

A Feature-Based Approach to Modeling Protein-DNA Interactions

Eilon Sharon and Eran Segal

Department of Computer Science,
Weizmann Institute of Science,
Rehovot, 76100, Israel.

{eilon.sharon, eran.segal}@weizmann.ac.il
WWW home page: <http://genie.weizmann.ac.il>

Abstract. Transcription factor (TF) binding to its DNA target site is a fundamental regulatory interaction. The most common model used to represent TF binding specificities is a *position specific scoring matrix* (PSSM), which assumes independence between binding positions. In many cases this simplifying assumption does not hold. Here, we present *feature motif models* (FMMs), a novel probabilistic method for modeling TF-DNA interactions, based on *Markov networks*. Our approach uses sequence *features* to represent TF binding specificities, where each feature may span multiple positions. We develop the mathematical formulation of our models, and devise an algorithm for learning their structural features from binding site data. We evaluate our approach on synthetic data, and then apply it to binding site and ChIP-chip data from yeast. We reveal sequence features that are present in the binding specificities of yeast TFs, and show that FMMs explain the binding data significantly better than PSSMs.

Key words: transcription factor binding sites, DNA sequence motifs, probabilistic graphical models, Markov networks, motif finder.

1 Introduction

Precise control of gene expression lies at the heart of nearly all biological processes. An important layer in such control is the regulation of transcription. This regulation is preformed by a network of interactions between transcription factor proteins (TFs) and the DNA of the genes they regulate. To understand the workings of this network, it is thus crucial to understand the most basic interaction between a TF and its target site on the DNA. Indeed, much effort has been devoted to detecting the TF-DNA binding location and specificities.

Experimentally, much of the binding specificity information has been determined using traditional methodologies such as footprinting, gel-shift analysis, Southwestern blotting, or reporter constructs. Recently, a number of high-throughput technologies for identifying TF binding specificities have been developed. These methods can be classified to two major classes, *in vitro* and *in vivo*

methods. In vitro methods can further be classified to methods that select high-affinity binding sequences for a protein of interest (Elnitski *et al.*[1]), and high-throughput methods that measure the affinities of specific proteins to multiple DNA sequences. Examples of the latter class of methods include protein binding microarrays [2] and microfluidic platforms [3], which claim to achieve better measurement of transient low affinity interactions. The in vivo methods are mainly based on microarray readout of either DNA adenine methyltransferase fusion proteins (DamID) or of chromatin immunoprecipitation DNA-bound proteins (ChIP-chip) [2]. However, despite these technological advances, distilling the TF binding specificity from these assays remains a great challenge, since in many cases the in vivo measured targets of a TF do not have common binding sites, and in other cases genes that have the known and experimentally determined site for a TF are not measured as its targets. For these reasons, the problem of identifying transcription factor binding sites (TFBSs) has also been the subject of much computational work [1].

The experimental and computational approaches above revealed TFBSs are short, typically 6-20 base pairs, and that some degree of variability in the TFBSs is allowed. For these reasons, the binding site specificities of TFs are described by a sequence *motif*, which should represent the set of multiple allowed TFBSs for a given TF. The most common representation for sequence motifs is the *position specific scoring matrix* (PSSM), which specifies a separate probability distribution over nucleotides at each position of the TFBS. The goal of computational approaches is then to identify the PSSM associated with each TF.

Despite its successes, the PSSM representation makes the very strong assumption that the binding specificities of TFs are position-independent. That is, the PSSM assumes that for any given TF and TFBS, the contribution of a nucleotide at one position of the site to the overall binding affinity of the TF to the site does not depend on the nucleotides that appear in other positions of the site. In theory, it is easy to see where this assumption fails. If instead of the PSSM representation, we allowed ourselves to assign probabilities to multiple nucleotides at multiple positions, then we could use the same number of parameters to specify the desired TF binding specificities. This observation lies at the heart of our approach (see Figure 1).

From the above discussion, it should be clear that the position-independent assumption of PSSMs is rather strong, and that relaxing this assumption may lead to a qualitatively better characterization of TF motifs. Indeed, recent studies revealed specific cases in which dependencies between positions may exist, [3]. In a more comprehensive study, Barash *et al.*[4] developed a Bayesian network approach to represent higher order dependencies between motif positions, and showed that these models predict putative TFBSs in ChIP-chip data with higher accuracy than PSSMs. However, the Bayesian network representation, due to its acyclicity constraints, imposes unnecessary restrictions on the motif structure, and its conditional probability distributions limit the number of dependencies that can be introduced between positions in practice, due to the exponential increase in the number of parameters introduced with each additional

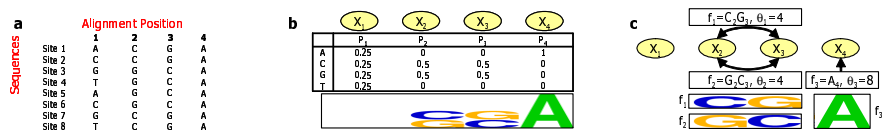


Fig. 1. Comparison between FMMs and PSSMs in a toy example of a TFBS with 4 positions. (a) Eight input TFBSs that the TF recognizes. (b) A PSSM for the input data in (a), showing its Markov network representation, probability distributions over each position, and sequence logo. Note that the PSSM assigns a high probability to CG and GC in positions 2 and 3 as expected by the input data, but it also undesirably assigns the same high probability to CC and GG in these positions. (c) An FMM for the input data in (a), showing the associated Markov network, with 3 features, and sequence logo. Note that features f_1 and f_2 assign a high probability to CG and GC in positions 2 and 3 but not to CC and GG in these positions, as desired.

dependency. While some of these issues may be addressed, e.g., using sparse conditional probability distribution representations, Bayesian networks are not the ideal and most intuitive tool for the task.

