Precise Physical Models of Protein-DNA Interaction from High-Throughput Data: Supplementary Material

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

Failure of Nucleotide Independence in *P. falciparum*

Table 1 shows typical intergenic DNA from *P. falciparum*. Low complexity DNA, such as long stretches of 'TTTTT...' and 'ATATAT...', occur with vastly more frequency than permitted by the independent nucleotide model required for the Berg von Hippel equivalence.

Non-Gaussian Intensity Ratios

Fig. 5 shows a histogram of PBM LIRs, along with a histogram of corresponding intensity ratios (i.e. LIRs exponentiated using base 2). One expects the bulk of these distributions to be caused by experimental noise, and while the bulk of the LIR distribution (Fig. 5*a*) appears surprisingly Gaussian (the physical reasons for this are unclear), the intensity ratio distribution (Fig. 5*b*) clearly has a very fat tail on the right. This calls into question the assumption of Gaussian noise in intensity ratios, implicitly made by Foat *et al.* (7) when fitting predicted intensity ratios to data using least squares. This is of concern because underlying distributions with heavy tails can cause outliers to dominate such χ^2 fits.

Energy Matrix Models

In our analysis we assume the simplest possible model for SDBE, one where each base within a contiguous TF binding site of length L contributes independently to the overall Gibbs free energy. While simplistic, this assumption has proven surprisingly accurate in a number of cases (1-3) (see ref. 4 for a situation in which couplings might be important). This SDBE is naturally represented by a $4 \times L$ "energy matrix" whose elements $\{M_{bl}\}$ (where $l \in \{1, \ldots, L\}$ and $b \in \{A, C, G, T\}$) constitute the model parameters θ. The energy assigned to a site $b_1 b_2 \cdots b_L$ is given by $\sum_l M_{b_l l}$, with lower energy corresponding to stronger binding. Because only energy differences matter, we generally fix the smallest matrix element $\min_b M_{bl}$ in each column l to zero.

Because such a TF model assigns energies to each of the $\sim 10^3$ possible overlapping sites in each intergenic region s_i, we need some way of collapsing all these energies into a discrete set of values $\{x_i\}$, which, if the TF model is accurate, will largely account for our experimental observations. For simplicity, we declare a sequence s_i to be bound ($x_i = 1$) if it contains at least one site with energy below some binding threshold and otherwise declare it to be not bound ($x_i = 0$). In our analysis we set this threshold to unity, thus fixing the overall scale of the energy matrix elements. While this hard energy cutoff is only a crude approximation to the more complicated physics of binding site occupancy, we find that it can account for real data surprisingly well. It also greatly speeds the necessary computations.

EMA Likelihood Calculation

For brevity, denote $E_{zx} \equiv E(z|x)$. Assuming each probability E_{zx} is restricted to the interval $[0, 1]$, the only *a priori* constraint on $\{E_{zx}\}$ is that, for every x, $\sum_z E_{zx} = 1$. Letting E denote the set of variables $\{E_{zx}\}$ which constitute the error model, we define the uniform prior on the space of error models as the maximum entropy distribution $p(E)$ subject only to the constraint that $p(E) = 0$ whenever, for some x, $\sum_z E_{zx} \neq 1$. Explicitly, this uniform prior is

$$p_U(E) \equiv \Gamma(m)^n \prod_x \delta(1 - \sum_z E_{zx}), \qquad \textbf{[1]}$$

where $\delta(\cdot)$ is the Dirac delta function, m and n are the number of possible values for z and x, respectively, and $\Gamma(m)^n = (m-1)!^n$ provides the necessary normalization for this distribution.

Now suppose we have data $\{z_i\}$, as well as a specific model which makes corresponding predictions $\{x_i\}$. Let c_{zx} denote the number of data points i for which $z_i = z$ and $x_i = x$. Also define $c_x \equiv \sum_z c_{zx}$. Using the uniform error model prior $p_U(E)$, the EMA likelihood (i.e. the likelihood averaged over all error models according to p_U) can be calculated analytically:

$$p_U(\{z_i\}|\{x_i\}) = \int dE \; p_U(E) \prod_{z,x} (E_{zx})^{c_{zx}} = \frac{\Gamma(m)^n \prod_{z,x} \Gamma(c_{zx} + 1)}{\prod_x \Gamma(c_x + m)}. \qquad \textbf{[2]}$$

This is the result quoted in Eq. **3** in the main text, which uses factorials in place of gamma functions, but is otherwise identical. As with most likelihood calculations, this depends on model parameters only through the specific predictions $\{x_i\}$ of the model. However, since all possible error models are considered with equal weight, no prior assumptions are made about which measurements z should result from which predictions x.

