Direct inference of protein–DNA interactions using compressed sensing methods

Mohammed AlQuraishi[a,b,c] **and Harley H. McAdams**[a,1]

[a]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA 94305; [b]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305; and [c]Department of Statistics, Stanford University, Stanford, CA 94305

Edited* by Stephen J. Benkovic, Pennsylvania State University, University Park, PA, and approved July 6, 2011 (received for review April 25, 2011)

Compressed sensing has revolutionized signal acquisition, by enabling complex signals to be measured with remarkable fidelity using a small number of so-called incoherent sensors. We show that molecular interactions, e.g., protein–DNA interactions, can be analyzed in a directly analogous manner and with similarly remarkable results. Specifically, mesoscopic molecular interactions act as incoherent sensors that measure the energies of microscopic interactions between atoms. We combine concepts from compressed sensing and statistical mechanics to determine the interatomic interaction energies of a molecular system exclusively from experimental measurements, resulting in a "de novo" energy potential. In contrast, conventional methods for estimating energy potentials are based on theoretical models premised on a priori assumptions and extensive domain knowledge. We determine the de novo energy potential for pairwise interactions between protein and DNA atoms from (*i*) experimental measurements of the binding affinity of protein–DNA complexes and (*ii*) crystal structures of the complexes. We show that the de novo energy potential can be used to predict the binding specificity of proteins to DNA with approximately 90% accuracy, compared to approximately 60% for the best performing alternative computational methods applied to this fundamental problem. This de novo potential method is directly extendable to other biomolecule interaction domains (enzymes and signaling molecule interactions) and to other classes of molecular interactions.

DNA motifs | structural biology | machine learning | protein–DNA binding | DNA binding sites

The foundation of molecular analyses of chemical and biological phenomena is the energy potential, a mathematical description of the energy of every possible interaction in a molecular system (Fig. 1*B*). The accuracy of computational and laboratory studies of phenomena ranging from pharmaceutical drug interactions and protein folding to material phase transitions and thin film growth is often limited by the accuracy of these energy potentials. Currently, potentials are inferred using a mixture of theoretical modeling and experimental data (Fig. 1*A*). "Physical potentials" rely on theoretical models to specify the potential's mathematical form and use experimental data to fit few model parameters (1). In contrast, "statistical potentials" fit many parameters to experimental data and use theoretical models for the expected statistics of interactions under randomness to infer a potential (2). In both approaches, theoretical models shape and constrain the inferred potential, resulting in a so-called parametric model. There are several drawbacks to this: (*i*) The a priori assumptions underlying the inferred potentials may be inaccurate. (*ii*) Substantial domain knowledge is required (often exceeding what is known). (*iii*) Potential modeling is lengthy and technically difficult. The theoretical development of some potentials has taken decades (3). To overcome these problems, potentials could in principle be determined strictly from experimental data without recourse to theoretical modeling by experimentally measuring the energies of all distinct interactions. In practice, direct measurement of interatomic potentials has been possible only for the simplest systems, due to a combinatorial explosion in the number of possible interactions that renders experiment-based inference intractable. We have developed a general method for the inference of "de novo" potentials that circumvents the experimental intractability barrier by exploiting recent discoveries in information theory known as compressed sensing. This approach results in a nonparametric potential that does not require an a priori assumption of a theoretical model, overcoming a fundamental limitation of both physical and statistical potentials.

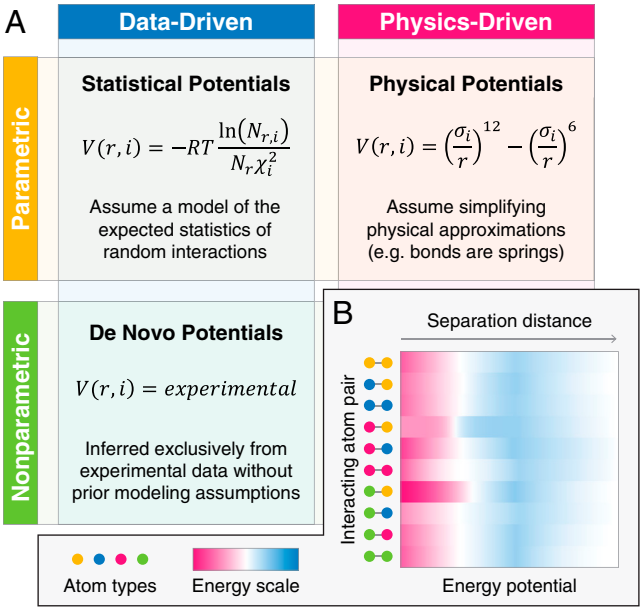

Fig. 1. Types of energy potentials. (*A*) A potential $V(r,i)$ mathematically specifies the energies of all microscopic interactions in a molecular system in terms of distance r and interaction type i. Conventional physical and statistical potentials are parametric mathematical models similar to the examples shown. Our de novo potentials are nonparametric; i.e., they do not assume a mathematical model. (*B*) A potential can be visualized as a heat map where the interaction energy of every atom pair as a function of the atoms' separation distance is represented by a color (pink: high potential energy, repulsive region; blue: low potential energy, attractive region).

Below, we demonstrate our method by applying it to the prediction of sequence-specific protein–DNA-binding interactions, a classic problem in molecular biology. Sequence-specific protein–DNA binding is a central phenomenon underlying transcriptional regulation of the cell in all organisms. Here, we describe a de novo potential for interatomic protein–DNA interactions and

Author contributions: M.A. designed research; M.A. performed research; M.A. analyzed data; and M.A. and H.H.M. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

See Commentary on page 14713.

[1]To whom correspondence should be addressed. E-mail: hmcadams@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1106460108/-/DCSupplemental.

use this potential to computationally predict the DNA-binding sites of proteins with near experimental accuracy. A schematic overview summarizing the key steps in our approach is shown in Fig. S1.

The crux of our method is a unique mathematical formulation that recasts the determination of potentials as a signal acquisition problem using compressed sensing techniques (4, 5). By exploiting a key property of compressed sensing (discussed below), we circumvent the experimental intractability of determining energy potentials. This formulation rests on three key observations: (*i*) In certain systems, e.g., biomolecular interactions, only a few types of interatomic interactions are energetically important, such as hydrogen bonds. (*ii*) Mesoscopic interactions, e.g., protein–DNA binding, can be viewed as sensors of the underlying microscopic potential. (*iii*) There is a correspondence between the mathematical formulation of logistic regression and the statistical mechanical concept of the canonical ensemble. Using these ideas, we formulate the determination of energy potentials as a tractable signal acquisition problem as described in *Methods*. Central to this approach is the distinction between microscopic and mesoscopic interactions. The term "mesoscopic interactions" refers to interactions between molecules or molecular complexes, whereas "microscopic interactions" refers to interactions at the atomic scale. This distinction is ultimately application-specific and depends on what interaction energies are to be determined and what experimental data are available. An essential requirement is that mesoscopic interactions be comprised of microscopic interactions, such that the energy of a mesoscopic interaction is the sum of the energies of its constituent microscopic interactions. The identities of the microscopic interactions constituting each mesoscopic interaction must also be known. When these requirements are met, the method can determine the energies of the microscopic interactions using observed energies or probabilities of the mesoscopic interactions as the experimental data. For protein–DNA interactions, the mesoscopic interactions are protein–nucleotide binding events represented by protein–DNA crystal structures with known binding energies or relative binding probabilities, and the microscopic interactions are pairwise, distance-dependent, contacts between protein atoms and nucleotide atoms [standard atom type categories are used (6), and distances are segmented into discrete bins whose widths are treated as model parameters as described in *SI Text* S1, Section 2]. Thus, each protein–nucleotide crystal structure characterizes a mesoscopic interaction whose constituent microscopic interactions are readily identified from the structure. Intraprotein and intra-DNA energies are ignored as protein–DNA interactions are our focus. We use these data together with the known energies or probabilities of protein–nucleotide binding events to infer the interatomic protein–DNA potentials, and we show that these de novo potentials can be used to predict protein–DNA-binding motifs with unprecedented accuracy.

Model

Overview. We outline our signal-based formulation of the potential inference problem here and derive it formally in *SI Text* S1, Section 1. Signal acquisition is comprised of three parts: the signal (conventionally an image), the sensor (camera photographing the image), and the sensor measurements (light intensities of the image) (Fig. 2*A*). A signal is represented by a vector whose elements are the signal intensities at different locations (e.g., different image pixels). Conventional sensing theory (Nyquist–Shannon) stipulates that signal acquisition requires twice as many

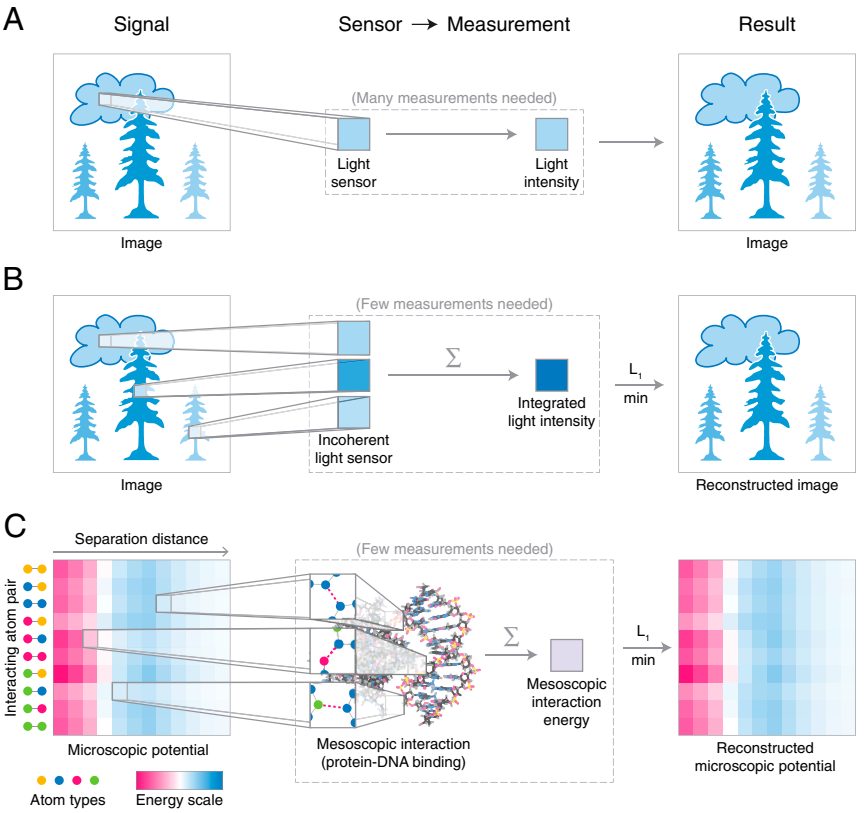

Fig. 2. Comparison of conventional sensing, compressed sensing, and de novo potential determination. (*A*) In conventional image sensing the intensities of all pixels are acquired directly. (*B*) In compressed sensing, the image is inferred using ℓ_1 minimization from a relatively small number of measurements that sum the signal intensity of multiple image pixels. (*C*) Potential determination as an application of compressed sensing. The potential is represented by a heat map of microscopic interaction energies ranging from repulsive (dark pink) to attractive (dark blue). The interatomic protein–DNA potential can be inferred by ℓ_1 minimization from measurements provided by a small number of sensors (protein–DNA structures + binding energies). See text for additional details and Fig. S2 for a more mathematical treatment.

