Precise physical models of protein–DNA interaction from high-throughput data

Justin B. Kinney, Gašper Tkačik, and Curtis G. Callan, Jr.*

Physics Department and Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544

Contributed by Curtis G. Callan, Jr., November 8, 2006 (sent for review September 30, 2006)

A cell's ability to regulate gene transcription depends in large part on the energy with which transcription factors (TFs) bind their DNA regulatory sites. Obtaining accurate models of this binding energy is therefore an important goal for quantitative biology. In this article, we present a principled likelihood-based approach for inferring physical models of TF–DNA binding energy from the data produced by modern high-throughput binding assays. Central to our analysis is the ability to assess the relative likelihood of different model parameters given experimental observations. We take a unique approach to this problem and show how to compute likelihood without any explicit assumptions about the noise that inevitably corrupts such measurements. Sampling possible choices for model parameters according to this likelihood function, we can then make probabilistic predictions for the identities of binding sites and their physical binding energies. Applying this procedure to previously published data on the *Saccharomyces cerevisiae* TF Abf1p, we find models of TF binding whose parameters are determined with remarkable precision. Evidence for the accuracy of these models is provided by an astonishing level of phylogenetic conservation in the predicted energies of putative binding sites. Results from *in vivo* and *in vitro* experiments also provide highly consistent characterizations of Abf1p, a result that contrasts with a previous analysis of the same data.

binding energy | likelihood | transcription factor | mutual information

Transcription factors (TFs) are central to the cell's ability to regulate gene expression (1). Any quantitative understanding of gene regulation will therefore require an accurate characterization of the specificity with which these proteins recognize their DNA target sites. This problem is usually phrased in one of two ways: can we find a statistical pattern (or "motif") that distinguishes TF binding sites from the genomic background or, alternatively, can we find a faithful representation of the TF's sequence-dependent binding energy (SDBE) to DNA? For a variety of reasons, the development of motif-finding algorithms addressing the statistical question has received much more attention than attempts to directly model TF binding energy. Although the statistical and the energetic pictures are related, they are not equivalent. Indeed, knowledge of binding energy is essential for an understanding of gene regulatory dynamics: a TF binds, not in response to a statistical P value, but to the physical affinity of a site. In this article we describe a method for inferring physical models of binding energy from genome-scale TF binding assays, such as ChIP–chip (2, 3) or protein binding microarrays (PBMs) (4). Instead of seeking a "best" model, our method finds ensembles of models with a high likelihood of being responsible for the data. This probabilistic approach allows direct comparison of results from different experiments and provides additional ways of making biologically relevant predictions.

To set a context for our proposal, we briefly compare and contrast the statistical and energetic approaches to TF specificity. Because high-throughput binding assays generally localize TFs to within only ~500 bp, one often faces the problem of predicting where, precisely, a TF binds within experimentally bound regions of DNA. The statistics-based algorithms that have been proposed to do this locate binding sites (see ref. 5 for a concise summary) typically proceed by positing a model for "background" DNA and then searching for statistical motifs describing short DNA sequences that appear more often in TF-bound regions than otherwise expected.

In a seminal work, Berg and von Hippel (6) [see also the work of Stormo (7) and Stormo and Fields (8)] proposed a method for estimating the SDBE of a TF from the sequence statistics of known binding sites, and this method is often used to estimate TF–DNA binding energies from the output of motif-finding programs. However, Berg and von Hippel's derivation assumes (*i*) that all of the binding sites of a given TF experience the same selection pressure on their energy and (*ii*) that such sites evolve out of random background DNA (see refs. 9 and 10 for a more general discussion of how binding site sequence, energy, and fitness relate to one another). In reality, different sites may need to have different binding energies for functional reasons, and the noncoding DNA of many organisms is clearly not random [*Plasmodium falciparum* is an instructive example: see supporting information (SI) Table 1]. Although the Berg–von Hippel formula provides a plausible connection between a TF's SDBE and the sequences of its target sites, it is unlikely to be exact or universally valid.

In fact, it should not be necessary to invoke such a connection at all in analyzing PBM and ChIP–chip data, because both assays probe the SDBE of the TF rather directly. This is clear for PBM experiments, where the TF is directly bound *in vitro* to dsDNA-spotted microarrays. Things are more subtle in ChIP–chip experiments where *in vivo* effects can modify TF binding (1): in yeast, chromatin can obscure binding sites and other transfactors can alter TF–DNA binding energy. Although such effects are biologically important, they generally vary from site to site, whereas the SDBE of the TF acts in the same way throughout the genome. Thus, for some purposes, it is possible to regard these *in vivo* effects as a type of "noise" that obscures, but does not negate, the influence of the TF's SDBE.

