Published online 20 November 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 22* ***12877–12887***
doi: 10.1093/nar/gkx1145

Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding

Jinsen Li[1], Jared M. Sagendorf[1], Tsu-Pei Chiu[1], Marco Pasi[2], Alberto Perez[3] and Remo Rohs[1,*]

[1]Computational Biology and Bioinformatics Program, Departments of Biological Sciences, Chemistry, Physics & Astronomy, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA, [2]Centre for Biomolecular Sciences and School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, UK and [3]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY 11794, USA

Received March 21, 2017; Revised October 2, 2017; Editorial Decision October 27, 2017; Accepted October 30, 2017

ABSTRACT

Uncovering the mechanisms that affect the binding specificity of transcription factors (TFs) is critical for understanding the principles of gene regulation. Although sequence-based models have been used successfully to predict TF binding specificities, we found that including DNA shape information in these models improved their accuracy and interpretability. Previously, we developed a method for modeling DNA binding specificities based on DNA shape features extracted from Monte Carlo (MC) simulations. Prediction accuracies of our models, however, have not yet been compared to accuracies of models incorporating DNA shape information extracted from X-ray crystallography (XRC) data or Molecular Dynamics (MD) simulations. Here, we integrated DNA shape information extracted from MC or MD simulations and XRC data into predictive models of TF binding and compared their performance. Models that incorporated structural information consistently showed improved performance over sequence-based models regardless of data source. Furthermore, we derived and validated nine additional DNA shape features beyond our original set of four features. The expanded repertoire of 13 distinct DNA shape features, including six intra-base pair and six inter-base pair parameters and minor groove width, is available in our R/Bioconductor package DNAshapeR and enables a comprehensive structural description of the double helix on a genome-wide scale.

INTRODUCTION

The binding of transcription factors (TFs) to DNA is a fundamental and crucial step for gene regulation. However, many mechanisms involving TF binding are still unknown (1,2). Basic questions remain as to how a TF selectively recognizes and binds to specific DNA sequences. Experimental protocols have been established to measure TF–DNA binding specificities quantitatively and to understand the underlying mechanisms. For example, protein-binding microarrays (PBMs) (3) and systematic evolution of ligands by exponential enrichment combined with massively parallel sequencing (SELEX-seq) (4) or high-throughput SELEX (HT-SELEX) (5) are widely used methods to probe *in vitro* TF–DNA binding quantitatively (6).

To analyze and interpret the large amounts of data that are obtained from high-throughput (HT) binding assays, researchers have proposed various models for TF–DNA binding specificities, the most widely used of which is the position weight matrix (PWM) (7). PWMs use position frequency matrices, created by counting the probability that a nucleotide will occur at each individual position of the binding site, while ignoring dependencies between nucleotide positions (7). This approach has been successfully used to approximate TF–DNA binding events and has been expanded to include interdependencies between adjacent nucleotides in predicting TF–DNA binding (8–10). Despite its success, researchers have pointed out limitations of the PWM model (11), leading to the development of alternative approaches (12).

One alternative representation of interdependencies between nucleotide positions considers that TFs not only recognize DNA sequences base-by-base through hydrogen bonds or other amino acid contacts but also favor certain DNA conformations (6). Many studies have verified the recognition of three-dimensional DNA structure for diverse TF families (13–26). DNA conformation can be represented by DNA shape features (27), which assign numerical values to translations and rotations between and within base pairs (bp) (28) or measure groove width between opposite phosphodiester backbones (29). These parameters can be calculated by different software tools, such

*To whom correspondence should be addressed. Tel: +1 213 740 0552; Fax: +1 213 821 4257; Email: rohs@usc.edu

as CURVES or 3DNA (30,31), and incorporated into quantitative models (13). Information on DNA shape can be extracted from either experimental studies or molecular simulations (32). For experimentally solved structures, X-ray crystallography (XRC) data provide much larger sequence coverage than nuclear magnetic resonance (NMR) spectroscopy data. Molecular simulations, such as Monte Carlo (MC) or Molecular Dynamics (MD) simulations (33–35), likewise provide a wide sequence coverage. However, a HT method is needed to obtain shape information for large numbers of sequences of arbitrary length or entire genomes.

In previous work, we used a sliding-window approach together with a query table for all unique pentamers generated from MC simulations to produce a predictive HT method of DNA shape (27). Using this method, we predicted four DNA shape features: minor groove width (MGW), propeller twist (ProT), helix twist (HelT) and Roll. Experimental structures have indicated that these four DNA shape features are among the most important structural characteristics for evaluating protein–DNA readout modes (10,29,36–39). However, given the intricacies of DNA structure, we hypothesized that these four DNA shape features might not suffice in describing the entire set of DNA recognition mechanisms. Therefore, we derived and validated nine additional DNA shape parameters, expanding our available repertoire to a total set of 13 features: six inter-bp parameters, six intra-bp parameters and MGW (Figure 1). We evaluated whether including 13 shape features in our quantitative models would enhance the accuracy of TF binding predictions beyond prior models based on just four shape features (13).

To evaluate the contributions of these additional shape features, we needed to compare model performances using MC-derived shape features to models based on features derived from other methods. With the recent publication of 1-μs MD simulations of all unique tetramers (40), we were able to extract the same DNA shape information from MD simulations and to compare models using these MD data to models including equivalent data extracted from MC simulations. Considering that MC and MD simulations are both computational prediction methods, we added an experimental reference and extracted DNA shape features from XRC data available in the Protein Data Bank (PDB) (41). Although the Nucleic Acid Database (42) contains the same structures for nucleic acids, we reported the PDB identifiers for consistency with our previous work (27). To summarize, DNA shape features used in this study were derived from MC or MD simulations and XRC experiments.

Because we derived DNA shape features from various sources, we showed the value of this additional information in studying several biological questions, which are routinely addressed based on DNA sequence methods (43). Previously, we found that using shape information from MC simulations improved prediction accuracy and reduced computational complexity compared to *k*-mer models (13,19,44). We sought to verify the robustness of models using additional shape features and to compare the performances of models based on MC-derived shape features to models using MD and XRC-derived shape features. Therefore, we evaluated and compared performances of regression models for TF binding specificity predictions incorporating 13 DNA shape feature categories in analogy to previous models using just four shape categories (13,16).

MATERIALS AND METHODS

Dataset selection and preprocessing

We used three different datasets, which were designed for detecting TF–DNA binding specificities and were derived from different experimental platforms: genomic-context PBM (gcPBM) (10), HT-SELEX (5) and SELEX-seq (4). The gcPBM data used here contained data for the human basic helix-loop-helix (bHLH) TF dimers Mad1/Max ('Mad'), Max/Max ('Max') and c-Myc/Max ('Myc') (available in the Gene Expression Omnibus [GEO] under accession number GSE59845) (13). The SELEX-seq data contained 21 different datasets for *Drosophila* Hox TFs in complex with their cofactor Extradenticle (Exd) (available under GEO accession number GSE65073) (4). The HT-SELEX data included 240 datasets for 215 TFs from 27 protein families (available in the European Nucleotide Archive [ENA] under study identifier PRJEB14744) (19).

