Improved models for transcription factor binding site identification using non-independent interactions

Yue Zhao, Shuxiang Ruan, Manishi Pandey and Gary D Stormo

Department of Genetics
Washington University School of Medicine
St. Louis, MO 63108
USA
Abstract

Identifying transcription factor (TF) binding sites is essential for understanding regulatory networks. The specificity of most TFs is currently modeled using position weight matrices (PWMs) that assume the positions within a binding site contribute independently to binding affinity for any site. Extensive, high-throughput quantitative binding assays let us examine, for the first time, the independence assumption for many TFs. We find that the specificity of most TFs is well fit with the simple PWM model, but in some cases more complex models are required. We introduce a binding energy model (BEM) that can include energy parameters for non-independent contributions to binding affinity. We show that in most cases where a PWM is not sufficient, a BEM that includes energy parameters for adjacent di-nucleotide contributions models the specificity very well. Having more accurate models of specificity greatly improves the interpretation of in vivo TF localization data, such as from ChIP-seq experiments.

Introduction

Transcription factor proteins (TFs) function by binding to specific sequences in the genome and activating or repressing the expression of their target genes. Identifying the sequences that each TF binds to can help map out transcriptional regulatory networks as well as predict how genetic variation may disrupt normal gene expression, which is often associated with disease. In vivo TF binding locations can be determined experimentally using techniques such as ChIP-chip (REN et al. 2000) or ChIP-seq (JOHNSON et al. 2007). In addition to its intrinsic specificity for DNA, in vivo TF binding preferences are influenced by other cellular factors such as the presence of cooperating or competing TFs as well as the local chromatin state. It is possible to gain mechanistic insight by comparing the intrinsic TF specificity measured in vitro, which reflects only the bi-molecular interaction between the TF and DNA, with in vivo binding locations, which are influenced by many other cellular factors. For example, in vivo binding to a genomic region without strong binding sites indicates that the TF is either binding indirectly (the TF itself is not bound to DNA, but is in complex with some
other factor that is bound to DNA) or bound to a weak site that is stabilized by a cooperative interaction with another factor (GORDAN et al. 2009).

Currently, the most widely used mathematical representation of TF specificity is the Position Weight Matrix (PWM) model (STORMO 2000). This model assumes the positions within the binding site are independent, the contribution at one position of the binding site to the overall affinity does not depend on the identity of nucleotides in other positions of the site. Despite the restrictions imposed by this strong independence assumption, the PWM model has been successfully used to identify TF binding sites (TFBS) in sets of co-expressed genes (BUSSEMAKER et al. 2001; ROTH et al. 1998; STORMO and HARTZELL 1989; TAVAZOIE et al. 1999) as well as model TF binding site evolution (BRADLEY et al. 2010; DONIGER and FAY 2007; MUSTONEN et al. 2008). Quantitative analysis of high-throughput binding data has also shown that PWMs are a good quantitative model for most TFs (ZHAO et al. 2009; ZHAO and STORMO 2011).

Despite the success of PWM based methods, it has long been recognized that TF-DNA interaction is highly complex structurally. Many cases of a single amino acid interacting with multiple bases simultaneously have been observed in crystal structures of TF-DNA complexes (LUSCOMBE et al. 2001), the overwhelming majority of which are neighboring bases. In addition to direct contact with bases, TFs can recognize DNA sequence indirectly through sequence specific DNA conformations, distortions or water-mediated contacts (ROHS et al. 2010; ROHS et al. 2009; SARAI and KONO 2005). For example, drastic DNA deformations have been observed in some TF-DNA complexes, including catabolite gene activator protein (CAP or CRP) and TATA binding protein (TBP) (KIM et al. 1993; SCHULTZ et al. 1991). The identity of bases in neighboring positions is particularly important for DNA deformation energy through their stacking interactions.

