Quantitative modeling of transcription factor binding specificities using DNA shape

Tianyin Zhou[a,1], Ning Shen[b,c,1], Lin Yang[a], Namiko Abe[d], John Horton[c,e], Richard S. Mann[d,f], Harmen J. Bussemaker[f,g], Raluca Gordân[c,e,2], and Remo Rohs[a,2]

[a]Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089; Departments of [b]Pharmacology and Cancer Biology and [e]Biostatistics and Bioinformatics and [c]Center for Genomic and Computational Biology, Duke University, Durham, NC 27708; Departments of [d]Biochemistry and Molecular Biophysics and [f]Systems Biology, Columbia University, New York, NY 10032; and [g]Department of Biological Sciences, Columbia University, New York, NY 10027

Edited by Steven Henikoff, Fred Hutchinson Cancer Research Center, Seattle, WA, and approved February 13, 2015 (received for review November 18, 2014)

DNA binding specificities of transcription factors (TFs) are a key component of gene regulatory processes. Underlying mechanisms that explain the highly specific binding of TFs to their genomic target sites are poorly understood. A better understanding of TF–DNA binding requires the ability to quantitatively model TF binding to accessible DNA as its basic step, before additional in vivo components can be considered. Traditionally, these models were built based on nucleotide sequence. Here, we integrated 3D DNA shape information derived with a high-throughput approach into the modeling of TF binding specificities. Using support vector regression, we trained quantitative models of TF binding specificity based on protein binding microarray (PBM) data for 68 mammalian TFs. The evaluation of our models included cross-validation on specific PBM array designs, testing across different PBM array designs, and using PBM-trained models to predict relative binding affinities derived from in vitro selection combined with deep sequencing (SELEX-seq). Our results showed that shape-augmented models compared favorably to sequence-based models. Although both *k*-mer and DNA shape features can encode interdependencies between nucleotide positions of the binding site, using DNA shape features reduced the dimensionality of the feature space. In addition, analyzing the feature weights of DNA shape-augmented models uncovered TF family-specific structural readout mechanisms that were not revealed by the DNA sequence. As such, this work combines knowledge from structural biology and genomics, and suggests a new path toward understanding TF binding and genome function.

protein–DNA recognition | statistical machine learning | support vector regression | protein binding microarray | DNA structure

The mechanisms by which transcription factors (TFs) bind to their genomic target sites and regulate gene expression are still not well understood (1, 2). In particular, it is still unknown why a given TF binds only to a subset of putative binding sites in the genome and how these targets are selected. Studies have identified multiple factors that play roles in achieving the DNA binding specificity of TFs in vivo, including cofactors, cooperativity, and chromatin accessibility (3).

A fundamental first step toward understanding TF binding is to describe the recognition of "naked" DNA by TFs in vitro. This process involves readout of the nucleotide sequence (4, 5) and three-dimensional (3D) DNA structure (6–8). Although DNA structural features have been discussed qualitatively as determinants of TF binding (9–11), quantitative models describing the impact of DNA shape on the strength of TF binding have received less attention.

Experimental high-throughput (HT) assays, such as protein-binding microarrays (PBMs) (12), measure the in vitro binding preferences of TFs to tens of thousands of different nucleotide sequences. Sequence-based modeling of in vitro TF binding specificities using HT data has been a topic of broad interest (13–16). Several methods for PBM data analysis have been developed. Recently, the performances of 26 sequence-based methods for predicting DNA binding specificity were assessed, based on PBM data for 66 mouse TFs (17) generated by the fifth dialogue for reverse engineering assessments and methods (DREAM5).

DNA sequence preferences are generally represented as position weight matrices (PWMs) (4) or position-specific affinity matrices (18). The original concept of the PWM assumed that a position within a binding site contributes to binding affinity independent of other positions (19). This concept has recently been expanded to include dinucleotide features that encode dependencies between adjacent nucleotide positions (e.g., ref. 20). Trinucleotides (21, 22) and higher-order *k*-mer features (23, 24), defined as all possible sequences of length *k*, have also been included in models of DNA sequence specificity. However, model complexity can increase dramatically when such *k*-mer features are used (21). Interdependencies between nucleotide positions originate from physical interactions between base pairs, which give rise to the 3D DNA structure. DNA-binding proteins, in turn, recognize the resulting DNA structure (10). Thus, the

Significance

Genomes provide an abundance of putative binding sites for each transcription factor (TF). However, only small subsets of these potential targets are functional. TFs of the same protein family bind to target sites that are very similar but not identical. This distinction allows closely related TFs to regulate different genes and thus execute distinct functions. Because the nucleotide sequence of the core motif is often not sufficient for identifying a genomic target, we refined the description of TF binding sites by introducing a combination of DNA sequence and shape features, which consistently improved the modeling of in vitro TF–DNA binding specificities. Although additional factors affect TF binding in vivo, shape-augmented models reveal binding specificity mechanisms that are not apparent from sequence alone.

