Published in final edited form as:
J Stat Phys. 2011 April ; 142(6): 1187–1205. doi:10.1007/s10955-010-0102-x.

Statistical Mechanics of Transcription-Factor Binding Site Discovery Using Hidden Markov Models

Pankaj Mehta,
Dept. of Physics, Boston University, Boston, MA, USA pankajm@bu.edu

David J. Schwab, and
Dept. of Molecular Biology and Lewis-Sigler Institute, Princeton University, Princeton, NJ, USA
dschwab@princeton.edu

Anirvan M. Sengupta
BioMAPS and Dept. of Physics, Rutgers University, Piscataway, NJ, USA
anirvans@physics.rutgers.edu

Abstract

Hidden Markov Models (HMMs) are a commonly used tool for inference of transcription factor (TF) binding sites from DNA sequence data. We exploit the mathematical equivalence between HMMs for TF binding and the "inverse" statistical mechanics of hard rods in a one-dimensional disordered potential to investigate learning in HMMs. We derive analytic expressions for the Fisher information, a commonly employed measure of confidence in learned parameters, in the biologically relevant limit where the density of binding sites is low. We then use techniques from statistical mechanics to derive a scaling principle relating the specificity (binding energy) of a TF to the minimum amount of training data necessary to learn it.

Keywords

Bioinformatics; Hidden Markov Models; One-dimensional statistical mechanics; Fisher information; Machine learning

1 Introduction

Biological organisms control the expression of genes using transcription factor (TF) proteins. TFs bind to regulatory DNA segments (6-20bp) called binding sites thereby controlling the expression of nearby genes. An important task in Bioinformatics is identifying TF binding sites from DNA sequence data. This poses a non-trivial pattern recognition problem, and many computational and statistical techniques have been developed towards this goal. The goal of these algorithms is to identify new binding sites starting from a known collection of TF binding sites. Many different types of algorithms exist including Position Weight Matrices (PWMs) [1, 2], biophysics-inspired algorithms [2, 3], Hidden Markov Models (HMMs) [4–6], and information theoretic algorithms [7].

In general, only a limited number of binding sites are known for a given TF. Thus, any algorithm must build a general classifier based on limited training data. This places constraints of the type of algorithms and classifiers that can be used. The end goal of all models is *generalization*—the ability to correctly categorize new sequences that differ from the training set. This is especially important since the training set is comprised of a small sample fraction of all possible sequences. Most algorithms create a (often probabilistic)

model for whether a particular DNA sequence is a binding site. The model contains a set of parameters, θ, that are fit, or learned, from training data.

All algorithms exploit the statistical differences between binding sites and background DNA in order to identify new binding sites. Two distinct factors contribute to how well one can learn θ, the size of the training data set and the specificity of the TF under consideration. Many TFs are highly specific. Namely, they bind strongly only to small subset of all possible DNA sequences which are statistically distinct from background DNA. Physically, this means that these TF have large binding energies for certain sequence motifs (binding sites) and low binding energies for random segments of DNA, i.e. "background" DNA. Other TFs are less specific and often exhibit non-specific binding to random DNA sequences. In this case, the statistical signatures that distinguish binding sites from background DNA are less clear. In general, the more training data one has and the more specific a TF, the easier it is to learn its binding sites.

This raises the natural question: how much data is needed to train an algorithm to learn the binding sites of a TF? In this paper, we explore this question in the context of a widely-used class of bioinformatic methods termed Hidden Markov Models (HMMs). We exploit the mathematical equivalence between HMMs for TF binding and the "inverse" statistical mechanics of hard rods in a one-dimensional disordered potential to derive a scaling principle relating the specificity (binding energy) of a TF to the minimum amount of training data necessary to learn its binding sites. Unlike ordinary statistical mechanics where the goal is to derive statistical properties from a given Hamiltonian, the goal of the "inverse" problem is to learn the Hamiltonian that most likely gave rise to the observed data. Thus, we are led to consider a well-studied physics problem [8]—the statistical mechanics of a one-dimensional gas of hard rods in an arbitrary external potential—from an entirely new perspective.

The paper is organized as follows. We start by reviewing the mapping between HMMs and the statistical mechanics of hard rods. We then introduce the Fisher Information, a commonly employed measure of confidence in learned parameters, and derive an analytic expression for the Fisher information in the dilute binding site limit. We then use this expression to formulate a simple criteria for how much sample data is needed to learn the binding sites of a TF of a given specificity.

2 HMMs for Binding Site Discovery

HMMs are powerful tools for analyzing sequential data [9, 10] that have been adapted to binding site discovery [5, 6]. HMMs model a system as a Markov process on internal states that are hidden and cannot be observed directly. Instead, the hidden states can only be inferred indirectly through an observable state-dependent output. In the context of binding site discovery, HMMs serve as generative models for DNA sequences. A DNA sequence is modeled as a mixture of hidden states—background DNA and binding sites—with a hidden state-dependent probability for observing a nucleotide (A, T, C, G) at a given location (see Fig. 1).

For concreteness, consider a TF whose binding sites are of length l. An HMM for discovering the binding sites can be characterized by four distinct elements (see Fig. 1) [10]:

1. $l + 1$ hidden states with state 0 corresponding to background DNA and states $j = 1 \ldots l$ corresponding to position j of a binding site.
2. 4 observation symbols corresponding to the four observable nucleotides $\alpha = A, T, C, G$.

3. The transition probabilities, $\{a_{ij}\}$ $(i, j = 0 \ldots l)$ between the hidden states which take the particular form shown in Fig. 1 with only $a_{j,j+1}$, a_{l0}, and a_{01} non-zero. In addition, for simplicity, we assume that binding sites cannot touch (i.e. a_{l1} 0). The generalization to the case where the last assumption is relaxed is straightforward.

4. The observation symbols probabilities $\{b_j(\alpha)\}$ for seeing a symbol $\alpha = A, C, G, T$ in a hidden state j. Often we will rewrite these probabilities in more transparent notation with $p_\alpha = b_0(\alpha)$, the probability of seeing base α in the background DNA, and, $p_{j\alpha}^{(bs)} = b_j(\alpha)$ the probability of seeing base α at position j in a binding site. Finally, denote the collection of all parameters of an HMM (a_{ij} and $b_j(\alpha)$) by the symbol θ.

A DNA sequence of length L, $\mathcal{S} = s_1 s_2 \ldots s_L$, is generated by an HMM starting in a hidden state q_1 as follows. Starting with $i = 1$, choose s_i according to $b_{qi}(s_1)$ and then switch to a new hidden state q_{i+1} using the switching probabilities a_{qiqi+j} and repeat this process until $i = L$. In this way, one can associate a probability, $p(\mathcal{S}|\theta)$, to each sequence $\mathcal{S}$, corresponding to the probability of generating $\mathcal{S}$ using an HMM with parameters θ. The goal of bioinformatic approaches is to learn the parameters θ from training data and use the result to predict new binding sites. Many specialized algorithms, often termed dynamic programming in the computer science literature, have been developed to this end [9, 10].

2.1 Mapping HMMs to the Statistical Mechanics of Hard Rods

Before discussing the mapping between HMMs and statistical mechanics, we briefly review the physics of a one-dimensional gas of hard rods in a disordered external field [8]. The system consists of hard rods—one-dimensional hard core particles—of length l in a spatially dependent binding energy $E(S_{xi})$, with x_i the location of the starting site, at a inverse temperature β, and a fugacity, z (i.e. chemical potential $\mu = \log z$). The equilibrium statistical mechanics of the system is determined by the grand canonical partition function obtained by summing over all possible configurations of hard rods obeying the hard-core constraint [8]. In addition, the pressure can be calculated by taking the logarithm of the grand canonical partition function. Since this model is one-dimensional and has only local interactions, many statistical properties can be calculated exactly using Transfer Matrix techniques. Consequently, variations of this simple hard rod model have been used extensively to model the sequence dependence of nucleosome positioning [11, 12].

