Learning Protein-DNA Interaction Landscapes by Integrating Experimental Data through Computational Models

Jianling Zhong[1], Todd Wasson[2], and Alexander J. Hartemink[1,3,⋆]

[1] Computational Biology & Bioinformatics, Duke University, Durham, NC 27708
[2] Lawrence Livermore National Laboratory, Livermore, CA 94550
[3] Department of Computer Science, Duke University, Durham, NC 27708
{zhong,amink}@cs.duke.edu

Abstract. Transcriptional regulation is directly enacted by the interactions between DNA and many proteins, including transcription factors, nucleosomes, and polymerases. A critical step in deciphering transcriptional regulation is to infer, and eventually predict, the precise locations of these interactions, along with their strength and frequency. While recent datasets yield great insight into these interactions, individual data sources often provide only noisy information regarding one specific aspect of the complete interaction landscape. For example, chromatin immunoprecipitation (ChIP) reveals the precise binding positions of a protein, but only for one protein at a time. In contrast, nucleases like MNase and DNase reveal binding positions for many different proteins at once, but cannot easily determine the identities of those proteins. Here, we develop a novel statistical framework that integrates different sources of experimental information within a thermodynamic model of competitive binding to jointly learn a holistic view of the *in vivo* protein-DNA interaction landscape. We show that our framework learns an interaction landscape with increased accuracy, explaining multiple sets of data in accordance with thermodynamic principles of competitive DNA binding. The resulting model of genomic occupancy provides a precise, mechanistic vantage point from which to explore the role of protein-DNA interactions in transcriptional regulation.

Keywords: protein-DNA interaction landscape, thermodynamic modeling, genomic data integration, competitive binding, COMPETE.

1 Introduction

As an essential component of transcriptional regulation, the interaction between DNA-binding factors (DBFs) and DNA has been studied extensively by experimentalists. To map genome-wide protein-DNA interactions, two basic categories of experimental techniques have been developed: chromatin immunoprecipitation

⋆ Corresponding author.

R. Sharan (Ed.): RECOMB 2014, LNBI 8394, pp. 433–447, 2014.

(ChIP) based methods (numerous studies in many organisms, but a few examples for yeast are [9, 18, 19]); and nuclease digestion based methods that profile chromatin with either DNase [11] or MNase [10]. To reveal high-resolution DNA interaction sites for a single antibody-targeted factor, ChIP methods can be used, especially the recently developed ChIP-exo methods [19] that use lambda exonuclease to obtain precise positions of protein binding. Nuclease digestion methods can be used to efficiently assay genome-wide DNA occupancy of all proteins at once, without explicit information about protein identities. These and other experimental efforts over the past decade have generated a large amount of data regarding the chromatin landscape and its role in transcriptional regulation. We now need computational models that can effectively integrate all this data to generate deeper insights into transcriptional regulation.

A popular set of computational models use this data to search for overrepresented DNA sequences bound by certain DBFs; these are often applied in the setting of motif discovery [4, 9, 15, 23]. More recently, models have been applied to DNase-seq data to identify 'digital footprints' of DBFs [3, 11, 14, 16]. However, many of these approaches share certain drawbacks. First, protein binding is typically treated as a binary event amenable to classification: either a protein binds at a particular site on the DNA sequence or it does not. However, both empirical and theoretical work has demonstrated that proteins bind DNA with continuous occupancy levels (as reviewed by Biggin [1]). Second, most computational methods model the binding events for one protein at a time instead of taking into consideration the interactions among different DBFs, especially nucleosomes. Although the work of Segal et al. [22], Kaplan et al. [12], and Teif and Rippe [24] are notable exceptions, these all consider small genomic regions and include only a few transcription factors (TFs); Segal et al. [22] ignored the role of nucleosomes altogether. Third, and most importantly, almost all current methods fail to integrate different kinds of datasets. This is insufficient because data from one kind of experiment only reveals partial information about the *in vivo* protein-DNA interaction landscape. For example, ChIP datasets only contain binding information for one specific protein under a specific condition; nuclease digestion datasets provide binding information for all proteins, but do not reveal the identities of the proteins; and protein binding microarray (PBM) experiments only look at sequence specificity of one isolated protein in an *in vitro* environment.

We previously published a computational model of protein-DNA interactions, termed COMPETE [25], that overcomes the first two drawbacks above by representing the competitive binding of proteins to DNA within a thermodynamic ensemble. Interactions between proteins and DNA are treated as probabilistic events, whose (continuous) probabilities are calculated from a Boltzmann distribution. COMPETE can easily include a large number of DBFs, including nucleosomes, and can efficiently profile entire genomes with single base-pair resolution. However, a major limitation of COMPETE is that it is a purely theoretical model of binding, based on thermodynamic first principles but not guided by data regarding *in vivo* binding events. Indeed, it is possible for COMPETE to

predict superfluous binding events that are inconsistent with observed data (see Supplemental Figure S1). It is therefore necessary to develop a new computational framework for jointly interpreting experimentally-derived data regarding genomic occupancy within a model built upon the thermodynamic foundation of COMPETE.

