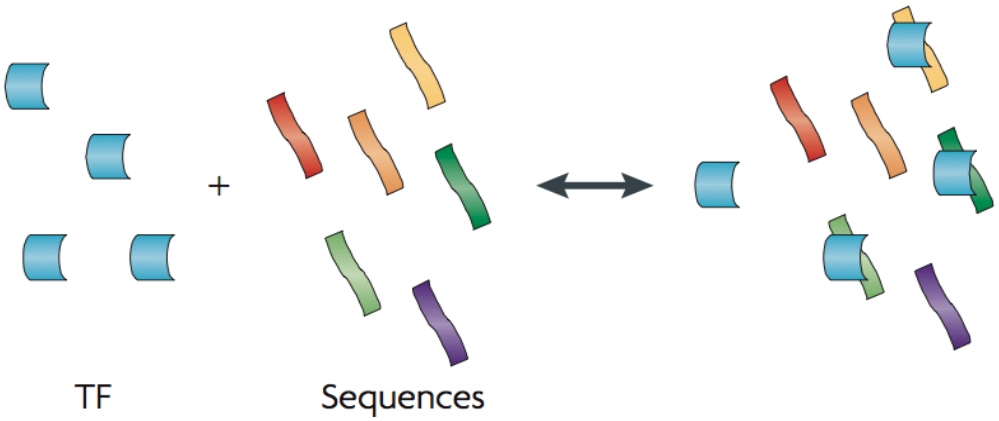


Figure 1.2 Affinity of TF for a particular sequence can be determined from isothermal binding curve, taken from (Stormo & Zhao, 2010)

Binding affinity for a particular sequence is useful if one wishes to determine if TF will be bond or free at a particular concentration. However, *in vivo* TF is exposed to the entire genome's worth of potential binding sites, the question is not so much whether the TF will be bound or free but which site on the genome it will be bond to. We are interested in the specificity of the TF, or the affinities for all potential binding sites. Experimentally, the major difficulty is one of scale. Direct measurement of TF specificity requires measurement of 4^L affinities for an L long binding site. For example, more than a million measurements must be made to characterize the specificity of a TF that recognizes a 10-long binding site. Even with high-throughput techniques, direct measurement of affinity for all sites is not practical. Instead, recently developed methods can measure binding probabilities of all sequences at a fixed TF concentration. This is illustrated in Figure 1.3: each line parallel to the [TF] axis is equivalent to the plot in figure 1.2. Instead of measuring 4^L binding curves, binding probabilities of TF for all sequences at a fixed [TF] can be measured in a single experiment. This is displayed as a line parallel to the binding energy axis. Appropriate data analysis methods can then be used to build a quantitative model that allows prediction of binding energies for all sequences (Zhao et al., 2009; Zhao & Stormo, 2011).



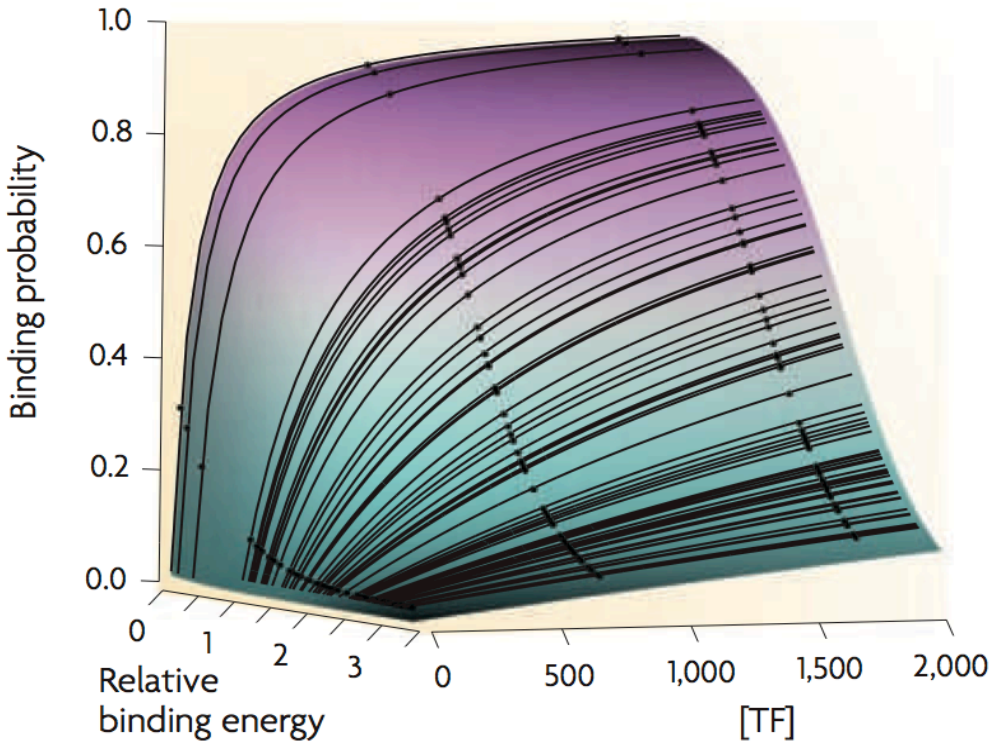


Figure 1.3 TF specificity. Each solid line is an isothermal binding curve (as in figure 1.2). Curves for sequences with different binding energies are plotted together in three dimensions. Lines parallel to the binding energy axis represent the type of data from high-throughput experiments. Taken from (Stormo & Zhao, 2010).

1.2 High-throughput Binding Experiments

Recent advances in microarray and sequencing technologies have made large-scale measurement of TF-DNA interaction possible. The following techniques are currently in use:

Microfluidics.

Mechanically induced trapping of molecular interactions (MITOMI) is a method that uses microfluidic devices to determine binding specificities in a relatively high-throughput manner (figure 1.4A), obtaining the binding affinities of a TF to a few hundred different DNA sites per device (Maerkl & Quake, 2007). TFs are synthesized *in vitro* inside microfluidic chambers, which are lined with antibodies to attach the TF to its surface. The chambers are also seeded with a specific DNA binding site at a specific concentration and with a fluorescent tag. Overall, the device contains hundreds of different DNA binding site sequences, each at multiple different concentrations. The total DNA binding site concentration in each chamber is determined by its fluorescent signal, and any DNA that is not bound to the TF that is attached to the surface is then flushed out. The amount of protein is determined by its fluorescent signal, and the amount of DNA bound to the TF is determined by the remaining DNA signal. By determining the fraction of DNA that is bound at several different DNA concentrations, the relative affinities of each different sequence can be determined. This does not, by itself, determine the absolute dissociation constant (K_d, inverse of K_i) between the TF and the binding site, but comparison to a

reference with known K_d allows determination of the absolute K_d for each binding site sequence.

Protein-binding microarrays.

Protein-binding microarrays (PBMs) are a technology developed in the last 10 years that has greatly increased the throughput for assessing the binding specificities of TFs (Bulyk et al., 2001; Mukherjee et al., 2004; Berger et al., 2006). As with microarrays for gene expression, this technology has made possible large-scale, high-throughput analyses to collect information that previously must be acquired on a gene-by-gene basis (Philippakis et al., 2008; Berger & Bulyk, 2009). The current version of PBM (figure 1.4B) uses arrays that contain 44,000 spots designed such that all possible ten-base-long DNA binding sites once on each array. This means that every eight-base-long sequence occurs 32 times, taking both orientations into account. A TF, either purified from cells or synthesized *in vitro*, is added to the array, which is then washed to remove nonspecific binding. The amount of protein binding to any specific DNA spot is determined with a fluorescent antibody to the protein.

Cognate Site Identifier.

Cognate Site Identifier (CSI) also uses arrayed DNA sequences to measure relative binding by TFs (Warren et al., 2006; Puckett et al., 2007). The major difference between PBMs and CSIs is that in CSIs, single-stranded DNAs are synthesized that fold back to form dsDNA binding sites, thereby eliminating the need for primer directed DNA synthesis to

generate dsDNA, which is required for PBMs (figure 1.4B). Current CSI arrays also include all possible ten-base-long sequences.

In vitro Selection

Using purified proteins to select high-affinity binding sites from random libraries *in vitro* is a very powerful technique. Although invented independently multiple times, the term SELEXX seems to be the most commonly used name (Oliphant et al., 1989; Blackwell & Weintraub, 1990; Tuerk & Gold, 1990; Wright et al., 1991). The general strategy is to create a library of potential binding sites, which may be from randomly synthesized DNA or created from genomic sequences. Both ends of the library sequences can have primer binding sites so that they can be amplified by PCR. Purified TF is added to the library of DNA sites and the bound and unbound sequences are separated by various means, such as gel filtration, filter binding or binding to immobilized protein (figure. 1.4C). Although higher affinity sites have a higher probability of being bound by the TF, after a single selection most of the bound sequences are still low affinity because they greatly exceed the number of high-affinity sequences. To increase the fraction of high-affinity sites, the bound fraction can be amplified and rebound and those steps repeated as many times as needed. Typically, after several rounds, the selected sites would be cloned and sequenced, often obtaining fewer than 100 independent sites (Fields et al., 1997). These methods are capable of determining important aspects of the binding specificity, including the consensus sequence and the relative variability in affinity for different bases at different positions within the binding sites.

Utilizing second generation sequencing technologies it is now possible to derive binding energy profiles from SELEX data efficiently using a method called high-throughput SELEX (Zhao et al., 2009) or bind-and-seq (Zykovich et al., 2009). An advantage of this approach is that the output (the number of counts observed for each sequence) is digital and there is a very large dynamic range. From a total set of hundreds of thousands or millions of individual sequences there will be many nonspecific sites, but they usually only occur once, whereas the highest affinity sites may occur thousands of times. From millions of reads one can estimate the binding model as well as nonspecific binding energies and the free-TF concentration after a single round of selection (Zhao et al., 2009). Performing additional rounds of selection may provide more information about specific segments of the energy distribution and may give more accurate models, particularly for low-specificity TFs or those with large non-independent contributions.

A recent publication (Jolma et al., 2010) has pushed this approach much further. Using tagged proteins the authors performed HT-SELEX from cell extracts (rather than purifying the TF) and by barcoding individual experiments they collected binding sites for several TFs in parallel. In total they obtained binding site data for 19 different TFs, many of which have low specificity and required multiple rounds of selection. This demonstrated that HT-SELEX could have very high throughput and generate enormous amounts of specificity data quite rapidly.

Bacterial one-hybrid selections

Bacterial one-hybrid (B1H) selections are not *in vitro* assays (unlike the methods described above) and can be used for any TF that can be cloned and expressed in *Escherichia coli*; this method has the advantage that the TF need not be purified or synthesized *in vitro* (Meng et al., 2005; Meng & Wolfe, 2006; Noyes et al., 2008). The approach uses a library of randomized binding sites upstream of a weak promoter that drives the expression of a selectable gene, typically the yeast *HIS3* (which encodes a component of the histidine biosynthesis pathway; the *E. coli* strain lacks the bacterial homologue) (figure. 1.4D). When the cells are grown in medium lacking histidine, expression of the *HIS3* gene is required for growth and the stringency of the selection can be modulated by the addition of 3-amino-1,2,4-triazole (3AT), an inhibitor of the *HIS3* enzyme. The TF is fused to the non-essential ω subunit of RNA polymerase, so that TF binding recruits RNA polymerase and increases promoter activity.

In earlier works, the binding sites from selected colonies were sequenced individually, typically obtaining about 50 sequences for each selection. Recently, high-throughput sequencing methods allows one to collect all of the cells on the plate and sequence them all *en masse,* obtaining millions of binding sites for each experiment (Christensen et al., 2011). The high-affinity sites lead to more *HIS3* expression and so allow the cells with those sites to grow the fastest, resulting in a higher number of sequence reads. Therefore, coupling B1H to high-throughput sequencing provides a digital read-out with a very large dynamic range, the highest affinity sites occurring hundreds to thousands of times

and the lowest affinity sites typically occurring only once or not at all. It can also be multiplexed so that the data from several experiments, for different TFs or under different selection stringencies can be obtained in parallel.

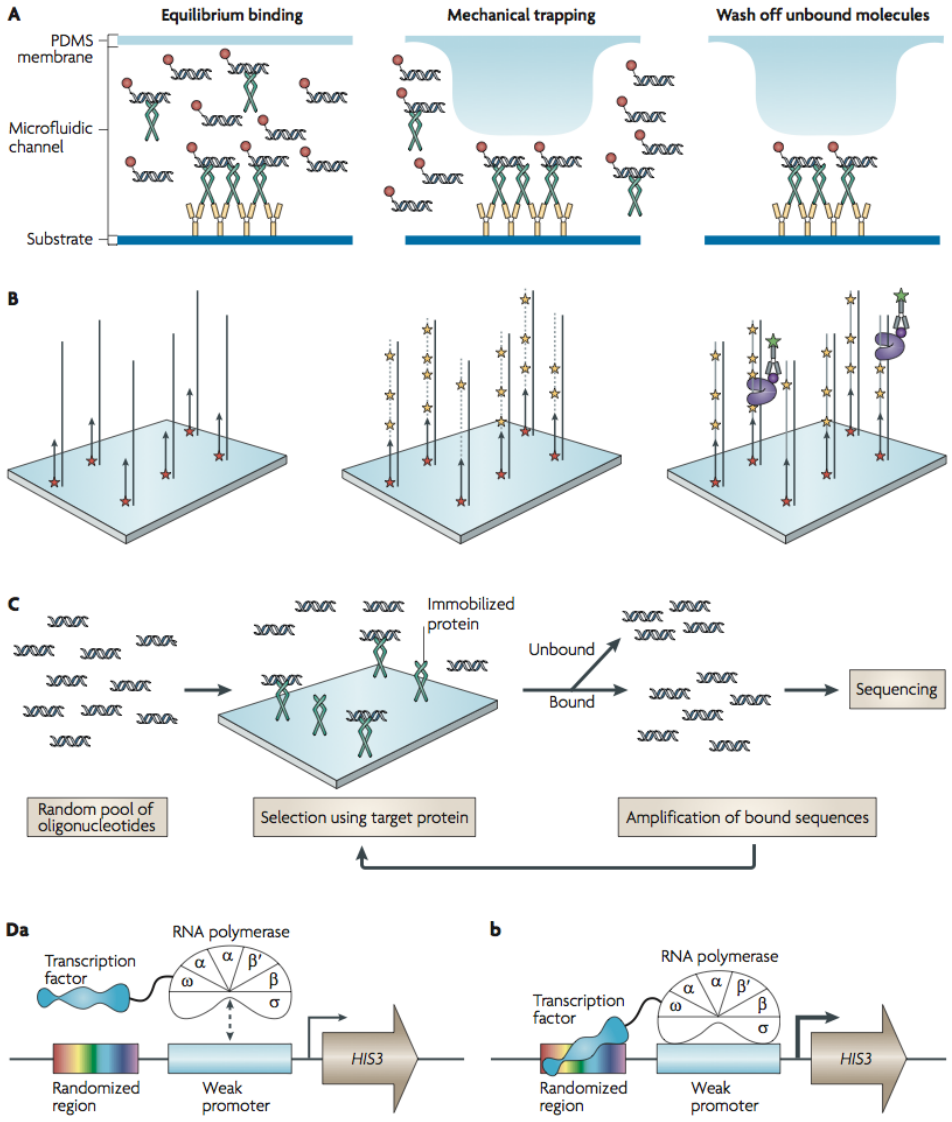


Figure 1.4 A) overview of MITOMI method B) overview of PBM method C) overview of High-throughput SELEX method D) overview of B1H method. Taken from (Stormo & Zhao, 2010)

1.3 Computational Modeling of Specificity

The binding of a TF to its binding site goes hand in hand with a favorable change in the free energy of the system (figure 2.1). The simplest model to represent the specificity of a TF is its consensus sequence: the DNA sequence with the highest affinity for the TF. However, it is not a good model of TF specificity since TFs, unlike restriction enzymes, recognize a set of sequences with varying affinity rather than just a single sequence (Stormo, 2000). A simple generalization of the consensus is the position weight matrix (PWM). In this model there is a score assigned to each possible base at each position in the binding site. The sum of the elements that correspond to a specific sequence gives a total score for that sequence. This allows the model to provide a score to all possible binding sites for the protein. The PWM scores can be interpreted statistically (Stormo & Hartzell, 1989; Lawrence & Reilly, 1990) or biophysically (Stormo et al., 1986; Berg & von Hippel, 1987; Heumann et al., 1994). Since I am mostly concerned with analysis of experimental data, this thesis will focus on the biophysical interpretation, although the statistical view is also very powerful, especially when the amount of data is small as it is able to incorporate phylogenetic conservation (Wang & Stormo, 2003; Sinha et al., 2004; Siddharthan et al., 2005).

In a classic paper (Berg & von Hippel, 1987), Berg and von Hippel introduced a theoretical framework for inferring a model of specificity from a set of functional TF binding sites. This framework has two aspects: 1) the thermodynamics of TF-DNA interaction and 2) a model of evolutionary selection. The central assumption is the existence of a critical energy E_c. Sites that bind weakly, $E_s > E_c$ are assumed to be nonfunctional while

all sites whose energy is less than E_c are assumed to be equally suited to be regulatory sites and therefore equally likely to be present in the dataset. Further assuming individual positions in a site are independent from each other, i.e., seeing a particular base at position m in the site does not affect the probability of observing another base elsewhere, the logics of the classical derivation of the Boltzmann distribution in statistical mechanics (Dill & Bromberg, 2002) can be used to arrive at the conclusion that frequencies of bases in each position of the binding sites follow the Boltzmann distribution. Mathematically, this can be written as:

$$\varepsilon_{b,m} = -\frac{1}{\lambda} ln \frac{n_{b,m}}{n_{o,m}} \tag{1.2}$$

Where $\varepsilon_{b,m}$ is the $\Delta\Delta G$ (see figure 2.1), $n_{b,m}$ is the number of times base b is observed in position m of the binding site and $n_{0,m}$ is the number of times the consensus base is observed in position m. λ is a scaling parameter related to the strength of selection.

From a biophysical point of view, the invocation of an evolutionary argument is unsatisfactory, especially for the interpretation of *in vitro* binding experiments. Making a different assumption, Heumann et al. (Heumann et al., 1994) derived the Boltzmann distribution in the limiting case of low TF concentration, but in a purely biophysical framework. Methods following the same logic have been developed to infer TF specificity from gene expression (Bussemaker et al., 2001) or quantitative *in vivo* binding data (Foat et al., 2006; Tanay, 2006). Another biophysical approach is to assume the low temperature

limit, where sites are either unoccupied or saturated (Djordjevic et al., 2003). In this case, the inference of TF specificity reduces to the problem of finding a classifier that maximally separates bound and unbound sequences. However, the inferred energies can still only be determined up to an arbitrary scaling parameter.

All of the biophysically motivated methods discussed above, as well as the vast majority of statistically based motif finding algorithms assume that positions within the binding site are mutually independent. Although the independent assumption have been repeatedly challenged (Wolfe et al., 2000; Man & Stormo, 2001; Bulyk et al., 2002; Maerkl & Quake, 2007; Badis et al., 2009), it is possible that PWM is still a good enough approximation of the "true" TF-DNA interaction model (Benos et al., 2002; Stormo & Zhao, 2007; Zhao & Stormo, 2011). On the one hand, a more complicated model will fit the observed data better. On the other hand, a more complicated model may overfit the data and perform poorly on data it has not seen before. Many existing computational methods can model position dependencies, using techniques such as Bayesian Networks (Barash et al., 2003), generalized weight matrix models (Zhou & Liu, 2004), permuted Markov models (Zhao et al., 2005) or Markov Networks (Sharon et al., 2008). Although these models are very powerful, they are also complex which means the large number of parameters required may not be supportable by available data. Zhou and Liu (Zhou & Liu, 2004) conducted a statistically rigorous analysis of the known TF binding sites in TRANSFAC database (Wingender et al., 2000). They searched for correlated position pairs and found that 25% of the data have statistically significant correlated positions.

With the availability of high-throughput TF-DNA interaction data, not only can one examine the fit of complex specificity models, it is also possible to avoid making any assumptions about TF concentration or temperature and estimate TF specificity and concentration directly from data. While the framework for analysis exists(Liu & Clarke, 2002; Granek & Clarke, 2005; Segal et al., 2008), computational methods that use this framework to analyze high-throughput binding data have not been developed.

In this thesis, I will describe methods appropriate to high-throughput SELEX as well as PBM data. I will use these methods to analyze the performance of PWM as well as more complex models that consider pairwise interactions to show that PWM is a good approximation for most TFs.

Chapter 2
Methods

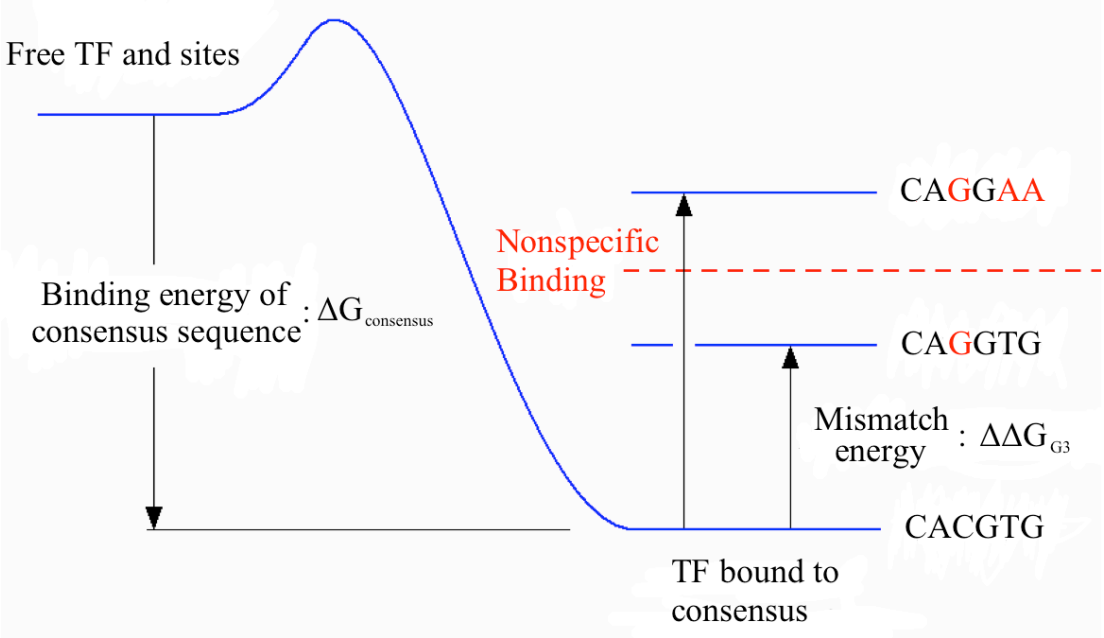


Figure 2.1 Energetics of TF-DNA recognition

The specificity model of a TF can be thought of as a function relating binding energy and the sequence: $E(S_i)$. I will use the term binding energy to refer to $\Delta\Delta G$, the difference between Gibbs free energy of complex formation for sequence S_i and the consensus sequence (see figure 2.1):

$$E(S_i) = \Delta\Delta G_i = \Delta G_i - \Delta G_{consenus}$$

Equation 1.1 can be re-written as:

$$P(s = 1 \mid S_i) = \frac{1}{1 + e^{\mathrm{E}(S_i) - \mu}}$$

(2.1)

where $\mu = \ln[TF] + \Delta G_{consensus}$

This is necessary because absolute binding energies (ΔG's) cannot be determined from the types of experiments analyzed in this thesis. The notation change amounts to defining the ground state to be the state where TF is bound to the consensus sequence, instead of the conventional ground state where both TF and sites are free (see figure 2.1).

2.1 Binding Energy Estimation from Aligned Sites

Model Specification

The binding energy can be decomposed into two, or more, modes of binding (Gerland et al., 2002; von Hippel, 2007). In the following analysis we assume two modes, non-specific binding that is independent of the sequence, and specific binding that varies with different sequences such that

$$e^{-E(S_i)} = e^{-E_{sp}(S_i)} + e^{-E_{ns}}$$

(2.2)

The specific binding component, $E_{sp}(S_i)$, could be a complex function of the sequence, perhaps even composed of multiple modes of binding. But for most of the analysis in this thesis I assume a simple additive energy function that can be represented as a

position weight matrix (PWM). This model requires an energy contribution, $\varepsilon(b,m)$, for each base, b, at position, m, in the binding site such that:

$$E_{sp}(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} \varepsilon(b,m) S_i(b,m) \tag{2.3}$$

where S_i(b,m) is an indicator variable with S_i(b,m) = 1 if base b occurs at position m of sequence S_i and S_i(b,m) = 0 otherwise. This model can be easily expanded to accommodate pairwise interactions between positions:

$$E_{sp}(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} e(b,m) S_i(b,m) + \sum_{m} \sum_{n} \sum_{b=A}^{T} \sum_{c=A}^{T} e(b,m,c,n) S_i(b,m,c,n) \tag{2.4}$$

where e(b,m,c,n) is the energy contribution of having base b at position AND base c at position n.