We now discuss the mapping between HMMs and a gas of hard rods. We start by showing that the observation symbol probabilities $b_j(\alpha)$ have a natural interpretation as a binding energy. Consider a DNA sequence $S = s_1 \ldots s_l$, with l the length of a binding site. Denote the corresponding hidden state at the j-th position of S by q_j. It is helpful to represent this sequence by a l by 4 matrix $S_{j\alpha}$ of DNA of length l where $S_{j\alpha} = 1$ if base $s_i = \alpha$ and zero otherwise. Denote the probability of generating S from background DNA as $P\left(S|\left\{q_j=0, j=1\ldots l\right\},\theta\right) = \prod_{j=1}^{l} b_0\left(s_j\right)$, and the probability of generating the same sequence within a binding site is $P(S|\{q_j = j, j = 1 \ldots l\}, \theta) = \prod_j b_j(s_j)$. Note that we can rewrite the ratio of these as probabilities as

$$\frac{P\left(S|\left\{q_j=j, j=1\ldots l\right\},\theta\right)}{P\left(S|\left\{q_j=0, j=1\ldots\right\},\theta\right)} = \prod_j \frac{b_i\left(s_j\right)}{b_0\left(s_j\right)} \equiv e^{-E(S)} \tag{1}$$

where we have defined a “sequence-dependent” binding energy

$$E(S) = \epsilon \cdot S = \sum_{\alpha j} \epsilon_{\alpha j} S^{j\alpha} \tag{2}$$

with

$$\epsilon_{\alpha j} \equiv -\log\left(\frac{b_j(\alpha)}{b_0(\alpha)}\right) = -\log\left(\frac{p_{j\alpha}^{(bs)}}{p_\alpha}\right). \tag{3}$$

Notice that the ratio (1) is of a Boltzmann form with a 'binding energy' that can be expressed in terms of a Position Weight Matrix (PWM), ϵ, related to the observation symbol probabilities (3).

Now consider a sequence $\mathcal{S} = s_1 s_2 \ldots s_L$ of length $L \gg l$. In this case, the probability of generating the sequence, $P(\mathcal{S}|\theta)$, is obtained by summing over all possible hidden state configurations. Notice that we can uniquely denote a hidden state configuration by specifying the starting positions within the sequence $\mathcal{S}$ of all the binding sites, $\{x_1 \ldots x_n\}$. The hardrod constraint means that the only allowed configuration are those where $|x_u - x_v| \geq l + 1$ for all u, v (the extra factor of 1 arises because $a_{l0} = 0$). Consequently, the probability of generating a sequence $\mathcal{S}$ is given by summing over all possible hidden state configurations

$$P(\mathcal{S}|\theta) = \sum_n \sum_{x_1 \ldots x_n} P(\mathcal{S}|\{x_1 \ldots x_n\}, \theta) P(\{x_1 \ldots x_n\}|\theta), \tag{4}$$

where $P(\{x_1 \ldots x_n\}|\theta)$ is the probability of generating an allowed hidden state configuration, $\{x_1, \ldots, x_n\}$ and we have factorized the probability using the fact that in a HMM, transition probabilities are independent of the observed output symbol. Furthermore, the ratio of $P(\{x_1 \ldots x_n\}|\theta)$ to the probability of generating a hidden-state configuration with no binding sites, $P(\varnothing|\theta)$ is just

$$\frac{P(\{x_1 \ldots x_n\}|\theta)}{P(\varnothing|\theta)} = z^n = e^{n\mu} \tag{5}$$

with the 'fugacity', z, given by

$$z = \frac{a_{01}}{(1 - a_{01})^{l+1}} \tag{6}$$

and $\mu = \log z$ the chemical potential. Combining (1), (5), and (4) yields

$$\frac{P(\mathcal{S}|\theta)}{C(\mathcal{S}, \theta)} = \mathcal{Z}(\mathcal{S}|\theta) \tag{7}$$

with

$$\mathcal{Z}(\mathcal{S}|\theta) = \sum_{n=0}^{L/l} \sum_{x_1 \ldots x_n} e^{-\Sigma x_1 \ldots x_n E(x_i)} z^n \tag{8}$$

and

$$C(\mathscr{S},\theta)=a_{00}^{L-1}\prod_{i,\alpha}p_{\alpha}^{\mathscr{S}_{i\alpha}} \tag{9}$$

where $E(x_i)$ is the binding energy, (1), for a sub-sequence of length l starting at position x_i of $\mathscr{S}$.

Notice that $\mathscr{Z}(\mathscr{S}|\theta)$ is the grand canonical partition function for a classical fluid of hard rods in an external potential [8]. The sequence-dependence PWM ε acts as an arbitrary external potential, and the switching rate a_{01} sets the chemical potential for binding. Thus, up to a multiplicative factor $C(\mathscr{S},\theta)$ that is independent of the emission probabilities for binding sites, an HMM is mathematically equivalent to a thermodynamic model of hard rods. Importantly, the amount of training data, L, plays the role of system size. Furthermore, the negative log-likelihood, $-\log\ P(\mathscr{S}|\theta)$ is, up to a factor of L just the pressure of the gas of hard rods [8]. In what follows, we exploit the relationship between system size and the quantity of training data to use insights from finite-size scaling to better understand how much data one needs to learn small differences. The relationship between HMMs and the statistical mechanics of hard rods is summarized in Table 1.

2.2 HMMs, Position-Weight Matrices, and Cutoffs

The matrix of parameters, $\varepsilon_{i\alpha}$, defined in (3), are often referred to in bioinformatics as the Position Weight Matrix (PWM) [1, 2]. PWMs are the most commonly used bioinformatic method for discovering new binding sites. In PWM-based approaches, sequences, S, whose binding energies, $E(S) = \varepsilon \cdot S$ are below some arbitrary threshold, are considered binding sites. This points to a major shortcoming of PWM based methods- namely the inability to learn a threshold directly from data. A major advantage of HMM models over PWM-only approached is that HMMs learn both a PWM, ε, and a natural "cutoff" through the chemical potential $\mu = \log z$ [6]. In terms of the corresponding hard-rod model, the probability, $P_{bs}(S)$, for a sequence, S, to be a binding site takes the form of a Fermi-function,

$$P_{bs}(S)=\frac{1}{1+e^{\epsilon\cdot S-\mu}}. \tag{10}$$

If one makes the reasonable assumption that a sequence S is a binding site if $P_{bs}(S) > 1/2$, we see that μ serves as a natural cut-off for binding site energies [6]. Thus, the switching probabilities a_{ij} of the HMM can be interpreted as providing a natural cut-off for binding energies through (6). This points to a natural advantage of HMMs over PWM-only approach, namely one learns the threshold binding energy for determining whether a sequence is a binding site self-consistently from the data. Thus, though in practice binding sites are dilute in the DNA and hard-rod constraints can often be neglected, it is still beneficial to use the full HMM machinery for binding site discovery.

3 Fisher Information & Learning with Finite Data

3.1 Fisher Information and Error-bars

In general, learning the parameters of an HMM from training data is a difficult task. Commonly, parameters of an HMM are chosen to maximize the likelihood of observed data, $\mathscr{S}$, through Maximum Likelihood Estimation (MLE), i.e. parameters are chosen so that

$$\widehat{\theta}=\underset{\theta}{\operatorname{argmax}}\ \ \mathscr{L}(\mathscr{S}|\theta)=\underset{\theta}{\operatorname{argmax}}\log\ P(\mathscr{S}|\theta). \tag{11}$$

Finding the global maxima is an extremely difficult problem. However, one can often find a local maximum in parameter space, $\widehat{\theta}$, using Expectation Maximization algorithms such as Baum-Welch [13]. In general, for any finite amount of training data, the learned parameters $\widehat{\theta}$ (even if they are a global maxima) will differ from the "true" parameters θ_T. The reason for this is that the probabilistic nature of HMMs leads to 'finite size' fluctuations so that the training data may not be representative of the data as a whole. These fluctuations are suppressed asymptotically as the training data size approaches infinity. For this reason, it is useful to have a measure of how well the learned parameters $\widehat{\theta}$ describe the data.

In the remainder of the paper, we assume there is enough training data to ensure that we can consider parameters in the neighborhood of the true parameters. The mapping between HMMs and the statistical mechanics of hard rods allows us to gain insight into the relationship between the amount of training data and the confidence in learned parameters. Recall that log-likelihood per unit volume, $\mathscr{L}(\mathscr{S}|\theta)/L$, is analogous to a pressure and the amount of training data is just the system size. From finite-size scaling in statistical mechanics, we know that as $L \to \infty$ the log-likelihood/pressure becomes increasingly peaked around its true value (see Fig. 2). In addition, we can approximate the uncertainty we have about parameters by calculating the curvature of the log-likelihood, $\partial^2_{AB}\mathscr{L}(\mathscr{S}|\theta)$, around $\widehat{\theta}$ where ∂_A denotes the derivative with respect to the A-th parameter.