Here, we propose a novel approach to modeling TFBS motifs, termed *feature motif models* (FMMs). Our approach is based on describing the set of sequence properties, or *features*, that are relevant to the TF-DNA interactions. Intuitively, the binding affinity of a given site to the TF increases as it contains more of the features that are important for the TF in recognizing its target site. In our framework, features may be binary (e.g., “C at position 2, and G at position 3”) or multi-valued (e.g., “the number of G or C nucleotides at positions 1-4”), and global features are also allowed (e.g., “the sequence is palindromic”). Each feature is assigned a statistical weight, representing the degree of its importance to the TF-DNA interaction, and the overall strength of a TFBS can then be computed by summing the contribution of all of its constituent features. We argue that this formulation captures the essence of the TF-DNA interaction more explicitly than PSSMs and other previous approaches. It is easy to see that our FMMs contains in it the PSSM description, since a PSSM can be described within our framework using four single nucleotide features per position.

In what follows, we provide the mathematical formulation of FMMs, and devise an algorithm for learning FMMs from TFBSs data. This problem is quite difficult, as it reduces to learning structure in Markov networks, a paradigm that is still poorly developed. We evaluate our approach in a controlled synthetic data setting, and demonstrate that we can learn the correct features even from a relatively small number of positive examples. Finally, we apply our method to real TFBSs for yeast TFs [5,6], and show several cases where our method better explains the observed TFBS data and identifies motif sequence features that span multiple positions. We identify global properties that are common to the DNA sequence specificities of most TFs: TFBSs have strong dependencies between positions; these dependencies mostly occur in the center of the site; and dependencies typically exist between nearby positions in the site.

2 The Feature Motif Model

We now present our approach for representing TF binding specificities. Much like in the PSSM representation, our goal is to represent commonalities among the different TFBSs that a given TF can recognize, and assign a different strength to each potential site, corresponding to the affinity that the TF has for it. The key difference between our approach and a PSSM is that we want to represent more expressive types of motif commonalities compared to the PSSM representation, in which motif commonalities can only be represented separately for each position of the motif. Intuitively, we think of a TF-DNA interaction as one that can be described by a set of sequence *features*, such as pairs or triplets of nucleotides at key positions, that are important for the interaction to take place: the more important features a specific site has, the higher affinity it will have for the TF.

One way to achieve the above task is to represent a probability distribution over the set of all sequences of the length recognized by the given TF. That is, for a motif of length L , we represent a probability distribution over all 4^L possible L -mer sequences. Formally, we wish to represent a joint probability distribution $P(X_1, \dots, X_L)$, where X_i is a random variable with domain $\{A, C, G, T\}$ corresponding to the nucleotide at the i -th position of the sequence. However, rather than representing this distribution using the prohibitively large number of $4^L - 1$ independent parameters, our goal is to represent this joint distribution more compactly in a way that requires many fewer parameters but still captures the essence of TF-DNA interactions. The PSSM does exactly this, but it forces the form of the joint distribution to be decomposable by positions. Barash *et al.*[4] presented alternative representations to the PSSM, using Bayesian networks, that allow for dependencies to exist across the motif positions. However, as discussed above, the use of Bayesian networks imposes unnecessary restrictions and is not natural in this context.

A more natural approach that can easily capture our above desiderata is the framework of undirected graphical models, such as Markov networks or log-linear models, which have been used successfully in an increasingly large number of settings. As it is more intuitive for our setting, we focus our presentation on log-linear models. Let $\mathcal{X} = \{X_1, \dots, X_L\}$ be a set of discrete-valued random variables. A *log-linear model* is a compact representation of a probability distribution over assignments to \mathcal{X} . The model is defined in terms of a set of *feature functions* $f_k(\mathbf{X}_k)$, each of which is a function that defines a numerical value for each assignment \mathbf{x}_k to some subset $\mathbf{X}_k \subset \mathcal{X}$. Given a set of feature functions $F = \{f_k\}$, the parameters of the log-linear model are weights $\theta = \{\theta_k : f_k \in F\}$. The overall joint distribution is then defined as:

$$P(\mathbf{x}) = \frac{1}{Z} \exp \left(\sum_{f_k \in F} \theta_k f_k(\mathbf{x}_k) \right), \text{ where } Z = \sum_{\mathbf{x} \in \mathcal{X}} \exp \left(\sum_{f_k \in F} \theta_k f_k(\mathbf{x}_k) \right) \quad (1)$$

is the *partition function* that ensures that the distribution P is properly normalized (i.e., $\sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x}) = 1$), and \mathbf{x}_k is the assignment to \mathbf{X}_k in \mathbf{x} . Although

we chose the log-linear model representation, we note that it is in fact equivalent to the Markov network representation, and the mapping between the two is straightforward. We now demonstrate how we can use this log-linear model representation in our setting, to represent feature-based motifs. We start by showing how PSSMs can be represented within this framework.

