Published in final edited form as:
Proteins. 2013 March ; 81(3): 426–442. doi:10.1002/prot.24201.

Three enhancements to the inference of statistical protein-DNA potentials

Mohammed AlQuraishi and **Harley H. McAdams***
Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305

Abstract

The energetics of protein-DNA interactions are often modeled using so-called statistical potentials, that is, energy models derived from the atomic structures of protein-DNA complexes. Many statistical protein-DNA potentials based on differing theoretical assumptions have been investigated, but little attention has been paid to the types of data and the parameter estimation process used in deriving the statistical potentials. We describe three enhancements to statistical potential inference that significantly improve the accuracy of predicted protein-DNA interactions: (i) incorporation of binding energy data of protein-DNA complexes, in conjunction with their X-ray crystal structures, (ii) use of spatially-aware parameter fitting, and (iii) use of ensemble-based parameter fitting. We apply these enhancements to three widely-used statistical potentials and use the resulting enhanced potentials in a structure-based prediction of the DNA binding sites of proteins. These enhancements are directly applicable to all statistical potentials used in protein-DNA modeling, and we show that they can improve the accuracy of predicted DNA binding sites by up to 21%.

Keywords

protein-DNA binding; energy potentials; structural biology; DNA binding sites; DNA motifs; machine learning; biophysics

INTRODUCTION

The binding of proteins to DNA is one of the fundamental molecular interactions in biology; it is central to transcriptional regulation, chromosome replication and segregation, DNA repair, and the formation of nucleosomes.[1,2] Molecular level mathematical models known as energy potentials are used to describe the energies of the molecular interactions underlying protein-DNA binding. Using energy potentials the binding energy of a protein-DNA complex is computed from its atomic structure by summing the energetic contributions of all the molecular interactions observed in the complex. This enables computation of the DNA-binding affinity of a protein, described by a consensus sequence or a position-weight matrix

*Correspondence to: Harley McAdams, Department of Developmental Biology, B300 Beckman Center, Stanford University, Stanford, CA 94305. hmcadams@stanford.edu.

(PWM), from its binding energy to different DNA sequences. Many methods have been developed that utilize this basic approach,[3–7] with recent algorithms showing increasing promise.[8,9]

Protein-DNA energy potentials range in resolution from the atomistic,[4] in which protein-DNA binding is defined in terms of pairwise contacts between protein and DNA atoms, to the residue-level,[10] in which pairwise or higher-order contacts between residues and basepairs are used. Energy potentials in common use also vary in the degree to which they rely on theoretical models versus experimental data for their derivation. “Physical potentials”[11] rely almost exclusively on physics-based theoretical models. We previously reported computation of a *de novo* potential that relies exclusively on experimental data to directly infer the potentials and the energies of molecular interactions.[8] In contrast, “statistical potentials,” the focus of this study, combine experimental data with modeling assumptions to derive an energy model.[12] A key advantage of statistical potentials is their ability to utilize prior physical knowledge about energy potentials, by loosely defining the mathematical form of the potential, while incorporating experimental data to derive the detailed features of the potential. In addition, their wide use in a broad range of biomolecular domains renders them critical in many molecular analysis and simulation pipelines.

The process for inferring statistical potentials begins with two ingredients: a mathematical model, known as the reference state, which captures the expected statistics of random interactions (described in Materials and Methods section), and a data set of experimental measurements to fit the unknown parameters of the model. These two ingredients are combined in a parameter fitting process to obtain the final statistical potential, which assigns a numeric energy value to every possible molecular interaction (e.g., every pairwise atomic contact) at every possible distance (typically discretized to 1–2 Å wide bins).

Many previous studies have focused on the reference state of a statistical potential and its effect on the final inferred potential, with recent models growing increasingly complex.[13] Less attention has been paid to the parameter fitting process and the types of data used in inferring the protein-DNA potentials. Here, we consider three enhancements to the approaches commonly used for inferring statistical potentials. First, we use experimental data on the relative binding affinity of protein-DNA complexes in conjunction with their atomic structures to infer statistical potentials. This is in contrast to the conventional approach that relies only on atomic structures of protein-DNA complexes. Second, we develop a “spatially-aware” parameter fitting process by permitting the statistical potential to vary as a function of position along the protein-DNA binding interface. Third, we employ an ensemble-based approach to the parameter fitting process, which blends a collection of statistical potentials into a single potential. These enhancements are not specific to any particular statistical potential, but instead exploit the biophysical details of protein-DNA binding to improve the quality of the statistical potentials. These enhancements are broadly applicable to the entire field of protein-DNA interactions.

We test these enhancements by incorporating them into the inference of three widely-used atomistic statistical potentials for protein-DNA modeling: the Quasichemical,[4] DFire,[12–15] and μ potentials.[16–18] We use the resulting potentials to predict the DNA binding affinity of

proteins from their cocomplex structures with DNA, and show that the enhanced potentials outperform the unenhanced potentials by up to 21% when predicting the consensus sequences and PWMs of DNA binding sites.

MATERIAL AND METHODS

Derivation of statistical potentials

Underlying theory—Statistical potentials are motivated by a principle from statistical mechanics, which stipulates that frequently observed interactions are energetically favorable.[12] By formalizing this principle mathematically, a statistical potential is constructed from the observed frequencies of molecular interactions in atomic structures of protein-DNA complexes. Exactly how to formalize this into the statistical potential is a key question. In particular, when an interaction is deemed to be more frequently observed than expected, a mathematical formulation of this expectation is necessary, and a baseline known as the "reference state" against which the observed frequencies can be compared must be specified. The resulting formula captures the expectation for the frequency of a molecular interaction based on the assumption that it is a random occurrence. Formally, any statistical potential must obey the following master equation[12]:

$$V(r,i,j) = -RT \ln \frac{N_{\text{obs}}(r,i,j)}{N_{\text{exp}}(r,i,j)}$$

Assuming an atomistic potential, the left hand side $V(r,i,j)$ is the energy of a pairwise interaction between two atoms of type i and type j occurring at distance bin r (atom types and distance bins are discussed in the next section). The right hand side is the mathematical form that the energy must conform to. R is the gas constant, T is the temperature, and $N_{\text{obs}}(r,i,j)$ is the empirically observed frequency of the interaction. These are all fixed quantities or ones inferred from the data. Alternative statistical potentials differ in their formulation of the $N_{\text{exp}}(r,i,j)$ term, which specifies the reference state. We consider three widely-used reference states in this study.

Reference states—The mathematical form of the term $N_{\text{exp}}(r,i,j)$ in the master equation arises from the underlying modeling assumptions of a statistical potential. We consider three statistical potentials that have been used for modeling protein-DNA energetics. The first is the quasichemical potential,[4] based on the assumption that the protein and DNA atoms are randomly shuffled in the reference state. A zero energy interaction is taken to be one such that its observed frequency is the same as its frequency when all the atoms in the structure are shuffled randomly. This is captured by the following reference state equation, where $N_{\text{obs}}(r)$ is the observed number of all interactions occurring at distance bin r, and $x_{\text{obs}}(k)$ is the observed fraction of all atoms that are of type k.

$$N_{\text{exp}}(r,i,j) = N_{\text{obj}}(r)\chi_{\text{obs}}(i)\chi_{\text{obs}}(j)$$

A second potential is DFire,[12–15] which is based on the assumption that the atoms are uniformly distributed and freely mixing in an ideal gas state in the reference state. Corrections are included to account for the finite nature of protein-sized spheres as well as the fact that interactions are short ranged and thus their energy tails off at sufficiently far distances. This is formalized by the following reference state:

$$N_{\text{exp}}(r,i,j)=N_{\text{obs}}(r_{\text{cut}},i,j)\left(\frac{d(r)}{d(r_{\text{cut}})}\right)^{\alpha}\left(\frac{\Delta d(r)}{\Delta d(r_{\text{cut}})}\right)$$

The terms $d(r)$ and $\Delta d(r)$ correspond to the midpoint and width of the distance bin r, respectively. α is an adjustable parameter, set to 1.61 in the DFire potential, which accounts for restricting atoms within protein-sized spheres. Finally, r_{cut} is the cut-off distance bin, that is the distance beyond which atoms are assumed to be noninteracting. To see that the interaction energy in a DFire potential diminishes beyond the cut-off distance, note that when $r = r_{\text{cut}}$, the energy $V(r_{\text{cut},i,j})$ is equal to:

$$V(r_{\text{cut}},i,j)=-RT\ln\frac{N_{\text{obs}}(r_{\text{cut}},i,j)}{N_{\text{obs}}(r_{\text{cut}},i,j)\left(\frac{d(r_{\text{cut}})}{d(r_{\text{cut}})}\right)^{\alpha}\left(\frac{\Delta d(r_{\text{cut}})}{\Delta d(r_{\text{cut}})}\right)}=0$$

A justification of the DFire potential can be found from basic principles of liquid-state statistical mechanics.[14] A third potential is the μ potential,[16–18] a generalization of the Go potential used in protein folding to make the native state the minimally frustrated global minima.[16] The μ potential is formulated as an interaction energy equation:

$$V(r,i,j)=\frac{-\mu(r)N_{\text{obs}}(r,i,j)+(1-\mu(r))(N_{\text{obs}}(i,j)-N_{\text{obs}}(r,i,j))}{\mu(r)N_{\text{obs}}(r,i,j)+(1-\mu(r))(N_{\text{obs}}(i,j)-N_{\text{obs}}(r,i,j))}$$

$N_{\text{obs}}(i,j)$ is the observed number of all interactions (within cut-off distance) occurring between atoms of type i and j. $\mu(r)$ is an adjustable, distance-bin specific parameter, ranging between 0 and 1. The value of $\mu(r)$ for a given distance bin balances between attraction and repulsion, with smaller values favoring a more repulsive potential and larger values favoring a more attractive one. A common approach to choosing $\mu(r)$, used in[4,18] and followed here, is to make the expected value of interaction energy over all possible interaction types zero for a given distance bin r. That is, for each r, $\mu(r)$ is chosen such that:

$$E_{i,j}[V(r,i,j)]=\sum_{i,j}V(r,i,j)p(r,i,j)=0$$

In the above equation $p(r,i,j)$ is the probability of observing atoms of type i and j interacting at distance bin r, assuming a uniform distribution over interaction types.