We derived equation (2.1) by considering a simple experiment where only a single sequence, S_i, is available for binding. It also holds true in the more general case where there are many different sequences all competing for binding to the TF. However, the interpretation μ is different between the simple and general case. In the simple experiment, TF not bound to S_i are simply free in solution, so $\mu = \ln[TF] + \Delta G_{consensus}$. In the general case, TF not bound to S_i could be bound to any of the other available sequences, so μ

corresponds to a free energy for the collection of all of the states with the TF not bound to S_i and $\Delta G_{consensus}$. We present an alternative derivation of equation (2.1) to further illustrate this point. Consider that at any given time a particular sequence, S_i, can be in one of three possible states: bound to the TF in the specific binding mode ($s=1_{sp}$); bound to the TF in the non-specific binding mode ($s=1_{ns}$); unbound by the TF ($s=0$). At equilibrium the probability of being in each state is determined by the energy of that state according to the Boltzmann distribution:

$$P(s = 1_{sp}|S_i) = \frac{e^{-E_{sp}(S_i)}}{e^{-\mu} + e^{-E_{sp}(S_i)} + e^{-E_{ns}}}$$

$$P(s = 1_{ns}|S_i) = \frac{e^{-E_{ns}}}{e^{-\mu} + e^{-E_{sp}(S_i)} + e^{-E_{ns}}}$$

$$P(s = 0|S_i) = \frac{e^{-\mu}}{e^{-\mu} + e^{-E_{sp}(S_i)} + e^{-E_{ns}}}$$

(2.5)

The overall probability of the sequence being bound ($s=1$) is the sum of the specific and non-specific binding probabilities. Using equations (2.2) and (2.5):

$$P(s = 1|S_i) = \frac{e^{-E_i}}{e^{-\mu} + e^{-E_i}}$$

(2.6)

which is equivalent to equation (2.1) but now for the general case of many sequences competing for the same pool of TF. It is worth noting that equation (2.6) simplifies if the TF is at very low concentration ($\mu \to -\infty$) then the probability of binding is directly proportional to the binding affinity (though the binding probability approaches 0). This simplification is the basis of the traditional log-odds model.

A HT-SELEX experiment can be modeled as a binding reaction with a pool of TF molecules and a large pool of different sequences, S_i ($1 \leq i \leq 4^L$ for the list of all possible sequences of length L), and with each sequence in proportion $P(S_i)$, which can be determined with high-throughput sequencing. At equilibrium the TF molecules are extracted from the reaction along with the DNA sequences bound to them. The bound DNA sequences are subjected to high-throughput sequencing to obtain a large collection of binding sites, with the proportion of each sequence being $P(S_i | s=1)$, which can be related to equation (2.1) using Bayes' rule:

$$P(S_i|s=1) = \frac{P(s=1|S_i)P(S_i)}{\sum_{j=1}^{4^L} P(s=1|S_j)P(S_j)} = \frac{\frac{e^{-E_i}}{e^{-\mu}+e^{-E_i}}P(S_i)}{\sum_{j=1}^{4^L}\frac{e^{-E_j}}{e^{-\mu}+e^{-E_j}}P(S_j)} \tag{2.7}$$

Given a large enough sample of binding sites this experimental procedure could provide good estimates of the binding free energy for each sequence in the initial pool. However, for typical lengths L and typical differences in binding energy this would require an extremely large number of binding sites, more than available even from current high-

throughput sequencing methods. By employing a model for the binding energy, such as equation (2.3), we can infer binding energies for sequences with limited or inaccurate measurements. Furthermore, having a model for the sequence dependence of the binding energy, instead of just a list of binding energies to different sequences, can be useful in understanding the physical interaction of the protein with the DNA and can facilitate the prediction of changes in binding energies for variant proteins (Benos et al., 2002).

Equation (2.1) was used by Djordjevic et al (Djordjevic et al., 2003) as the starting point in the development of their QPMEME method. However, QPMEME makes the additional assumption that all observed sequences are bound with probability close to 1 (the zero temperature approximation) which prevents it from making use of the quantitative data generated by the HT-SELEX method in which many of the observed sites after one round of selection have low, even non-specific, binding affinity. A direct comparison with our approach is not possible because QPMEME fails to find a solution on datasets containing many low affinity sequences. The TRAP algorithm (Roider et al., 2007) used an equivalent model to estimating total occupancy in ChIP-chip experiments. TRAP assumes the specific energy model (PWM) is known and only estimated μ from the data, whereas we attempt to learn both the energy model and μ simultaneously.

This completes the description of the model. By substituting equation (2.3 or 2.4) into equation (2.2), and that into equation (2.7), we obtain the relationship between the statistics of observed binding sites, $P(S_i|s=1)$, and the binding energy E_i of each sequence S_i.

Maximum likelihood parameter estimation

Given a collection of N bound sequences, we model the relationship between N_i, the number of occurrences of each sequence S_i in this collection, and $\tilde{N}_i = N\ P(S_i \mid s=1)$, the number of occurrences of S_i predicted by the model, as:

$$N_i = \widetilde{N}_i + \epsilon \tag{2.8}$$

where ϵ is a measurement error due to sequencing error as well as the stochastic nature of the sampling. For simplicity we assume ε is a zero-mean Gaussian random variable with standard deviation σ, although other error models are possible (Kinney et al., 2007). For any set of parameters $\theta = \{PWM, \mu, E_{ns}\}$, the likelihood function, or the probability of the data given parameters is:

$$P(data \mid \theta) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\widetilde{N}_i - N_i)^2}{2\sigma^2}} \tag{2.9}$$

Maximizing the likelihood function (2.9) with respect to θ is equivalent to minimizing the negative log of the likelihood function. Dropping terms not dependent on θ, we obtain maximum likelihood estimate of θ by minimizing objective function:

$$\sum_i \left(\widetilde{N}_i - N_i\right)^2 \tag{2.10}$$

This is a non-linear parameter estimation problem and we minimize equation (2.10) using the Levenberg-Marquardt algorithm implemented in minpack (More, 1977). A practical issue is the calculation of the denominator of equation (2.7), the partition function. For longer values of L the naive approach of enumerating over all sequences becomes too computationally expensive. We deal with this situation by rewriting equation (2.7) as:

$$P(S_i|s=1) = \frac{P(s=1|S_i)P(S_i)}{\sum_j P\left(s=1\middle|E_j\right)P(E_j)} \tag{2.11}$$

where E_j is a particular energy level. Instead of summing over all 4L sequences, equation (2.11) allows us to sum over a user-defined number of energy levels (default is 16384) with some loss of accuracy due to the discretization. This does not solve the problem by itself, merely shifts it from enumerating all sequences to the calculation of the energy distribution $P(E_i)$. The naive method of calculating $P(E_j)$ is to compute binding energy for all sequences and $P(E_j)$ is simply the fraction of sequences having energy level E_j.

A more efficient method is possible under the PWM energy model by taking advantage of its independence structure. Each position in the PWM can represented by a probability generating function, possibly with coefficients to account for unequal priors. The distribution of energies defined by the entire PWM is obtained by multiplying the generating functions for each position (Staden, 1989). This polynomial multiplication can be performed very efficiently with a Fast Fourier Transform (FFT) (Cormen et al., 1990). By default, I use FFT approximation when length of binding sites is greater than 10. The algorithm described

above is implemented in an R (R Core Development Team, 2011) program called BEEML (Binding Energy Estimates using Maximum Likelihood).

2.2 Binding energy estimation from PBMs

Calculation of binding probability

Equilibrium binding probability of a site s calculated according to equation (2.1). Since many TFs binds to palindromic binding sites, binding probability to a position j of the probe, with sequence $S_{i,}$ is calculated as:

$$P(j) = P(S_i) + \big(1 - P(S_i)\big)P(\overline{S_i}) \tag{2.12}$$

This accounts for the fact that TF cannot bind to both strands $(S_i, \overline{S_i})$ at the same position of the probe at the same time. This is avoids double counting for TFs with palindromic binding sites. For computational simplicity we ignore the case of overlapping binding sites because for most PWMs, and typical values of μ, it is very unlikely that multiple good binding sites occur in an overlapping fashion on the de Bruijn sequence used to generate PBM probe sequences.

Estimation of position effect

The position of the binding site within a probe significantly influence the signal intensity, with binding sites located further away from the glass slides giving stronger signal

(figure 2.2). We estimate this position effect empirically using top n (n=25 by default) 8mers with highest median intensities. We assume the intensity of probes containing these 8mers are entirely due to the presence of these 8mers and variations in the intensities of probes containing the same 8mer is due to differences in the distance of 8mer occurrence from the glass slide. We find that false positives are usually not a problem if we limit ourselves to the top 25 8mers.

The effect of position j, $F_{pos}(j)$, is estimated as:

$$F_{pos}(j) = \frac{\langle \frac{I_{avg}(S_{i,j})}{I_{avg}(S_i)} \rangle}{\sum_{k=1}^{L} \langle \frac{I_{avg}(S_{i,k})}{I_{avg}(S_i)} \rangle} \tag{2.13}$$

where L is the length of variable region on the probe, $I_{avg}(S_{i,k})$ is the average intensity of probes containing sequence S_i in position k of the probe and $I_{avg}(S_i)$ is the average intensity of all probes containing sequence S_i, in any position. The angled brackets denote averaging over top n 8mers.

The position effect for PBM experiment of mouse TF Plagl1 (pleomorphic adenoma gene-like 1) as estimated by equation 2.13 is shown in figure 2.2. The pattern shown is typical of PBM experiments. The linear loss of signal as binding site location is moved closer to the glass slide is probably due to surface effect as well as crowding. There are a number of possible explanations for the drastic loss of signal at the top of the probe: it is possible

primer extension did not go to completion; thermal fluctuations could fray the double strand DNA. The most likely explanation is the loss of non-specific interaction outside the binding site required to stabilize TF-DNA complex.

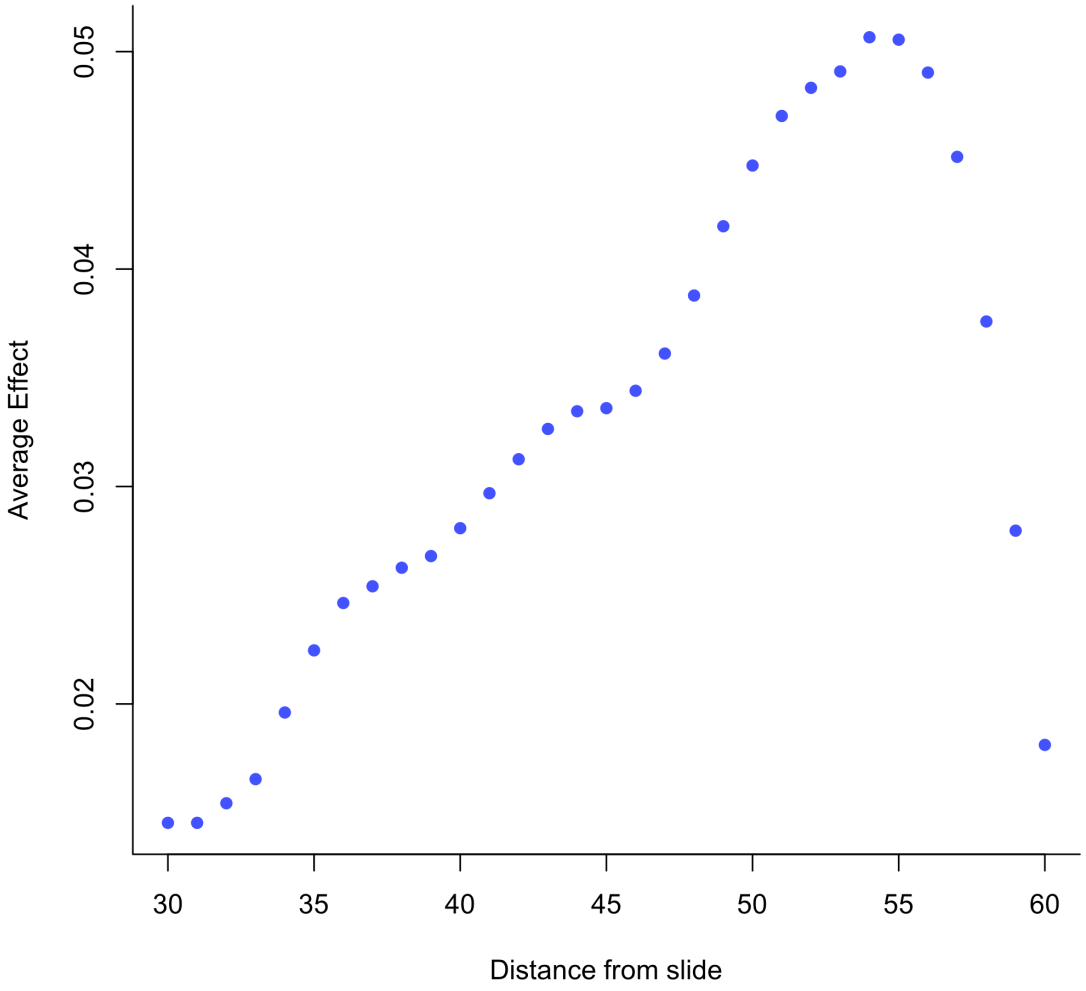


Figure 2.2 Estimated position effect from Plagl1 PBM experiment (Badis et al., 2009), according to equation 2.13. Y-axis shows the average intensities of probes containing binding sites at the position indicated on the x-axis, normalized such that values on the y-axis sum to 1

PBM probe binding probability

Binding probability to the probe i is calculated as:

$$F(i) = \sum_{j=1}^{L} P(j)\, F_{pos}(j) \tag{2.14}$$

where L is length of the variable region on the probe. A more elaborate model of binding that takes overlapping binding sites into account can be used, (Rajewsky et al., 2002; Drawid et al., 2009) but since the variable region is fairly short (35bp), there is usually only one strong binding site per probe and we expect the approximation of equation (2.14) to hold.

Background effect estimation

For sequence specific TFs, the intensity of most probes in a PBM experiment is due to microarray background rather than TF binding. We estimate the distribution of background intensity by symmetrizing the lower half of the observed background peak. Probe intensities are binned (200 bins by default), probability that probe intensities in bin i is generated by TF binding is given by:

$$W_i = \frac{O_i - B_i}{O_i} \tag{2.15}$$

where O_i is the observed number of probes in bin i, B_i is the expected number of probes in bin i from background distribution.

Objective function

Values of free parameters for the PWM ε and μ are determined by minimizing objective function:

$$O(\varepsilon,\mu) = \sum_i W_i \left(Y_i - a - cF(i)\right)^2 + \lambda \sum_{b=A}^{T} \sum_{m=1}^{l} \varepsilon(b,m)^2 \tag{2.16}$$

Where W_i is the probability intensity of probe i is due to background fluorescence, Y_i is z-transformed intensity of probe i, F(i) is the TF binding probability for the probe, calculated as in equation (2.14), ε is the PWM energy model, a and c are parameters used to scale model predicted probe occupancy to PBM fluorescence units. The second term in equation (2.16) is a regularization term designed to prevent PWM parameters from growing too large, with λ controlling the strength of penalization. This form of penalization is equivalent to Maximum a Posteriori estimation of parameter values when the prior distributions of parameters are zero-mean Gaussians with the same variance (Bishop, 2007). By default λ = 0.01.

Minimization is performed using Levenberg-Marquardt algorithm implemented in minpack (More, 1977). Since it is a local optimization algorithm, I use 25 sets of random initial parameters to seed the optimization and report the result with best fit.

Chapter 3
Results and discussion

3.1 Simulations

The purpose of the following simulation studies is to compare the biophysical model used by BEEML with the low concentration approximation implicit in the log-odds model used by practically all motif-finding algorithms. The goal is to identify when the low concentration assumption of the log-odds model breaks down and how it affects the estimation of model parameters.

We performed simulation using the half-site of the Mnt protein to test the BEEML method. Mnt is a repressor from phage P22 for which the binding affinities to all single base variants of the preferred binding sequence have been measured experimentally (Fields et al., 1997; Stormo & Fields, 1998). We use the convention that the preferred base in each position is assigned an energy of 0 and all other values are positive and represent the difference in binding free energy, $\Delta\Delta G$ relative to the preferred base, attributed to each of the other bases (see figure 2.1). Figure 3.1 demonstrates the selection process, with separate panels representing prior distribution, likelihood function and posterior distribution. Figure 3.1A shows the distribution of binding energies over all 7-long sequences for the half-site

energy matrix of Mnt. Figure 3.1B plots the probability of drawing a sequence with a specific energy, from equation (2.1), for three different values of μ in which the probability of the binding to the preferred sequence (with $E_i = 0$) is 0.03, 0.3 and 0.9. Figure 3.1C shows the posterior distribution of binding energies which is the normalized product of the plots in Figures 3.1A and 3.1B, as in equation (2.11). This plot does not use a non-specific binding energy but that is employed in some of the simulations described later. Including E_{ns} has the effect of essentially truncating the distribution at that point and all of probability density that would have been higher accumulates at $E_i = E_{ns}$.

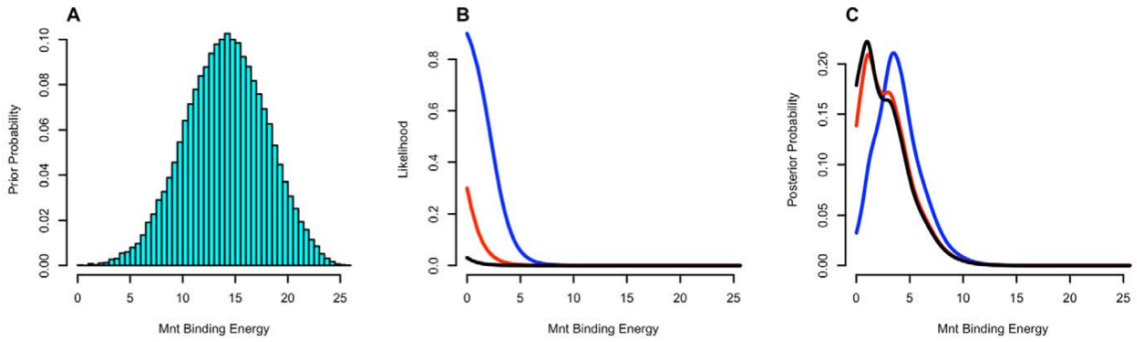


Figure 3.1. Effect of μ on binding probabilities. (A) Prior distribution of binding energy for Mnt half-site, with equiprobable background frequency. (B) Binding probability as function of binding energy, according to equation (2.1). Colors correspond to values of μ, Black: μ = 23.48, Red: μ = 20.85, Blue: μ = 2.2. These values were chosen such that binding probabilities of the consensus sequence are 0.03, 0.3 and 0.9, respectively. No non-specific binding energy is used. (C) Posterior distribution of binding energy, that is, the distribution of energies of the selected sequences. Taken from (Zhao et al., 2009)

Using various values of μ and E_{ns} and setting $P(S_i)$ to be constant (equiprobable background distribution) 100,000 sites were drawn for each simulation according to equation (2.11). I used BEEML and the standard log-odds model (equation 1.2 with $\lambda = 1$) to estimate PWMs from the sampled sites. Figure 3.2 shows the performance of BEEML at predicting the true binding probabilities in the Mnt simulations for several different values of μ (figure 3.2A–C) and E_{ns} (figure 3.2D–F). Each graph shows the true probabilities for all sequences and the predicted probabilities obtained by BEEML and also using a standard log-odds approach where the probabilities of each base at each position are taken directly from the observed sites. As expected, both methods give very accurate predictions of binding probabilities when μ is low. At higher values of μ, when the highest affinity sites approach saturation, the log-odds method is much worse at predicting the binding probabilities. Even when the preferred site is bound with $p = 0.3$ (figure 3.2B), which is less than half saturated, there is a substantial difference in accuracy of predicted binding probability. At $p = 0.9$ for the preferred site (figure 3.2C), the predictions from the log-odds method are wrong by about a factor of 2, whereas the BEEML predictions are very accurate. Many TF binding sites in vivo are likely to function at near saturation, especially those regulated by repressors, and inaccurate models for the binding probabilities can lead to very large increases in the number of false positive predictions of regulatory sites (Djordjevic et al., 2003; Roider et al., 2007; Homsi et al., 2009).

Similar results are obtained for variations of E_{ns}. When E_{ns}=13.8 (which corresponds to a 10^6-fold ratio of non-specific binding affinity compared to the preferred binding site, figure 3.2D) both methods give accurate predictions of binding probabilities. But when it is reduced to 11.5 (ratio of 10^5, figure 3.2E) the log-odds method is less accurate, and when it is reduced to 9.2 (ratio of 10^4, figure 3.2F) the log-odds predictions are wrong by about a factor of 2, whereas the BEEML predictions are still very accurate because it explicitly account for that parameter whereas the log-odds method cannot.

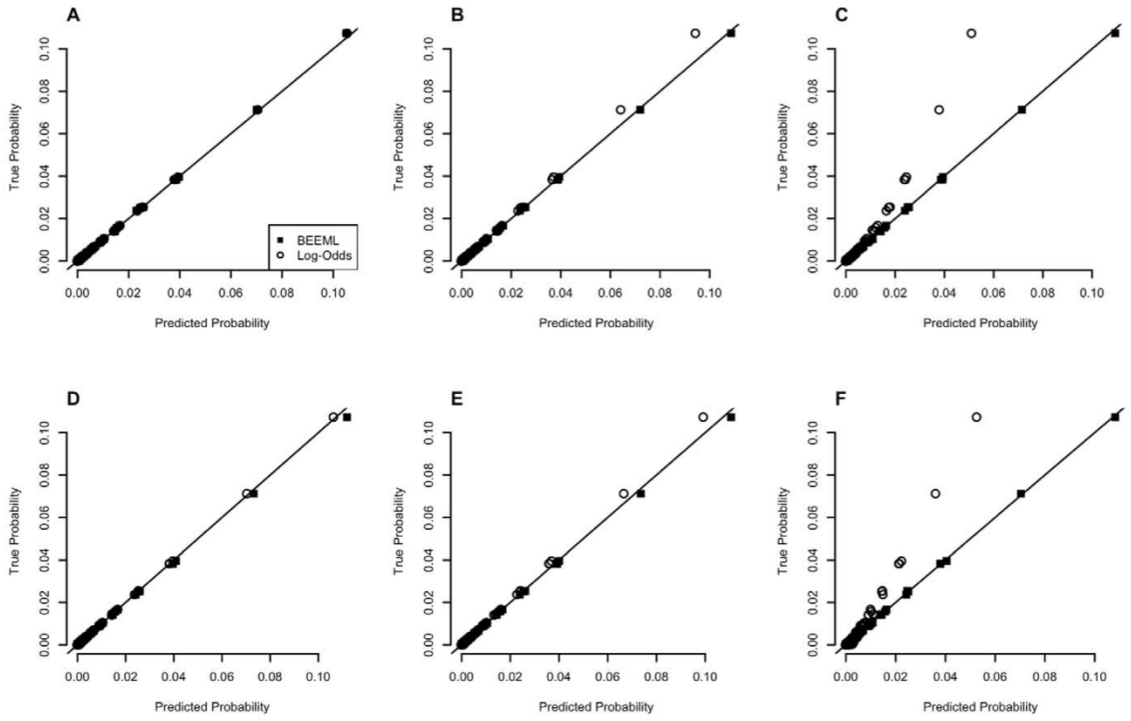


Figure 3.2. Examples of Simulation Results. Top Panel (A–C): Effects of m. Non-specific energy was set to 30 so as to have negligible effect on binding. (A) μ = 23.48 (B) μ = 20.85 (C) μ = 2.2. Bottom Panel (D–F): Effects of E_{ns} at low concentration limit. μ was set to -100. (D) E_{ns} = 13.82 (E) E_{ns} = 11.51 (F) E_{ns} = 9.21. These values were chosen such that the relative K_i of consensus sequence to non-specific binding is (D) 1,000,000 (E) 100,000 (F) 10,000. Taken from (Zhao et al., 2009)

3.2 Analysis of MITOMI binding data

Figure 3.3 shows results for BEEML analysis of the basic helix-loop-helix TF MaxA binding affinity data measured by Mechanically Induced Trapping of Molecular Interactions (Maerkl & Quake, 2007). Figure 3.3A comes directly from quantitative binding data where the measured binding energies are plotted against the predictions assuming that multi-position variants show the additive energy changes of the individual base changes. As Maerkl and Quake point out, this additive assumption is not very accurate and the fit between the observed and predicted binding energies has only $r^2 = 0.57$. Figure 3.3B plots the predictions from BEEML which estimates $E_{ns} \approx 3$ (much lower than the values used in the simulations of figure 3.2) and finds the best overall additive parameters, which together lead to an improved $r^2 = 0.84$. Figure 3.3C goes one step further and estimates maximum likelihood parameters for nearest neighbor contributions to the binding energy. Using these adjacent di-nucleotide parameters increases the fit to $r^2 = 0.96$, which is essentially within the measurement error.