This intuition can be formalized for MLE using the Cramer-Rao bound which relates the covariance of estimated parameters to the Fisher Information (FI) Matrix, $\mathscr{I}_{AB}(\theta)$, defined by

$$[\mathscr{I}(\theta)]_{AB} = -E_\theta\left[\partial^2_{AB}\mathscr{L}(\mathscr{S}|\theta)\right] \tag{12}$$

where $E_\theta[g(\mathscr{S})] = \sum_{\mathscr{S}} p(\mathscr{S}|\theta)\, g(\mathscr{S})$ ([9] and see Appendices A, B and C). An important property of the Fisher information is that it provides a bound for how well one can estimate the parameters of the likelihood function by placing a lower bound on the covariance of the estimated parameters. The Cramer-Rao bound relates the Fisher information to the expected value of an unbiased estimator, $E_\theta\left[\widehat{\theta}(\mathscr{S})\right]$, and the covariancematrix of the estimator,

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \equiv E_\theta\left[\left(\widehat{\theta}_A(\mathscr{S}) - \theta_A\right)\left(\widehat{\theta}_B(\mathscr{S}) - \theta_B\right)\right],$$

through the inequality

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right] \geq [\mathscr{I}(\theta)]^{-1}. \tag{13}$$

For MLE, the Cramer-Rao bound is asymptotically saturated in the limit of infinite data. Thus, we expect the Fisher Information to be a good approximation for $\mathrm{Cov}_\theta\left[\widehat{\theta}\right]$ when the amount of training data is large. In the limit of large data, the pressure, or equivalently $\mathscr{L}(\mathscr{S}|\theta)$, "self-averages" and we can ignore the expectation value (12). Thus, to leading order in L, one can approximate the covariance matrix as

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \approx [\mathscr{I}(\theta)]^{-1}_{AB} \approx -\left[\partial^2_{AB}\mathscr{L}(\mathscr{S}|\theta)\right]^{-1}, \tag{14}$$

in agreement with intuition from finite size scaling. The previous expression provides a way to put error bars on learned parameters. However, in practice we seldom have access to the "true" parameters θ that generated the observed sequences. Instead, we only know the

parameters learned from the training data, $\widehat{\theta}$. Thus, one often substitutes $\widehat{\theta}$, our best guess for the parameters θ in (14).

3.2 Fisher Information as Correlation Functions

It is worth noting that the expression above, in conjunction with the mapping to the hard-rod model, allows us to calculate error bars directly from data. In particular, we show below that the Fisher information can be interpreted as a correlation function and thus can be calculated using Transfer Matrix techniques. It is helpful to reframe the discussion above in the language of the statistical mechanics of disordered systems. Recall that up to a normalization constant, $C(\mathcal{S},\theta)$, in the corresponding hard-rod model $P(\mathcal{S}|\theta)$ is the grand canonical partition function, $\mathcal{L}(\mathcal{S}|\theta)/L$ is the pressure, and the amount of training data, L, is just the size of the statistical mechanical system (see Table 1 of main text). When L is large, we expect that the sequence $\mathcal{S}$ self-averages and the Fisher Information is related to the second derivative of the log-likelihood of the observed data,

$$[\mathcal{I}(\theta)]_{AB} \approx -\partial^2_{AB}\mathcal{L}(\mathcal{S}|\theta). \tag{15}$$

Thus, aside from the normalization $C(\mathcal{S},\theta)$, the Fisher Information can be calculated from the second derivative of the pressure. From the fluctuation dissipation theorem, we conclude that the Fisher information can be expressed in terms of connected correlation functions. In particular, let $\mathcal{O}_A$ be the operator conjugate to the A-the parameter, θ_A in the partition function $\mathcal{Z}_{\mathcal{S}}(\theta)$. The Fisher Information then takes the form

$$[\mathcal{I}(\theta)]_{AB} = \langle \mathcal{O}_A \mathcal{O}_B \rangle_c - \partial_{AB} \quad \log \quad C(\mathcal{S},\theta), \tag{16}$$

where $\langle \mathcal{O}_A \mathcal{O}_B \rangle_c = \langle \mathcal{O}_A \mathcal{O}_B \rangle - \langle \mathcal{O}_A \rangle \langle \mathcal{O}_B \rangle$ and

$$\langle \mathcal{O} \rangle = \frac{\sum_{n=0}^{L/l} \sum_{x_1 \ldots x_n} e^{-E(S_{x_i})} z^n \mathcal{O}}{\mathcal{Z}_{\mathcal{S}}(\theta)} \tag{17}$$

Note that these correlation functions can be calculated directly from the data using Transfer Matrix techniques without resorting to more complicated methods.

In general, the background statistics of the DNA are known and the parameters one wishes to learn are the switching rates, a_{ij}, and symbol observation probabilities, $b_j(\alpha)$. In practice, it is often more convenient to work with the fugacity, z, rather than the switching rate (see Table 1). The operator conjugate to the fugacity is n, the number of binding sites. Consequently,

$$[\mathcal{I}(\theta)]_{zz} = \left\langle (n - \langle n \rangle)^2 \right\rangle + \partial_{zz} \quad \log \quad C(\mathcal{S},\theta). \tag{18}$$

Thus, the uncertainty in the switching rates is controlled by the fluctuations in binding site number, as is intuitively expected. One can also derive the conjugate operators for the emission probabilities $b_j(\alpha)$ and/or the sequence dependent "binding energies" $\varepsilon_{j\alpha}$ (see Table 1) via a straight forward calculation (see calculations in sections below).

The expression (16) provides a computationally tractable way to calculate the Fisher Information and, consequently, the covariance matrix $\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB}$. Not only can we learn the maximum likelihood estimate for parameters, we can also put 'error bars' on the MLE. We emphasize that in general, this requires powerful, computationally intensive techniques.

However, by exploiting transfer matrix/ dynamic programming techniques, the correlation functions (16) can be computed in polynomial time. This result highlights how thinking about HMMs in the language of statistical mechanics can lead to interesting new results.

4 Analytic Expression Using a Virial Expansion

In general, calculating the log-likelihood $\mathscr{L}(\mathscr{S},\theta)$ analytically is intractable. However, we can exploit the fact that binding sites are relatively rare in DNA and perform a Virial expansion in the density of binding sites, ρ, or in the HMM language, the switching rate from background to binding site $(a_{01} \ll 1)$. This is a good approximation in most cases. For example, for the NF-κB TF family, a_{01} was recently found to be of order 10^{-2}–10^{-4} [6] Thus, to leading order in ρ, we can ignore exclusion effects due to overlap between binding sites and write the partition function of the hard-rod model as

$$\mathscr{Z}_{\mathscr{S}}(\theta) = \sum_{n=0} \prod_{x_1 \ldots x_n} e^{-E(x_i)} z^n \approx \prod_{\sigma=1}^{L} \left(1 + z e^{-E(S^\sigma)}\right) + \sigma\left(\rho^2\right) \tag{19}$$

where $E(S^\sigma)$ is the binding energy, (2), for a hard-rod bound to a sequence, S^σ, of length l starting at position σ on the full DNA sequence $\mathscr{S}$. The corrections due to steric exclusion are higher order in density and thus can be ignoredSto leading order. Thus, the log-likelihood takes the simple form

$$\mathscr{L}(\mathscr{S}|\theta) \approx \sum_{\sigma} \log\left(1 + z e^{-E(S^\sigma)}\right) - \log \quad C(\mathscr{S},\theta), \tag{20}$$

where $C(\mathscr{S},\theta)$ is a normalization constant.

Notice the log-likelihood (20) is a sum over the free-energies of single particles in potentials given by the observed DNA. For long sequences where $L \gg 1$, we expect on average $N = La_{01}$ binding sites, and $L - N$ background DNA sequences in the sum. In this case, we expect that the single particle energy self-averages and we can replace the sum by the average value of the single-particle free energy in either background DNA or a binding site. In particular, we expect that

$$\mathscr{L}(\mathscr{S}|\theta) \approx N\left\langle \log\left(1 + z e^{-E(S)}\right)\right\rangle_{bs} + (L - N)\left\langle \log\left(1 + z e^{-E(S)}\right)\right\rangle_{bg} - \log \quad C(\mathscr{S},\theta) \tag{21}$$

where $\langle H(S)\rangle_{bg}$ and $\langle H(S)\rangle_{bs}$ are the expectation value of $H(S)$ for sequences S of length l drawn from the background DNA and binding site distributions, respectively.