Representation—In addition to the choice of the reference state, statistical potentials vary in their representation of molecular interactions and spatial distance. Statistical potentials are a list of assignments of numeric energy values to molecular interactions, and a molecular

interaction can range from a pairwise contact between two atoms to a complex multiway interaction involving multiple atoms or residues. In the potentials considered here, a molecular interaction is defined as a pairwise contact between a DNA atom and a protein atom within a specified distance range. Intra-DNA and intra-protein interactions are ignored. Protein and DNA atoms are typed, that is considered to be distinct, based on an existing atom typing scheme that identifies 37 DNA atom types and 27 protein atom types.[4] This typing scheme takes into consideration the chemical identity of the atom and its local chemical moiety. Contact distances are categorized into bins of fixed width. If there are D distance bins, then the model contains a total of $37 \times 27 \times D$ distinct possible interactions, or interaction types, to which the statistical potential must assign a numeric energy value.

The first distance bin is 3 Å in width, ranging from 0 to 3 Å. A more granular binning approach would be preferable, as the interaction energy is likely to be rapidly changing in this regime, but the paucity of available data effectively makes this an obligatory choice. All subsequent bins are 0.5 Å in width, extending up to the final cut-off distance. Thus the second bin ranges from 3 to 3.5 Å, the third from 3.5 to 4 Å, and so on. This binning scheme was shown to yield the best-performing statistical potentials in previous studies,[4] and it insures that the majority (~87%) of interaction types are observed at least once in our data set (see next section). To account for interaction types with zero counts, a count of 1 is added to all interaction types (Laplace smoothing). Finally, since interatomic interactions are short-ranged, a cut-off distance is chosen beyond which all interactions are assumed to have zero energy. This also improves the computational tractability of the problem. As the optimum cut-off distance is not known *a priori*, it is treated as a parameter to be fit on the data.

Training and data sets—Once the reference state of the potential and its representation are chosen, the numerical potential can be inferred by computing all the terms in the master equation. Terms of the form $N_{\text{obs}}(r,i,j)$ and $x_{\text{obs}}(i)$ correspond to empirical frequencies of interactions, which we obtain from an experimental data set containing atomic structures of protein-DNA complexes of sufficient resolution. By determining all interatomic pairwise distances in every available complex, approximate values of $N_{\text{obs}}(r,i,j)$ are directly computed. We obtained a set of protein-DNA complex structures from the Protein Data Bank[19] by searching for X-ray crystal structures that contain a helix-turn-helix (HTH) domain and DNA molecules. We focus on HTH proteins as they are the most widely distributed family of DNA-binding proteins, occurring in all biological kingdoms, with a large number of crystallized structures. HTH proteins include virtually all bacterial transcription factors and about 25% of human transcription factors.[20] From this initial data set, we removed structures that shared the same sequence of amino acids within a 10 residue window of the recognition α helix. We chose this criterion due to the dominant role that recognition α helices play in effecting the sequence specificity of HTH proteins, and the fact that HTHs with otherwise highly similar sequences may still exhibit differential DNA binding properties.[21,22] In addition, we removed complexes with pathologies such as a large number of missing heavy atoms in the published structure. This resulted in a data set of 63 protein-DNA complexes (Table I). On the basis of this data set, we trained and tested the energy potentials in three distinct configurations. In the “full” configuration, the entire data

set was divided into nine nonoverlapping cross-validation (CV) sets. The model was trained on each of the nine CV sets and tested against the subset of data outside the CV set. This helps prevent overfitting. Overall model performance was taken as the average of the model's performance on each of the nine test sets. In the "pruned" configuration, the data set was further pruned by eliminating structures with greater than 25% sequence identity. This resulted in a set of 36 complexes, which were divided into six nonoverlapping CV sets. Training and testing was done in the same way as in the full configuration. While the pruned data set does not capture the differential binding affinity of closely related HTHs, it strongly guards against overfitting on similar protein-DNA complexes. Finally in the "homeodomain" configuration, we divided the full data set into two non-overlapping sets. The first set contained all nonhomeodomain HTHs (45 structures), and served as the training set. The second set contained all homeodomain HTHs (18 structures), and served as the testing set. By eliminating all homeodomains from the training set, this configuration helps test the transferability of the energy potentials to previously unseen domains. Furthermore, the sequence identity between any two proteins in the training and test sets did not exceed 25%, which made the homeodomain configuration useful for guarding against overfitting. A listing of all CV sets and the structures they contain is shown in Table II.

The potential function contains two additional meta-parameters that must be fit: (i) the combined term RT, which we varied from 20^{-3} to 20^{3} using 60 logarithmic increments, and (ii) the cut-off distance, which we varied from 3 to 15 Å in steps of 0.5 Å. For both of these metaparameters the values that maximize model performance on the training set in each configuration were chosen.

Prediction of protein-DNA binding affinity

After inferring a statistical potential we used it for structure-based prediction of protein-DNA binding affinity.[3,4,6,7] We used the atomic structure of the protein complexed with DNA (typically an X-ray crystal complex or one derived through structural modeling) to identify all molecular interactions in the complex by computing the pairwise distances between all atoms. The binding energy of the complex was then computed by summing the individual energetic contributions from the molecular interactions. Formally, if I is the set of all interactions identified in a structure, e_i is the energetic contribution of interaction i, C_i is the number of times interaction i is observed in the structure, then the binding energy of the structure is $\Delta G = \sum_{i \in I} e_i C_i$.

The relative binding affinity of a protein to two different DNA sequences can be evaluated by computing the binding energy of the protein to those two sequences. This is done by mutating the DNA sequence *in silico* while keeping the protein fixed. We used the 3DNA software package for mutating DNA,[23,24] which maintains the backbone atoms of the DNA molecule but replaces the basepair atoms in a way that is consistent with the backbone orientation of the DNA. To make the problem computationally tractable, we assumed that the energetics for any DNA position in the binding interface could be computed separately from other positions, a common assumption in the field.[3–7] While this assumption may be violated in many DNA-binding proteins,[25] the scarcity of data available precludes using

more complex models. Using the computed binding energies, the Boltzmann formula[26] then provides the probability of observing nucleotide *m* at position *n*, denoted by $p_{\mathrm{m}}^{(n)}$:

$$p_{\mathrm{m}}^{(n)} = \frac{e^{-\Delta G_{\mathrm{m}}^{(n)}}}{\sum_{k \in \{A,C,G,T\}} e^{-\Delta G_{k}^{(n)}}}$$

$\Delta G_{\mathrm{m}}^{(n)}$ is the binding energy of the structure when position *n* is mutated to nucleotide *m*. Performing this computation for every position *n* and every nucleotide *m* yields a set of probabilities $\{p_{\mathrm{m}}^{(n)}\}_{1 \leq n \leq N, m \in \{A,C,G,T\}}$. These probabilities are used to identify the most likely DNA sequence that the protein will bind to, by choosing the most probable nucleotide at every position, that is, the predicted consensus binding sequence. Alternatively, the probabilities of alternative bases yield the PWM. We compared the performance of alternative consensus sequence prediction algorithms in terms of the normalized Hamming distance, which is the fraction of incorrectly predicted bases in the consensus sequence. For evaluating performance of PWM predictions, we used the symmetric Kullback–Leibler divergence (SKLD), the most common metric used in comparison of PWMs[27] to measure how close two PWMs are:

$$D_{\mathrm{SKL}}(P \| Q) = \frac{1}{N} \sum_{n=1}^{N} \sum_{m \in \{A,C,G,T\}} p_{\mathrm{m}}^{(n)} \ln \frac{p_{\mathrm{m}}^{(n)}}{q_{\mathrm{m}}^{(n)}} + q_{\mathrm{m}}^{(n)} \ln \frac{q_{\mathrm{m}}^{(n)}}{p_{\mathrm{m}}^{(n)}}$$

P and *Q* are the experimentally-determined PWM and the predicted PWM, respectively, and $p_{\mathrm{m}}^{(n)}$ and $q_{\mathrm{m}}^{(n)}$ are the probabilities of observing nucleotide *m* at position *n* in *P* and *Q*, respectively.

Enhancement 1: training with binding affinity data

Previously reported statistical protein-DNA potentials have been trained using only atomic structures of protein-DNA complexes, from which empirical counts of the form $N_{\mathrm{obs}}(r,i,j)$ are obtained. This approach does not take advantage of experimental biochemical data, especially the experimentally-determined PWMs, which characterize a protein's observed DNA binding specificity. Combining structural data with biochemical data for protein-DNA binding site prediction has been a longstanding objective.[6,28–30] We hypothesized that if biochemical data were incorporated into the training process, more accurate empirical frequencies can be derived from the same number of atomic structures.

Theoretical considerations—In conventional statistical potentials, the empirical frequencies of molecular interactions are inferred by counting the number of times a given interaction is observed across protein-DNA complexes in the data set. This implicitly assumes that all complexes are equally likely, since the weight given to an occurrence of an interaction is the same across all complexes. Specifically, when computing a quantity such as $N_{\mathrm{obs}}(r,i,j)$, the quantity is incremented by one every time an interaction involving atom

types *i* and *j* is observed at distance *r*, regardless of which complex the interaction is observed in. But this is inconsistent with the standard statistical mechanical theory of protein-DNA binding,[26] in which the binding affinity of a protein to different DNA sequences is taken to be positively correlated with the occurrence frequency of the corresponding protein-DNA complexes in biological systems. By treating all complexes as occurring with equal frequency, the inference procedure could be introducing significant bias into the calculations, especially if the binding energy of a protein to different DNA sequences varies greatly.

To address this shortcoming, we use the extensive body of existing biochemical data on the binding affinity of protein-DNA complexes. For our problem, we focus on experimentally-determined protein-DNA PWMs, but quantitative information from other biochemical interactions could also be used. PWMs specify a probability distribution over alternative DNA sites for a given protein. The theoretical rationale behind PWMs is the statistical mechanical concept of the occupancy state.[26] The probabilities specified by a PWM for a protein-DNA interaction represent, to a first approximation, the observed frequency of the protein binding to each DNA sequence, relative to its binding at alternative sites. This suggests incorporating PWM data into the training process, by weighting protein-DNA complexes containing the same protein but different DNA sequences differently, according to their experimental probability of occurrence specified by their corresponding PWM. When computing empirical counts such as $N_{\text{obs}}(r,i,j)$, an interaction observed in a given complex is only given a fractional count, equal to that of the complex's probability of occurrence amongst different DNA sequences.