One can also bias the error model prior toward a specific error model E^*_{zx} through a simple generalization. Using some set of nonnegative weights $\{W_x\}$, we define the Dirichlet

error model prior

$$p_D(E) \equiv \mathcal{N} \prod_x \delta(1 - \sum_z E_{zx}) \prod_z (E_{zx})^{W_x E^*_{zx}}, \qquad \textbf{[3]}$$

which requires the normalization constant

$$\mathcal{N} = \frac{\prod_x \Gamma(W_x E^*_x + m)}{\prod_{z,x} \Gamma(W_x E^*_{zx} + 1)} \ . \qquad \textbf{[4]}$$

Given the counts $\{c_{zx}\}$, we can again compute the EMA likelihood explicitly using this Dirichlet piror:

$$p_D(\{z_i\}|\{x_i\}) = \int dE \ p_D(E) \prod_{z,x} (E_{zx})^{c_{zx}} = \mathcal{N} \frac{\prod_{z,x} \Gamma(c_{zx} + W_x E^*_{zx} + 1)}{\prod_x \Gamma(c_x + W_x E^*_x + m)} \ . \qquad \textbf{[5]}$$

Up to a multiplicative constant, this is exactly what one gets by computing the EMA likelihood with the uniform error model prior after "spiking" each count c_{zx} with $W_x E^*_{zx}$ additional data points. Indeed, when $W_x = 0$ for all x, one recovers the uniform EMA likelihood. In the other limit, when all $W_x \to \infty$, one obtains

$$p_D(\{z_i\}|\{x_i\}) \longrightarrow \prod_{z,x} (E^*_{zx})^{c_{zx}} \ , \qquad \textbf{[6]}$$

which is the likelihood one gets by using the single error model $E(z|x) = E^*_{zx}$ without any averaging. Dirichlet error model priors thus provide a convenient way of interpolating between an analysis using a single, well defined error model and an analysis, which is completely agnostic about error models. As shown in the main text, however, the agnostic approach can lead to a very informative analysis when copious amounts of data are available.

EMA Likelihood and Mutual Information

For any error model prior $p(E)$, the EMA likelihood can be parsed to reveal a striking connection with mutual information. Our considerations are based on the identity (equivalent to Eq. **4** in the main text),

$$p(\{z_i\}|\{x_i\}) = \int dE \ p(E) \prod_{z,x} (E_{zx})^{c_{zx}} = \exp N[I(z;x) - H(z) - \Delta], \qquad \textbf{[7]}$$

where $I(z;x)$ is the empirical mutual information between the N observations $\{z_i\}$ and the N model predictions $\{x_i\}$, $H(z)$ is the empirical entropy of the observations, and Δ is a correction factor that vanishes as $N \to \infty$ under very general considerations.

To show this explicitly, define the joint distribution $f(z,x) \equiv c_{zx}/N$ of z and x, along with the marginal distributions $f(z) \equiv \sum_x f(z,x)$, $f(x) \equiv \sum_z f(z,x)$ and the conditional distribution $f(z|x) \equiv f(z,x)/f(x)$. By the standard definitions,

$$I(z;x) = \sum_{z,x} f(z,x) \ln \frac{f(z,x)}{f(z)f(x)} \qquad \textbf{[8]}$$

and

$$H(z) = -\sum_{z} f(z) \ln f(z). \qquad \textbf{[9]}$$

Substituting these definitions into the right hand site of Eq. **7**, one finds

$$\Delta = -\frac{1}{N} \ln \int dE\ p(E) e^{-ND(f||E)} \quad \text{with} \quad D(f||E) \equiv \sum_{z,x} f(z,x) \ln \frac{f(z|x)}{E(z|x)}\ . \qquad \textbf{[10]}$$

The correction Δ therefore captures all information about our choice of error model prior and the fact that we have only finite data. Note: $D(f||E)$ is the Kullback-Leibler divergence between the empirical distribution $f(z|x)$ and the error model $E(z|x)$. Since the KL divergence is always non-negative, so is Δ.