We preprocessed raw data from gcPBM and SELEX-seq experiments using methods from (13), and HT-SELEX data using methods from (19), which essentially involved PWM-based sequence alignment and trimming. After preprocessing, we collected more than 10 000 sequences per TF for gcPBM data, and thousands of sequences per TF for SELEX-seq and HT-SELEX data. Each sequence was assigned a relative binding affinity score, which was measured in the respective experiment. Log-affinity scores were used in this study.

DNA shape prediction and query table generation

We sought to use DNA shape information derived from three different data sources (MC and MD simulations and XRC experiments) in regression algorithms for predicting TF–DNA binding specificities. Different HT predictions of DNA shape features were obtained by generating query tables from different data sources. Our previous work used a pentamer query table generated from MC data (27); however, the available MD data did not cover all 512 unique pentamers. To ensure complete sequence coverage and to enable comparisons between data sources, we generated three tetramer query tables from (i) the 1-μs MD data (40), (ii) an XRC dataset used for validation in our previous work (27) (Table S2) and (iii) MC data (27). The new MC tetramer query table was generated based on the same MC simulation data as used in our previously generated pentamer query table (see Supplementary Materials and Methods and (27) for details on the MC simulation protocol), by mining data for all tetramers instead of pentamers. For MD data, simulations were available for 39 oligomers with sequences designed so that all 136 unique tetramers were covered with an average occurrence of 3.9 (see Supplementary Materials and Methods and (40) for details on the MD simulation protocol). The XRC data provided an average occurrence of each tetramer of 44.6, and the MC data provided an average coverage of 249.8 for each unique tetramer. Although the PDB contains additional DNA structures, we used the original XRC dataset to be consistent with our

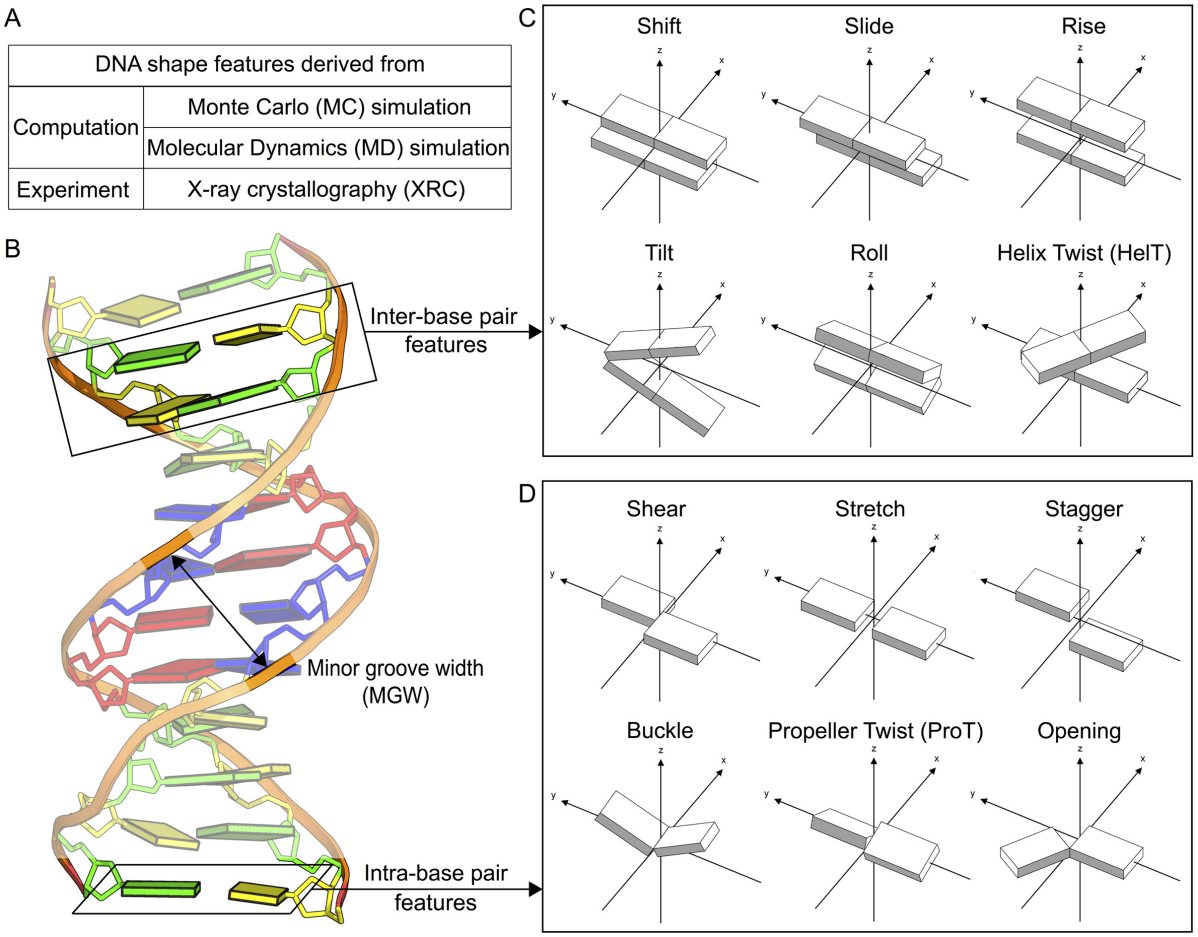

Figure 1. Introduction of an expanded repertoire of 13 DNA shape features. (**A**) We used three methods, including MC and MD simulations and XRC, to derive DNA shape features. (**B**) Schematic representation of a DNA fragment (PDB ID: 1BNA taken from the Protein Data Bank) with definition of MGW, inter-bp and intra-bp parameters. (**C**) Schematic representation of all inter-bp DNA shape features used in this research. Each long brick represents a bp. Translations and rotations of bp are shown in top and bottom row, respectively. (**D**) Schematic representation of all intra-bp DNA shape features used in this research. Each short brick represents a base. Translations and rotations of bases are shown in top and bottom row, respectively.

original DNAshape method (27). XRC data were filtered for unusual deformations, and chemically modified structures were removed as previously described (27). The different coverage of tetramers in the MD, XRC and MC datasets indicates why only our MC-based DNAshape method provided sufficient coverage for all 512 unique pentamers (27).

Each query table provided 13 DNA shape features, including six inter-bp or bp-step parameters (HelT, Rise, Roll, Shift, Slide and Tilt), six intra-bp or bp parameters (Buckle, Opening, ProT, Shear, Stagger and Stretch), and MGW. We implemented a sliding window algorithm (27) that uses a sliding pentamer window to acquire numerical shape values from the MC-derived pentamer query table (available for download at http://rohslab.usc.edu/DNAshape+/) and to combine the values into a feature vector. For example, considering a sequence of length n, when the sliding window begins at position i, we use the query table and find the shape feature values for position i+2 (for bp parameters) and positions i+1 and i+2 (for bp-step parameters) using the first pentamer. Then, we move the sliding window to position i+1 and use the next pentamer until we reach the end of the query sequence. We determine bp-step parameters from overlapping pentamers using the arithmetic average of values derived from two adjacent pentamers. The algorithm works similarly on tetramer query tables based on a tetramer query window, with the exception that a tetramer is assigned two central bp parameters and one central bp-step parameter.