In addition to structural analysis, detailed biochemical studies of specific proteins have also shown dependencies between adjacent positions (BERGER et
al. 2006; Bulyk et al. 2002; Man and Stormo 2001). Interactions between non-
adjacent bases are possible (Jacobson et al. 1997), but they appear to be much
less common than interactions between adjacent positions (Luscombe et al.
2001).

Statistical analyses of collections of known binding sites have also offered
evidence of interactions between positions within binding sites for some TFs.
Several groups (Barash 2003; Tomovic and Oakeley 2007; Zhou and Liu 2004)
have analyzed collections of TF binding sites in the TRANSFAC and JASPAR
databases (Matys et al. 2006; Portales-Casamar et al. 2010) and found
statistically significant correlated positions. Although these studies successfully
identified the existence of correlated positions within TFBS, the type of binding
data then available imposed severe limits on their analyses. The most serious
problem was the small number of known binding sites available. For example, on
average, only 30 binding sites per TF were used in the Tomovic and Oakeley
study (Tomovic and Oakeley 2007). Another problem is that TFBSs collected in
databases such as TRANSFAC (Matys et al. 2006) and JASPAR (Portales-
Casamar et al. 2010) generally do not include affinity data, a sequence is simply
labeled as a binding site. This binarization results in a loss of information and is
especially problematic in the setting of small sample sizes. Finally, because of
the non-linearity between binding probability and binding affinity, the non-
independence in base frequencies in different positions may be observed even
when the binding energy contributions are independent (Djordjevic et al. 2003;
Homsí et al. 2009; Zhao et al. 2009). Therefore analyses that specifically model
the binding energy contributions of each base at each position can better
address the issue of whether they act independently.

Recently, experimental techniques for high-throughput, quantitative
measurements of TF binding specificity have been developed (Stormo and Zhao
2010). The wealth of data generated by these high-throughput experiments
provides us with an opportunity to answer questions left open by previous
analyses. One such unanswered question is the average effect size of
interactions between positions. Even though statistically significant interactions
may exist between positions within the binding site of a TF, if their effect size is small then the PWM model may still provide a good approximation of the true TF specificity (BENOS et al. 2002). However, if the effect size is large then more complex statistical models, requiring additional parameters, must be used to adequately model the specificity of the TF. This has important practical implications as the vast majority of existing bioinformatics software use the PWM model. Models designed to accommodate complex interactions within a binding site have been developed (BARASH 2003; KING and ROTH 2003; SHARON et al. 2008; STORMO 2011; STORMO et al. 1986; ZHANG and MARR 1993; ZHOU and LIU 2004) but have not seen wide adoption.

In this paper, we report the results of a quantitative analysis of more than 400 TF specificity data obtained using the Universal Protein Binding Microarray (PBM) technology (BERGER and BULYK 2009; BERGER et al. 2006) and available in the UniPROBE database (ROBASKY and BULYK 2011). We use the BEEMIL-PBM (ZHAO and STORMO 2011) program to parameterize specificity models of varying complexity and find that improvements from incorporating interactions between positions is usually small, though there are some significant exceptions. Moreover, we find that interactions between neighboring bases are stronger than interactions between non-neighboring bases.

**Material and Methods**

**Model of TF specificity**

We use an equilibrium model of binding (DJORDJEVIC et al. 2003; ZHAO et al. 2009) where the probability that DNA sequence $S_i$ is bound by the TF is given by

$$P(S_i \text{ is bound }) = \frac{1}{1+e^{E(S_i)-\mu}} \quad (1)$$

where $E(S_i)$ is the free energy of the TF binding to $S_i$ and $\mu$ is the chemical potential which is related to the TF concentration.
We introduce the Binding Energy Model (BEM) as a vector of energy contributions, \( \bar{E} \). For any sequence, \( S_i \), the binding energy predicted by the model is 
\[
E(S_i) = \bar{E} \cdot \bar{S}_i,
\]
where \( \bar{S}_i \) is the vector encoding of the sequence \( S_i \) that can include whatever features of the sequence are relevant to its binding energy (Schneider et al. 1984; Sharon et al. 2008; Stormo 2011; Stormo et al. 1986).