In the case of the uniform error model prior p_U, Δ becomes

$$\Delta_U = \frac{1}{N} \ln \left[\frac{\prod_x (m-1+c_x)! c_x^{-c_x}}{(m-1)!^n \prod_{z,x} c_{zx}! c_{zx}^{-c_{zx}}} \right] = \frac{1}{N} \ln \prod_x \prod_{i=1}^{m-1} (i+c_x) + \frac{\ln AB}{N} \qquad \textbf{[11]}$$

where $A = (m-1)!^{-n} \leq 1$ and

$$B = \left[\frac{\prod_x c_x!}{\prod_{z,x} c_{zx}!} \right] \prod_{z,x} \left(\frac{c_{zx}}{c_x} \right)^{c_{zx}} \qquad \textbf{[12]}$$

is a term in the binomial expansion of

$$1 = \prod_x \left(\sum_z \frac{c_{zx}}{c_x} \right)^{c_x}, \qquad \textbf{[13]}$$

and so is ≤ 1 as well. The last term in Eq. **11** is therefore negative, giving

$$\Delta_U \leq \frac{1}{N} \sum_{i=1}^{m-1} \sum_x \ln(i+c_x). \qquad \textbf{[14]}$$

Concavity of the logarithm implies

$$\sum_{i=1}^{m-1} \sum_x \ln(i+c_x) \leq (m-1)n \ln \left(\frac{m}{2} + \frac{N}{n} \right), \qquad \textbf{[15]}$$

and thus,

$$0 \leq \Delta_U \leq \frac{(m-1)n}{N} \ln\left(\frac{m}{2} + \frac{N}{n}\right). \quad \textbf{[16]}$$

These are loose bounds, but they show explicitly that $\Delta_U \to 0$ as $N \to \infty$. Similarly, $\Delta_D \to 0$ as $N \to \infty$ for the Dirichlet prior in Eq. **3**, and generally it appears that $\Delta \to 0$ as $N \to \infty$ if $p(E)$ is continuous and nonzero almost everywhere on the space of properly normalized error models (though we do not provide a proof here).

The MCMC Algorithm

To implement MCMC, one must first define the prior distribution $p(\theta)$ one places on the space of models Ω. This choice of prior is largely arbitrary, so to speed up computations we choose to define $p(\theta)$ in terms of the projection of a uniform prior on the space of models $\tilde{\Omega}$ where $\tilde{\theta} \in \tilde{\Omega}$ consists of an energy matrix whose elements $\{\tilde{M}_{bl}\}$ can take on values anywhere from 0 to 1 independent of each other, and an energy cutoff $\mu \in [0, L]$. Any such $\tilde{\theta} = (\{\tilde{M}_{bl}\}, \mu)$ that makes nontrivial predictions can be projected to a normalized model $\theta = \{M_{bl}\} \in \Omega$, having an energy cutoff of 1 and matrix elements satisfying $\min_b M_{bl} = 0$ for all l, without changing the predictions $\{x_i\}$. This is done by performing the following shifts and rescalings on energy matrix elements (which have no effect on whether any given sequence is or is not below the energy cutoff):

$$M_{bl} = \frac{\tilde{M}_{bl} - \min_c \tilde{M}_{cl}}{\mu - \sum_{l'} \min_c \tilde{M}_{cl'}}. \quad \textbf{[17]}$$

The advantage of doing MCMC on $\tilde{\Omega}$ instead of Ω is the added ease with which the algorithm can explore parameter space. On the other hand, it is convenient to work in Ω when analyzing MCMC results because all the "gauge freedoms" have been removed, forcing models which make similar predictions to have similar parameters.

MCMC starts from a seed model $\tilde{\theta}_0 \in \tilde{\Omega}$, then wanders from model to model so that the resulting chain of models encountered along the way, $\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_T$, is distributed according to the posterior distribution which, because we use a uniform prior on $\tilde{\Omega}$, is essentially the EMA likelihood $p(\{z_i\}|\tilde{\theta})$. This is done as follows. In each step t of the Markov chain, the parameters of the current model $\tilde{\theta}_{t-1}$ are perturbed slightly to give a new model $\tilde{\theta}'_{t-1}$. The perturbations we allow are

- Adding a small normally distributed number to one of the matrix elements $\tilde{M}_{bl}$.
- Adding a small normally distributed number to the energy cutoff μ.
- Adding the same small normally distributed number to the energy cutoff μ and all energy matrix elements $\tilde{M}_{bl}$ in some specific column l.