Here, we develop just such a method: a general framework for combining both a thermodynamic model for protein-DNA interactions (along the lines of COMPETE) and a new statistical model for learning from experimental observations regarding those interactions. Information from different experimental observations can be integrated to infer the actual thermodynamic interactions between DBFs and a genome. In this particular study, we demonstrate the use of this framework by integrating paired-end micrococcal nuclease sequencing (MNase-seq) data, which reveals information about the binding occupancy of both nucleosomes and smaller (subnucleosomal) factors. Our framework also integrates protein binding specificity information from PBM data and produces a more accurate and realistic protein-DNA interaction landscape than COMPETE alone, along with a mechanistic explanation of MNase-digested fragments of different sizes. The cross-validated performance of our framework is significantly higher than several baselines to which we compared it. Our framework is flexible and can easily incorporate other data sources as well, and thus represents a general modeling framework for integrating multiple sources of information to produce a more precise view of the interaction landscape undergirding transcriptional regulation.

2 Methods

2.1 Modeling Protein-DNA Interaction

We model the binding of DBFs (e.g., transcription factors and nucleosomes) to DNA along a probabilistic continuum, and we incorporate explicit competition between different DBFs. The ensemble average of the probability with which a particular DBF binds a specific position of the sequence can be derived from thermodynamic principles. To calculate this average probability, consider a specific binding configuration i from the ensemble, where i can be viewed as an instantaneous snapshot of the dynamic competition between DBFs for binding sites along the genome. Following the Boltzmann distribution, the unnormalized probability w_i of configuration i can be shown to be $w_i = \prod_{t=1}^{N_i} X_t \times P(S_t, E_t|DBF_t)$, where t is an index over the N_i DBF binding sites in configuration i. To simplify notation, we have treated each unbound nucleotide as being bound by a special kind of 'empty' DBF. In the above expression, X_t denotes a weight associated with DBF t, while S_t and E_t denote the start and end position of the DBF binding site, respectively. $P(S_t, E_t|DBF_t)$ is the probability of observing the DNA sequence between S_t and E_t, given that DBF t is bound there. If we use p_i to denote the probability of configuration i after normalization by the partition function, we can write the probability that DBF t binds at a specific

position j as $\sum_{i \in I(t,j)} p_i$, where $I(t, j)$ is the subset of binding configurations in the ensemble that have DBF t bound at sequence position j.

This model can be formulated analogously to a hidden Markov model (HMM) [17], in which the states correspond to the binding of different DBFs and the observations are the DNA sequence. The various probabilities, along with the partition function, can then be calculated efficiently using the forward-backward algorithm. For transcription factors, we have chosen to represent $P(S_t, E_t|DBF_t)$ using a position weight matrix (PWM), but more sophisticated models can also be used (e.g., relaxing positional independence, or based on energies rather than probabilities [26]). Regardless, binding models from different sources and of different forms can be easily incorporated into our model, generating the appropriate states and sequence emission probabilities. We use the curated PWMs from Gordân et al. [7], derived from *in vitro* PBM experiments, as the input protein binding specificities and consider them fixed (though our framework also could allow them to be updated).

The analogues of HMM transition probabilities in our model are the DBF weights, but these are not constrained to be probabilities. To allow this flexibility, we adopt a more general statistical framework called a Boltzmann chain [21] which can be understood as a HMM that allows the use of any positive real numbers for these weights. Because of the analogy with an HMM, we henceforth refer to these DBF weights as 'transition weights' and denote them collectively as a vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_D)$, where D is the number of different kinds of DBFs. We treat the D elements of $\boldsymbol{X}$ as free parameters, and we will fit them using experimentally-derived genomic data.

We should note that the DBF transition weights in a Boltzmann chain are sometimes called 'concentrations'. However, it is important to point out that these transition weights are not the same as bulk cellular protein concentrations, of the kind that can sometimes be measured experimentally [5]. Bulk cellular protein concentrations are not necessarily indicative of the availability of a DBF to bind DNA, because they do not account for phenomena like subcellular localization or extra-nuclear sequestration, protein activation through post-translational modification or ligand or co-factor binding, or the number of DBFs already bound to DNA. In contrast, our transition weights correspond to nuclear concentrations of active proteins that are free and available to bind DNA. In this sense, our weight parameters are more reasonably interpreted not as cellular concentrations but rather as the chemical potentials of the DBFs for interacting with the genome.