measurements as the length of this vector for complete recovery of the signal. However, the compressed sensing (CS) framework has shown that under certain conditions, far fewer measurements are necessary when the signal is inferred using ℓ_1 minimization (5). We exploit this property to circumvent the combinatorial explosion noted earlier that causes experimental intractability. The CS technique requires two conditions for applicability (5): (*i*) The signal must be nearly sparse; i.e., most vector elements must have negligible intensity. (*ii*) The sensors must be incoherent; i.e., they measure the integrated intensity of multiple signal vector elements (Fig. 2*B*), and the set of vector elements sensed must be highly variable (ideally, random) between sensors (*SI Text* S1, Section 1). Also, the identity of the vector elements sensed by each sensor must be known. We reformulate potential determination as a CS problem by treating the interatomic potential as the signal we wish to acquire, with mesoscopic interactions as the sensors and mesoscopic interaction energies as the measurements (Fig. 2*C*). The signal's vector is comprised of the energies of all distinct microscopic interactions, with different vector elements corresponding to different microscopic interactions and signal intensity corresponding to interaction energy. In the protein–DNA application, we treat distinct combinations of protein atoms, nucleotide atoms, and distance bins as distinct interactions (*SI Text* S1, Section 2). This leads to a combinatorial explosion in the number of possible interactions, causing the signal's vector to be extremely long (up to approximately 50,000 elements) (*SI Text* S1, Section 2). However, the vector will be nearly sparse because most interactions are energetically negligible (7). In our crystal structure dataset, we found that only 9% of interaction energies were nonnegligible (*Results*). This satisfies the first condition. Regarding the second condition, the energies of mesoscopic interactions are incoherent measurements, because (*i*) they are the summed energies of the microscopic interactions and thus integrate the intensity of multiple vector elements, and (*ii*) the set of microscopic interactions present in each mesoscopic interaction is highly variable as discussed below. Because the vector elements sensed by each measurement must be known, the microscopic interactions comprising each mesoscopic interaction must be known. Protein–DNA crystal structures provide a dataset of mesoscopic interactions whose constituent microscopic interactions are identified from the positions of the protein and nucleotide atoms in the contact regions of the structure. We used a set of 63 such nonredundant structures (Dataset S1), combined with their measured binding affinities, as the dataset for the de novo potential determination described below. The nonredundancy of these structures ensures that each mesoscopic interaction samples a different set of microscopic interactions, because the intrinsic variability of the structures due to their varying spatial conformations and different amino acid compositions results in high variability in the microscopic interactions that constitute each mesoscopic interaction. (The degree of incoherence is quantified later in *Results*). Now, because we have recast potential determination as a CS problem, only a small number of incoherent measurements, i.e., experimentally characterized protein–nucleotide binding events, are needed. This circumvents the experimental combinatorial explosion problem cited earlier.

Mathematical Formulations. We show that ℓ_1-regularized linear regression (8) infers potentials from mesoscopic interaction energies in *SI Text* S1, Section 1 (see also Fig. S2). We also derive a probability-based formulation that uses the relative probability of a mesoscopic interaction within a collection of possible alternative interactions (e.g., alternative DNA sites where the protein binds) as experimental data. This collection must form a canonical ensemble, i.e., a set of physical states in which the energy may vary, but the volume, temperature, and number of particles are fixed. Multiple distinct canonical ensembles can be used to infer a single potential (e.g., multiple protein–DNA complexes can be used to infer a single protein–DNA potential). We derive this formulation using a constrained version of ℓ_1-regularized multinomial logistic regression (9) (*SI Text* S1, Section 1). In our protein–DNA application, a collection of protein–nucleotide complexes in which the protein is fixed and individual nucleotides are varied forms a canonical ensemble, and the protein's relative binding probabilities to different nucleotides are the experimental data. These probabilities are obtained from experimentally determined position weight matrices (PWMs) of protein binding sites or from consensus binding sequences (by assuming that consensus nucleotides bind with 100% probability).

Results

Application to Prediction of Protein Binding Sites. We have used the probability-based formulation to determine the protein–DNA potential of helix-turn-helix (HTH) proteins and predict their consensus binding sequences and PWM motifs. 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 structures in the Protein Data Bank (10). HTH proteins include virtually all bacterial transcription factors and about 25% of human transcription factors (11). For the prediction of consensus binding sequences, the potential is inferred using probabilities derived from the consensus sequences of protein–DNA structures in a dataset reserved for training the algorithm (*SI Text* S1, Section 3). A separate set of protein–DNA structures is used to test predictions made with the inferred potential. For each protein–DNA structure in the test set, every DNA sequence position is mutated in silico to every possible pair of nucleotides, and the relative binding affinities of the mutated structures are computed (*SI Text* S1, Section 3). In silico mutagenesis was carried out using the 3DNA software package (12, 13), which maintains the backbone atoms of the DNA molecule, but replaces the base pair atoms in a way that is consistent with the backbone orientation in the crystal. We assume independence of DNA positions and repeat this process for every position. The most probable nucleotides at all positions are predicted with 12.9% error, compared to 42.1% error by the leading alternative method (Table 1, Baseline model). For the more complex problem of predicting quantitative PWMs, we determine the potential using probabilities derived from published experimentally determined PWMs of the 63 protein–DNA structures in the dataset (Dataset S2 and *SI Text* S1, Section 3). Compared to leading physical and statistical potentials (6, 14–16), our de novo potential method produces the best PWM score (Table 1) on the symmetric Kullback–Leibler divergence (SKLD) metric (*SI Text* S1, Section 3). Note that the second and third best performing potentials require consensus binding sequences as input. Providing that input significantly simplifies the problem, whereas our method infers the consensus binding sequences.

Generalizations. We also consider two generalizations that relax physical constraints, yielding pseudopotentials that perform better in practice. First, we mathematically transform the regression inputs to improve their statistical and numerical properties (*SI Text* S1, Section 2). Second, we infer distinct potentials for interactions occurring in different regions of the HTH–DNA-binding interface, motivated by the observation that binding affinity is strongest in the core region of the binding interface and gets progressively weaker away from the core region (17) (*SI Text* S1, Section 2 and Fig. S3). We tested these generalizations individually and in combination (see Table 1). The consensus sequence predictions are slightly improved (10.1% vs. 12.9% error), but the improvement in PWM prediction is dramatic (SKLD of 1.699 vs. 1.960), larger than the gain obtained in going from the Quasichemical (6) to the DNAPROT (16) algorithm (2.248 vs. 1.991), which accounts for intra-DNA interactions and requires consensus sequences. The positive impact of this

Table 1. Performance of de novo potential and other leading potentials

Potential	Type	Intra-DNA interactions	Prediction quality: Consensus sequence error	Prediction quality: PWM symmetric KL divergence
Random model	N/A	N/A	75%	3.335
Methods requiring consensus sequences				
Rosetta (12)	physical	yes	N/A	2.632
Cumulative contacts (13)	statistical	no	N/A	2.033
DNAPROT (14)	statistical	yes	N/A	1.991
Methods not requiring consensus sequences				
DNAPROT* (14)	statistical	no	60.20%	3.279
Rosetta* (12)	physical	no	50.80%	2.719
Quasichemical (6)	statistical	no	42.10%	2.248
Our methods (do not require consensus sequences)				
Baseline	de novo	no	12.90%	1.96
Transformed inputs	de novo	no	10.20%	1.861
Region specific	de novo	no	13.70%	1.792
Both generalizations	de novo	no	10.10%	1.699

Performance is assessed based on predictions of consensus sequences and PWMs averaged over the 63 structures in the dataset. For consensus sequence prediction, error is measured by the percentage of incorrectly predicted bases. For PWM predictions, the average SKLD over all DNA positions is reported (lower is better). A random model in which all DNA base pairs are assumed to be equally likely is also shown for reference.
*Only the direct readout components of potentials are used in those tests because they do not require consensus sequences as input.

second generalization on PWM prediction, but not on consensus sequence prediction, results because consensus sequences do not capture the relative binding strength of protein–DNA interactions for alternative DNA sequences. Fig. 3 shows a bar chart of the accuracy of all 63 predictions made using our de novo potential with both generalizations, along with representative best, average, and worst case predictions of consensus sequences and PWMs. Fig. 4 compares predictions for the proteins where the de novo algorithm exhibited the greatest improvements relative to other methods.

Characterization of Best Performing Model. As discussed earlier, a collection of sensors must be incoherent to ensure high-quality reconstruction of the underlying potential with compressive sensing methods, and the potential must be sufficiently sparse relative to the number of available measurements (*SI Text* S1, Section 1). To determine the degree to which these requirements are satisfied by the potentials derived from the protein:DNA complexes in our dataset, we consider the best performing baseline model, applied without the two generalizations discussed in the previous section. This model produced an energy potential with a total of 2,997 unique microscopic interactions using 1.3-Å wide distance bins and 5.9-Å cutoff distance (*SI Text* S1, Section 2). The total number of sensors in the dataset is 592, as each protein–DNA crystal structure yields multiple sensors because we make the common assumption of independence between DNA base pair positions. As previously noted, accurate inference is still possible despite having a smaller number of sensors than the number of unique microscopic interactions, if the potential is sufficiently sparse. Of the 2,997 unique interactions, only 270 have nonzero energy, suggesting that our dataset will yield accurate potentials. In fact, it is likely that the best performing choices of binning width and cutoff distance used for the baseline model represent the optimal trade-off between spatial resolution and statistical power.

An additional way to address the suitability of the dataset for potential inference is based on the consideration of all the pairwise angles between the sensor vectors in the dataset. The distribution of the absolute values of the cosines of these pairwise angles (Fig. S4) characterizes the incoherence of the sensors (18). The mean and median values of this distribution are 0.081 and 0.041, respectively. These values are significantly lower than 1, thus indicating that the set of sensors comprised by the

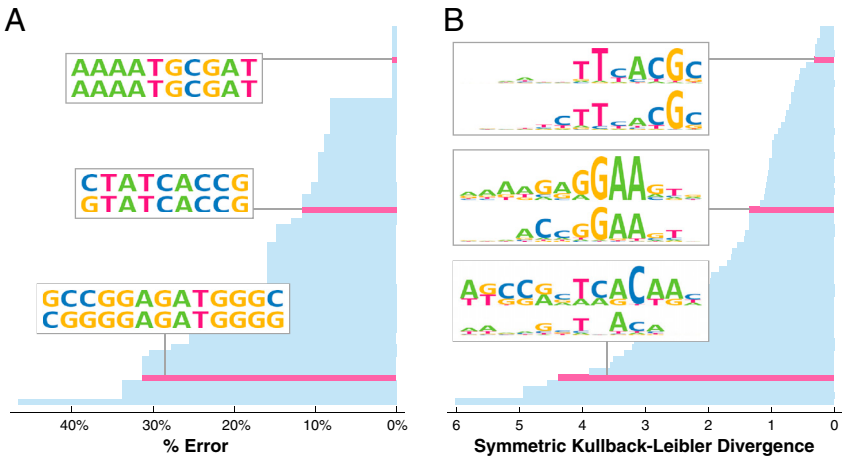

Fig. 3. Representative performance of de novo potential in predicting DNA-binding sites of 63 proteins. (*A*) Bar chart of the errors (fraction of incorrect bases) in consensus sequences predicted using de novo potential method. Each bar represents a single prediction made by the algorithm, with shorter 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*). (*B*) SKLD scores (lower is better) for PWM predictions, with insets comparing experimentally determined PWMs (*Top*) to predictions (*Bottom*).