While affinity measurements of single base variants did not lead to very accurate models, including contributions from non-specific binding resulted in a very good fit, demonstrating that additivity is a good assumption for the specific component of binding energy. The additive model may already be sufficient for many purposes, with the addition

of nearest neighbor energy contributions; we obtained a model that fit the data almost perfectly. If this holds true for more TFs, then the task of measuring TF specificity can be greatly simplified. For example, if the binding site is 10 long and all non-additive interactions were confined between neighboring positions, then instead of measuring affinity of all 4^{10} = 1,048,576 possible 10-long sequences, one can obtain the same information by only measuring the affinity of 112 sequences (the consensus sequence + 30 single base variants + 81 interaction parameters), a > 9,000 fold reduction in effort.

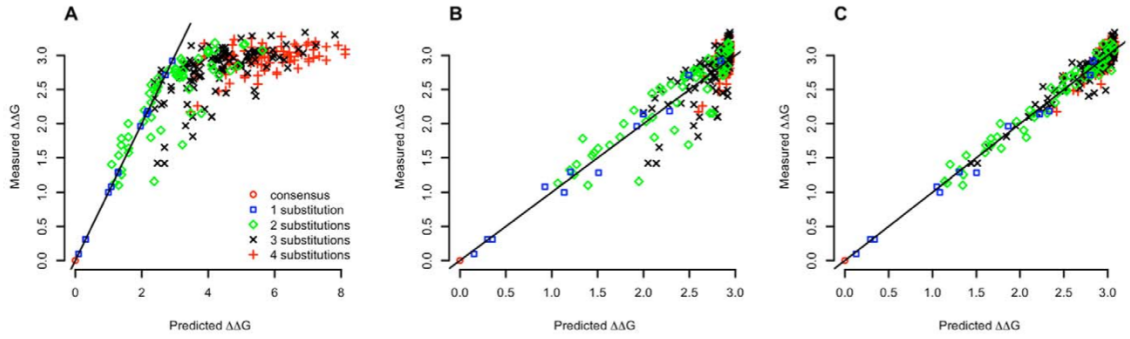


Figure 3.3. BEEML analysis of MITOMI (Maerkl & Quake, 2007) data. (A) Fit of point-estimate of binding energy as done in the original paper (B) BEEML fit with PWM energy model and non-specific energy parameter (C) BEEML fit with position specific di-nucleotide energy model and non-specific energy parameter. Taken from (Zhao et al., 2009)

3.3 Analysis of HT-SELEX Data

I next used BEEML to analyze high-throughput sequencing data from a single round of selection with C2H2 zinc-finger TF Zif268 (Zhao et al., 2009). The sequences of the initial library showed a small bias in the composition on the synthetic strand: A = 24.5%; C = 21.0%; G=27.2%; T=27.4%. Prior probabilities of sequences, $P(S_i)$, were estimated based on the mononucleotide composition. It is possible to measure $P(S_i)$ directly by sequencing the initial library more deeply, but in these experiments we only obtained about 200,000 sequences from each library, too few to estimate the frequencies of all 4^{10} ($\sim 10^6$) 10-mers. Since no significant higher-order biases were observed we expect that the frequencies of all 10-mers in the initial library to be well approximated based on the mononucleotide composition.

An initial BEEML model based on all of the selected binding sites was used to determine the most likely orientation of each site and whether it was entirely within the 10bp randomized region or overlapped the fixed sequences. Sites that were determined to overlap the fixed regions were eliminated from further analysis and the remaining sequences were reanalyzed by BEEML. As expected, because of the slight compositional bias and the G-rich consensus for zif268: GCGTGGGCGT (Liu & Stormo, 2005), more sites were selected in the "top" orientation than in the reverse. When computing the likelihood we sum over

binding in both orientations. Figure 3.4 shows the observed and predicted counts for all of the sequences in the selected set based on the BEEML model and also for a model obtained using BioProspector (Liu et al., 2001), a motif discovery program designed for this type of data. From the total of 259,704 sites, BioProspector built a model based on only 28,046 (10.8%) sites, but obtained a model that is similar to the known zif268 binding model. While BioProspector identifies the known consensus sequence and the PWM it finds is similar to previously published ones for zif268 (Liu & Stormo, 2005), its quantitative predictions are much worse than those from the BEEML model (r^2 = 0.74 for BioProspector, r^2 = 0.92 for BEEML). Not only are the non-specific and low affinity sites, which are the majority after only a single round of selection, better predicted by BEEML, but the high affinity, near-consensus sites are predicted much more accurately and with very little scatter compared to the BioProspector predictions. BEEML also returns estimates of μ=1.98 and E_{ns}=12.37. The predicted non-specific binding ratio of ~10^5 fold less than to the consensus sequence is in the range typical for many TFs. The estimate of μ predicts that the consensus sites should be about 88% bound which is reasonable because, even though DNA is in 100-fold excess over protein in these experiments, most of the DNA sequences will have only non-specific affinity. This makes the experiment similar to the simulation depicted in Figure 3.2C and highlights the importance of the biophysical model instead of the log-odds approach. Because we are estimating only 32 parameters (30 for the PWM, and μ and E_{ns}) and have >

10^5 binding sites, we do not expect any over-fitting but to verify that is the case we performed a 10-fold cross-validation where we determined the parameters based on a random sample of 90% of the sequences and measured the fit to the remaining 10%. Indeed, we find that $r^2 = 0.91$ on those samples.

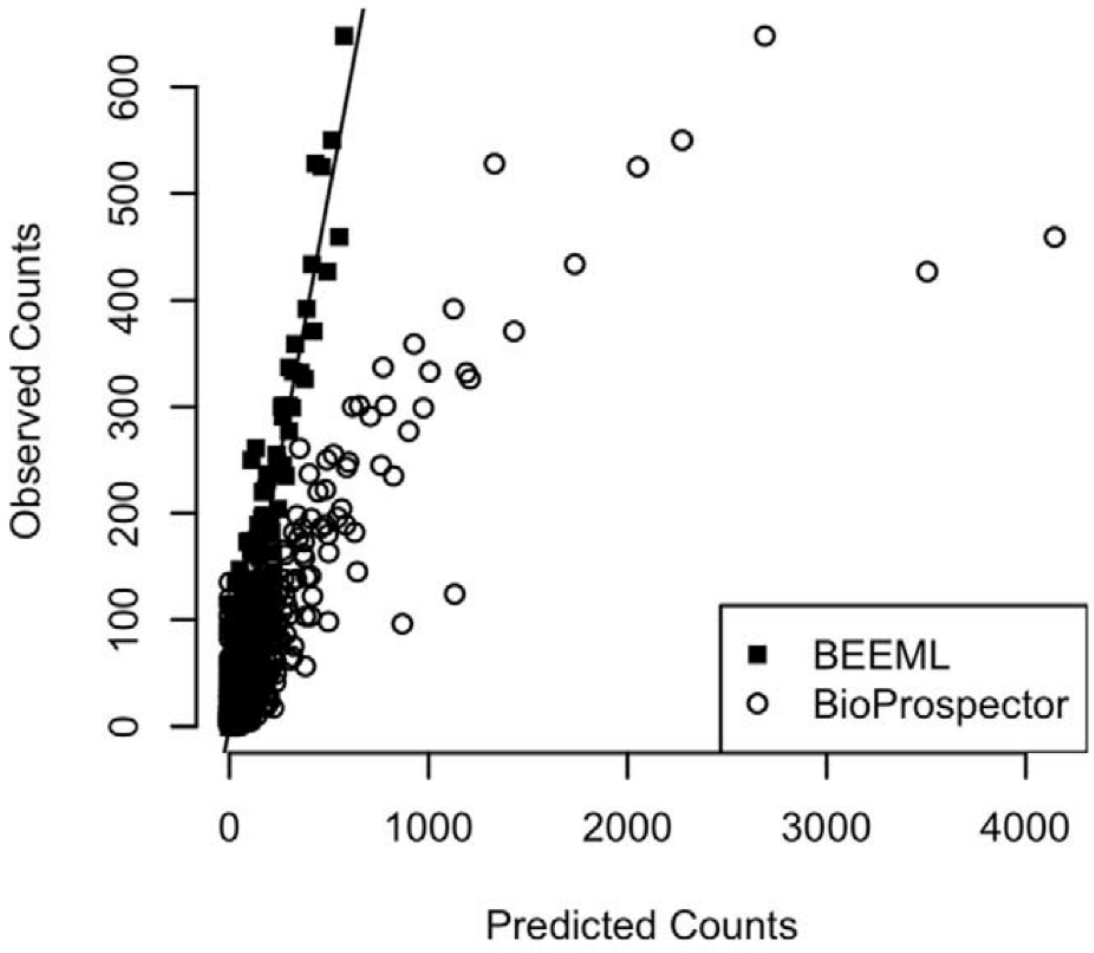


Figure 3.4. Fit of BEEML and BioProspector model to HT-SELEX data. Taken from (Zhao et al., 2009)

Comparison of BEEML and Other Biophysical Approaches

Probabilistic models for binding site recognition, such as the fairly standard log-odds method, are popular because of their simplicity, intuitive appeal and because they can be easily implemented in motif discovery algorithms. But they suffer from over-simplification of the underlying model, not just the typical additivity assumption which is known to be an approximation, a good one sometimes and other times not (Benos et al., 2002), but also because it ignores the non-linear relationship between binding energy and site statistics which is especially pronounced when high affinity binding sites approach saturation. A biophysical model (Gerland et al., 2002; Djordjevic et al., 2003) captures the non-linear dependence of the binding probability on the energy and can easily incorporate multiple modes of binding, even beyond the specific and non-specific contributions that we employed in this study. It can easily incorporate non-additive, or higher order, contributions of the sequence to the binding energy, as we demonstrated on the MaxA data.

Djordjevic et al (Djordjevic et al., 2003) developed a quadratic programming (QP) method to estimate binding energy parameters from example binding sites and demonstrated that the resulting model could make many fewer false positive predictions on genome sequences. But QP is still limited in the kinds of data for which it works well. It assumes a "zero temperature" limit for the binding probability so that sites either bind or not, rather than have a specific probability of binding. It functions like a support vector machine trained

on only positive examples and is very sensitive to any outliers or noisy data. For these reasons it works well on collections of high affinity sites but its performance is degraded with any background or non-specific binding, and the quality of the model decreases rapidly as low quality, or even low affinity, data is added. BEEML doesn't suffer from those limitations because it models the complete distribution including non-specific binding so that the more data available the better it works, even if most of the sequences are non-specific. The algorithm is more complex and slower than QP, but still reasonably fast even for long sites when using the FFT to estimate the partition function.

3.4 Analysis of PBM data

Protein binding microarray (PBM) is a technique that measures the binding of TFs to double-stranded DNA arrays that currently contain all possible 10-long binding sites and so provides enormous information about the specificity of the TF (Bulyk et al., 2001; Mukherjee et al., 2004; Berger et al., 2006). All of the published data sets from Martha Bulyk's group are available in an online database, UniPROBE (Newburger & Bulyk, 2009; Robasky & Bulyk, 2011). UniPROBE currently contains specificities for 404 TFs from *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Mus musculus* and *Homo sapiens* belonging to a variety of structural classes.

A recent PBM study of mouse TFs (Badis et al., 2009) concluded that the energetics of TF-DNA recognition are highly complex: 41 out of the 104 TFs studied had clear secondary binding preferences not captured by the primary PWM and 89 out of 104 TFs were better represented by a linear combination of multiple PWMs than a single PWM. However, three different methods were used to obtain PWMs in this study, each method being superior to the others on some datasets, indicating that none of the methods can be optimal at determining the PWM parameters. As noted by the authors, it is possible that the insufficiency of their PWMs is not due to the complexity of TF-DNA recognition, but rather the algorithms used for parameter estimation. Before abandoning the idea that specificity can be largely explained with simple models, it is critical to assess the fitness of optimal PWMs.

Factors confounding analysis of Badis et al (Badis et al., 2009)

In a PBM experiment, a purified, epitope-tagged TF is applied to a double-strand DNA microarray. The degree of binding to each probe on the microarray is quantified by the application of a labeled antibody specific to the epitope tag. In theory, signal intensity of a probe should be directly proportional to the probability of TF binding to the sequence of that probe. In practice, however, the relationship is not so straightforward due to a number of factors such as background signal, position effect and influence of flanking sequences. We

have found that these factors significantly confound current analysis methods, such as 8mer enrichment analysis (Berger et al., 2006).

A number of factors complicate the relationship between probe sequence and intensity in a PBM experiment. One such factor is the background intensity. In a typical experiment, the majority of the probes have low background signal while a small subset of probes, containing high affinity binding sites, display high signal intensity. An example histogram of probe intensities for mouse TF Esrra (Estrogen related receptor, alpha) is shown in figure 3.5. The variations in probe intensity observed in the large low intensity peak is most likely dominated by background signal rather than differences in specific binding by the TF.

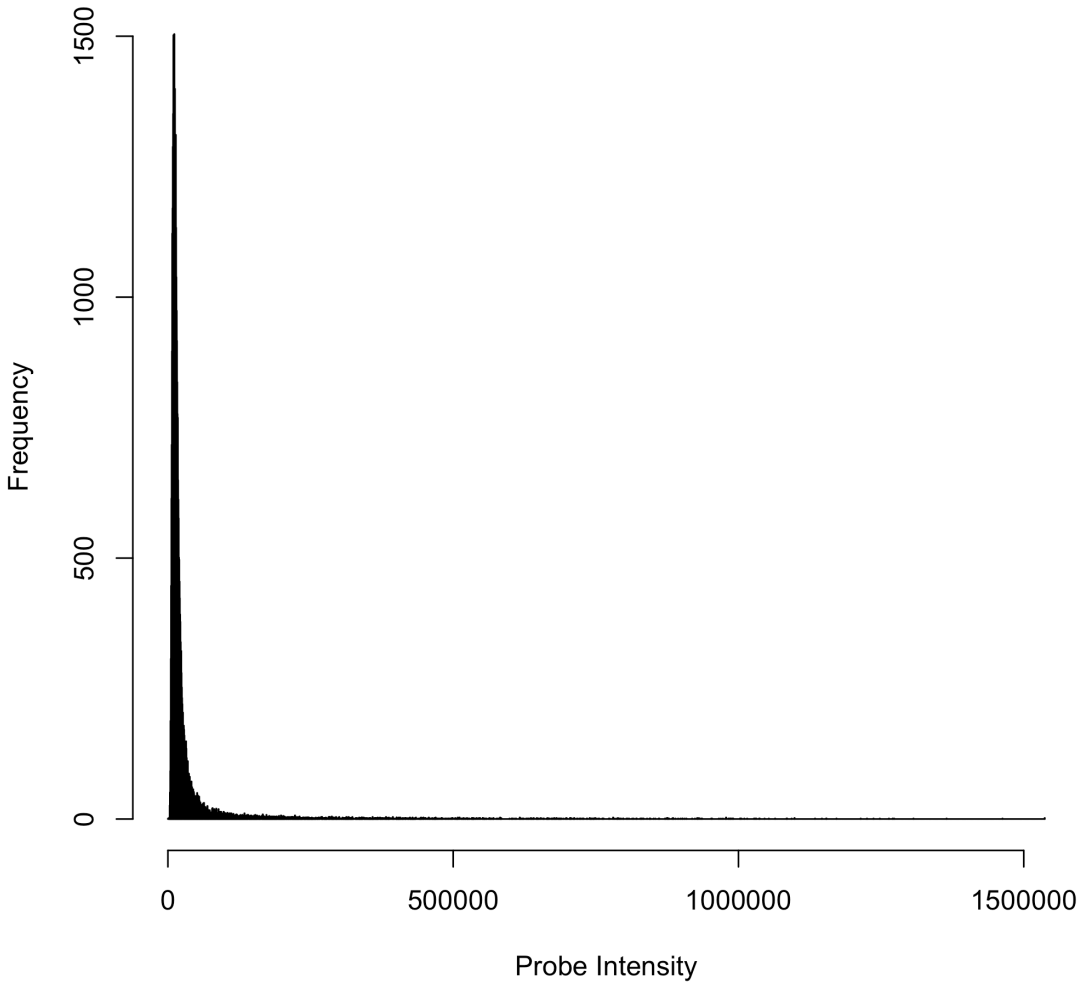


Figure 3.5 Histogram of probe intensities from PBM measurement of mouse TF Esrra (Badis et al., 2009). As a sequence specific TF, Esrra only binds strongly to a small subset of probes on the PBM, which shows up as the long right tail in the histogram.

Berger et al. (Berger et al., 2006) identified three other factors that influence the observed signal: flanking sequence, position and orientation of the binding site within the probe. Flanking sequence is important because the length of the variable region on each probe is 35 base pairs, much longer than the length of a typical binding site (6-10 bp). Using a series of controls, Berger et al. (Berger et al., 2006) demonstrated that probes containing sites farther away from the slide gave higher intensity signals (also see figure 2.2). A smaller effect also exists for the orientation of binding sites.

Currently, the standard method for estimating affinity from noisy probe intensities is to use multiple probes and hope the noise will average out (Berger & Bulyk, 2009). Typically, analysis is performed at the level of 8-long sequences (8mers): the median intensity of all probes containing a particular 8mer is considered as a measure of the TF's preference for that 8mer. Median intensities of 8mers can be standardized as 8mer z-score, which allows one to compare specificities of different TFs (Badis et al., 2008; Wei et al., 2010; Lam et al., 2011). The ranks of median intensities can also be transformed into 8mer enrichment scores (E-scores), which is a measure of the ability of an 8mer to function as a classifier to distinguish "bound" probes from "unbound" probes. The somewhat arbitrary division of probes into bound and unbound sets, as well as the use of rank bases statistics makes E-score more robust but less sensitive.

E-scores were used as proxies for affinities by Badis et al. (Badis et al., 2009) and serve as the basis of many of the observations of the non-independence of positions in the binding site. The use of E-scores to estimate TF specificity has many drawbacks: First, much of the quantitative binding information is lost since only ranks of median intensities are used. Second, this analysis attempts to estimate 32,896 parameters (8mer E-scores), each from only 32 probe intensities. Given there are only ~44,000 probes on the array, this is a difficult task that practically guarantees many of the parameters are not well estimated.

This difficulty is illustrated in figure 3.6. Badis et al. (Badis et al., 2009) stated, on the basis of E-scores, that Esrra has a strong preference for CAAGGTCA or AGGGGTCA, but not CGGGGTCA or CAGGGTCA. While the median probe intensities are consistent with this conclusion, it is apparent that the intensities of probes containing the same 8mer are highly variable. For example, intensities of probes containing the consensus sequence CAAGGTCA span the entire range of PBM signal, likely due to the effects of the confounding factors discussed above. The high level of variability in probe intensities indicates that it is not practical to accurately estimate all 32,896 8mer affinities directly from the data.

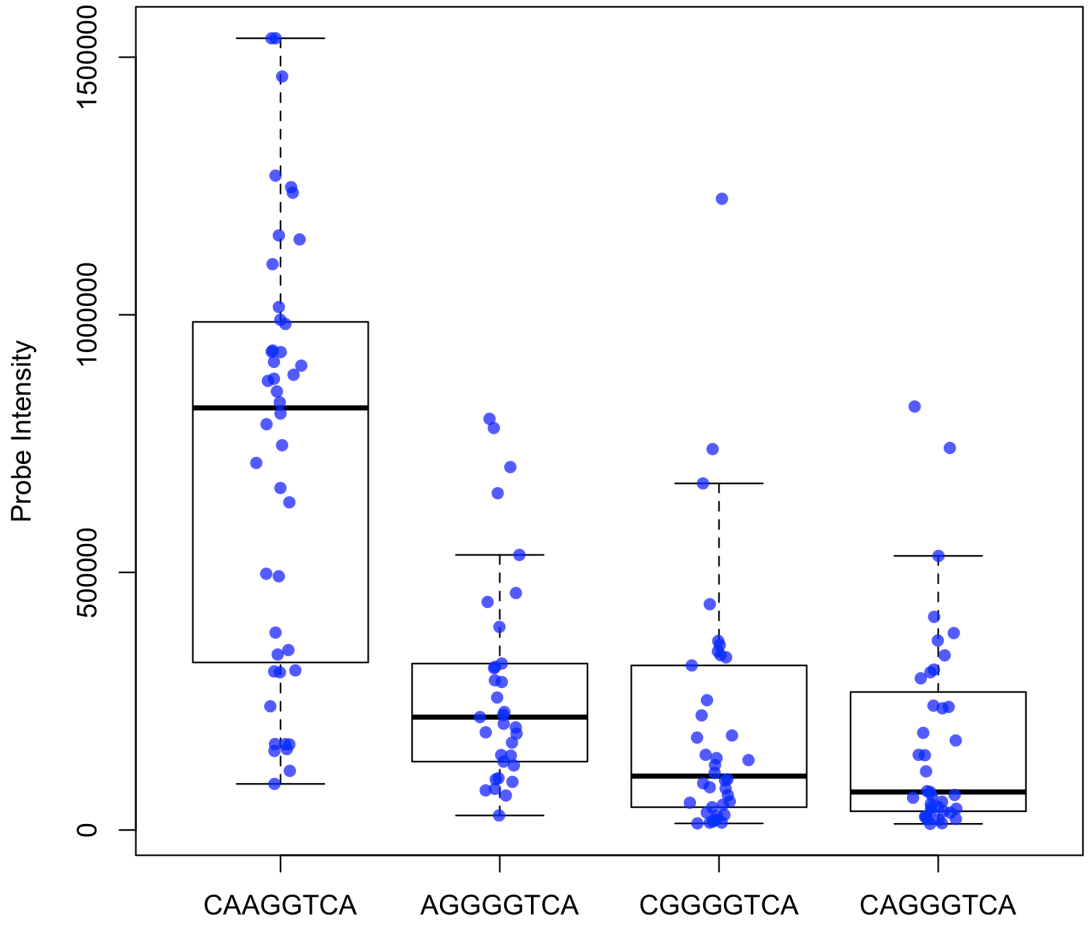


Figure 3.6 Intensities of probes containing particular 8mers in Esrra PBM experiment. Each point is the intensity of a probe containing the 8mer listed on x-axis.

A more serious problem with E-score analysis is that it is vulnerable to the influence of flanking sequences. Low affinity sequences that partially overlap the binding site tend to appear on the same probes as high affinity 8mers, often resulting in artificially high E-scores. This problem can be demonstrated using yeast TF Pho4 as an example. Pho4 is a well-studied TF that binds to a core CACGTG motif. The quantitative specificity of Pho4p has been measured using two different techniques: PBM (Zhu et al., 2009) and Mechanically Induced Trapping Of Molecular Interactions (MITOMI) (Maerkl & Quake, 2007). There are 136 8mers from Pho4 PBM data in Zhu et al. (Zhu et al., 2009) that would be considered to be high-affinity sequences by the criterion of E-score ≥ 0.45 commonly used in PBM analysis. Comparison with MITOMI measured binding energies of these 8mers (figure 3.7) shows that while the E-score of the majority of 8mers are consistent with MITOMI energies, a substantial minority (44 out of 136) of 8mers with high E-scores have high binding energies (low affinity) according to MITOMI measurements.

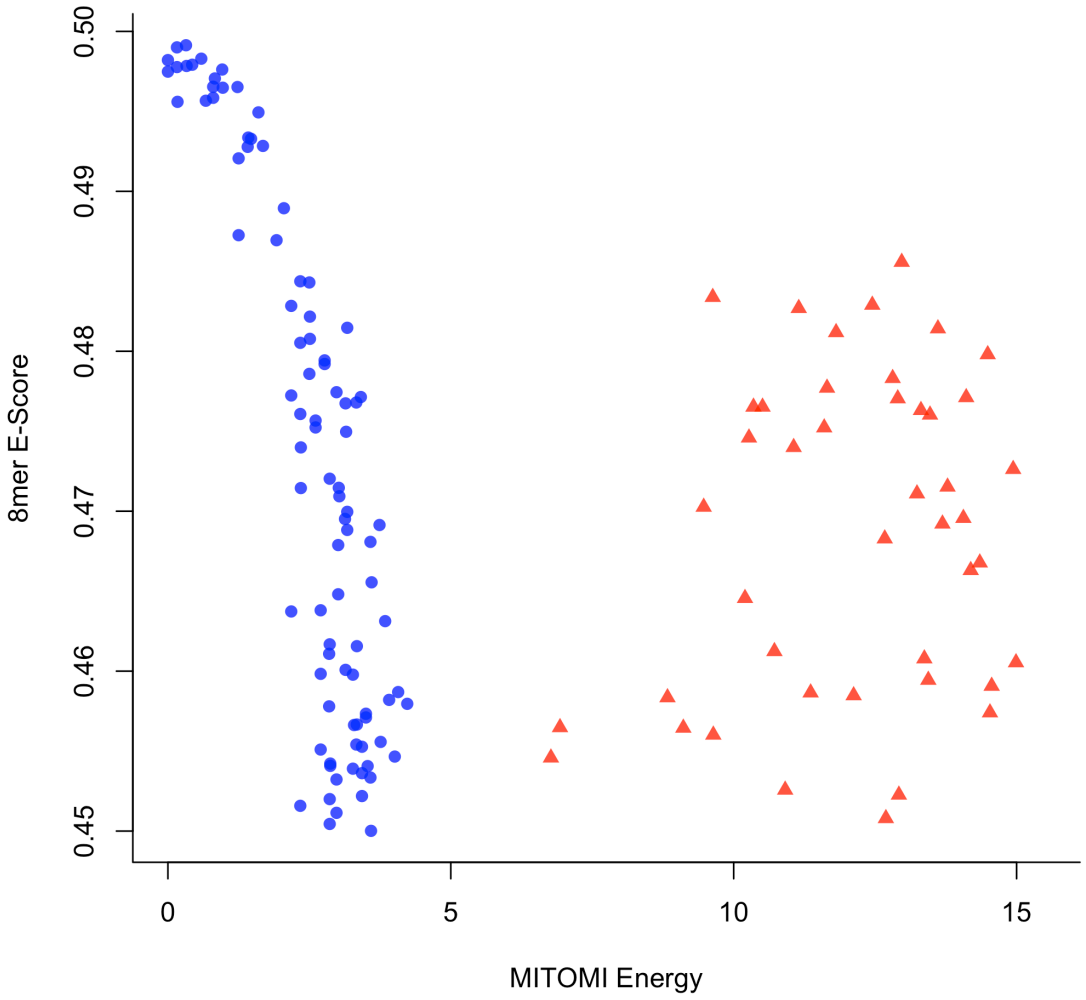


Figure 3.7 Comparison of binding energies measured by MITOMI and E-scores of all 8mers with E-scores ≥ 0.45 in Pho4 PBM. Blue circles represent high-affinity sequences measured by both methods; red triangles represent those sequences with low affinity as measured by MITOMI but high affinity according to PBM.