Introduction of additional DNA shape features

We validated the nine additional DNA shape features that were introduced in this study by comparing the HT predictions with equivalent features derived from XRC, following a protocol described for the DNAshape method (27). We calculated Spearman's rank correlation coefficients for the comparison with experimental data (Supplementary Table S1). Whereas the MC and MD simulations were performed for unbound DNA fragments, the XRC dataset included DNA conformations from protein–DNA complexes due to the scarcity of experimental structures for free DNA molecules (32). We removed structures with deformations due to crystal packing effects and other deformations, as previously described (27). We chose to use Spearman's rank

correlation as a criterion because it captures the pattern between minima and maxima in shape features and is less sensitive to fluctuations in actual values of shape parameters, which can be affected by crystal packing artifacts or protein-induced deformations (27).

The general concept of DNA shape includes conformational flexibility. Certain DNA sequences are obviously more flexible than others, which can influence TF binding. To capture this effect, we used the standard deviation (SD) of each DNA shape feature as an approximation of DNA flexibility. Instead of calculating SD values along a simulation trajectory for the single occurrence of a pentamer, we pooled all of the pentamers of the same identity together to generate SD values. This approach has the limitation that not all pentamers occurred with the same frequency in our dataset. Nevertheless, the approach provided a single SD value for each shape feature in a given sequence environment, independent of different occurrences of a pentamer or tetramer in our dataset. For each bp-step parameter, one SD value was derived to prevent averaging between two SD values.

Implementing statistical regression models

Feature vectors for a given sequence were generated in the following manner. Each type of nucleotide was assigned one of four binary variable vectors (mononucleotide or 1mer model): A was encoded as '1 0 0 0', C as '0 1 0 0', G as '0 0 1 0', and T as '0 0 0 1' (Figure 2). Following the same scheme, dinucleotides were encoded by one of 16 binary vectors (dinucleotide or 2mer model): AA was encoded as '1' followed by 15 '0' values and so on. In general, a k-mer model requires $O(4^k)$ binary features to encode the sequence (see Supplementary Materials and Methods).

DNA shape features were predicted with our DNAshape method (27), normalized and concatenated with the encoded sequence feature into a feature vector (see Figure 2 and Supplementary Materials and Methods). For each TF dataset, which contains a list of aligned sequences and their corresponding binding affinity scores, each sequence was used to create a representative feature vector. Based on the resulting feature matrix and corresponding binding affinity scores, we applied a multiple linear regression (MLR) model with L2 regularization to prevent overfitting (45). MLR models with L2 regularization have the following loss function for minimizations using a closed-form solution:

$$\ell = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\intercal}\boldsymbol{\omega}\right)^2 + \lambda \parallel \boldsymbol{\omega} \parallel_2^2$$

where, y_i represents the i-th observed binding affinity score, $\boldsymbol{x}_i$ represents the i-th feature vector, $\boldsymbol{\omega}$ represents feature weights and λ is the L2 regularization parameter. To minimize ℓ, a closed-form solution can be derived as:

$$\hat{\boldsymbol{\omega}} = (\boldsymbol{X}^{\intercal}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\ \boldsymbol{X}^{\intercal}\boldsymbol{y}$$

where, $\boldsymbol{I}$ is an identity matrix that has the size of the number of features in the feature vector. The L2 regularization parameter λ in loss function ℓ penalizes large $\boldsymbol{\omega}$ values to prevent overfitting. To apply this regression model, each dataset was separated into 10-fold training and testing data. Models were trained on 90% of the data and tested on the remaining 10%. This procedure was repeated ten times until each tenth of the data was assigned to the training data. To select the best regularization parameter λ, another 10-fold cross validation was performed on each fold of the training data. After optimizing λ, we applied the model to all training data and calculated feature weights. With the regularization parameter and weights, we applied the model to the test data and retrieved predicted binding affinity scores. To improve robustness of the models, rather than randomly dividing data into 10-fold, we divided data into 10-fold with similar binding affinity score distributions by selecting one of the 10 entries into a fold based on the sorted binding affinity-score list. Each model was trained separately for each individual TF dataset because the derived binding affinities (representing the response variable) could be incomparable for different experiments or each individual TF.

Prediction accuracy and performance assessment

Accuracies of the predicted binding affinity scores were determined by using the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where, y_i, $\hat{y}_i$ and $\bar{y}$ represent the observed, predicted and averaged observed binding affinity scores, respectively. When comparing performances between different sources of DNA shape data, we directly compared R^2 results to see if there was a significant difference in model performance. In nearly every dataset, the sample size was much larger than the length of our feature vector; therefore, an adjusted R^2 was not introduced. R^2 is an indicator of performance for each TF dataset. We calculated the weighted-average R^2, which considers the number of available sequences to derive an overall performance for certain datasets or TF families. We assumed that datasets with greater numbers of sequences could be used to predict binding affinity scores more accurately than datasets with fewer sequences, assuming that the binding affinity scores were measured with similar systematic errors.

In addition to R^2, we used mean squared error (MSE) to achieve rigorous model assessment. In regression models, MSE is computed based on the number of statistical degrees of freedom (sample size minus the number of features) and, therefore, will penalize long feature vectors (see Supplementary Materials and Methods for equation). For example, if model B has a higher R^2 but does not have a lower MSE than model A, then one cannot claim that model B performs better than model A.

RESULTS

Performance comparisons of TF binding models derived using structural data from different sources

MGW, ProT, HelT and Roll were used in our previous studies (13,27) and served as a reference for comparison of the three new query tables, now tetramer-based and derived from different data sources. R^2 values

