BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 20 2014, pages 2868–2874
doi:10.1093/bioinformatics/btu408

Genome analysis

Advance Access publication June 27, 2014

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]Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, [2]Knowledge Systems and Informatics, Lawrence Livermore National Laboratory, Livermore, CA 94550 and [3]Department of Computer Science, Duke University, Durham, NC 27708, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Transcriptional regulation is directly enacted by the interactions between DNA and many proteins, including transcription factors (TFs), 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 partial information regarding one aspect of the complete interaction landscape. For example, chromatin immunoprecipitation (ChIP) reveals the binding positions of a protein, but only for one protein at a time. In contrast, nucleases like MNase and DNase can be used to reveal binding positions for many different proteins at once, but cannot easily determine the identities of those proteins. Currently, few statistical frameworks jointly model these different data sources to reveal an accurate, holistic view of the *in vivo* protein–DNA interaction landscape.

Results: 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.

Availability and implementation: The C source code for COMPETE and Python source code for MCMC-based inference are available at http://www.cs.duke.edu/~amink.

Contact: amink@cs.duke.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received and revised on May 16, 2014; accepted on June 20, 2014

1 INTRODUCTION

As an essential component of transcriptional regulation, the interaction between DNA-binding factors (DBFs) and DNA has been studied extensively. To map genome-wide protein–DNA interactions experimentally, two basic categories of techniques have been developed: ChIP-based methods (numerous studies in many organisms, but a few examples for yeast are Harbison *et al*., 2004; Ren *et al*., 2000; Rhee and Pugh, 2011); and nuclease digestion-based methods that profile chromatin with either DNase (Hesselberth *et al*., 2009) or MNase (Henikoff *et al*., 2011). ChIP methods can be used to reveal high-resolution DNA interaction sites for a single antibody-targeted factor, especially the recently developed ChIP-exo methods (Rhee and Pugh, 2011) 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, but 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 these data to generate deeper insights into transcriptional regulation.

A popular set of computational models use these data to search for overrepresented DNA sequences bound by certain DBFs; these are often applied in the setting of motif discovery (Foat *et al*., 2006; Harbison *et al*., 2004; MacIsaac *et al*., 2006; Tanay, 2006). More recently, models have been applied to DNase-seq data to identify 'digital footprints' of DBFs (Chen *et al*., 2010; Hesselberth *et al*., 2009; Luo and Hartemink, 2012; Pique-Regi *et al*., 2011). 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 (2011)]. Second, most computational methods model the binding events for one kind of protein at a time instead of taking into consideration the interactions among different kinds of DBFs, especially nucleosomes. Although the work of Kaplan *et al*. (2011), Segal *et al*. (2008) and Teif and Rippe (2012) are notable exceptions, these all consider small genomic regions and include only a few TFs; Segal *et al*. (2008) did not consider the role of nucleosomes. Third, and most importantly, almost all current methods fail to integrate different kinds of datasets. This is suboptimal because data from one kind of experiment only reveal partial information about the *in vivo* protein–DNA interaction landscape. For example, ChIP datasets only contain binding information for one specific protein under one specific condition; nuclease digestion datasets provide binding information for all proteins, but do not reveal the

*To whom correspondence should be addressed.

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 developed a computational model of protein–DNA interactions, termed COMPETE (Wasson and Hartemink, 2009), 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 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 (Supplementary Fig. S1). It is therefore necessary to develop a new computational framework for jointly interpreting experimentally derived data regarding genomic occupancy within a model built on the thermodynamic foundation of COMPETE.

Here, we develop just such a method: a general framework that combines a thermodynamic model for protein–DNA interactions and a new statistical model for learning from experimental observations regarding those interactions. Information from different experimental observations can be integrated to infer the thermodynamic interactions between DBFs and a genome. In this particular study, we demonstrate the use of this framework by integrating paired-end MNase-seq data, which reveal 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 with 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. TFs 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 to 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, 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. To simplify notation, we have treated each unbound nucleotide as being bound by a special kind of 'empty' DBF. 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) (Rabiner, 1989), 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 TFs, 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 (Weirauch *et al.*, 2013)]. 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.* (2011), 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 (Saul and Jordan, 1995), which can be understood as an 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 (Ghaemmaghami *et al.*, 2003). 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 extranuclear 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.* (2011). 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 the fragments into long (140–200 bp) and short (0–100 bp) 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 TFs.