If the only relevant features are which bases occur at each position within the binding site, then \( \bar{E} \) will be a PWM with the characteristic that each element is an energy contribution (Djordjevic et al. 2003; Zhao et al. 2009). PWMs have been developed using a variety of techniques. The first use was as a discriminant function learned to separate sequences into two classes (Stormo et al. 1982). Staden introduced a probabilistic version of the PWM (Staden 1984) and a variety of other, mostly ad hoc, methods have also been used. Under certain conditions the probabilistic model is equivalent to an energy model (Berg and von Hippel 1987; Heumann et al. 1994; Lassig 2007; Stormo and Fields 1998), but those conditions are violated at high concentrations of the TF (Djordjevic et al. 2003; Zhao et al. 2009). By using an energy model directly and including the TF concentration as a separate parameter, \( \mu \), one can model the non-linear relationship between the logarithm of binding probability and the binding free energy described in equation (1).

When the energy contributions of each position are independent, \( \bar{E} \cdot \bar{S}_i \) is explicitly:
\[
E(S_i) = \sum_{b=0}^{L} \sum_{m=1}^{L} \epsilon(b,m) \cdot S_i(b,m)
\]
where \( L \) is the length of the binding site, \( \epsilon(b,m) \) are the energy contributions of base \( b \) at position \( m \) and \( S_i(b,m) \) 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 (Stormo 2000; Stormo et al. 1982). If we find that the positions are not independent we can include pairwise interactions between adjacent positions by adding interaction terms to the energy function such that \( \bar{E} \cdot \bar{S}_i \) is:
\[
E(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} \epsilon(b, m)S_i(b, m) + \sum_{m=1}^{L-1} \sum_{n=m+1}^{L} \sum_{b=A}^{T} \sum_{c=A}^{T} \epsilon(b, m, c, n)S_i(b, m, c, n)
\]

where \( \epsilon(b, m, c, n) \) is the energy contribution of having base \( b \) at position \( m \) AND base \( c \) at position \( n \). Since the single base contributions are included explicitly, in the first set of sums, the di-nucleotide contributions explicitly represent the deviations from additivity of the individual bases, the energy residuals not captured by the single base contributions. Higher-order models can be constructed similarly and in each case can be represented as \( \vec{E} \cdot \vec{S}_i \) where both vectors include the relevant elements that contribute to the binding energy.

**Encoding of DNA sequence**

In learning the energy parameters we use the WYK encoding scheme (STORMO 2011), which uses the minimum number of parameters required of the model and enforces the clear separation of the contributions of the individual bases from the contributions of the interactions between the bases. This encoding has the further advantage, especially for some machine learning and optimization methods (STORMO 2011), that all sequence vectors have the same Euclidian length \( |\vec{S}_i| = (\vec{S}_i \cdot \vec{S}_i)^{1/2} \) so the energy of a sequence is proportional to the cosine of the angle between it and the energy vector \( E_i \equiv E(S_i) = \vec{E} \cdot \vec{S}_i = |\vec{E}| |\vec{S}_i| \cos \theta \) where \( \theta \) is the angle between the vectors). For convenience of display and interpretation the models are converted to standard ACGT encoding as described above.

**Scoring sequences with Binding Energy Models**

Given a binding energy model \( \vec{E} \) the program BEMSER scans a sequence and determines the predicted binding energy at every location within the sequence (including both orientations if desired). If \( \vec{E} \) contains only independent base contributions (it is a PWM) then BEMSER returns the same scores as would be obtained using PatSer (HERTZ and STORMO 1999), a commonly used program for scanning sequences with a PWM. But if \( \vec{E} \) also contains energy contributions from combinations of bases, BEMSER uses those contributions as
well to score each potential site (subsequence) in the sequence. BEMSER as well as all of the BEMs are available for download from http://stormo.wustl.edu/TF-BEMs.