Model training—To incorporate PWMs into the training process, we modify the process described earlier. Since PWMs by definition are based on the assumption of independence of basepair positions; we assume that each position in a protein-DNA complex can be treated separately. For each base position, *in silico* structural mutants are generated using 3DNA[23,24] to mutate the basepair to include all four possibilities. The independence assumption makes it irrelevant for purposes of the binding energy computation whether two basepair positions come from the same or from different protein-DNA complexes, so we treat each basepair position as a distinct complex, as if it were bound to a distinct protein. This implies that the collection of all of the *in silico* mutants of an individual basepair position constitutes an ensemble whose total weight should be 1, in accord with the assumption that complexes with different proteins are given equal weight. Note that the weight does not necessarily have to be 1, but it has to be the same for all ensembles. Within each of the ensembles, the distribution of weightings among the four *in silico* mutants is determined by the experimentally-derived PWMs. In the development of our data-set, we searched primary sources and online repositories such as TRANSFAC[31] for experimentally-verified DNA binding sites for each of the proteins in our data set. More than 45 repositories and primary sources were used,[7,31–74] providing experimental evidence in the form of footprints, SELEX experiments, and dsDNA microarray assays. The experimentally-verified DNA binding sites for each protein were then aligned and combined into a single PWM, by setting the probability of a given nucleotide in the PWM to its frequency as observed in the set of DNA binding sites. The resulting PWMs serve as the gold standard in our tests. While

we believe these PWMs to be generally accurate, they are ultimately subject to the biases and errors introduced by the experiments from which they are derived. Finally, we align each PWM with its corresponding protein-DNA structural complex,[8] so that for every basepair position in every protein-DNA complex in the data set, we have an empirical probability distribution over the four nucleotides. We assign these probability distributions as weights to the *in silico* mutants.

Enhancement 2: spatially-aware metaparameter fitting

The conventional approach for deriving statistical protein-DNA potentials infers a single potential with a single choice of parameter values. In this section, we generalize this approach to inferring position-specific parameters, i.e. parameters that vary depending on the basepair position (along the protein-DNA binding interface) in which the molecular interaction is occurring. We focus on the cut-off distance and *RT* metaparameters.

Biophysical motivation—Energy potentials are typically defined as interaction energies that are independent of the absolute spatial position within the binding site of the protein-DNA complex, since the particular position is assumed to have no bearing on the energy. However, if the interaction's spatial position is affected by the local physico-chemical environment, any known properties of this environment can be exploited when modeling interaction energies. This is currently done on very small scales, for example by treating two carbon atoms as distinct atom types depending on their chemical moiety, as in the C_1' and C_2' atoms of DNA.[4] In this case, the carbon atom's local chemical moiety is the consistent physico-chemical environment.

For protein-DNA potentials, we generalized this notion to a larger spatial context, by deriving a potential where the interaction energies can vary as a function of position along the protein-DNA binding interface. We treated position in a coarse-grained fashion, by considering distinct basepair positions in a DNA molecule as distinct positions, with no finer spatial distinction. This approach requires that a given relative position along the protein-DNA binding interface exhibits a consistent physico-chemical environment across different protein-DNA complexes, for example consistent steric constraints. We hypothesized that this is the case when a set of proteins employ the same binding modality for docking into DNA, as is often true for members of the same protein family. This property was previously exploited in modeling the binding of zinc finger proteins to DNA.[5] We conjectured that a position-specific potential may also be suitable for the HTH family, since (i) the sequence-specificity of HTHs is largely mediated by interactions between the DNA and the recognition α-helix of the HTHs[75] and (ii) the relative orientation of the two core α helices that make up the HTH domain is conserved across HTH families, despite the broad structural diversity of HTH domains.[75–77]

Information-theoretic motivation—To further investigate the validity of a position-specific potential, we examined our curated data set of HTH-DNA complexes. In this data set, we structurally aligned the complexes so that the DNA molecules and the protein recognition helices are superimposed. Using this alignment we can define a unified coordinate system across all the structures that consists of 13 DNA basepair positions (Fig.

1). In this coordinate system a given position refers to a specific set of DNA basepairs, one in each HTH-DNA complex, in which the same spatial context exists across all HTH-DNA complexes.

To test whether these individual spatial contexts represent a meaningful distinction in terms of protein-DNA binding, we searched for characteristics of the binding behavior that consistently vary across the 13 basepair positions. One such characteristic is the Shannon entropy associated with each position. For a discrete random variable *X* with *n* possible values $\{x_1, \ldots, x_2\}$, the Shannon entropy is:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log\left(p(x_i)\right)$$

where *p* is the probability mass function associated with *X*. The Shannon entropy of a random variable is a measure of its unpredictability. In the context of protein-DNA binding, every basepair position in the binding interface is represented by a random variable that takes the values {A, C, G, T}, and the Shannon entropy associated with a position corresponds to its sequence-specificity with lower values implying higher specificity. For each of the 13 DNA basepair positions, we have a collection of empirical probability distributions from the experimental PWMs in our data set. Each such empirical distribution identifies the Shannon entropy at a given position. The empirical distribution of the Shannon entropies is tight for a given DNA basepair position, but varies across positions in a very consistent manner (Fig. 2). At the 9th position in our 13-basepair coordinate system, the Shannon entropy is lowest, presumably corresponding to the position of closest proximity of the DNA to the protein, since this position makes the largest number of interatomic contacts with the protein and is thus likely to be highly specific. Moving away from this position, the average Shannon entropy increases monotonically, reaching nonspecificity at the ends of the 13-basepair window. This too is expected, as the protein is able to make fewer contacts the farther away the position is from the center of the window. This observation further supports the notion of position-specific potentials, since it indicates that certain properties are invariant for a given DNA basepair position, but are variable across positions.

Basic metaparameter fitting—We fitted the cut-off distance and *RT* metaparameters in a position-dependent manner. While, in principle, it is possible to train the entire statistical potential in a position-specific manner, that is all the empirical frequencies in the master equation, the paucity of available data makes this difficult as we discuss later. Jointly, the two metaparameters describe a landscape of energy potentials in which the height of the landscape corresponds to prediction performance (Fig. 3). Our intuition about the position-specificity of these parameters is supported empirically, since the landscape changes dramatically as a function of position along the protein-DNA binding interface (Fig. 4).

We fit the metaparameters position-specifically as follows: First, the potential was trained by computing the empirical frequencies of interatomic interactions using the entire training set. This was done using all the parameter values of the cut-off distance and *RT* considered in the nonposition specific case. Next, the training set was split into 13 groups, one

corresponding to each DNA basepair position in the unified coordinate system shown in Fig. 1. Since we treat basepair positions independently, this corresponds to grouping the training data along basepair positions. The performance of the potential was then tested separately for each of the 13 groups, and the metaparameter values yielding the best performing model were selected, resulting in 13 sets of metaparameter values and 13 distinct energy potentials. These 13 potentials taken together constitute a position-specific potential. When using this potential to predict the binding energy of a protein-DNA complex, the approach described above in "Prediction of Protein-DNA Binding Affinity" is followed with a slight modification. In the earlier section, we defined the binding energy of the complex as $\Delta G = \sum_{i \in I} e_i C_i$, where e_i is the energetic contribution of interaction i, C_i is the number of times interaction i is observed in the structure, and I is the set of all interaction types. Now, we perform the energy calculations in a position-specific manner, by using the appropriate potential for each DNA basepair position in the protein-DNA complex. Denoting the number of times interaction i occurs in position n by $C_i^{(n)}$, and letting $e_i^{(n)}$ be the energy of interaction i in the nth potential, the binding energy becomes:

$$\Delta G = \sum_{n=1}^{13} \sum_{i \in I} e_i^{(n)} C_i^{(n)}$$

Using this approach, the regular parameters of the master equation are fit using the entire training set, but the metaparameters are fit using data from each basepair position separately.

Windowed metaparameter fitting—When fitting metaparameters in a position-specific manner, the amount of data available for fitting each position is reduced by a factor of 13 relative to the nonposition specific fitting. To maintain the position-specificity of the metaparameter fitting procedure but utilize a larger amount of data, we also considered a windowed approach in which data from adjacent basepair positions were pooled. Unlike in the previous section where the data used in fitting meta-parameters came from a single basepair position, this produces a window of size $2w + 1$ around the basepair position. For example, when fitting the parameters of the energy potential for position n, we evaluated the potential's performance on the training data from positions $n - w$ to position $n + w$, and choose the set of parameters that maximize performance on those data points (when the window extends beyond the DNA sequence it is truncated). This alleviates the problem of limited data, but maintains the position-specificity of the model. Fitting the metaparameters in this way introduces a new metaparameter w that controls the window size. Fitting w using only the training data would always result in its being assigned the value zero, since in that scenario the metaparameters will be trained and tested on the exact same data set, and thus no other value of w can even in principle perform better than $w = 0$. Consequently, the value of the metaparameter w, which ranges from 0 to 12, is fit on the test set.

Enhancement 3: ensemble-based metaparameter fitting

Prior approaches to fitting the metaparameters of energy potentials find a single set of values that maximize prediction performance.[3–8] In this section, we consider an approach in which

an ensemble of metaparameter values are chosen that are well-performing but that do not necessarily achieve the absolute maximum performance. This metaparameter values ensemble is used to derive an ensemble of energy potentials, which in turn are used to predict an ensemble of consensus sequences or PWMs. A single consensus sequence or PWM can then be derived using a weighted average of the individual members of the ensemble. Such ensemble-based methods have found considerable success in machine learning,[78,79] as they are robust against overfitting.

