12604–12617 Nucleic Acids Research, 2020, Vol. 48, No. 22 Published online 2 December 2020
doi: 10.1093/nar/gkaa1134

Integrating genome sequence and structural data for statistical learning to predict transcription factor binding sites

Pengpeng Long [1], Lu Zhang [1], Bin Huang [1], Quan Chen [1,2,*] and Haiyan Liu [1,2,3,*]

[1] School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China, [2] Hefei National Laboratory for Physical Sciences at the Microscale, Hefei, Anhui 230026, China and [3] School of Data Science, University of Science and Technology of China, Hefei, Anhui 230026, China

Received June 22, 2020; Revised September 18, 2020; Editorial Decision November 04, 2020; Accepted November 10, 2020

ABSTRACT

We report an approach to predict DNA specificity of the tetracycline repressor (TetR) family transcription regulators (TFRs). First, a genome sequence-based method was streamlined with quantitative *P*-values defined to filter out reliable predictions. Then, a framework was introduced to incorporate structural data and to train a statistical energy function to score the pairing between TFR and TFR binding site (TFBS) based on sequences. The predictions benchmarked against experiments, TFBSs for 29 out of 30 TFRs were correctly predicted by either the genome sequence-based or the statistical energy-based method. Using *P*-values or *Z*-scores as indicators, we estimate that 59.6% of TFRs are covered with relatively reliable predictions by at least one of the two methods, while only 28.7% are covered by the genome sequence-based method alone. Our approach predicts a large number of new TFBs which cannot be correctly retrieved from public databases such as FootprintDB. High-throughput experimental assays suggest that the statistical energy can model the TFBSs of a significant number of TFRs reliably. Thus the energy function may be applied to explore for new TFBSs in respective genomes. It is possible to extend our approach to other transcriptional factor families with sufficient structural information.

INTRODUCTION

Transcription factors (TFs) are key players in gene regulation. Thanks to the rapid development of DNA sequencing technologies, a large number of TFs can be identified from the many sequenced genomes of various organisms (1–3). Knowing the DNA sequences specifically recognized by these TFs is important for identifying the TFs' target genes and for elucidating gene regulatory networks (4,5). Experimental approaches can provide comprehensive information about the DNA sequences recognized by particular TFs (6–13). However, such experimental analyses have only covered a small fraction of TFs identifiable from known genome sequences. It is thus of great interest to develop computational methods to predict TF-binding DNA sequences (TF binding sites or TFBSs) with available sequence (14–27) and structural data (28–33).

As one of the largest TF families in bacteria (34), the tetracycline repressor (TetR) family of regulators (TFRs) are involved in many important cellular processes such as antibiotic resistance (35) and biofilm formation (36). Up to now, the TFBSs of only dozens of TFRs have been experimentally known (37), and genome sequence-based TFBS predictions have been reported for more but still a limited number (24). TFRs usually function as homodimers, each monomer recognizing a specific half-palindromic DNA sequence via its DNA binding domain (DBD). The DBDs of TFRs are of highly similar structures, each DBD containing a structurally conserved helix-turn-helix (HTH) motif, which constitutes the part of the protein that directly interacts with the DNA (24). Thus, it is the amino acid sequence of the HTH motif of a TFR that determines the TFR's DNA specificity at the level of half palindromes. A model that correctly captures this relationship between amino acid and nucleotide sequences can facilitate the TFBS prediction for TFRs.

In this work, we introduce a model in which three different aspects, including genome sequences, structural information of protein-DNA complexes, and statistical learning, are integrated to enable sequence-based TFBS predictions for a significant number of TFRs. As briefly introduced below, although these aspects have been important ingredients of different published methods for TFBS prediction, they are usually separately involved in different studies, not in a single approach.

[*] To whom correspondence should be addressed. Email: hyliu@ustc.edu.cn
Correspondence may also be addressed to Quan Chen. Email: chenquan@ustc.edu.cn

In principle, TFBSs may be predicted by quantifying the specific DNA-protein interactions based on structures of protein-DNA complexes (28–32). However, the applicability of structure-based predictions is still strongly inhibited by the relative scarcity of high quality structural data as well as by inaccuracies of current computational models for quantifying molecular interactions (28).

Structure-independent methods for TFBS predictions have been proposed to consider solely genome sequences, using ideas of phylogenetic footprinting (14–16,21–24). Basically, such predictions are made based on two hypotheses. The first is that the TFBSs recognized by a particular TF are likely to be present or enriched within certain genome regions given the locations of the TF or its targeted genes in genome DNA. The second is that these TFBSs are likely to be more conserved than their surrounding sequences. These assumptions allows TFBSs to be proposed as shared sequence motifs across different regions in the same genome or in different genomes containing TFs sharing DNA specificity (15). For a significant number of TFs from prokaryotic organisms, their TFBS motifs are enriched in the genome regions near the genes encoding the TFs themselves, allowing predictions to be made solely based on the sequences of genomes containing TFs anticipated to share DNA specificity. These ideas have been exploited by Francke *et al*. (22) to predict TFBSs in *Lactobacillus plantarum*, by Yan *et al*. (23) to predict TFBSs in *Geobacter sulfurreducens*. and by Yu *et al*. (24) to predict TFBSs of the tetracycline repressor (TetR) family of repressors.