4.1 Maximum Likelihood Equations via the Virial Expansion

We now derive the Maximum-Likelihood equations (MLE) within the Virial expansion to the log-likelihood (20). Recall from (11) that the Maximum Likelihood estimator is the set of parameters most likely to generate the data. Thus, we can derive MLE by taking the first derivatives of the log-likelihood and setting the expressions to zero. Consider first the MLE for the binding energy matrix $\varepsilon_{i\alpha}$. Since $C(S, \theta)$ is independent of the binding energy, we focus only on the first term of (20). Define the matrix $S_{i\alpha}$ which is one if position i has base α and zero otherwise. The MLE can be derived by taking the first derivative

$$\partial_{\epsilon_{i\alpha}}\left[\sum_{\sigma}\log\left(1+ze^{-E(S^{\sigma})}\right)+\sum_{i}\lambda_i\left(\sum_{\alpha}p_{\alpha}e^{-\epsilon_{i\alpha}}-1\right)\right]$$
$$=\partial_{\epsilon_{i\alpha}}\left[\sum_{\sigma}\log\left(1+ze^{-E(S^{\sigma})}\right)+\sum_{i}\lambda_i\left(\sum_{\alpha}p_{\alpha i}^{(bs)}-1\right)\right] \quad (22)$$

where λ_i are Lagrange multipliers that ensure proper normalization of probabilities. Explicitly taking the derivative, using probability conservation, and noticing that $\sum_{\alpha}S_{i\alpha}=1$ gives

$$\frac{\Sigma_{\sigma}f_{z,\epsilon}(S^{\sigma})S_{i\alpha}}{\Sigma_{\sigma}f_{z,\epsilon}(S^{\sigma})}=p_{\alpha}e^{-\epsilon_{i\alpha}}=p_{i\alpha}^{(bs)} \quad (23)$$

where

$$f_{z,\epsilon}(S^{\sigma})=\frac{1}{1+z^{-1}e^{E(S^{\sigma})}} \quad (24)$$

is the Fermi-Dirac distribution function.

We can also derive the MLE corresponding to the fugacity. The fugacity depends explicitly on the normalization constant $C(\mathscr{S},\theta)$. Note that in HMMs, $C(\mathscr{S},\theta)=a_{00}^{L}\prod_{\sigma}p_{\alpha}^{\mathscr{S}_{i\alpha}}$ and ensures probability conservation. Since $z = a_{01}/(1 - a_{01})^{l+1}$, to leading order in a_{01}, naively $\log C(\mathscr{S},\theta)\sim L\log(1-z)$. However, choosing this normalization explicitly violates probability conservation in the corresponding HMM because we have truncated the Virial expansion for the log-likelihood at first order and consequently allowed unphysical configurations. Since deriving the MLEs requires probability conservation, we impose by hand that the normalization has the z dependence,

$$\log C(\mathscr{S},\theta)\sim L\log(1+z). \quad (25)$$

With this normalization, the log-likelihood (20) becomes analogous to that for a mixture model where the sequences S^{σ} are drawn from background DNA or binding sites. With this choice of $C(S, \theta)$ the MLE equations can be calculated in a straightforward manner by taking the derivative of (20) with respect to z (see Appendix B) to get

$$\sum_{\sigma}f_{z,\epsilon}(S^{\sigma})=\frac{Lz}{1+z}. \quad (26)$$

4.2 Fisher Information via the Virial Expansion

One can also derive an analytic expressions for the Fisher information within the Virial expansion. Generally, the background observation probabilities $b_0(a) = p_a$ are known and the HMM parameters, θ, to be learned are the observation symbol probabilities in binding sites, $b_j(\alpha)=p_{j\alpha}^{(bs)}$ and the switching probability a_{01}. Technically, it is easier to work with the corresponding parameters of the hard-rod model, the binding energies ε_{ia} and the fugacity, z. Note that probability conservation and (3) imply that only three of the ε_{ia} ($a = A, C, G$) are independent. A straight forward calculation (see Appendices A, B and C) yields

$$\left[\mathscr{I}(\theta)^{-1}\right]_{i\alpha,j\beta}\approx N\left\langle A_{i\alpha,j\beta}\right\rangle_{bs}+(L-N)\langle A_{i\alpha},j\beta\rangle_{bg} \quad (27)$$

with

$$A_{i\alpha,i\beta}=[f_{z,\epsilon}(S)]^2\left[\delta_{\alpha\beta}S_{i\alpha}S_{i\beta}+\frac{p_{i\alpha}^{(bs)}p_{i\beta}^{(bs)}}{p_{iT}^{(bs)^2}}S_{iT}S_{iT}\right] \tag{28}$$

and for $i \neq j$,

$$A_{i\alpha,(j\neq i)\beta}=-f_{z,\epsilon}(S)(1-f_{z,\epsilon}(S))\left[S_{i\alpha}-\frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}}S_{iT}\right]\left[S_{j\beta}-\frac{p_{j\beta}^{(bs)}}{p_{jT}^{(bs)}}S_{jT}\right]$$

where, as above, $f_{z,\epsilon}(S)$ is the Fermi-Dirac distribution function

$$f_{z,\epsilon}(S)=\frac{1}{1+z^{-1}e^{E(S)}}. \tag{29}$$

One also has (see Appendix B)

$$\left[\mathcal{I}(\theta)^{-1}\right]_{i\alpha,z}\approx N\langle C_{i\alpha}\rangle_{bs}+(L-N)\langle C_{i\alpha}\rangle_{bg}, \tag{30}$$

with

$$C_{i\alpha}=\frac{1}{z}f_{z,\epsilon}(S)(1-f_{z,\epsilon}(S))S_{i\alpha}, \tag{31}$$

and

$$\left[\mathcal{I}(\theta)^{-1}\right]_{z,z}\approx N\langle D\rangle_{bs}+(L-N)\langle D\rangle_{bg}, \tag{32}$$

with

$$D=\frac{f_{z,\epsilon}(S)}{z^2}-\frac{1}{(1+z)^2}. \tag{33}$$

The expressions (27), (30), and (33) depend only on $\varepsilon_{i\alpha}$ and thus can be used to calculate the expected error in learned parameters as a function of training data using only the Position Weight Matrix (PWM) of a transcription factor and a rough estimate of the switching probability a_{01} or equivalently the fugacity z. The explicit dependence on base T reflects the fact that not all the elements of the PWM are independent.

5 Scaling Relation for Learning with Finite Data

An important issue in statistical learning is how much data is needed to learn the parameters of a statistical model. The more statistically similar the binding sites are to background DNA (i.e the smaller the binding energy of a TF), the more data is required to learn the model parameters. The underlying reason for this is that the probabilistic nature of HMMs means that the training data may not be representative of the data as a whole. Intuitively, it is clear that in order to be able to effectively learn model parameters, the training data set should be large enough to ensure that “finite-size” fluctuations resulting from limited data cannot mask the statistical differences between binding sites and background DNA. To address this question, we must consider PWMs learned from strictly random data. As the size of the

training set is increased, the finite-size fluctuations are tamed. Our approach is then, in a sense, complementary to looking for rare, high-scoring sequence alignments which become *more* likely as L increases in random data [14]. Of course, estimations based on random data neglect non-trivial structure of real sequences [15].

5.1 Maximum Likelihood and Jeffreys Priors

Within the Maximum Likelihood framework, the probability that one learn a ML estimator, $\widehat{\theta}$, given that the data is generated by parameters θ, can be approximated by a Gaussian whose width is related to the Fisher information using a Jeffreys prior [16],

$$P\left(\widehat{\theta}\right) \propto \sqrt{|\mathscr{I}(\theta)|}e^{-\left(\widehat{\theta}-\theta_0\right)[\mathscr{I}(\theta)]^{-1}\left(\widehat{\theta}-\theta_0\right)}. \tag{34}$$

As expected, the width of the Gaussian is set by the covariance matrix for $\widehat{\theta}$, and is related to the second derivative of the log-likelihood through (14). Since the log-likelihood—in analogy with the pressure (times volume) of the corresponding hard-rod gas—is an extensive quantity, an increase in the amount of training data L means a narrower distribution for the learned parameters $\widehat{\theta}$ (see Fig. 2). When L is large, the inverse of the Fisher information is well approximated by the Jacobian of the log-likelihood, (14). In general, the Jacobian is a positive semi-definite, symmetric square matrix of dimension n, with n the number of parameters needed to specify the position weight-matrix and fugacity for a single TF. In most cases, n is large and typically ranges from 24–45, with the exact number equal to three time the length of a binding site.