There are good reasons to believe that high E-scores of these low affinity 8mers (red triangles in figure 3.7) are due to artifacts in the E-score analysis rather than preferential TF binding. First, the sequences of these low affinity 8mers do not contain the CACGTG core motif, but rather a partial match, CACGT, on one of the strands. Second, comparison of MITOMI energies and Z-transformed 8mer median intensities (figure 3.8) demonstrates that many of these low affinity 8mers with high E-scores have low Z-scores, suggesting enrichment calculation, rather than high signal intensities, is responsible for observed high E-scores. Third, MITOMI binding energies can be used in the BEEML framework to accurately predict PBM probe intensities (taking binding site position and protein concentration have been taken into account). Figure 3.9 shows the 8mer medians of predicted probe intensities using MITOMI energies are in good agreement with PBM data, with r^2 value of 0.70. Further, median probe intensities of all 136 8mers calculated using MITOMI energies, which includes the contribution of flanking sequences, are consistent with their E-scores (figure 3.10). Taken together, it is clear that the discrepancies between MITOMI and PBM measurements shown in figure 3.7 are due to artifacts introduced in the E-score analysis. As figure 3.9 shows, despite the fact that very different techniques were use to measure TF specificity, quantitative binding data produced by MITOMI and PBM are in good agreement. It is also clear that 8mer enrichment analysis conducted by Badis et al.

(Badis et al., 2009) suffers from numerous inaccuracies and their conclusion that the energetics of TF-DNA recognition is highly complex must be re-visited.

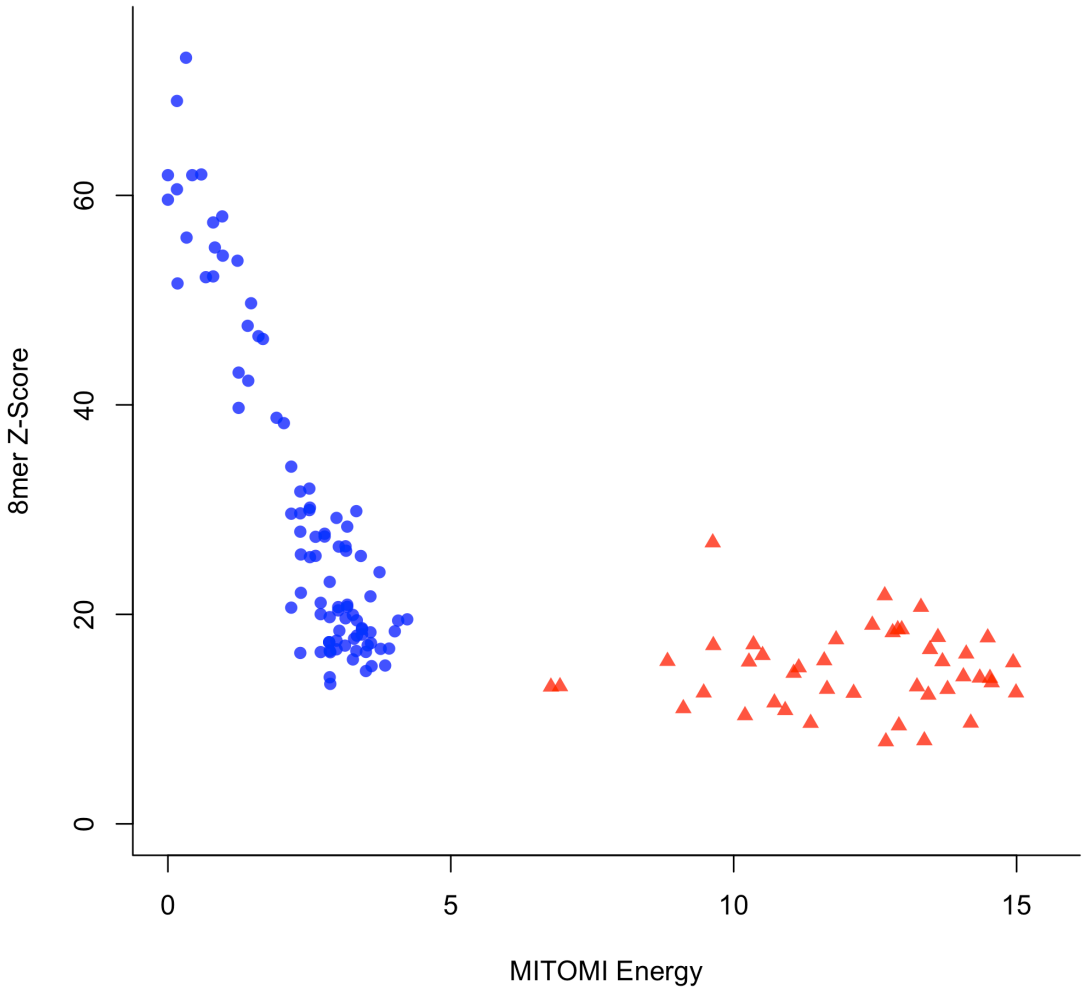


Figure 3.8 Comparison of binding energies measured by MITOMI and E-scores for the same 8mer sequences as figure 3.7

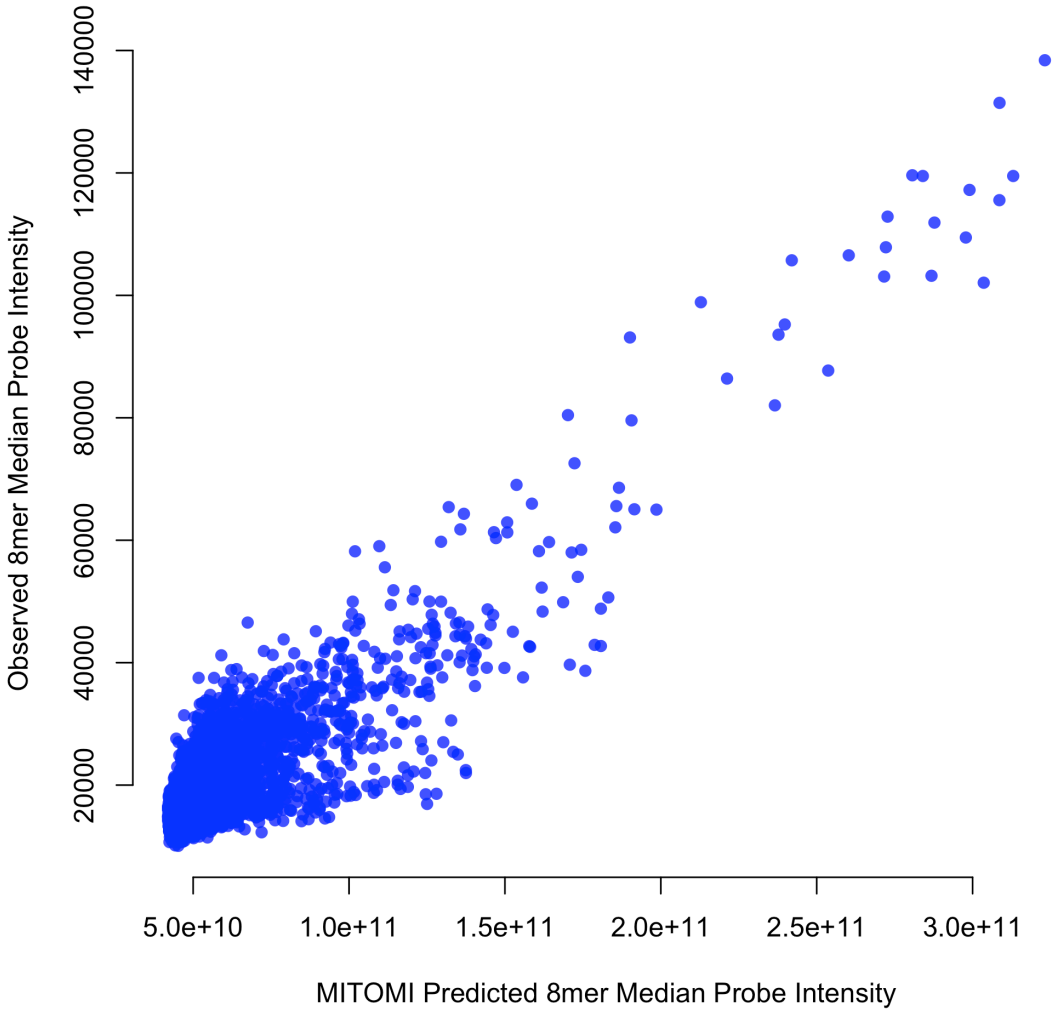


Figure 3.9 MITOMI measured binding energies can be used to predict PBM probe intensities with high accuracy. Probe intensities were calculated using equation 2.14

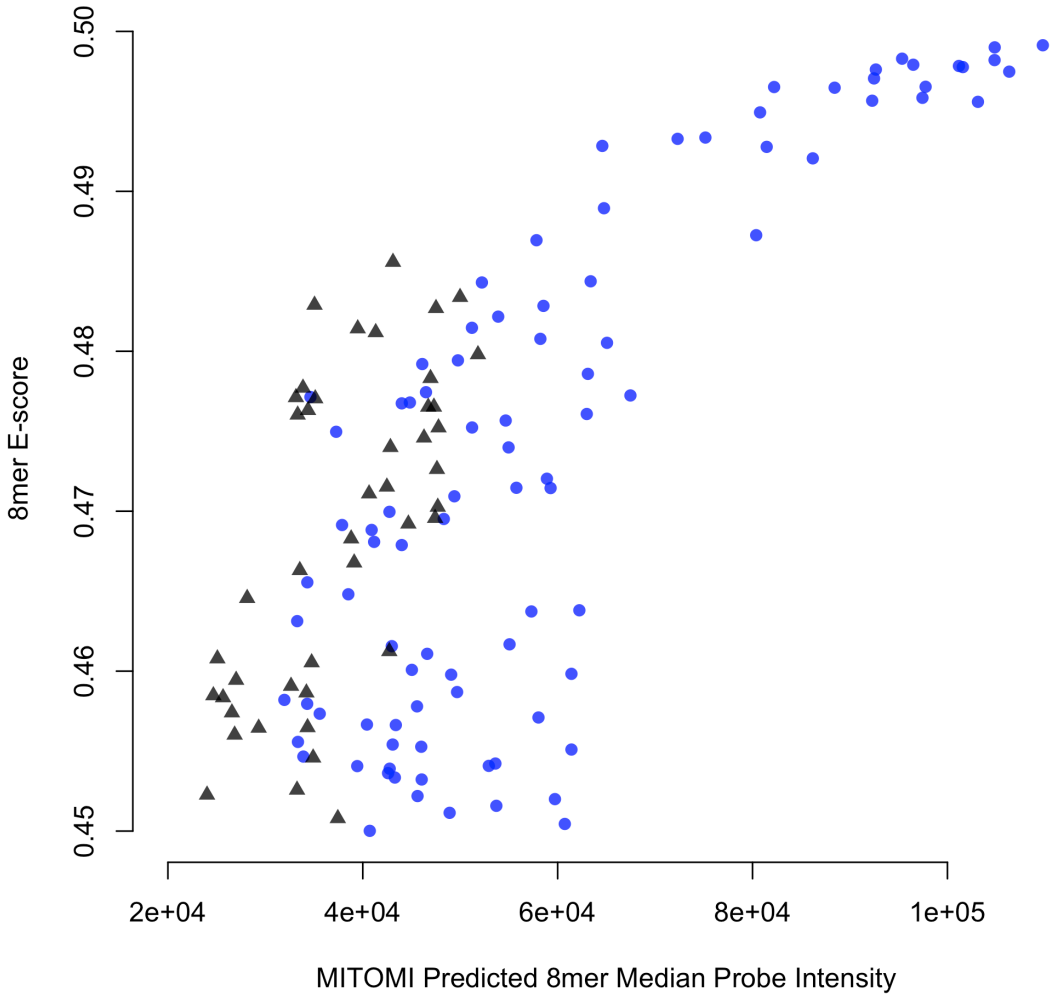


Figure 3.10 Predicted 8mer median probe intensities based on MITOMI binding energies are consistent with E-scores. The difference between figure 3.10 and figure 3.7 is that contributions of flanking sequences on the probe are included in figure 3.10

BEEML analysis of PBM data

The BEEML approach to PBM analysis is to estimate the position and background effects from the data first, then perform weighted regression to parameterize a model of binding energy, explicitly taking these biases into account. This offers several benefits. First, using a model drastically reduces the number of parameters required: a 10- long PWM only requires 30 parameters. This represents a 1000 fold reduction over 8mer enrichment analysis, which attempts to estimate TF affinity for all 8-long sequences. Second, having a model of specificity allows us to test hypotheses about the binding mechanism. For example, if the performance of the palindromic model, where the parameters of the half-sites are constrained to equal to each other, is comparable to the full model where all parameters are allowed to vary then it is likely that the TF binds DNA as a homo-dimer. Third, all of the data are used to estimate each parameter, improving accuracy. Finally, by using a model to calculate TF binding probability for the entire probe, the influence of flanking sequence that confound the current analysis is explicitly included.

Our algorithm, BEEML-PBM (Binding Energy Estimation by Maximum Likelihood for Protein Binding Microarrays) extends the existing algorithm BEEML (Zhao et al., 2009) to estimate models of TF specificity by weighted regression on PBM data. PBM signal intensity is modeled as a convolution of background effect, position effect and equilibrium binding probability to the probe sequence. Using BEEML-PBM, we find that the simple

PWM model of specificity performs very well for most transcription factors. This simplicity has important implications for our understanding of the molecular basis of TF specificity and demonstrates the importance of the analysis method in the interpretation of high-throughput data.

An example of model-based analysis of yeast factor Pho4 PBM data (Zhu et al., 2009) is shown in figures 3.11 – 3.16. The LOGO representation (Schneider & Stephens, 1990) of the 10-long PWM fitted by BEEML-PBM is shown in figure 3.11. Since Pho4 is a basic helix loop helix TF that is known to dimerize(Shimizu et al., 1997), BEEML-PBM was also used to fit a palindromic model, where the parameters of the half-sites are constrained to equal to each other. LOGO of the fitted palindromic model is shown in figure 3.12. The performances of full and palindromic models are shown in figures 3.13 and 3.14, respectively. The fact that the palindromic model fits the data nearly as well as the full model (palindromic model has r^2 of 0.755 vs. 0.762 for the full model) despite it having only half the parameters of the full model provides strong support for the hypothesis that Pho4p binds DNA as a homodimer (Shimizu et al., 1997) in PBM experiments.