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 (Fig. 1D 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. We performed a sensitivity analysis and observed that our results are largely unaffected across a broad range of these parameters (Supplementary Fig. S2). We also note that MNase is known to prefer to cut A/T compared with G/C. We assessed the severity of this well-known bias (Supplementary Figs. S3 and S4) and observed that it does not affect our final results, primarily because we are not using profiles of the total number of cuts at each genomic position, but rather using the full fragments to generate profiles of fragment coverage; while the former might be 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 *Saccharomyces cerevisiae* TFs. However, our choice of TFs 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 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.* (2006). 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.* (2006) (at $P<0.005$ and conservation level 3).

Having selected our 42 TFs, we next chose a set of promoter regions that, according to MacIsaac *et al.* (2006) (at $P<0.005$ and conservation level 3), are bound exclusively by those TFs. For this study, we focus on 81 such promoter regions. We 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 200 bp downstream of each TATA element. For the other (non-divergent) promoters, we extracted MNase-seq data 500 bp upstream of the TATA element (or 100 bp upstream of the end of the upstream gene, whichever is smaller), and 200 bp downstream of the TATA element. Locations of TATA elements were taken from Rhee and Pugh (2012).

2.4 Incorporating MNase-seq data through an objective function

We model MNase-seq data through a pseudo-likelihood function. To calculate the function, we process the COMPETE output TF binding probabilities as follows: the binding probability of each 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 probabilities of all TFs (truncating values >1) and smooth the track using a Gaussian kernel of bandwidth B_c to get a composite TF binding profile (Fig. 1C). We process the occupancy profile in such a way for two reasons: (i) the resolution of the short fragment coverage does not distinguish protection from adjacent proteins, and (ii) 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

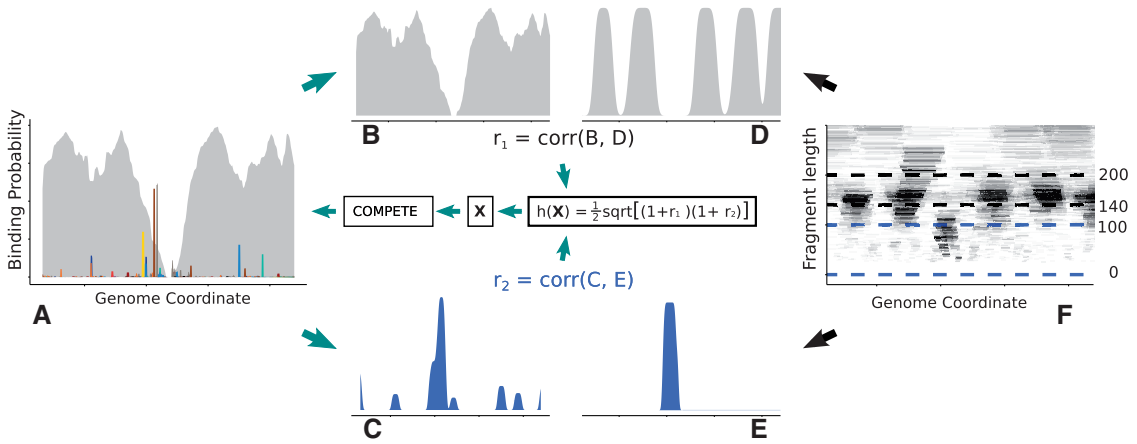

Fig. 1. Overview of how the objective function is evaluated and iteratively optimized. (**A**) Predicted probability that each particular DBF binds at a given genome position, as calculated by COMPETE, given current DBF weights. We then separate these probabilities into two profiles: (**B**) predicted 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 with a short fragment coverage profile. Similarly, we separate the observed MNase-seq fragments (**F**) into long (140–200 bp) and short (0–100 bp) 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 with 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 with 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

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, as the model already takes nucleosome padding into consideration. We also capture the effect of the pre-initiation complex (PIC) on the MNase-seq profile by adding an empirical PIC protection shape to the predicted binding profile (Supplementary Text and Supplementary Fig. S5).