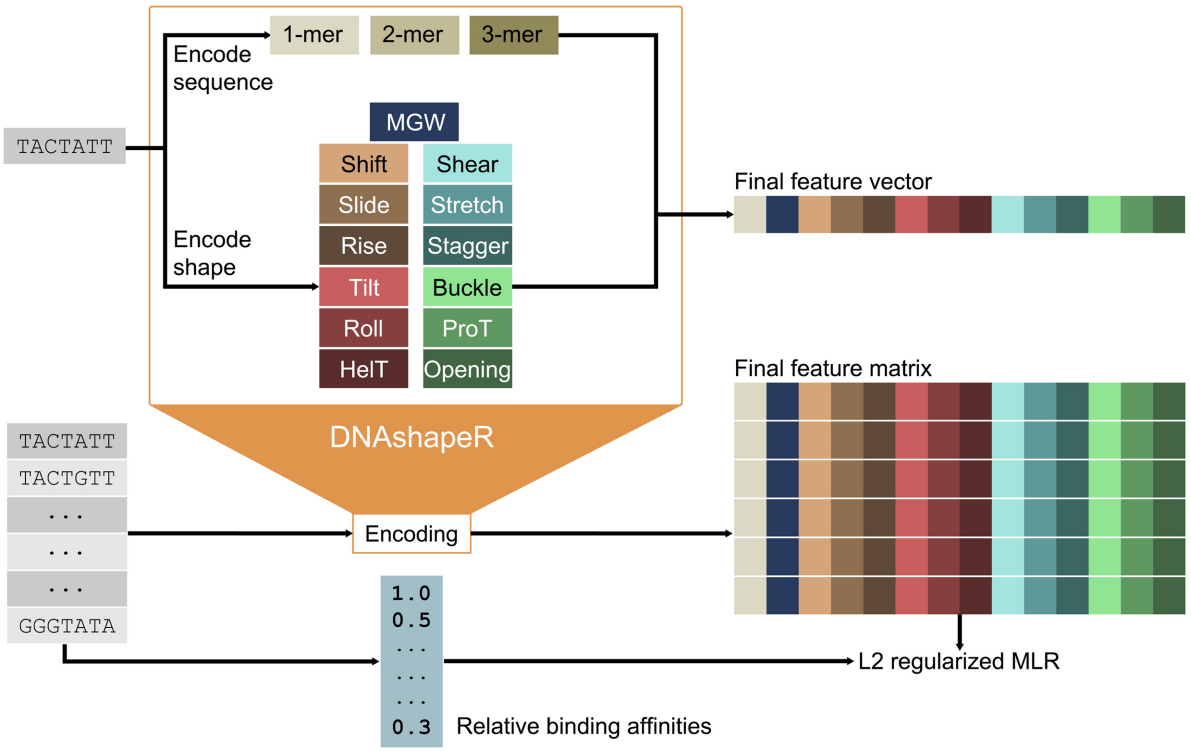

Figure 2. Work flow to encode DNA sequence and shape, which were used to train L2-regularized MLR models that predicted TF binding specificities. DNA sequences were encoded into sequence and shape features. Any combination of sequence and 13 shape features could be chosen. DNA sequence and corresponding relative binding affinities were acquired from experimental data. Encoded DNA sequences (final feature matrix) and corresponding binding affinity scores were used in training an MLR model. To select L2-regularization parameters, 10-fold cross-validation was used. All datasets were divided into 10-fold training and test datasets.

for predicted binding scores based on the three tetramer tables (derived from MC, MD or XRC data, denoted 1mer+4shape$_{MC4}$, 1mer+4shape$_{MD4}$ and 1mer+4shape$_{XRC4}$ models, respectively) and the MC-derived pentamer table (1mer+4DNAshape model) compared to the sequence-only 1mer model were plotted for each TF (Figure 3). R^2 values for the four models that included shape features were significantly greater than values for the 1mer model. Inclusion of shape information substantially improved R^2 values compared to the 1mer model for the gcPBM data (bHLH family), SELEX-seq data (homeodomain family), and HT-SELEX data (multiple TF families).

For each bp, we needed four binary variables to encode the 1mer sequence and four numerical values to encode shape features. A typical TF–DNA binding dataset contains DNA sequences of 12 bp in length. Thus, we performed MLR with 96 features for each dataset. To ensure satisfactory prediction accuracy, large amounts of data were required. The three datasets we used all had more than 1000 entries available for training; thus, additional filtering was not performed. In practice, if a dataset had an R^2 value below 0, it was removed. Fortunately, the three datasets that we used (gcPBM, SELEX-seq and HT-SELEX) all were of high quality.

We compared performances of models combining the 1mer sequence with the four shape features from different data sources (Figure 4A and Supplementary Figure S1). Small differences in R^2 values were observed between DNAshape (MC- and pentamer-based), MC, MD and XRC (all tetramer-based). However, when we removed the 1mer sequence features, which are crucial for prediction due to hydrogen bonds and direct contacts between amino acids and the bases, larger differences in the prediction accuracies of models from different sources started to emerge. An explicit comparison of R^2 values (Supplementary Figure S2) indicated that XRC data provided lower-quality structural information than the two computational approaches (MC and MD simulations). This finding became even more apparent when we considered the weighted average over R^2 values for each dataset based on their sequence quantities. This approach allowed us to compare the prediction accuracy across the three datasets (Figure 4B) or across all different TF families (Supplementary Figure S3). Results of plotting the MSE values of one model against those of another model (Supplementary Figure S4) confirmed these conclusions.

Whereas shape features greatly benefitted models based on gcPBM datasets, even in the absence of sequence features, this was not the case for SELEX-seq and HT-SELEX datasets. For these two datasets, using only shape features was not comparable to using only the 1mer sequence features. However, the 1mer sequence+shape models continuously outperformed the 1mer sequence-only models for all three datasets, in agreement with previous studies using MC data (13,19).