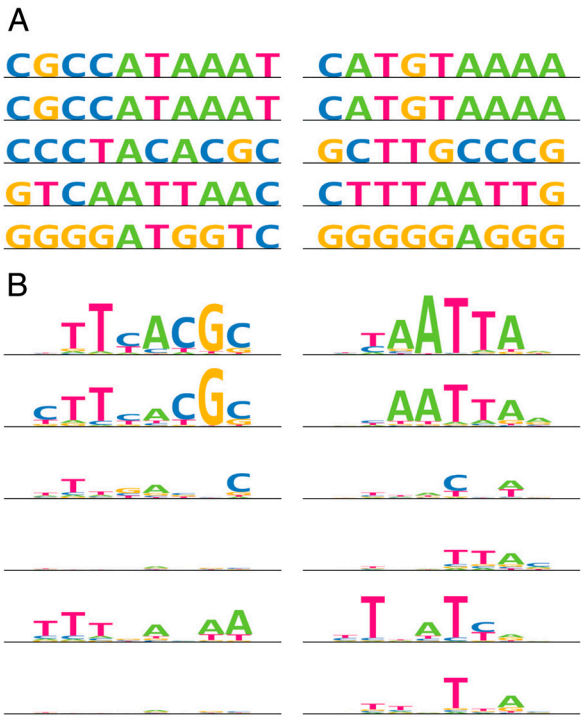

Fig. 4. Examples highlighting significant improvement in prediction quality between de novo potential and other leading potentials. (*A*) Experimental and predicted consensus sequences for the *Drosophila melanogaster* Ultrabithorax Hox protein (*Left*) and the *Saccharomyces cerevisiae* MATα2 (*Right*) protein are shown. (*Top* to *Bottom*) Experimental, de novo potential, Rosetta (direct readout only), Quasichemical, and DNAPROT (direct readout only). (*B*) Experimental and predicted PWMs for the *Homo sapiens* Pax6 Paired domain (*Left*) and *D. melanogaster* Engrailed homeodomain (*Right*) are shown. (*Top* to *Bottom*) Experimental, de novo potential, Rosetta, Cumulative Contacts, Quasichemical, and DNAPROT.

protein:DNA crystal structures in our dataset has good incoherence properties (18, 19).

Discussion

Potential for Improving Performance. The protein–DNA-binding site predictions we report are an application of our de novo potential inference method, and they exhibit a dramatic improvement over the leading alternative methods, with predictions within the experimental error of the PWMs for at least half of the cases studied. Although the accuracy depicted in Fig. 3 is substantially better than achieved with alternative methods, the predictions for the proteins in the lower quarter of Fig. 3 need to be improved. The 63 protein–DNA complexes in our database may provide biased or insufficient coverage of some of the microscopic interactions. Or, there could be significant variance in the quality of the crystal structures determined for the different complexes in the database that we curated from the Protein Data Base (10). Additionally, prediction errors might reflect the effects of other mechanisms that affect the shape and accessibility of the DNA in vivo so that the structure of the crystallized complex differs from the in vivo structure. Also, some transcription factors have been observed to bind DNA with two or more distinct motifs (20), and the alternate motifs would be missing from our dataset.

Principled Selection of Crystallization Targets. As in other CS application domains, the accuracy of the inferred potentials depends on the characteristics of the sensor matrix: in this case, the collection of protein–DNA structures available. As discussed above, quantitative measures such as coherence can be used to assess the suitability of a sensor matrix for compressive sensing (5, 18, 19). These measures can provide a principled framework for selecting additional crystallization targets that will maximally enhance the sensing performance of a protein–DNA structural dataset. We showed above that our current dataset has good incoherence properties, yet we expect that the addition of more protein–DNA crystal structures, specifically chosen to yield a sensor matrix with even lower coherence, will yield more accurate energy potentials and better binding site predictions.

Advantages over Statistical Potentials. Although statistical potentials and our de novo potentials both use experimental datasets to derive the final energy potential, de novo potentials are nonparametric; i.e., they do not assume an underlying mathematical form. In contrast, statistical potentials rely on experimental datasets to fit a parametric model with a fixed mathematical form. De novo potentials overcome additional limitations specific to statistical potentials. First, although statistical potentials utilize only atomic data such as crystal structures to fit their parameters, de novo potentials combine structural information and experimental binding data into a single formulation for inference. In the field of protein–DNA-binding site prediction, combining these types of data has been a long-standing objective (21–24). Second, statistical potentials implicitly assume that all structures come from the same canonical ensemble. This oversimplification ignores the chain connectivity and amino acid composition of proteins, and it is thought to be the cause of common anomalies observed in statistical potentials (2). In de novo potentials, this assumption is eliminated in the energy-based formulation, and it is relaxed significantly in the probability-based formulation, so that only subsets of the data are assumed to form canonical ensembles (*SI Text* S1, Section 1). Third, by assigning equal weight to microscopic interactions observed in different structures, statistical potentials implicitly assume that different structures have the same binding or formation energy. This is not the case, as different protein–DNA complexes are known to have different binding affinities. From a statistical mechanical standpoint, high affinity complexes correspond to more frequent mesoscopic interactions than low affinity complexes, requiring that the underlying microscopic interactions be given proportionally greater weight. De novo potentials take this into account, as the binding energies of mesoscopic interactions are an explicit part of the formulation. Fourth, there is no theoretical assurance that as more data are added for the statistical potential fitting process, the inferred energies will ultimately converge to the true underlying interaction energies. The same observation holds for physical potentials. In contrast, for de novo potentials, if the formal requirements of compressed sensing are satisfied, then additional sensors in the dataset will lead to ever closer estimates of the interaction energies.

De Novo Potentials in Other Molecular Interaction Domains. The important insight that underlies our de novo methodology is that experimental datasets relating to mesoscopic interactions in a wide range of fields can be cast as incoherent measurements of microscopic interactions in the compressed sensing framework. In these cases, powerful compressed sensing methods can be used to determine de novo potentials. In the current protein–DNA application, microscopic interactions were defined to always involve one protein atom and one DNA atom, thus neglecting intra-DNA interactions. If microscopic interactions are defined to include noncovalent contacts between two DNA atoms, then the indirect readout component of protein–DNA interactions can also be modeled, capturing intra-DNA interactions.

Similarly, to infer a potential for protein–protein interactions or for protein folding, noncovalent contacts between protein atoms would be treated as the microscopic interactions. However, interactions need not be restricted to those involving pairs of atoms. Interactions involving multiple atoms, as well as coarse-grained potentials in which the "atoms" of the systems are

residues for example, could be used. For mesoscopic measurements of protein–protein interactions, biochemical data on protein–protein binding kinetics can be used. The Protein–Protein Interaction Thermodynamic Database (PINT) is a database of such measurements (25). For mesoscopic measurements of protein folding, the kinetics or mean folding times of proteins are necessary. Although such measurements are difficult to obtain, significant progress in experimental techniques has been made recently (26–28). The advent of these recent experimental techniques promises to make such measurements more readily available in the future.

ACKNOWLEDGMENTS. R. Altman, G. Bejerano, J. Boyd Kozdon, A. Deacon, S. Hong, V. Pande, K. Sachs, and L. Shapiro provided helpful comments. We thank E. Candes, T. Hastie, and M. Levitt for insightful discussions, A. Morozov for helpful advice on the Rosetta protein–DNA module, and K. Arya and G. Cooperman for customizing the DMTCP checkpointing software for our purposes. Wolfram Research provided the Mathematica software environment necessary for the analyses performed. This work was supported by Department of Energy Office of Science Grant DE-FG02-05ER64136 (to H.H.M.). We used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract DE-AC02-05CH11231. M.A. was supported by the Stanford Genome Training Program (Grant T32 HG00044 from the National Human Genome Research Institute).

1. Jorgensen WL, Tirado-Rives J (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci USA* 102:6665–6670.
2. Zhou Y, Zhou HY, Zhang C, Liu S (2006) What is a desirable statistical energy function for proteins and how can it be obtained? *Cell Biochem Biophys* 46:165–174.
3. Ponder JW, Case DA (2003) Force fields for protein simulations. *Adv Protein Chem* 66:27–85.
4. Chartrand R, Baraniuk RG, Eldar YC, Figueiredo MAT, Tanner J (2010) Introduction to the issue on compressive sensing. *IEEE J Sel Top Signal Process* 4:241–243.
5. Candes EJ, Wakin MB (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25:21–30.
6. Donald JE, Chen WW, Shakhnovich EI (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res* 35:1039–1047.
7. Brändén C-I, Tooze J (1999) *Introduction to Protein Structure* (Garland, New York), 2nd Ed, p xiv.
8. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 58:267–288.
9. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.
10. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
11. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet* 10:252–263.
12. Lu XJ, Olson WK (2003) 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res* 31:5108–5121.
13. Lu XJ, Olson WK (2008) 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* 3:1213–1227.
14. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* 33:5781–5798.
15. Morozov AV, Siggia ED (2007) Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci USA* 104:7068–7073.
16. Angarica VE, Perez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics* 9:436.
17. Wintjens R, Rooman M (1996) Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J Mol Biol* 262:294–313.
18. Tropp JA (2008) On the conditioning of random subdictionaries. *Appl Comput Harmon Anal* 25:1–24.
19. Candes EJ, Romberg J (2006) Quantitative robust uncertainty principles and optimally sparse decompositions. *Found Comput Math* 6:227–254.
20. Badis G, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324:1720–1723.
21. Mirny LA, Gelfand MS (2002) Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res* 30:1704–1711.
22. Hoglund A, Kohlbacher O (2004) From sequence to structure and back again: Approaches for predicting protein-DNA binding. *Proteome Sci* 2:3.
23. Eisen M (2005) All motifs are NOT created equal: Structural properties of transcription factor-DNA interactions and the inference of sequence specificity. *Genome Biol* 6:P7.
24. Moroni E, Caselle M, Fogolari F (2007) Identification of DNA-binding protein target sequences by physical effective energy functions: Free energy analysis of *lambda* repressor-DNA complexes. *BMC Struct Biol* 7:61.
25. Kumar MD, Gromiha MM (2006) PINT: Protein-Potein Interactions Thermodynamic Database. *Nucleic Acids Res* 34:D195–198.
26. Vendruscolo M, Paci E (2003) Protein folding: Binging theory and experiment closer together. *Curr Opin Struct Biol* 13:82–87.
27. Oliveberg M, Wolynes PG (2005) The experimental survey of protein-folding energy landscapes. *Q Rev Biophys* 38:245–288.
28. Mello CC, Barrick D (2004) An experimentally determined protein folding energy landscape. *Proc Natl Acad Sci USA* 101:14102–14107.

Supporting Information

AlQuraishi and McAdams 10.1073/pnas.1106460108

SI Methods

SI Methods describe the general methodology for de novo potential determination using compressed sensing, the determination of protein–DNA potentials, and the prediction of protein–DNA-binding sites and testing methodology. There are three sections. Section I contains the derivation of the general methodology for de novo potential determination using compressed sensing. The formulation in Section I is not specific to a particular application, but is applicable to a range of biological and chemical systems. The mathematical notation and terminology used throughout is in Section I.

This work brings together disparate concepts from information theory, statistical mechanics, and structural biology. The following background references may be useful to the reader:

- Linear and logistic regression (1).
- Compressed sensing (2, 3).
- Statistical mechanical ensembles (4, 5).
- Structural basis of protein–DNA interactions (6, 7).

Section II describes the application of the general methodology from Section I to the determination of a de novo protein–DNA potential. We reformulate the abstract constructs of the general methodology to the specifics of protein–DNA interactions and introduce several modifications that exploit the unique properties of protein–DNA interactions. Our choices for meta-parameters and implementation details are also described in Section II.