Representing PSSMs. Recall that a PSSM defines independent probability distributions over each of the L positions of the motif. To represent PSSMs in our model, we define 4 features f_{iJ} for each position that indicate whether a specific nucleotide $J \in \{A, C, G, T\}$ exists at a specific position $1 \leq i \leq L$ of the TFBS. We associate each feature with a weight θ_{iJ} that is equal to its marginal log probability over all possible TFBSs. It is easy to show that putting this into Equation 1 defines the exact same probability distribution as of the PSSM, and that the *partition function* as defined in Equation 1 is equal to 1 in this case.

Representing Feature Motifs. Given a TF that recognizes TFBSs of length L , our feature-based model represents its motif using the log-linear model of Equation 1, where each feature f_k corresponds to a sequence property that may be defined over multiple positions. As an example for a feature, consider the indicator function: ‘C’ at position 2 and ‘G’ at position 3, as in Figure 1c. This feature illustrates our ability to define features over multiple positions. We note, that continuous and even global features (such as G/C content) can easily be defined within our model. We then associate each feature with a weight, θ_k , that defines its importance to the TF-DNA binding affinity. Given a sequence, we can now compute its probability using Equation 1, which boils down to summing the value of all the features present in the sequence, each multiplied by its respective weight parameter, and exponentiating and normalizing this resulting sum. Intuitively, this model corresponds to identifying which of the features that are important for the TF-DNA interaction are present in the sequence, and summing their contributions to obtain the overall affinity of the TF to the site. This intuitive model is precisely the one we set out to obtain.

3 Learning Feature Motif Models

In the previous section, we presented our feature-based model for representing motifs. Given a collection of features F , our method uses the log-linear model to integrate them, as in Equation 1. As we showed, the standard PSSM model can be represented in our framework. However, our motivation in defining the model was to allow for integration of other features, that may span multiple positions. A key question is how to select the set of features for a given model. In this section, we address this problem. Since log-linear models are equivalent to Markov networks, our problem essentially reduces to structure learning in Markov networks. This problem is quite difficult, since even the simpler problem of estimating the parameters of a fixed model does not have an analytical closed form solution. Thus, the solutions proposed for this problem have been various heuristic searches, which incrementally modify the model by adding and deleting features to it in some predefined scheme [7, 8].

We now present our algorithm for learning a feature-based model from TFBSs data. Our approach follows the Markov network structure learning method of Lee *et al.*[8]. It incrementally introduces (or selects) features using the *grafting* method of Perkins *et al.*[9]. We first present the simpler task of estimating the parameters of a given model, as this is a sub-problem that we need to solve when searching over the space of possible network structures.

3.1 Parameter Estimation

For the parameter estimation task, we assume that we are given as input a dataset $D = \{\mathbf{x}[1], \dots, \mathbf{x}[N]\}$ of N *aligned* i.i.d TFBSs, each of length L , and a model \mathcal{M} defined by a set of sequence features $F = \{f_1, \dots, f_k\}$. Our goal is to find the parameter vector $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_k\}$ that specifies a weight for each feature $f_i \in F$, and maximizes the log-likelihood function:

$$\log P(D | \boldsymbol{\theta}, \mathcal{M}) = \sum_{i=1}^N \log P(\mathbf{x}[i] | \boldsymbol{\theta}, \mathcal{M}) = \sum_{i=1}^N \sum_{f_k \in F} \theta_k f_k(\mathbf{x}[i]_k) - N \log Z \quad (2)$$

where $\mathbf{x}[i]_k$ corresponds to the nucleotides of the i -th TFBS at the positions relevant to feature k , and Z is the partition function as in Equation 1. It can easily be shown that the gradient of Equation 2 is:

$$\frac{\partial \log P(D | \boldsymbol{\theta}, \mathcal{M})}{\partial \theta_k} = \sum_{i=1}^N \sum_{f_k \in F} f_k(\mathbf{x}[i]_k) - N \frac{1}{Z} \frac{\partial Z}{\partial \theta_k} \quad (3)$$

Although no closed-form solution exists for finding the parameters that maximize Equation 2, the objective function is concave, and we can thus find the optimal parameter settings using numerical optimization procedures such as gradient ascent or conjugate gradient [10]. We now deal with optimizing Equation 2.

3.2 Optimization of the Objective Function

Applying numerical optimization procedures such as gradient ascent requires the computation of the objective function and the gradient with respect to any of the θ_k parameters. Although the fact that the objective function is concave, and that both the function and its gradient have simple closed forms may make the parameter estimation task look simple, in practice the computing them may be quite expensive. The reason is that the second terms of both the function and the gradient involve evaluating the partition function, which requires, in a naive implementation, summing over 4^L possible TFBSs sequences.