One should therefore be able to model this energy directly from ChIP–chip and PBM data without calling on any assumptions about binding site evolution or the statistics of background DNA. Foat *et al.* (11) have recently developed such an approach: they postulate a simple parametrized model of the TF's SDBE and then find the parameters that best fit experimental data. Although it is a big step in the right direction, the Foat *et al.* analysis leaves some things to be desired: it makes the implicit assumption that measurement noise is Gaussian, an assumption that is not always supported by the data (see SI Fig. 5); more importantly, it returns only the best set of parameters, rather

Author contributions: J.B.K., G.T., and C.G.C. designed research; J.B.K., G.T., and C.G.C. performed research; J.B.K., G.T., and C.G.C. analyzed data; and J.B.K. and C.G.C. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: TF, transcription factor; SDBE, sequence-dependent binding energy; PBM, protein binding microarray; EMA, error-model-averaged; MCMC, Markov chain Monte Carlo; LIR, log intensity ratio; HF, hit fraction.

*To whom correspondence should be addressed. E-mail: ccallan@princeton.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0609908104/DC1.

© 2006 by The National Academy of Sciences of the USA

than giving the full range of parameters consistent with given observations. In this article, we address these problems and in the process develop a simple but general method for analyzing TF–DNA binding data.

Our basic method is to use experimental data to specify a probability distribution on the parameters of a model of the TF's SDBE. By model we mean a function, involving parameters denoted collectively by θ, that takes a potential binding site sequence as input and gives an energy as output. The distribution on models is then used to make predictions (e.g., for binding site energies) that have mean values and variances. We think of the data as a set of values $\{z_i\}_{i=1}^N$ (such as observed fluorescence intensities) obtained for the N regions of DNA probed by the experiment. Because of experimental noise, particular model parameters θ will produce the observed data with a probability $p(\{z_i\}|\theta)$ (also referred to as the "likelihood"), which can be computed if the statistics of the noise are known. Given a prior distribution $p(\theta)$ on allowable model parameters, we can use Bayes' theorem to turn this into a posterior distribution on model parameters, given the observed data:

$$p(\theta|\{z_i\}) \propto p(\{z_i\}|\theta)p(\theta). \quad [1]$$

Unfortunately, traditional methods of computing $p(\{z_i\}|\theta)$ require a quantitative model of the experimental noise (or "error model"), something that is not usually available; moreover, using the wrong error model will generally lead to incorrect inferences. To deal with this problem, we take the further Bayesian step of averaging $p(\{z_i\}|\theta)$ over the space of all possible error models to obtain an "error-model-averaged" (EMA) likelihood that can be explicitly evaluated and used in Eq. **1**. Our major result is that this relaxed version of likelihood still allows real data to constrain energy models: in effect the data are used to determine both the energy model and the error model.

To make predictions by using $p(\theta|\{z_i\})$ we use a Markov chain Monte Carlo (MCMC) algorithm to generate a large ensemble of models $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$, sampled according to this distribution. Θ is then used to give concrete probabilistic answers to questions about TF binding behavior conditioned on the experimental data. We demonstrate this approach on ChIP–chip (3) and PBM (4) studies of the yeast TF Abf1p and find that the data determine the parameters of simple binding models with remarkably low statistical uncertainty. This finding contrasts with the commonly held view that high-throughput experiments can give only rough characterizations of TF specificity. Although the microarray data are noisy, the sheer number of regions probed, along with the large amount of DNA sequence in each region, allows for a precise characterization of SDBE. We have found the same to be true for some other broad-acting yeast TFs (data not shown).

Results

We first present the results of a likelihood analysis of the PBM data of Mukherjee *et al.* (4) for the yeast TF Abf1p. In this assay, epitope-tagged TFs were bound directly to dsDNA spotted on a glass microarray and then visualized with fluorescent antibodies. The fluorescent intensity observed for each of the ~6,000 microarray spots (representing virtually all of the intergenic regions of *Saccharomyces cerevisiae*) was then normalized by the amount of DNA in each spot. After averaging the data over replicates and further processing, Mukherjee *et al.* reported log intensity ratios (LIRs) for $N = 5{,}812$ of these intergenic sequences.

Our goal was to find models of Abf1p–DNA binding that were likely to have produced these LIRs. In *Methods*, we derive an expression (Eq. **3**) for the EMA likelihood of different energy models for a given set of binding assay data. Because this equation applies to discretized data, we began our analysis by binning the N intergenic sequences $\{s_i\}_{i=1}^N$ according to their LIRs into "z-bins." We chose 20 sequences per bin, for a total of $m = 292$ bins. The bin size is, of course, arbitrary, but the results of our analysis were found to depend only weakly on this choice (see *SI Text* and SI Fig. 6). Each sequence s_i was thus assigned an integer z_i identifying the bin into which it was placed. This set of intergenic sequences $\{s_i\}$ and their corresponding bin numbers $\{z_i\}$ constituted the sole input to the rest of our analysis.