2.2 Using Paired-End MNase-Seq Data as a Measure of Genomic Occupancy Level of DNA-Binding Proteins

We used paired-end MNase-seq data from Henikoff et al. [10]. Based on their protocol, the length of the sequencing fragments correspond roughly to the size of the protein protecting that part of the DNA; the number of fragments mapping to the location correlates with the binding strength or occupancy. Therefore, to measure the level of occupancy of different DNA binding proteins, we separate

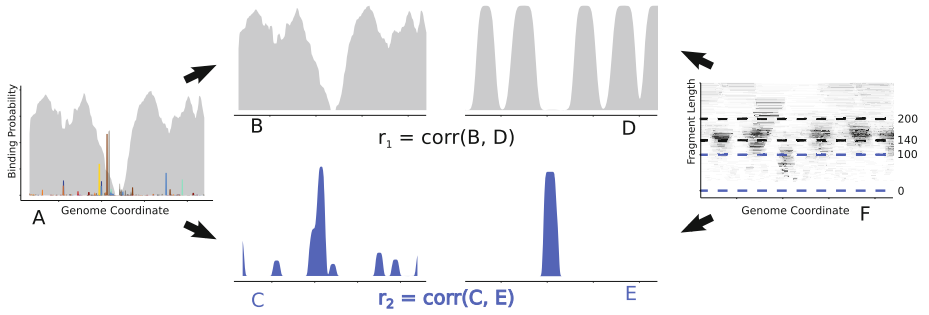

Fig. 1. Overview of objective function evaluation. (A) Predicted probability that each particular DBF binds at a given genome position, as calculated by COMPETE, given current DBF weights. We then separate these probabilities into two profiles: (B) predicted nucleosome binding profile and (C) predicted composite TF binding profile in which protein identities have been removed; the latter is smoothed to make it comparable to a short fragment coverage profile. Similarly, we separate the observed MNase-seq fragments (F) into long (140–200bp) and short (0–100bp) fragments, which are summed to produce measures of coverage. (D) Total long fragment coverage is processed into a large protein binding profile, which is compared to predicted nucleosomal binding, arriving at Pearson correlation r_1. (E) Total short fragment coverage is processed into a small protein binding profile, which is compared to predicted composite TF binding, arriving at Pearson correlation r_2. For this promoter, the quantity h that appears in our objective function (the pseudo-likelihood) is simply the geometric mean of the two correlations, after they are rescaled to lie in the interval $[0, 1]$: $h = \frac{1}{2}\sqrt{(1 + r_1) \times (1 + r_2)}$. The complete pseudo-likelihood over all promoters is then optimized with respect to the DBF weights using the inference method described below.

the fragments into long (140–200bp) and short (0–100bp) fragment groups and count the number of fragments in each group that cover a specific genomic location (called long and short fragment coverage, respectively). The long fragment coverage is used as a measure of the occupancy of large protein complexes, which are mainly nucleosomes, while the short fragment coverage is used as a measure of the occupancy of smaller proteins, which are mainly transcription factors.

To reduce noise in the MNase-seq data, we process the noisy fragment data into binding profiles through thresholding and smoothing. We define two thresholds: a bottom threshold T_b and a top threshold T_t. Coverage values that are below T_b are converted to 0, while those above T_t are converted to 1; coverage values between the two thresholds are normalized linearly to $[0, 1]$. We then smooth the track using a Gaussian kernel of bandwidth B_m. We process long and short fragment coverage data separately to get the large and small protein binding profiles, respectively (Figure 1 D and E). We choose $T_b = 200$ and $T_t = 500$, with $B_m = 10$ for short fragment coverage and $B_m = 30$ for long fragment coverage. These values give satisfying results in terms of reducing noise while retaining clear peaks. Because the choices are a little arbitrary, we performed a sensitivity analysis and observed that our results are largely unaffected across

a broad range of these parameters (see Supplemental Figure S2). We also note that MNase is known to prefer to cut A/T compared to G/C. We assessed the severity of this well-known bias and observed that it does not affect our final results (see Supplemental Figure S3). This is primarily because we are not using profiles of the total number of cuts at each genomic position, but rather using the full fragments (available as a result of paired-end sequencing) to generate profiles of fragment coverage; while the former would be highly sensitive to MNase bias, the latter is relatively insensitive to the small fluctuations in fragment end locations introduced by MNase bias.

2.3 Selecting a Subset of TFs and Promoter Regions

Our framework has the capability to include all *S. cerevisiae* transcription factors. However, our choice of transcription factors is limited by available high quality binding preference data. In addition, adding more TFs increases the dimensionality of the parameter space and therefore the computation time required to explore the space. In this study, we chose a set of 42 TFs with available high quality binding preference data. These TFs cover a wide range of cellular functions, including the widely-studied transcriptional regulators Reb1, Rap1, and Abf1 (possessing some chromatin remodeling activity), TFs involved in pheromone response (Ste12 and Tec1), TFs involved in stress response (like Msn4), and TFs involved in cell cycle regulation (Fkh1, Mbp1, and so forth). We also included some TFs, like Pho2 and Phd1, that regulate a large number of genes according to MacIsaac et al. [15]. While these 42 do not represent all yeast TFs, they are collectively responsible for 66% of the genome-wide protein-DNA interactions reported by MacIsaac et al. [15] (at p-value < 0.005 and conservation level 3).

Having selected our 42 TFs, we next chose a set of promoter regions that, according to MacIsaac et al. [15] (at p-value < 0.005 and conservation level 3), seem to be bound exclusively by those TFs. For this study, we focus on 81 such promoter regions, and extracted MNase-seq data for these loci as follows. If the promoter is divergently transcribed, we extracted the MNase-seq data between the two TATA elements, plus 200bp downstream of each TATA element. For the other (non-divergent) promoters, we extracted MNase-seq data 500bp upstream of the TATA element (or 100bp upstream of the end of the upstream gene, whichever is smaller), and 200bp downstream of the TATA element. Locations of TATA elements were taken from Rhee and Pugh [20].