For promoter region m, we calculate two correlations: the Pearson correlation $r_{1,m}$ between the nucleosome binding profile and the MNase-seq large protein 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

$$L(\boldsymbol{X}) = \prod_{m=1}^{M} h_m(\boldsymbol{X})$$

where $h_m(\boldsymbol{X}) = \frac{1}{2}\sqrt{(1+r_{1,m}) \times (1+r_{2,m})}$, the geometric mean of the two rescaled correlations for promoter region m (an example is shown in Fig. 1). Note that $h_m(\boldsymbol{X})$ depends on the vector of DBF weights $\boldsymbol{X}$. 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, as 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 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 (Ghaemmaghami *et al.*, 2003). 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 (2005)]: TF Sig1 has the highest log K_d value of –2.5 and TF Asg1 has the lowest log K_d value of –7.6.

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.

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 9-fold cross-validation: parameters were trained on 72 promoter regions using MCMC and we used the mean of the MCMC samples as trained DBF weights $\hat{\boldsymbol{X}}$; we then calculated $h(\hat{\boldsymbol{X}})$ values for the nine held out 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: 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 8 K_d, 16 K_d, 32 K_d or 64 K_d.

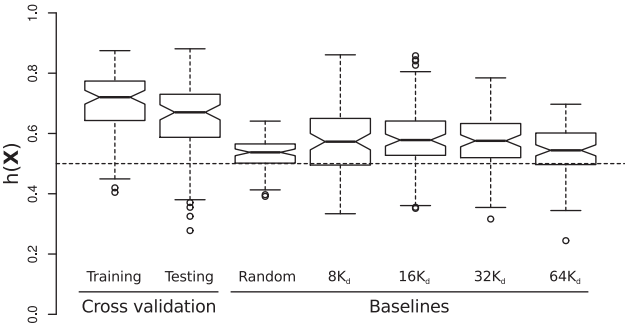

Fig. 2. Comparison of cross-validated inference performance with various baselines. Data from the 81 promoter regions were split into nine equal parts. A standard 9-fold cross-validation procedure was applied: 72 promoter regions were used as training data to obtain trained DBF weights $\hat{\boldsymbol{X}}$; we then calculated $h(\hat{\boldsymbol{X}})$ values of the nine held out promoter regions (testing results). 'Cross validation training' considers the $h(\hat{\boldsymbol{X}})$ values for each promoter when used as training data. 'Cross validation testing' shows the $h(\hat{\boldsymbol{X}})$ values for each promoter when used as testing data. Uniformly randomly 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

As Figure 2 shows, our learned model outperforms all five baselines significantly. Note that $h(\boldsymbol{X}) = 0.5$ indicates no correlation on average between the model predictions and observed data. Median performance for the random baseline is still >0.5 even with uninformed TF transition weights; 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 9-fold cross-validation, in comparison with the corresponding MNase-seq binding profile tracks (see Supplementary Fig. S4 for raw coverage and Supplementary Fig. S6 for additional comparisons between composite predicted profiles and processed MNase-seq fragment coverages). These examples span the full spectrum of