For the phylogenetic footprinting approach to yield reliable predictions, multiple homologous TFs sharing DNA specificity with the TF of interest must exist in sequenced genomes, and their TFBSs should exist in close proximity to the TFs' encoding genes. In addition, the sequences surrounding the actual TFBS motifs in different genomes must be far more variable than the TFBS motifs themselves. Otherwise the false positive rate would be high (25). For many TFs, these conditions are not met by the available data and reliable predictions cannot be made solely based on genome sequences.

Another promising type of methods suitable for large scale TFBS prediction is statistical learning or machine learning (38–41). By supervised learning, computational models are trained using datasets comprising TFs with known TFBSs. The trained models can be applied to make predictions for TFs that are different from but still closely related to the training TFs. Using experimentally characterized DNA binding properties and guided by structural models, Anton *et al*. (38,39) defined and trained support vector machines (SVM) to predict the DNA binding preferences of C_2H_2 Zinc finger domains. Khamis *et al*. (40) trained random forest models to summarize experimentally known DNA sequence preferences of eukaryotic TFs. More recently, Alipanahi *et al*. (42) trained deep learning models to predict sequence specificities of RNA binding and DNA binding proteins based on experimental binding data. In general, machine learning models are limited mainly by available training data. To train a model suited for TFs of diverged amino acid sequences and DNA specificity, a large number of TFs with known TFBS sequences are needed. In addition, the training TFBS nucleotide sequences need to be pre-aligned correctly for most machine learning models so that the nucleotide types at individual positions can be encoded by the correct components of sequence-encoding vectors (40) (although the deep learning method used by Alipanahi *et al*. (42) did not consider pre-aligned nucleotide sequences, it did require a large amount of experimental data for each individual TF so that sequence patterns of TFBSs could be extracted and used by the underlying neural networks). When the TFBS sequences of different training TFs are not similar enough and do not observe well-defined sequence patterns, the lack of correct DNA sequence alignments becomes a major obstacle for machine learning. Because of this, most current machine learning models have been trained for small groups of TFs, for each of which available experimental data were sufficient to allow its TFBS sequences to be aligned according to sequence similarity or sequence patterns.

In the current work, genome sequences, structural data, and statistical learning are integrated in a single approach to enable the prediction of TFBS sequences based on the amino acid sequence of the DBDs of TFRs, which are available in large numbers from sequenced prokaryotic genomes. The first part of this approach is a streamlined version of genome sequence-based TFBS prediction (15,19). This automated version enables large scale predictions and produces *P*-values to measure the significance of individual prediction results. In the second part of the approach, the diverse TFBS nucleotide sequences predicted with small *P*-values are aligned together based on the structures of the various TFR-DNA complexes. In this step, the small number of DNA motifs in the protein-DNA complexes are first aligned together according to the structure alignments between the DNA-binding motifs of the proteins. Then these DNA motifs from the protein-DNA complexes of known structures are separately used as seeding sequences in psi-BLAST-like (43) sequence search and alignment processes against the TFBSs predicted with low uncertainty by the genome sequence-based method. This leads to a unified set of TFBS DNA sequences jointed with DBD amino acid sequences, with all sequences properly aligned, and each TFBS sequence correctly paired with its cognate DBD sequence. In the last part of our approach, this unified set of aligned sequences are used to train a statistical energy defined as a function of both the amino acid sequence of the DBD HTH motif and the nucleotide sequence of the TFBS motif. To verify our model, besides using pre-existing experimental data, we perform new experiments on some TFRs to test their binding with respective TFBSs predicted from their upstream genome sequences. For several TFRs, high-throughput Spec-Seq experiments (11–13) are carried out to verify the TFBS sequence profiles represented by the statistical energy function.

MATERIALS AND METHODS

Figure 1 shows the overall framework of the computation and analysis steps of the current study. Details of these steps are described below and in Supplementary Methods.