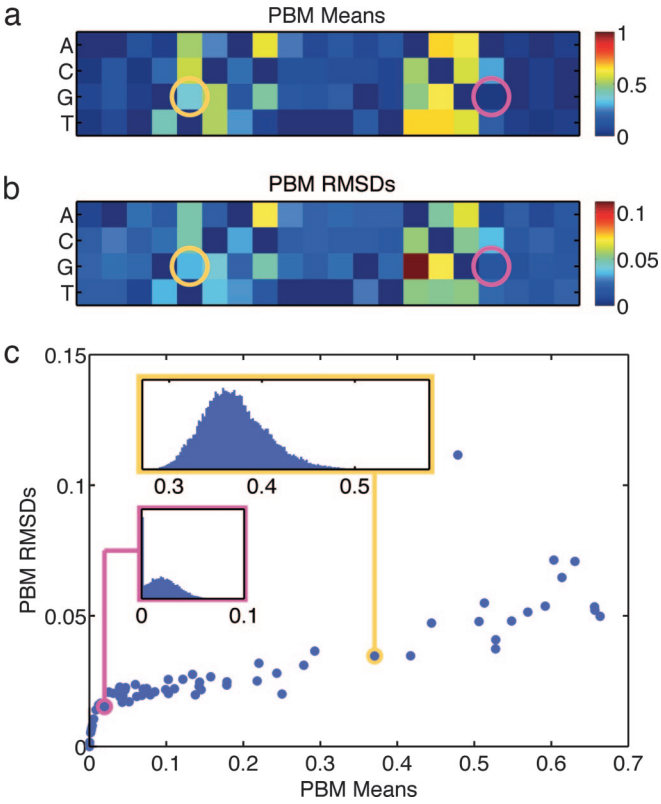

Fig. 1. PBM data determines Abf1p energy model parameters with surprising precision. (*a* and *b*) Mean and rmsd of energy matrix elements over the Θ_{PBM} ensemble of Abf1p models. All-blue columns contribute little to TF binding specificity. Overall, these matrices match the known Abf1p motif RTCRYNNNNNACG well. (*c*) Scatter plot of matrix element means versus rmsds: all rmsds are small compared with the binding cutoff of 1. (*Insets*) Marginal distributions of two representative matrix elements, circled in corresponding color in *a* and *b*.

The parametrized energy model we chose for Abf1p was a 4×20 "energy matrix" where each base in a site of length 20 contributes additively to the overall binding energy. We use this energy matrix to classify sites as "bound" (having a substantial TF occupancy in the experiment) if their energies lie below some threshold μ; otherwise they are classified as "unbound." The energy baseline was fixed by setting the lowest element in each column to zero, and the overall scale was fixed by setting $\mu = 1$. The elements of this matrix are the model parameters θ. See SI *Text* for more details.

A specific energy matrix classifies each region as bound if it contains at least one bound site, and otherwise classifies it as unbound. The numbers of bound and unbound regions in each z-bin suffice to calculate the posterior probability of the matrix $p(\theta|\{z_i\})$ using Eq. **3** of *Methods*. We performed multiple MCMC runs on the Abf1p data of Mukherjee *et al.* (4) to obtain an ensemble of 4×10^4 matrices (which we denote by Θ_{PBM}) sampled according to this distribution. Appropriate tests were used to verify MCMC convergence (see *SI Text* and Fig. 7).

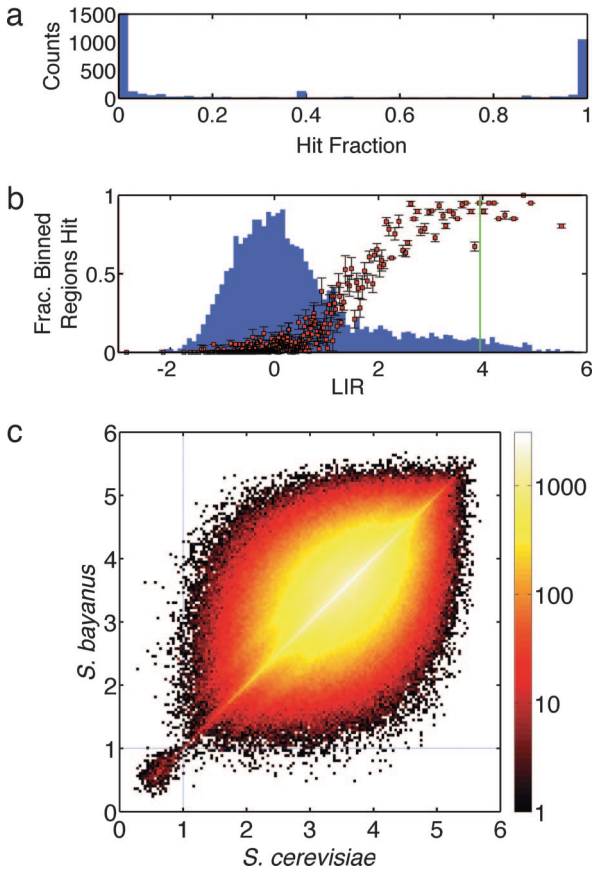

Fig. 2. Predictions derived from the ensemble of Abf1p energy models. (*a*) Histogram showing the Θ_{PBM} HF of all 20-bp sites in the intergenic DNA of *S. cerevisiae*. The distribution is strongly bimodal, efficiently separating bound sites (HF near 1) from unbound sites (HF near 0). The left-most bin contains the vast majority of sites and has been truncated for readability. (*b*) Mean fraction of regions in each *z*-bin declared bound by models in Θ_{PBM}, plotted against the mean LIR of those regions. Error bars show the rmsd variation in this fraction from model to model. The distribution of experimentally determined LIRs is shown in the background for reference. The green line is the threshold used by Mukherjee *et al.* (4) to identify bound regions. (*c*) 2D histogram of mean energies assigned by Θ_{PBM} to ungapped orthologous intergenic site pairs in *S. cerevisiae* and *S. bayanus*. Sites lying below the $E = 1$ binding cutoff (dashed lines) in one species are highly likely to fall below this cutoff in the other species.