Statistical motivation—As previously described, the cut-off distance and *RT* metaparameters can be viewed as describing a landscape of energy potentials in which the height of the landscape at a given point corresponds to the prediction performance of the respective energy potential (Fig. 3). In particular, the lowest point in this landscape corresponds to lowest error, that is the best performing potential. In our training and testing approach, in which the data set is split into multiple CV sets, we could assess the stability of this landscape as the training data set was changed. The landscape as a whole was robust across the training sets (full configuration), but the lowest point in the landscape was highly variable between training sets [Fig. 5(A)]. This variability suggests that the precise location of the best performing model, where location corresponds to the choice of metaparameter values, is an artifact of the data set used for training. Consequently, a more robust approach to deriving energy potentials is to use the entire landscape of metaparameters, which exhibits little variability across training sets.

Ensemble-based fitting—Utilizing the entire metaparameter landscape of a statistical potential entails the use of an ensemble approach as follows: instead of picking the single lowest point in the landscape to select the metaparameter values for one energy potential, we pick a collection of points, according to some criteria, and then we use the resulting ensemble of energy potentials to predict DNA binding sites. We choose the m best performing models to generate m putative predictions (consensus sequences or PWMs). Since eventually a single prediction is required, we take a weighted average of the m predictions. Formally for the case of PWMs, if p is the final probability of a certain basepair at a certain position, and p_i is the probability of the same basepair at the same position as predicted by the ith best performing model, then:

$$p=\sum_{i=1}^{m} w_i p_i$$

where w_i is the weight assigned to the ith model. A model's weight is equal to the normalized inverse of its score on the training data. Formally, letting sc_i be the score of the ith model (Hamming distance or SKLD), then:

$$w_i=\frac{\frac{1}{sc_i}}{\sum_{j=1}^{m}\frac{1}{sc_j}}$$

This approach makes use of the *m* best performing models in the landscape to make predictions, so that we hedge against choosing a single unstable set of metaparameter values. In this approach, *m* is a metaparameter similar to the *w* metaparameter introduced in the second enhancement. We fit *m* by choosing the value that maximizes prediction performance on the training data.

RESULTS

Prediction of Consensus Sequences

We first tested the impact of the respective enhancements by evaluating the performance of statistical potentials in predicting consensus sequences, using the normalized Hamming distance (see Materials and Methods section). We tested every potential using all possible combinations of the three enhancements on all testing configurations. Table III summarizes the results. The performance impact of each enhancement is somewhat dependent on the testing configuration. In the full configuration, every enhancement individually improved every potential, except for the combination of the μ potential with quantitative binding affinity data. This is confirmed by examining the best performing version of each potential; the Quasichemical and DFire potentials exhibit their best performance when all three enhancements are enabled, while the μ potential performs best when spatially-aware metaparameter fitting and ensemble-based metaparameter fitting are enabled. Training potentials with quantitative binding affinity data is unlikely to improve their performance in predicting consensus sequences, since consensus sequences only encode the most likely DNA binding sequence, ignoring the probabilistic and quantitative nature of protein-DNA interactions. In fact, training with quantitative binding affinity data may decrease performance (as observed for the μ potential), since it lowers the certainty with which proteins are predicted to bind their consensus sequences.

The situation is the same with the pruned testing configuration, except for the Quasichemical potential which only improves when using spatially-aware metaparameter fitting. The degraded performance of the Quasichemical potential when combined with the other enhancements may be due to the lack of sufficient statistics in the pruned testing configuration; this is supported by its performance in the homeodomain configuration, which has greater statistical power than the pruned configuration but maintains the same sequence distance between the training and testing sets. More generally, in the homeodomain testing configuration, binding affinity data and spatially-aware metaparameter fitting improve the performance of all potentials tested, when enabled individually and in combination, with the best performance of each potential achieved when both enhancements are enabled. These results suggest that ensemble-based metaparameter fitting results in potentials that are specific to the domain family used in training, and thus while it generally improves performance within the domain family, it also limits the transferability of the potential.

The potential most improved by the enhancements was the DFire potential. When used with all three enhancements enabled, the percent reduction in Hamming distance is 20.8%, 21.7%, and 16.4% for the full, pruned, and homeodomain testing configurations, respectively. The enhancement that provided the most improvement when applied

individually was spatially-aware metaparameter fitting, which on average (across testing configurations) reduced the percent error in Hamming distance by 8.9%, 8.3%, and 4.8% for the DFire, μ, and Quasichemical potentials, respectively. In addition, synergistic effects were observed when multiple enhancements were combined. For the Quasichemical and DFire potentials, the contributions are almost perfectly additive, suggesting that the enhancements are independently contributing to the improved performance of the potentials. In absolute terms, the Quasichemical potential with all three enhancements enabled performed the best (63.2% accuracy) on the full testing configuration. A bar chart showing the normalized Hamming distance scores for all predictions made using this potential is shown in Figure 6(A). The bar chart comprehensively captures the performance of the best performing potential for predicting consensus sequences, highlighting its predictions in the best, average, and worst cases.

Prediction of PWMs

We next tested the impact of the three enhancements on predicting PWMs. As in the consensus sequence case, we tested all possible combinations of the three enhancements on all testing configurations. The results are summarized in Table IV. In general all three enhancements improved results regardless of the potential and testing configuration used, with potentials achieving their best performance when all three enhancements are enabled. However, there are some important exceptions. Training with binding affinity data significantly worsened performance for the Quasichemical potential in the pruned and homeodomain testing configurations. This suggests that for the Quasichemical potential, training with binding affinity data led to overfitting. It is unclear why only the Quasichemical potential suffers from this problem. Using binding affinity data with the μ potential slightly worsened performance in the pruned configuration, but that is likely due to weak statistical power, as it improved performance in both the full and homeodomain testing configurations.

As in the case of predicting consensus sequences, the potential most improved by the enhancements was the DFire potential. When used with all three enhancements enabled, the percent reduction in SKLD was 7.5%, 7.0%, and 11.7% for the full, pruned, and homeodomain testing configurations, respectively. The enhancement that provided the most improvement when applied individually was also spatially-aware metaparameter fitting, which on average (across testing configurations) reduced the SKLD by 5.6%, 3.9%, and 3.4% for the μ, Quasichemical, and DFire potentials, respectively. As with the consensus sequence prediction, combining multiple enhancements gave nearly additive improvements, suggesting that the enhancements are independently contributing to improved performance. In absolute terms, the μ potential with all three enhancements enabled performed the best (SKLD of 1.998) on the full testing configuration. A bar chart showing the SKLD scores for all predictions made using this potential is shown in Figure 6(B). The figure captures the performance of the best performing potential for predicting PWMs, highlighting its predictions in the best, average, and worst cases.

To further ascertain the individual contributions of the three enhancements, we examined whether an enhancement improved or worsened results in each of the instances in which it

was used. For a given statistical potential and testing configuration, each enhancement was used four times for consensus predictions and four times for PWM predictions. Figure 7 summarizes the number of times an enhancement improved or worsened results for each potential and testing configuration. Spatially-aware metaparameter fitting performed the best, improving results in every test it was subjected to. Training with binding affinity data and ensemble-based metaparameter fitting showed more mixed results, generally improving performance but occasionally degrading it.

Comparison to *De Novo* Potentials

The results obtained when combining all three enhancements represent state of the art performance for statistical potentials in predicting consensus sequences and PWMs. We previously reported a *de novo* potential[8] that outperforms the statistical potentials tested here over the same test data set (0.101 Hamming distance and 1.699 SKLD). An important advantage of statistical potentials over other types of potentials, including *de novo* potentials, is their synthesis of prior modeling assumptions with experimental data. While inaccurate modeling assumptions can cause statistical potentials to perform poorly, accurate modeling assumptions can enable statistical potentials to use less data than *de novo* potentials, since the modeling assumptions help constrain the space of models to consider. Furthermore, our focus in this study is not the absolute performance of the statistical potentials tested, but their relative improvement after the three enhancements are applied. The improvement reported here for the Quasichemical potential in the full testing configuration (21% reduction in Hamming distance and 7.7% reduction in SKLD) is comparable to the largest difference between the unenhanced potentials tested (DFire vs. μ potential, 24.9% difference in Hamming distance and 12.3% difference in SKLD). As new statistical potentials are developed, the three enhancements described here can be immediately applied to improve performance.

DISCUSSION

We analyzed the behavior of each enhancement individually to determine how each enhancement affects model performance. We focused on the Quasichemical potential applied to PWM prediction on the full testing configuration to streamline the analysis, but the conclusions are generalizable to other performance metrics and to other statistical potentials.

Spatially-Aware Metaparameter Fitting

Spatially-aware metaparameter fitting had the largest impact on prediction performance; however, the magnitude of the impact depended on the window size parameter w discussed in Materials and Methods section. If w is too small, the statistical power of the fitting process is too weak to infer an accurate model. On the other hand, if w is too large, the inferred potential loses its spatial-awareness. Figure 8(A) shows the performance of the Quasichemical potential as a function of w. The choice of $w = 1$, corresponding to a three basepair window size, produced the best trade-off between statistical power and model complexity.

We examined two quantities to examine the basis of this behavior. The first is the variance, across CV sets, of the optimal values for the two metaparameters *RT* and cut-off distance. If the model is robust, we expect that the optimal metaparameter values for all CV sets would be the same. On the other hand, if the model is not robust, for example if it is overfitting on each CV set because of low statistical power, then we expect the metaparameter values to change dramatically between CV sets. We further expect the variance across CV sets to decrease as a function of *w*, as statistical power is increased when *w* is increased. In Figure 8(B), this expectation is validated, with the variance of optimal metaparameter values generally decreasing as *w* increases. In general, we want to minimize this variance.

The second quantity we examined is the variance in optimal metaparameter values across DNA basepair positions. In our approach the metaparameters are fit separately for each DNA basepair position. The variance across positions indicates how specialized the position-specific potentials are to their respective spatial contexts. Higher variance corresponds to higher model complexity, and we expect this value to decrease as *w* increases, since increasing *w* pools more and more of the data for each DNA basepair position, lowering model complexity. In Figure 8(C) we plot the variance across positions as a function of *w*, and as expected the value of the variance decreases as *w* increases. In general, we want to maximize this variance to create the most position-specific potentials possible.