Author contributions: T.Z., R.G., and R.R. conceived the study; T.Z. designed the sequence-shape feature vectors and implemented and executed the SVR analyses; N.S. analyzed PBM and SELEX-seq data; L.Y. conceived the second-order shape features and analyzed crystal structures; N.A. performed and analyzed SELEX-seq experiments; J.H. performed gcPBM experiments; R.S.M. and H.J.B. analyzed SELEX-seq data; H.J.B. conceived validation tests; R.G. and R.R. directed the study; and T.Z., N.S., R.G., and R.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession nos. GSE59845 and GSE60200).

See Commentary on page 4516.

[1]T.Z. and N.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: rohs@usc.edu or raluca.gordan@duke.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1422023112/-/DCSupplemental.

large number of *k*-mer features necessary for encoding interdependencies between nucleotide positions could potentially be replaced by a smaller number of structural features.

Recently, we developed technology enabling augmentation of the nucleotide sequence with 3D DNA shape features predicted using a pentamer-based model built from all-atom Monte Carlo simulations of DNA structures (25). Four distinct shape features—Minor Groove Width (MGW), Propeller Twist (ProT), Roll, and Helix Twist (HelT)—had been shown to be important for protein–DNA recognition in specific cases (22, 26–28). However, to date, a systematic and comprehensive survey of the value of shape-based models of DNA recognition has been lacking.

Here, we used HT protein–DNA binding data for 68 mammalian TFs from different structural classes to develop and evaluate DNA-binding specificity models based on different combinations of sequence- and shape-based features, including mononucleotide (1-mer), dinucleotide (2-mer), and trinucleotide (3-mer) identity, as well as the DNA shape features of MGW, ProT, Roll, and HelT. We used support vector regression (SVR) (29) with linear kernel to train regression models for mapping DNA sequences to measurements of binding affinity. Then, we evaluated these models using four different approaches.

First, we used 10-fold cross-validation on a diverse set of universal PBM (uPBM) data for 65 mouse TFs (17) and genomic context PBM (gcPBM) data for three human basic helix–loop–helix (bHLH) TFs. Second, in a manner similar to ref. 17, we trained our models using PBM data from one uPBM array design and tested the models using data from a different uPBM array design (see *SI Methods* for details). Third, we tested the ability of our PBM-derived models to predict data generated using a different experimental platform that provides quantitative measurements of TF–DNA binding affinity. In particular, for one of the TFs in our study (the human TF Max), we generated independent in vitro data using systematic evolution of ligands by exponential enrichment combined with massively parallel sequencing (SELEX-seq) (30). We then used our PBM-derived binding specificity models to predict the relative binding affinities measured by SELEX-seq. Finally, using feature weights for the best-performing models, we identified structural mechanisms that are used on a TF family-specific basis for achieving DNA binding specificity.

Results and Discussion

DNA Shape-Augmented Models Outperform Sequence-Based Models. TF binding sites were described in our quantitative specificity models based on feature vectors containing the distinct set of features of a specific model at each nucleotide position. Fig. 1*A* shows a shape-augmented model that combines 1-mer sequence features with the four DNA shape features, MGW, ProT, Roll, and HelT. Features of the *k*-mers and of DNA shape are substantially different in nature. Specifically, the *k*-mer features are binary categorical attributes that characterize hydrogen bonds and other direct contacts between the protein and the base pairs in the major groove (6). In contrast, the DNA shape features are continuous attributes that reflect properties of the DNA structure and capture interactions in the minor groove (10). Given these differences, these two feature types may potentially describe different mechanisms by which a TF achieves its DNA binding specificity.

When tested on the uPBM data from the DREAM5 dataset (17), our shape-augmented (1mer+shape) model outperformed the sequence-only (1mer or PWM) model on 56 of the 65 TFs that were tested (Fig. 1*B*). The only exceptions to this finding were TFs with low-quality data ($R^2 < 0.25$) and, to a lesser extent, three zinc finger TFs. To include high-quality TF datasets, we generated gcPBM data for the human bHLH TFs Mad1 (also known as Mxd1), Max, and c-Myc (Fig. S1), based on a previously described experimental protocol (21, 22). For these gcPBM data, we observed even larger improvements in R^2 when DNA shape features were incorporated into the binding specificity models (Fig. 1*B*).