Our results are generally in line with the qualitative motif RTCRYNNNNNACG known for Abf1p (12). However, inspection of Θ_{PBM} shows that the parameters of the Abf1p energy matrix are determined with remarkable precision. The mean and rmsd of each matrix element across all models in Θ_{PBM} are shown in Fig. 1 *a* and *b*. A scatter plot of the same data is given in Fig. 1*c* along with the marginal Θ_{PBM} distribution for two representative matrix elements. Matrix elements that differ significantly from zero generally have uncertainties much smaller than their means, and their distributions have a roughly Gaussian shape (Fig. 1*c Upper Inset*). Other elements are consistently assigned the lowest energy in their respective columns and are thus set by our normalization to be precisely zero in much of the ensemble (e.g. Fig. 1*c Lower Inset*). The matrix elements are determined with a degree of precision that makes it possible to see meaningful structure even in the center of the binding site, in positions that contribute little to overall TF specificity. That this precision is not an artifact of overfitting was verified by showing that parameter distributions derived from disjoint halves of the data provide consistent predictions for the energy matrix. We also verified that different choices for the matrix width led to consistent parameter distributions (see *SI Text* and SI Figs. 8 and 9).

The model ensemble Θ_{PBM} provides a direct way of predicting putative binding sites. For any 20-bp sequence of DNA, we can determine the fraction of models θ in Θ_{PBM} that "hit" that site (i.e., assign it an energy < 1). Fig. 2*a* histograms this "hit fraction" (HF) for every possible 20-bp site in the intergenic DNA of *S. cerevisiae* [as defined in Kellis *et al.* (12)]. The plot shows that the distribution of HFs is strongly bimodal (1,469 sites have HF $> 50\%$, whereas 1,182 of these also have HF $> 90\%$). We adopt HF $> 50\%$ as a plausible criterion for predicting a site to be bound.

Fig. 2*b* plots the mean fraction of sequences in each *z*-bin declared to be bound by models in Θ_{PBM} against the mean LIR of those sequences. Error bars show the variation in these predictions across different models in Θ_{PBM}. A histogram of the actual measured LIRs is shown in the background. The most striking feature of this plot is the sigmoidal relationship between model predictions and measured LIRs, showing a rapid transition from mostly not bound to mostly bound sequences at the beginning of the heavy tail of the LIR distribution. Whereas its general shape is exactly as expected, this outcome is in no way predetermined: Eq. **3** places no *a priori* weight on models that make similar predictions for sequences in neighboring *z*-bins or that declare sequences with large LIR to be bound. The consistency between the shape of this scatter plot and our physical expectation is an independent confirmation of the validity of EMA likelihood.

The Abf1p models in Θ_{PBM} appear to account for the data to within Mukherjee *et al.*'s (4) estimate of the experimental error. The green line in Fig. 2*b* indicates the LIR cut used by Mukherjee *et al.* to define bound regions. Of the 186 regions passing this cut, the experimenters estimated that 7–9% were false positives. We find that 167 (89.8%) of these regions have HF $> 90\%$, 18 (9.7%) have HF $< 10\%$, and only 1 region has an intermediate HF. So, although energy matrix models are simplistic, they appear to account for this Abf1p PBM data about as well as one could hope for any model, regardless of sophistication.

Mukherjee *et al.* (4) were obliged to adopt a stringent threshold to minimize false positives, in the process rejecting an unknown number of bound sequences. Our model-based approach, by contrast, can tease out regions that are likely to be bound, regardless of where they lie in the raw data distribution. In this case, we find 840 regions with HF $> 50\%$ (and therefore bound by our criterion) and with LIR below the cut chosen by Mukherjee *et al*. (4). In short, we find many more bound regions lying below the experimenters' threshold than lying above it.

This is a strong statement and one for which one would like independent confirmation. The best evidence would come from the direct *in vitro* measurement of large numbers of putative site binding energies. Energy measurements for Abf1p have been carried out for a small number of sites (13), but the results agree neither with our predictions nor with the analysis of Mukherjee *et al.* (4) or Lee *et al.* (3). It is possible that these measurements are in error, and we believe further *in vitro* studies are necessary to resolve this issue. Recently developed high-throughput techniques for direct measurement of binding affinities give promise of providing data of the quality and scope needed to test these models in quantitative detail (S. Quake, personal communication).