The two quantities just described represent the tradeoff between statistical power and model complexity. In Figure 8(D) we plot their quotient as a function of *w*. We want to maximize this quotient as that simultaneously maximizes the second quantity and minimizes the first. As shown, a value of $w = 1$ does that, consistent with Figure 8(A) in which $w = 1$ maximized prediction performance.

Ensemble-Based Metaparameter Fitting

As with spatially-aware metaparameter fitting, the performance impact of ensemble-based metaparameter fitting depends on the value of its metaparameter *m* (see Materials and Methods section), which controls the number of models used in the ensemble. Furthermore, similar to the behavior of the metaparameter *w*, small values of *m* yield ensemble models that are too complex and prone to overfitting, while large values of *m* yield ensemble models that are not sufficiently complex. Figure 9 depicts the performance of the Quasichemical potential as a function of *m*. The overall pattern is that small and large values of *m* yield poor performance, with an optimal valley around a middle-ranged value of $m = 40$. As discussed in the Materials and Methods section, the shape of the metaparameter landscape as a whole is robust across training sets, while the single point that minimizes the landscape is variable [Figure 5(A)]. In Figure 5(B) we plot the locations of the optimum meta-parameter values for the top 40 performing models for the parameter landscapes. Consistent with our expectations, the collection of 40 points as a whole is robust, explaining the improved performance obtained.

Limitations of Training With Binding Affinity Data

The incorporation of PWM data in the inference process of energy potentials enables us to account for the varying binding affinity of a single protein to different DNA sequences, but it does not account for the varying binding affinity of different proteins to the same DNA sequence. This limitation is effectively forced on our model due to the relative unavailability of quantitative experimental data on binding energies of protein-DNA complexes. However, when such data becomes available, the approach we outlined in the Materials and Methods section can be extended to incorporate them. Another drawback of the use of PWMs is the independence assumption, which forces protein-DNA complexes to be energetically uncoupled across different basepair positions. This limitation can also be overcome with the availability of additional experimental data, which comprehensively sample a protein's relative binding affinity to every possible DNA sequence.[80,81] Use of such data will enable assignment of an experimental probability to every distinct DNA sequence. Finally, we note that the approach outlined in Materials and Methods section can also be extended to the derivation of statistical potentials for protein–protein interactions and protein folding, contingent on the availability of the necessary experimental data.

Machine Learning and Energy Potentials

We have shown that improvements in the accuracy of statistical potentials do not have to come from improvements in the mathematical models (reference states), which underlie statistical potentials. Instead, by focusing on the inference process used in deriving statistical potentials, an entirely complementary approach for improving their accuracy is available. This approach uses concepts from machine learning, such as model robustness, to guide the inference of statistical potentials, and it is amenable to continual improvement as new discoveries are made in machine learning. In addition, the enhancements we describe can be applied to other types of potentials, such as physical potentials, and other molecular domains such as protein–protein interactions.

Acknowledgments

The authors thank the anonymous reviewers for their helpful feedback, and they also thank K. Arya and G. Cooperman for customizing the DMTCP check pointing software for our purposes. Wolfram Research provided the Mathematica software environment necessary for the analyses performed.

Grant sponsor: U.S. Department of Energy Office of Science; Grant number: DE-FG02-05ER64136; Grant sponsor: National Human Genome Research Institute, the Stanford Genome Training Program; Grant number: T32 HG00044; Grant sponsor: the National Energy Research Scientific Computing Center, Office of Science of the U.S. Department of Energy; Grant number: DE-AC02-05CH11231

References

1. Latchman, DS. Gene control. Vol. 16. New York: Garland Science; 2010. p. 430
2. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Molecular biology of the cell. Vol. 33. New York: Garland Science; 2008. p. 1601p. 1690
3. Angarica VE, Perez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B. Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinform. 2008; 9:436.
4. Donald JE, Chen WW, Shakhnovich EI. Energetics of protein-DNA interactions. Nucleic Acids Res. 2007; 35:1039–1047. [PubMed: 17259221]

5. Kaplan T, Friedman N, Margalit H. Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol. 2005; 1:e1. [PubMed: 16103898]
6. Moroni E, Caselle M, Fogolari F. Identification of DNA-binding protein target sequences by physical effective energy functions: free energy analysis of *lambda* repressor-DNA complexes. BMC Struct Biol. 2007; 7:61. [PubMed: 17900341]
7. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res. 2005; 33:5781–5798. [PubMed: 16246914]
8. AlQuraishi M, McAdams HH. Direct inference of protein–DNA interactions using compressed sensing methods. Proc Natl Acad Sci U S A. 2011; 108:14819–14824. [PubMed: 21825146]
9. Pande VS. (Compressed) sensing and sensibility. Proc Natl Acad Sci U S A. 2011; 108:14713–14714. [PubMed: 21873202]
10. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. Proteins. 1999; 35:114–131. [PubMed: 10090291]
11. Jorgensen WL, Tirado-Rives J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc Natl Acad Sci U S A. 2005; 102:6665–6670. [PubMed: 15870211]
12. Zhou Y, Zhou HY, Zhang C, Liu S. What is a desirable statistical energy function for proteins and how can it be obtained? Cell Biochem Biophys. 2006; 46:165–174. [PubMed: 17012757]
13. Zhao HY, Yang YD, Zhou YQ. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics. 2010; 26:1857–1863. [PubMed: 20525822]
14. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Prot Sci. 2002; 11:2714–2726.
15. Xu BS, Yang YD, Liang HJ, Zhou YQ. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. Prot Struct Funct Bioinform. 2009; 76:718–730.
16. Kussell E, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. Proc Natl Acad Sci U S A. 2002; 99:5343–5348. [PubMed: 11943859]
17. Chen WW, Shakhnovich EI. Lessons from the design of a novel atomic potential for protein folding. Prot Sci. 2005; 14:1741–1752.
18. Hubner IA, Deeds EJ, Shakhnovich EI. High-resolution protein folding with a transferable potential. Proc Natl Acad Sci U S A. 2005; 102:18914–18919. [PubMed: 16365306]
19. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000; 28:235–242. [PubMed: 10592235]
20. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009; 10:252–263. [PubMed: 19274049]
21. Gajiwala KS, Burley SK. Winged helix proteins. Curr Opin Struct Biol. 2000; 10:110–116. [PubMed: 10679470]
22. Mo Y, Vaessen B, Johnston K, Marmorstein R. Structure of the elk-1-DNA complex reveals how DNA-distal residues affect ETS domain recognition of DNA. Nat Struct Biol. 2000; 7:292–297. [PubMed: 10742173]
23. Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucl Acids Res. 2003; 31:5108–5121. [PubMed: 12930962]
24. Lu XJ, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008; 3:1213–1227. [PubMed: 18600227]
25. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen XY, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324:1720–1723. [PubMed: 19443739]

26. Berg OG, Vonhippel PH. Selection of DNA-binding sites by regulatory proteins—statistical-mechanical theory and application to operators and promoters. J Mol Biol. 1987; 193:723–743. [PubMed: 3612791]
27. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951; 22:79–86.
28. Mirny LA, Gelfand MS. Structural analysis of conserved base pairs in protein-DNA complexes. Nucleic Acids Res. 2002; 30:1704–1711. [PubMed: 11917033]
29. Hoglund A, Kohlbacher O. From sequence to structure and back again: approaches for predicting protein-DNA binding. Proteome Sci. 2004; 2:3. [PubMed: 15202939]
30. Eisen M. All motifs are NOT created equal: structural properties of transcription factor-DNA interactions and the inference of sequence specificity. Genome Biol. 2005; 6:7.
31. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34(Database issue):D108–D110. [PubMed: 16381825]
32. Kazakov AE, Cipriano MJ, Novichkov PS, Minovitsky S, Vinogradov DV, Arkin A, Mironov AA, Gelfand MS, Dubchak I. RegTrans-Base—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Res. 2007; 35:D407–D412. [PubMed: 17142223]
33. Halfon MS, Gallo SM, Bergman CM. REDfly 2. 0: an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*. Nucleic Acids Res. 2008; 36(Database issue):D594–D598. [PubMed: 18039705]
34. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010; 38(Database issue):D105–D110. [PubMed: 19906716]
35. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2009; 37:D77–D82. [PubMed: 18842628]
36. Munch R, Hiller K, Barg H, Heldt D, Linz S, Wingender E, Jahn D. PRODORIC: prokaryotic database of gene regulation. Nucleic Acids Res. 2003; 31:266–269. [PubMed: 12519998]
37. Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Contreras-Moreira B, Segura-Salazar J, Muniz-Rascado L, Martinez-Flores I, Salgado H, Bonavides-Marti-nez C, Abreu-Goodger C, Rodriguez-Penagos C, Miranda-Rios J, Morett E, Merino E, Huerta AM, Trevino-Quintanilla L, Collado-Vides J. Regulon DB (version 6. 0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res. 2008; 36(Database issue):D120–D124. [PubMed: 18158297]
38. Sierro N, Makita Y, de Hoon M, Nakai K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. Nucleic Acids Res. 2008; 36(Database issue):D93–D96. [PubMed: 17962296]
39. Jagannathan V, Roulet E, Delorenzi M, Bucher P. HTPSELEX—a database of high-throughput SELEX libraries for transcription factor binding sites. Nucleic Acids Res. 2006; 34(Database issue):D90–D94. [PubMed: 16381982]
40. Down TA, Bergman CM, Su J, Hubbard TJ. Large-scale discovery of promoter motifs in *Drosophila melanogaster*. PLoS Comput Biol. 2007; 3:e7. [PubMed: 17238282]
41. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grote-wold E. AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol. 2006; 140:818–829. [PubMed: 16524982]
42. Bulow L, Engelmann S, Schindler M, Hehl R. AthaMap, integrating transcriptional and post-transcriptional data. Nucleic Acids Res. 2009; 37(Database issue):D983–D986. [PubMed: 18842622]
43. Kumar MDS, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006; 34:D204–D206. [PubMed: 16381846]