Since algorithms for learning Markov networks usually require computation of the partition function, this problem was intensively researched. Although in some cases the structure of the features may be such that we can decompose the computation to achieve efficient computation, in the general case it can be shown to be an NP-hard problem and hence requires approximation. Here we

suggest a novel strategy of optimizing the objective function. We first use the (known) observation that the gradient of Equation 2 can also be expressed in terms of features expectations. Specifically, since

$$\frac{1}{Z} \frac{\partial Z}{\partial \theta_k} = \frac{\sum_{\mathbf{x} \in \mathcal{X}} f_k(\mathbf{x}_k) \exp\left(\sum_{f_k \in F} \theta_k f_k(\mathbf{x}_k)\right)}{\sum_{\mathbf{x} \in \mathcal{X}} \exp\left(\sum_{f_k \in F} \theta_k f_k(\mathbf{x}_k)\right)} = E_{P \sim \theta}(f_k(\mathbf{x}_k)), \quad (4)$$

we can rewrite Equation 3 as:

$$\frac{\partial \log P(D | \boldsymbol{\theta}, \mathcal{M})}{\partial \theta_k} = \sum_{i=1}^N \sum_{f_k \in F} f_k(\mathbf{x}[i]_k) - N E_{P \sim \theta}(f_k(\mathbf{x}_k)). \quad (5)$$

We further observed that since Equation 2 is a concave function, its absolute directional derivative along any given line in its domain is also a concave function. We used this observation to use the conjugate gradient function optimization algorithm [10] in a slightly modified version: Although the gradient that was given to the algorithm was indeed as in Equation 5, the function value along every line search step of the algorithm was the absolute directional derivative along this line. For example, at the line search step along direction \mathbf{y} our function $F^*(\boldsymbol{\theta}, \mathbf{y})$ value is: $F^*(\boldsymbol{\theta}, \mathbf{y}) = \langle \nabla \log P(D | \boldsymbol{\theta}, \mathcal{M}), \mathbf{y} \rangle$

Following the above strategy allows us to optimize Equation 2 without computing its actual value. Specifically, it means that we can optimize our objective without computing the partition function. Instead, the problem reduces to evaluating feature expectations, a special case of inference in Markov networks, that can be exactly computed using algorithms such as *loopy belief propagation* [11]. The ability of these algorithms to give an exact result depends on the underlying network structure. As the network structure becomes more complex, the algorithms need to use approximations. Since this family of algorithms can also approximate the partition function, our method will be similar to methods that evaluate the partition function when the network structure allows for exact inference. However, as the error bounds for approximate inference are better characterized than the error bounds of partition function estimations, it is possible that our approach may work better under conditions that require approximation.

3.3 Learning the Features

In Section 3.1, we developed our approach for estimating the feature parameters for a fixed model in which the feature set F is defined. We now turn to the more complex problem of automatically learning the set of features from aligned TFBSs data. This problem is an instance of the more general problem of learning the structure of Markov networks from data. However, quite surprisingly, although Markov networks are used in a wide variety of applications, there are very few effective algorithms for learning Markov network structure from data.

In this paper we followed the Markov network structure learning approach suggested by Lee *et al.*[8]. This approach extends the learning approach of

Perkins *et al.*[9] to learning the structure of Markov network using the L_1 -Regularization over the model parameters. To incorporate the L_1 -Regularization into our model we need to introduce a *Laplacian* parameter prior over each feature, leading to the modified objective function:

$$\log P(D, \boldsymbol{\theta} \mid \mathcal{M}) = \log P(D \mid \boldsymbol{\theta}, \mathcal{M}) + \log P(\boldsymbol{\theta} \mid \mathcal{M}) \quad (6)$$

where $P(\boldsymbol{\theta} \mid \mathcal{M}) = \left(\frac{\alpha}{2}\right)^{|F|} \exp\left(-\sum_{f_k \in F} \alpha |\theta_k|\right)$ and $\log P(D \mid \boldsymbol{\theta}, \mathcal{M})$ is the data likelihood function as in Equation 2. Taking the log of this parameter prior and eliminating constant terms, we arrive at the final form of our objective function:

$$\log P(D, \boldsymbol{\theta} \mid \mathcal{M}) = \sum_{i=1}^N \sum_{f_k \in F} \theta_k f_k(\mathbf{x}[i]_k) - N \log Z - \alpha \sum_{f_k \in F} |\theta_k| \quad (7)$$

It is easy to see that this modified objective function is also concave in the feature parameters $\boldsymbol{\theta}$ and we can thus optimize it using the same conjugate gradient procedure described in Section 3.1. We then follow the *grafting* approach of adding features in a stepwise manner. In each step, the algorithm first optimizes the objective function relative to the current set of active features F , and then adds the inactive feature $f_i \in \neg F$ with the maximal gradient at $\theta_i = 0$. Using an L_1 -Regularized concave function provides a stopping criteria to the algorithm that leads to the global optimum [9]. The L_1 -Regularization has yet another desirable quality for our purpose, as it has a preference for learning sparse models with a limited number of features [8]. It has long been known to have a tendency towards learning sparse models, in which many of the parameters have weight zero [12] and theoretical results show that it is useful in selecting the features that are most relevant to the learning task [13]. Since the *grafting* feature addition method is a heuristic, it seems reasonable that features that were added at an early stage may become irrelevant at later stages, and hence get a zero weight. We thus introduce an important difference from the method of Lee *et al.*, by allowing the removal of features that become irrelevant.