Section III contains a description of the use of the protein–DNA potentials described in Section II to predict protein–DNA-binding sites. We detail our structure-based approach to protein–DNA-binding site prediction, the dataset used for training and testing, and the quantitative metrics used to compare results between our de novo potential method and previously published methods.

I. General De Novo Potential Determination Using Compressed Sensing. This section introduces our general methodology for inferring potentials using compressed sensing. First, we show how the energies of microscopic interactions can be determined from the measurements of the interaction energies or relative probabilities of mesoscopic interactions by reformulating potential inference as one of two possible regression problems: (*i*) linear regression for when the measurements are energies of mesoscopic interactions or (*ii*) multinomial logistic regression when the measurements are the relative probabilities of mesoscopic interactions. Next, having reformulated the potential determination problem as a regression problem, we show how compressed sensing methods can be employed to infer a de novo potential using a relatively small number of these experimental measurements.

Notation and preliminary assumptions. I is the set of all possible microscopic interactions. A microscopic interaction may be as simple as two atoms existing within a predefined distance of one another or as complex as a system of multiple molecules in conjunction with the solvent. Note that an interaction may depend on the distance of the interacting elements, in which case the same elements interacting at different distances would correspond to different interactions in our formulation. Distance may be defined in terms of discrete distance ranges, or bins. For each microscopic interaction $i \in I$, we denote its energy by e_i.

K is the set of mesoscopic interactions. Intuitively, a mesoscopic interaction $k \in K$ is a large-scale interaction where many microscopic interactions combine to create the larger-scale interaction. For example, a protein binding DNA is a mesoscopic interaction comprised of all the energetic interactions occurring between protein and DNA atoms. Formally, we associate with every mesoscopic interaction $k \in K$ a vector $\mathscr{C}_k \in \mathbb{Z}^{*|I|}$ that contains for each $i \in I$ a count $\mathscr{C}_{k,i}$ of the number of times that microscopic interaction i is observed in mesoscopic interaction k. For microscopic interactions unobserved in $k \in K$, their corresponding counts are set to 0.

We denote the energy of a mesoscopic interaction k by E_k and define it as

$$E_k = \sum_{i \in I} e_i \mathscr{C}_{k,i}.$$

This formalizes the notion of a mesoscopic interaction by defining the energy of a mesoscopic interaction k to be the sum of the energies of its constituent microscopic interactions.

Inference using mesoscopic interaction energies. Given a set of mesoscopic interactions K for which we know the constituent microscopic interactions, and a corresponding set of $\{E_k\}_{k \in K}$, linear regression (1) can be used to infer the energies $\{e_i\}_{i \in I}$ of the microscopic interactions. The mapping is direct (Fig. S2). We treat E_k as the response variable y_k, and $\mathscr{C}_{k,i}$ as the input variable $x_{k,i}$. In linear regression,

$$y_k = \sum_{i \in I} \beta_i x_{k,i};$$

thus the inferred coefficients $\{\beta_i\}_{i \in I}$ will be the microscopic interaction energies $\{e_i\}_{i \in I}$. Alternatively in matrix notation, we set the energy vector $E = \{E_k\}_{k \in K}$ to equal the response vector $y = \{y_k\}_{k \in K}$, and the counts matrix $\mathscr{C} = \{\mathscr{C}_{k,i}\}_{k \in K, i \in I}$ to equal the design matrix $X = \{x_{k,i}\}_{k \in K, i \in I}$, to obtain the standard linear regression relationship $y = X\beta$. Within the compressed sensing framework, y is the set of compressive measurements, X is the sensor matrix (set of sensors), and β, the underlying microscopic potential, is the signal.

Inference using relative probabilities of mesoscopic interactions. If the absolute energies of mesoscopic interactions are unknown, but their relative probabilities are known, then a constrained version of multinomial logistic regression can be shown to infer the energies of the microscopic interactions. A key feature of our formulation is the use of a specific type of statistical mechanical ensemble, the canonical ensemble (4). An ensemble is a collection of physical states, for example, the different structural conformations of a protein. In canonical ensembles, the temperature, volume, and number of particles are assumed identical across the states, but the energy may vary between states. In our formulation, different mesoscopic interactions correspond to different states of a canonical ensemble. For the protein–DNA-binding problem, this corresponds to the protein binding to alternate DNA sequences with different binding energies. The probabilities of finding a system in each of the different states of a canonical ensemble follow the Boltzmann distribution (4). We can exploit this property to reformulate potential determination as an instance of multinomial logistic regression. In this formulation,

the relative probabilities of mesoscopic interactions are used as experimental data. By relative probabilities we mean that the probability of a mesoscopic interaction only has to be normalized with respect to two or more mesoscopic interactions within the same canonical ensemble. Multiple distinct canonical ensembles can be used, and so not all mesoscopic interactions need to come from the same canonical ensemble. We first derive the general formulation, and then we show its formal equivalence to multinomial logistic regression under appropriate constraints.

Formally, let K be the set of mesoscopic interactions for which we have measurements (described below). We induce a partition $\{K_1,\cdots,K_L\}$ on K such that mesoscopic interactions within a block K_l come from the same canonical ensemble, and there are L distinct canonical ensembles. In other words, for a fixed $l \leq L$, for all $k \in K_l$, the temperature, volume, and number of particles are assumed fixed. Note that k is now being used to refer to a mesoscopic interaction within a given canonical ensemble.

For each canonical ensemble K_l, we require measurements corresponding to the relative probabilities of every measured mesoscopic interaction $k \in K_l$ in the ensemble. In other words, the measured probabilities $\{p_k\}_{k\in K_l}$ obey

$$\sum_{k\in K_l} p_k = 1.$$

For all $l \leq L$. Because by definition all mesoscopic interactions $k \in K_l$ come from the same canonical ensemble, then the measured probability p_k of a given mesoscopic interaction k is described by the Boltzmann distribution (4):

$$p_k = \frac{1}{Z_l} e^{-\alpha_l E_k},$$

where E_k is the energy of the mesoscopic interaction as previously defined, Z_l is the partition function of the canonical ensemble K_l, and α_l is the inverse temperature parameter for the ensemble (we are not following the standard convention of using β for the inverse temperature to avoid confusion with the notation for regression coefficients).

Given a set of mesoscopic interactions K for which we know the constituent microscopic interactions, and a partitioning of the mesoscopic interactions such that the probabilities $\{p_k\}_{k\in K_l}$ are known for all $l \leq L$, we can infer the energies of the microscopic interactions. Recall the previous definition of the energy E_k, and obtain

$$p_k = \frac{1}{Z_l} e^{-\alpha_l E_k} = \frac{1}{Z_l} e^{-\alpha_l \sum_{i\in I} e_i \mathscr{C}_{k,i}}.$$

Expanding the partition function yields

$$p_k = \frac{e^{-\alpha_l \sum_{i\in I} e_i \mathscr{C}_{k,i}}}{\sum_{j\in K_l} e^{-\alpha_l \sum_{i\in I} e_i \mathscr{C}_{j,i}}}.$$

Multiplying the numerator and denominator by $e^{\alpha_l \sum_{i\in I} e_i \mathscr{C}_{k,i}}$, we obtain

$$p_k = \frac{e^{-\alpha_l \sum_{i\in I} e_i \mathscr{C}_{k,i}}}{\sum_{j\in K_l} e^{-\alpha_l \sum_{i\in I} e_i \mathscr{C}_{j,i}}} \frac{e^{\alpha_l \sum_{i\in I} e_i \mathscr{C}_{k,i}}}{e^{\alpha_l \sum_{i\in I} e_i \mathscr{C}_{k,i}}} = \frac{1}{\sum_{j\in K_l} e^{\alpha_l \sum_{i\in I} e_i (\mathscr{C}_{k,i} - \mathscr{C}_{j,i})}}.$$

This yields a system of equations that can be solved for $\{e_i\}_{i\in I}$. The inverse temperature parameters $\{\alpha_l\}_{1\leq l\leq L}$ may either be treated as free parameters or they may be the experimentally known temperatures of the mesoscopic interactions.

Instead of directly solving the above systems of equations, we recast the problem as a constrained version of multinomial regression. For this formulation, the inverse temperature parameters have to be known a priori. For biological applications we assume that interactions occur under "standard biological conditions" (7). Furthermore, without loss of generality, we assume that the set K is partitioned into blocks of equal size of cardinality M and indexed such that the first M mesoscopic interactions are in K_1, the second M mesoscopic interactions are in K_2, and so forth, yielding $\{k\}_{1\leq k\leq M} = K_1$, $\{k\}_{M+1\leq k\leq 2M} = K_2,\cdots,\{k\}_{M(L-1)+1\leq k\leq ML} = K_L$. No loss of generality results from such a partitioning because zero-probability mesoscopic interactions (e.g., all atoms are in the same position) can be used to pad canonical ensembles to be of equal size. Using this partitioning, each canonical ensemble K_l can then be treated as a single data point in an M-class multinomial regression problem, with the regression input vectors $\{X_l \in \mathbb{R}^{I(M-1)}\}_{1\leq l\leq L}$ having the form

$$X_l = \{\alpha_l(\mathscr{C}_{Ml,i} - \mathscr{C}_{M(l-1)+m,i})\}_{i\in I, 1\leq m\leq M-1}$$

Expanded:

$$X_l = \{\alpha_l(\mathscr{C}_{Ml,1} - \mathscr{C}_{M(l-1)+1,1}),\cdots,\alpha_l(\mathscr{C}_{Ml,I} - \mathscr{C}_{M(l-1)+1,I}),\alpha_l(\mathscr{C}_{Ml,1} - \mathscr{C}_{M(l-1)+2,1}),\cdots,\alpha_l(\mathscr{C}_{Ml,I} - \mathscr{C}_{Ml-1,I})\}.$$

And the design matrix is $\{X_l\}_{1\leq l\leq L} = X \in \mathbb{R}^{L\times I(M-1)}$. The output vectors $\{Y_l \in \mathbb{R}^{M-1}\}_{1\leq l\leq L}$ have the following form:

$$Y_l = \{p_{M(l-1)+m}\}_{1\leq m\leq M-1}.$$

Expanded:

$$Y_l = \{p_{M(l-1)+1},\cdots p_{Ml-1}\}.$$

And the output matrix is $\{Y_l\}_{1\leq l\leq L} = Y \in \mathbb{R}^{L\times(M-1)}$. Using this formulation, the multinomial logistic regression problem becomes

$$Y_{l,m} = \frac{e^{X_l\beta_m}}{1 + \sum_{n=1}^{M-1} X_l\beta_n} = \frac{e^{\sum_{p=1}^{I(M-1)} \beta_{m,p} X_{l,p}}}{1 + \sum_{n=1}^{M-1} e^{\sum_{p=1}^{I(M-1)} \beta_{n,p} X_{l,p}}},$$

where $\beta_{m,p}$ is the pth coefficient for the mth class in the regression coefficient matrix, X_l is indexed as previously described, and $Y_{l,m}$ is the output of the mth class for the lth data point. The Mth class is treated as the comparison category in the regression. We now introduce two constraints that, when enforced, make the statistical mechanical model derived earlier and multinomial logistic regression equivalent. The two constraints are

$$\forall\, m \in [1,M-1], p \notin [I(m-1)+1, Im]\colon \beta_{m,p} = 0$$
$$\forall\, m,n \in [1,M-1], i \in [1,I]\colon \beta_{m,I(m-1)+i} = \beta_{n,I(n-1)+i}.$$