Although direct energy data are lacking, phylogenetic analysis provides alternative evidence in support of our binding energy predictions. Using the intergenic alignments of Kellis *et al.* (12), we identified all pairs of ungapped orthologous intergenic 20-bp sequences in *S. cerevisiae* and *Saccharomyces bayanus*. We then computed the mean predicted binding energy (using Θ_{PBM}) of each site in *S. cerevisiae*, as well as that of its ortholog. A scatter plot of the resulting energies (Fig. 2*c*) reveals a large and well separated population of sites whose putative energies lie below

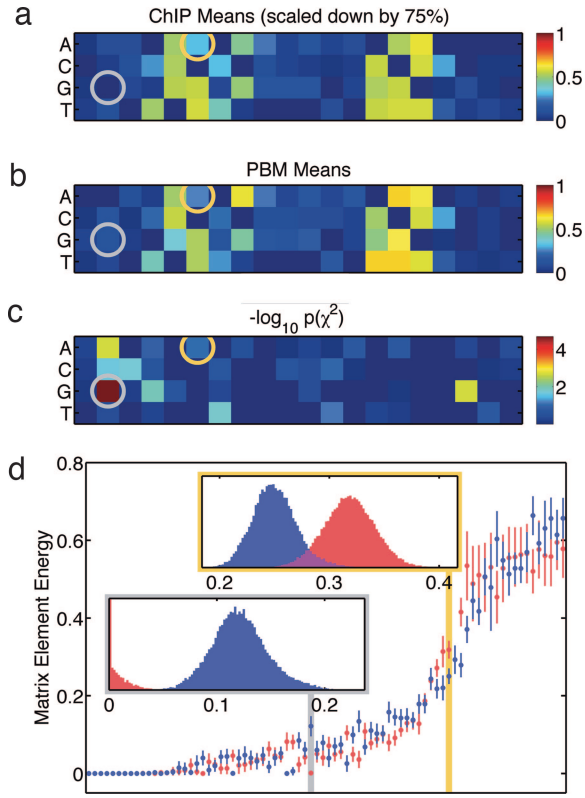

Fig. 3. Comparison of Θ_{PBM} and Θ_{ChIP} parameter distributions. (*a*) Mean values of rescaled matrix elements in Θ_{ChIP}. (*b*) Mean matrix elements in Θ_{PBM} (same as Fig. 1*a*). (*c*) χ^2 *P* values quantifying the element-by-element consistency of the Θ_{ChIP} and Θ_{PBM} distributions. (*d*) Mean and rmsd uncertainty of each matrix element according to the Θ_{PBM} (blue) and rescaled Θ_{ChIP} (red) distributions. Elements are arranged from left to right in order of increasing mean. (*Insets*) Raw MCMC histograms show the values obtained for the matrix elements circled in *a*–*c* and highlighted below each *Inset* in *d*. The Θ_{ChIP} matrix element distribution (*Lower Inset*) has most of its weight at precisely 0; the corresponding histogram has been truncated at this bin.

1 (i.e., are predicted to be bound) in both genomes: of the 676 *S. cerevisiae* sites with energy < 1, a total of 501 (74%) have an *S. bayanus* ortholog whose energy also lies < 1. By contrast, sites with energy > 1 in either genome tend to have orthologs with highly randomized energy. In short, the large majority of alignable sites predicted to be bound by our models have strongly conserved putative energies. The fact that our models are found directly from *in vitro* binding data provides a compelling case that they also describe the free energy of Abf1p–DNA binding *in vivo*, and that this binding energy plays a major role in determining which sites have biological function.

Next, we performed a similar analysis of Lee *et al.*'s ChIP–chip data (3) to determine whether it gives a description of Abf1p consistent with that obtained from the PBM data. In these ChIP–chip experiments, TFs were cross-linked *in vivo* to their binding sites, after which TF-bound fragments of DNA were isolated, amplified, labeled, and hybridized in competition with reference DNA to a ssDNA microarray of yeast intergenic regions. The enrichment observed in each microarray spot was characterized by an "*X*-statistic" based on the single array error model of Hughes *et al.* (14). Assuming a Gaussian distribution for these *X*-statistics in the absence of TF binding, Lee *et al.* reported an enrichment *P* value for each region.

We assigned these probed sequences to *z*-bins according to their *P* values (equivalently, according to their *X*-statistics). MCMC analysis then gave an ensemble Θ_{ChIP} of 4×10^4 matrix

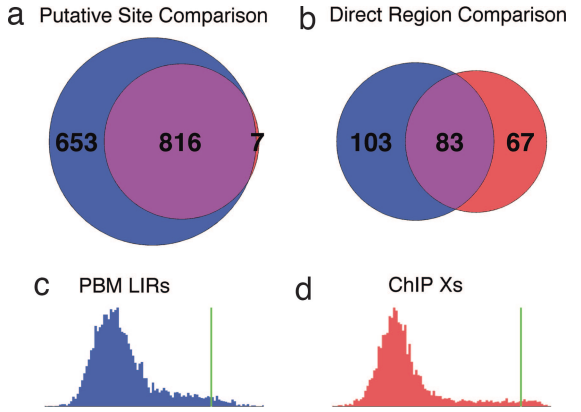

Fig. 4. EMA likelihood analysis leads to compatible binding site predictions from different experimental data sets, whereas the more standard method of thresholding the experimental signal leads to substantial disagreement. (*a*) The 20-bp intergenic *sites* in *S. cerevisiae* having Θ_{ChIP} HF $> 50\%$ (red) are a nearly perfect subset of those with Θ_{PBM} HF $> 50\%$ (blue). (*b*) In contrast, the intergenic regions selected by Mukherjee *et al.*'s (4) LIR threshold on PBM data (blue) overlap poorly with those selected by Lee *et al.*'s (3) (*P* value threshold on ChIP–chip data (red). (*c* and *d*) The thresholds chosen by the experimenters are indicated by the green lines on the experimental LIR histogram of PBM data in *c* and on the *X*-statistic histogram of ChIP–chip data in *d*.