4 Experimental Results

We now present an experimental evaluation of our approach. We first use synthetic data to test whether our method can reconstruct sequence features that span multiple positions when these are present, and then compare the ability of our approach to learn binding specificities of yeast TFs to that of PSSMs.

4.1 Synthetic Data

To evaluate our models in a controlled setting, we manually created three FMMs (Figure 2) of varying weights and features, and learned both PSSM and FMMs from TFBSs that we sampled from them. We evaluated the learned models by computing the log-likelihood that the learned models assign to a test set of 10,000

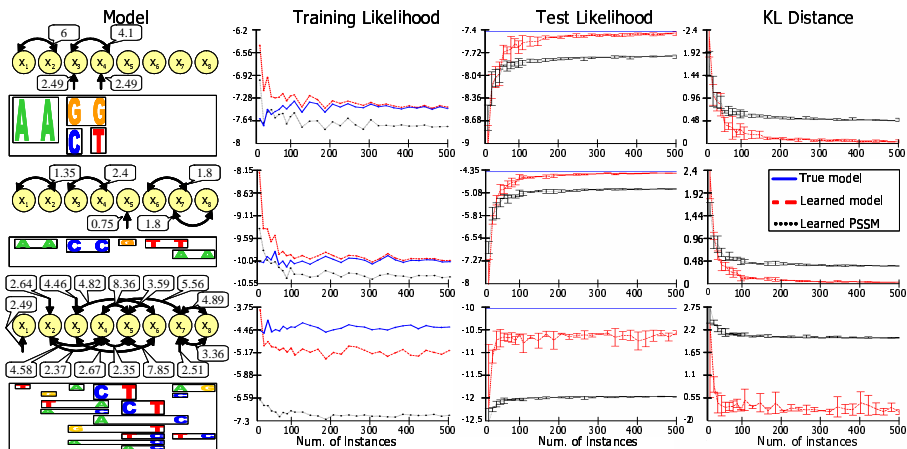


Fig. 2. Evaluation of our approach on synthetic data. Results are shown for three manually constructed model, from which we drew samples and constructed FMMS and PSSMs. For each model, shown is its Markov network and sequence logo (left), training and test log-likelihood (average per instance for the true model, and learned FMM and PSSM) and KL-distance of the learned FMM and PSSM models from the true model.

unseen TFBSs sampled from the true model, and by computing the Kullback Leibler (KL) distance between distributions of the true and learned models.

We evaluated two specific aspects of our approach: the minimum number of samples needed for learning FMMS, and the dependency of the learning on the prior weighting parameter, α . In all experiments, we limited the FMM to structures that allow exact inference using belief propagation algorithm [11]. While this poses constraints on the underlying network, learning more complex models also gave good performance, since the most important feature were still learned. We repeated each experiment setting 3 times.

We first tested the effect of the prior weight parameter α on the quality of the learning reconstruction. To this end, we varied α in the range of 10^{-6} to 100, while using a fixed number of 500 input sequences. The results showed that in the range tested, the best reconstruction performance was achieved for $\alpha = 0.1$. While smaller values tend to allow over fitting, higher values pose harsh constraints on the leaned model.

Second, we estimated the minimum number of samples needed for learning FMMS, by sampling different training set sizes in the range of 10-500. As can be seen in Figure 2, for all three cases, our model reconstructs the true model with high accuracy even with a modest number of 50 input TFBSs, reconstructs the true model nearly perfectly with 100 or more samples. As expected, since the true model includes dependencies between positions, our model significantly outperforms the PSSM in these cases even when only 20 input sites were used. In these experiments, we fixed the prior weight parameter to 0.1. Examining the learned features, we found that for a sample size of 20 or more, only features that

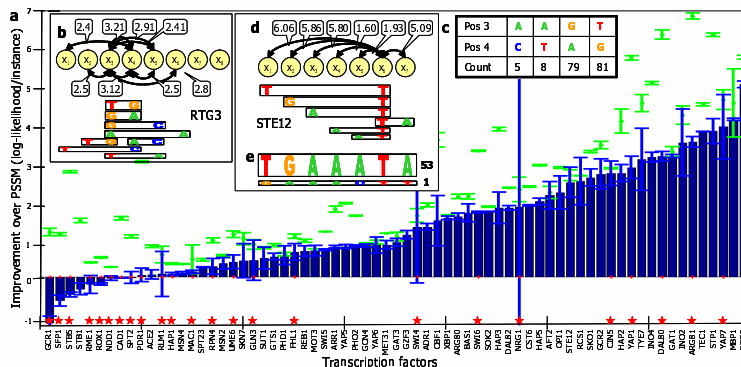


Fig. 3. Evaluating our approach on real TFBSs from yeast. (a) Train (green points) and test log-likelihood (blue bars), shown as the mean and standard deviation improvements in the average log-likelihood per instance compared to a PSSM. Each model was learned from the TFBSs reported by MacIsaac06 et al. in a 5-fold cross validation scheme. Models that were constrained to allow exact inference are marked with a red star. (b) Markov network representation of the dominant features of the FMM model learned for RTG3. (c) Sequence counts for positions 3 and 4 of the input TFBSs of RTG3. (d,e) Same as (b,c), for strong feature relations learned for the STE12 TF.