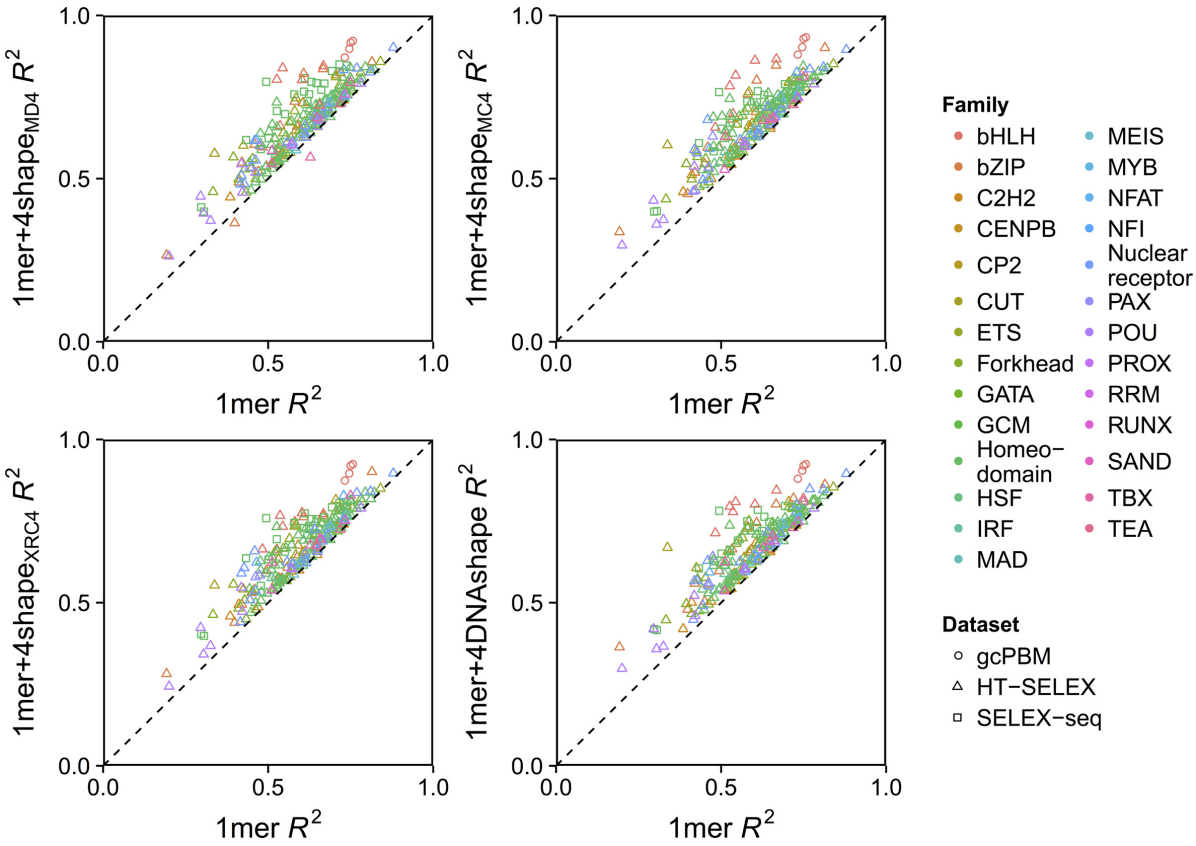

Figure 3. Direct comparison of R^2 values between 1mer+4shape versus 1mer models. As an indicator of the accuracy of predicted TF binding specificity using the trained MLR model, R^2 was computed on the test dataset. Shape features were predicted based on tetramer query tables derived from (**A**) MC data (1mer+4shape$_{MC4}$ model), (**B**) MD data (1mer+4shape$_{MD4}$ model) and (**C**) XRC data (1mer+4shape$_{XRC4}$ model) and (**D**) a pentamer query table derived from MC data (1mer+4DNAshape model). 1mer indicates that DNA sequences were encoded as mononucleotide occurrences. 1mer+4shape indicates that DNA sequences were encoded as 1mer features augmented by four DNA shape features (MGW, ProT, HelT and Roll).

Model performance improves as the number of DNA shape features increases

We explored whether the additional nine DNA shape features (Buckle, Opening, Rise, Shear, Shift, Slide, Stagger, Stretch and Tilt) improved existing binding specificity models. These DNA structural features were previously defined (30,46) but not available in HT predictions (27,47). After validating these features using Spearman's rank correlation coefficients (Supplementary Table S1), we added each of the 13 features to models individually, using the sequence-only model as a baseline. Figure 5A presents the performance differences (in ΔR^2) between the 1mer+1shape, 1mer+2shape and sequence-only 1mer models. By sorting the shape features based on their average performance with four different data sources (MC, MD, XRC tetramer-based and MC pentamer-based), we revealed the order of importance for each feature when added to a 1mer model (Figure 5A). Regardless of which feature we added, the performance of any 1mer+1shape model was improved compared to the 1mer model. The next-most-informative feature would be the one whose inclusion in the best 1mer+1shape model provided the greatest gain in performance over the 1mer+1shape model. The resulting model after adding the second shape feature became the 1mer+2shape model (Figure 5A). Although feature importance among structural parameters varied depending on the experimental dataset that was used, the effect of adding another shape feature into the sequence-only models was consistently beneficial.

When we continued to add shape features to the best 1mer+2shape model, the average performance increased further (shown versus number of shape features in the model in Figure 5B). Here, we used DNAshape as our standard data source to choose the next-most-informative feature in each round. Although model performances varied among different experimental datasets, adding increasingly more shape features to the 1mer model led to a general upward trend in performance. The MSE results, representing model performance (Supplementary Figure S5), supported this finding.

Our results indicate that models trained on gcPBM data outperformed models trained on SELEX-seq and HT-SELEX data (Figure 5B; Supplementary Figures S5 and S6). This finding is not surprising given the higher quality of gcPBM data due to inclusion of 15 bp flanking a binding site 5′ and 3′ of its core (10). To demonstrate the importance of including information on longer binding sites or flanking regions, we tested our models on additional data, a widely used universal PBM (uPBM) dataset, generated by the fifth dialogue for reverse engineering assessments and methods

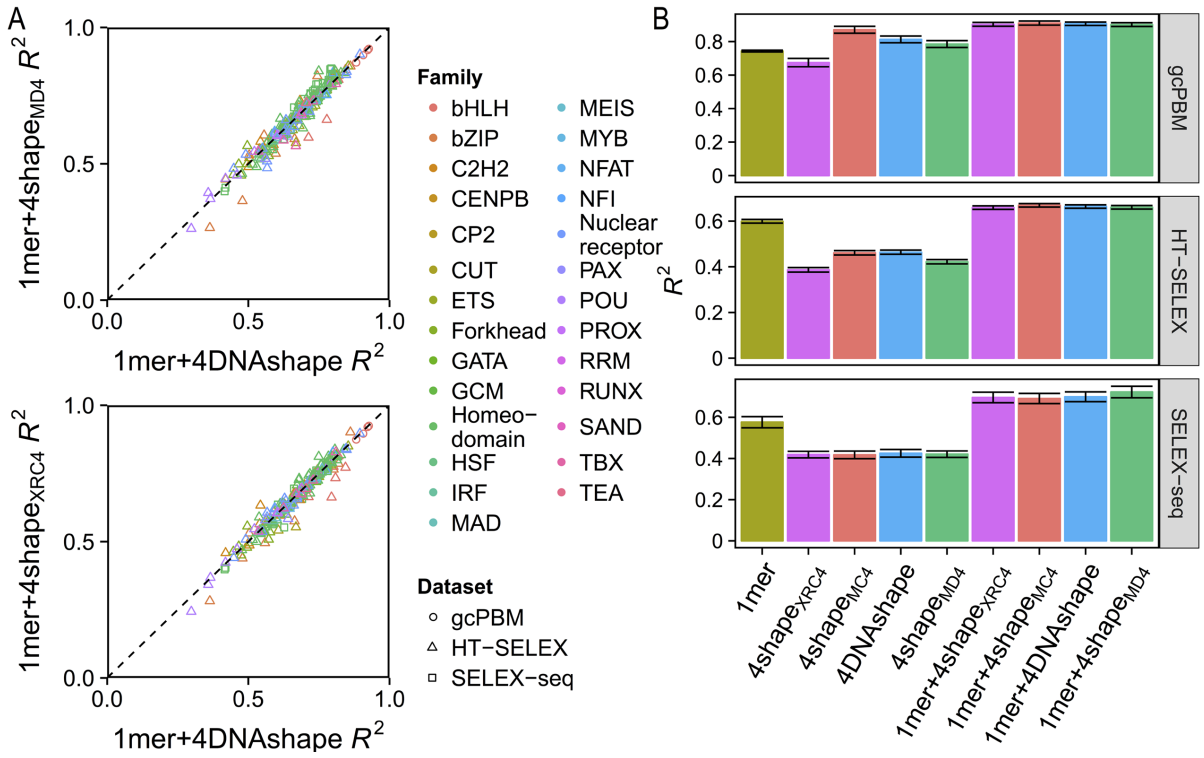

Figure 4. (**A**) Direct comparison of R^2 values between 1mer+4shape models. Points near the diagonal suggest similar performance of compared models. (**B**) Summary of weighted-average performance for different models. Shape features used here were MGW, ProT, HelT and Roll. Error bars were calculated based on standard errors of the mean. Performances were divided into three groups based on different experimental methods (see Supplementary Figure S3 for groups based on TF families). Identical color was used for models using the same source of shape features.