**Parameterization of TF specificity model**

The models are all determined from PBM data available from the UniPROBE database (ROBASKY and BULYK 2011). All models were parameterized using the BEEML-PBM algorithm, which has been described in detail previously (ZHAO and STORMO 2011). Briefly, BEEML-PBM employs non-linear regression to obtain the parameters of the model, $E$ and $\mu$, that maximize the fit to the data.

**Evaluation of model performance on PBM data**

Given a binding energy model the fluorescence of each probe is predicted by summing the predicted binding probabilities for each position in the probe sequence in both orientations. Because the individual probe measurements are somewhat noisy, we determine the median fluorescence intensity for all 8mers, each of which occurs 32 times on the array (counting both orientations). We assess the performance of different models by determining the square of the correlation coefficient ($r^2$) between the predicted and measured 8mer intensities. When only one array exists for a specific TF this is useful in determining how much better fit is obtained (increase in $r^2$) for models that include di-nucleotide contributions from those that assume independent base contributions. Because of the large number of 8mers, even very small changes in $r^2$ are statistically significant ($p<0.05$), but may make almost no difference in prediction accuracy.

A subset of 147 TFs, from the complete set of 401 available in UniPROBE, were assayed on two independent arrays with different probe sequences, but each containing all possible 10mers (BERGER and BULYK 2009). In those cases we can assess how well a model obtained from one array predicts the 8mer intensities of the other array for the same TF; we use the average of the two cross-predictions as the performance of the model. The squared correlation coefficient between the observed 8mer intensities for the two arrays is referred to
as the replicate reproducibility. In previous work we considered models whose performance was at least 90% of the replicate reproducibility to provide very good fits to the data (ZHAO and STORMO 2011). But it should be noted that it is possible, in fact common, for the model performance to exceed the replicate reproducibility as we showed in the supplemental materials of that paper. This can occur when one of the two datasets is considerably noisier than the other. The replicate reproducibility can then be fairly low and the predictions using the model from the good array have low $r^2$ on the noisier array. But we showed that the noisy array is still capable of generating fairly good models with much higher $r^2$ on the good array. The average of the two model $r^2$ values can then exceed the replicate reproducibility.

**Analysis of ChIP-seq data**

The DNA binding domain of Hepatocyte nuclear factor 4, alpha (Hnf4a) is completely conserved between human and mouse. Since Badis et al. (BADIS et al. 2009) used only the DNA binding domain of mouse Hnf4a to determine its *in vitro* binding specificity, we used this information to analyze Verzi et al. (VERZI et al. 2010) ChIP-seq data even though it was carried out in human cell lines. Hnf4a ChIP-seq data was downloaded from Gene Expression Omnibus (accession number GSM575227) from the study conducted by Verzi et al. (VERZI et al. 2010) and we used the peak location and summit information as provided. Binding site location analysis was conducted by aligning all peaks by their annotated summit and scoring genomic sequences within 200 bp on each side of the summit using each of the different models for predicting the binding sites. For each model, the predicted lowest energy site within that window was considered the binding site for that peak. In the vast majority of cases the predicted best binding site was the same for all models within that 400 bp window. We then determined the number of reads that overlap that predicted binding site after extending each read by 100 bp on each side. While the choices of 400 and 200 bp windows are somewhat arbitrary they are consistent with typical fragmentation sizes for ChIP-seq experiments. Most importantly, since most of the predicted sites are the same for all of the models, we can directly compare their predicted relative affinities for
each identified peak without the confounding effect of different models predicting different binding sites. For each model the predicted relative binding affinities for the predicted binding site within each peak were calculated as $e^{(E_{\text{con}}-E_{\text{site}})}$ where $E_{\text{con}}$ is the energy of lowest possible site (the consensus sequence for the model) and $E_{\text{site}}$ is the energy of the site being being considered.