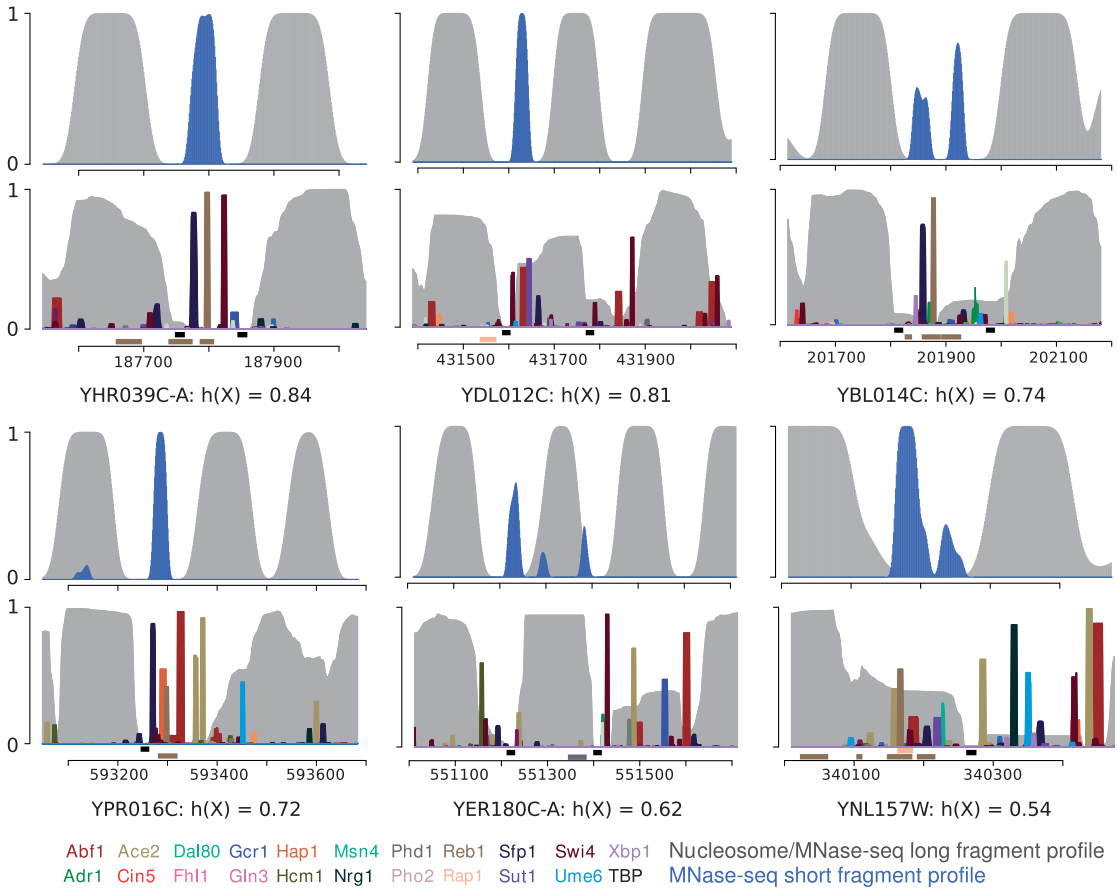

Fig. 3. Predicted binding profiles versus MNase-seq binding profiles. For six promoter regions in our 81 promoter set, we plot the predicted binding profiles when they were evaluated as testing data. We also indicate reported binding sites from ChIP-exo (Rhee and Pugh, 2011) underneath the predicted binding profiles; these have the same color as the corresponding TF's binding probability. No binding event is reported by MacIsaac *et al.*, 2006, ($P<0.001$ and conservation level 3) for these promoter regions

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 data and are much more consistent with the data compared with random baseline predictions (Supplementary Fig. S1).

One difficulty in interpreting high-throughput nuclease digestion data is identifying DBFs 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 (MacIsaac *et al.*, 2006) but are captured in ChIP-exo experiments (Rhee and Pugh, 2011; Fig. 3).

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 mild 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 Supplementary Fig. S4 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.* (2003) 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

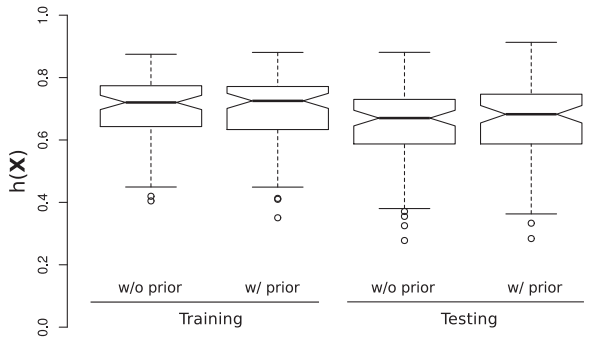

Fig. 4. Comparison of cross-validation performance with and without prior information regarding measured cellular protein concentration. Performance for each promoter is measured by the geometric mean $[h(\hat{X})]$ of the two Pearson correlations defined in Figure 1. Each boxplot shows the performance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