Label the A-th component of θ by θ_A. Then, the probability distribution (34) can also be used to derive a distribution for the Mahalanobis distance [17]

$$\widehat{r}^2 = -\sum_{A,B}\left[\widehat{\theta}_A - \theta_{0A}\right]\left.\frac{\partial^2 \mathscr{L}(\mathscr{S}|\theta)}{\partial\theta_A\partial\theta_B}\right|_{\theta_0}\left[\widehat{\theta}_B - \theta_{0B}\right]. \tag{35}$$

The Mahalanobis distance is a scale-invariant measure of how far the learned parameters $\widehat{\theta}$ are from the true parameters θ. Intuitively, it measures distances in units of standard deviations. Furthermore, the Mahalanobis distance scales linearly with the amount of data/system size L since it is proportional to log-likelihood $\mathscr{L}(\mathscr{S}|\theta_0)$ (i.e. pressure times volume). By changing variable to the eigenvectors of the Jacobian, normalizing by the eigenvalues, and integrating out angular variables, one can show that (34) yields the following distribution for $\hat{r}$,

$$P(\widehat{r}) \propto \widehat{r}^{n-1}e^{-\widehat{r}^2} = e^{-\widehat{r}^2+(n-1)\log\widehat{r}}. \tag{36}$$

When n is large, we can perform a saddle-point approximation for r around its maximum value,

$$\widehat{r}_* = \sqrt{(n-1)/2}. \tag{37}$$

Writing $\hat{r} = \hat{r}_* + \delta\hat{r}$, one has

$$P(\delta\widehat{r}) \approx e^{-(n-1)/2+\log(n-1)/2}e^{-\delta\widehat{r}^2}. \tag{38}$$

Thus, for large n, almost in all cases the learned parameters $\widehat{\theta}$ will be peaked sharply around a distance, $\widehat{r}_*^2 = (n-1)/2$, with a width of order 1. This result is a general property of large-dimensional Gaussians and will be used below.

5.2 Scaling Relation for Learning with Finite Data

We now formulate a simple criteria for when there is enough data to learn the binding sites of a TF characterized by a PWM ε. We take as our null hypothesis that the data was generated entirely from background DNA (i.e. the true parameters are $\varepsilon_0 = 0$ and $z_0 = z$) and require enough data so that the probability of learning $\widehat{\epsilon}=\epsilon$ be negligible. In other words, we want to make sure that there is enough data so that there is almost no chance of learning $\widehat{\epsilon}=\epsilon$ for data generated entirely from background DNA, $\varepsilon_0 = 0$. From (38), we know that for large n, with probability almost 1 due to finite size fluctuations, any learned $\widehat{\epsilon}$ will lie a Mahalanobis distance, $r^2(\widehat{\epsilon}, z) = (n-1)/2$ away from the true parameters. Thus, we require enough data so that

$$r^2(\epsilon, z) \equiv L\tilde{r}^2(\epsilon, z) \geq (n-1)/2, \tag{39}$$

with $\tilde{r}^2(\varepsilon, z)$ defined by the first equality. An explicit calculation of the left hand side of (39) yields (see Appendices A, B and C)

$$\tilde{r}^2(z, \epsilon) = \frac{z^2}{(1+z)^2}\left[\sum_i \bar{\epsilon}_i^2 + \frac{\bar{\epsilon}_i^2}{p_T}\right] \tag{40}$$

where we have defined

$$\bar{\epsilon}_i^{\gamma} = \sum_{\alpha=A,C,G} p_{i\alpha}\epsilon_{i\alpha}^{\gamma}, \quad \gamma=1,2. \tag{41}$$

Together, (39) and (40) define a criteria for how much data is needed to learn the binding sites of a TF with PWM (binding energy), ε, whose binding sites occur in background DNA with a fugacity z. Notice that (40) contains terms that scale as the square of the energy difference, indicating that it is much easier to learn binding sites with a few large differences than many small differences.

6 Discussion

In this paper, we exploited the mathematical equivalence between HMMs for TF binding and the "inverse" statistical mechanics of hard rods in a one-dimensional disordered potential to investigate learning in HMMs. This allowed us to derive a scaling principle relating the specificity (binding energy) of a TF to the minimum amount of training data necessary to learn its binding sites. Thus, we were led to consider a well-studies physics problem [8]—the statistical mechanics of a one-dimensional gas of hard rods in an arbitrary external potential—from an entirely new perspective.

In this paper, we assumed that there was enough data so that we could focus on the neighborhood of a single maximum in the Maximum Likelihood problem. However, in principle, for very small amounts of data, the parameter landscape has the potential to be glassy and possess many local minima of about equal likelihood. However in our experience, this does not seem to be the case in practice for most TFs. In the future, it will be interesting to investigate the parameter landscape of HMMs in greater detail to understand when they exhibit glassy behavior.

The work presented here is part of a larger series of works that seeks to use methods from "inverse statistical mechanics" to study biological phenomenon [18–21]. Inverse statistical mechanics inverts the usual logic of statistical mechanics where one starts with a microscopic Hamiltonian and calculates statistical properties such as correlation functions.

In the inverse problem, the goal is to start from observed correlations and find the Hamiltonian from which they were most likely generated. In the context of binding site discovery, considering the inverse statistical problem allows us to ask and answer new and interesting questions about how much data one needs to learn the binding sites of a TF. In particular, it allows us to calculate error bars for learned parameters directly from data and derive a simple scaling relation between the amount of training data and the specificity of TF encoded in its PWM.

Our understanding of how the size of training data affects our ability to learn the parameters in inverse statistical mechanics is still in its infancy. It will be interesting to see if the analogy between finite-size scaling in the thermodynamics of disordered systems and learning in inverse statistical mechanics holds in other systems, or if it is particular to the problem considered here. More generally, it will be interesting to see methods from physics and statistical mechanics yield new insights about large data sets now being generated in biology.

Acknowledgments

We would like to thank Amor Drawid and the Princeton Biophysics Theory group for useful discussion. This work was partially supported by NIH Grants K25GM086909 (to PM) and R01HG03470 (to AMS). DS was partially supported by DARPA grant HR0011-05-1-0057 and NSF grant PHY-0957573. PM would also like to thank the Aspen Center for Physics where part of this work was completed.

Appendix A: Covariance Matrix and Fisher Information

The Fisher information is a commonly employed measure of how well one learns the parameters, θ, of a probabilistic model from training data, $\mathcal{S}$. In our context, $\mathcal{S}$ is the observed DNA sequence and θ are the parameters of the HMM for generating DNA sequences. The Fisher information matrix, $\mathcal{I}_{AB}(\theta)$, is given in terms of the log-likelihood, $\mathcal{L}(\mathcal{S}|\theta) = \log\ p(\mathcal{S}|\theta)$, by

$$[\mathcal{I}(\theta)]^{-1}_{AB} \equiv E_\theta[\partial_A \mathcal{L}(\mathcal{S}|\theta)\,\partial_B \mathcal{L}(\mathcal{S}|\theta)] = \sum_{\mathcal{S}} p(\mathcal{S}|\theta)\,\partial_A \mathcal{L}(\mathcal{S}|\theta)\,\partial_B \mathcal{L}(\mathcal{S}|\theta), \tag{A.1}$$

where E_θ denotes the expectation value averaged over different data sets generated using the parameters θ and ∂_A denotes the partial derivative with respect to the A-th component of θ.