44. Yellaboina S, Ranjan S, Chakhaiyar P, Hasnain SE, Ranjan A. Prediction of DtxR regulon: identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*. BMC Microbiol. 2004; 4:38. [PubMed: 15447793]
45. Franks AH, Griffiths AA, Wake RG. Identification and characterization of new DNA-replication terminators in *Bacillus subtilis*. Mol Microbiol. 1995; 17:13–23. [PubMed: 7476199]
46. Griffiths AA, Wake RG. Search for additional replication terminators in the *Bacillus subtilis* 168 chromosome. J Bacteriol. 1997; 179:3358–3361. [PubMed: 9150236]
47. Griffiths AA, Andersen PA, Wake RG. Replication terminator protein-based replication fork-arrest systems in various *Bacillus* species. J Bacteriol. 1998; 180:3360–3367. [PubMed: 9642188]
48. Sugisaki H, Kanazawa S. New restriction endonucleases from *Flavo-bacterium okeanokoites* (FokI) and *Micrococcus luteus* (MluI). Gene. 1981; 16:73–78. [PubMed: 6282705]
49. Falvey E, Grindley NDF. Contacts between gamma-delta-resolvase and the gamma-delta-res site. Embo J. 1987; 6:815–821. [PubMed: 3034611]
50. Moskowitz IP, Heichman KA, Johnson RC. Alignment of recombination sites in Hin-mediated site-specific DNA recombination. Genes Dev. 1991; 5:1635–1645. [PubMed: 1885005]
51. Rosandic M, Paar V, Basar I, Gluncic M, Pavin N, Pilas I. CENP-B box and pJ alpha sequence distribution in human alpha satellite higher-order repeats (HOR). Chromosome Res. 2006; 14:735–753. [PubMed: 17115329]
52. Tronche F, Yaniv M. HNF1, a homeoprotein member of the hepatic transcription regulatory network. Bioessays. 1992; 14:579–587. [PubMed: 1365913]
53. Liston DR, Johnson PJ. Analysis of a ubiquitous promoter element in a primitive eukaryote: early evolution of the initiator element. Mol Cell Biol. 1999; 19:2380–2388. [PubMed: 10022924]
54. Shen WF, Montgomery JC, Rozenfeld S, Moskow JJ, Lawrence HJ, Buchberg AM, Largman C. AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. Mol Cell Biol. 1997; 17:6448–6458. [PubMed: 9343407]
55. Kostelidou K, Thomas CM. The hierarchy of KorB binding at its 12 binding sites on the broad-host-range plasmid RK2 and modulation of this binding by IncC1 protein. J Mol Biol. 2000; 295:411–422. [PubMed: 10623535]
56. Garcia-Castellanos R, Mallorqui-Fernandez G, Marrero A, Potempa J, Coll M, Gomis-Ruth FX. On the transcriptional regulation of methicillin resistance-MecI repressor in complex with its operator. J Biol Chem. 2004; 279:17888–17896. [PubMed: 14960592]
57. Colloms SD, van Luenen HG, Plasterk RH. DNA binding activities of the Caenorhabditis elegans Tc3 transposase. Nucleic Acids Res. 1994; 22:5548–5554. [PubMed: 7838706]
58. Prakash P, Yellaboina S, Ranjan A, Hasnain SE. Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of *Mycobacterium tuberculosis* open reading frames. Bioinformatics. 2005; 21:2161–2166. [PubMed: 15746274]
59. Wilson DS, Guenther B, Desplan C, Kuriyan J. High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. Cell. 1995; 82:709–719. [PubMed: 7671301]
60. Hughes KT, Gaines PCW, Karlinsey JE, Vinayak R, Simon MI. Sequence-specific interaction of the Salmonella Hin recombinase in both major and minor grooves of DNA. EMBO J. 1992; 11:2695–2705. [PubMed: 1628628]
61. Hoey T, Levine M. Divergent homeo box proteins recognize similar DNA sequences in Drosophila. Nature. 1988; 332:858–861. [PubMed: 2895896]
62. White CE, Winans SC. The quorum-sensing transcription factor TraR decodes its DNA binding site by direct contacts with DNA bases and by detection of DNA flexibility. Mol Microbiol. 2007; 64:245–256. [PubMed: 17376086]
63. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004; 431:99–104. [PubMed: 15343339]
64. Chen SF, Gunasekera A, Zhang XP, Kunkel TA, Ebright RH, Berman HM. Indirect readout of DNA sequence at the primary-kink site in the CAP-DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. J Mol Biol. 2001; 314:75–82. [PubMed: 11724533]

65. Koudelka GB, Lam CY. Differential recognition of OR1 and OR3 by bacteriophage 434 repressor and Cro. J Biol Chem. 1993; 268:23812–23817. [PubMed: 8226917]
66. Koudelka GB, Harrison SC, Ptashne M. Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. Nature. 1987; 326:886–888. [PubMed: 3553960]
67. Schumacher MA, Lau AOT, Johnson PJ. Structural basis of core promoter recognition in a primitive eukaryote. Cell. 2003; 115:413–424. [PubMed: 14622596]
68. Smale ST, Jain A, Kaufmann J, Emami KH, Lo K, Garraway IP. The initiator element: a paradigm for core promoter heterogeneity within metazoan protein-coding genes. Cold Spring Harb Symp Quant Biol. 1998; 63:21–31. [PubMed: 10384267]
69. Lo K, Smale ST. Generality of a functional initiator consensus sequence. Gene. 1996; 182:13–22. [PubMed: 8982062]
70. Javahery R, Khachi A, Lo K, Zenziegregory B, Smale ST. DNA-sequence requirements for transcriptional initiator activity in mammalian-cells. Mol Cell Biol. 1994; 14:116–127. [PubMed: 8264580]
71. Huerta AM, Francino MP, Morett E, Collado-Vides J. Selection for unequal densities of sigma(70) promoter-like signals in different regions of large bacterial genomes. Plos Genet. 2006; 2:1740–1750.
72. Fischer SEJ, van Luenen HGAM, Plasterk RHA. Cis requirements for transposition of Tc1-like transposons in *C. elegans*. Mol Gen Genet. 1999; 262:268–274. [PubMed: 10517322]
73. Rodgers DW, Harrison SC. The complex between phage 434 repressor DNA-binding domain and operator site OR3: structural differences between consensus and non-consensus half-sites. Structure. 1993; 1:227–240. [PubMed: 8081737]
74. van Luenen HGAM, Plasterk RHA. Target site choice of the related transposable elements Tc1 and Tc3 of Caenorhabditis elegans. Nucleic Acids Res. 1994; 22:262–269. [PubMed: 8127662]
75. Wintjens R, Rooman M. Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. J Mol Biol. 1996; 262:294–313. [PubMed: 8831795]
76. Suzuki M, Gerstein M. Binding geometry of alpha-helices that recognize DNA. Prot Struct Funct Genet. 1995; 23:525–535.
77. Pabo CO, Nekludova L. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol. 2000; 301:597–624. [PubMed: 10966773]
78. Bell RM, Koren Y. Lessons from the netflix prize challenge. SIGKDD Explor Newsl. 2007; 9:75–79.
79. Dietterich TG. Ensemble methods in machine learning. Lect Notes Comput Sci. 2000; 1857:1–15.
80. Bulyk ML, Gentalen E, Lockhart DJ, Church GM. Quantifying DNA-protein interactions by double-stranded DNA arrays. Nat Biotechnol. 1999; 17:573–577. [PubMed: 10385322]
81. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev Genet. 2010; 11:751–760. [PubMed: 20877328]

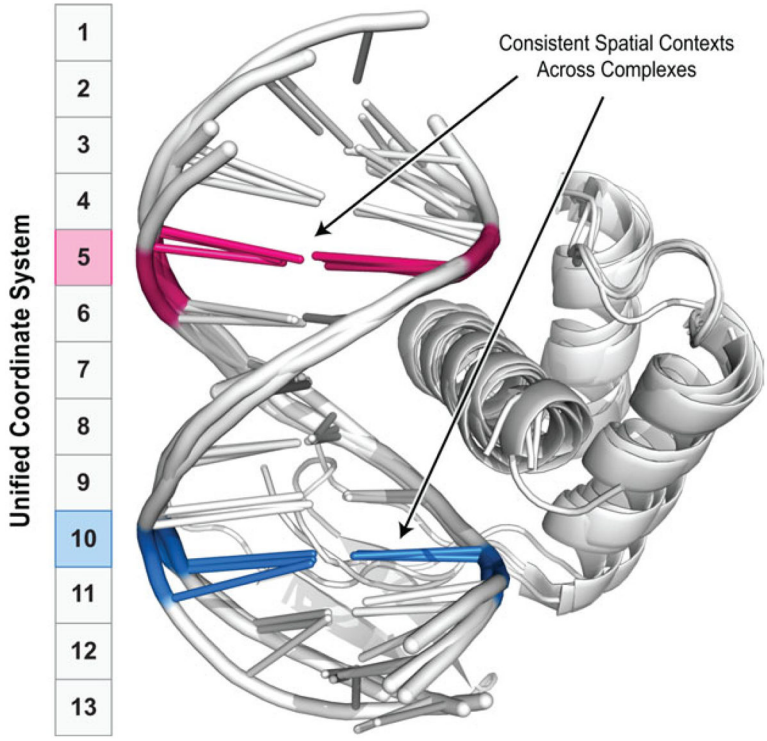

Figure 1.

Unified coordinate system across HTH-DNA complexes. On the basis of the structural alignment of the 63 HTH-DNA complexes in our data set, we created a unified coordinate system consisting of 13 DNA basepair positions (shown schematically as a column of numbered squares). A position in this coordinate system corresponds to a set of basepairs, one in each of the 63 HTH-DNA complexes, with a consistent spatial context. The DNA and HTH domains of four such complexes are shown, with positions 5 and 10 highlighted in pink and blue, respectively. Because of variability in length, some complexes have fewer than 13 positions.