The first constraint leads to cancellations that yield

$$Y_{l,m} = \frac{e^{\sum_{p=I(m-1)+1}^{Im} \beta_{m,p} X_{l,p}}}{1 + \sum_{n=1}^{M-1} e^{\sum_{p=I(n-1)+1}^{In} \beta_{n,p} X_{l,p}}}.$$

Using the second constraint and simplifying the notation by setting $\beta_{m,I(m-1)+i} = \beta'_i$ for all m and i, we obtain

$$Y_{l,m} = \frac{e^{\sum_{i=1}^{I} \beta_i' X_{l,I(m-1)+i}}}{1 + \sum_{n=1}^{M-1} e^{\sum_{i=1}^{I} \beta_i' X_{l,I(n-1)+i}}}.$$

Inserting the values for X_l as previously defined,

$$Y_{l,m} = \frac{e^{\sum_{i=1}^{I} \beta_i' \alpha_l (\mathscr{E}_{Ml,i} - \mathscr{E}_{M(l-1)+m,i})}}{1 + \sum_{n=1}^{M-1} e^{\sum_{i=1}^{I} \beta_i' \alpha_l (\mathscr{E}_{Ml,i} - \mathscr{E}_{M(l-1)+n,i})}}.$$

Dividing the numerator and denominator by the expression in the numerator we obtain

$$Y_{l,m} = 1 \Big/ \left(e^{\sum_{i=1}^{I} \beta_i' \alpha_l (\mathscr{E}_{M(l-1)+m,i} - \mathscr{E}_{Ml,i})} + \sum_{n=1}^{M-1} e^{\sum_{i=1}^{I} \beta_i' \alpha_l (\mathscr{E}_{Ml,i} - \mathscr{E}_{M(l-1)+n,i} - \mathscr{E}_{Ml,i} + \mathscr{E}_{M(l-1)+m,i})} \right)$$

$$Y_{l,m} = 1 \Big/ \left(e^{\sum_{i=1}^{I} \beta_i' \alpha_l (\mathscr{E}_{M(l-1)+m,i} - \mathscr{E}_{Ml,i})} + \sum_{n=1}^{M-1} e^{\sum_{i=1}^{I} \beta_i' \alpha_l (\mathscr{E}_{M(l-1)+m,i} - \mathscr{E}_{M(l-1)+n,i})} \right)$$

$$Y_{l,m} = 1 \Big/ \left(\sum_{n=1}^{M} e^{\alpha_l \sum_{i=1}^{I} \beta_i' (\mathscr{E}_{M(l-1)+m,i} - \mathscr{E}_{M(l-1)+n,i})} \right).$$

Finally by letting k be the mth mesoscopic interaction of the lth partition of K, and noting that $\forall\ l$: $K_l = [M(l-1)+1, Ml]$, we can rewrite the above expression as

$$Y_{l,m} = \frac{1}{\sum_{j \in K_l} e^{\alpha_l \sum_{i=1}^{I} \beta_i' (\mathscr{E}_{k,i} - \mathscr{E}_{j,i})}},$$

which yields the sought equivalence of the statistical mechanical model to multinomial logistic regression, where $p_k = Y_{l,m}$ and the inferred coefficient $\beta_i' = e_i$ for all $i \in I$. Thus using constrained logistic regression with the previously defined input vectors $\{X_l \in \mathbb{R}^{I(M-1)}\}_{1 \le l \le L}$ and output vectors $\{Y_l \in \mathbb{R}^{M-1}\}_{1 \le l \le L}$ will yield the energies of all microscopic interactions.

This is a key result in our formulation. Once we recognize that the de novo potential determination problem can be cast as a regression problem, we can consider applying the methods of compressed sensing (2) to the regression problem.

Applying compressed sensing to de novo potential determination. Compressed sensing is a theoretical and applied framework that enables the recovery of a signal with high accuracy from relatively few measurements. In the context of our regression formulation, the signal is the set of microscopic interaction energies $\{e_i\}_{i \in I}$, the sensors are mesoscopic interactions such as protein–DNA or protein–protein binding events, and the measurements are the energies or relative probabilities of the mesoscopic interactions. Compressed sensing enables the use of relatively few measurements to infer the energies of the microscopic interactions underlying the mesoscopic interactions.

The application of compressed sensing to regression is achieved by imposing an ℓ_1 regularization penalty on the regression problem (8). Several compressed sensing approaches have been developed to accomplish this task. Here, we utilize the lasso penalty for linear regression (8) and its generalization for logistic regression (9). Other possibilities include the Dantzig selector (10) for linear regression and suitable generalizations for logistic regression (11).

Our formulation casts potential determination as a compressed sensing problem. Accurate inference of energy potentials is ultimately dependent on two important characteristics of the particular application. These two characteristics are the sparsity of the microscopic potential to be inferred and the coherence of the mesoscopic measurements used for inference. The compressed sensing literature contains many theoretical guarantees about the expected number of measurements required, denoted by m, given a signal of length n with s nonzero entries and a sensor matrix X with coherence μ (the design matrix in regression). The majority of these results concern sensor matrices with some degree of randomness, such as the guarantees in Candes and Plan (12). In our specific application, the sensor matrix is fixed and defined by the set of protein:DNA structures in the dataset. Theoretical guarantees for fixed matrices are discussed in the papers by Candes and Romberg (13) and Tropp (14).

In general, most results relate m to an increasing function of n, s, and μ. In our context n is the total number of possible microscopic interactions, s is the number of microscopic interactions $I_0 \subset I$ whose energies are nonnegligible, i.e., for all $i \subset I_0$, $e_i \gg 0$, and μ is the coherence of the sensor matrix X. Thus, if only a small number of the microscopic interactions are nonnegligible (i.e., s is small), then only a small number of measurements are needed for accurate inference of the potential. The role that μ plays is slightly more subtle (the formal definition of μ can be found in ref. 2). Intuitively, coherence relates to the diversity of microscopic interactions that constitute each mesoscopic interaction. The greater the variety of microscopic interactions participating in each experimentally characterized mesoscopic interaction, the smaller the coherence. Thus, to minimize the number of experimental measurements needed, one needs to minimize the coherence of the sensor matrix. This implies that the mesoscopic interactions must incorporate as wide a range of microscopic interactions as possible. However, the degree to which the coherence of mesoscopic interactions can be controlled is dependent on the application. For biological systems, the choice of biomolecules comprising the set of mesoscopic interactions directly impacts the coherence of the sensor matrix.

In particular, for the protein–DNA application, the coherence of the sensor matrix is dependent on the set of protein–DNA complexes used as mesoscopic interactions. For a given set of protein–DNA complexes whose atomic structures have been determined, the coherence μ can be computed from the counts of atomic interactions present in the structures (2). This provides a principled approach for choosing new protein–DNA complexes for crystallization, based on their predicted atomic structures and the resulting value for μ. By carefully selecting protein–DNA complexes that yield the smallest predicted coherence of the sensor matrix, the accuracy of the inferred protein–DNA potential can be increased with a small number of targeted crystallization experiments.

II. De Novo Determination of Protein–DNA Potential. In this section the general methodology derived in section I will be applied to the specific problem of determining protein–DNA potentials. Several different approaches to potential determination will be discussed. We first delineate the details common to all inferred protein–DNA potentials and then describe the baseline potential and modified potentials that relax physical constraints.

Common details of all protein–DNA potentials. In all protein-DNA potentials in this section, the microscopic interactions are defined as pairwise contacts between DNA and protein atoms at specified distance ranges. Intra-DNA interactions, known as indirect readout, are ignored, as are intraprotein interactions. We use the atom classification scheme of the Quasichemical potential (15), which takes into consideration the chemical identity of

the atom as well as its local moiety. In this scheme there are 37 DNA atom types and 27 protein atom types. Contact distances are categorized into bins of fixed width (see *Metaparameters* below). Thus, a model with D distance bins contains a total of $37 \times 27 \times D$ interaction types. The potential inference problem is to determine the energy of each of these microscopic interaction types using experimental measurements of mesoscopic interactions, specifically protein–DNA-binding events. As described below, these binding events will be based on protein–DNA crystal structures and their associated binding energies.

For this inference problem, the multinomial logistic regression approach derived in Section I is used. The canonical ensembles are created by fixing the protein and varying the DNA sequence in silico by substitution of the alternative nucleic acids at each site in the interaction region as described in the next paragraph. We make the conventional assumption of independence of the DNA base pair positions in the interaction region, and we treat each set of protein–DNA base pair complexes derived from separate protein–DNA crystal structures as a separate canonical ensemble. Thus each ensemble is comprised of four mesoscopic interactions, corresponding to one protein binding to each of the four possible nucleotide base pairs. Previous studies have provided justification for treating such ensembles as canonical ensembles (16).

The experimental data we use to infer the microscopic energies for the protein–DNA-binding problem are crystallized protein:DNA structures in conjunction with their experimentally determined DNA-binding motifs. A single protein–DNA-binding event corresponds to a protein–DNA base pair complex. For the regression input vectors, the counts of the interaction types as defined above are required for each protein–DNA-binding event. We obtain these counts from the atomic coordinates of protein and DNA atoms in the X-ray crystal structures of protein:DNA complexes. For each protein:DNA crystal complex, we generate in silico structural mutants so that every base pair in each DNA position is in silico mutated to every possible alternative nucleotide. (For example, if there is an "A" at a particular position, we construct the three alternative structures with "T," "G," and "C" at that position.) Thus, for each protein:DNA complex with a DNA sequence of length N, we obtain a total of $4N$ structures which are grouped into N canonical ensembles as previously described.

For the regression output vectors, the relative probabilities of the mesoscopic interaction events within each canonical ensemble are required. These relative probabilities correspond to the probabilities with which a protein binds the four DNA nucleotides at a given position. When inferring a protein–DNA potential to predict consensus sequences from the crystal structures, we set the probability of binding to the nucleotide observed in the structure to 1 and the probability of binding to any other nucleotide to 0. When inferring a protein–DNA potential for predicting position weight matrices (PWMs), the probabilities are derived from experimentally determined PWMs that are manually curated for each protein:DNA complex in our dataset (see *Dataset* below). A PWM is a set of probability distributions, one for each base pair position in the DNA-binding site, over the four possible DNA nucleotides. The probability we assign to each protein–DNA-binding event is the weight of that event given in the experimental PWM.

All regressions are regularized using the lasso penalty (see *Applying compressed sensing to de novo potential determination* above). Additionally, we assume that all protein–DNA interactions occur under standard biological conditions so that there is a common temperature across all ensembles. Because the temperature is fixed, it can be factored out of the inferred microscopic energies, and so we set it equal to 1 to simplify the calculation.

Metaparameters. All protein–DNA potentials include two spatial metaparameters that require fitting: binning width and cutoff distance. Binning width corresponds to the resolution at which contact distances are discretized. Cutoff distance is the distance beyond which interactions are ignored. A third parameter, known as the regularization parameter λ (1), adjusts the expected degree of sparsity of the microscopic potential (see *Applying compressed sensing to de novo potential determination* above). Because the sparsity of the microscopic potential is not known a priori, this parameter must also be fitted.

To find the optimal combination of binning width, cutoff distance, and λ, we vary these three metaparameters and test the resulting potential on a criteria that minimizes prediction error (see *Testing Methodology* in Section III). Binning width is varied from 0.6 to 3 Å, in steps of 0.1 Å. Cutoff distance is varied from 2 to 20 Å, in steps equal to the binning width. For λ, the maximum value considered is the smallest value such that all input features in the model are weighted 0. The minimum value considered is equal to 20^{-3} times the maximum value. This is a standard approach that ensures that sufficiently small values are considered to include the regime where the model overfits (1).