well as experimental measurement error, we use a truncated normal prior for log transition weights with a large standard deviation (SD) of 2 (so 1 SD 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*. (2003) to molar concentration using a yeast cell volume of 5×10^{-14} L (Bryan *et al*., 2010). 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 use this prior information, we observe no change in training performance and a marginal increase in testing performance (median $h(\hat{X})$ increases by 0.013; Fig. 4). Such an insignificant result could arise for multiple reasons: (i) 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, (ii) the noisy physiological measurements of both cellular concentration and cell volume means that the measurements we used might not be accurate, or (iii) the weak prior that we used in the model because the measured concentrations are not trusted to be 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 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 (Rhee and Pugh, 2012). 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 (Biggin, 2011; Segal *et al*., 2008; Tanay, 2006). Our framework's predictions agree with this viewpoint: 72% of the binding events in our predicted profiles have a binding probability less 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 (Rhee and Pugh, 2011), 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 (2012) reported several distinct *in vivo* motifs. However, the Phd1 PWM we used in our framework comes from *in vitro* data (Zhu *et al*., 2009) and does not match the *in vivo* DNA-binding specificity of Phd1 reported by Rhee and Pugh (2012). Similarly, for Reb1, Rhee and Pugh (2012) reported 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 used. 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, as the assay cannot distinguish between direct protein–DNA interaction and indirect interaction (Gordân *et al*., 2009). 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*. (2012) 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 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 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 because of 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 valuable for integrating different pieces of information to infer a more precise view of the protein–DNA interactions that govern transcriptional regulation.

ACKNOWLEDGMENTS

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

Funding: This work was funded in part by grants from NIH (P50 GM081883-01) and DARPA (HR0011-09-1-0040) to A.J.H.

Conflict of interest: none declared.

REFERENCES

Biggin,M. (2011) Animal transcription networks as highly connected, quantitative continua. *Dev. Cell*, **21**, 611–626.

Bryan,A.K. *et al.* (2010) Measurement of mass, density, and volume during the cell cycle of yeast. *Proc. Natl Acad. Sci. USA*, **107**, 999–1004.

Chen,X. *et al.* (2010) A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. *Bioinformatics*, **26**, i334–i342.

Foat,B. *et al.* (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

Ghaemmaghami,S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Gordân,R. *et al.* (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res.*, **19**, 2090–2100.

Gordân,R. *et al.* (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol.*, **12**, R125.

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

Harbison,C. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Henikoff,J. *et al.* (2011) Epigenome characterization at single base-pair resolution. *Proc. Natl Acad. Sci. USA*, **108**, 18318–18323.

Hesselberth,J. *et al.* (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

Kaplan,T. *et al.* (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet.*, **7**, e1001290.

Lickwar,C.R. *et al.* (2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, **484**, 251–255.

Luo,K. and Hartemink,A.J. (2013) Using DNase digestion data to accurately identify transcription factor binding sites. In: Altman,R. *et al.* (eds) *Pacific Symposium on Biocomputing 2013 (PSB13)*. World Scientific, New Jersey, pp. 80–91.

MacIsaac,K. *et al.* (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Pique-Regi,R. *et al.* (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.

Rabiner,L. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Rhee,H. and Pugh,B. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.

Rhee,H. and Pugh,B. (2012) Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, **483**, 295–301.

Saul,L. and Jordan,M. (1995) Boltzmann chains and hidden Markov models. In: Tesauro,G. (ed.) *Advances in Neural Information Processing Systems*. Vol. 7, The MIT Press, Cambridge, MA, pp. 435–442.

Segal,E. *et al.* (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**, 535–540.

Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.*, **16**, 962–972.

Teif,V. and Rippe,K. (2012) Calculating transcription factor binding maps for chromatin. *Brief. Bioinform.*, **13**, 187–201.

Wasson,T. and Hartemink,A.J. (2009) An ensemble model of competitive multi-factor binding of the genome. *Genome Res.*, **19**, 2101–2112.

Weirauch,M.T. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.

Zhu,C. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.