appeared in the true model were learned with significant weight. Our results thus show that we can successfully learn FMMs, even in a realistic setting in which only a limited number of input TFBSs is available.

4.2 Identifying Binding Features of Yeast TFs

Having validated our approach on synthetic data, we next applied it to TFBSs data for yeast TFs. Our goal is to identify whether FMMs can better describe the sequence specificities of yeast TFs. As input to our method, we used the high-quality TFBSs data reported by MacIsaac *et al.*[6]. This dataset consists of 16371 regulatory TF-binding site interactions, where each interaction reported is one in which the TF is bound to the promoter region containing the TFBS as determined by the ChIP-chip assays of Harbison *et al.*[5], and the TFBS has a good match to the PSSM reported for the corresponding TF. While this dataset is quite comprehensive, it is in fact a very stringent test for our method, since each reported TFBS is required to have a relatively good match to the PSSM, a property that we do not necessarily expect from sequences that are well explained by our feature motif models.

We used a five fold cross validation scheme to test whether FMMs can better explain yeast TFBSs. We took each of the 69 TFs and learned a model from the training set. Models of length greater than 8 were constrained to allow exact inference as in Section 4.1. As a measure of success, we computed for each motif, the average and standard deviation of the five test sets average log likelihood.

Using this criterion, we compared the results of applying our model to that of applying the PSSM model to the same input data. The results are shown in Figure 3(a). As can be seen, FMMs better explained the TFBSs data of 60 of the 69 TFs (86%). In 34 of the 69 (49%) cases, the probability that our model gave to each TFBS in the test data was, on average, more than twice the probability assigned by the PSSM. We note that although the results of the constrained model were slightly weaker (66%, and 33% respectively) they are still relatively good. Taken together, these results demonstrate that TFBSs data can be better characterized by feature motif models compared to PSSMs, and that the position independent assumption of the PSSM model does not hold in many cases and can thus poorly represent the binding affinities of many TFs.

We next turned to examine the actual features that we identified and test their biological significance. To this end, we first examined the models learned for each of the 69 TFs, by extracting the dominant features learned and observing the counts of these features in the original input TFBS data. Two examples of such a model examination are shown in Figure 3(b-e). The leucine zipper TF RTG3, an activator of the TOR growth pathway, represents one case in which our model provides insight into its binding specificity, and in which we can clearly understand why the PSSM model fails. For this TF, our model assigns a probability that, on average per test-set TFBS, is more than 20 times greater than the corresponding probability assigned by the PSSM. Examining the dominant features of the model reveals that the two most dominant features were defined over positions 3 and 4. Each one of these features gives high weight to either “GA” or “TG” at these positions. Strikingly, the counts of these two features in the original input data were 79 and 81 (out of 173 BSs), respectively. Clearly, the PSSM model completely misses this. These results suggest that RTG3 may have two distinct types of TFBSs, one with a “TG” in positions 3 and 4 and another with “GA” in these positions. This hypothesis is consistent with a study by Rothermel *et al.*[14] showing that RTG3 contains at least two independent activation domains, which may interact with different co-factors, leading to two different binding modes.

The STE12 transcription factor, an activator of the mating or pseudohyphal growth pathways, is another intriguing example where our model provides insight into the specificity of the corresponding TF. Of all the 994 TFBSs of STE12 in the input data, 54 have a ‘T’ in position 6. Of these, 53 have the exact full TFBS of ‘TGAAATA’. In other words, if a ‘T’ appears in position 6 of the TFBS, it fully determines the remaining basepairs of the site. As can be seen in Figure 3(d,e), our model captures this property, by learning six features with high weights that each contained a ‘T’ in position 6, and one of the other positions as the second position. This result is consistent with reports in the literature that the specificity of STE12 can change, depending on its interaction with other regulators [15]. This TFBS is also an example where a simple Bayesian network representation of the site would not be able to compactly represent the site, since position 6 would have to be a parent of each of the other positions, thereby placing constraints (due to acyclicity) on the types of features that could be learned between the

positions when ‘T’ is not present in position 6, and in any case requiring many parameters for the representation. A mixture model, which is one of the options presented Barash *et al.*[4] would work here, but learning it from the data might be challenging.