Primary and secondary PWMs for Hnf4a obtained by the Seed-and-Wobble method (BERGER et al. 2006) were downloaded from the UniPROBE database (ROBASKY and BULYK 2011) and converted to energy PWMs. Positions with low information content were trimmed off so both PWMs were 8-long, the same size as the BEEML models. The energy of an 8mer was calculated as the minimum of the energies predicted by primary and secondary PWMs. If longer models were used the predicted relative affinities would either stay the same (in cases where the predicted site had the consensus base in the additional position) or decrease (if it had a non-consensus base). So by trimming the UniPROBE models to be 8-long, the differences in the predicted relative affinities between different models is minimized.

Results

Protein binding microarray (PBM) technology uses double-stranded DNA microarrays to measure the binding of TFs to many sequences in a highly parallel fashion (BERGER et al. 2006; BULYK et al. 2001; MUKHERJEE et al. 2004). In the current design (BERGER and BULYK 2009; BERGER et al. 2006) all possible 10 nucleotide long binding sites (10mers) are contained in the sequences of microarray probes. In a recent PBM study, Badis et al. (BADIS et al. 2009) measured the binding specificity of 104 mouse TFs and found many cases where the best PWM model they obtained seemed inadequate to represent TF specificity. In fact the authors invoked a model where TFs could use alternative modes of binding, each represented by independent PWMs, to explain the PBM data. We developed a non-linear regression method, BEEML-PBM, to estimate PWM parameters from PBM data (ZHAO and STORMO 2011). Using BEEML-PBM,
we showed that, contrary to the conclusions of Badis et al. (BADIS et al. 2009), the PWM model provides a good quantitative model of specificity for most of the TFs in their study.

Despite the good performance of the PWM model in general, there were cases where the simple PWM performed poorly. For example, an 8-long BEEML-PBM PWM for Hnf4a trained on the data from one array is able to predict the median 8mer intensities of probes on the test array with only an $r^2 = 0.55$ (Fig. 1a,b), much less than the experimental reproducibility between the training and testing data ($r^2=0.82$). A model that includes interactions between all adjacent positions, which requires an additional 63 parameters over the PWM, results in significantly improved performance ($r^2 = 0.81$). We also tested the performance of all 28 possible 8-long models that include a single pairwise interaction between two positions and found that a model with interaction terms only between positions 4 and 5 is able to achieve an $r^2 = 0.78$ (Fig. 1c,d). This model appears to capture most of the relevant features of Hnf4a binding and includes only 9 more parameters than the PWM.

To determine the biological significance of the position dependence observed in vitro, we compared different specificity models of Hnf4a learned from PBM data with in vivo binding data from the ChIP-seq experiment conducted by Verzi et al. (VERZI et al. 2010). Fig. 2a shows the primary and secondary PWMs obtained for Hnf4a from the UniProbe database (ROBASKY and BULYK 2011). Fig. 2b shows that the vast majority of best binding sites for the ChIP-seq peaks have very low predicted relative affinity using the UniProbe PWMs. Those PWMs are highly specific and the vast majority of peaks do not contain consensus sites for either the primary or secondary motif, or even for variations from the consensus that are predicted to be of high relatively affinity. Fig. 2c shows that more of the peaks have higher predicted relative affinity using the PWM obtained by BEEML-PBM on the same training data (Fig. 1a), consistent with the previous results that those PWMs fit the in vitro PBM binding data better than the PWMs obtained using the Seed-and-Wobble algorithm (ZHAO and STORMO 2011). Fig. 2d shows that the energy model that includes di-nucleotide energy contributions between
positions 4 and 5 (Fig. 1c) has even more predicted high affinity binding sites. Fig. 2e shows the cumulative frequency of the number of reads at increasing relative binding affinities for each of the three models.