Model solving. We perform ℓ_1-regularized multinomial logistic regression using the R-based glmnet package (9). Solutions for the constrained regression problem we derived earlier are feasible for the unconstrained regression problem. Consequently, the constraints are not enforced in the implementation as the unconstrained optimization is more computationally efficient and enforcing the constraints is unlikely to improve performance.

Variants of protein–DNA potentials. The previous subsection described details common to all protein–DNA potentials. We infer three variants of protein–DNA potentials in this work. This subsection will describe the specific choices and motivations for each variant.

Variant 1: The baseline. The baseline protein–DNA potential uses differences of counts of microscopic interactions as the regression input vectors, as previously derived in *Inference using relative probabilities of mesoscopic interactions*. Our formulation also requires that one state within each canonical ensemble, where the states correspond to the four nucleotides, is used as a comparison category. We arbitrarily choose guanine for this purpose. See *Inference using relative probabilities of mesoscopic interactions* for more details.

Variant 2: Transformed inputs. The "transformed inputs" protein–DNA potential uses ratio of counts of microscopic interactions as the regression input vectors. Specifically, the regression input vectors $\{X_l \in \mathbb{R}^{IM}\}_{1 \le l \le L}$ have the following form:

$$X_l = \left\{ \alpha_l \left(\frac{\mathscr{C}_{M(l-1)+m,i}}{\sum_{n=1}^{M} \mathscr{C}_{M(l-1)+n,i}} \right) \right\}_{i \in I, 1 \le m \le M}.$$

Expanded:

$$X_l = \left\{ \alpha_l \left(\frac{\mathscr{C}_{M(l-1)+1,1}}{\sum_{n=1}^{M} \mathscr{C}_{M(l-1)+n,1}} \right), \cdots, \alpha_l \left(\frac{\mathscr{C}_{M(l-1)+1,I}}{\sum_{n=1}^{M} \mathscr{C}_{M(l-1)+n,I}} \right), \right.$$
$$\left. \alpha_l \left(\frac{\mathscr{C}_{M(l-1)+2,1}}{\sum_{n=1}^{M} \mathscr{C}_{M(l-1)+n,1}} \right), \cdots, \alpha_l \left(\frac{\mathscr{C}_{Ml,I}}{\sum_{n=1}^{M} \mathscr{C}_{M(l-1)+n,I}} \right) \right\}.$$

In this protein–DNA potential all elements of the regression input vectors are also standardized such that their mean is zero and variance is one. Input vectors with this property typically exhibit better statistical properties (1).

Variant 3: Region-specific potentials. In the "region-specific" protein–DNA potential, the energies of microscopic interactions are allowed to vary as a function of position along the protein–DNA-binding interface. The motivation behind this generalization will be addressed first, followed by a description of how region specificity is incorporated into the protein–DNA potential.

Potentials typically define interaction energies independently of the absolute spatial position of the interaction. This is a desirable property when the position has no bearing on the energetics. However, if the position of the interaction is indicative of a consistent physico-chemical environment, properties of this environment can be exploited when modeling the energies of interactions. This is usually exploited on a small scale, for example, by treating two carbon atoms as distinct atom types depending on their chemical moiety, as in the C1′ and C2′ atoms of DNA (15). In that case, the carbon atom's local chemical moiety is the consistent physico-chemical environment, and it is used in modeling the system to influence the interaction energies of the atoms, by treating carbon atoms with distinct chemical moieties as distinct atom types. For protein–DNA potentials, we asked whether this notion can be generalized to a larger spatial context, by deriving a potential where the microscopic interaction energies are allowed to vary as a function of position along the protein:DNA-binding interface. A key requirement for such a potential to work is that any given position along the binding interface must exhibit a consistent physico-chemical environment across structures, for example, to have essentially the same steric constraints. For the protein–DNA potential, we hypothesized that this is the case when a set of proteins employs 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 (17). We conjectured that a region-specific model may also be suitable for the helix-turn-helix (HTH) family, because (*i*) the sequence specificity of HTHs is largely mediated by interactions between the DNA and the recognition α helix of HTHs (18) 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 (18–20).

To accomplish this, we structurally aligned all 63 HTH:DNA complexes in our dataset (see *Dataset*) so that the DNA molecules are superimposed and the variation in the orientation and position of the recognition helices is minimized. Formally, if rmsd_{DNA} is the rmsd between the backbone carbon atoms of two aligned DNA molecules and rmsd_{HTH} is the rmsd between the C^{α} atoms of two aligned HTH recognition helices, then we solve the following optimization problem for all pairwise comparisons of HTH:DNA complexes:

$$\min_{\text{alignments}} \ \text{rmsd}_{\text{HTH}}$$

$$\text{subject to } \text{rmsd}_{\text{DNA}} < \delta,$$

where δ is a parameter of the algorithm (a value of 2 Å is used throughout). Any HTH:DNA complex in the dataset can be used as a baseline for a multiple alignment, so we select the complex that minimizes the average rmsd_{HTH} to all other complexes. The resulting multiple alignment, shown in Fig. S3, produces a unified coordinate system along the HTH:DNA-binding interface such that the physico-chemical environment at a given spatial position is comparable across all structures.

To exploit this alignment in the determination of the protein–DNA potential, the resolution of the coordinate system is reduced to the level of individual DNA base pairs. In other words, the number of distinct positions in the coordinate system is made equal to the number of distinct DNA base pair positions in the HTH-DNA-binding interface. In this way, a DNA base in one HTH:DNA complex is comparable with exactly one other DNA base in every other complex. In total there are 13 distinct DNA base pair positions, encompassing the largest observed HTH:DNA-binding interfaces in our dataset.

To incorporate region specificity into the inferred protein–DNA potential, the elements of every regression input vector are duplicated 13 times. Formally, the regression input vectors have the following form:

$$X_l^{(n)} = \{\alpha_l(\mathscr{C}_{Ml,i} - \mathscr{C}_{M(l-1)+m,i})\}_{i\in I, 1\le m\le M-1} \cup \{1\{r=n\}\alpha_l(\mathscr{C}_{Ml,i} - \mathscr{C}_{M(l-1)+m,i})\}_{i\in I, 1\le m\le M-1, 1\le r\le 13}.$$

In the above formulation, the original set of elements in the input vectors are left unchanged, and the nth duplicate set in an ensemble representing the nth DNA base pair is also left unchanged. All other vector elements are set to 0. $X_l^{(n)}$ is a canonical ensemble that corresponds to a binding event at the nth DNA basepair position, and $1\{r=n\}$ is the indicator function testing whether r equals n. This encoding of the regression input vectors allows the inference algorithm to model the shared interaction energies common across all positions, as well as capture position-specific adjustments to the core potential.

III. Application of De Novo Protein–DNA Potentials to the Prediction of Protein–DNA-Binding Sites. In this section the de novo protein–DNA potentials described in Section II are used to predict the DNA-binding sites of proteins. This is accomplished by first using the inferred protein–DNA potentials to compute the binding energies of the protein:DNA structures, and then using those computed energies to predict the DNA-binding sites of proteins. We first describe the structure-based approach to binding site prediction and then detail the testing methodology and metrics used.

Structure-based prediction of protein DNA-binding sites. Structure-based methods for the prediction of protein–DNA-binding sites follow a somewhat standardized approach when predicting the DNA-binding affinity of proteins (15, 17, 21–23). Starting with a protein:DNA X-ray crystal structure, structure-based methods transform the structure into a set of microscopic interactions (in our case protein atom to DNA atom contacts using the atom-type categories discussed earlier). In some approaches, the structure is first relaxed using a molecular mechanics force field (24), but we do not employ this step in our approach. Once the set of microscopic interactions is identified, the overall binding energy of a protein:DNA structure is computed by adding up the individual energetic contributions from all the microscopic interactions observed in the structure. How the interaction energies are defined depends on the choice of potential used; in our case, we use the de novo protein–DNA potentials described in section II to compute the energies. Formally, if I is the set of all interactions observed in a structure, e_i is the energetic contribution of interaction i, $\mathscr{C}_i$ is the number of times interaction i is observed in the structure, then the binding energy of the structure, denoted by ΔG, is

$$\Delta G = \sum_{i\in I} e_i \mathscr{C}_i.$$

The relative 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 fixing the protein in the protein:DNA complex and mutating the DNA sequence in silico as described next. To make this problem computationally tractable, it is often assumed that the energetics for one DNA position in the binding interface can be computed separately from other positions (15, 17, 21–23). We adopt this assumption. Doing so simplifies the computation by requiring for a DNA sequence of length N only $4N$ energetic calculations, where each base pair

in the DNA is mutated in silico to every possible nucleotide independently from other base pairs, and the binding energy of each of these in silico mutagenized protein:DNA structures is computed. Using the computed binding energies, the Boltzmann formula (16) can then be used to compute the probability of observing nucleotide m at position n, denoted by $p_m^{(n)}$:

$$p_m^{(n)} = \frac{e^{-\beta \Delta G_m^{(n)}}}{\sum_{k \in \{A,C,G,T\}} e^{-\beta \Delta G_k^{(n)}}},$$

where $\Delta G_m^{(n)}$ denotes the binding energy of the structure when position n is mutated to nucleotide m, and β is the inverse temperature parameter that is set to unity in our computations. Performing this computation for every position n and every nucleotide m yields the predicted PWM for the DNA-binding sites of the protein. When the consensus sequence is to be predicted, the most probable nucleotide at every position is chosen as the consensus nucleotide.

Note that our approach does not require the energy computation step because the probabilities can be computed directly using the inferred regression model. This is done by first converting each position in a protein:DNA structure into the appropriate regression input vector, as we did previously for the potential inference step, and then using the regression coefficient matrix to compute the binding probabilities. We use this approach for all of our DNA-binding site predictions, including the computations that produce the results in Figs. 3 and 4 and Table 1 of the main text.

Testing methodology. Quality metrics. We used two metrics to test the quality of predictions made by our de novo potentials applied to the protein–DNA-binding site prediction problem. For consensus binding sites, we compute the fraction of bases correctly predicted by the algorithm for a given protein. To count as correct a predicted base must exactly match the identity of the base in the protein:DNA X-ray structure.

To test the performance of the algorithm in predicting position weight matrices for protein–DNA-binding sites, we compare the algorithm's predictions to published experimentally determined PWMs. To compare two PWMs, a distance measure over their probability distributions is required to assess the "closeness" of the prediction. The most commonly used such measure is the symmetric Kullback–Leibler divergence (SKLD) (25). The SKLD between two PWMs P and Q is defined as

$$\mathrm{SKLD}(P,Q) = \frac{1}{N}\sum_{n=1}^{N}\sum_{m \in \{A,C,G,T\}} p_m^{(n)} \ln\frac{p_m^{(n)}}{q_m^{(n)}} + q_m^{(n)} \ln\frac{p_m^{(n)}}{q_m^{(n)}},$$

where $p_m^{(n)}$ and $q_m^{(n)}$ are the probabilities of observing nucleotide m at position n in P and Q, respectively.

Dataset. We obtained a set of HTH:DNA complex structures from the Protein Data Bank (26) by searching for X-ray crystal structures that contain an HTH domain and DNA molecules. We considered complexes to be redundant if they shared the same sequence of amino acids within a 10-residue window of the recognition α-helix, and we retained only one representative for such complexes. We chose this criterion due to the dominant rule 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 (27, 28). In addition, we removed complexes with pathologies such as a large number of missing heavy atoms in the published structure. The resulting dataset is comprised of 63 nonredundant HTH:DNA complexes (Dataset S1). For each of the HTH:DNA complexes in this dataset, a PWM was derived based on experimental data curated from multiple sources (Dataset S2) (23, 29–72).