The Fisher information can also be expressed as a second derivative of the log-likelihood function. This follows from differentiating both sides of the equation

$$\sum_{\mathcal{S}} e^{\mathcal{L}(\mathcal{S}|\theta)} = 1 \tag{A.2}$$

with respect to θ_A and θ_B which yields the expression

$$\sum_{\mathcal{S}} e^{\mathcal{L}(\mathcal{S}|\theta)}\partial_A \mathcal{L}(\mathcal{S}|\theta)\,\partial_B \mathcal{L}(\mathcal{S}|\theta) + \sum_{\mathcal{S}} e^{\mathcal{L}(\mathcal{S}|\theta)}\partial^2_{AB}\mathcal{L}(\mathcal{S}|\theta) = 0. \tag{A.3}$$

Comparing with (A.1), we see that the Fisher information can also be expressed as

$$[\mathcal{I}(\theta)]^{-1}_{AB} = -E_\theta\left[\partial^2_{AB}\mathcal{L}(\mathcal{S}|\theta)\right] = -\sum_{\mathcal{S}} p(\mathcal{S}|\theta)\,\partial^2_{AB}\mathcal{L}(\mathcal{S}|\theta). \tag{A.4}$$

An important property of the Fisher information is that it provides a bound for how well one can estimate the parameters of the likelihood function. As discussed in the main text, the parameters of a HMM can be estimated from an observed sequence, $\mathcal{S}$, using a Maximum Likelihood Estimator (MLE), $\widehat{\theta}(\mathcal{S})$, defined as

$$\widehat{\theta}(\mathcal{S}) \equiv \arg \max_{\theta} \mathcal{L}(\mathcal{S}|\theta). \tag{A.5}$$

The Cramer-Rao bound relates the Fisher Information to the expected value of the estimator

$$E_\theta\left[\widehat{\theta}(\mathcal{S})\right] = \sum_{\mathcal{S}} \widehat{\theta}(\mathcal{S})\, p(\mathcal{S}|\theta), \tag{A.6}$$

and the covariance matrix of the estimator,

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \equiv E_\theta\left[\left(\widehat{\theta}_A(\mathcal{S}) - \theta_A\right)\left(\widehat{\theta}_B(\mathcal{S}) - \theta_B\right)\right].$$

For a multidimensional estimator, the Cramer-Rao bound is given by

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \geq \sum_{C,D} \partial_B E_\theta\left(\widehat{\theta}_C\right)\left[\mathcal{I}(\theta)\right]_{CD} \partial_B E_\theta\left(\widehat{\theta}_D\right). \tag{A.7}$$

To gain intuition, it is worth considering the special case where the estimator is unbiased, $E_S\left[\widehat{\theta}(S)\right] = \theta$, in which case the Cramer-Rao bound simply reads

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \geq \left[\mathcal{I}(\theta)^{-1}\right]_{AB}. \tag{A.8}$$

Thus, the Fisher information gives a fundamental bound on how well one can learn the parameters of our HMM.

For MLEs, the Cramer-Rao bound is asymptotically saturated in the limit of infinite data. Thus, we expect the Fisher Information to be a good approximation for $\mathrm{Cov}_\theta\left[\widehat{\theta}\right]$ when the length, L, of the DNA sequences, S, from which we learn parameters is long. In this case,

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \approx \left[\mathcal{I}(\theta)\right]^{-1} \tag{A.9}$$

The previous expressions provide a way to put error bars on learned parameters. However, in practice we never have access to the "true" parameters θ that generated the observed sequences. Instead, we only know the parameters learned from the training data, $\widehat{\theta}$. Thus, we make the additional approximation

$$\left[\mathrm{Cov}_\theta\left(\widehat{\theta}\right)\right]_{AB} \approx \left[\mathcal{I}(\theta)\right]^{-1} \approx \left[\mathcal{I}\left(\widehat{\theta}\right)\right]^{-1}. \tag{A.10}$$

Appendix B: Calculation of Fisher Information using a Virial Expansion

B.1 PWM Dependent Elements

We now calculate the Fisher information for a HMM for binding sites from a single binding site distribution using the Virial expansion. We are interested in the Fisher information for

the parameters $\varepsilon_{i\alpha}$ (the energies in the corresponding Position Weight Matrix). An important complication is that not all the $\varepsilon_{i\alpha}$ are independent. In particular, we have

$$\sum_{\alpha} p_{\alpha} e^{-\epsilon_{i\alpha}} = \sum_{\alpha} p_{\alpha i}^{(bs)} = 1 \tag{B.1}$$

Thus, there are only three independent parameters at each position in the binding site. Let us choose ε_{iT} to depend on the other three energies. Rearranging the equation above, one has that

$$\epsilon_{iT} = -\log\left(\frac{1 - \sum_{\alpha \neq T} p_{\alpha} e^{-\epsilon_{\alpha i}}}{p_T}\right) \equiv g_{\epsilon_{iT}} \tag{B.2}$$

Taking the first derivative of (20) with respect to $\varepsilon_{i\alpha}$ with $\alpha \neq T$ yields

$$\frac{\partial \mathcal{L}(\mathcal{S}|\theta)}{\partial \epsilon_{i\alpha}} = -\sum_{\sigma} f_{z,\epsilon}(S^{\sigma}) \left[S_{i\alpha}^{\sigma} + \frac{\partial g_{\epsilon_{iT}}}{\partial \epsilon_{i\alpha}} S_{iT}^{\sigma} \right] \tag{B.3}$$

Taking the second derivative yields

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\mathcal{S}|\theta)}{\partial \epsilon_{i\alpha} \partial \epsilon_{j\beta}} = & \sum_{\sigma} f_{z,\epsilon}(S^{\sigma})\left(1 - f_{z,\epsilon}(S^{\sigma})\right) \left[S_{i\alpha}^{\sigma} + \frac{\partial g_{\epsilon_{iT}}}{\partial \epsilon_{i\alpha}} S_{iT}^{\sigma} \right] \left[S_{j\beta}^{\sigma} + \frac{\partial g_{\epsilon_{jT}}}{\partial \epsilon_{j\beta}} S_{jT}^{\sigma} \right] \\ & - \delta_{ij} S_{iT}^{\sigma} f_{z,\epsilon}(S^{\sigma}) \frac{\partial^2 g_{\epsilon_{iT}}}{\partial \epsilon_{i\alpha} \partial \epsilon_{i\beta}} \end{aligned} \tag{B.4}$$

We can simplify the expressions further by noting

$$\frac{\partial g_{\epsilon_{iT}}}{\partial \epsilon_{i\alpha}} = -\frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} \tag{B.5}$$

$$\frac{\partial^2 g_{\epsilon_{iT}}}{\partial \epsilon_{i\alpha} \epsilon_{i\beta}} = \left[\delta_{\alpha\beta} \frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} + \frac{p_{i\alpha}^{(bs)} p_{i\beta}^{(bs)}}{\left(p_{iT}^{(bs)}\right)^2} \right] \tag{B.6}$$

Plugging in these expressions into (B.4) yields

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\mathcal{S}|\theta)}{\partial \epsilon_{i\alpha} \partial \epsilon_{j\beta}} = & \sum_{\sigma} f_{z,\epsilon}(S^{\sigma})\left(1 - f_{z,\epsilon}(S^{\sigma})\right) \left[S_{i\alpha}^{\sigma} - \frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} S_{iT}^{\sigma} \right] \left[S_{j\beta}^{\sigma} - \frac{p_{j\beta}^{(bs)}}{p_{jT}^{(bs)}} S_{jT}^{\sigma} \right] \\ & - \delta_{ij} S_{iT}^{\sigma} f_{z,\epsilon}(S^{\sigma}) \left[\delta_{\alpha\beta} \frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} + \frac{p_{i\alpha}^{(bs)} p_{i\beta}^{(bs)}}{\left(p_{iT}^{(bs)}\right)^2} \right] \end{aligned} \tag{B.7}$$

The Fisher information is obtained in the usual way from

$$\left[\mathcal{I}(\theta)\right]^{-1}_{\epsilon_{i\alpha}\epsilon_{j\beta}} = -\frac{\partial^2 \mathcal{L}(\mathcal{S}|\theta)}{\partial \epsilon_{i\alpha} \partial \epsilon_{j\beta}} \tag{B.8}$$