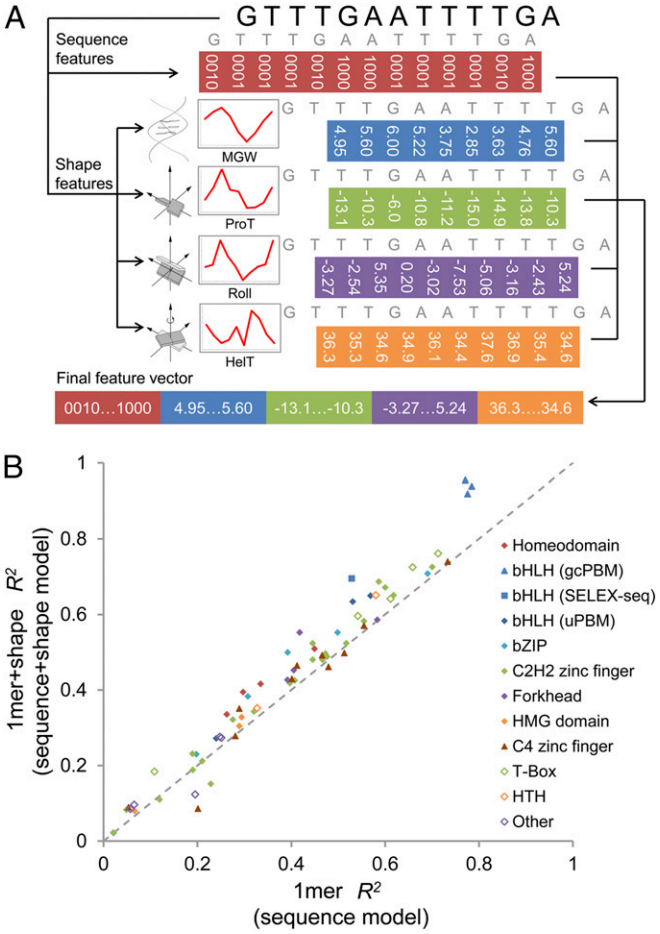

Fig. 1. Design of the sequence+shape feature vector, and TF family-specific performance comparison of binding specificity predictions. (*A*) The feature vector used in the 1mer+shape model combined binary features for the sequence (1-mers) with continuous values for the DNA shape features (MGW, ProT, Roll, and HelT). In addition, second-order shape features were also used throughout this study (see *SI Methods* for details). (*B*) Performance comparison for different TF families tested in this study. DNA shape contributed to the DNA binding specificities of all homeodomain and bHLH TFs in the uPBM, gcPBM, and SELEX-seq datasets, consistent with previous work on these TF families (9, 22, 30, 31, 34).

We observed a similar improvement when we used SELEX-seq data that we generated for the human TF Max.

Considering the uPBM, gcPBM, and SELEX-seq data together, we found that the 1mer+shape model led to consistent improvements for all homeodomain and bHLH TFs (Fig. 1*B*). This observation is in accordance with our previous finding that DNA shape readout plays an important role for these TF families (22, 30, 31). The results also indicate that predictions of binding specificities for other TF families (e.g., bZIP and Forkhead TFs) can benefit from adding DNA shape information to the models. The results indicate that only a subset of zinc finger proteins recognize DNA shape, which is likely due to their modular DNA binding with one finger contacting three base pairs (bp) in a largely independent manner from the binding of adjacent fingers (6). However, because the DREAM5 data were of lower quality (in terms of R^2) for some members of these families, conclusions on readout mechanisms used by these TF families will require additional data from HT binding assays.

The 1mer+shape model implemented in this study used additional second-order shape features to account for dependencies between structural features at adjacent nucleotide positions, such

as the formation of A-tracts, defined as stretches of at least four As and Ts without TpA bp steps (10), where MGW narrowing is more pronounced due to several adjacent A/T bp (6). These second-order shape features were the product terms for the same DNA shape feature category at two adjacent nucleotide positions (MGW and ProT) or base pair step positions (Roll and HelT). The combined use of the second-order DNA shape features with first-order shape features and 1-mer features improved the prediction accuracy compared with the 1mer+first order shape model (Fig. S2*A*).

Models Using 2-mers and 3-mers Contain Implicit DNA Shape Information. We also tested dinucleotide- and trinucleotide-based models using 1-mer, 2-mer, and 3-mer features as predictors. When tested on the DREAM5 uPBM data, the performances of the 1mer+2mer and 1mer+2mer+3mer models were, on average, very close to the performance of the 1mer+shape model (Fig. 2 *A* and *B*). This observation was not surprising because 2-mers and 3-mers partially capture the effect of the DNA shape variation on binding. Specifically, 2-mers describe stacking interactions between adjacent bp, and 3-mers represent short structural elements, such as A-tracts, which form distinct structures (10). Furthermore, for certain TFs, models that used DNA shape features alone, without any explicit information about base identity, were more accurate than models based on sequence alone (Fig. 2*C*).

The DNA shape features used in our study were generated based on pentamer query tables derived from thousands of all-atom Monte Carlo simulations, and they were validated with X-ray and NMR structures (25). To rule out the possibility that the use of pentamer-based independent features in itself could explain the enhanced performance of our shape-based models, we independently randomized the association between pentamer identities and values for each of the four shape features, which intentionally breaks the relationships between pentamers. The R^2 values obtained with the randomized shape tables were significantly lower compared with those for our DNA shape predictions (Fig. 2*D*), demonstrating that the predictive power of DNA shape was not simply due to the inclusion of pentamer-based features.