2.4 Incorporating MNase-seq Data through an Objective Function

We model MNase-seq data through a pseudo-likelihood function, conditioned on COMPETE outputs. To calculate the pseudo-likelihood function, we process the COMPETE output TF binding probabilities as following: the binding probability of each COMPETE output TF binding event is expanded to a flanking region of C_e bp, and is then dropped linearly to 0 for another C_r bp; we then sum the

expanded binding probability of all TFs (truncating values larger than 1) and smooth the track using a Gaussian kernel of bandwidth B_c to get a composite TF binding profile (Figure 1C). We process the occupancy profile in such a way for two reasons: (a) the resolution of the short fragment coverage does not distinguish protection from adjacent proteins, and (b) MNase does not completely digest all unprotected DNA, leaving some additional nucleotides flanking any TF's actual binding site. We choose $C_e = C_r = B_c = 10$, though, as with the threshold and bandwidth parameters discussed above, varying the specific values tends to have only small effects on the model predictions. We do not process the nucleosome profile predicted by COMPETE since the model already takes nucleosome padding into consideration.

For promoter region m, we calculate two correlations: the Pearson correlation $r_{1,m}$ between the nucleosome binding profile and the MNase-seq long fragment coverage profile, and the Pearson correlation $r_{2,m}$ between the composite TF binding profile and the MNase-seq small protein coverage profile. The complete pseudo-likelihood function we seek to maximize is defined as:

$$L(\boldsymbol{X}) = \prod_{m=1}^{M} h_m(\boldsymbol{X}) \quad \text{where} \quad h_m(\boldsymbol{X}) = \frac{1}{2}\sqrt{(1 + r_{1,m}) \times (1 + r_{2,m})}.$$

Note that $h_m(\boldsymbol{X})$, which depends on the vector of DBF weights $\boldsymbol{X}$, is the geometric mean of the two rescaled correlations for promoter region m (an example is shown in Figure 1). In this study, $M = 81$.

2.5 Inference Method

We use Markov chain Monte Carlo (MCMC) to explore a posterior distribution based on the pseudo-likelihood function. However, since correlation measures the overall goodness of fit for many genomic locations at once, our pseudo-likelihood function is much flatter than typical likelihood functions. This property can be useful in preventing overfitting, but it also imposes some difficulty for parameter inference. To alleviate this concern, and allow for more efficient MCMC exploration, we apply a temperature parameter τ to each dimension of the search space in order to concentrate the mass of $L(\boldsymbol{X})$ around its modes. We apply a possibly different temperature to each dimension (i.e., each element of the vector $\boldsymbol{X}$) because the pseudo-likelihood in one dimension may be more or less flat than in others. We base our choice of temperature parameter on the MCMC acceptance rate, and empirically set τ for each dimension to be one of $\{0.1, 0.05, 0.01, 0.002\}$. Note that none of these choices change the local maxima of our objective function in any way; they simply may make convergence more efficient.

As for the prior over $\boldsymbol{X}$, a nice feature of our framework is that we can use non-uniform priors if there is reason to do so; later, we explore the possibility of including mildly informative priors for certain TFs where measurements of cellular concentrations in *S. cerevisiae* are available [5]. However, when no relevant information is available, a uniform prior distribution is a natural choice. In what follows, we use a uniform prior over $[-10, 2]$ for log transition weights of

TFs and a uniform prior over $[0, 3]$ for the log transition weight of nucleosomes. Such values are chosen based on the range of TF dissociation constants at their respective optimal binding sites (K_d, as defined and computed by Granek and Clarke [8]). Sig1 has the highest log K_d value of -2.5 and Asg1 has the lowest log K_d value of -7.6. Empirical observations also show that MCMC never produced samples outside these ranges.

In our Gibbs-style MCMC, each iteration consists of an update for each of the transition weight parameters in the model. On a commodity computer cluster, we could compute roughly 25 such iterations per hour.

2.6 Incorporating Pre-initiation Complexes

The pre-initiation complex (PIC) assembles at nucleosome-free promoter regions and facilitates transcription initiation and regulation. PICs compete with other DBFs for binding sites when they are assembled around TATA or TATA-like elements (henceforth referred to as TATA boxes, for simplicity). To account for this competition, we calculate the TATA-binding protein (TBP) binding probability in our model using the DNA binding specificity derived from Rhee and Pugh [20]. Because of the degenerate nature of the TBP binding motif, we amend our model to allow this competition to occur only at TATA boxes (essentially, we set the transition weight for TBP to be 0 at all sequence locations except TATA boxes).