We next examined the 147 TFs in the UniPROBE database (Robasky and Bulyk, 2011) for which replicate data are available. Fig. 3a shows that the PWM model is unable to explain more than 90% of the reproducibility for 25 out of these 147 TFs (17%). Note that the models obtained using BEEML-PBM are often better at predicting probe intensities than the reproducibility between arrays. This is because good models can be obtained even from noisy experimental data, as demonstrated in the Foxa2 example of Supplement Figure S4 of Zhao and Stormo (ZHAO and STORMO 2011). For most of the TFs that are not well modeled by simple PWMs, predictive performance was substantially improved with the addition of interaction terms between adjacent positions (Fig. 3b), indicating that the majority of interactions not captured by the PWM are between adjacent positions.

For a more global perspective, we compared the fit of PWMs with those of nearest neighbor and random interaction models for all 401 TFs in the UniPROBE database (Robasky and Bulyk 2011). Nearest neighbor BEMs include the independent base contributions to binding (as in a PWM) as well as interaction energies between all adjacent positions. In this model, an L-long binding site requires $3L + 9(L-1)$ independent parameters, approximately 4 times as many as a simple PWM model that requires $3L$ parameters. The random interaction model is defined as a PWM as well as interaction energies for the same number, $L-1$, of 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 comparing the nearest neighbor to the 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 for many of the TFs, the performance of the random interaction model, which has the same number of parameters as the nearest neighbor model, gives an
Fig. 4a shows the distribution of increases in $r^2$ for all of the nearest neighbor BEMs compared to the PWMs. Fig. 4b plots the PWM performance versus the nearest neighbor di-nucleotide models for all 401 TFs. For most TFs, addition of interaction parameters did not substantially improve the fit; for less than 15% of TFs is the increase in $r^2 > 0.06$. Furthermore, the nearest neighbor model always outperformed the 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 fails to find optimal parameter values, resulting in few points falling below the main diagonal of Fig 4b.

We examined the nearest neighbor model performances for different TF structural classes (Fig. 4c-f). 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 relatively small gains in performance, but there are several cases where nearest neighbor interactions are important (Fig. 4c). This pattern holds for true for the zinc finger class, which has 89 members including C2H2, C4, C6 and GATA zinc finger domains (Fig. 4d). This includes the Hnf4a example analyzed in detail above. The 25 TFs of the zipper class, including the basic leucine zipper (bZIP) and the basic helix-loop-helix (bHLH) domains, appear to have benefitted the most from the inclusion of nearest neighbor interactions (Fig. 4e). By contrast, none of the 24 HMG (High Mobility Group) TFs benefitted substantially from the additional parameters (Fig. 4f). While there are data showing non-independence between positions for at least some HMG proteins ([JAUCH et al. 2011]), those appear to be relatively minor contributions overall, as found previously for several zinc finger proteins ([BENOS et al. 2002; BULYK et al. 2002]).

Discussion
Our quantitative analysis of more than 400 in vitro quantitative TF specificities generated by PBM technology demonstrates that explicitly including interactions within the binding site generally results in relatively small improvements in performance. With a few exceptions, the PWM model provides a good approximation for TF specificity. Consistent with available structural information, when interactions between positions are important most of them are found to occur between adjacent positions in the binding site. Some TF families are more likely to require interaction models than others. In particular the bZIP and bHLH families are commonly fit much better by including adjacent di-nucleotide energy contributions, consistent with previous information (BERGER et al. 2006; MAERKL and QUAKE 2007; NUTIU et al. 2011; STORMO and ZHAO 2007; ZHAO et al. 2009).

Improved specificity models that are based on in vitro binding data can be very useful for assessing how consistent in vivo location data are with the expected binding sites. When predicted genomic binding sites are not observed in ChIP-seq data one can usually assume that those locations are not accessible. But when binding is observed in locations without predicted binding sites, or with only very low predicted affinity sites, that implies either indirect or cooperative binding mediated through some other factor(s) that bind directly to the DNA (GORDAN et al. 2009). Such indirect and cooperative binding events can lead to the discovery of interacting TFs that coordinately control gene expression. But to be confident about which ChIP-seq peaks are not due to direct binding one needs an accurate model for the specificity of the TF. As we show for Hnf4a, different models can lead to quite different conclusions about which peaks contain predicted high affinity sites. For both the BEEML-PBM models, with and without di-nucleotide contributions, a much greater fraction of the ChIP-seq reads can be explained by direct binding, with the di-nucleotide model explaining the most.