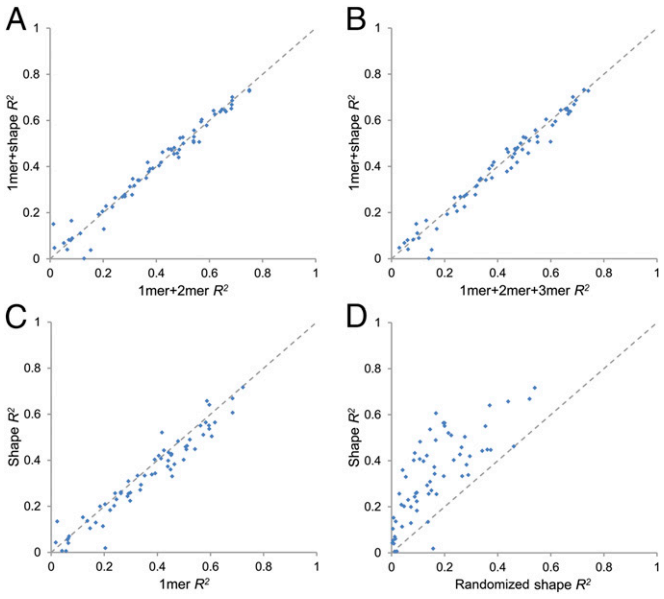

Fig. 2. Performance of various models on uPBM data for 65 mouse TFs. (*A* and *B*) Using R^2 as a measure for prediction accuracy, the performance of the shape-augmented model (1mer+shape) was compared with the performances of sequence-based (*A*) 1mer+2mer and (*B*) 1mer+2mer+3mer models. (*C*) Performance comparison of the shape-only model to the sequence-only (1mer) model. (*D*) Performance comparison of the shape-only model to a model augmented by randomized shape features.

In the aforementioned analyses of 65 mouse TFs using uPBM data from the DREAM5 study (17), we used 10-fold cross-validation to evaluate model performance, based on a single uPBM dataset for each TF. In an alternative approach, we evaluated our DNA shape-augmented models by training the models on uPBM data from one array design and using them to predict binding data obtained from a different array design, similar to the procedure used by ref. 17. The results of the cross-array testing were fully consistent with the results of the 10-fold cross-validation tests (Fig. S2 *B*–*E*).

DNA Shape-Augmented Models Can Accurately Predict TF Binding Data Across Experimental Platforms. To ensure that our regression models of TF binding specificity are capturing real signals and not experimental biases intrinsic to the PBM technology, we performed a cross-platform analysis. Specifically, we trained our models on gcPBM data and tested them on quantitative SELEX-seq data generated for the human TF Max (see *Methods* and *SI Methods*). Briefly, the Max SELEX-seq experiment was carried out as described previously (30), with two rounds of selection, and the data were used to compute relative binding affinities for 12-mer DNA sequences. Next, sequence- and shape-based regression models trained on our Max gcPBM data were used to predict the binding levels of Max to DNA targets identified by SELEX-seq (Fig. 3*A*). Given that the two experimental platforms (gcPBM and SELEX-seq) provided different measurements for TF binding (i.e., binding intensity and relative binding affinity, respectively), we used Spearman's rank correlation coefficients to assess the accuracy of our predictions. When tested on SELEX-seq data, the 1mer+shape model outperformed all of the sequence-based models, including the 1mer+2mer+3mer model (Fig. 3 *B*–*D* and Fig. S3 *A*–*C*). The improvement is modest, but nonetheless statistically significant (Fig. 3*B*).

Replacing *k*-mer with DNA Shape Features Reduces the Dimensionality of the Feature Space. The assessment of different models required not only a comparison of prediction accuracy but also of model complexity, which relates to the required size of the training data and the computational cost (Fig. 4*A*). Compared with the sequence-based 1mer+2mer and 1mer+2mer+3mer models, the 1mer+shape model contained fewer features, which translated into fewer model parameters. For each nucleotide position, 4 features were introduced to encode 1-mer identity, 16 features to encode 2-mer identity, 64 features to encode 3-mer identity, and 8 features to encode DNA shape. Therefore, the 1mer+shape, 1mer+2mer, and 1mer+2mer+3mer models used a total of 12, 20, and 84 features, respectively, per nucleotide position (Fig. 4*A*). The finite sequence length required a slight end-effect adjustment of these numbers of features per nucleotide position.

To assess how models of different complexity performed on smaller datasets, we used the gcPBM data that we generated for the human bHLH TFs to assemble datasets of decreasing sample size. As expected, the performance of all models decreased with decreasing sample size (Fig. 4*B* and Fig. S4). This decreasing trend was consistently more pronounced for sequence-based models (1mer+2mer and 1mer+2mer+3mer) than for shape-based models (1mer+shape and shape-only). This finding suggests that shape features more efficiently captured the binding specificities of the studied TFs than *k*-mer features.

Analysis of the gcPBM data for human bHLH TFs demonstrated that the 1mer+shape model was superior to the 1mer+2mer and 1mer+2mer+3mer models (Fig. 4*B* and Fig. S4). This difference was much more pronounced for the higher-quality gcPBM than for the lower-quality uPBM data (Fig. 2 *A* and *B*), likely because the gcPBM data contained less positional bias and provided information on the genomic flanking regions (22).