models. Results of a single-ensemble analysis of Θ_{ChIP} were similar to those of Θ_{PBM} (see *SI Text* and SI Figs. 10–14). Note that, by integrating over all possible error models, we avoided having to model the noise contributions from each individual step in the ChIP–chip protocol. We also avoided having to estimate *in vivo* contributions to the experimental noise, such as the fraction of valid binding sites likely to be obscured by chromatin.

Although the ChIP–chip and PBM results are quite similar, the energy matrix elements derived from Θ_{ChIP} are systematically larger than those of Θ_{PBM}. Our procedure, though, produces energy matrices artificially scaled so that the energy cutoff is equal to 1. Thus, when comparing ensembles, we are free to rescale the matrices in one ensemble so as to bring them into accord with those of the other. The resulting difference between the rescaled cutoffs has a natural interpretation: binding site occupancy, which we approximate by a step function at the energy cutoff, should vary with TF concentration and may well differ between experiments; the energy matrix itself, on the other hand, reflects an intrinsic property of the TF molecule that should, in principle, not vary between experiments (if other factors, such as ion concentration and pH, are kept at similar levels). In the case at hand, we found that rescaling the Θ_{ChIP} energy cutoff to 0.75, while keeping the Θ_{PBM} cutoff at 1, brought the Θ_{ChIP} energy matrix elements into close agreement with those of Θ_{PBM}. Fig. 3 *a* and *b* illustrates this close agreement between the mean matrix elements in the two ensembles. Fig. 3*d* provides a direct comparison of the Θ_{ChIP} and Θ_{PBM} distributions for each matrix element. In most cases, values that could plausibly have been drawn from either the Θ_{ChIP} or Θ_{PBM} distribution can be identified (illustrated in Fig. 3*d Upper Inset*, which shows the raw Θ_{ChIP} and Θ_{PBM} histograms for the orange-circled matrix element in Fig. 3 *a* and *b*). Although the two histograms are not identical, they overlap enough that a matrix element value consistent with both distributions can be found.

In *SI Text* we argue that a simple χ^2 test provides a valid way of quantifying such consistency. The resulting χ^2 *P* values for each matrix element are shown in Fig. 3*c*, with lower *P* values corresponding to poorer consistency between ensemble distributions. There are a few matrix elements for which Θ_{ChIP} and Θ_{PBM} give inconsistent distributions by this test (the red and

yellow squares in Fig. 3*c*). The raw histograms for the most inconsistent matrix element (the red square in Fig. 3*c*), plotted in Fig. 3*d Lower Inset*, display the poor overlap between the two distributions. Although the inconsistency is significant (even accounting for multiple hypotheses), it occurs outside of the binding site proper and may not have much practical impact. One way to assess the impact of such discrepancies is to compare the binding site predictions of Θ_{ChIP} and Θ_{PBM}. If the matrices in both ensembles were identical, the sites selected by the smaller (ChIP–chip) cutoff would be a subset of the sites selected by the larger (PBM) cutoff. Fig. 4*a* shows that this is in fact nearly the case: of the 823 Θ_{ChIP}-predicted sites, all but 7 are predicted by Θ_{PBM}. We stress that these two sets of predictions were made by using data from different experimental platforms and performing separate analyses involving no free parameters.

Our consistency analysis contrasts with that of Mukherjee *et al.* (4), who compared their putatively bound regions (identified by a stringent LIR threshold; see Fig. 4*c*) with those of Lee *et al.* (3) (identified by a stringent *P* value threshold; see Fig. 4*d*) and obtained the Venn diagram reproduced in Fig. 4*b*. Although the overlap between the two sets of regions is certainly significant, neither is even approximately a subset of the other. This problem cannot be avoided by choosing lower thresholds, as that would result in more false-positive predictions. Even though both data sets are of high quality, this discrepancy is an inevitable result of experimental noise, *in vivo* effects on binding, etc. However, by "bottlenecking" PBM and ChIP–chip data through a simple parametric model, one obtains almost entirely consistent pictures of Abf1p specificity.

Discussion

A quantitative understanding of biological systems will require the ability to deduce quantitative models from data in a transparent and principled way. Not only must optimal model parameters be found, but the confidence one has in the values of these parameters must be characterized and parameters determined from different data sets must be compared for consistency. For these and other reasons, we believe likelihood inference using MCMC, an approach that has had enormous success in physics, should become an important tool in 21st-century biology.

Modern biology, however, faces challenges very different from those of physics. Although physicists often take great care to understand and minimize experimental noise, high-throughput assays have sources of noise that may vary from system to system or even from laboratory to laboratory and may be impractical to characterize quantitatively. The central point of this article is that, given enough data, one can precisely characterize interesting biological phenomena without having to model uninteresting effects, such as the experimental errors. We believe this is of particular importance for the study of TFs, although it may find application in other areas of quantitative biology as well.