In these analyses we have considered a model to be a good fit to the PBM data if it can capture more than 90% of the reproducible variance of the experiment. However, some of the datasets are fairly noisy and it is possible that
cleaner data would show that even the interaction models do not capture the specificity well. For example, the Hnf4a PBM data has reproducibility between the two arrays of only $r^2=0.82$. The BEM with interaction terms between only positions 4 and 5 predicts the test array data with $r^2=0.78$, which is nearly all of the reproducible variance. But if additional data had higher consistency between experiments, it is possible that additional terms would be required to obtain an adequate model. We can only claim that, given the current experimental datasets, in most cases simple PWMs fit the data quite well and in most of the remaining cases an extended BEM, with energy terms for adjacent di-nucleotides, captures most of the remaining variance.

Acknowledgements

We thank all members of the Stormo lab for helpful comments on this work. Funding was provided by NIH grant HG00249 to GDS.

References


Figure Legends

Figure 1. Binding energy model including interactions makes more accurate predictions of in vitro binding specificity than the PWM for Hnf4a. A) Graphical representation of Hnf4a binding energies estimated from PBM data under the PWM model (supplemental figure 1). Negatives of binding energy (in units of RT) are plotted on the y-axis. Energies are normalized such that the average energy at each position is 0. This energy logo is equivalent to the “affinity logo” from (FOAT et al. 2006). B) Performance of model shown in A) on test PBM data. C) Binding energy model estimated from the same training data but including interaction energies between positions 4 and 5 (supplemental figure 2). D) Performance of the energy model including interactions on test PBM data.

Figure 2. Hnf4a energy model including interactions makes more accurate predictions of in vivo ChIP-seq binding data. A) Primary and secondary binding energy models for Hnf4a obtained by UniPROBE (ROBASKY and BULYK 2011) (supplemental figure 3). B) Primary and secondary models predicted affinity of the best site under each peak vs. number of reads overlapping the best predicted sites. C) BEEML-PBM PWM predicted affinities of best sites vs number of reads. D) energy model including interaction between positions 4 and 5 predicted affinities of best sites vs. number of reads. E) Cumulative frequency (fraction of total reads) for each model at increasing predicted relative affinities.

Figure 3. Comparisons of energy model predictions and experimental reproducibility. A) The PWM energy model is unable to explain more than 90% of the reproducibility for 25 out of these 147 TFs. B) Predictive performance was substantially improved with the addition of interaction terms between adjacent positions, indicating that the majority of interactions not captured by the PWM are between adjacent positions.

Figure 4. Comparison of the fit of PWM, nearest neighbor and random interaction energy models. A) histogram of the improvement in r^2 for the nearest neighbor interaction models compared to the PWMs. B-F) The performance of PWM model is plotted on the x-axis, performances of nearest neighbor (circles) and random interaction (triangles) models are plotted on the y-axis. B) Comparisons for all 401 TFs in UniPROBE database. C) Comparisons for the 209 helix-turn-helix TFs in the database. D) Comparisons for the 89 zinc finger TFs, including C2H2, C4, C6 and GATA zinc finger domains. E) Comparisons of 25 zipper class TFs, including basic leucine zipper (bZIP) domain and helix-loop-helix (bHLH) domain. F) Comparisons of the 24 HMG (High Mobility Group) TFs.
Figure 1C
Figure 2A
Figure 2E
Figure 3B
Figure 4A

Improvement from neighboring di-nucleotides
Figure 4B
Figure 4C

Helix-Turn-Helix, 209 TFs

PWM + Pairwise Interactions Performance

PWM performance

Nearest Neighbors
Random Pairs
Zinc Finger, 89 TFs

Figure 4D
Zipper Type, 25 TFs

Figure 4E
HMG Box, 24 TFs

Figure 4f