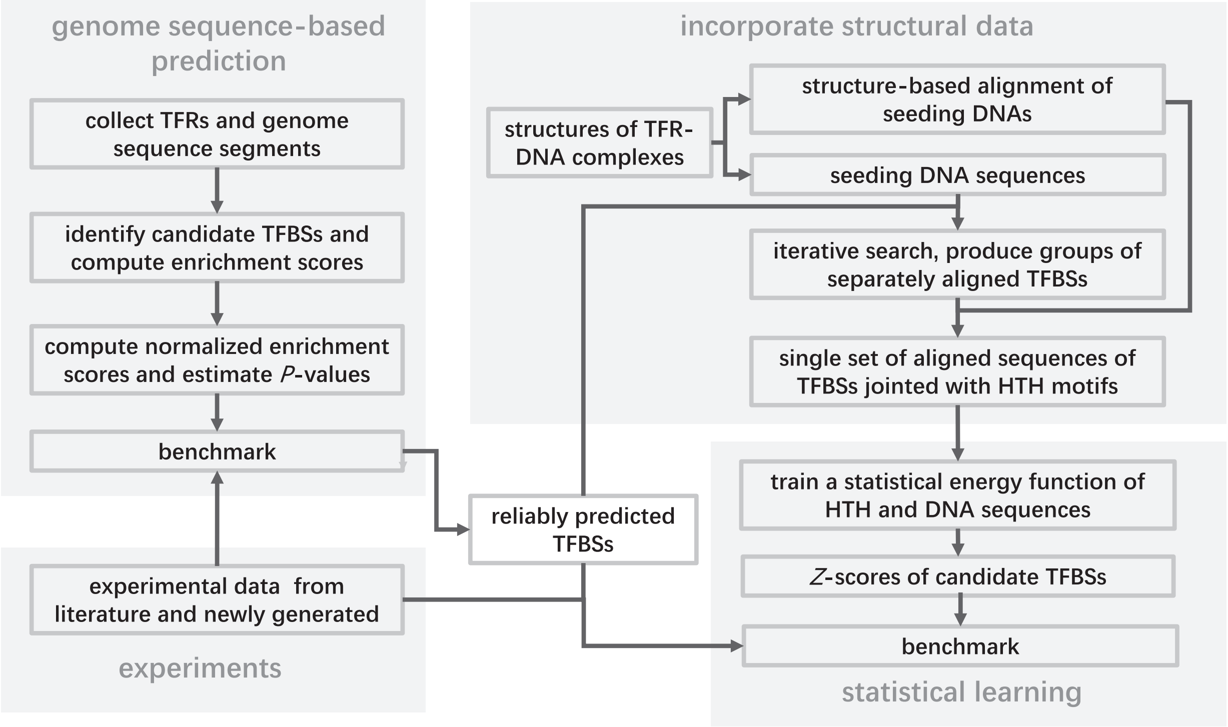

genome sequence-based prediction
collect TFRs and genome sequence segments
identify candidate TFBSs and compute enrichment scores
compute normalized enrichment scores and estimate P-values
benchmark
experimental data from literature and newly generated
experiments
reliably predicted TFBSs
incorporate structural data
structures of TFR-DNA complexes
structure-based alignment of seeding DNAs
seeding DNA sequences
iterative search, produce groups of separately aligned TFBSs
single set of aligned sequences of TFBSs jointed with HTH motifs
train a statistical energy function of HTH and DNA sequences
Z-scores of candidate TFBSs
benchmark
statistical learning

Figure 1. An overall framework of the computational and analysis steps.

A streamlined workflow to predict TFBSs based on genome sequences

TFR dataset and genome sequence degeneracy weights. According to previous studies (24,44), the TFBSs of TFRs are enriched in the genome sequence segments upstream the genes encoding TFRs. We will refer to the upstream genome sequence segments as GSSs. A dataset of >197 000 TFRs paired with respective GSSs have been defined (see Supplementary Methods). Some of the different GSSs may be of highly similar overall sequences, potentially causing problems of high false positive rates for motif enrichment-based TFBS predictions (25). To offset this effect, each GSS i has been assigned a numerical GSS degeneracy weight $w_i^g = 1/n_i$, where n_i is the number of GSSs in the dataset that are globally similar to i(specifically, a BLASTN search of the GSS dataset is performed with GSS i as the query and n_i is the number of hits with alignment coverage > 50% and E-value $< 10^{-15}$).