B.1.1 Simplified Equations for i = j

When $i = j$, we can simplify the equations above using the ML equations (23) and noting that $S_{i\alpha} S_{i\beta} = \delta_{\alpha\beta} S_{i\alpha}$ and $S_{i\alpha} S_{iT} = 0$. Using the expressions above yields

$$\begin{aligned}\frac{\partial^2 \mathscr{L}(\mathscr{S}|\theta)}{\partial \epsilon_{i\alpha}\partial \epsilon_{i\beta}} = & \sum_{\sigma} f_{z,\epsilon}(S^{\sigma})\left(1 - f_{z,\epsilon}(S^{\sigma})\right)\left[S_{i\alpha}\delta_{\alpha\beta} + \frac{p_{i\alpha}^{(bs)} p_{i\beta}^{(bs)}}{\left(p_{iT}^{(bs)}\right)^2} S_{iT}^{\sigma}\right] \\ & - S_{iT}^{\sigma} f_{z,\epsilon}(S^{\sigma})\left[\delta_{\alpha\beta}\frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} + \frac{p_{i\alpha}^{(bs)} p_{i\beta}^{(bs)}}{\left(p_{iT}^{(bs)}\right)^2}\right]\end{aligned} \tag{B.9}$$

From the MLE (23), we know that

$$\begin{aligned}\sum_{\sigma} f_{z,\epsilon}(S^{\sigma}) S_{iT}^{\sigma}\frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} &= \left[\sum_{\sigma} f_{z,\epsilon}(S^{\sigma}) S_{iT}^{\sigma}\right]\left[\sum_{\sigma'} f_{z,\epsilon}\left(S^{\sigma'}\right) S_{i\alpha}^{\sigma'}\right] / \left[\sum_{\sigma'' \in BS} f_{z,\epsilon}\left(S^{\sigma''}\right) S_{iA}^{\sigma''}\right] \\ &= \sum_{\sigma} f_{z,\epsilon}(S^{\sigma}) S_{i\alpha}^{\sigma}\end{aligned} \tag{B.10}$$

Plugging this into the equations above yields

$$\frac{\partial^2 \mathscr{L}(\mathscr{S}|\theta)}{\partial \epsilon_{i\alpha}\partial \epsilon_{i\beta}} = -\sum_{\sigma}\left[f_{z,\epsilon}\left(S^{\sigma'}\right)\right]^2\left[\delta_{\alpha\beta} S_{i\alpha}^{\sigma} S_{i\beta}^{\sigma} + \frac{p_{i\alpha}^{(bs)} p_{i\beta}^{(bs)}}{\left(p_{iT}^{(bs)}\right)^2} S_{iT}^{\sigma} S_{iT}^{\sigma}\right] = -\sum_{\sigma} A_{ii}(S^{\sigma}) \tag{B.11}$$

This is the operator A_{ii} in the main text.

B.1.2 Simplified Equations for i ≠ j

In this case, we know that

$$\begin{aligned}\frac{\partial^2 \mathscr{L}(\mathscr{S}|\theta)}{\Delta \epsilon_{i\alpha}\Delta \epsilon_{j\beta}} &= \sum_{\sigma} f_{z,\epsilon}(S^{\sigma})\left(1 - f_{z,\epsilon}(S^{\sigma})\right)\left[S_{i\alpha}^{\sigma} - \frac{p_{i\alpha}^{(bs)}}{p_{iT}^{(bs)}} S_{iT}^{\sigma}\right]\left[S_{j\beta}^{\sigma} - \frac{p_{j\beta}^{(bs)}}{p_{jT}^{(bs)}} S_{jT}^{\sigma}\right] \\ &= -\sum_{\sigma} A_{i\alpha,j\beta}(S^{\sigma})\end{aligned} \tag{B.12}$$

This is the operator A_{ij} in the main text.

B.2 Fugacity Dependent Elements

We start by calculating the elements $\mathscr{I}_{i\alpha,z}$. As before, within the virial expansion

$$\mathscr{L}(\mathscr{S}|\theta) \approx \sum_{\sigma} \log\left(1 + z e^{-\epsilon\cdot\sigma}\right) - L \quad \log(1+z). \tag{B.13}$$

Thus, we have

$$\frac{\partial \mathscr{L}}{\partial z} = \sum_{\sigma}\frac{1}{z} f_{z,\epsilon,z}(S^{\sigma}) - L\frac{1}{(1+z)} = \sum_{\sigma}\frac{e^{-\epsilon\cdot S^{\sigma}}}{1 + z e^{-\epsilon\cdot S^{\sigma}}} - L\frac{1}{(1+z)}. \tag{B.14}$$

Taking the second derivate with respect to $\varepsilon_{i\alpha}$ yields

$$\begin{aligned}\frac{\partial^2 \mathscr{L}}{\partial \epsilon_{i\alpha} z} &= -\sum_{\sigma}\frac{\left(S_{i\alpha} - \frac{p_{i\alpha} S_{iT}^{\sigma}}{p_{iT}}\right) e^{-\epsilon\cdot S^{\sigma}}}{\left(1 + z e^{-\epsilon\cdot S^{\sigma}}\right)^2} \\ &= -\sum_{\sigma}\frac{1}{z} f_{z,\epsilon}(S^{\sigma})\left(1 - f_{z,\epsilon}(S^{\sigma})\right)\left(S_{i\alpha} - \frac{p_{i\alpha} S_{iT}^{\sigma}}{p_{iT}}\right)\end{aligned} \tag{B.15}$$

Furthermore, one has

$$\frac{\partial^2 \mathscr{L}}{\partial z^2} = -\sum_{\sigma} \frac{1}{z^2} [f_{z,\epsilon}(S^{\sigma})]^2 + L \frac{1}{(1+z)^2}. \tag{B.16}$$

B.3 Relating Expressions to Those in Main Text

The equations in the main text follow by noting that the sum over σ can be replaced by a sum over expectation value over sequences S^{σ} drawn from the binding site distribution and background DNA. For an arbitrary function, $H(S)$, of a sequence S of length l,

$$\sum_{\sigma} H(S^{\sigma}) = \sum_{\sigma \in BS} H(S^{\sigma}) + \sum_{\sigma \in BG} H(S^{\sigma}) \tag{B.17}$$

$$\approx N \langle H(S) \rangle_{bs} + (L - N) \langle H(S) \rangle_{bg}, \tag{B.18}$$

with N the expected number of binding sites in a sequence of length L, and where $\langle H(S) \rangle_{bg}$ and $\langle H(S) \rangle_{bs}$ are the expectation value of $H(S)$ for sequences S of length l drawn background DNA and binding site distributions, respectively. Combining (B.18) and (14) with the expressions above yields the equations in the main text.

Appendix C: Derivation of the Scaling Relationship

To derive the scaling relationship, we must calculate the quantity

$$r^2(z,\epsilon) = -\sum_{ij\alpha\beta} \epsilon_{i\alpha} \frac{\partial^2 \mathscr{L}(\mathscr{S}|\theta)}{\partial \epsilon_{i\alpha} \partial \epsilon_{j\beta}} \epsilon_{j\beta} - \sum_{i\alpha} \epsilon_{i\alpha} \frac{\partial^2 \mathscr{L}(\mathscr{S}|\theta)}{\partial \epsilon_{i\alpha} \partial_z} z - \frac{\partial^2 \mathscr{L}}{\partial^2 z} z^2, \tag{C.1}$$

where all the second derivatives are evaluated at $\varepsilon = 0$. Plugging (B.7), (B.15), and (B.16) into the expression above, one has

$$\begin{aligned} r^2(z,\epsilon) = & \sum_{\sigma,i} f_{z,\epsilon=0}(S^{\sigma}) \frac{S^{\sigma}_{iT}}{p_T} \left[\bar{\epsilon}_i^2 + \frac{\bar{\epsilon}_i^2}{p_T} \right] \\ & - \sum_{\sigma,i,j} f_{z,\epsilon=0}(S^{\sigma}) (1 - f_{z,\epsilon=0}(S^{\sigma})) \left[\sum_{\alpha} \epsilon_{i\alpha} S^{\sigma}_{i\alpha} - \frac{\bar{\epsilon}_i S^{\sigma}_{iT}}{p_T} \right] \left[\sum_{\beta} \epsilon_{j\beta} S^{\sigma}_{i\alpha} - \frac{\bar{\epsilon}_j S^{\sigma}_{jT}}{p_T} \right] \\ & + \sum_{\sigma,i,\alpha} f_{z,\epsilon=0}(S^{\sigma}) (1 - f_{z,\epsilon=0}(S^{\sigma})) \frac{\epsilon_{i\alpha}}{z} \left(S_{i\alpha} - \frac{p_{i\alpha} S^{\sigma}_{iT}}{p_{iT}} \right) \\ & - \sum_{\sigma} [f_{z,\epsilon=0}(S^{\sigma})]^2 + L \frac{Z^2}{(1+z)^2}, \end{aligned} \tag{C.2}$$