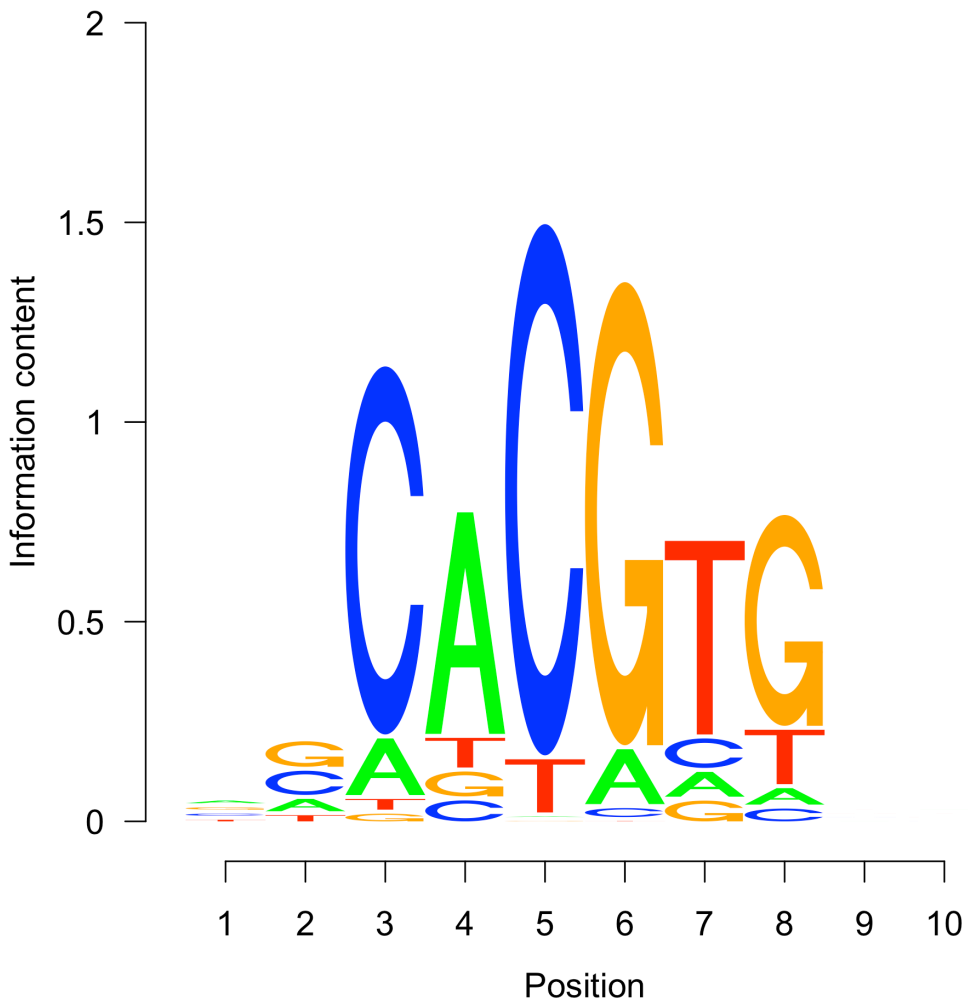


Figure 3.11 LOGO representation of BEEML-PBM fitted full model for Pho4

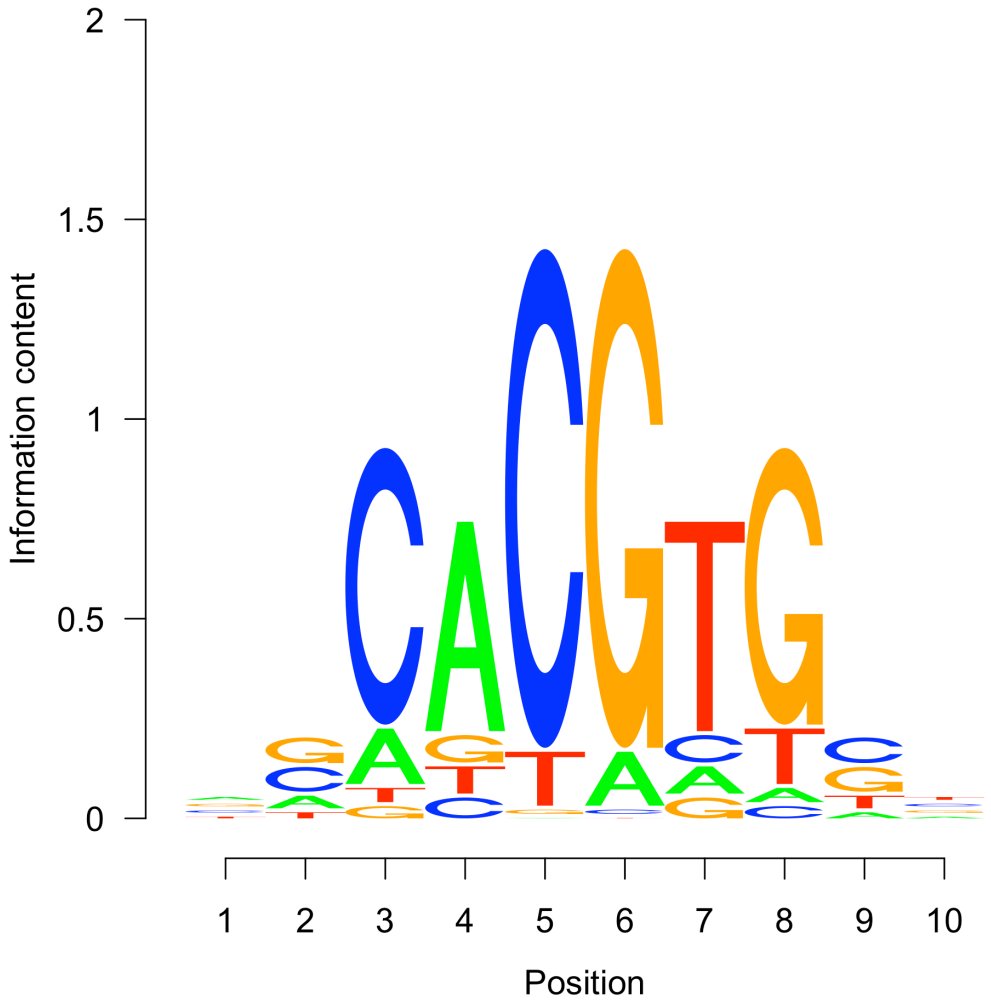


Figure 3.12 LOGO representation of BEEML-PBM fitted palindromic model for Pho4

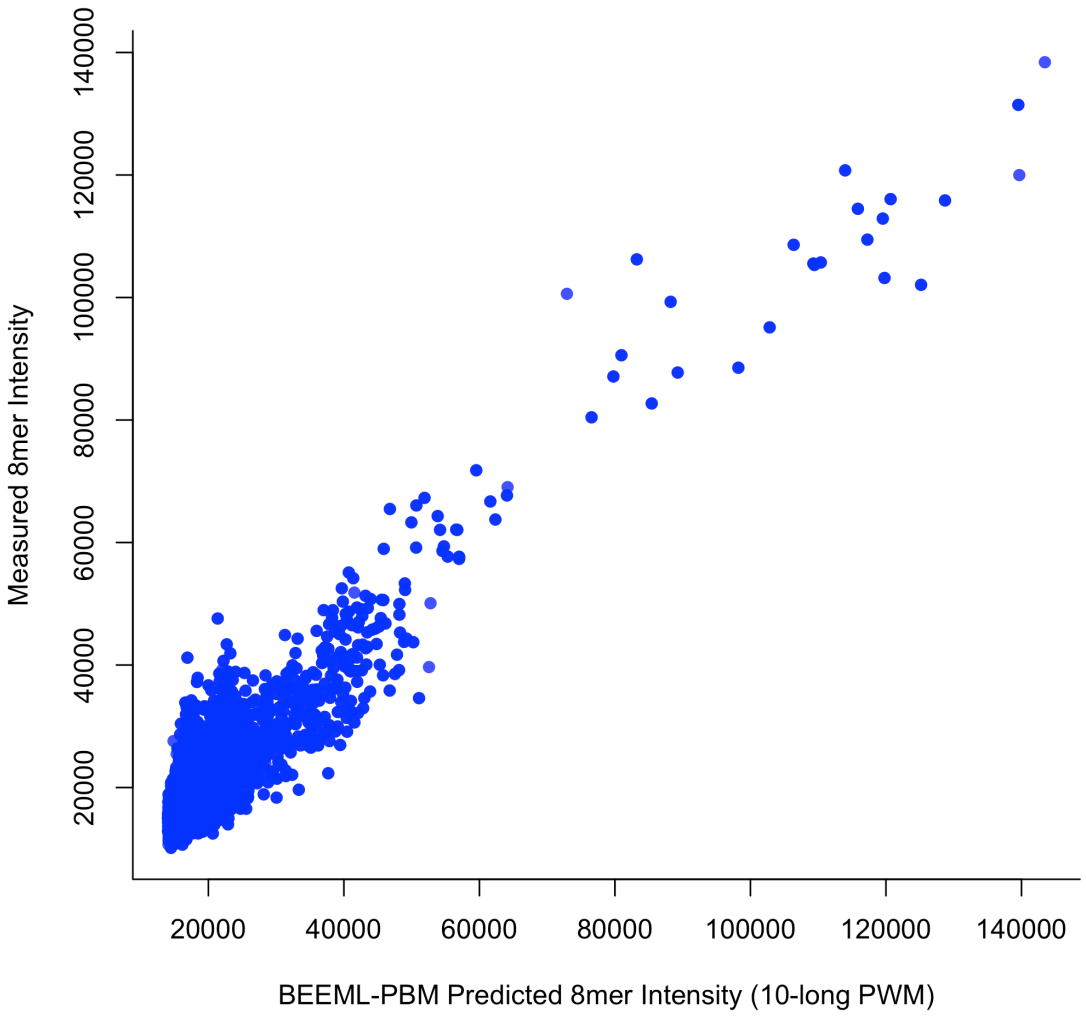


Figure 3.13 Performance of 10-long Pho4 PWM on replicate array, $R^2 = 0.762$

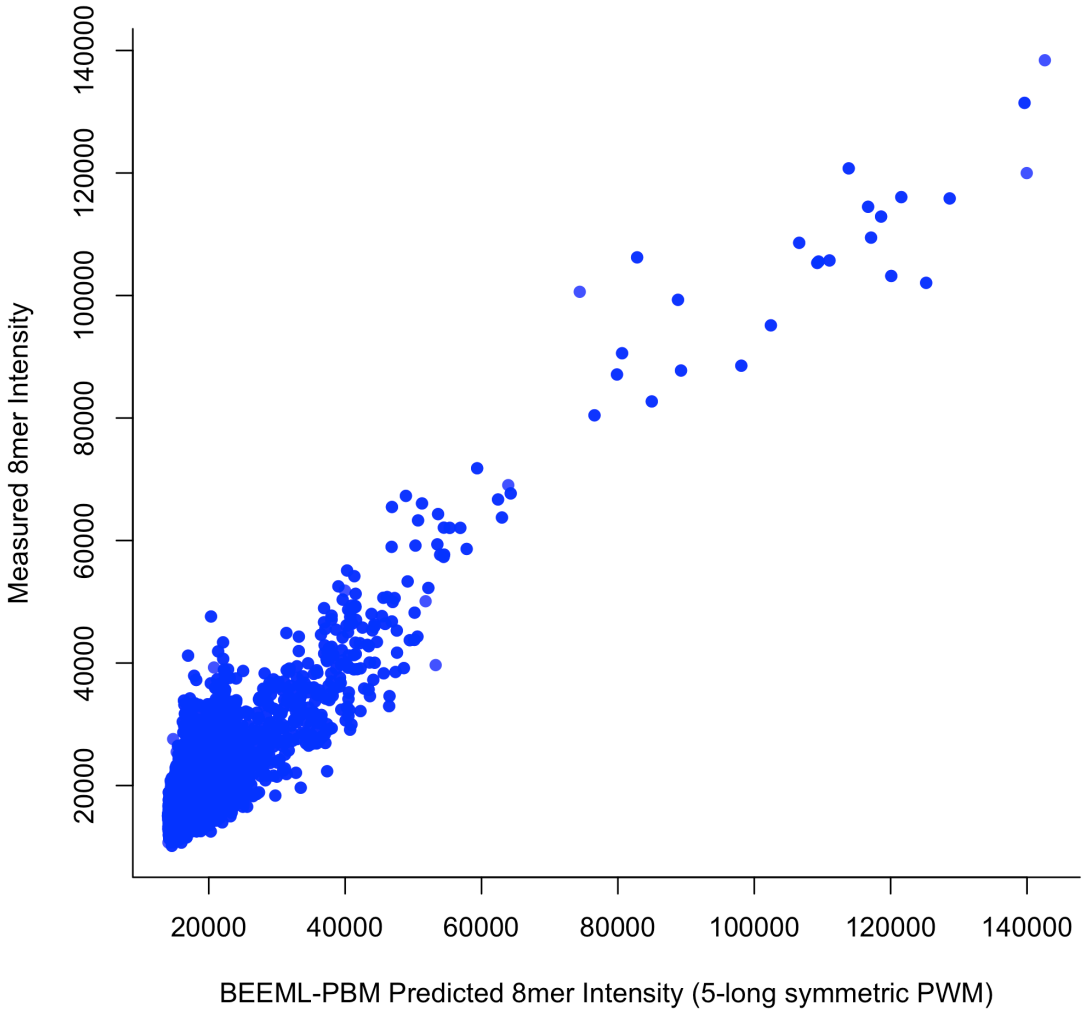


Figure 3.14 Performance of palindromic Pho4 PWM on replicate array, $r^2 = 0.755$

We next compared binding energies estimated from PBM data (Zhu et al., 2009) with those measured by Maerkl and Quake (Maerkl & Quake, 2007) for yeast TFs Pho4 and Cbf1. Although measured by completely different experimental techniques, binding energies estimated by BEEML-PBM are consistent with MITOMI results, with r^2 values of 0.94 and 0.88 respectively (figures 3.15 and 3.16), further establishing the ability of BEEML-PBM to obtain accurate binding energies from PBM data.

Although Cbf1 PWM parameters estimated by BEEML-PBM are directly proportional to those measured in MITOMI experiment, they do not agree at high binding energies. MITOMI binding energies saturates at 3 kcal/mol while BEEML-PBM energies go up to 6. There are two possible explanations: overfitting by BEEML-PBM or difference in dynamic range between MITOMI and PBM techniques. It is not possible to definitively determine the cause in the absence of independent data.

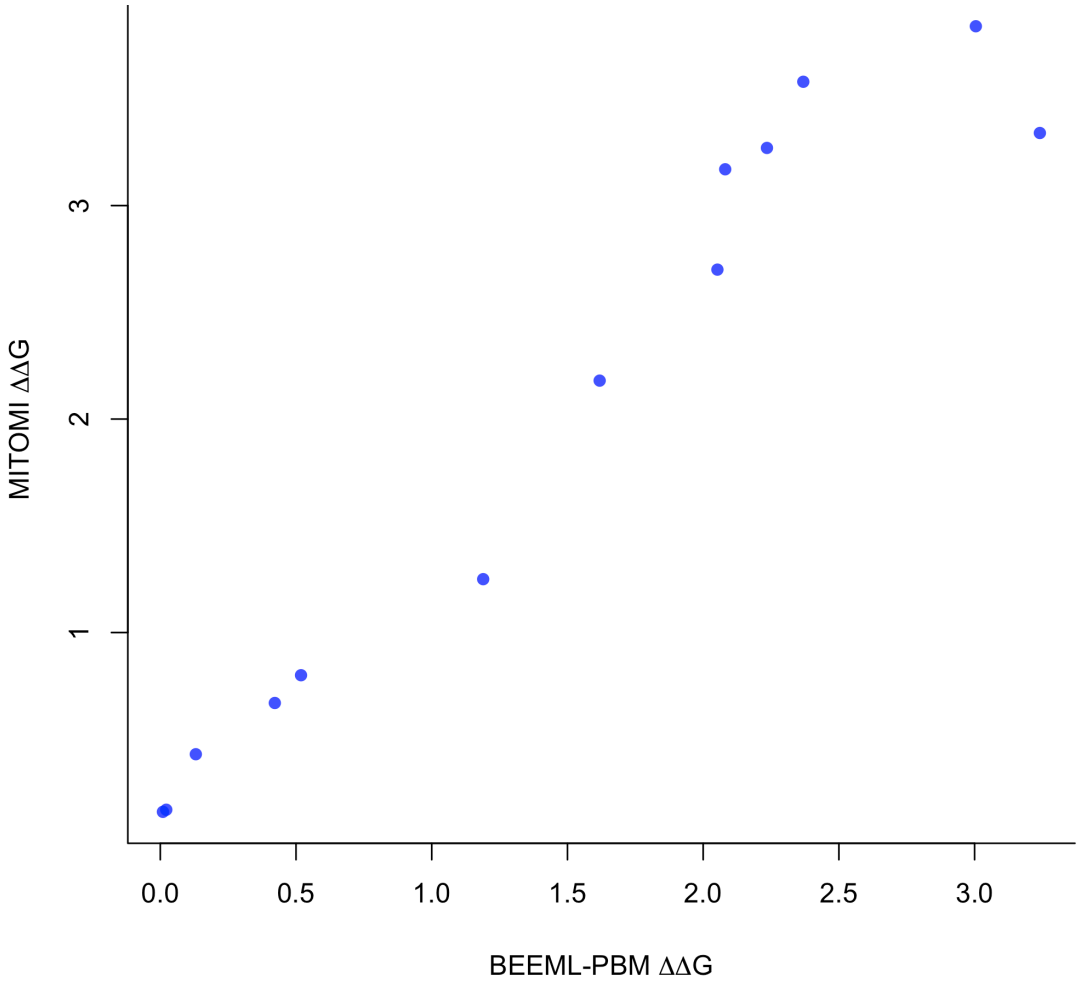


Figure 3.15 Comparison of binding energies measured by MITOMI and those estimated by BEEML-PBM from PBM data for TF Pho4, $r^2 = 0.94$

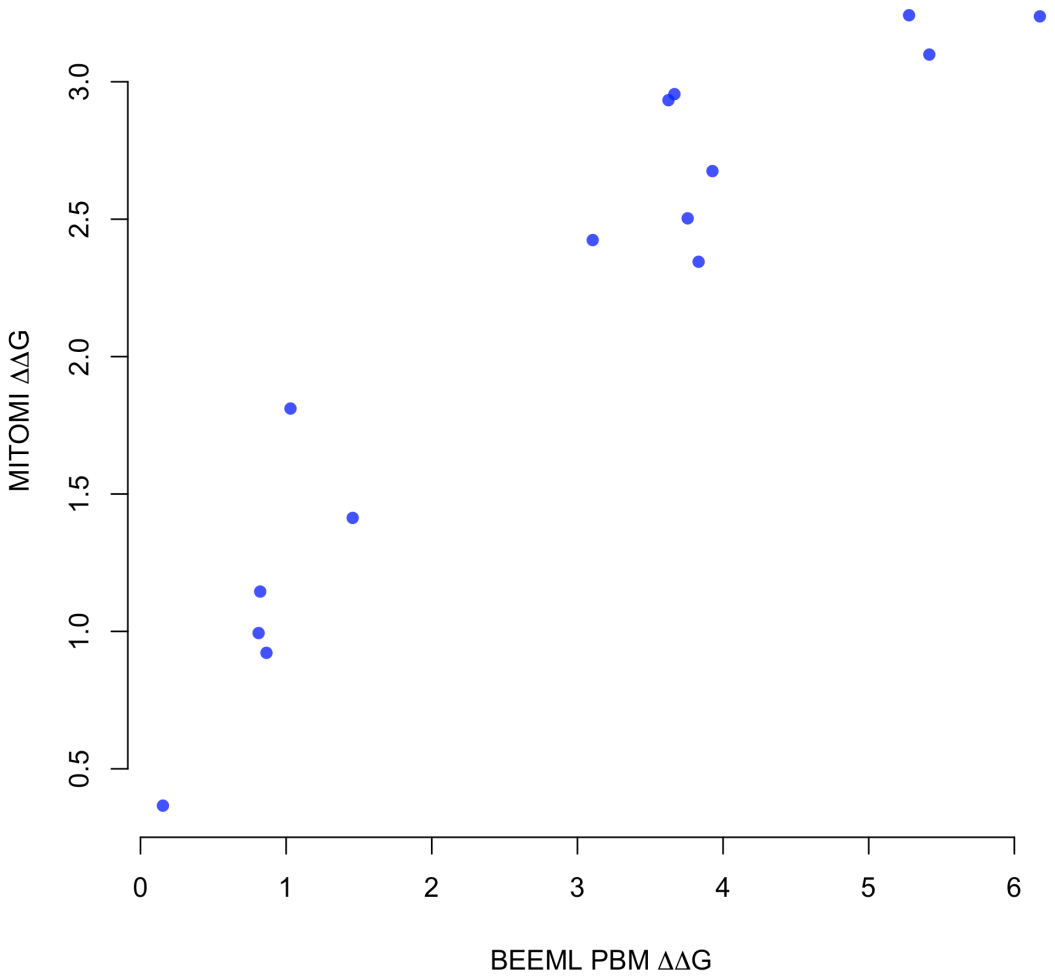


Figure 3.16 Comparison of binding energies measured by MITOMI and those estimated by BEEML-PBM from PBM data for TF Cbf1, $r^2 = 0.88$

PWMs fitted by BEEML-PBM are also able to capture subtle differences between factors with similar specificities. In a survey of mouse homeodomain TF specificities, Berger et al. (Berger et al., 2008) found that some factors in the Lhx (LIM homeobox) family: Lhx2, Lhx3 and Lhx4 have clear, systematic differences in preferences for moderate and low affinity binding sites even though they all bind to the same high affinity sequence TAATTA. BEEML-PBM PWMs are able to recapitulate these differences (figure 3.17 and 3.18), demonstrating that information contained in Lhx factor 8mer intensities can be compressed into the simple PWM model, which is able to capture the important aspects of their sequence specificities.

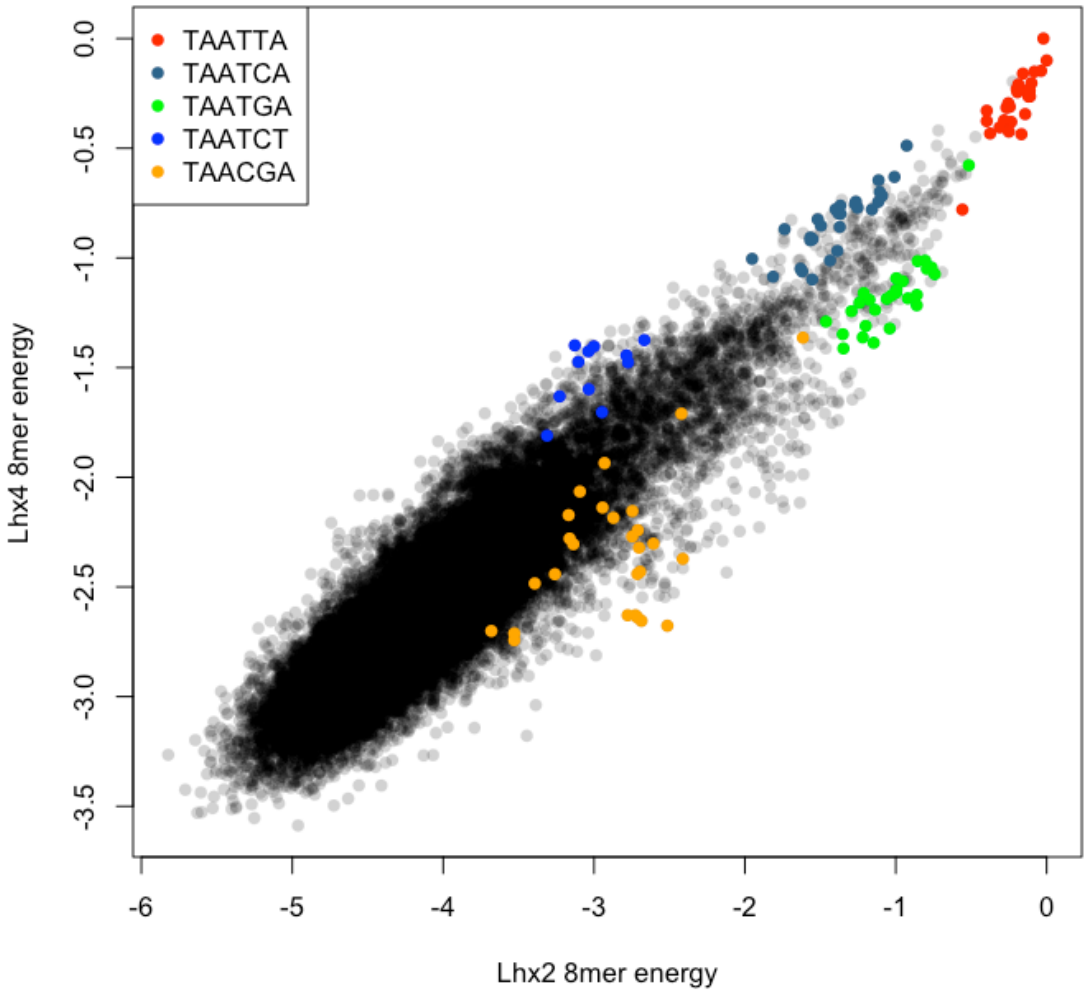


Figure 3.17 Comparison of binding energies according to Lhx2 and Lhx4 BEEML-PBM PWMs. 8mers containing different 6mer sequences are colored to show the systematic differences found by Berger et al. (Berger et al., 2008) is recovered by BEEML-PBM.

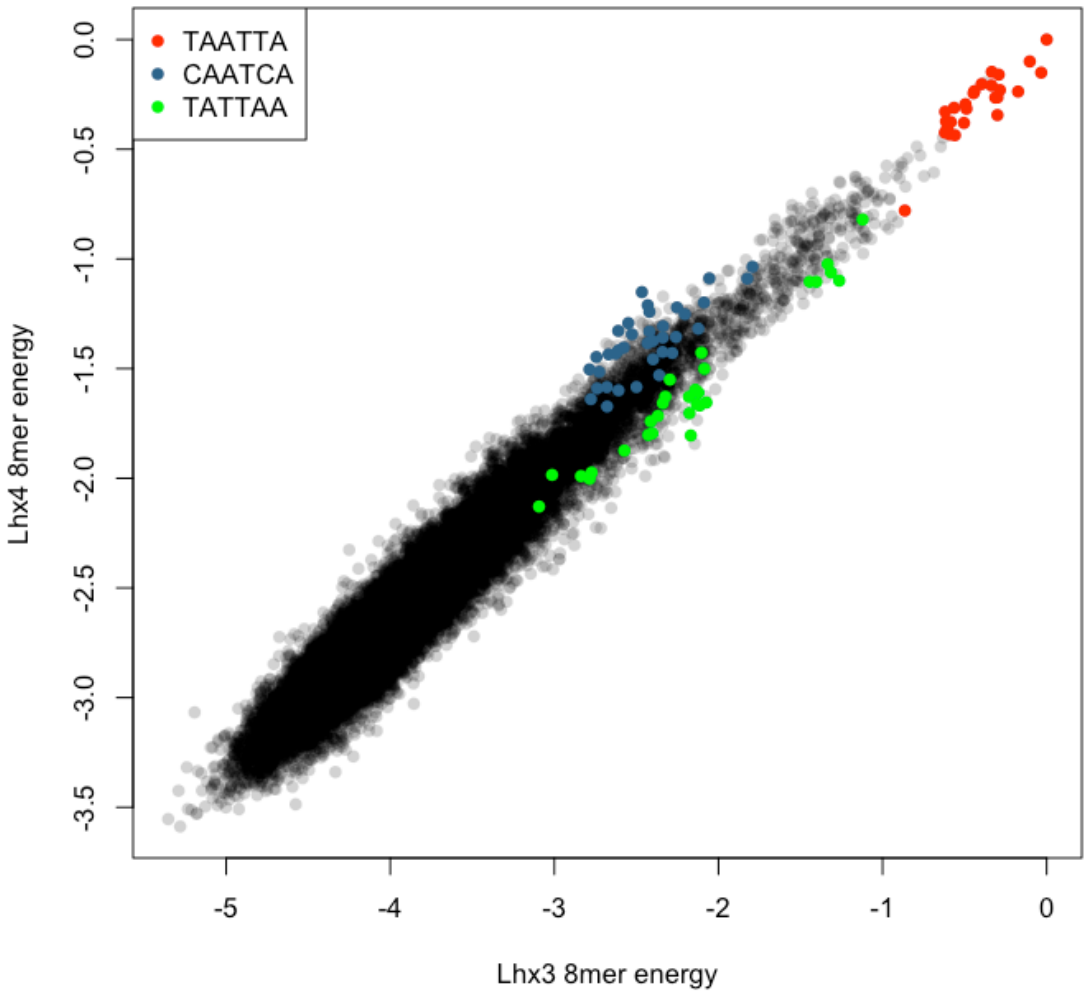


Figure 3.18 Comparison of binding energies according to Lhx3 and Lhx4 BEEML-PBM PWMs. 8mers containing different 6mer sequences are colored to show the systematic differences found by Berger et al. (Berger et al., 2008) is recovered by BEEML-PBM.

3.5 PWM is a good approximation for most TFs

PBM experiments are usually performed using two arrays with different probe sequences, but both contain all possible 10-long binding sites. We use the PWM trained on one array to predict probe intensities on the other array to assess the performance of the PWM. This is then compared with experimental reproducibility between the two arrays to determine if PWM is a good approximation for TF specificity.

We use BEEML-PBM to estimate all free parameters (PWM, chemical potential and scaling parameters) from the training array and use them to predict probe intensities of the test array using equation (2.14). Quality of prediction is measured by r^2 between predicted and measured 8mer median intensities. PWM performance is calculated as the average r^2 of training on array 1, test on array 2 and vice versa. Experimental reproducibility is simply the r^2 between 8mer median intensities measured on array 1 and array 2.

Although 8mer median intensities are problematic as measures of binding affinity, they serve as a useful measure of how much of the observed sequence-dependent binding variation is experimentally reproducible. We find that a single BEEML-PBM PWM is usually sufficient to provide excellent quantitative descriptions of PBM data. An example of this is shown in figure 3.19 for mouse factor Plagl1 (pleomorphic adenoma gene-like 1), where the PWM estimated from replicate 1 performs very well on replicate 2 data, $r^2 = 0.91$. By

contrast, the primary PWM found by Badis et al. (Badis et al., 2009) is unable to capture much of Plagl1 binding specificity, with r^2 =0.47 (figure 3.20), leading them to conclude that multiple PWMs are required. The BEEML-PBM PWM is qualitatively different from the primary PWM identified by Badis et al. (Badis et al., 2009) (figure 3.21); given the high level of performance achieved by a single BEEML-PBM PWM it is likely that the need for multiple PWMs identified by Badis et al. (Badis et al., 2009) is due to suboptimal parameterization rather than the complexity of Plagl1 DNA recognition.