(DREAM5) (43). This dataset only contained information on 8–10 variable bp centered at the core of the binding site (see Supplementary Materials and Methods; Supplementary Figure S6). Therefore, the data were of much lower quality than the SELEX-seq or HT-SELEX-derived experimental data.

It was important to determine whether our inclusion of nine additional shape features in the models directly influenced the ability to classify different TF families. We selected the top-affinity binding sites in HT-SELEX data among multiple TF families and encoded them using our concatenated feature vector. We applied principal component analysis (PCA), rather than an MLR model, on these vectors to evaluate whether the additional shape features could help in separating different TF families. The results revealed that introducing nine additional shape features was beneficial for classifying TF families (Supplementary Figure S7).

Performance comparison of shape-augmented models versus *k*-mer-based models

Encoding a sequence of length N into k-mer features requires $N \times 4^k$ features. The number of required features will dramatically increase as k increases, especially if k is >3. We compared performances of our 1mer+13shape models with those of 2mer (dinucleotide) models. We also compared performances of the 1mer+13shape models complemented by SDs for every shape feature with performances of the 3mer (trinucleotide) models (Figure 6). Including SDs might be a way to represent DNA conformational flexibility, which seems to be an important property of DNA binding sites.

The number of features used in 1mer+13shape+ SD models (30 features per nucleotide position) was still far below the number of features per nucleotide used in 3mer models (64 features per nucleotide position) (Figure 6). Reduction of the computational cost compared to k-mer sequence models is a major advantage of augmenting the 1mer model with shape features (13). We analyzed whether, at a similar computational cost, the 1mer+Nshape models can outperform k-mer models. Our study revealed that the 1mer+13shape model (17 features per nucleotide position) outperformed 2mer and 3mer models for gcPBM datasets (Figure 6). When considering computational cost (feature quantity), sequence+shape models performed consistently better (Supplementary Figure S8). However, 1mer+13shape models did not outperform 3mer models for the SELEX-seq and HT-SELEX datasets, due to the lack of DNA shape information at the 5′ and 3′ terminals of each sequence. The effect of adding information from a single 3mer at the end was shown in (19) to boost the performance of 1mer+shape models. Therefore, if we could find a feasible method to predict shape features at the terminal ends of DNA sequences, the performance of 1mer+shape models would be higher.

The performance also increased when we included SDs in our models, suggesting that SDs can potentially be used to model DNA flexibility (Figure 6). However, we did not compute the SD values from the trajectory of a simulation

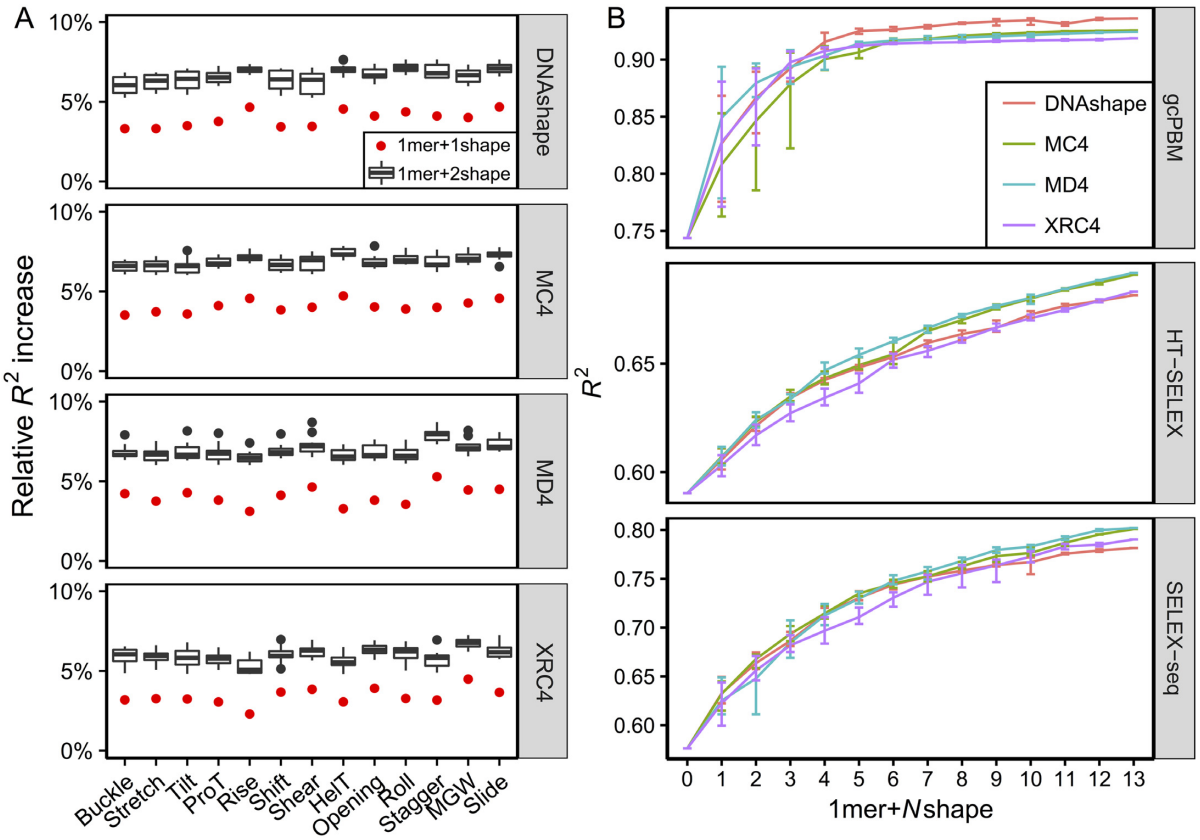

Figure 5. (**A**) Box and dot plots showing performance gain after adding one or two shape features into the 1mer model. Red dots represent the performance of adding one shape feature (1mer+1shape model). Box plots illustrate the performance of adding two shape features (1mer+2shape model), including the feature indicated on the *x*-axis. Black dots are outliers of the box plots. Shape features were sorted based on their average performance of four data sources. (**B**) Plots generated after repeatedly adding shape features into the previous model. In each round, the most-informative feature was added based on performance of using the DNAshape query table. Model performance was evaluated by using a weighted mean R^2 among all datasets in each experimental category. Error bars indicate maximum and minimum performance when *N* shape features were added. DNAshape is pentamer-based; MD, MC and XRC data are tetramer-based due to limitations in sequence coverage for the MD and XRC methods.