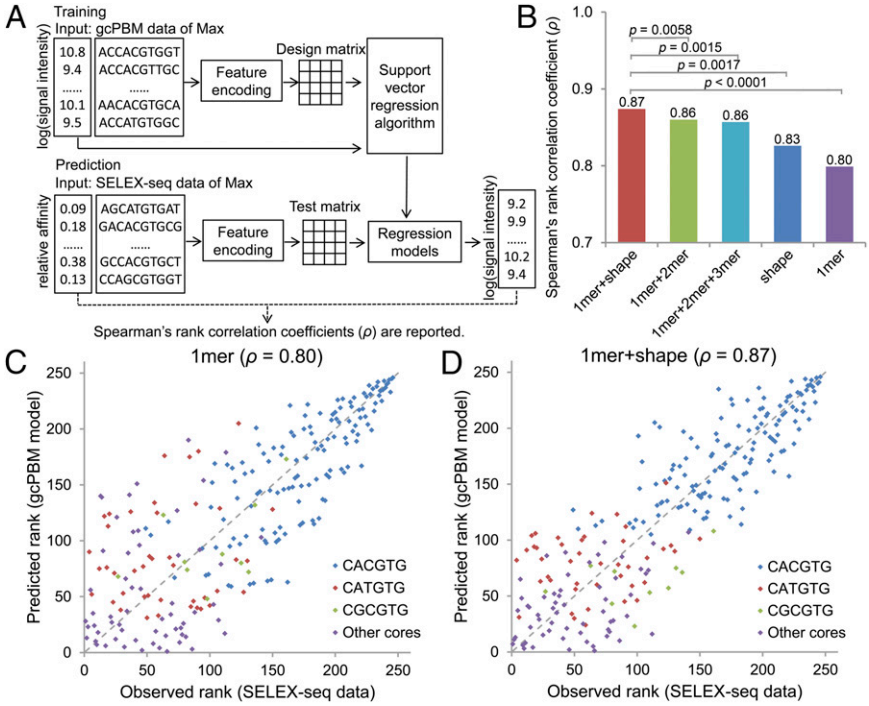

Fig. 3. Performance of binding specificity models across experimental platforms. (*A*) Flowchart illustrating that Max–DNA binding specificity models were trained on natural logarithms of fluorescence binding intensities measured by a gcPBM experiment and used to predict the binding level of DNA targets derived from a SELEX-seq experiment for the same TF. (*B*) Performance of various models for cross-platform predictions based on Spearman's rank correlation coefficients between observed SELEX-seq relative binding affinities and predicted gcPBM signal intensities. The *P* values indicate that the improvement in prediction accuracy using the 1mer+shape model is significant compared with the sequence-based models. (*C* and *D*) Scatter plots of predicted versus observed binding site ranks, showing the performance of the (*C*) 1mer and (*D*) 1mer+shape models trained on gcPBM data and tested on SELEX-seq data. Here, higher ranks represent higher-affinity binding sites.

The positive impact of a smaller number of features on the model prediction accuracy for small sample sizes was also demonstrated by using sequence models augmented with only one category of DNA shape features, instead of all four. The inclusion of only one DNA shape feature further reduced the number of features (Fig. 5*A*). When tested on gcPBM data for Max with smaller sample sizes (10 randomly generated samples for each size), the prediction accuracies of the 1mer+Roll and 1mer+ProT models dropped at a slower rate compared with the 1mer+shape and 1mer+2mer+3mer models, which required many more features (Fig. 5*B*).

DNA Shape-Augmented Models Reveal TF Family-Specific Readout Mechanisms. We trained regression models using mononucleotide identities (i.e., 1-mer features) and individual DNA shape features (i.e., MGW, ProT, Roll, or HelT), to determine which shape features are most important for predicting the binding preferences of TFs from different structural families. For the bHLH TF Max, data for models using sequence and a single DNA shape category suggested that Roll and ProT were the dominant structural determinants of the DNA binding specificity. MGW and HelT were of lesser importance for the DNA readout of this specific TF (Fig. 5*B*). The inclusion of 1-mer features ensured that the nucleotide identities necessary for base readout did not indirectly influence the impact of the DNA shape features.

Analysis of the feature weights for 1mer+first order shape models indicated that our approach has the potential to reveal readout mechanisms on a TF family-specific basis. Second-order shape features were not included in the models, to simplify the interpretation of the shape feature weights. Feature weights for Roll in the 1mer+first order shape model varied between large positive and large negative weights within the core enhancer box

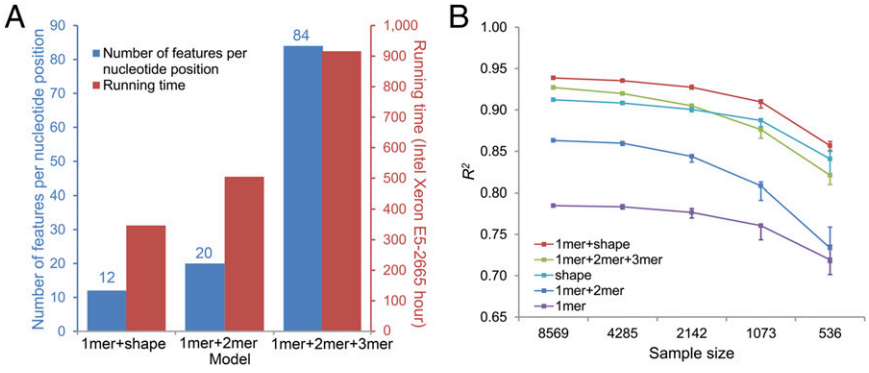

Fig. 4. Comparison of various models using gcPBM data for human Max TF. (*A*) The number of required features (per nucleotide position) was correlated with the average running time for the training and testing of different models for Max–DNA binding, based on gcPBM data. (*B*) Performances of the sequence- and shape-based models for Max–DNA binding as the sample size was decreased.