Given the new parameters $\tilde{\theta}'_{t-1}$, we compute the new posterior probability $p(\{z_i\}|\tilde{\theta}'_{t-1})$. If all parameters remain within their allowed range, we set $\tilde{\theta}_t = \tilde{\theta}'_{t-1}$ (i.e. accept the new parameters) with probability

$$\min\left(1, \frac{p(\{z_i\}|\tilde{\theta}'_{t-1})}{p(\{z_i\}|\tilde{\theta}_{t-1})}\right). \quad \textbf{[18]}$$

Otherwise we set $\tilde{\theta}_t = \tilde{\theta}_{t-1}$ (i.e. reject the new parameters) and try again. These computations are sped up greatly by restricting, for many consecutive steps, matrix perturbations to elements in two specific columns, then choosing two different columns to focus on, etc. In this way we build a large sample of properly distributed models which are projected onto Ω for analysis.

MCMC Sampling and Convergence

In the main text we discuss the results of MCMC sampling on Mukherjee *et al.*'s Abf1p PBM data (5) (with 20 intergenic sequences per z-bin) and Lee *et al.*'s Abf1p ChIP-chip data (6) (with 50 intergenic sequences per z-bin). For each of these data sets, 10 MCMC runs were started with a seed model derived from the known qualitative Abf1p recognition motif NNNRTCAYTNNNNACGWNNN by assigning an energy of 0 to allowed nucleotides and 1 to disallowed nucleotides. We then let each run go for 5×10^6 steps, and every 1,000th model visited by MCMC in each run was recorded. The last 4,000 models recorded for each of the 10 runs were then concatenated to give a representative ensemble Θ_{PBM} (or Θ_{ChIP}) of 4×10^4 models.

MCMC was seeded with the known Abf1p motif only to reduce the amount of time it took for the algorithm to "burn-in" to the true posterior distribution. Other than the particular alignment of matrix elements within the 20-bp window, this choice of starting point does not affect the distribution of models found after burn-in has occurred. In particular, we have been able to achieve burn-in *de novo* through simulated annealing, starting from seed models with randomly chosen parameters (data not shown).

Evidence for convergence of our MCMC routines is presented in Fig. 7*a* and 7*c* where we plot, for PBM and ChIP-chip data respectively, the mean intrarun variance (the mean over the 10 runs of the variance within each MCMC run) against the interrun variance (the variance across models in the 10 MCMC runs concatenated together) for the 80 energy matrix elements. The scatter plot of these two numbers should lie along the diagonal if the different MCMC runs thoroughly sampled the same distribution, and should lie above the diagonal if the MCMC runs did not have time to converge. This computation was done for different numbers of models (20, 100, 1,000, 5,000) taken from the beginning of each run and the results for each of these sample sizes (distinguished by color) are shown. The scatter plot collapses convincingly to the diagonal by the time the sample size reaches 5,000

models, and we take this as evidence that MCMC indeed provided a thorough sample of the true marginal distribution of each matrix element in both data sets.

Another way to assess MCMC convergence is to look at the per-datum log likelihood achieved by models from each MCMC run as a function of the order in which these models were sampled. For the MCMC runs on PBM and ChIP-chip data, shown in Fig. 7*b* and 7*d* respectively, the per-datum log likelihood saturated by the time the 50th model was recorded. The fact that no large jumps in the per-datum log likelihood occurred in any of the runs after that point strongly suggests that all MCMC runs quickly found the bulk of the posterior distribution and had ample time to sample it before being terminated.

χ^2 Test for Distribution Consistency

In the course of this work we needed to assess the consistency of independently obtained MCMC distributions for model parameters describing the same TF, either inferred from different data sets (as in *Results*) or by MCMC sampling using different matrix widths or data quantizations. We perform this assessment as follows:

Suppose we have well sampled model ensembles $\Theta_1, \ldots, \Theta_K$, generated by different MCMC runs. Let μ_{blk} and σ^2_{blk} denote the mean and variance of matrix element M_{bl} in the distribution Θ_k. Let us also assume, for these purposes, that all underlying distributions are Gaussian. The value M^*_{bl} of M_{bl} that maximizes the joint likelihood over all K distributions is then the value that minimizes the χ^2 statistic

$$\chi^2_{bl} = \sum_k \frac{(M^*_{bl} - \mu_{blk})^2}{\sigma^2_{blk}}. \qquad \textbf{[19]}$$