(see 'Methods and Materials' section). Therefore, further improvements in deriving SDs or including DNA flexibility are needed. In general, our findings indicate that information can be learned from varying conformational flexibilities of different DNA segments, and that flexibility is a very important feature in protein–DNA binding. The PCA results further support this conclusion (Supplementary Figure S9).

DNAshapeR bioconductor package for additional shape features

To make the additional DNA shape information broadly available, we extended our DNAshapeR package (47) built in R/Bioconductor. The software package is now not only able to predict the previously provided four shape features (MGW, Roll, ProT and HelT), but can also generate DNA shape predictions for the additional nine DNA shape features derived from an MC-derived pentamer query table. To use the additional shape features, the user must input additional parameters into the R function. The package and its expanded manual are available at Bioconductor (http://www.bioconductor.org/packages/devel/bioc/html/DNAshapeR.html) and GitHub (http://tsupeichiu.github.io/DNAshapeR/). The software can predict, plot, and encode values of 13 DNA shape features for any number of DNA sequences or entire genomes.

DISCUSSION

In this study, we applied shape-augmented machine-learning models to predict TF binding specificities and compared the effects of DNA shape features from different data sources (MC and MD simulations and XRC experimental data). We used three representative experimental datasets, acquired by gcPBM, SELEX-seq and HT-SELEX HT binding assays. The datasets contained tens or hundreds of TFs and offered variable data qualities. All datasets had comparable prediction accuracies, as indicated by previous analyses (4,13,19,45). Our shape-augmented regression models outperformed models without shape information. Regardless of the source, structural information improved the accuracy of predictions of TF binding specificities. When we combined sequence with four shape features, acquired from DNAshape (pentamer), MC (tetramer), MD (tetramer) and XRC (tetramer), all provided significant improvements compared to 1mer sequence-only models. Given the consistency of our results, we conclude that, at

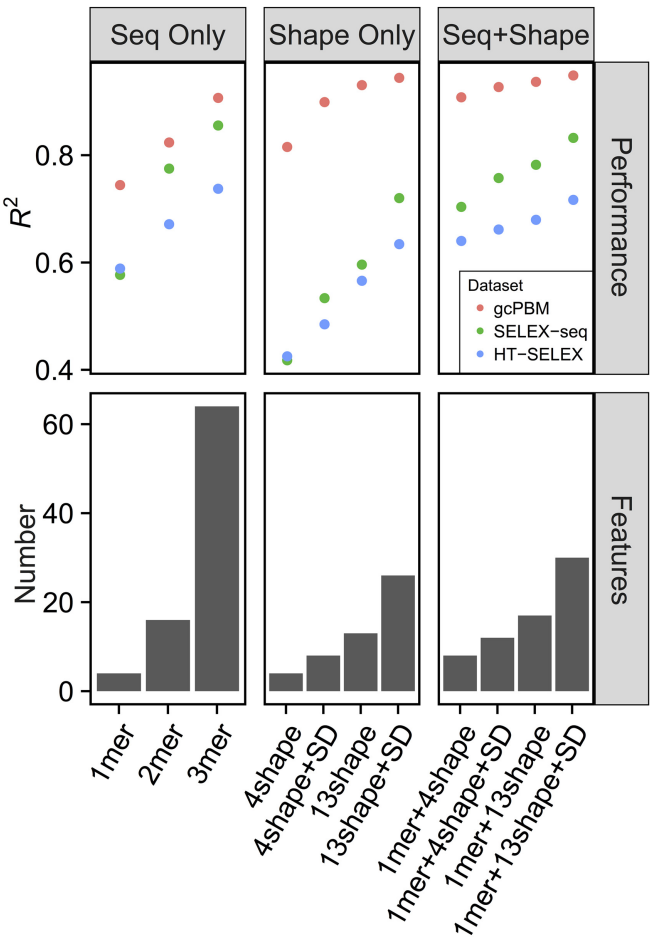

Figure 6. Summary of weighted average performances (R^2) between different models and the feature numbers used in the model. SD was the standard deviation of corresponding shape feature values. All shape features in this figure were generated from the MC query table. The four shape features were MGW, ProT, HelT and Roll. The 13 shape features were MGW, six inter-bp and six intra-bp parameters. Adding more shape features increased the feature number (computational cost). See Supplementary Figure S8 for a direct comparison between feature quantity and model performance.

least in predicting TF binding specificities, adding DNA shape information from any of the MD, MC or XRC data sources will produce improved quantitative models.

We tested the prediction performances of nine additional DNA inter- and intra-bp shape parameters. These shape features, predicted based on MC simulation, can be validated on a large number of XRC structures. Adding more DNA shape features to the feature matrix produced performance gains, although saturation was reached for a larger set of shape features. Performance gains were also visible when more DNA shape features derived from MD simulations or XRC data were added. Generally speaking, adding four shape features to the sequence-only models is the key step in improving model performance. If datasets have a large number of sequences, adding more shape features to the model is always favorable; otherwise, one should opt to use as small of a set of shape features as possible. If computational cost is a consideration, sequence+shape models are always preferable over *k*-mer models. Including additional DNA shape features might also enhance shape-augmented thermodynamic modeling approaches (48) and methods for *in vivo* TF binding site prediction (14,18).

We also evaluated the prediction performance after adding SD values for each DNA shape feature. SD values were calculated for each pentamer and each DNA shape category based on the multiple occurrences of that pentamer in our MC-derived dataset. Although our SD values were not derived from a simulation trajectory for a single pentamer occurrence, our results showed that using SD values, which can be considered as a surrogate of DNA flexibility, is a promising approach in predicting TF–DNA binding. However, one should exercise caution when using SD values because they are currently impossible to validate with XRC data. Further studies are required to identify the role of conformational flexibility in TF binding specificity. After submission of this manuscript, an article was published that reported the derivation of additional features describing DNA structure and flexibility from MD simulations using a different protocol (49). Likely due to that protocol's limitations (see Supplementary Methods and Materials), models using these alternate MD-derived features (49) were outperformed by the models based on any of the three datasets used in this study (MC, MD or XRC data) (Supplementary Figures S11 and 12).

A limitation of our expanded repertoire of DNA shape features is that we did not include phosphodiester backbone features, which are an important subset of DNA shape features. Backbone features are difficult to validate through experimental structures because variations are not well captured at limited XRC resolution or NMR accuracy. Further validation is required to unlock the use of backbone features. Another limitation is that we calculated each DNA shape feature based on ensemble averages, which may not capture the actual distribution. For certain shape features (e.g. backbone dihedrals and HelT), the distribution might be bimodal (50) and computing the mean might not be the best way to encode DNA shape information.