Furthermore, the high-throughput platforms currently available for probing TF binding appear, at least for broad-acting TFs, to provide sufficient data for such analyses. In our study of Mukherjee *et al.*'s (4) PBM data for Abf1p, we were essentially able to infer the values of 352 independent parameters describing both the TF and the experimental error model (although the error model parameters were integrated out analytically). Nonetheless, most of the 80 energy matrix elements describing the TF were determined to a precision of $\leq 5\%$ of the functional range. This precision is not caused by overfitting; it simply reflects the large amount of information contained in the data.

Our analysis is meant only as a proof-of-principle demonstration and could be extended in many ways. Perhaps our most unrealistic assumption is that TF binding sites are either bound or not bound. This was done primarily to keep the analysis simple and speed up the MCMC algorithm, but a more physically realistic computation of region occupancy, such as that used by the program GOMER (15), could be used instead. The assumption that each nucleotide contributes independently to the TF's SDBE (see ref. 16 for a critique) could also be relaxed by allowing couplings between positions. Indeed, a great advantage of likelihood inference is the way it accommodates such refinements without changing the conceptual basis of the analysis. The only obstacle one faces when implementing such changes is the need for computational power.

Another important aspect of likelihood inference is the possibility of refining quantitative models by using data from multiple experiments. For example, if two data sets $\{z_i\}$ and $\{z'_j\}$ are available for some TF, one can perform a likelihood analysis by using the product of each data set's likelihood as the combined likelihood of both data sets, i.e., $p(\{z_i\},\{z'_j\}|\theta) = p(\{z_i\}|\theta)p(\{z'_j\}|\theta)$. Extending this to more than two data sets is straightforward, and data from different experimental platforms may be combined in this way. For example, ChIP–chip data might be combined with low-throughput EMSA measurements and high-throughput *in vitro* data by using synthetic DNA probes (17–19). Also, any knowledge one has about the experimental errors in any of these experiments may be used, either by assuming an explicit error model or biasing the error model prior (see *SI Text*). The combination of data from different experiments is a powerful analysis technique, and our method should help facilitate its application to high-throughput biological data.

Methods

Software and Results. The Θ_{PBM} and Θ_{ChIP} matrix ensembles are available on request. The software used in this analysis, including our MCMC algorithm, was written in Matlab and C++ and is available on request.

Likelihood With Unknown Error Models. Suppose an experiment assigns a value z_i to each DNA region s_i ($i = 1, \ldots, N$) and let $x_i = \theta(s_i)$ be the corresponding prediction made by the TF binding model for a particular choice of parameters θ. The z_i and x_i can be quite different quantities: a model might predict whether or not a TF is likely to be bound to a particular DNA sequence in thermal equilibrium, while an experiment might observe the fluorescence intensity of the corresponding spot on a microarray. Because of experimental noise, the two quantities are related by an error model $E(z|x)$, giving the probability of observing z when the state predicted by the model is really x. We make the simplifying assumption that the error model is the same for all regions probed in any given experiment, although different experiments may be subject to different error models.

We want to find specific parameters θ such that the N model predictions $\{x_i\}_{i=1}^N$ account well for the measurements $\{z_i\}$, i.e., give a large value to the likelihood $p(\{z_i\}|\theta)$. Our analysis relies on three further assumptions: (*i*) There is a "correct" set of parameters θ that accurately describes the TF's behavior in the experiment. (*ii*) The likelihood of the data $\{z_i\}$ depends on the model parameters θ only through the model predictions $\{x_i\}$. Thus, $p(\{z_i\}|\theta) = p(\{z_i\}|\{x_i\})$. (*iii*) The experimental results for each sequence are independent, so that $p(\{z_i\}|\{x_i\}) = \Pi_i\, p(z_i|x_i)$.

Our method is most simply implemented if both the observations and the predictions are discrete. Accordingly, we group the $N \sim 6{,}000$ DNA sequences in a binding assay into $\sim$100–300 equipopulated z-bins on the basis of observed fluorescences; at the same time the model predictions assign each sequence to an x-bin (bound or not bound). If we know the error model, we can then write an explicit expression for the likelihood:

$$p(\{z_i\}|\theta) = \prod_{i=1}^{N} E(z_i|x_i) = \prod_{z,x} E(z|x)^{c_{zx}}, \qquad [2]$$

where c_{zx} is the number of regions assigned simultaneously to bin z and bin x. This expression depends on the parameters θ only through the way regions are assigned to x-bins. Because error models applicable to high-throughput biological experiments are usually unknown, in practice we cannot evaluate Eq. **2**. We propose to deal with this problem by averaging over the space of all error models with some reasonable prior $p(E)$. The explicit expression we obtain for the EMA likelihood under a uniform prior on the error models is (see *SI Text* for more details):

$$p(\{z_i\}|\theta) = \int dE\, p(E) \prod_{z,x} E(z|x)^{c_{zx}} = \frac{(m-1)!^n \prod_{z,x} c_{zx}!}{\prod_x \left(m - 1 + \sum_z c_{zx}\right)!}, \quad [3]$$

where m and n, respectively, denote the number of possible values that z and x can take on. We stress that this equation for EMA likelihood is completely general and may be applied to any situation in which both the data and the model predictions are quantized. Note that the specific numerical values of the experimental data and model predictions serve only to cluster sequences together and are otherwise unused in this analysis. Thus, monotonic reparametrizations of these values do not affect our results.