This statistic is minimized by the weighted average of means

$$M^*_{bl} = \frac{\sum_k \mu_{blk}/\sigma^2_{blk}}{\sum_k 1/\sigma^2_{blk}}. \qquad \textbf{[20]}$$

This is the best solution to the problem of satisfying all the distributions at once, but it may not be a good solution: if χ^2_{bl} is too large, then M^*_{bl} is improbable according to at least some of the distributions $\{\Theta_k\}$. For the different distributions for M_{bl} to be consistent with each other, this minimized χ^2_{bl} should not be much larger than its expected value in the χ^2 distribution with K degrees of freedom. The p-value, or the probability of finding values $> \chi^2_{bl}$, therefore provides a convenient element-by-element diagnostic of the consistency of multiple MCMC runs (e.g., Fig. 3*c*). Since the Gaussian assumption is generally not accurate, these p-values should not be interpreted too literally. They do, however, provide a useful and easy-to-compute diagnostic.

Robustness of EMA Likelihood Analysis

To test for over-fitting in Θ_{PBM}, we divided Mukherjee *et al.*'s (5) Abf1p data into two randomly chosen halves: A and B. For each half of the data, we partitioned the probed sequences into equi-populated z-bins containing 20 sequences each and ran separate MCMC samplings. The mean matrix elements in the two resulting ensembles, $\Theta_{PBM,A}$ and $\Theta_{PBM,B}$ are shown in Figs. 8*a* and 8*b*. Clearly these means are very similar despite being inferred from two completely disjoint sets of data. Even much of the fine structure in the low specificity regions of the binding site is similar. The χ^2 consistency p-values shown in Fig. 8*c* confirm this, revealing no significant discrepancy between the two parameter distributions for any of the matrix elements.

We also tested whether or not our results were sensitive to how we binned the data. Fig. 6*a* and 6*b* shows the matrix element means deduced from MCMC runs performed with 20 and 50 sequences per z-bin, respectively. Again, these means are very similar, and the χ^2 consistency p-values shown in Fig. 6*c* confirm the absence of significant discrepancies in any of the matrix elements. However, we note that MCMC sampling using 50 sequences per bin was much less efficient, and multiple MCMC replicas were needed to obtain consistent estimates of the posterior distribution.

We further tested the sensitivity of our predictions to the choice of matrix width. MCMC runs were performed on Mukherjee *et al.*'s (5) Abf1p PBM data using matrix widths ranging from 14 to 26; all enough to encompass the primary Abf1p binding site. The MCMC ensembles for the shorter matrices tended to have larger matrix elements, but this is expected because the lack of additional positive energy contributions in the flanking bases should result in a lower energy cutoff; when this cutoff is normalized to 1, matrix elements across the entire energy matrix become artificially large. When checking these MCMC ensembles for consistency, we therefore allowed an arbitrary rescaling for each. The resulting rescaled energy matrix means from each ensemble are shown in Fig. 9. Chi-squared consistency p-values for these seven different ensembles are shown at the top (only for matrix elements shared by all MCMC ensembles) and reveal no significant discrepancies.

Analysis of ChIP-chip Data

Figs. 10 and 11 (mirroring Figs. 1 and 2) show our analysis results for Θ_{ChIP}, which was inferred from Lee *et al.*'s (6) Abf1p ChIP-chip data (using 50 regions per z-bin instead of 20). These results are very similar to those obtained for Mukherjee *et al.*'s (5) PBM data. Fig. 12 reveals a lack of over fitting in Θ_{ChIP}, and Figs. 13 and 14 demonstrate the insensitivity of these results to the level of z-bin quantization and the choice of matrix width. We note that, for unknown reasons, MCMC sampling was much more efficient on this data than on PBM data when coarser z-bin quantizations were used. This allowed us to test quantizations ranging from 20 to 200 regions per z-bin.

References

1. Takeda Y, Sarai A, Rivera VM (1989) *Proc Natl Acad Sci USA* 86:439-443.

2. Sarai A, Takeda Y (1989) *Proc Natl Acad Sci USA* 86:6513-6517.

3. Liu X, Clarke ND (2002) *J Mol Biol* 323:1-8.

4. Hogan ME, Austin RH (1987) *Nature* 329:263-266.

5. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML (2004) *Nat Genet* 36:1331-1339.

6. 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.

7. Foat BC, Morozov AV, Bussemaker HJ (2006) *Bioinformatics* 22:141-149.