CONCLUSION

Our work demonstrated that DNA shape features are important in TF binding, regardless of the source used to acquire the structural information (MC, MD or XRC data). To the best of our knowledge, our study is the first to derive DNA structural features on a HT basis from multiple different data sources and to test these features in statistical machine-learning applications. Even when derived from different data sources, structural information consistently improved prediction of TF binding specificity. Furthermore, we provided evidence that sequence+shape models, especially models using the expanded repertoire of 13 DNA shape features, offer advantages over *k*-mer models in terms of performance, computational cost and interpretability. For the benefit of the community, the expanded repertoire of 13 DNA shape features has been made available for use in our R package, DNAshapeR, at http://bioconductor.org/packages/release/bioc/html/DNAshapeR.html.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge the Banff International Research Station (BIRS) workshop 15w5167 at Casa Matemática Oaxaca, which initiated this collaborative work, and the Ascona B-DNA Consortium (ABC) for providing the tetramer query table for DNA shape features derived from 1-μs MD simulations of DNA fragments that contain all unique tetramers. See Supplementary Data for detailed author contributions.

FUNDING

Andrew J. Viterbi Fellowship (to J.L.); USC Graduate School, Research Enhancement Fellowship and Manning Endowed Fellowship (to T.P.C.); National Institutes of Health [R01GM106056, U01GM103804 to R.R.; R01HG003008 to R.R., in part]; Alfred P. Sloan Foundation (to R.R.); European Union's H2020 MuG project [676556 to M.P.]. Funding for open access charge: National Institutes of Health [R01GM106056].
Conflict of interest statement. None declared.

REFERENCES

1. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.
2. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.
3. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
4. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
5. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
6. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
7. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
8. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.
9. Sharon,E., Lubliner,S. and Segal,E. (2008) A feature- based approach to modeling protein–DNA interactions. *PLoS Comput. Biol.*, **4**, e1000154.
10. Gordân,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
11. Zabet,N.R. and Adryan,B. (2015) Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Res.*, **43**, 84–94.
12. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
13. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordân,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
14. Yang,J. and Ramsey,S.A. (2015) A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites. *Bioinformatics*, **31**, 3445–3450.
15. Krietenstein,N., Wal,M., Watanabe,S., Park,B., Peterson,C.L., Pugh,B.F. and Korber,P. (2016) Genomic nucleosome organization reconstituted with pure proteins. *Cell*, **167**, 709–721.
16. Abe,N., Dror,I., Yang,L., Slattery,M., Zhou,T., Bussemaker,H.J., Rohs,R. and Mann,R.S. (2015) Deconvolving the recognition of DNA shape from sequence. *Cell*, **161**, 307–318.
17. Villa,R., Schauer,T., Smialowski,P., Straub,T. and Becker,P.B. (2016) PionX sites mark the X chromosome for dosage compensation. *Nature*, **537**, 244–248.
18. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.
19. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.
20. Schöne,S., Jurk,M., Helabad,M.B., Dror,I., Lebars,I., Kieffer,B., Imhof,P., Rohs,R., Vingron,M., Thomas-Chollier,M. *et al.* (2016) Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat. Commun.*, **7**, 12621.
21. Shakked,Z., Guzikevich-Guerstein,G., Frolow,F., Rabinovich,D., Joachimiak,A. and Sigler,P.B. (1994) Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature*, **368**, 469–473.
22. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, doi:10.1186/gb-2000-1-1-reviews001.
23. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell*, **8**, 937–946.
24. Locasale,J.W., Napoli,A.A., Chen,S., Berman,H.M. and Lawson,C.L. (2009) Signatures of protein-DNA recognition in free DNA binding sites. *J. Mol. Biol.*, **386**, 1054–1065.
25. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
26. Dantas Machado,A.C., Saleebyan,S.B., Holmes,B.T., Karelina,M., Tam,J., Kim,S.Y., Kim,K.H., Dror,I., Hodis,E., Martz,E. *et al.* (2012) Proteopedia: 3D visualization and annotation of transcription factor-DNA readout modes. *Biochem. Mol. Biol. Educ.*, **40**, 400–401.
27. Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56–W62.
28. Olson,W.K., Bansal,M. and Burley,S.K. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
29. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
30. Lavery,R. and Sklenar,H. (1989) Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.*, **6**, 655–667.
31. Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
32. Rohs,R., West,S.M., Liu,P. and Honig,B. (2009) Nuance in the double-helix and its role in protein–DNA recognition. *Curr. Opin. Struct. Biol.*, **19**, 171–177.
33. Dixit,S.B., Beveridge,D.L., Case,D.A., Cheatham,T.E. 3rd, Giudice,E., Lankas,F., Lavery,R., Maddocks,J.H., Osman,R., Sklenar,H. *et al.* (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.*, **89**, 3721–3740.
34. Rohs,R., Sklenar,H. and Shakked,Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, **13**, 1499–1509.

35. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
36. Chen,Y., Zhang,X., Dantas Machado,A.C., Ding,Y., Chen,Z., Qin,P.Z., Rohs,R. and Chen,L. (2013) Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. *Nucleic Acids Res.*, **41**, 8368–8376.
37. Chang,Y.P., Xu,M., Dantas Machado,A.C., Yu,X.J., Rohs,R. and Chen,X.S. (2013) Mechanism of origin DNA recognition and assembly of an initiator-helicase complex by SV40 large tumor antigen. *Cell Rep.*, **3**, 1117–1127.
38. Dantas Machado,A.C., Zhou,T., Rao,S., Goel,P., Rastogi,C., Lazarovici,A., Bussemaker,H.J. and Rohs,R. (2015) Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics*, **14**, 61–73.
39. Li,J., Dantas Machado,A.C., Guo,M., Sagendorf,J.M., Zhou,Z., Jiang,L., Chen,X., Wu,D., Qu,L., Chen,Z. *et al.* (2017) Structure of the forkhead domain of FOXA2 bound to a complete DNA consensus site. *Biochemistry*, **56**, 3745–3753.
40. Pasi,M., Maddocks,J.H., Beveridge,D., Bishop,T.C., Case,D.A., Cheatham,T., Dans,P.D., Jayaram,B., Lankas,F., Laughton,C. *et al.* (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.*, **42**, 12272–12283.
41. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
42. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical J.*, **63**, 751–759.
43. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
44. Ma,W., Yang,L., Rohs,R. and Noble,W.S. (2017) DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics*, **33**, 3003–3010.
45. Yang,L., Zhou,T., Dror,I., Mathelier,A., Wasserman,W.W., Gordân,R. and Rohs,R. (2014) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **42**, D148–D155.
46. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
47. Chiu,T.-P., Comoglio,F., Zhou,T., Yang,L., Paro,R. and Rohs,R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211–1213.
48. Peng,P.-C. and Sinha,S. (2016) Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Res.*, **44**, e120.
49. Andrabi,M., Hutchins,A.P., Miranda-Saavedra,D., Kono,H., Nussinov,R., Mizuguchi,K. and Ahmad,S. (2017) Predicting conformational ensembles and genome-wide transcription factor binding sites from DNA sequences. *Sci. Rep.*, **7**, 4071.
50. Dans,P.D., Pérez,A., Faustino,I., Lavery,R. and Orozco,M. (2012) Exploring polymorphisms in B-DNA helical conformations. *Nucleic Acids Res.*, **40**, 10668–10678.