Validation. To test the performance of our models, we used 9-fold cross-validation. The 63 complexes were split into 9 groups of 7 complexes each. Each group yielded a testing configuration where 7 complexes comprise the test set, and the remaining 56 the training set. The model was trained on the training set and tested on the test set using all 9 testing configurations, and the average was taken over all test sets. Because some metaparameters vary the number of degrees of freedom in the model, fitting them strictly on the training set was not possible, as the algorithm would always maximize the number of degrees of freedom. Consequently the 3 metaparameters were fit by finding the value that minimizes average error over all test sets. Results are shown in Figs. 3 and 4 and Table 1 in the main text.

Assessment of other methods. Our objective was to create a consistent testing environment for all methods considered for comparison with our algorithm. When original code implementations were available, they were used; otherwise, the method was reimplemented, as for the Quasichemical potential (15). Recommended default settings were chosen when applicable, and all parameter settings fit using the original methods' datasets were left unchanged. For the Rosetta potential, the energetic terms were set to the following values: $\mathrm{fa_{atr}} = 0.947733$, $\mathrm{fa_{rep}} = 0.577238$, $\mathrm{hb_{sc}} = 1.596235$, $\mathrm{gb_{elec}} = 0.203353$, $\mathrm{fa_{sol}} = 0.507356$, $\mathrm{dna_{bs}} = 0.1$, and $\mathrm{dna_{bp}} = 0.1$. These values are based on the fitted parameters from ref. 23. Depending on the original datasets used for training each algorithm, some methods may have an advantage over other methods if their training dataset included structures that were also in our curated test sets. Nonetheless, given the heterogeneous set of methods tested and their reliance on different types of input data, the methodology used provided a consistent and objective comparison.

When testing the consensus sequence predictions of the Rosetta (23) and DNAPROT (21) algorithms, their indirect readout components had to be disabled as they explicitly rely on the consensus sequence itself as input. For Rosetta, this was done by setting the dna_bs and dna_bp terms to zero weight, to eliminate the indirect readout component from the potential. For DNAPROT, this was done by running it with the setting "-p′ -P -1 -e -c -D 0′", which disables DNAPROT's indirect readout component and its Cumulative Contacts (73) component, both of which rely on the consensus sequence as input.

1. Hastie T, Tibshirani R, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York), 2nd Ed, p xxii.
2. Candes EJ, Wakin MB (2008) An introduction to compressive sampling. *IEEE Signal Process Mag* 25:21–30.
3. Donoho DL (2006) Compressed sensing. *IEEE Trans Inf Theory* 52:1289–1306.
4. McQuarrie DA (2000) *Statistical Mechanics* (University Science Books, Sausalito, CA), p xii.
5. Reif F (1965) *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, New York), p x.
6. Brändén C-I, Tooze J (1999) *Introduction to Protein Structure* (Garland, New York), 2nd Ed, p xiv.
7. Berg JM, Tymoczko JL, Stryer L (2007) *Biochemistry* (Freeman, New York), 6th Ed.
8. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 58:267–288.
9. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 33:1–22.
10. Candes E, Tao T (2007) The Dantzig selector: Statistical estimation when p is much larger than n. *Ann Stat* 35:2313–2351.
11. James GM, Radchenko P (2009) A generalized Dantzig selector with shrinkage tuning. *Biometrika* 96:323–337.
12. Candes EJ, Plan Y (2010) A probabilistic and RIPless theory of compressed sensing. *IEEE Trans Inf Theory* (in press).
13. Candes EJ, Romberg J (2006) Quantitative robust uncertainty principles and optimally sparse decompositions. *Found Comput Math* 6:227–254.

14. Tropp JA (2008) On the conditioning of random subdictionaries. *Appl Comput Harmon Anal* 25:1–24.
15. Donald JE, Chen WW, Shakhnovich EI (2007) Energetics of protein-DNA interactions. *Nucleic Acids Res* 35:1039–1047.
16. Berg OG, Vonhippel PH (1987) Selection of DNA-binding sites by regulatory proteins—statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193:723–743.
17. Kaplan T, Friedman N, Margalit H (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol* 1:e1.
18. Wintjens R, Rooman M (1996) Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J Mol Biol* 262:294–313.
19. Suzuki M, Gerstein M (1995) Binding geometry of alpha-helices that recognize DNA. *Proteins* 23:525–535.
20. Pabo CO, Nekludova L (2000) Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition? *J Mol Biol* 301:597–624.
21. Angarica VE, Perez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics* 9:436.
22. Moroni E, Caselle M, Fogolari F (2007) Identification of DNA-binding protein target sequences by physical effective energy functions: Free energy analysis of *lambda* repressor-DNA complexes. *BMC Struct Biol* 7:61.
23. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* 33:5781–5798.
24. Jorgensen WL, Tirado-Rives J (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc Natl Acad Sci USA* 102:6665–6670.
25. Hastie T (1987) A closer look at the deviance. *Am Stat* 41:16–20.
26. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28:235–242.
27. Gajiwala KS, Burley SK (2000) Winged helix proteins. *Curr Opin Struc Biol* 10:110–116.
28. Mo Y, Vaessen B, Johnston K, Marmorstein R (2000) Structure of the elk-1-DNA complex reveals how DNA-distal residues affect ETS domain recognition of DNA. *Nat Struct Biol* 7:292–297.
29. Matys V, et al. (2006) TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34:D108–110.
30. Kazakov AE, et al. (2007) RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res* 35: D407–D412.
31. Halfon MS, Gallo SM, Bergman CM (2008) REDfly 2.0: An integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*. *Nucleic Acids Res* 36:D594–598.
32. Portales-Casamar E, et al. (2010) JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 38:D105–110.
33. Newburger DE, Bulyk ML (2009) UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 37:D77–D82.
34. Munch R, et al. (2003) PRODORIC: Prokaryotic database of gene regulation. *Nucleic Acids Res* 31:266–269.
35. Gama-Castro S, et al. (2008) RegulonDB (version 6.0): Gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 36:D120–124.
36. Sierro N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: A database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* 36:D93–96.
37. Jagannathan V, Roulet E, Delorenzi M, Bucher P (2006) HTPSELEX—a database of high-throughput SELEX libraries for transcription factor binding sites. *Nucleic Acids Res* 34:D90–94.
38. Down TA, Bergman CM, Su J, Hubbard TJ (2007) Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput Biol* 3:e7.
39. Palaniswamy SK, et al. (2006) AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* 140:818–829.
40. Bulow L, Engelmann S, Schindler M, Hehl R (2009) AthaMap, integrating transcriptional and post-transcriptional data. *Nucleic Acids Res* 37:D983–986.
41. Kumar MDS, et al. (2006) ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res* 34:D204–D206.
42. Yellaboina S, Ranjan S, Chakhaiyar P, Hasnain SE, Ranjan A (2004) Prediction of DtxR regulon: Identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*. *BMC Microbiol* 4:38.
43. Franks AH, Griffiths AA, Wake RG (1995) Identification and characterization of new DNA-replication terminators in *Bacillus subtilis*. *Mol Microbiol* 17:13–23.
44. Griffiths AA, Wake RG (1997) Search for additional replication terminators in the *Bacillus subtilis* 168 chromosome. *J Bacteriol* 179:3358–3361.
45. Griffiths AA, Andersen PA, Wake RG (1998) Replication terminator protein-based replication fork-arrest systems in various *Bacillus* species. *J Bacteriol* 180:3360–3367.
46. Sugisaki H, Kanazawa S (1981) New restriction endonucleases from *Flavobacterium okeanokoites* (FokI) and *Micrococcus luteus* (MluI). *Gene* 16:73–78.
47. Falvey E, Grindley NDF (1987) Contacts between gamma-delta-resolvase and the gamma-delta-res site. *EMBO J* 6:815–821.
48. Moskowitz IP, Heichman KA, Johnson RC (1991) Alignment of recombination sites in Hin-mediated site-specific DNA recombination. *Genes Dev* 5:1635–1645.
49. Rosandic M, et al. (2006) CENP-B box and pJ alpha sequence distribution in human alpha satellite higher-order repeats (HOR). *Chromosome Res* 14:735–753.
50. Tronche F, Yaniv M (1992) HNF1, a homeoprotein member of the hepatic transcription regulatory network. *Bioessays* 14:579–587.
51. Liston DR, Johnson PJ (1999) Analysis of a ubiquitous promoter element in a primitive eukaryote: early evolution of the initiator element. *Mol Cell Biol* 19:2380–2388.
52. Shen WF, et al. (1997) AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol Cell Biol* 17:6448–6458.
53. Kostelidou K, Thomas CM (2000) 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* 295:411–422.
54. Garcia-Castellanos R, et al. (2004) On the transcriptional regulation of methicillin resistance—MecI repressor in complex with its operator. *J Biol Chem* 279:17888–17896.
55. Colloms SD, van Luenen HG, Plasterk RH (1994) DNA binding activities of the Caenorhabditis elegans Tc3 transposase. *Nucleic Acids Res* 22:5548–5554.
56. Prakash P, Yellaboina S, Ranjan A, Hasnain SE (2005) Computational prediction and experimental verification of novel IdeR binding sites in the upstream sequences of Mycobacterium tuberculosis open reading frames. *Bioinformatics* 21:2161–2166.
57. Wilson DS, Guenther B, Desplan C, Kuriyan J (1995) High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA. *Cell* 82:709–719.
58. Hughes KT, Gaines PCW, Karlinsey JE, Vinayak R, Simon MI (1992) Sequence-specific interaction of the Salmonella Hin recombinase in both major and minor grooves of DNA. *EMBO J* 11:2695–2705.
59. Hoey T, Levine M (1988) Divergent homeo box proteins recognize similar DNA sequences in Drosophila. *Nature* 332:858–861.
60. White CE, Winans SC (2007) 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* 64:245–256.
61. Harbison CT, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.
62. Chen SF, et al. (2001) 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* 314:75–82.
63. Koudelka GB, Lam CY (1993) Differential recognition of OR1 and OR3 by bacteriophage 434 repressor and Cro. *J Biol Chem* 268:23812–23817.
64. Koudelka GB, Harrison SC, Ptashne M (1987) Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature* 326:886–888.
65. Schumacher MA, Lau AOT, Johnson PJ (2003) Structural basis of core promoter recognition in a primitive eukaryote. *Cell* 115:413–424.
66. Smale ST, et al. (1998) The initiator element: A paradigm for core promoter heterogeneity within metazoan protein-coding genes. *Cold Spring Harb Symp Quant Biol* 63:21–31.
67. Lo K, Smale ST (1996) Generality of a functional initiator consensus sequence. *Gene* 182:13–22.
68. Javahery R, Khachi A, Lo K, Zenziegregory B, Smale ST (1994) DNA-sequence requirements for transcriptional initiator activity in mammalian-cells. *Mol Cell Biol* 14:116–127.
69. Huerta AM, Francino MP, Morett E, Collado-Vides J (2006) Selection for unequal densities of sigma(70) promoter-like signals in different regions of large bacterial genomes. *PloS Genet* 2:1740–1750.
70. Fischer SEJ, van Luenen HGAM, Plasterk RHA (1999) Cis requirements for transposition of Tc1-like transposons in C. elegans. *Mol Gen Genet* 262:268–274.
71. Rodgers DW, Harrison SC (1993) The complex between phage 434 repressor DNA-binding domain and operator site OR3: structural differences between consensus and non-consensus half-sites. *Structure* 1:227–240.
72. van Luenen HGAM, Plasterk RHA (1994) Target site choice of the related transposable elements Tc1 and Tc3 of Caenorhabditis-Elegans. *Nucleic Acids Res* 22:262–269.
73. Morozov AV, Siggia ED (2007) Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci USA* 104:7068–7073.