Rhee and Pugh [20] report that core PICs (TBP-associated factors and general transcription factors) assemble approximately 40bp downstream of TATA boxes. The MNase digestion data used here also show an enrichment of short fragments coverage at the same location. Therefore, we approximate the PIC protection by adding the same MNase short fragment coverage shape (scaled by the probability of TBP binding) to the predicted small protein binding probability downstream of the TATA box (see Supplemental Figure S4 for details).

3 Results

3.1 Overall Inference Performance Evaluated by Cross Validation

We randomly split our 81 promoter regions into nine equal sets and performed a standard nine-fold cross validation: parameters were trained on 72 promoter regions using MCMC and we used the average of MCMC samples as trained DBF weights $\hat{\boldsymbol{X}}$; we then calculated $h(\hat{\boldsymbol{X}})$ values for the held out nine promoter regions. Figure 2 shows boxplots of $h(\hat{\boldsymbol{X}})$ values of all the training and testing promoter regions from all the folds of cross validation. We compare the performance to five baselines: (a) average performance when log transition weights are drawn 1000 times uniformly under the prior; or setting the nucleosome transition weight to 35 and TF transition weights to either (b) 8 K_d, (c) 16 K_d, (d) 32 K_d, or (e) 64 K_d.

As Figure 2 shows, our learned model outperforms all five baselines by a significant amount. Note that $h(\boldsymbol{X}) = 0.5$ indicates no correlation on average between

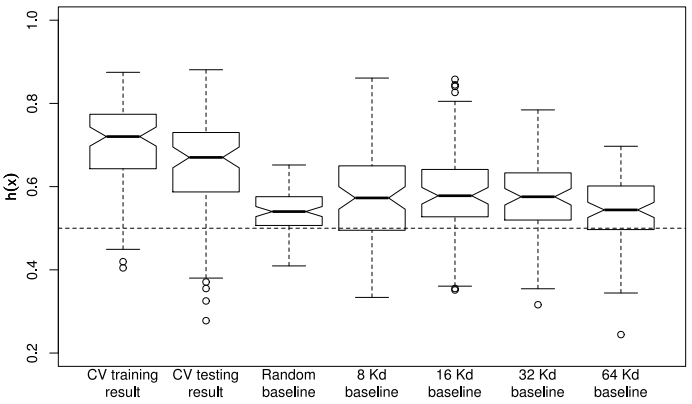

Fig. 2. Comparison of cross validated inference performance to various baselines. Data from the 81 promoter regions were split into nine equal parts. A standard nine-fold cross validation procedure was applied: 72 promoter regions were used as training data to obtain trained DBF weights $\hat{\boldsymbol{X}}$; we then calculated $h(\hat{\boldsymbol{X}})$ values of the held out nine promoter regions (testing results). 'CV training result' considers the $h(\hat{\boldsymbol{X}})$ values for each promoter when used as training data. 'CV testing result' shows the $h(\hat{\boldsymbol{X}})$ values for each promoter when used as testing data. Uniformly drawn TF transition weights and different multiples of K_d are used as baseline comparisons. Variance is reduced in the random baseline case because each result is the average of 1000 random samples.

the model predictions and observed data. We observe that median performance for the random baseline is still larger than 0.5 even though the TF transition weights are uninformed guesses; this is because the model's emission parameters (derived from *in vitro* experimental data regarding TF and nucleosome binding specificity) are highly informed.

3.2 A Mechanistic Explanation for Paired-End MNase-seq Data

Owing to *in vitro* experiments, our model has knowledge about inherent DBF sequence specificities. The thermodynamic interaction and competition between these DBFs are accounted for by COMPETE. By adding information about *in vivo* DBF binding occupancy levels present in MNase-seq data, our framework can now infer a DBF binding landscape that provides a mechanistic explanation for the observed data.

Figure 3 illustrates examples of predicted binding profiles for each DBF in six promoter regions in the test sets of the nine-fold cross validation, in comparison with the corresponding MNase-seq binding profile tracks (see Supplemental Figure S3 for raw coverage and Supplemental Figure S5 for additional comparisons between composite predicted profiles and processed MNase-seq fragment coverages). These examples span the full spectrum of our framework performance, from strong performance to weak performance. In all cases, our predictions for the TF binding profiles provide a good or fair explanation for the MNase-seq