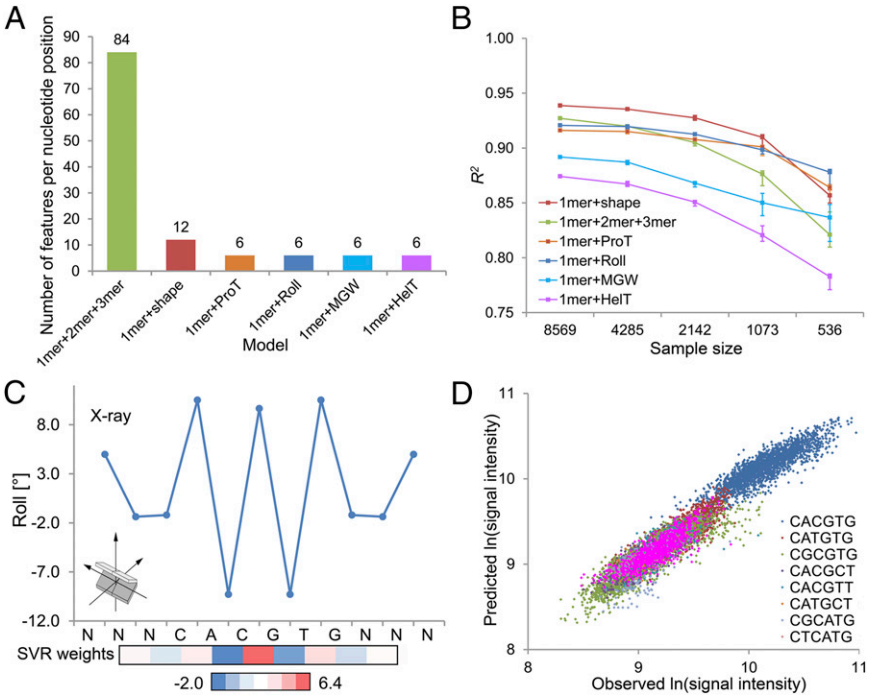

Fig. 5. Insights into TF-specific readout mechanisms derived from shape-augmented binding specificity models. (*A*) Number of features per nucleotide position introduced in different models. Models that included only one shape feature further reduced the total number of features compared with the 1mer+2mer+3mer and 1mer+shape models. (*B*) The single-shape-feature models 1mer+ProT and 1mer+Roll performed better than the 1mer+shape and 1mer+2mer+3mer models on smaller datasets. (*C*) Feature weights for Roll (heat map) derived from the 1mer+first order shape model using SVR accurately reflected the Roll characteristics (plot) observed in the cocrystal structure of the ternary Max homodimer/DNA complex [Protein Data Bank identifier (PDB ID) 1HLO] (32). (*D*) The CACGTG E-box in the cocrystal structure was the highest-affinity core among the nine observed E-box cores. Although other cores were present in the gcPBM data for Max, the SVR feature weights correctly reflected the Roll features of the CACGTG core observed in the cocrystal structure (32).

(E-box) binding site (Fig. 5*C*). These feature weights accurately reflected the pattern between large positive and large negative Roll angles observed in a cocrystal structure of the Max–Max/DNA complex (32). The DNA target of this ternary complex contained the CACGTG E-box, which is the highest-affinity core present in the gcPBM dataset (Fig. 5*D*). These observations indicate that despite the presence of multiple E-box cores, there exists a specific Roll pattern that is preferred by the Max homodimer. Moreover, the findings show that a degeneracy of DNA sequence and shape, similar to the degeneracy of protein sequence and structure (33), constitutes an important characteristic of TF binding sites (31).

Among the remaining DNA shape features, compared with the weights for MGW and HelT, the weights for ProT derived from the same Max 1mer+first order shape model agreed better with the cocrystal-derived structural parameters (Fig. S5). This observation was consistent with the prediction accuracies of the different models of Max binding specificity (Fig. 5*B*). We also analyzed the feature weights for MGW derived from uPBM data for homeodomain TFs. We found that the 1mer+first order shape model accurately reflected the known MGW preferences of homeodomain TFs (Fig. S5*D*). The negative feature weights in the second half of the TAAT core-binding site indicated a preference for a narrow MGW in this region, which is consistent with previous X-ray (9), SELEX-seq (30), and uPBM (34) data.