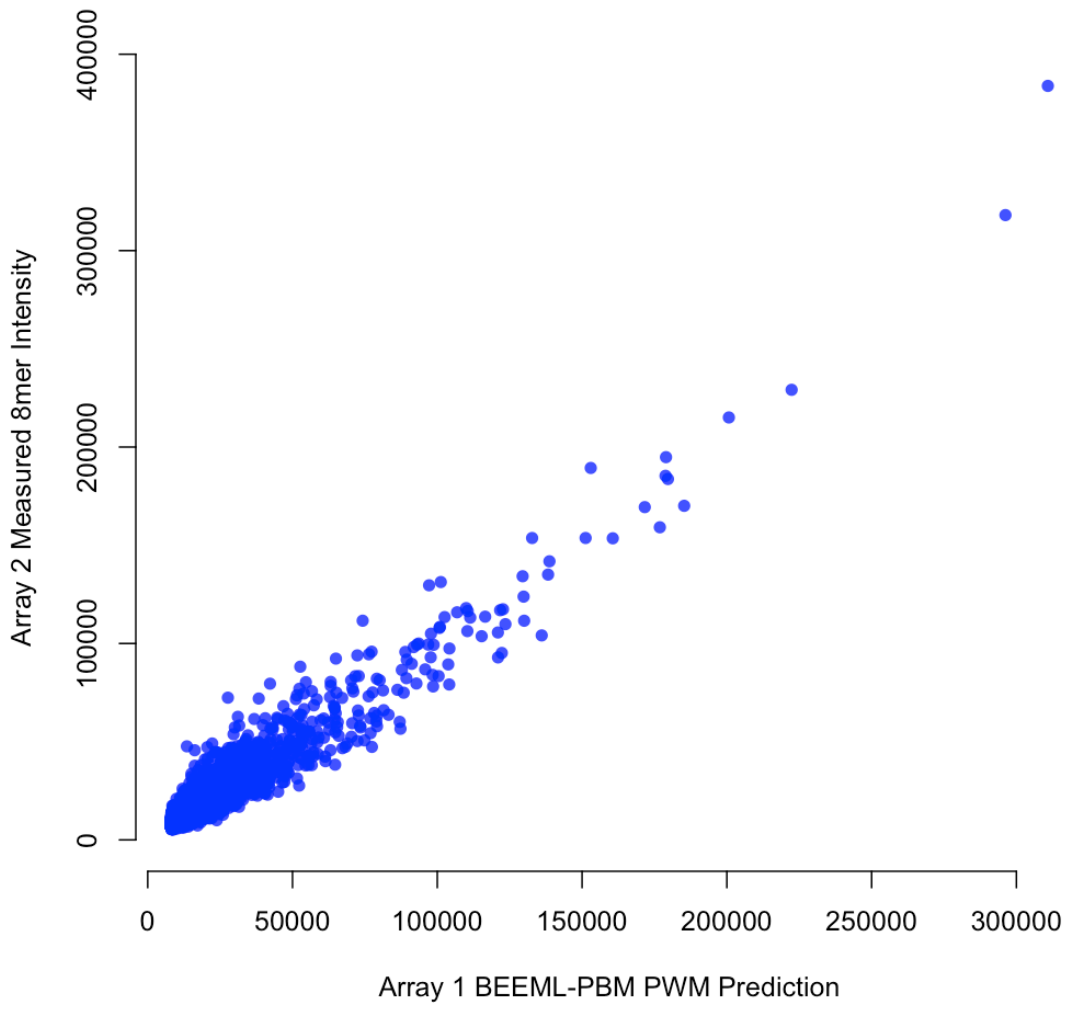


Figure 3.19 BEEML-PBM PWM trained on Plagl1 replicate 1 predicts replicate 2 8mer median intensities well, $r^2 = 0.91$

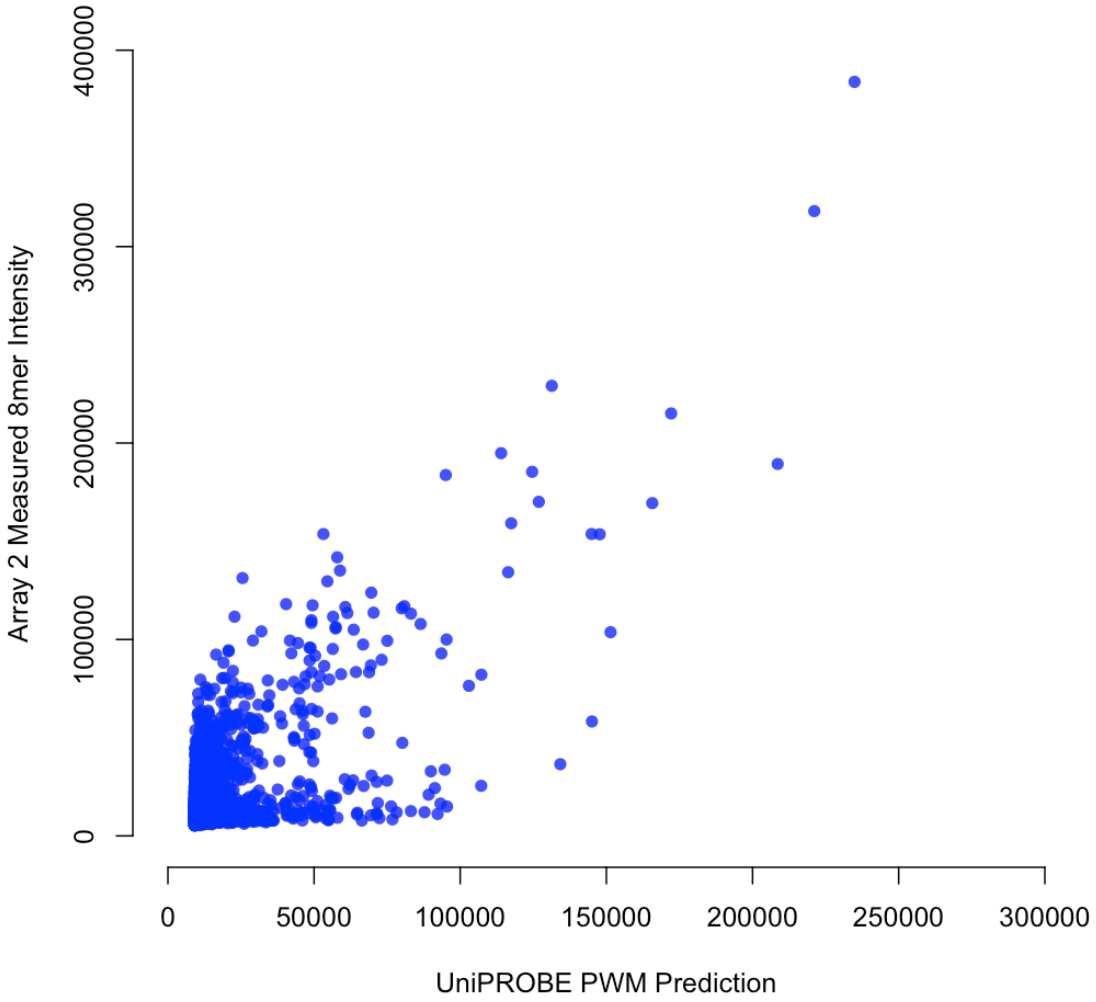


Figure 3.20 UniPROBE PWM for Plagl1 obtained by Badis et. al. (Badis et al., 2009) does not fit the data well, $r^2 = 0.47$

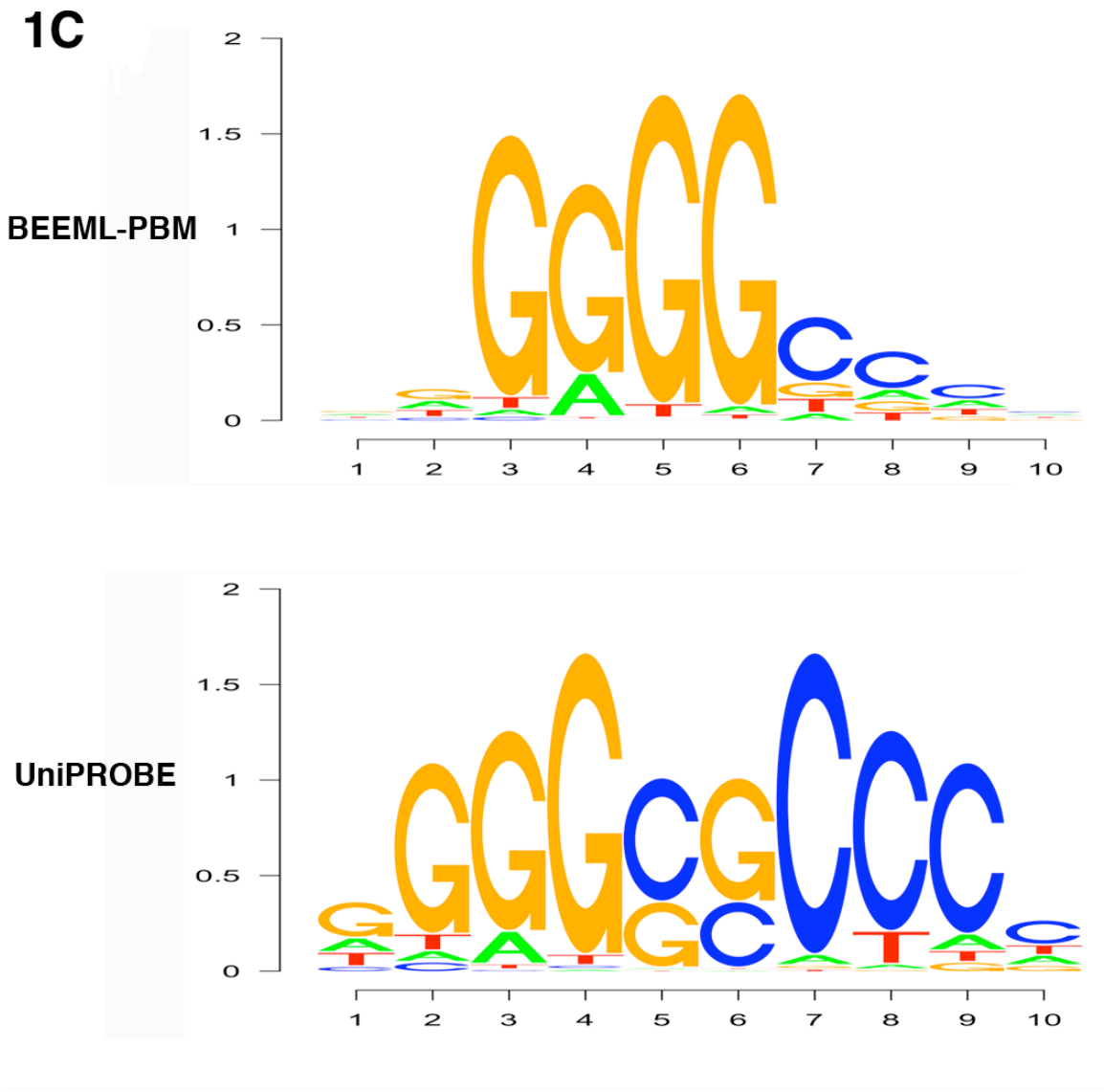


Figure 3.21 PWMs obtained by BEEML-PBM and UniPROBE for Plagl1 are qualitatively different.

Badis et. al. (Badis et al., 2009) identified 41 TFs as having clear secondary binding preferences. Figure 3.22 shows that in all but 7 cases, a single PWM explains more than 90% of the experimental variability for these factors. In some cases, PWM performances are better than experimental reproducibility, likely due to the non-linearity that may be introduced if different TF concentrations were used in replicate PBM experiments. Figure 3.23 demonstrates that for these 41 TFs, a single BEEML-PBM PWM usually performs as well as, and sometimes better than, a combination of primary and secondary PWMs in the UniPROBE database. Figure 3.24 shows that in all of the 104 PBM datasets of Badis et al. (Badis et al., 2009), the PWMs obtained by the BEEML-PBM method fit the replicate data better than the UniPROBE primary PWMs, in many cases very much better. Badis et al. (Badis et al., 2009)validated binding to secondary motifs of six TFs by electrophoretic mobility shift assay (EMSA). We find that the BEEML-PBM PWMs are usually shorter than the PWMs found by Badis et al. (Badis et al., 2009), and that our PWMs are often consistent with the EMSA results. For example, the consensus sequence of the BEEML PWM for TF Foxj3 is AAACA, which can be found on both primary (GTAAACAA) and secondary (CAAAACAA) probes. However, there are also a few cases, such as Hnf4a, where the single PWM model is clearly insufficient to capture TF binding specificity.

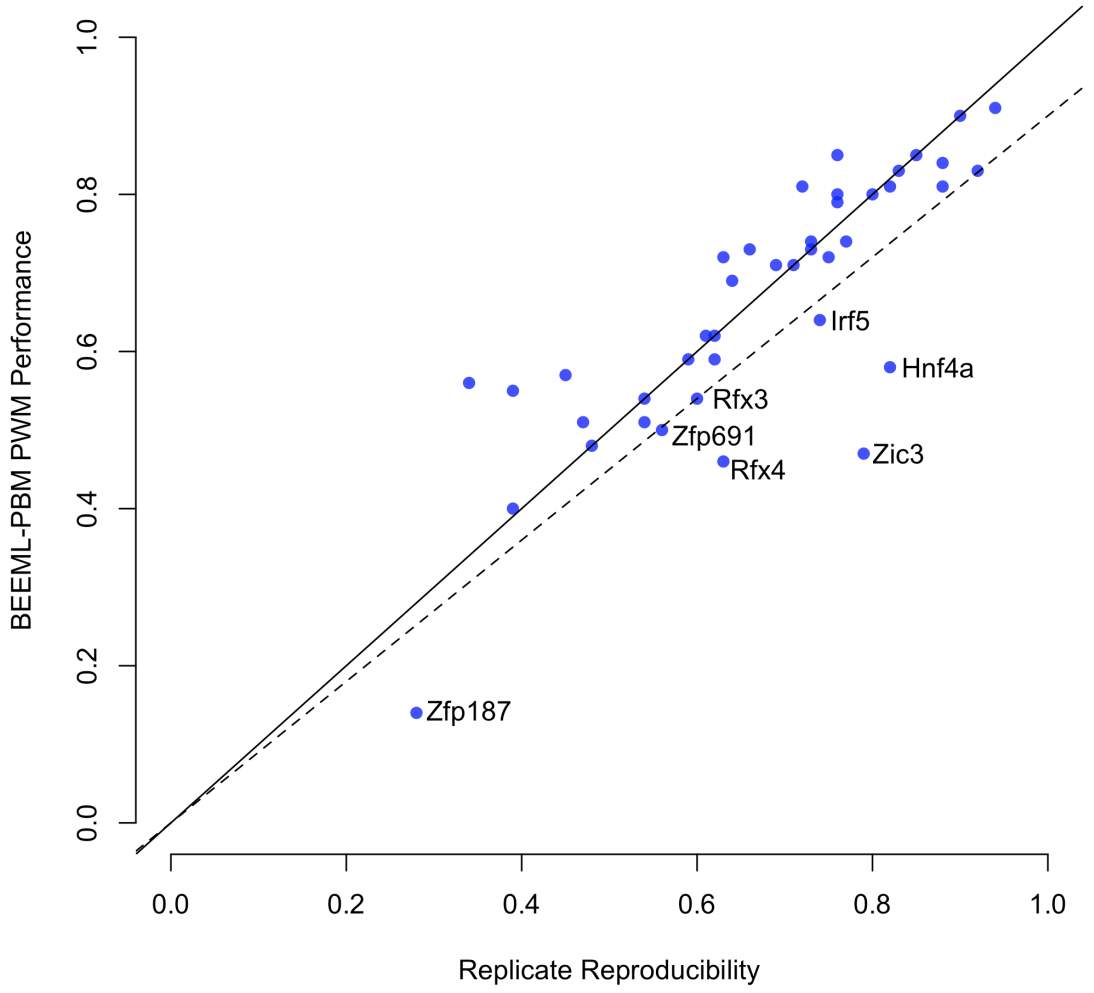


Figure 3.22 A single BEEML-PBM PWM explains "secondary motif" phenomenon. In all but 7 cases, BEEML-PBM PWM captured more than 90% of experimentally reproducible variability of the 41 TFs claimed to have secondary binding modes by (Badis et al., 2009). Dashed line marks 90%.

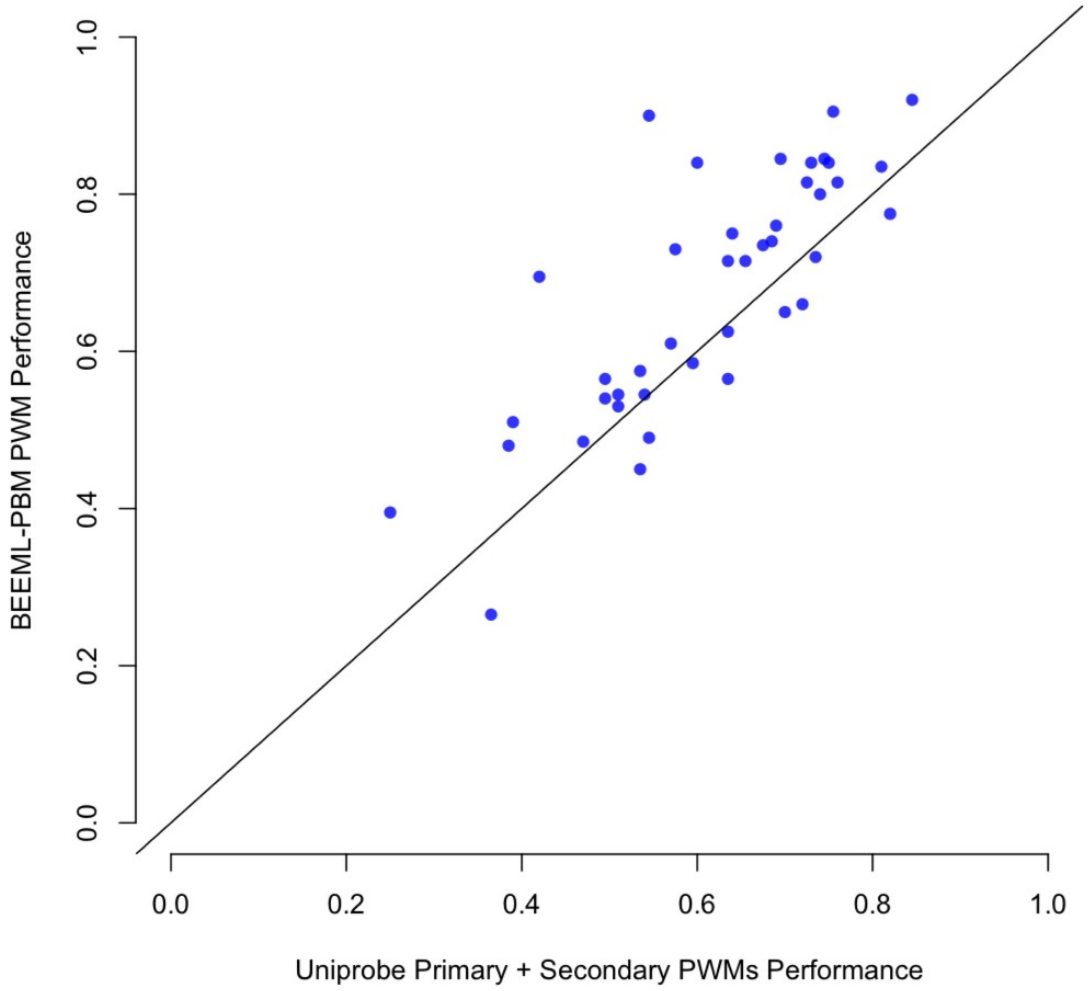


Figure 3.23 A single BEEML-PBM PWM usually outperforms a combination of primary and secondary PWMs for the 41 TFs claimed to have secondary binding modes by Badis et al. (Badis et al., 2009)

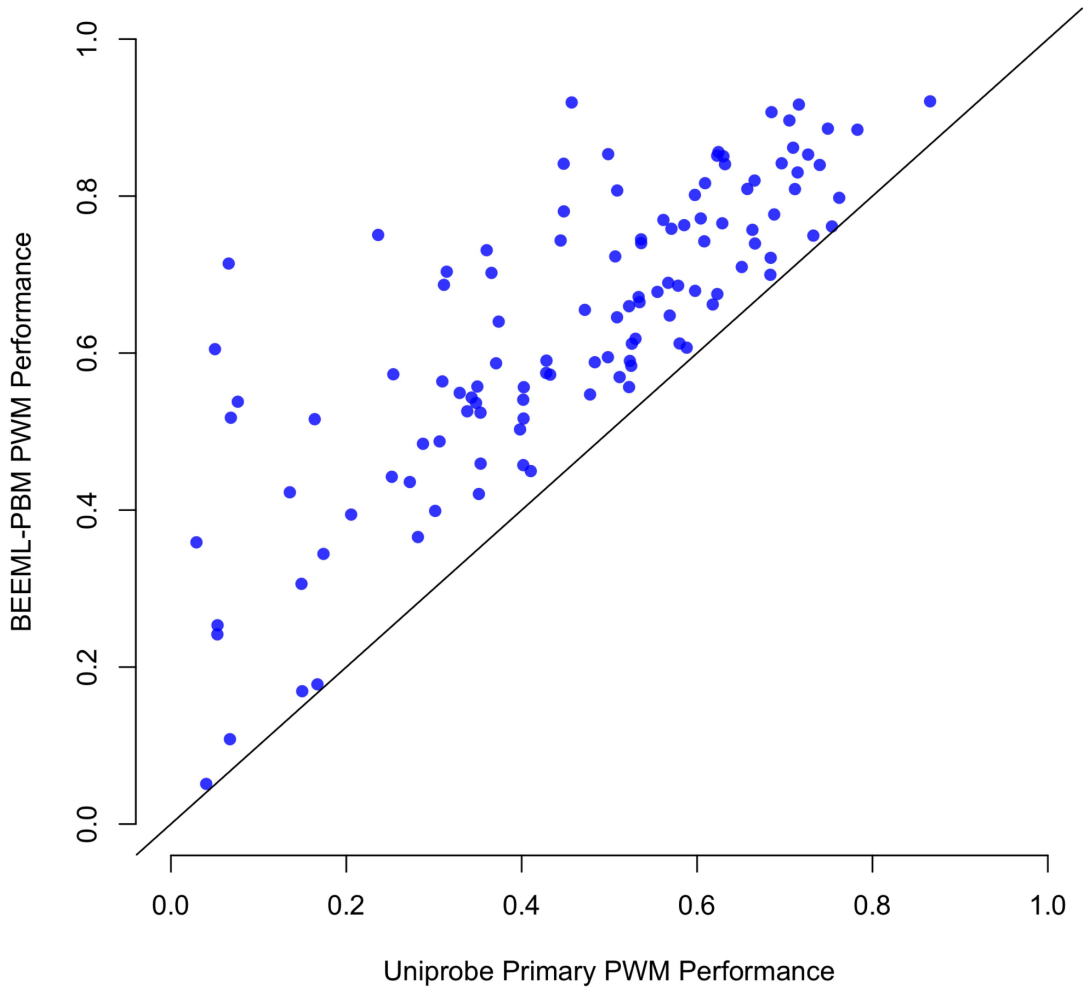


Figure 3.24 BEEML-PBM PWMs outperforms primary PWMs from Badis et al. (Badis et al., 2009) for all 105 TFs in that dataset.

3.6 Analysis of pair-wise dependencies in binding sites

Although the PWM model is a good approximation for the quantitative specificity of many TFs, the true interaction is often more complex. There are many examples in the literature of a single amino acid binding to multiple bases simultaneously (Luscombe et al., 2001), the overwhelming majority of which are neighboring bases. There are also examples where the TF backbone is flexible enough such that a single amino acid can be positioned to form hydrogen bonds with bases in different positions (Lamoureux & Glover, 2006). These types of complex interactions are strongly dependent on the conformation of DNA, which is also sequence-dependent. The use of DNA conformation as a recognition mechanism often called "indirect readout" or "conformational recognition" (Drew & Travers, 1985; Otwinowski et al., 1988). Drastic DNA deformation have been observed in for some TF-DNA complexes, including catabolite gene activator protein (CAP) and TATA binding protein (TBP) (Kim et al., 1993; Schultz et al., 1991). Identity of bases in neighboring positions are particularly important for DNA deformation energy through their stacking interactions.

In addition to structural analysis (Dickerson, 1998; Luscombe et al., 2001; Olson et al., 1998; Suzuki & Yagi, 1995; Werner et al., 1996), detailed biochemical studies of specific

proteins have also shown dependencies between adjacent positions (Bulyk et al., 2002; Man & Stormo, 2001). Interactions between non-adjacent bases are possible (Jacobson et al., 1997), but they appear to be much more rare than interactions between adjacent positions (Luscombe et al., 2001). Statistical analyses of collections of known binding sites have also offered evidences of correlations between non-adjacent positions (Sharon et al., 2008; Zhou & Liu, 2004).

Pairwise interactions can be easily incorporated into BEEML framework (equation 2.4). Higher order interactions can also be included, but they are impractical due to the large number of parameters required. Even when the model is restricted to pairwise interactions, the number of parameters can be large: for a 8-long binding site, there are 28 possible distinct pairs of positions, each requiring 9 parameters, for a total of 252 parameters that must be estimated in addition to the PWM. Since interactions occur mostly between adjacent positions, perhaps it is sufficient to consider only interactions between nearest neighbor positions. In this case, 7 pairs of adjacent positions need to be considered for a total of 63 parameters.

An example of nearest neighbor BEEML analysis is shown in figures 3.25-27 for C4 zinc finger TF Hnf4a (Hepatic nuclear factor 4, alpha, also see figure 3.22) from Badis et al. dataset (Badis et al., 2009). Figure 3.25 shows that an 8-long PWM trained on array 1 is unable to accurately predict 8mer median intensities on the test array ($r^2 = 0.55$). The model

including nearest neighbor interactions performed much better, achieving a r2 of 0.78 (figure 3.26), close to the experimental reproducibility of 0.82 (figure 3.27).