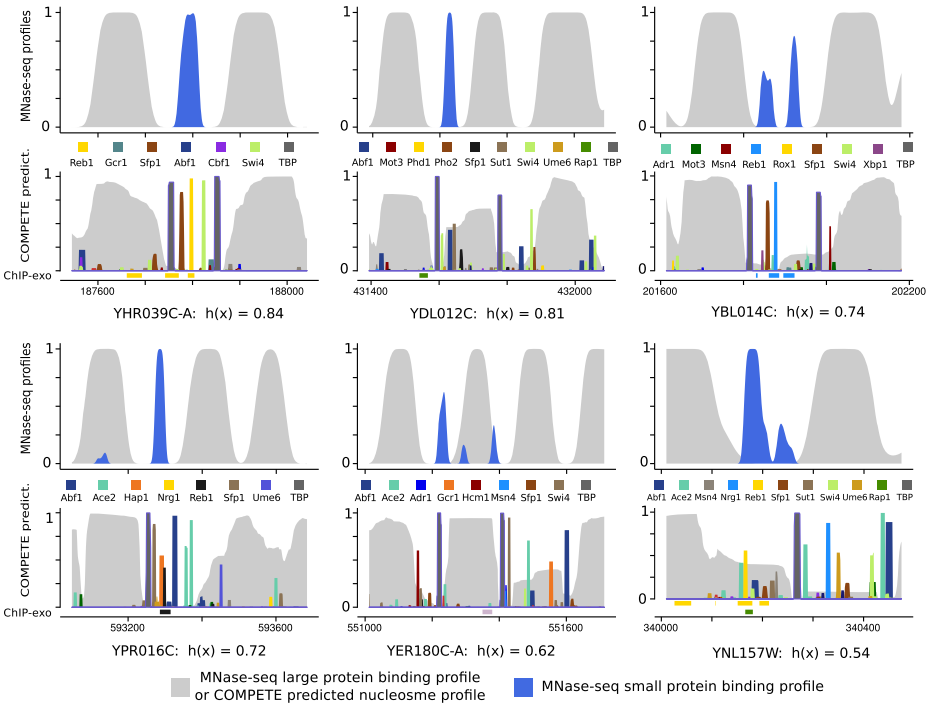

Fig. 3. Predicted binding profiles versus MNase-seq binding profiles. For six promoter regions in our 81 promoter set, we plot the predicted binding profiles when they were evaluated as testing data. We also indicate reported binding sites from ChIP-exo [19] underneath the predicted binding profiles; these have the same color as the corresponding TF's binding probability. No binding event is reported by MacIsaac et al. [15, p-value < 0.001 and conservation level 3] for these promoter regions.

data and are much more consistent with the data compared to random baseline predictions (see Supplemental Figure S1), considering the simplicity of our approach and the complexity of the problem.

One difficulty in interpreting high-throughput nuclease digestion data is identifying the binding proteins at read-enriched regions. Traditional motif matching is not satisfactory when there are multiple potentially overlapping motifs, nor can it assess the strength of protein binding. In contrast, our framework provides a principled interpretation for the data in terms of distinct binding events, each with its own probability of occurrence based on evaluating the probability of every possible binding configuration in the ensemble. This is demonstrated, for example, in the YDL012C and YPR016C promoter regions. Our approach can also capture weak binding events, such as the Reb1 binding events in the YPR016C and YNL157W promoter regions, which are missed in ChIP-chip experiments [15] but are captured in ChIP-exo experiments [19] (Figure 3; reported ChIP-exo binding sites are indicated underneath the predicted TF binding

landscape; no binding event is reported by MacIsaac et al. [15, p-value < 0.001 and conservation level 3] for these promoter regions).

Our predictions of nucleosome binding profiles match the data well in spite of the fact that nucleosome positioning is less precise than TF positioning. The predictions reflect the intrinsic uncertainty about nucleosome positioning related to their mobility and only mildly sequence preferences, especially when the MNase-seq large protein binding profile is more noisy, as in the promoter regions of YBL014C and YNL157W (Figure 3; see Supplemental Figure S3 for raw coverage).

3.3 Incorporating Measurements of Protein Concentration through Prior Distributions

We have demonstrated that our framework can achieve good performance using non-informative priors. However, the framework could potentially perform better by incorporating prior information when it is available. For instance, Ghaemmaghami et al. [5] measured cellular protein concentrations using Western blots in *S. cerevisiae* during log phase growth. As discussed above, although cellular protein concentrations are not precisely equivalent to the transition weights we are estimating, the two still might be expected to loosely correlate with one another. We can therefore use these measurements to construct weak prior distributions for the corresponding DBF transition weights. To account for the loose correlation between the two, as well as experimental measurement error, we use a truncated normal prior for log transition weights with a large standard deviation of 2 (so a standard deviation in each direction corresponds to multiplying the weight by 1/100 or 100, respectively). We calculate the mean for this normal prior by converting measurements from Ghaemmaghami et al. [5] to molar concentration using a yeast cell volume of 5×10^{-14}L [2]. The resulting prior means are in the range of -8 to -6 in log scale. Note that nine of the 42 TFs in our model do not have measurements available, and thus their priors remain uniform, as described above.

When we utilize this prior information, we observe no change in training performance and a marginal increase in testing performance (median $h(\hat{\boldsymbol{X}})$ increases by 0.013; Figure 4). Such an insignificant result could arise for multiple reasons: (a) the aforementioned difference between cellular concentration and the model's transition weights means that the information provided by the measured concentrations might not even be relevant; (b) the noisy physiological measurements of both cellular concentration and cell volume means that the measurements we used might not be quite accurate; or (c) the weak prior we utilized in the model because the measured concentrations are not trusted to be very precise means that the objective function landscape might change only slightly.