Our analysis across diverse TF families shows that a limited set of local DNA shape features improves binding specificity predictions. This approach introduces a coarse representation of the double helix. Although the conformation, flexibility, and energetics of DNA are properties that are known to affect transcription factor binding (35), our simplified models use DNA shape features that have been reported to be important for protein–DNA binding (3) and that are available on a genome-wide basis (36). Unlike atomistic modeling of protein–DNA binding, which is a fine-grained description necessary for explicit energy calculations (37), our coarse representation encodes the energy landscape of protein–DNA binding implicitly. For example, large positive Roll values of TpA bp steps encode weak stacking interactions and thus high conformational flexibility (6). Negative ProT values in A-tracts encode the possibility of interbase-pair hydrogen bonds in the major groove resulting in rigid elements (6). In our previous work on the DNA binding of the *Escherichia coli* Fis protein, we found that DNA shape features used in our models were predictive of binding affinities, as the MGW is indicative of how much the DNA needs to be bent in order for the binding to occur (25, 38). Here, we further demonstrate this point by showing that for the human bHLH Max TF DNA shape profiles of high-affinity binding sites based on mechanically induced trapping of molecular interactions (MITOMI) binding assays (39) substantially differ from shape profiles of low-affinity, nonspecific DNA binding sites (Fig. S6).

Conclusion

Quantitative Models Combining DNA Sequence and Shape Contribute to the Understanding of TF–DNA Binding. In this study, we have shown that statistical machine learning models of TF–DNA binding specificity consistently benefit from augmenting sequence-based models with features encoding interactions between nucleotide positions across a diverse panel of TFs. The improvement in model performance can be achieved either with *k*-mer or with DNA shape features. Although adding any interaction term (i.e., 2-mer, 3-mer, or shape features) improved the modeling of DNA binding specificities, more-efficient models were obtained by using DNA shape, which represents interactions with a smaller number of features.

DNA shape integrates the complex interdependencies between multiple positions of a binding site. This integration is achieved implicitly, without any explicit knowledge of individual interdependencies. In this way, the incorporation of DNA shape reduces the number of required parameters while providing a compelling mechanistic explanation for why dinucleotides and trinucleotides can increase the accuracy of motif descriptions. Despite the lower accuracy for some of the DREAM5 data, our results show that quantitative models derived from SVR analyses using HT sequence data can reveal specific mechanisms of protein–DNA recognition on a TF family basis and can contribute to the understanding of TF binding to the genome. Compared to sequence-based methods, the combination of

DNA sequence and shape provides a fundamentally different approach for understanding TF binding specificities, which can be broadly applied to TFs from different structural classes.

Methods

Encoding of DNA Sequence or *k*-mer Features. For each nucleotide position in a given DNA sequence of length L, the k-mer feature at that position was encoded as a binary vector of length 4^k, where a value of 1 represents the occurrence of a particular k-mer starting at that position. The k-mer features for a given DNA sequence of length L were encoded by concatenating the k-mer feature vectors at each nucleotide position, resulting in a binary vector of length $4^k(L-k+1)$. See *SI Methods* for the resulting 1-mer, 2-mer, and 3-mer feature vectors.

Encoding of DNA Shape Features. For a given DNA sequence of length L, four DNA shape features were predicted by an HT method trained on Monte Carlo simulations of 2,121 different DNA fragments of 12–27 bp in length (25). These structural features included two nucleotide parameters (MGW and ProT) and two bp step parameters (Roll and HelT). The total length of the final DNA shape feature vector was $(8L-32)$, due to the use of four first-order and four second-order shape features, and the unavailability of values at two positions at each end (31). See *SI Methods* for the resulting shape feature vectors and the source code used to compute DNA sequence features and shape features for putative TF binding sites. The generated feature vectors can be used directly to train and test SVR models with the publicly available library of support vector machines (LIBSVM) toolkit (40).

Performance Evaluation of Sequence- and Shape-Based Models. After preprocessing and feature encoding, PBM data for each TF were transformed into a matrix. The first column of this matrix contained the natural logarithm of the fluorescence signal intensities of the PBM probes, and the remaining columns contained the encoded features. For each DNA shape characteristic, first- and second-order DNA shape features were normalized to values between 0 and 1. The ε-SVR algorithm (41) with linear kernel, implemented in the LIBSVM toolkit (40), was used to train regression models for predicting the natural logarithm of the PBM signal intensities (response variable) based on the encoded features (see *SI Methods* for details).

To obtain unbiased performance estimates of the regression models on each dataset, a nested 10-fold cross-validation procedure was implemented. First, each dataset was randomly partitioned into 10 equally sized subsets. Each subset was used for testing while the other nine subsets were used for training. Thus, our models were always tested on data not included in the training process. For each TF dataset, the squared Pearson correlation coefficient R^2 between the predicted and observed values of the response variables for all DNA sequences in that dataset was reported.

PBM and SELEX-seq Binding Assays. The gcPBM experiments for the human TF dimers Mad1 (Mxd1)–Max, Max–Max, and c-Myc–Max (Mad, Max, and Myc, respectively) were performed essentially as described previously (21). The gcPBM data were preprocessed to filter out sequences containing more than one putative TF binding site (see *SI Methods*). After the filtering step, 6,927 probes for Mad, 8,569 probes for Max, and 7,535 probes for Myc were obtained. See also *SI Methods* for a description of the preprocessing of the uPBM data for the 66 mouse TFs (17), which resulted in 65 analyzable datasets.

A SELEX-seq experiment for the human TF dimer Max–Max was carried out as described previously (30). After two rounds of SELEX, the relative affinities of all 12-mers were calculated, and the data were preprocessed similarly to the uPBM data (see *SI Methods*).