where we have defined

$$\bar{\epsilon}_i^{\gamma} = \sum_{\alpha = A,C,G} p_{i\alpha} \epsilon_{i\alpha}^{\gamma}, \tag{C.3}$$

with $\gamma = 1, 2$ and used the fact that $p_{iT} = p_T$ when $\varepsilon = 0$. When L is large, we can replace the sum over σ by an expectation value in background DNA,

$$\frac{1}{L} \sum_{\sigma} \to \langle \rangle. \tag{C.4}$$

Furthermore,

$$\begin{aligned}\langle S_{iT}\rangle &= p_{iT}\\ \langle S_{i\alpha}S_{j\beta}\rangle &= p_{i\alpha}p_{j\beta}\left(1-\delta_{ij}\right)+p_{i\alpha}\delta_{ij}\delta_{\alpha\beta}\\ \langle S_{i\alpha}S_{jT}\rangle &= p_{i\alpha}p_{jT}\left(1-\delta_{ij}\right)\\ \langle S_{iT}S_{jT}\rangle &= p_{iT}p_{jT}\left(1-\delta_{ij}\right)+p_{iT}\delta_{ij}.\end{aligned} \tag{C.5}$$

Plugging these expressions into (C.2), noting that the third term averages to zero, and simplifying yields

$$r^2(z,\epsilon)=\frac{Lz^2}{(1+z)^2}\left[\sum_i \bar{\epsilon}_i^2+\frac{\bar{\epsilon}_i^{\,2}}{p_T}\right] \tag{C.6}$$

Finally, it is often helpful to define a rescaled version of $r^2(z, \epsilon)$ that makes the dependence of L explicit,

$$\tilde{r}^2(z,\epsilon)\equiv\frac{r^2(\epsilon,z)}{L} \tag{C.7}$$

References

1. Berg OG, von Hippel P. Trends Biochem. Sci. 1988; 13:207. [PubMed: 3079537]
2. Stormo G, Fields D. Trends Biochem. Sci. 1998; 23:109. [PubMed: 9581503]
3. Djordjevic M, Sengupta AM, Shraiman BI. Genome Res. 2003; 13:2381. [PubMed: 14597652]
4. Rajewsky N, Vergassola M, Gaul U, Siggia E. BMC Bioinform. 2002; 3
5. Sinha S, van Nimwegen E, Siggia ED. Bioinformatics. 2003; 19:292.
6. Drawid A, Gupta N, Nagaraj V, Gelinas C, Sengupta A. BMC Bioinform. 2009; 10:208.
7. Kinney JB, Tkaik G, Callan CG. Proc. Natl. Acad. Sci. USA. 2007; 104:501. [PubMed: 17197415]
8. Percus J. J. Stat. Phys. 1976; 15
9. Bishop, C. Pattern Recognition and Machine Learning. 2006.
10. Rabiner L. Proc. IEEE. 1989; 257
11. Schwab DJ, Bruinsma R, Rudnick J, Widom J. Phys. Rev. Lett. 2008; 100:228105. [PubMed: 18643465]
12. Morozov, A.; Fortney, K.; Gaykalova, DA.; Studitsky, V.; Widom, J.; Siggia, E. 2008. arXiv: 0805.4017
13. Baum LE, Petrie T, Soules G, Weiss N. Ann. Math. Stat. 1970; 41:164.
14. Olsen, R.; Bundschuh, R.; Hwa, T. Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology. 1999. p. 211
15. Tanay A, Siggia E. Genome Biol. 2008; 9:37.
16. Jeffreys H. Proc. R. Soc. Lond. Ser. A, Math. Phys. Sci. 1946; 186:453. [PubMed: 20998741]
17. Mahalanobis P. Proc. Natl. Inst. Sci. India. 1936; 2:49–55.
18. Mora T, Walczak A, Bialek W, Callan CG. Proc. Natl. Acad. Sci. USA. 2010; 107:5405. [PubMed: 20212159]
19. Schneidman E, Berry M, Segev R, Bialek W. Nature. 2006; 440:1007. [PubMed: 16625187]
20. Halabi N, Rivoire O, Leibler S, Ranganathan R. Cell. 2009; 138:774. [PubMed: 19703402]
21. Weigt M, White R, Szurmant H, Hoch J, Hwa T. Proc. Natl. Acad. Sci. USA. 2009; 106:67. [PubMed: 19116270]

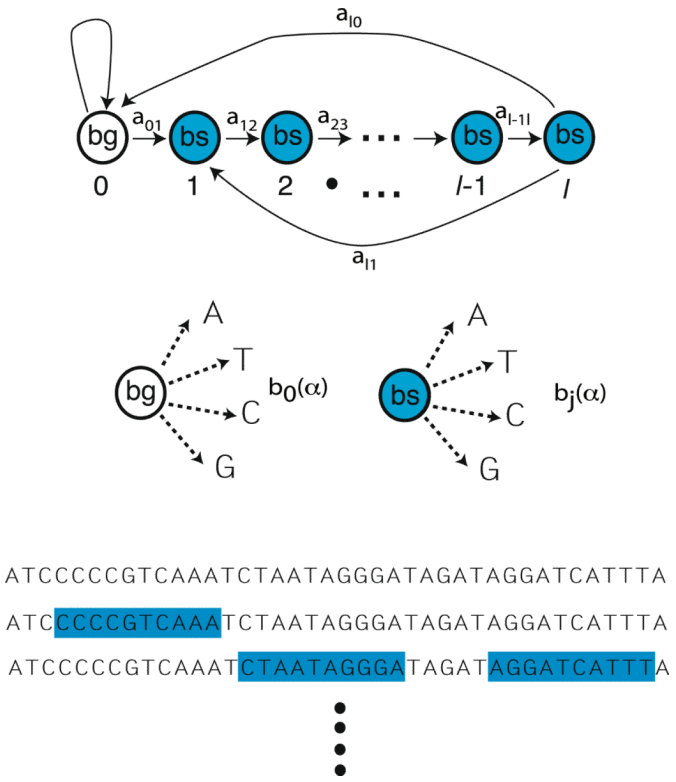

Fig. 1.

Hidden Markov Model for binding sites of size *l*. (*Top*) There are *l* + 1 hidden states, with state 0 background DNA and state *j* corresponding to position *j* = 1 . . . *l* in a binding site. The HMM is described by a Markov process with transition probabilities give by a_{ij}. (*Middle*) Each state *j* in an HMM is characterized by an observation symbol probability b_j (*k*), for the probability of seeing symbol *k* = *A, T , C, G* in a state *j*. (*Bottom*) A given sequence of DNA is composed of binding sites and background DNA

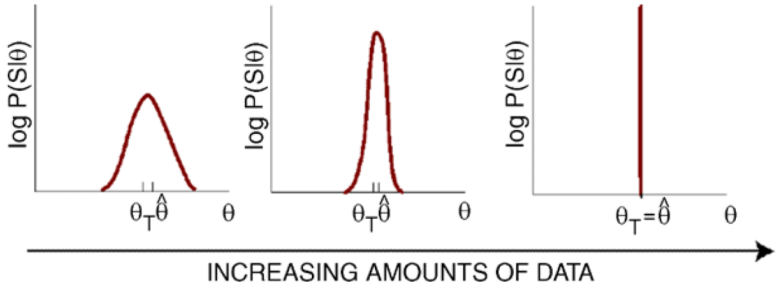

Fig. 2.

The log-likelihood becomes peaked around true parameters with increasing data, analogous to finite-size scaling of pressure times volume (logarithm of the grand canonical partition function) in the corresponding statistical mechanical model

Table 1

Relationship between HMMs and the statistical mechanics of hard-rods

	HMMs	**Hard-rods**
L	Size of training data	System size
S_j^{α}	Nucleotide sequence	Disorder
$b_j(\alpha)$	Symbol Probability	Binding Energy
a_{ij}	Switching rates	Fugacity
$P(S\|\theta)$	Probability	Partition Function
$\log P(S\|\theta)$	Log-likelihood	Pressure
$[I(\theta)]_{ij}$	Fisher Information	Correlation Functions
	Dynamic Prog.	Transfer Matrices
	Expectation Maximization	Variational Methods