It is interesting to note that Eq. **3** may be rewritten (see *SI Text*) as:

$$\ln p(\{z_i\}|\theta) = N[I(z;\, x) - H(z) - \Delta], \quad [4]$$

where $I(z;x)$ is the empirical mutual information (20) between $\{z_i\}$ and $\{x_i\}$ (and thus depends on θ), $H(z)$ is the empirical entropy of $\{z_i\}$ (and does not depend on θ), and Δ is a nonnegative correction, accounting for finite data and the choice of error model prior $p(E)$. For a large class of error model priors, including the uniform prior used to compute Eq. **3**, Δ vanishes as N becomes large. In this limit, the per-datum log likelihood becomes, up to an additive constant, the mutual information between model predictions and data. The emergence of mutual information helps explain how one can meaningfully evaluate likelihood without prior knowledge of which model predictions should correspond to which data: the best model will make predictions that provide the most information about the experimental results, regardless of how this correspondence is realized.

Sampling Model Space. The primary goal of our analysis is to characterize the distribution of model parameters θ specified by the data $\{z_i\}$ by using the posterior distribution in Eq. **1** with likelihood specified by Eq. **3**. To do this we used MCMC, a powerful computational method for sampling from such distributions (21). An essential feature of MCMC is that it does not require that one know how to normalize the distribution being sampled, which frees us from having to estimate the proportionality constant in Eq. **1**.

MCMC starts from a seed model θ_0 and stochastically wanders from model to model in such a way that the ensemble $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ of models visited is eventually distributed according to a desired distribution (in our case Eq. **1**). The expected value of any θ-dependent quantity $q(\theta)$, given the observations $\{z_i\}$, is then estimated by the ensemble average:

$$\langle q \rangle \equiv \int d\theta\, q(\theta) p(\theta|\{z_i\}) \approx \frac{1}{T} \sum_{t=1}^{T} q(\theta_t), \quad [5]$$

where $q(\theta)$ might be the value of a particular matrix element, the square deviation of that element from its mean, or a binary variable describing whether or not a particular site is bound. See *SI Text* for a detailed discussion of the particular MCMC algorithm used in this analysis.

We thank Manuel Llinás and William Bialek for generous advice and constructive criticism; William Press, Steven Quake, and Jason Lieb for valuable criticism of this manuscript; Michael Buck, Casey Bergman, Neil Clarke, Thomas Gregor, Michael Lässig, Saeed Tavazoie, Martin Vingron, and the participants of the 2006 Otto Warburg Summer School for helpful discussions; and Martin Jones for helping produce Fig. 2*c*. C.G.C. and J.B.K. were supported by National Institutes of Health Grant P50GM071508. C.G.C. was also supported by Department of Energy Grant DE-FG02-91ER40671. G.T. was supported by the Burroughs-Wellcome Fund.

1. Ptashne M, Gann A (2002) *Genes and Signals* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY).
2. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, *et al.* (2000) *Science* 290:2306–2309.
3. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.* (2002) *Science* 298:799–804.
4. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML (2004) *Nat Genet* 36:1331–1339.
5. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, *et al.* (2005) *Nat Biotechnol* 23:137–144.
6. Berg OG, von Hippel PH (1987) *J Mol Biol* 193:723–750.
7. Stormo GD, Fields DS (1998) *Trends Biochem Sci* 23:109–113.
8. Stormo GD (2000) *Bioinformatics* 16:16–23.
9. Mustonen V, Lässig M (2005) *Proc Natl Acad Sci USA* 102:15936–15941.
10. Berg J, Willmann S, Lässig M (2004) *BMC Evol Biol* 4:42.
11. Foat BC, Morozov AV, Bussemaker HJ (2006) *Bioinformatics* 22:141–149.
12. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) *Nature* 423:241–254.
13. Beinoravčiūtė-Kellner R, Lipps G, Krauss G (2005) *FEBS Lett* 579:4535–4540.
14. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.* (2000) *Cell* 102:109–126.
15. Granek JA, Clarke ND (2005) *Genome Res* 6:R87.
16. Benos PV, Bulyk ML, Stormo GD (2002) *Nucleic Acids Res* 30:4442–4451.
17. Bulyk ML, Gentalen E, Lockhart DJ, Church GM (1999) *Nat Biotechnol* 17:573–577.
18. Bulyk ML, Huang X, Choo Y, Church GM (2001) *Proc Natl Acad Sci USA* 98:7158–7163.
19. Warren CL, Kratochvil NC, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips GN, Jr, Ansari AZ (2006) *Proc Natl Acad Sci USA* 103:867–872.
20. Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York), pp 12–49.
21. Gilks WR, Richardson S, Spiegelhalter DJ (1996) *Markov Chain Monte Carlo in Practice* (Chapman & Hall, New York), pp 1–16.