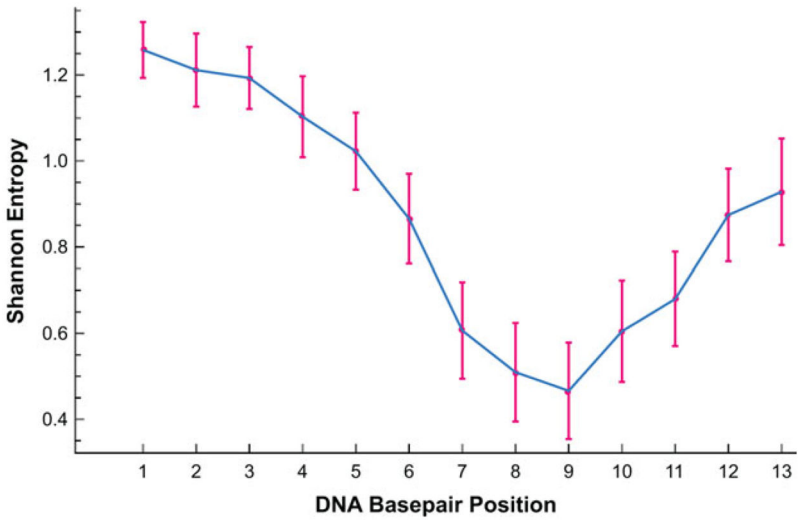

Figure 2.

Shannon entropy as a function of DNA basepair position in unified coordinate system. Blue line depicts the mean Shannon entropy at each DNA basepair position as derived from our experimental data set. Pink bars represent 95% confidence intervals. A consistent pattern is evident where the highest information content is concentrated around position 9, which is generally closest to the recognition helix, and decreases monotonically when moving away from the center toward the edges. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

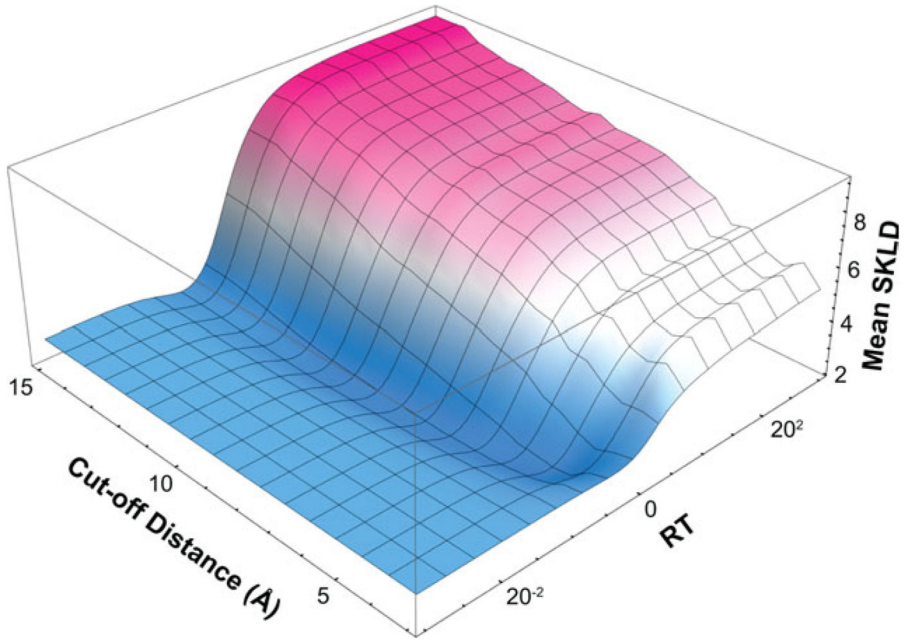

Figure 3.

Metaparameter landscape of energy potentials. The two values parameterizing an energy potential (cut-off distance and *RT*) are plotted along the vertical and horizontal axes, respectively. The height and color of the landscape represent the performance of the energy potential, as measured by mean SKLD, at a given choice of parameter values (blue is best performance, pink is worst performance). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

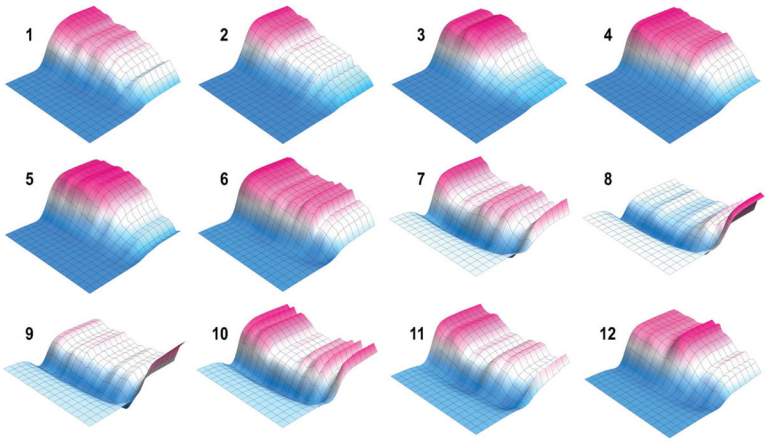

Figure 4.

Variability in metaparameter landscape as a function of basepair position. A sequence of metaparameter landscapes, for positions 1 through 12 in the unified coordinate system, is shown for spatially-aware energy potentials. The landscapes are largely constant for positions 1 through 6 and 11 through 13 (not shown), but change dramatically in the middle, likely due to the increased contact made by bases in those positions. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

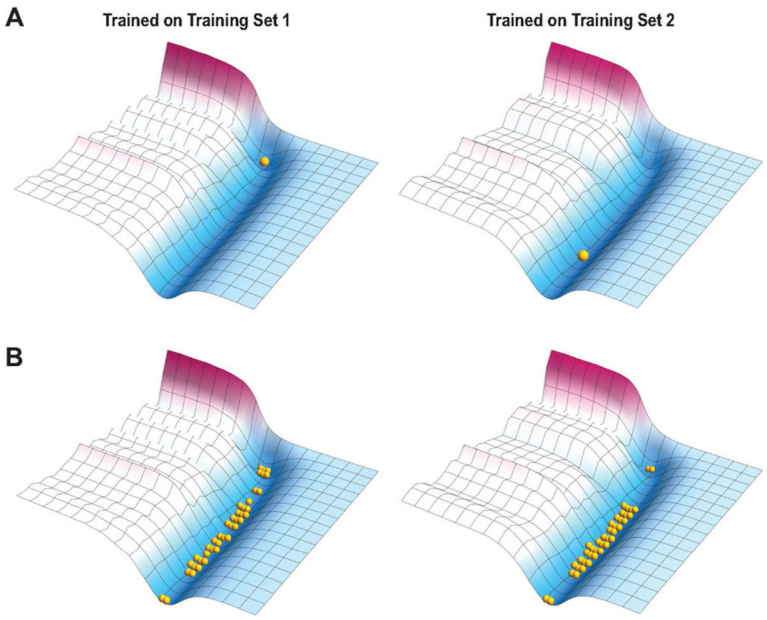

Figure 5.

Variability in metaparameter landscapes across training sets in the full configuration. The metaparameter landscapes for two potentials trained on two different training sets are shown. (**A**) The lowest point (best performing parameter choices) is shown as a yellow sphere. While the overall shape of the landscape is stable between training sets, the exact lowest point is highly variable. (**B**) The top 40 performing models are shown as yellow spheres. Unlike the single top-performing model, the locations of the 40 models as a whole are robust across training sets. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

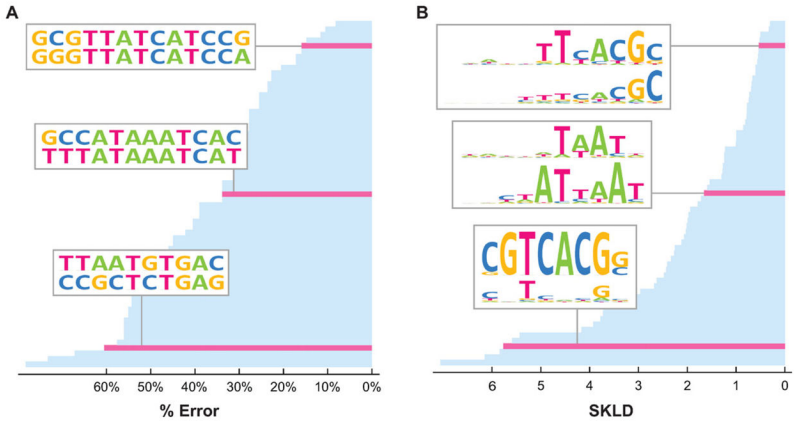

Figure 6.

Bar chart showing prediction performance. Each bar represents a single prediction made by the algorithm, with shorter bars corresponding to better predictions. Highlighted examples (pink bars) represent best, average, and worst cases, with insets comparing experimentally-determined consensus sequences (top) to predictions (bottom). (**A**) Bar chart showing the Hamming distance (fraction of incorrect bases) in consensus sequences predicted in the full testing configuration using the Quasichemical potential with all three enhancements enabled. (**B**) Bar chart showing the SKLD scores (lower is better) for PWMs predicted in the full testing configuration using the μ potential with all three enhancements enabled. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 7.

Individual performance of enhancements. The individual impact of each enhancement (TBA, training with binding affinity; SAF, spatially-aware metaparameter fitting, EBF, ensemble-based metaparameter fitting) on every statistical potential (Quasichemical, DFire, and μ) and testing configuration [full, pruned, and Homeodomain (Hd)] is summarized visually by a two-color bar, where the blue portion indicates the number of times (out of eight instances) that an enhancement improved performance, and the pink portion indicates the number of times an enhancement worsened performance. The eight instances arise because each enhancement was tested individually, in combination with each of the other two enhancements separately, and in combination with both of the other two enhancements together, on the Hamming distance and SKLD metric, for a total of eight tests. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

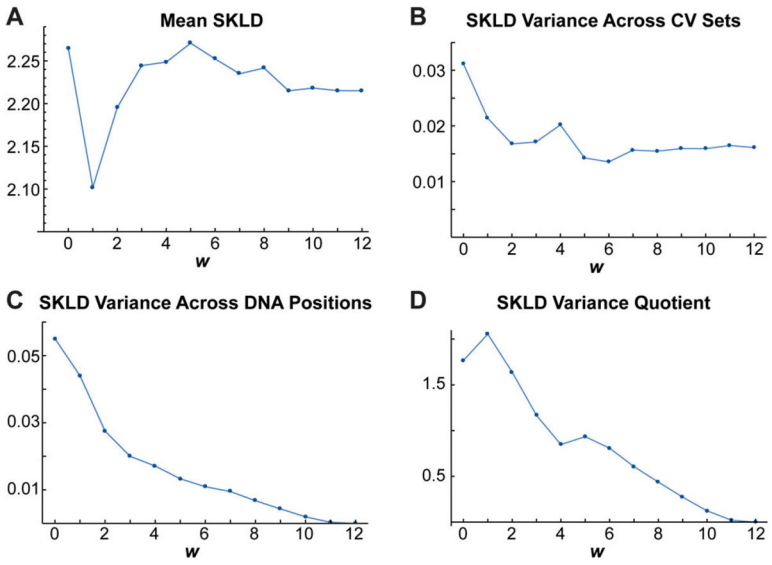

Figure 8.