4 Discussion

We show that integrating information from experimental data within a general framework built on a thermodynamic ensemble model of competitive factor

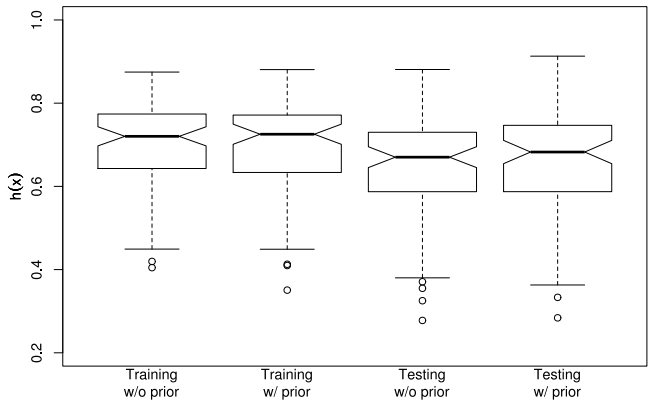

Fig. 4. Comparison of cross validation performance with and without prior information regarding measured cellular protein concentration. Performance for each promoter is measured by the geometric mean ($h(\hat{\boldsymbol{X}})$) of the two Pearson correlations defined in Figure 1. Each boxplot shows the performance summary of the 81 promoter regions across all the cross validation trials.

binding can improve the accuracy of inferred protein-DNA interactions, providing a more biologically plausible view of the protein-DNA interaction landscape. Such a landscape gives a mechanistic explanation for observed paired-end MNase-seq fragments through various protein binding events, each with its own probability of occurrence. Many of those binding events are weak binding events that are typically missed in other modeling methods, but are captured in our framework; these weaker binding events are also supported by higher resolution experimental data where available [20]. These weak binding events are important: It has been reported that low affinity protein-DNA interactions may be involved in fine-tuning transcriptional regulation and are common along the genome [1, 22, 23]. Our framework's predictions agree with this viewpoint: 72% of the binding events in our predicted profiles have a probability lower than 0.5. Our framework could thus form an important basis for future computational work that connects transcriptional activity with the protein-DNA interaction landscape.

Our framework does not successfully predict a few TF binding events reported by high resolution ChIP-exo experiments [19], most notably some of the binding sites for Phd1 and Reb1. We believe the primary reason is occasional mismatches between our input TF PWMs and these proteins' actual *in vivo* DNA-binding specificities. For Phd1, Rhee and Pugh [20] report several distinct *in vivo* motifs. However, the Phd1 PWM we used in our framework comes from *in vitro* data [27] and does not match the *in vivo* DNA-binding specificity of Phd1 reported by Rhee and Pugh [20]. Similarly, for Reb1, Rhee and Pugh [20] report that 40% of Reb1 binding sites are so-called 'secondary binding sites', with motifs that deviate from the TTAGGC consensus of the *in vitro* PWM we are using.

This mismatch in DNA binding specificity may account for much of the discrepancy between our predicted profiles and reported binding sites. However, some caution should be taken when interpreting *in vivo* ChIP data, since the assay cannot distinguish between direct protein-DNA interaction and indirect interaction [6]. We also note that our current framework only includes a subset of all yeast TFs. Some unexplained short fragment coverage peaks, such as those in the YBL014C promoter region, could indicate the binding of DBFs that are not in our set. These and other discrepancies may have an impact on our overall inference, resulting in missing binding events (or possibly even superfluous binding events, because of the competition that is inherent in our model).

In the promoters of YNL157W and YDL012C, our predictions do not include Rap1 binding events even though they are reported in ChIP-exo experiments. However, we believe this results from the nature of Rap1 binding: Lickwar et al. [13] report that Rap1 binding on non-ribosomal protein promoters, like the two mentioned above, is highly dynamic and involves fast turnover. Such binding events are possibly captured in ChIP experiments because of cross-linking, but may be difficult to observe in an MNase-based digestion experiment if the latter does not involve a cross-linking step. Incidentally, the two ChIP-determined Rap1 binding events are not close to MNase-seq small fragment coverage peaks. One possible use of our framework for extending the results shown here would be to incorporate data from ChIP-based experiments and use the framework to estimate parameters that reflect information from both kinds of data.

We designed our model to take advantage of published data on PIC positions [20]. Such data provides additional protein occupancy information that is likely the result of mechanisms beyond TBP-related protein sequence specificity and protein competition. We observed that adding PICs allowed the nucleosome free regions to agree better with the MNase fragment data, because the PICs both enhance the exclusion of nucleosomes and explain some of the small MNase fragments downstream of the TATA-like element (so that TFs are not needed to provide that explanation).

We also demonstrate the use of prior information in our framework through incorporating measured bulk cellular protein concentration. The model performance improved marginally, which can be interpreted two ways. On the one hand, it is reassuring that one need not have measured cellular protein concentrations in order to perform effective inference. The fact that our uniform priors work as well as having priors informed by measured concentrations means that the measured concentrations available currently are not critical for good performance. However, that said, it is also reassuring that our framework has the ability to incorporate this sort of prior information when available because we anticipate such data will only improve. As measurement technologies enable us to move from bulk cellular concentrations toward nuclear concentrations of active TFs, we anticipate that the ability to incorporate prior information will become more useful, if not for achieving better results then perhaps at least for more rapid convergence toward optima when we move to higher-dimensional inference (e.g., more TFs).