ACKNOWLEDGMENTS. The authors thank Rosa Di Felice for helpful comments on the project and Matt Weirauch for providing uPBM data. This work was supported by the National Institutes of Health [Grants R01GM106056 and U01GM103804 (to R.R.), R01HG003008 (to H.J.B. and R.R.), U54CA121852 (to H.J.B.), R01GM058575 (to R.S.M.), and F32GM099160 (to N.A.)], the National Science Foundation [Grant MCB-1412045 (to R.G.)], and the Pharmaceutical Research and Manufacturers of America Foundation (R.G.). R.R. and R.G. are Alfred P. Sloan Research Fellows. The open access publication charges were defrayed through the National Science Foundation [Grant MCB-1413539 (to R.R.)].

1. Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: From properties to genome-wide predictions. *Nat Rev Genet* 15(4):272–286.
2. Levo M, Segal E (2014) In pursuit of design principles of regulatory sequences. *Nat Rev Genet* 15(7):453–468.
3. Slattery M, et al. (2014) Absence of a simple code: How transcription factors read the genome. *Trends Biochem Sci* 39(9):381–399.
4. Stormo GD, Zhao Y (2010) Determining the specificity of protein-DNA interactions. *Nat Rev Genet* 11(11):751–760.
5. Stormo GD (2013) Modeling the specificity of protein-DNA interactions. *Quantitative Biology* (Springer, Berlin), Vol 1, pp 115–130.
6. Rohs R, et al. (2010) Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* 79:233–269.
7. Kim S, et al. (2013) Probing allostery through DNA. *Science* 339(6121):816–819.
8. Watson LC, et al. (2013) The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nat Struct Mol Biol* 20(7):876–883.
9. Joshi R, et al. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* 131(3):530–543.
10. Rohs R, et al. (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461(7268):1248–1253.
11. White MA, Myers CA, Corbo JC, Cohen BA (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the *cis*-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci USA* 110(29):11952–11957.
12. Berger MF, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24(11):1429–1435.
13. Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22(14):e141–e149.
14. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* 29(6):480–483.
15. Orenstein Y, Shamir R (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res* 42(8):e63.
16. Badis G, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324(5935):1720–1723.
17. Weirauch MT, et al.; DREAM5 Consortium (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* 31(2):126–134.
18. Bussemaker HJ, Foat BC, Ward LD (2007) Predictive modeling of genome-wide mRNA expression: From modules to molecules. *Annu Rev Biophys Biomol Struct* 36:329–347.
19. Stormo GD (2000) DNA binding sites: Representation and discovery. *Bioinformatics* 16(1):16–23.
20. Zhao Y, Ruan S, Pandey M, Stormo GD (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* 191(3):781–790.
21. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordân R (2013) Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics* 29(13):i117–i125.
22. Gordân R, et al. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Reports* 3(4):1093–1104.
23. Sharon E, Lubliner S, Segal E (2008) A feature-based approach to modeling protein-DNA interactions. *PLOS Comput Biol* 4(8):e1000154.
24. Agius P, Arvey A, Chang W, Noble WS, Leslie C (2010) High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLOS Comput Biol* 6(9).
25. Zhou T, et al. (2013) DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res* 41(web server issue):W56–W62.
26. Lazarovici A, et al. (2013) Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci USA* 110(16):6376–6381.
27. Chen Y, et al. (2013) Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Res* 41(17):8368–8376.
28. Chang YP, et al. (2013) Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. *Cell Reports* 3(4):1117–1127.
29. Vapnik VN (1995) *The Nature of Statistical Learning Theory* (Springer, New York).
30. Slattery M, et al. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* 147(6):1270–1282.
31. Yang L, et al. (2014) TFBSshape: A motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res* 42(database issue):D148–D155.
32. Brownlie P, et al. (1997) The crystal structure of an intact human Max-DNA complex: New insights into mechanisms of transcriptional control. *Structure* 5(4):509–520.
33. Kosloff M, Kolodny R (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins* 71(2):891–902.
34. Dror I, Zhou T, Mandel-Gutfreund Y, Rohs R (2014) Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Res* 42(1):430–441.
35. von Hippel PH (2007) From "simple" DNA-protein interactions to the macromolecular machines of gene expression. *Annu Rev Biophys Biomol Struct* 36:79–105.
36. Chiu TP, et al. (2015) GBshape: A genome browser database for DNA shape annotations. *Nucleic Acids Res* 43(database issue):D103–D109.
37. Liu LA, Bradley P (2012) Atomistic modeling of protein-DNA interaction specificity: Progress and applications. *Curr Opin Struct Biol* 22(4):397–405.
38. Hancock SP, et al. (2013) Control of DNA minor groove width and Fis protein binding by the purine 2-amino group. *Nucleic Acids Res* 41(13):6750–6760.
39. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science* 315(5809):233–237.
40. Chang CC, Lin CJ (2011) LIBSVM: A Library for Support Vector Machines. *ACM Trans Intell Syst Technol* 2(3):27.
41. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. *Neural Information Processing Systems 9* (MIT Press, Cambridge, MA), pp 155–161.