To further and globally characterize the biological significance of the feature motif models learned, we took the dominant features of each of the 69 models learned, and partitioned the TFBSs into two sets, based on the presence of each of the features. By mapping the sites back to the promoters in which they were identified, we could partition the genes regulated by the each TF into genes that have TFBS of the TF and have the examined feature, and genes that don’t have such TFBS. We used the hypergeometric distribution to compute a p -value for an enrichment of the partition to various features. In all enrichment tests we took $p < 0.01$ to be significant, corrected the results by FDR, and presented the best enrichment for each TF. We first tested for enrichment in functional categories from the Gene Ontology (GO) database. The results are shown in Figure 4(a). These results suggest that particular features of the TFBS of each TF may be important for its ability to regulate one specific class of genes. Second, we ran the same enrichment tests using a database of 346 protein-DNA interactions that we compiled from 10 different CHIP-chip studies. The top enrichment in this case, shown in Figure 4(b), suggest hypotheses on the cooperation between other proteins and specific types of the TFBS of the TF as characterized by the enriched feature. Since the data include protein-DNA interactions measured in various conditions [5], some enrichments represent TFBSs that are bound by the corresponding TF only in some conditions.

Finally, we used our resulting models to gain insights into the global properties of binding specificities of all yeast transcription factors. To this end, we collected all the dominant features that we learned across all 8 length models, and computed the average weight of features that were learned between each pair of positions of the TFBS. The comparison of this average weight for each combination of positions is shown in Figure 4(c). Intriguingly, although this average represents many different TFs, two prominent signals emerge. First, the strongest dependencies between features exist between features positioned in the center of the site. Second, nearby positions tend to have a higher dependency compared to dependencies that exist between distant positions. From these results, we compiled a general ‘consensus’ model for representing the dependencies between positions in the TFBSs of the yeast transcription factors, shown in Figure 4(d). Thus, our model provides insights into global properties that are characteristics of TFBS specificities across all yeast TFs.

4.3 Application of FMM to Motif Finder

As a natural extension of our FMM approach, we integrated our FMM model into a basic motif finder application. Our motif sampler takes as input a set of positive sequences, and a set of negative sequences. The algorithm searches for a motif of length L that maximizes the sum of the log-probabilities of the best TFBSs for each positive input sequence. The algorithm works in an iterative manner. It first

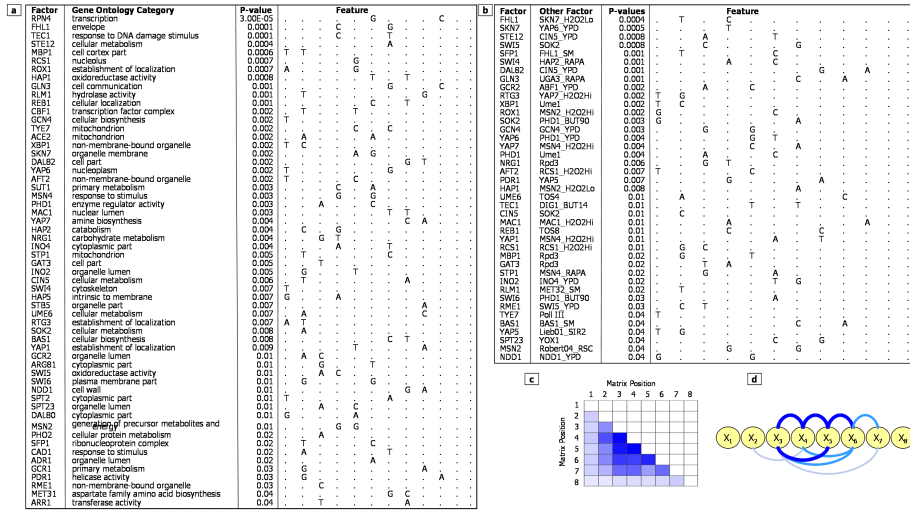


Fig. 4. Biological significance of FMMS. (a) TFBSs of yeast TFs with particular features are enriched for specific GO functional categories. (b) Same as (a) for enrichment in protein-DNA interactions that we compiled from 10 different studies. (c) Average weight of features that span 2 positions, across FMM models learned for all yeast TFs with $L = 8$ (d) ‘Consensus’ properties of correlations between positions in the sequence specificities of TFs, compiled based on (c).

searches for a sequence of length L that maximizes the ratio between fraction of positive sequences and negative sequences that contain it up to one mismatch. It then initializes a model from these L -length sequences that appear in the positive set. Following this initialization, we then use the Expectation Maximization (EM) algorithm to optimize the model. In the ‘E’ step the motif finder selects the maximum likelihood TFBS from each positive sequence, while in the ‘M’ step it learns a new model from these selected sequences. The algorithm stops after convergence is reached or after a maximal number of ‘EM’ steps. After finding a motif, the algorithm removes from each sequence the TFBS with the highest likelihood, and then searches for a new motif.