With adequately fitted parameters, our framework has the potential to perform *in silico* simulation for various environmental conditions by changing the protein concentrations. For example, we could simulate *in silico* heat shock by increasing the concentration of heat shock response factors in our model. We could also investigate how certain single nucleotide polymorphisms (SNP) affect the overall protein-DNA interaction landscape, not just at the site of the SNP but propagating to the surrounding region due to altered competition.

This work represents a first step toward a more general framework. By specifying probabilistic distributions appropriate for other kinds of experiments—like ChIP-seq, FAIRE-seq, or DNase-seq—the framework can integrate other sources of data through a joint likelihood. As more and larger-scale sequencing projects are carried out, such a framework will prove extremely valuable for integrating different pieces of information to infer a more precise view of the protein-DNA interactions that govern transcriptional regulation.

Supplemental information is available from `http://www.cs.duke.edu/~amink/`

Acknowledgments. We would like to thank Jason Belsky, Kaixuan Luo, Yezhou Huang, and Michael Mayhew for helpful discussions and comments.

References

[1] Biggin, M.: Animal transcription networks as highly connected, quantitative continua. Developmental Cell 21(4), 611–626 (2011)
[2] Bryan, A.K., Goranov, A., Amon, A., et al.: Measurement of mass, density, and volume during the cell cycle of yeast. Proceedings of the National Academy of Sciences 107(3), 999–1004 (2010)
[3] Chen, X., Hoffman, M., Bilmes, J., et al.: A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. Bioinformatics 26(12), i334–i342 (2010)
[4] Foat, B., Morozov, A., Bussemaker, H.: Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics 22(14), e141–e149 (2006)
[5] Ghaemmaghami, S., Huh, W.K., Bower, K., et al.: Global analysis of protein expression in yeast. Nature 425(6959), 737–741 (2003)
[6] Gordân, R., Hartemink, A.J., Bulyk, M.: Distinguishing direct versus indirect transcription factor-DNA interactions. Genome Research 19(11), 2090–2100 (2009)
[7] Gordân, R., Murphy, K., McCord, R., et al.: Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biology 12(12), R125 (2011)
[8] Granek, J., Clarke, N.: Explicit equilibrium modeling of transcription-factor binding and gene regulation. Genome Biology 6(10), R87 (2005)
[9] Harbison, C., Gordon, D., Lee, T., et al.: Transcriptional regulatory code of a eukaryotic genome. Nature 431(7004), 99–104 (2004)
[10] Henikoff, J., Belsky, J., Krassovsky, K., et al.: Epigenome characterization at single base-pair resolution. Proceedings of the National Academy of Sciences 108(45), 18318–18323 (2011)

[11] Hesselberth, J., Chen, X., Zhang, Z., et al.: Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature Methods 6(4), 283–289 (2009)
[12] Kaplan, T., Li, X.Y., Sabo, P., et al.: Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genetics 7(2), e1001290 (2011)
[13] Lickwar, C.R., Mueller, F., Hanlon, S.E., et al.: Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature 484(7393), 251–255 (2012)
[14] Luo, K., Hartemink, A.J.: Using DNase digestion data to accurately identify transcription factor binding sites. In: Pacific Symposium on Biocomputing, pp. 80–91. World Scientific (2013)
[15] MacIsaac, K., Wang, T., Gordon, D., et al.: An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7, 113 (2006)
[16] Pique-Regi, R., Degner, J.F., Pai, A.A., et al.: Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research 21(3), 447–455 (2011)
[17] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
[18] Ren, B., Robert, F., Wyrick, J., et al.: Genome-wide location and function of DNA binding proteins. Science 290(5500), 2306–2309 (2000)
[19] Rhee, H., Pugh, B.: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147(6), 1408–1419 (2011)
[20] Rhee, H., Pugh, B.: Genome-wide structure and organization of eukaryotic preinitiation complexes. Nature 483(7389), 295–301 (2012)
[21] Saul, L., Jordan, M.: Boltzmann chains and hidden Markov models. Advances in Neural Information Processing Systems, pp. 435–442. MIT Press (1995)
[22] Segal, E., Raveh-Sadka, T., Schroeder, M., et al.: Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451(7178), 535–540 (2008)
[23] Tanay, A.: Extensive low-affinity transcriptional interactions in the yeast genome. Genome Research 16(8), 962–972 (2006)
[24] Teif, V., Rippe, K.: Calculating transcription factor binding maps for chromatin. Briefings in Bioinformatics 13(2), 187–201 (2012)
[25] Wasson, T., Hartemink, A.J.: An ensemble model of competitive multi-factor binding of the genome. Genome Research 19(11), 2101–2112 (2009)
[26] Weirauch, M.T., Cote, A., Norel, R., et al.: Evaluation of methods for modeling transcription factor sequence specificity. Nature Biotechnology 31(2), 126–134 (2013)
[27] Zhu, C., Byers, K., McCord, R., et al.: High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Research 19(4), 556–566 (2009)