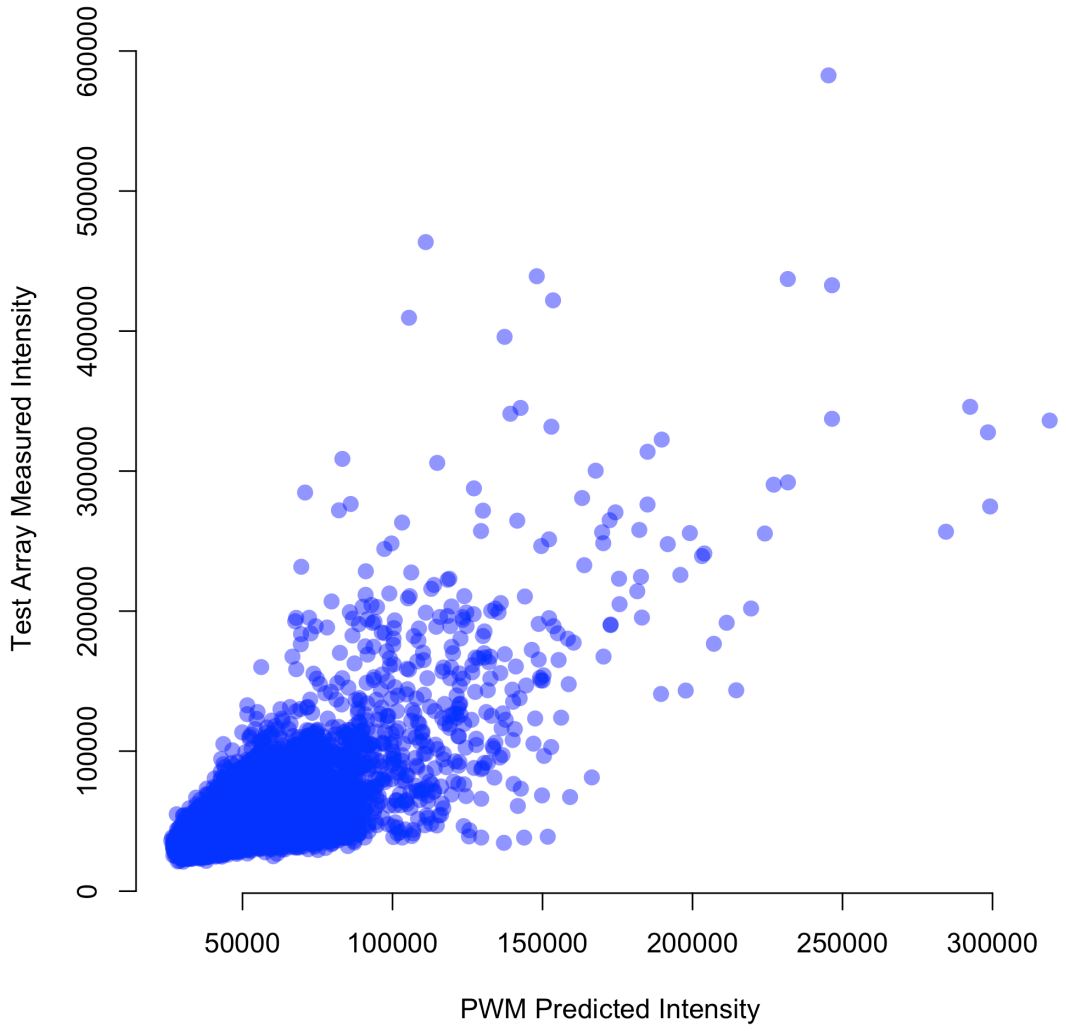


Figure 3.25 PWM is unable to accurately model Hnf4a specificity, $r^2 = 0.55$

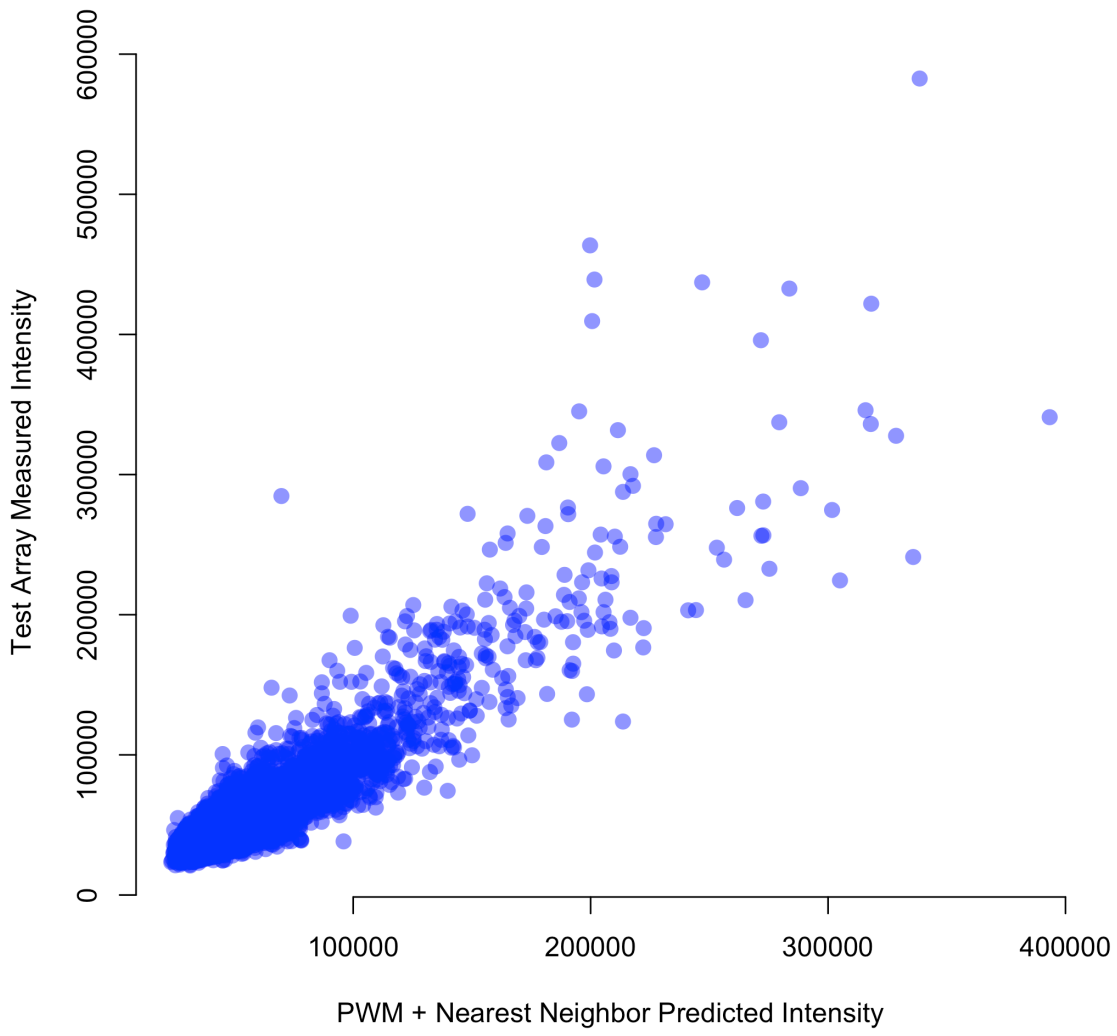


Figure 3.26 Energy model that include both PWM and nearest neighbor interactions captures Hnf4a specificity, r2 = 0.78

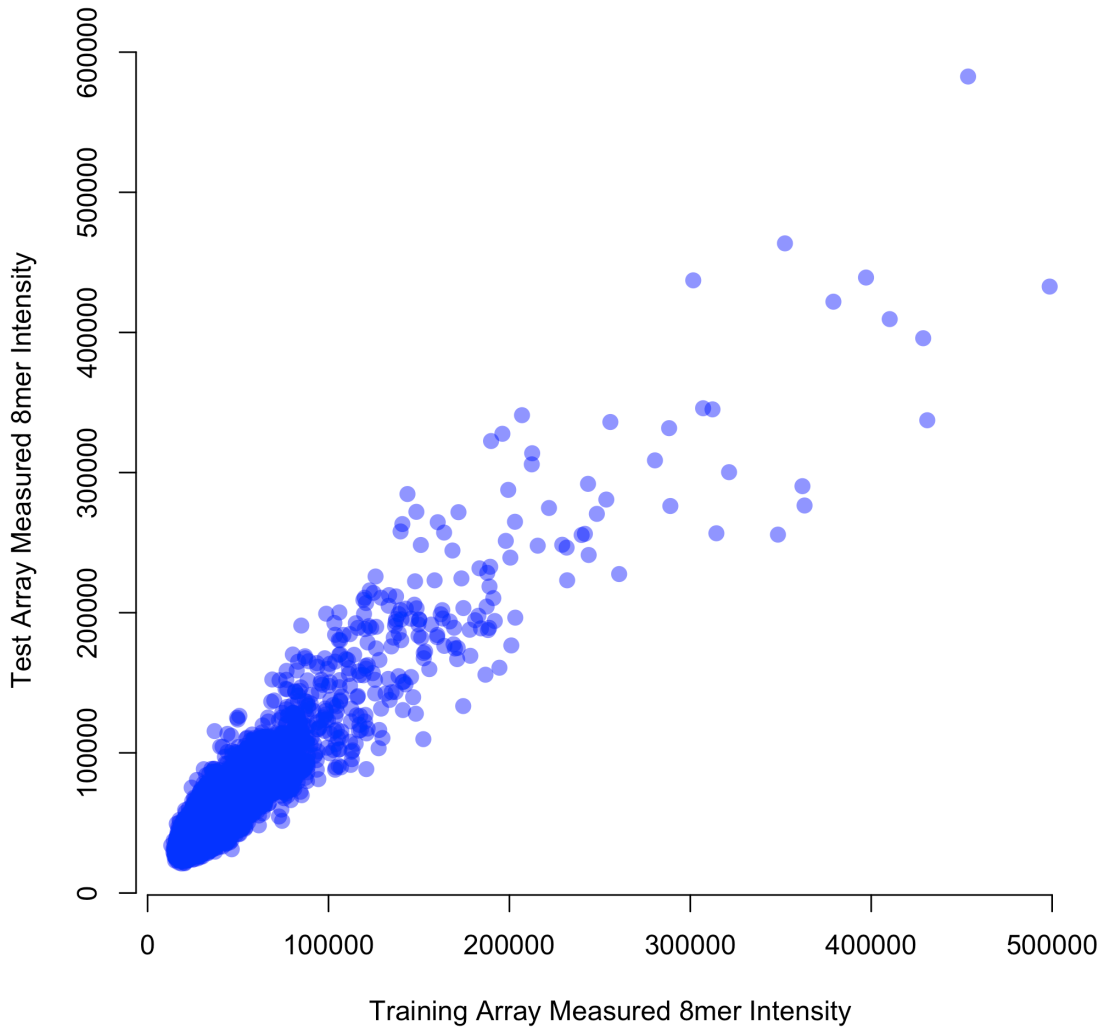


Figure 3.27 Experimental reproducibility of Hnf4a PBMs. Y-axis is the same as in figures 3.25 and 3.26, X-axis is the median 8mer intensity of the training array, $r^2 = 0.82$

Figure 3.28 shows that the PWM model is unable to explains more than 90% variance for 25 out of 147 TFs in the UniPROBE database (Robasky & Bulyk, 2011) with replicate data available. Predictive performance was substantially improved with the addition of nearest neighbor interactions (figure 3.29), indicating that the majority of interactions not captured by PWM are between adjacent positions.

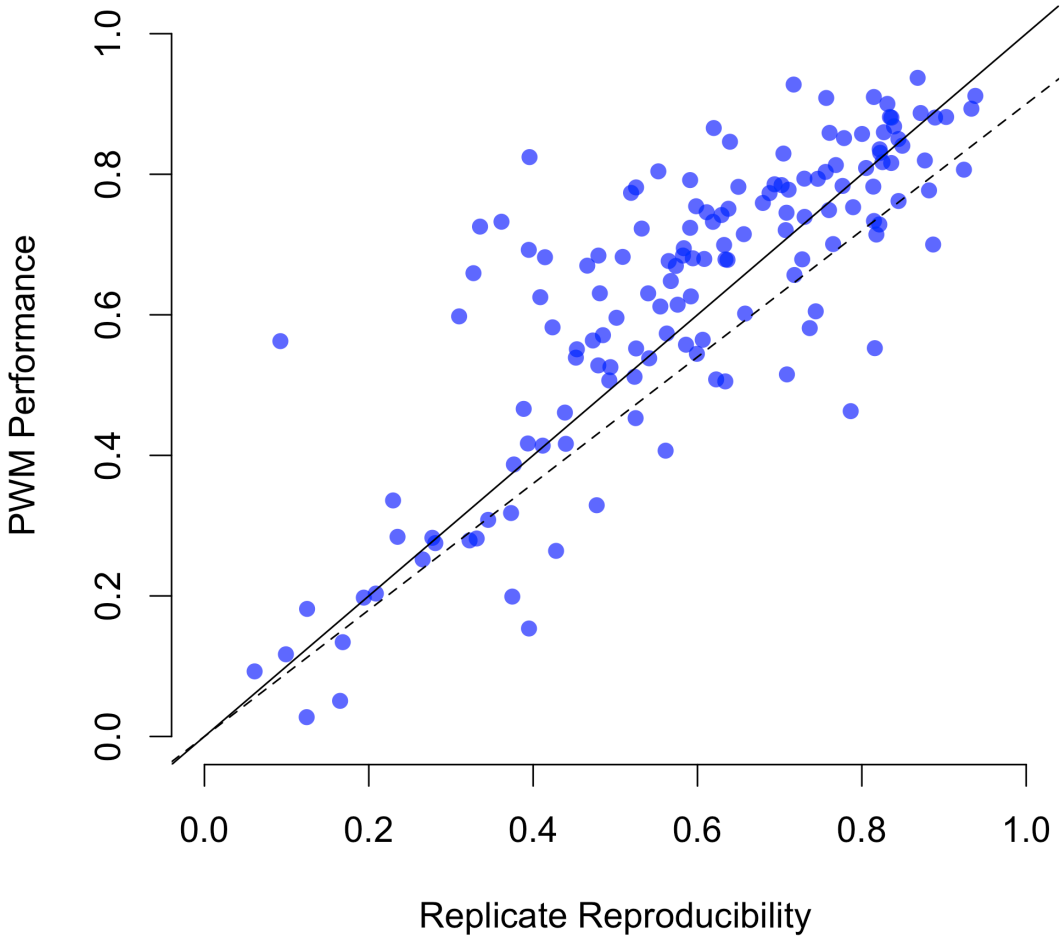


Figure 3.28 PWM model performs well for most TFs but insufficient for many

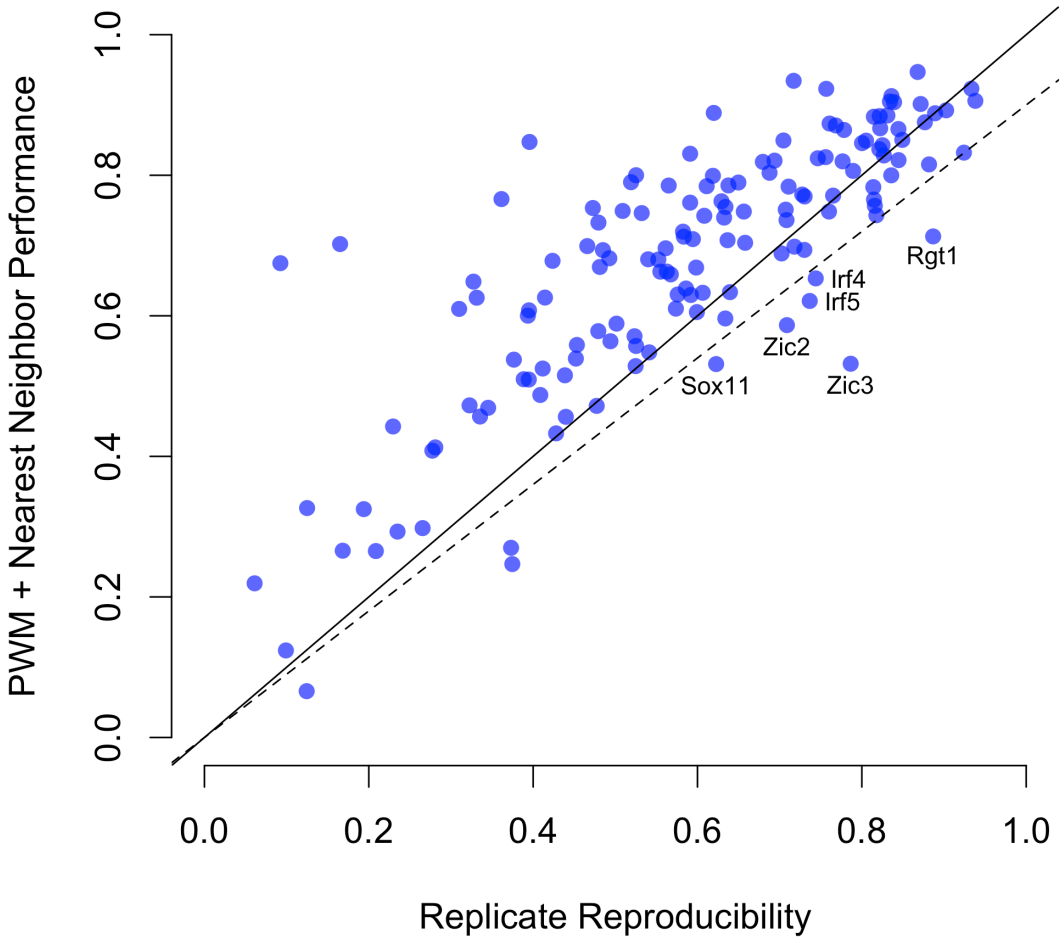


Figure 3.29 most deviations from strict additivity can be accounted for by the addition of nearest neighbor interactions

For a more global perspective, I compared the fit of PWM with those of nearest neighbor and random interaction models for all 401 TFs in the UniPROBE database (Robasky & Bulyk, 2011). Random interaction model include PWM as well as interaction energies for randomly chosen non-adjacent position pairs in the binding site and its performance is calculated as the average of 10 random interaction assignments. The reason for including random interaction model is two fold: first, it allows us to assess the importance of nearest neighbor interactions. Second, since replicate data is not available, the performance of random interaction model, which has the same number of parameters as nearest neighbor model, gives an indication of the extent to which performance gain is simply due to these models having more parameters than PWM.

Figure 3.30 shows the result of this comparison. For most TFs, addition of interaction parameters did not substantially improve the fit. Furthermore, nearest neighbor model always outperformed random interaction model, demonstrating the importance of nearest neighbor interactions. Although nearest neighbor and random interaction models have more parameters and therefore should always outperform PWM model, the local optimization procedure used by BEEML sometimes fail to find optimal parameter values, resulting in some points falling below the main diagonal of figure 3.30.

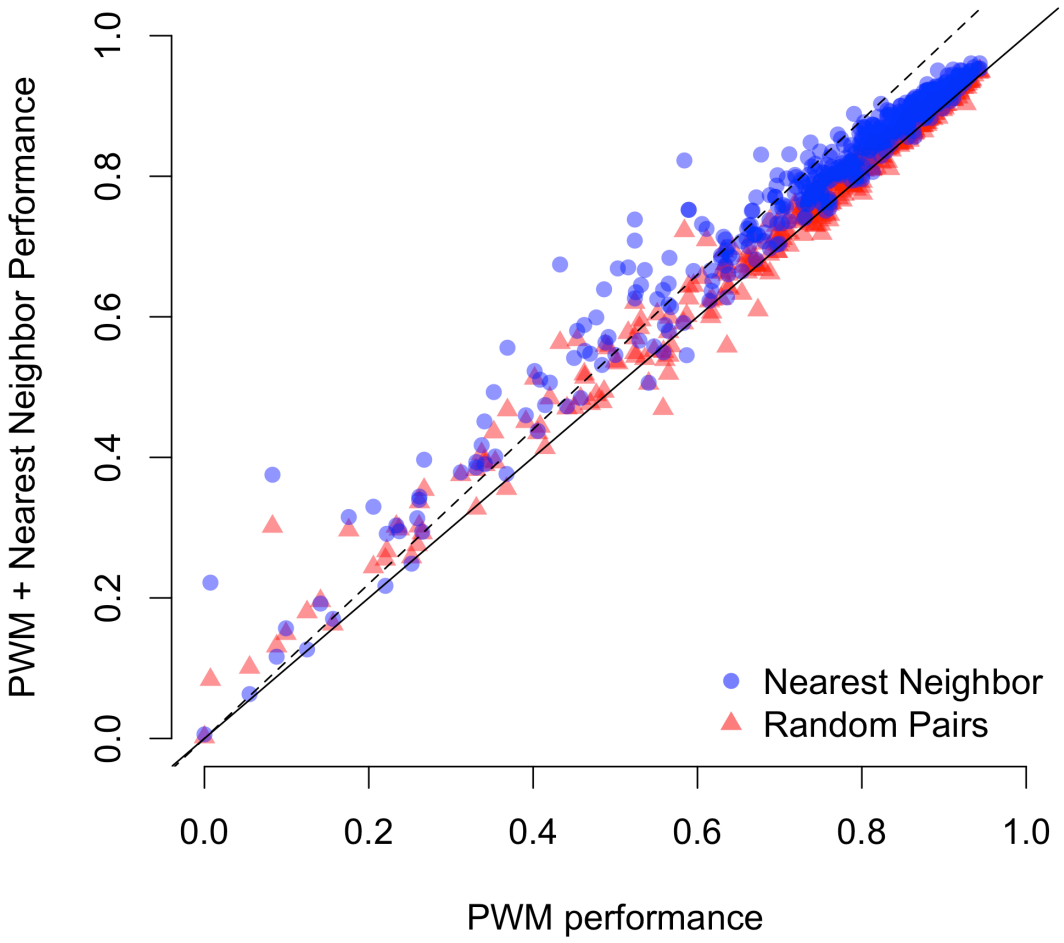


Figure 3.30 Nearest neighbor model outperforms PWM and random interaction model

I next examined nearest neighbor model performances for different TF structural classes (Figures 3.31-34). There are 209 helix-turn-helix TFs in the dataset, including homeodomain and winged helix-turn-helix such as ETS domain TFs. Addition of interaction parameters typically resulted in relative small gains in performance, but there are many cases where nearest neighbor interactions are important (Figure 3.31). This pattern holds for true for zinc finger class, which has 89 members, including C2H2, C4, C6 and GATA zinc finger domains (Figure 3.32). The 25 TFs of zipper class, including basic leucine zipper domain and helix-loop-helix domain, appear to have benefitted the most from the inclusion of nearest neighbor interactions (figure 3.33). By contrast, none of the 24 HMG (High Mobility Group) TFs benefitted substantially from the additional parameters (figure 3.34).

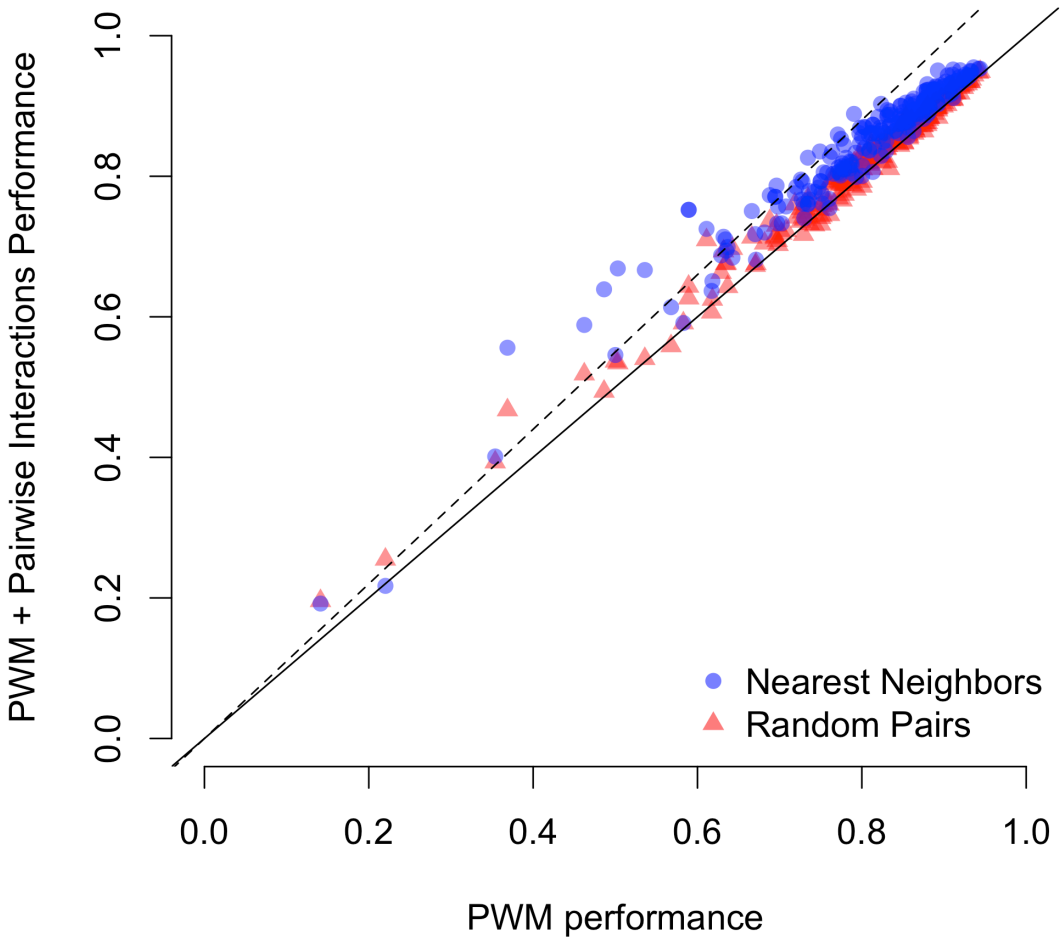


Figure 3.31 Comparison of nearest neighbor and random interaction model with PWM fit for 209 helix-turn-helix TFs (homeodomain, winged helix-turn-helix domains)