Although our motif finder does not yet integrate all the state of the art methodologies for motif finding, we use it to provide an example for the potential of using FMMS instead of PSSMs for the motif finding task. Specifically, we took the 177 sets of at least 25 sequences each, that bind a transcription factor under a specific condition according to the data of Harbison *et al.*[5] as positive sets, and the rest of the sequences as negative sets. We used a 5-fold cross validation scheme to evaluate the motif finder using either FMM or PSSM as the motif model. In these runs we used half of the background as input and half for evaluation. We evaluated the performance of the results by evaluating the sum of log-probabilities of the best TFBS for both the positive sequence test set, and for the held out set of background sequences, and compared the difference

of the two. For this evaluation, we used the best of motifs number 2 to 5 that the motif finder outputs. As the results presented in Figure 5 show, even with this relatively basic motif finder, in 133 of the 177 (75%) positive sets tested, we found motifs that gave better average likelihood on the held out positive test set compared to the PSSMs that were learned. Since we used the same framework for learning both the FMM and PSSM models, these results show the potential of our FMM models for the motif finding task, and suggest that combining them within advanced motif finding schemes may yield improved results.

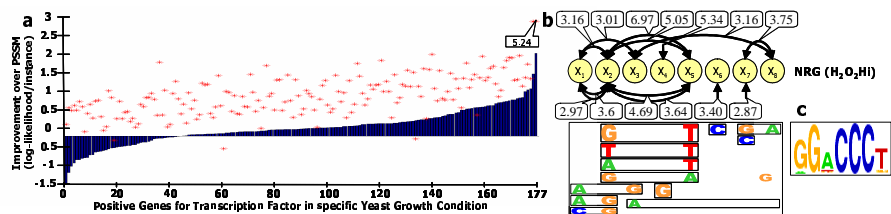


Fig. 5. Motif finder results. (a) The difference between the test average log-likelihood and the background average log-likelihood for the best FMM model (stars), and the difference between this value and the similar value using the best PSSM model (bars).

5 Conclusions

In this paper we presented *feature motif models* (FMMs), a novel probabilistic method for modeling the binding specificities of transcription factors. We presented the mathematical foundations of FMMs and showed their advantage over PSSMs in learning motifs from both synthetic and real data. We demonstrated the benefits of using undirected graphical models (Markov networks) for representing important features of TF binding specificities, and suggested a methodology to learn such features from both aligned and unaligned input sequences. We also suggested a methodology for optimizing the objective function, that may give better performance under settings that require approximation.

There are several directions for refining and extending our approach. First, expanding the network structure in which we perform exact inference, and improving our approximate inference abilities, will greatly increase the power of our models. Second, integrating our model into a state of the art (rather than basic) motif finder algorithms may allow us to improve upon existing approaches to the task. Finally, using our models as an improved basic building block, we can integrate it into higher level regulatory models (e.g., [16]) and obtain a much better quantitative understanding of the underlying principles of transcriptional regulatory networks.

Acknowledgments We thank Tali Sadka for useful discussions. This research was supported by the Israel Science Foundation (Grant No. 645/06), and by

ENFIN, a Network of Excellence funded by the European Commission within its FP6 Programme, contract number LSHG-CT-2005-518254.

References

1. Laura Elnitski, Victor X. Jin, P.J.F., Jones, S.J.: Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res* **16**(12) (2006) 1455–64
2. Bulyk, M.L.: Dna microarray technologies for measuring protein-dna interactions. *Current Opinion in Biotechnology* **17** (2006) 1–9
3. Maerkl, S.J., Quake, S.R.: A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**(5809) (2007) 233–236
4. Barash, Y., Elidan, G., Friedman, N., Kaplan, T.: Modeling dependencies in protein-dna binding sites. *RECOMB* (2003)
5. Harbison *et al.*: Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004) (2004) 99–104
6. MacIsaac, K., Wang, T., Gordon, D., Gifford, D., Stormo, G., Fraenkel, E.: An improved map of conserved regulatory sites for *saccharomyces cerevisiae*. *BMC Bioinformatics* **7** (2006) 113
7. Della Pietra, S., Della Pietra, V.J., Lafferty, J.D.: Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(4) (1997) 380–393
8. Lee, S.I., Ganapathi, V., Koller, D.: Efficient structure learning of Markov networks using L1-regularization. *NIPS* (2007)
9. Perkins, S., Lacker, K., Theiler, J.: Grafting: fast, incremental feature selection by gradient descent in function space. *J. Mach. Learn. Res.* **3** (2003) 1333–1356
10. Minka, T.P.: Algorithms for maximum-likelihood logistic regression. Technical Report 758, Carnegie Mellon University (2001)
11. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Generalized belief propagation. In: *NIPS*. (2000) 689–695
12. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58**(1) (1996) 267–288
13. Ng, A.: Feature selection, l1 vs. l2 regularization, and rotational invariance. In: *Twenty-first International Conference on Machine Learning (ICML)*. (2004)
14. Rothermel, B., Thornton, J., Butow, R.: Rtg3, a basic helix-loop-helix/leucine zipper protein that functions in mitochondrial-induced changes in gene expression, contains independent activation domains. *JBC* **272** (1997) 19801–7
15. Zeitlinger, J., Simon, I., Harbison, C., Hannett, N., Volkert, T., Fink, G., Young, R.: Program-specific distribution of a transcription factor dependent on partner transcription factor and mapk signaling. *Cell* **113**(3) (2003) 395–404
16. Segal, E., Yelensky, R., Koller, D.: Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19**(Suppl 1) (2003) S273–82