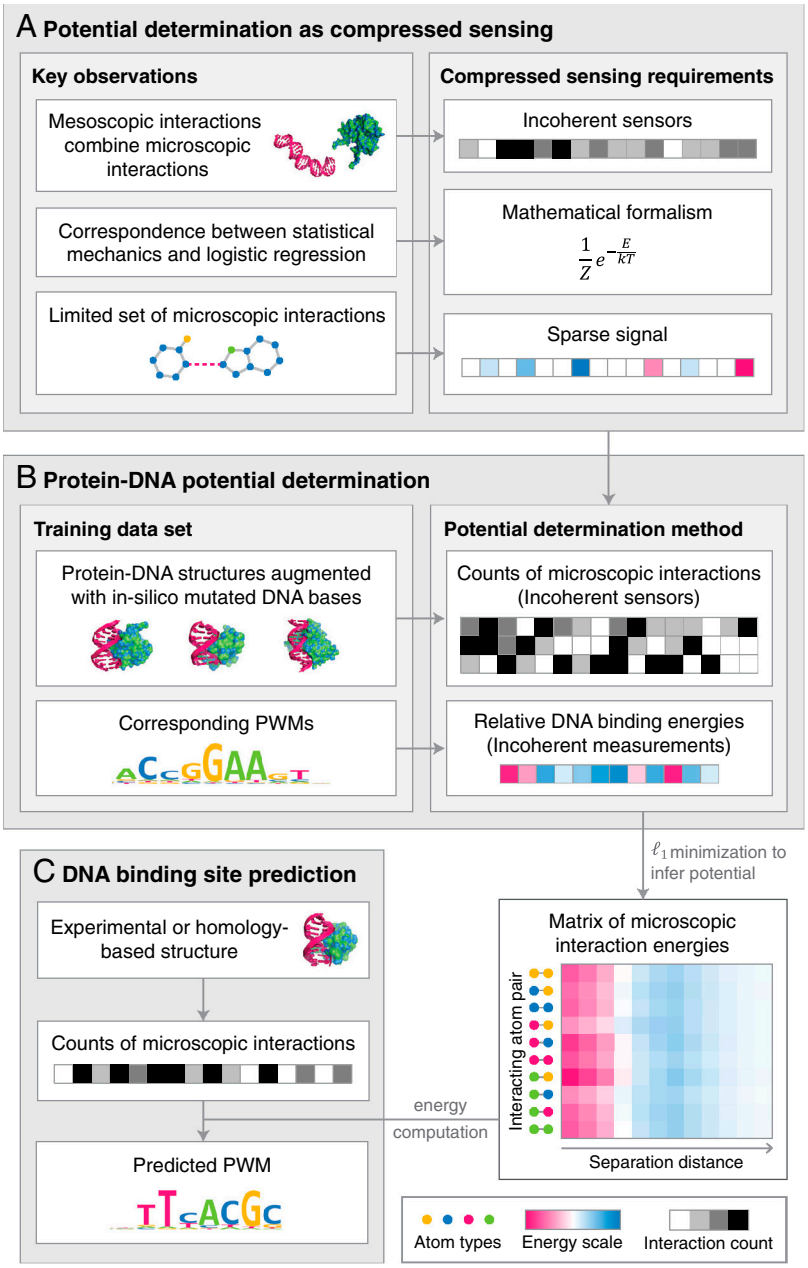

Fig. S1. Overview. (*A*) Three key observations are combined to cast the determination of energy potentials from crystal structures of molecular complexes as a compressed sensing problem: (*i*) Mesoscopic molecular interactions arise from a large number of microscopic interactions, enabling molecular complexes to be treated as incoherent sensors. (*ii*) The correspondence between the mathematics of statistical mechanics and logistic regression facilitates a compressed sensing formulation that uses relative binding probabilities as measurements. (*iii*) Biomolecular interactions are mediated by a small number of energetically dominant atomic-level interactions, resulting in a sparse vector of microscopic interaction energies (sparse signal). These three observations satisfy the requirements of compressed sensing and underlie our method for inferring energy potentials exclusively from experimental data. (*B*) Determination of de novo potential of protein–DNA interactions. Two types of training data are used in inferring the potential: (*i*) protein–DNA crystal structures, which serve as incoherent sensors of the microscopic interactions in a protein-DNA-binding event, and (*ii*) experimental protein–DNA-binding probabilities, which serve as incoherent measurements of the binding energies of the mesoscopic protein–DNA-binding events. Because the crystal structures of protein–DNA complexes are typically determined only for a single DNA sequence, whereas proteins bind with varying affinity to a family of sequences, we augment the dataset with structures in which the cognate DNA sequence is in silico mutated to alternative DNA sequences. The microscopic interaction energies for each distinct pair of protein atom and DNA atom are inferred from these datasets. (*C*) Using the inferred microscopic protein-DNA potential, the PWMs characterizing the DNA-binding sites of other proteins can be computed.

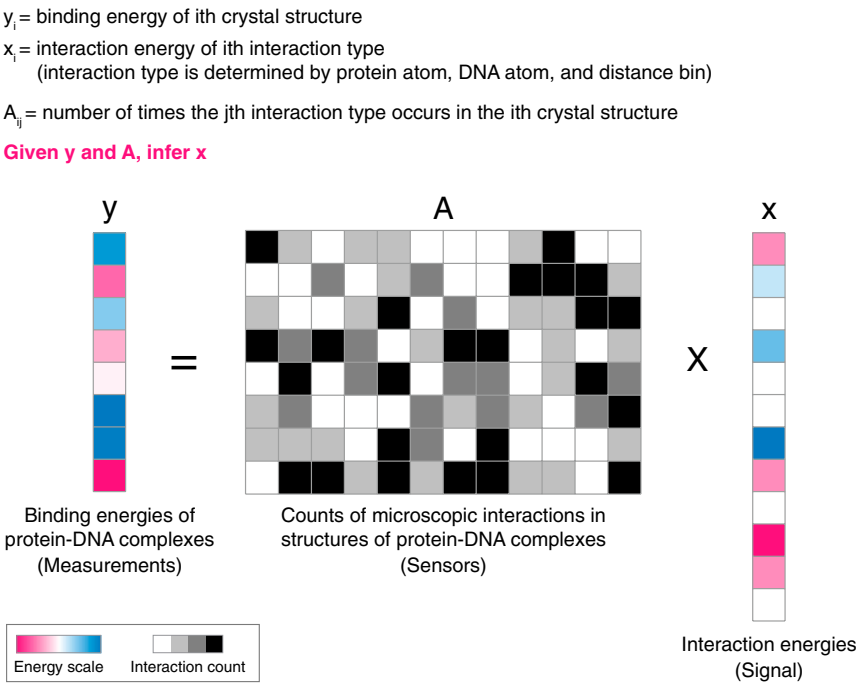

Fig. S2. Casting potential determination as a compressed sensing problem. In compressed sensing, the objective is to infer an unknown signal, represented by the vector *x*, from a set of measurements represented by the vector *y*. The relationship between the measurements and the signal is assumed to satisfy the equation $y = A \times x$; i.e., each measurement y_i is formed by the inner product of a row A_i with the vector *x*. Thus each row of *A* represents a distinct sensor vector. The number of measurements available (i.e., the length of *y*) is typically much smaller than the length of the signal vector *x*, which would make the equation $y = A \times x$ impossible to solve in general. However, for a sparse signal vector *x*, compressed sensing techniques enable inference of the original signal. In the protein–DNA application, the signal is the vector of microscopic interaction energies (protein–DNA potential), where each entry in the signal vector *x* corresponds to the energy of an interaction between a protein atom type and a DNA atom type within a discrete distance bin. Each row of *A* arises from a distinct protein–DNA crystal structure, and each column corresponds to a distinct type of microscopic interaction (combination of protein atom type, DNA atom type, and distance bin). An element A_{ij} of *A* encodes the number of times (indicated by gray levels) that the *j*th interaction type occurs in the *i*th crystal structure. The set of measurements *y* are the experimental binding energies of the corresponding protein–DNA complexes. Because the binding energy of a complex is the sum of the energies of all microscopic interactions in the complex, it is equal to the inner product between the row of *A* that encodes the complex and the signal vector *x*.

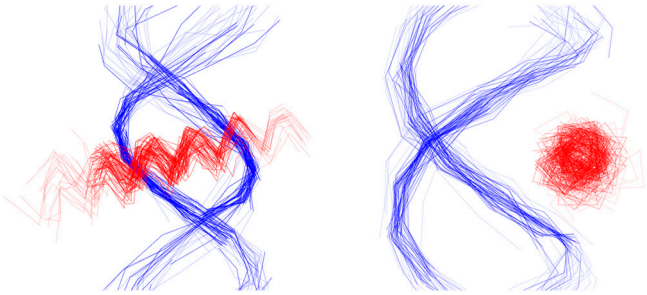

Fig. S3. Two views of 63 structurally aligned HTH:DNA complexes. The C_α traces of recognition α helices are shown in red and $C_{3'}$ traces of DNA helices are shown in blue.

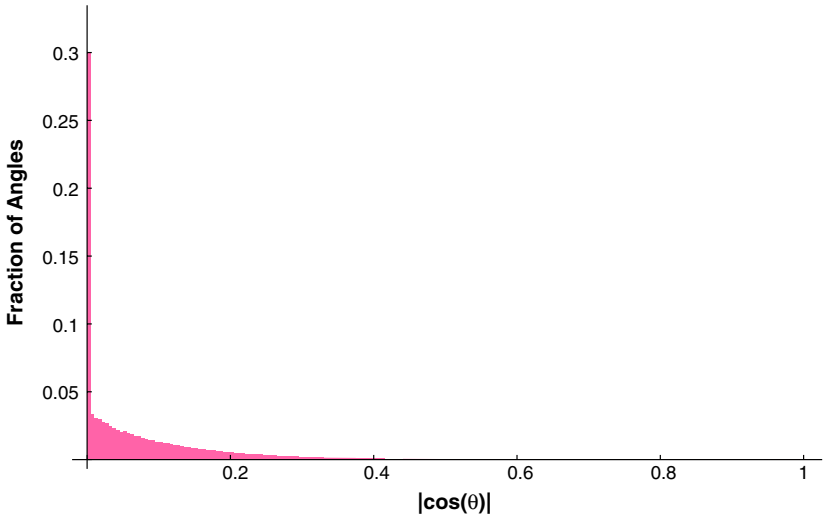

Fig. S4. Distribution of the absolute values of the cosines of angles between sensor vectors. For every pair of sensor vectors derived from the dataset and used in the best performing "Baseline" model (see Table 1 in the main text), the acute angle θ was calculated and used to compute the histogram of $|\cos\theta|$ values shown. This distribution characterizes the incoherence of the sensor matrix used. For an effective sensor matrix, it is desirable to have as many of the values be close to 0 as possible. The observed distribution suggests that the dataset used can act as an effective sensor matrix, with mean and median values of 0.081 and 0.041 (in units of $|\cos\theta|$), respectively.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)