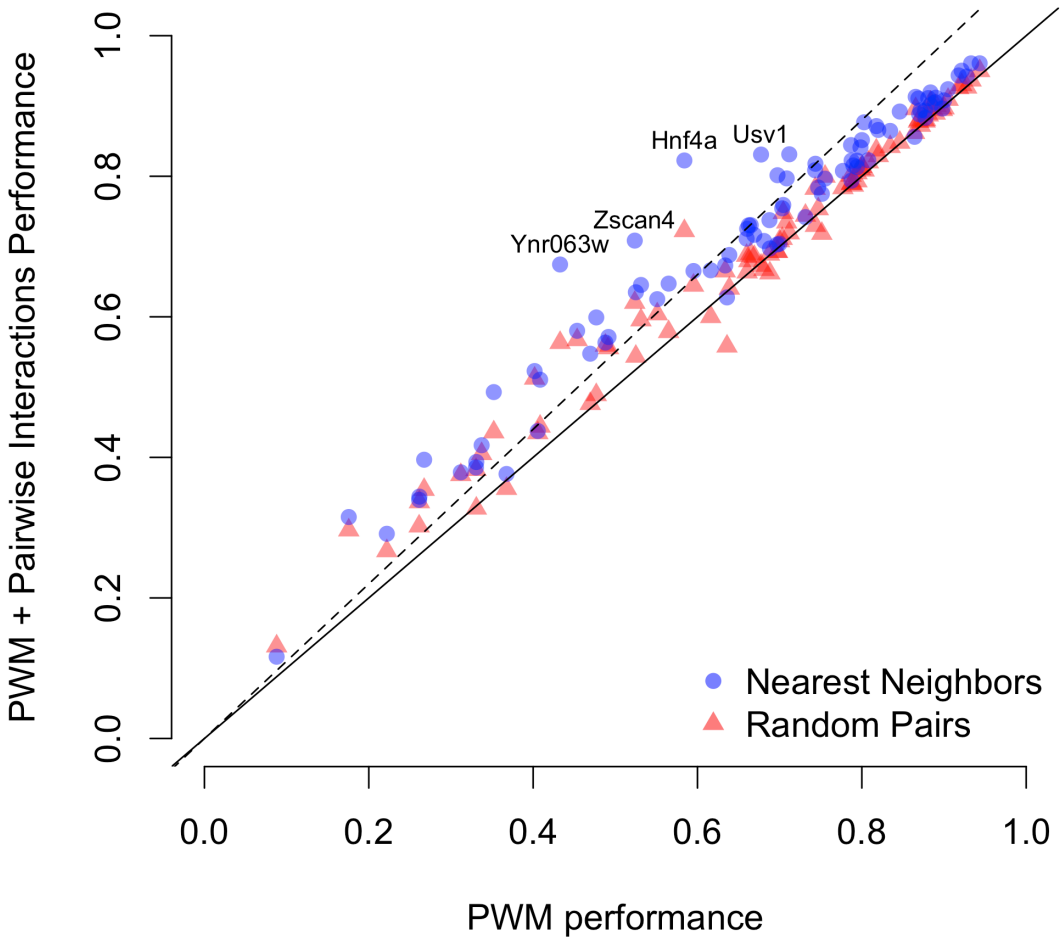


Figure 3.32 Comparison of nearest neighbor and random interaction model with PWM fit for 89 zinc finger TFs (C2H2, C4, C6, GATA zinc finger domains)

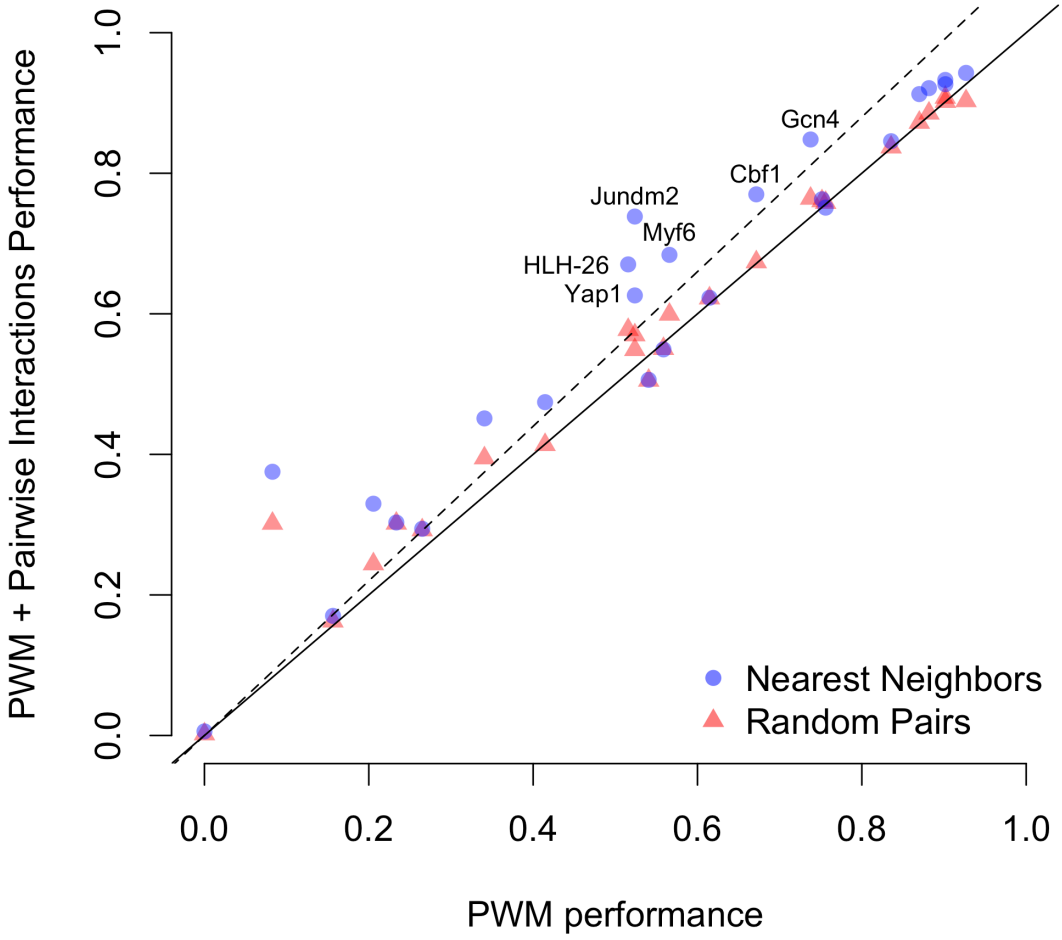


Figure 3.33 Comparison of nearest neighbor and random interaction model with PWM fit for 25 zipper-type TFs (basic leucine zipper and helix-loop-helix domains)

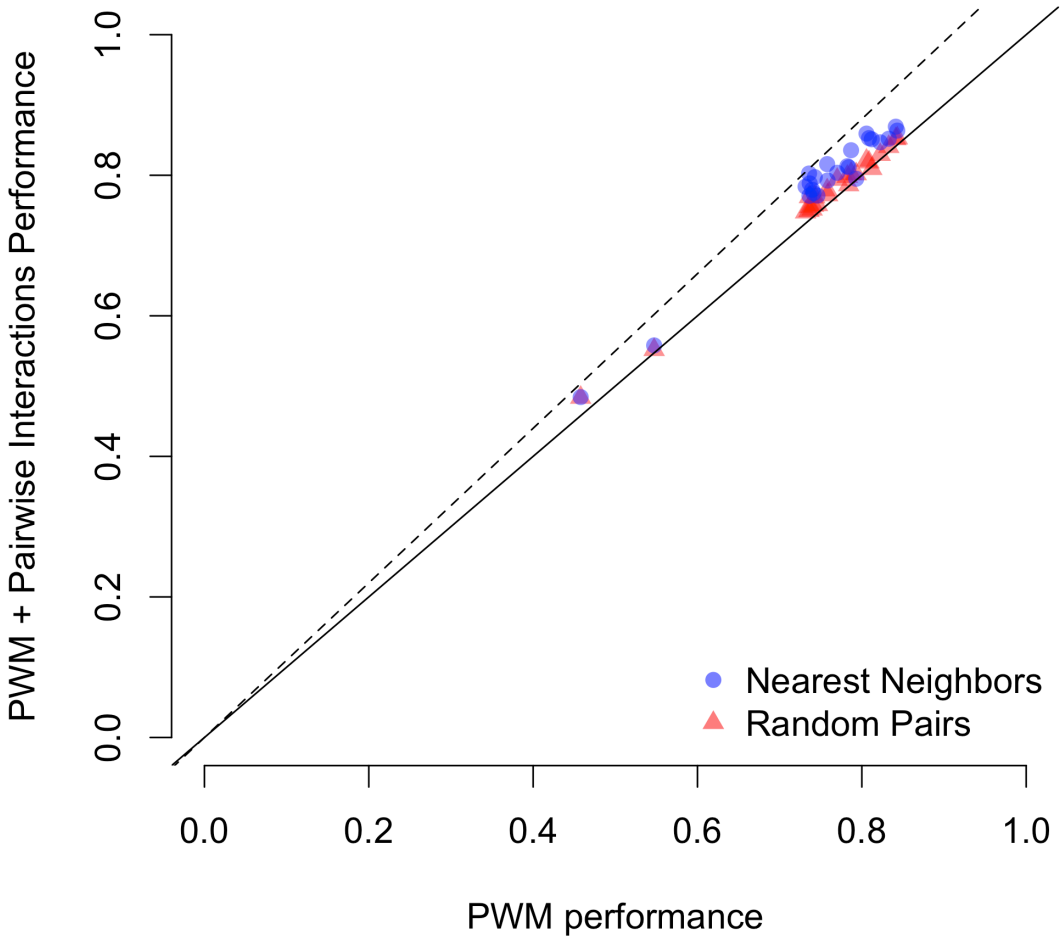


Figure 3.34 Comparison of nearest neighbor and random interaction model with PWM fit for 24 HMG TFs (High Mobility Group domains, including SOX family TFs)

3.7 Conclusion and future directions

The rapid development of high-throughput experimental techniques has resulted in unprecedented amounts of *in vitro* quantitative binding data. In addition to provide insights into TF specificity, this wealth of data has enhanced our understanding of *in vivo* TF binding (Gordân et al., 2009; Wei et al., 2010), transcriptional regulatory networks (Grove et al., 2009; Rowan et al., 2010) as well as regulatory mutations involved in disease (Alibés et al., 2010; Hirsch et al., 2010). However, existing ad hoc analysis methods suffers from a number of drawbacks, leading to conclusion that TF-DNA recognition is highly complex.

In this thesis, I developed a statistical analysis method BEEML (Binding Energy Estimation by Maximum Likelihood) that is based on sound biophysical principles. BEEML energy model is flexible and can be extended to include features important for TF binding, such as non-specific binding or interactions between positions. I have shown that BEEML method can be used to analyze quantitative binding data from a variety of high-throughput experimental methods. Using BEEML, I demonstrated that the PWM is a good model for most TF specificities, indicating the energetics of TF-DNA recognition is simple.

A number of improvements can make BEEML an even better tool for the study of TF specificity. First, better statistics. The arbitrary cutoff of 90% variance explained was useful to show that simple models performed well, but a more principled approach will be

needed to identify which interactions are significant. Statistically, this can be framed as the problem of variable selection, a trade off involving goodness of fit on one hand and model complexity on the other hand. A variety of existing methods can be used for this task, including Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (Schwarz, 1978) (BIC), Bayes factor(Zhou & Liu, 2004) and cross-validation based methods. However, the current implementation of BEEML is very computationally intensive (it takes ~6 hours to estimate the parameters of a single nearest neighbor model). Moreover, the parameter values estimated by BEEML are not guaranteed to optimal, which makes model selection difficult.

Second, more complex motif models need to be developed to model specific mechanisms of TF recognition. For example, many TFs bind as dimers, sometimes with variable spacing between halfsites. The halfsites themselves can be direct repeats or palindromic. There are existing computational methods for handling variable length between halfsites (Bi & Rogan, 2004; Cardon & Stormo, 1992), but they need to be adopted to work with quantitative data.

Third, it is possible that modeling DNA conformation explicitly can help us understand TF specificity in more detail. Although our understanding of how the three-dimensional structure of DNA is related to sequence remains incomplete (Peters & Maher, 2010), existing knowledge of sequence-dependent DNA conformation features can be

incorporated into models (O'Flanagan et al., 2005; Steffen et al., 2002). Inclusion of sequence-dependent DNA conformation information in BEEML analysis can help to explain the indirect readout component of TF specificity.

References

1. Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716--723.

2. Alibés, A., Nadra, A. D., De Masi, F., Bulyk, M. L., Serrano, L., & Stricher, F. (2010) Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. *Nucleic Acids Res* 38, 7422-31.

3. Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., & Bulyk, M. L. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720-3.

4. Badis, G., Chan, E. T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C. D., Gossett, A. J., Hasinoff, M. J., Warren, C. L., Gebbia, M., Talukder, S., Yang, A., Mnaimneh, S., Terterov, D., Coburn, D., Li Yeo, A., Yeo, Z. X., Clarke, N. D., Lieb, J. D., Ansari, A. Z., Nislow, C., & Hughes, T. R. (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* 32, 878-87.

5. Barash, Y., Elidan, G., Friedman, N., & Kaplan, T. (2003) Modeling dependencies in protein-DNA binding sites. Proceedings of the seventh annual international conference on Research in computational molecular biology, 28--37.

6. Benos, P. V., Bulyk, M. L., & Stormo, G. D. (2002) Additivity in protein-DNA interactions: how good an approximation is it?. *Nucleic Acids Res* 30, 4442-51.

7. Benos, P. V., Lapedes, A. S., & Stormo, G. D. (2002) Is there a code for protein-DNA recognition? Probab(ilistical)ly. *Bioessays* 24, 466-75.

8. Berg, O. G., Winter, R. B., & von Hippel, P. H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* 20, 6929-48.

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

10. Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Peña-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., Khalid, F., Zhang, W., Newburger, D., Jaeger, S. A., Morris, Q. D., Bulyk, M. L., & Hughes, T. R. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* 133, 1266-76.

11. Berger, M. F. & Bulyk, M. L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* 4, 393-411.

12. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, 3rd, P. W., & Bulyk, M. L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24, 1429-35.

14. Betz, J. L., Sasmor, H. M., Buck, F., Insley, M. Y., & Caruthers, M. H. (1986) Base substitution mutants of the lac operator: in vivo and in vitro affinities for lac repressor. *Gene* 50, 123-32.

15. Bi, C. & Rogan, P. K. (2004) Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucleic Acids Res* 32, 4979-91.

16. Bishop, C. M. Pattern Recognition and Machine Learning. Springer-Verlag, New York, 2007.

17. Blackwell, T. K. & Weintraub, H. (1990) Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection.

Science 250, 1104-10.

18. Bulyk, M. L., Huang, X., Choo, Y., & Church, G. M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A* 98, 7158-63.

19. Bulyk, M. L., Johnson, P. L. F., & Church, G. M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res* 30, 1255-61.

20. Bussemaker, H. J., Li, H., & Siggia, E. D. (2001) Regulatory element detection using correlation with expression. *Nat Genet* 27, 167-71.

21. Cardon, L. R. & Stormo, G. D. (1992) Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments. *J Mol Biol* 223, 159-70.

22. Christensen, R. G., Gupta, A., Zuo, Z., Schriefer, L. A., Wolfe, S. A., & Stormo, G. D. (2011) A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. *Nucleic Acids Res* , .

23. Church, G. M., Sussman, J. L., & Kim, S. H. (1977) Secondary structural complementarity between DNA and proteins. *Proc Natl Acad Sci U S A* 74, 1458-62.

24. Cormen, T. H., Leiserson, C. E., & L, R. R. (1990) Polynomials and the FFT, Introduction to Algorithms, Chapter 32. Cambridge, MA: The MIT Press.

25. Dickerson, R. E. (1998) DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res* 26, 1906-26.

26. Dill, K. A. & Bromberg, S. Molecular Driving Forces: Statistical Thermodynamics in Chemistry & Biology. Garland Science, New York, 2002.

27. Djordjevic, M., Sengupta, A. M., & Shraiman, B. I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res* 13, 2381-90.

28. Drawid, A., Gupta, N., Nagaraj, V. H., Gélinas, C., & Sengupta, A. M. (2009) OHMM: a Hidden Markov Model accurately predicting the occupancy of a transcription factor with a self-overlapping binding motif. *BMC Bioinformatics* 10, 208.

29. Drew, H. R. & Travers, A. A. (1985) DNA bending and its relation to nucleosome positioning. *J Mol Biol* 186, 773-90.

30. Fields, D. S., He, Y., Al-Uzri, A. Y., & Stormo, G. D. (1997) Quantitative specificity of the Mnt repressor. *J Mol Biol* 271, 178-94.

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

32. Fried, M. & Crothers, D. M. (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res* 9, 6505-25.

33. Galas, D. J. & Schmitz, A. (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5, 3157-70.

34. Garner, M. M. & Revzin, A. (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res* 9, 3047-60.

35. Garvie, C. W. & Wolberger, C. (2001) Recognition of specific DNA sequences. *Mol Cell* 8, 937-46.

36. Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G., & Wüthrich, K. (1994) Homeodomain-DNA recognition. *Cell* 78, 211-23.

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

38. Gordân, R., Hartemink, A. J., & Bulyk, M. L. (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res* 19, 2090-100.

39. Granek, J. A. & Clarke, N. D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol* 6, R87.

40. Grove, C. A., De Masi, F., Barrasa, M. I., Newburger, D. E., Alkema, M. J., Bulyk, M. L., & Walhout, A. J. M. (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* 138, 314-27.

41. Harrison, S. C. (1991) A structural taxonomy of DNA-binding domains. *Nature* 353, 715-9.

42. Heumann, J. M., Lapedes, A. S., & Stormo, G. D. (1994) Neural networks for determining protein specificity and multiple alignment of binding sites. *Proc Int Conf Intell Syst Mol Biol* 2, 188-94.

43. Hirsch, H. A., Iliopoulos, D., Joshi, A., Zhang, Y., Jaeger, S. A., Bulyk, M., Tsichlis, P. N., Shirley Liu, X., & Struhl, K. (2010) A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases. *Cancer Cell* 17, 348-61.

44. Homsi, D. S. F., Gupta, V., & Stormo, G. D. (2009) Modeling the quantitative specificity of DNA-binding proteins from example binding sites. *PLoS One* 4, e6736.

45. Jacobson, E. M., Li, P., Leon-del-Rio, A., Rosenfeld, M. G., & Aggarwal, A. K. (1997) Structure of Pit-1 POU domain bound to DNA as a dimer: unexpected arrangement and flexibility. *Genes Dev* 11, 198-212.

46. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., & Taipale, J. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* 20, 861-73.

47. Jordan, S. R. & Pabo, C. O. (1988) Structure of the lambda complex at 2.5 A resolution: details of the repressor-operator interactions. *Science* 242, 893-9.

48. Kim, Y., Geiger, J. H., Hahn, S., & Sigler, P. B. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature* 365, 512-20.

49. Kinney, J. B., Tkacik, G., & Callan, Jr, C. G. (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci U S A* 104, 501-6.

50. Lam, K. N., van Bakel, H., Cote, A. G., van der Ven, A., & Hughes, T. R. (2011) Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. *Nucleic Acids Res* , .

51. Lamoureux, J. S. & Glover, J. N. M. (2006) Principles of protein-DNA recognition revealed in the structural analysis of Ndt80-MSE DNA complexes. *Structure* 14, 555-65.

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

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

54. Liu, X., Brutlag, D. L., & Liu, J. S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* , 127-38.

55. Liu, X. & Clarke, N. D. (2002) Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding

affinities. *J Mol Biol* 323, 1-8.

56. Luscombe, N. M., Austin, S. E., Berman, H. M., & Thornton, J. M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol* 1, REVIEWS001.

57. Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* 29, 2860-74.

58. Luscombe, N. M. & Thornton, J. M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* 320, 991-1009.

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

60. Man, T. K. & Stormo, G. D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res* 29, 2471-8.

61. Mandel-Gutfreund, Y., Schueler, O., & Margalit, H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J Mol Biol* 253, 370-82.

62. Matthews, B. W. (1988) Protein-DNA interaction. No code for recognition. *Nature* 335, 294-5.

63. Meng, X., Brodsky, M. H., & Wolfe, S. A. (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat Biotechnol* 23, 988-94.

64. Meng, X. & Wolfe, S. A. (2006) Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat Protoc* 1, 30-45.

65. More, J. J. (1977) The Levenberg-Marquardt Algorithm: Implementation and Theory. *Lecture Notes in Mathematics*, Springer Berlin 1978 105-116.

66. Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., & Bulyk, M. L. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 36, 1331-9.

67. Newburger, D. E. & Bulyk, M. L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 37, D77-82.

68. Noyes, M. B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M. H., & Wolfe, S. A. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res* 36, 2547-60.

69. O'Flanagan, R. A., Paillard, G., Lavery, R., & Sengupta, A. M. (2005) Non-additivity in protein-DNA binding. *Bioinformatics* 21, 2254-63.

70. Oliphant, A. R., Brandl, C. J., & Struhl, K. (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol* 9, 2944-9.

71. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M., & Zhurkin, V. B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 95, 11163-8.

72. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., & Sigler, P. B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* 335, 321-9.

73. Pabo, C. O. & Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?. *J Mol Biol* 301, 597-624.

74. Peters, 3rd, J. P. & Maher, L. J. (2010) DNA curvature and flexibility in vitro and in vivo. *Q Rev Biophys* 43, 23-63.

75. Philippakis, A. A., Qureshi, A. M., Berger, M. F., & Bulyk, M. L. (2008) Design of compact, universal DNA microarrays for protein binding microarray experiments. *J Comput Biol* 15, 655-65.

76. Puckett, J. W., Muzikar, K. A., Tietjen, J., Warren, C. L., Ansari, A. Z., & Dervan, P. B. (2007) Quantitative microarray profiling of DNA-binding molecules. *J Am Chem Soc* 129, 12310-9.

77. Rajewsky, N., Vergassola, M., Gaul, U., & Siggia, E. D. (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 3, 30.

78. Riggs, A. D., Bourgeois, S., & Cohn, M. (1970) The lac repressor-operator interaction. 3. Kinetic studies. *J Mol Biol* 53, 401-17.

79. Robasky, K. & Bulyk, M. L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 39, D124-8.

80. Roider, H. G., Kanhere, A., Manke, T., & Vingron, M. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23, 134-41.

81. Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M., & Bourne, P. E. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res* 39, D392-401.

82. Rowan, S., Siggers, T., Lachke, S. A., Yue, Y., Bulyk, M. L., & Maas, R. L. (2010) Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. *Genes Dev* 24, 980-5.

83. Sarai, A. & Takeda, Y. (1989) Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci U S A* 86, 6513-7.

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

85. Schultz, S. C., Shields, G. C., & Steitz, T. A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science* 253, 1001-7.

86. Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics* 6, 461--464.

87. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., & Gaul, U. (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* 451, 535-40.

88. Sharon, E., Lubliner, S., & Segal, E. (2008) A feature-based approach to modeling protein-DNA interactions. *PLoS Comput Biol* 4, e1000154.

89. Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y., & Hakoshima, T. (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J* 16, 4689-97.

90. Siddharthan, R., Siggia, E. D., & van Nimwegen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1, e67.

91. Sinha, S., Blanchette, M., & Tompa, M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 5, 170.

92. Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 5, 89-96.

93. Steffen, N. R., Murphy, S. D., Tolleri, L., Hatfield, G. W., & Lathrop, R. H. (2002) DNA sequence and structure: direct and indirect recognition in protein-DNA binding.

Bioinformatics 18 Suppl 1, S22-30.

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

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

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

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

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

99. Stormo, G. D. & Zhao, Y. (2007) Putting numbers on the network connections. *Bioessays* 29, 717-21.

100. Suzuki, M. & Yagi, N. (1995) Stereochemical basis of DNA bending by transcription factors. *Nucleic Acids Res* 23, 2083-91.

101. Takeda, Y., Sarai, A., & Rivera, V. M. (1989) Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc Natl Acad Sci U S A* 86, 439-43.

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

103. R Development Core Team, (2011) R: A Language and Environment for Statistical Computing. 2011.

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

105. von Hippel, P. H. (2007) From "simple" DNA-protein interactions to the macromolecular machines of gene expression. *Annu Rev Biophys Biomol Struct* 36, 79-105.

106. Wang, T. & Stormo, G. D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19, 2369-80.

107. Warren, C. L., Kratochvil, N. C. S., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., Phillips, Jr, G. N., & Ansari, A. Z. (2006) Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci U S A* 103, 867-72.

108. Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., Yan, J., Talukder, S., Turunen, M., Taipale, M., Stunnenberg, H. G., Ukkonen, E., Hughes, T. R., Bulyk, M. L., & Taipale, J. (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* 29, 2147-60.

109. Werner, M. H., Gronenborn, A. M., & Clore, G. M. (1996) Intercalation, DNA kinking, and the control of transcription. *Science* 271, 778-84.

110. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüss, M., Reuter, I., & Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28, 316-9.

111. Winter, R. B., Berg, O. G., & von Hippel, P. H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* 20, 6961-77.

112. Winter, R. B. & von Hippel, P. H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Escherichia coli repressor--operator interaction: equilibrium measurements. *Biochemistry* 20, 6948-60.

113. Wolfe, S. A., Nekludova, L., & Pabo, C. O. (2000) DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* 29, 183-212.

114. Wright, W. E., Binder, M., & Funk, W. (1991) Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Mol Cell Biol* 11, 4104-10.

115. Zhao, X., Huang, H., & Speed, T. P. (2005) Finding short DNA motifs using permuted Markov models. *J Comput Biol* 12, 894-906.

116. Zhao, Y., Granas, D., & Stormo, G. D. (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5, e1000590.

117. Zhao, Y. & Stormo, G. D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* , .

118. Zhou, Q. & Liu, J. S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* 20, 909-16.

119. Zhu, C., Byers, K. J. R. P., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., Saulrieta, K., Smith, Z., Shah, M. V., Radhakrishnan, M., Philippakis, A. A., Hu, Y., De Masi, F., Pacek, M., Rolfs, A., Murthy, T., Labaer, J., & Bulyk, M. L. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19, 556-66.

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