Analysis of spatially-aware metaparameter fitting. (**A**) The SKLD performance of the spatially-aware Quasichemical potential on predicting PWMs is plotted as a function of the window length metaparameter *w*. An optimal region around $w = 1$ can be observed, which appears to balance the trade-off between model complexity and statistical power. (**B**) The variance in the PWM prediction performance of the spatially-aware Quasichemical potential across CV sets is plotted as a function of the window length metaparameter *w*. Generally this variance appears to decrease as *w* increases. (**C**) The variance in the PWM prediction performance of the spatially-aware Quasichemical potential across DNA basepair positions is plotted as a function of the window length metaparameter *w*. The variance decreases as *w* increases. (**D**) The variance in SKLD scores across DNA basepair positions over the variance in SKLD scores across CV sets for the spatially-aware Quasichemical potential is plotted as a function of the window length metaparameter *w*. The quotient is maximized when $w = 1$. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

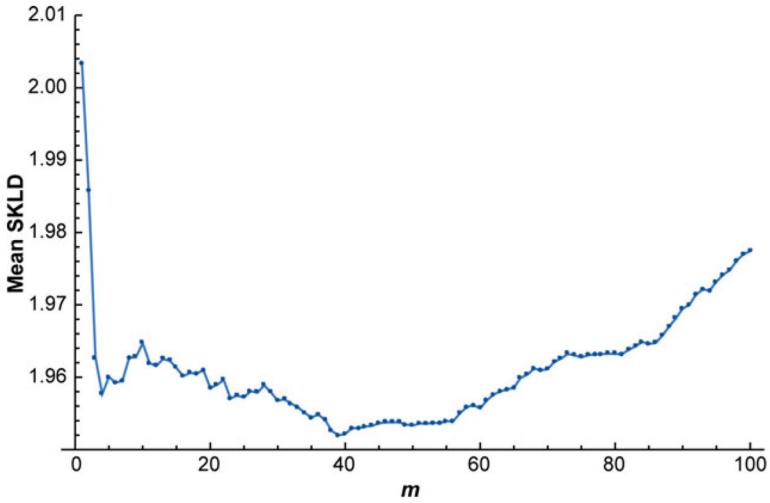

Figure 9.

PWM prediction performance of the Quasichemical potential on predicting PWMs is plotted as a function of the metaparameter *m*. An optimal valley around $m = 40$ is observed. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table I

List of PDB Structures that Comprise the Data Set

ID	PDB ID	Chain ID	Recognition helix residues
1	1AWC	A	371–383
2	1AU7	A	44–51
3	1AU7	A	142–157
4	1B72	A	244–262
5	1B8I	A	142–157
6	1B8I	B	245–260
7	1BC8	C	55–70
8	1BL0	A	41–52
9	1BL0	A	91–102
10	1CF7	A	55–68
11	1CF7	B	113–132
12	1D5Y	A	34–47
13	1DDN	A	38–50
14	1DU0	A	41–57
15	1DUX	C	56–68
16	1E3O	C	43–53
17	1E3O	C	141–157
18	1EFA	A	16–25
19	1F4K	A	53–67
20	1FJL	A	42–63
21	1FOK	A	104–116
22	1GDT	A	172–180
23	1GXP	A	192–206
24	1HCR	A	172–180
25	1HLV	A	119–130
26	1HLV	A	38–48
27	1IC8	A	140–150
28	1IC8	A	260–273
29	1IG7	A	141–159
30	1IGN	A	538–552
31	1JE8	A	183–198
32	1JGG	A	141–159
33	1K61	A	172–189
34	1K78	A	132–140
35	1K78	A	62–75
36	1K78	B	386–396
37	1L3L	A	201–217
38	1LE8	A	110–124
39	1LE8	B	172–188

ID	PDB ID	Chain ID	Recognition helix residues
40	1LMB	3	44–51
41	1LQ1	A	208–226
42	1O3S	A	179–193
43	1PDN	C	47–60
44	1PER	L	28–36
45	1PP7	U	79–90
46	1PUE	E	227–240
47	1PUF	A	245–268
48	1PUF	B	276–294
49	1R71	A	181–190
50	1RIO	H	408–424
51	1RZR	A	15–24
52	1SAX	A	41–55
53	1TC3	C	236–244
54	1U78	A	92–103
55	1U8R	A	37–51
56	2CGP	A	180–192
57	2HDD	A	42–57
58	3CRO	L	28–36
59	3HDD	A	42–57
60	6CRO	A	27–36
61	6PAX	A	117–130
62	6PAX	A	47–60
63	9ANT	A	42–58

Some PDB files contain multiple HTH domains which were treated as separate structures.

Table II

List of Testing Configurations

Configuration	CV Set	IDs of Structures in CV Set
Full	1	2, 3, 4, 7, 21, 49, 53
	2	13, 17, 18, 29, 32, 34, 54
	3	12, 25, 26, 31, 41, 45, 59
	4	15, 16, 20, 22, 30, 50, 61
	5	1, 6, 11, 37, 47, 55, 62
	6	27, 39, 42, 43, 44, 51, 60
	7	10, 14, 19, 24, 36, 52, 58
	8	5, 8, 38, 46, 48, 57, 63
	9	9, 23, 28, 33, 35, 40, 56
Pruned	1	3, 4, 12, 22, 37, 61
	2	16, 27, 49, 50, 54, 56
	3	11, 23, 30, 33, 52, 58
	4	21, 24, 28, 41, 45, 55
	5	1, 40, 51, 53, 60, 62
	6	9, 10, 19, 25, 26, 31
Homeodomain	Training	1, 2, 7, 8, 9, 10, 11, 12, 13, 15, 16, 18, 19, 21, 22, 23, 24, 25, 26, 27, 30, 31, 34, 35, 36, 37, 40, 41, 42, 43, 44, 45, 46, 49, 50, 51, 52, 53, 54, 55, 56, 58, 60, 61, 62
	Test	3, 4, 5, 6, 14, 17, 20, 28, 29, 32, 33, 38, 39, 47, 48, 57, 59, 63

The protein-DNA structures used in every CV set are indicated by their IDs (see Table I).

Table III

Assessment of Three Statistical Potentials in Predicting Consensus Sequences of DNA Binding Sites

Potential	Enhancements			Hamming Distance		
Quasichemical	TBA	SAF	EBF	Full	Pruned	Hd
				0.416	0.433	0.495
	■			0.411	0.453	0.461
		■		0.395	**0.393**	0.495
			■	0.414	0.438	0.500
	■	■		0.375	0.430	**0.456**
	■		■	0.401	0.455	0.480
		■	■	0.390	0.398	0.485
	■	■	■	**0.368**	0.423	0.475
DFire						
				0.538	0.585	0.446
	■			0.482	0.488	0.373
		■		0.463	0.510	0.446
			■	0.533	0.585	0.456
	■	■		0.431	0.460	**0.368**
	■		■	0.463	0.488	0.382
		■	■	0.475	0.502	0.456
	■	■	■	**0.426**	**0.458**	0.373
M						
				0.404	0.455	0.549
	■			0.458	0.498	0.461
		■		0.375	0.418	0.495
			■	0.405	0.457	0.554
	■	■		0.412	0.465	**0.446**
	■		■	0.431	0.507	0.500
		■	■	**0.371**	**0.418**	0.500
	■	■	■	0.399	0.455	0.461

The potentials are tested with three enhancements described in the text (TBA, training with binding affinity; SAF, spatially-aware metaparameter fitting; EBF, ensemble-based metaparameter fitting). Shaded cells in the table indicate the enhancement that is active for that row. For each potential tested, the performance is evaluated using the normalized Hamming distance (Materials and Methods section) on the full, pruned, and Homeodomain (Hd) testing configurations. Lower numbers indicate better performance. The score of the best performing version of each potential is highlighted in bold.

Table IV

Comprehensive Assessment of Three Statistical Potentials in Predicting PWMs of DNA Binding Sites

Potential	Enhancements			Sym. Kullback-Leibler Divergence		
Quasichemical	TBA	SAF	EBF	Full	Pruned	Hd
				2.300	2.180	2.389
	■			2.251	2.490	3.096
		■		2.216	2.100	**2.284**
			■	2.324	2.111	2.383
	■	■		2.141	2.437	2.968
	■		■	2.192	2.457	2.833
		■	■	2.235	**2.064**	2.368
	■	■	■	**2.123**	2.396	2.788
DFire						
				2.539	2.552	1.952
	■			2.470	2.425	1.774
		■		2.440	2.458	1.899
			■	2.510	2.465	1.911
	■	■		2.374	2.411	1.737
	■		■	2.442	2.410	1.844
		■	■	2.414	2.412	1.892
	■	■	■	**2.348**	**2.373**	**1.724**
μ						
				2.226	2.252	2.338
	■			2.220	2.260	2.236
		■		2.042	2.082	2.312
			■	2.127	2.113	2.419
	■	■		2.019	2.165	2.200
	■		■	2.084	2.230	2.233
		■	■	2.015	**2.053**	2.317
	■	■	■	**1.998**	2.176	**2.134**

The potentials are tested with three enhancements described in the text (TBA, training with binding affinity; SAF, spatially-aware metaparameter fitting; EBF, ensemble-based metaparameter fitting). Shaded cells in the table indicate that the enhancement is active for that row. For each potential tested, the performance is evaluated using the SLKD metric (Materials and Methods section) on the full, pruned, and Homeodomain (Hd) testing configurations. Lower numbers indicate better performance. The score of the best performing version of each potential is highlighted in bold.