Candidate TFBSs and enrichment scores. For each TFR, a set of candidate TFBSs were identified as short palindromic fragments (44) within its GSS using the Palindrome program (45) (see Supplementary Methods and Supplementary Figure S1). Following the idea of phylogenetic footprinting (15,16), we assume that the probability for a candidate palindromic fragment to be the actual TFBS of a query TFR is proportional to an enrichment factor. Straightforwardly, this factor can be determined as the number of occurrences of similar DNA motifs in the GSSs of other TFRs that are expected to share the same TFBS specificity as the query (TFRs sharing TFBS specificity are determined based on their amino acid sequence similarity, especially in the HTH motif of the DBD domain, see below). This idea has been implemented in the following quantitative enrichment scores. First, as TFRs of low overall sequence similarity rarely have similar DBD sequences to bind similar TFBSs, to increase computational efficiency, we limit subsequent analyses to a subset of homologous TFRs that share more than 30% amino acid sequence identity with the query TFR. For each such homolog i, the amino acid sequence of its HTH motif is compared with that of the query TFR to determine n_s, the number of substituted residues within the HTH motif (substitutions at a few chosen positions were not counted in n_s as these positions were found to be unimportant for DNA specificity, see Supplementary Methods and Supplementary Figure S2). For convenience, the number n_s is transformed into an empirical protein similarity weight w_i^p using a tableted function (Supplementary Table S1), the largest possible value of w_i^p of 1.0 corresponding to zero substitutions and the smallest value of 5x10^{-5} corresponding to 14 substitutions. The GSSs of the homologs were incrementally collected into an enrichment set of GSSs for the given query, starting from those of the homologs of the largest w_i^p, until the accumulated total w_i^g of GSSs in the set had reached a value of 20. Then for every candidate TFBS x of the query TFR, an absolute enrichment score $e_{abs}(x)$ was determined by comparing the candidate TFBS with each of the palindromic fragments contained in the enrichment GSS set and using the following formula,

$$e_{abs}(x) = \frac{\sum_{i \in collected\ GSSs} \theta_i(x) w_i^p w_i^g}{\sum_{i \in collected\ GSSs} w_i^p\ w_i^g}, \qquad (1)$$

where w_i^g and w_i^p are respective GSS degeneracy weights and protein similarity weights defined above, and $\theta_i(x)$ represents the number of palindromic fragments contained in GSS $i$that are similar to x. The criterion for similarity between two palindromic fragments is that at least 4 (for shorter palindromes of half lengths no more than 6) or 5 (for longer palindromes of half-lengths longer than 6) of the nucleotides in the half palindromic regions should be the same. If by this criterion the candidate TFBS x occurred in similar forms multiple times in the query TFR's GSS, the scores of those similar motifs were added into the score for x.

Normalization of the enrichment scores. The absolute enrichment scores as calculated by formula (1) depend on heuristic parameter choices. In addition, they depend on the lengths and nucleotide compositions of the candidate TFBSs, and are thus not directly comparable between different candidate TFBSs. To address these issues, the absolute enrichment factors were subjected to the following normalization treatment: a controlling set of 60 000 palindromic fragments were extracted from random genome segments in bacterium genomes contributing to the TFR dataset, and a normalization factor $N(x)$ for a candidate fragment x was determined as

$$N(x) = \frac{max(0,\ m(x) - 1) + 1}{60000}\ , \tag{2}$$

where $m(x)$ is the number of palindromic fragments in the controlling set that are similar to x according to the same criteria as those used to determine $\theta_i(x)$. The normalized enrichment score was calculated as

$$e_N(x)\ = e_{abs}(x)/N(x). \tag{3}$$

Mapping the normalized enrichment scores to P*-values.* The normalized enrichment score $e_N(x)$ was further gauged using a reference distribution of normalized enrichment scores calculated against a set of randomly selected GSSs instead of the set of GSSs selected according to the sequence similarity of the HTH motifs. This step mapped the $e_N(x)$ values to P-values. To estimate the reference distribution, the highest score associated with each of 70 000 randomly chosen TFRs was determined against a randomly selected GSS set. A histogram of the resulting scores, noted as e_0, was obtained. The P-value for predicting the candidate TFBS x to be the true TFBS of the corresponding TFR was defined as the probability $P(e_0\ \geq\ e_N(x))$ in the reference distribution.

Building a statistical model by combining genome sequence-based predictions and structural data

Using structural data to obtain an extended and unified set of aligned TFBSs. Overall, the TFBSs for thousands of TFRs could be predicted with high significance (small P-values). However, these TFBSs are highly divergent in their nucleotide sequences because the corresponding TFRs have diverged HTH amino acid sequences and distinguished DNA specificities. It is not feasible to obtain a unified set of aligned TFBSs covering extensive TFRs by just aligning the diverse DNA sequences. To overcome this difficulty, we used the structural alignments between different TFR proteins in complexes with DNA to bridge the DNA sequence alignments. More specifically, 16 non-redundant TFR-DNA complexes (Supplementary Table S2) from the Protein Data Bank (PDB) (46) were considered. After aligning the HTH motifs according to structures, core regions of the DNA fragments that directly interact with the aligned HTH motifs could also be well-aligned (see results). Thus the TFBS of these TFRs, although not being able to be aligned against each other reliably based on their DNA sequences, could be unambiguously aligned based on the structural alignments between these core regions. Subsequently, each of the TFBSs in the protein-DNA complexes was used as a seeding sequence to retrieve a group of similar TFBSs from the genome-sequence prediction results associated with small P-values, using an incremental approach similar to psi-BLAST (43) (see Supplementary Methods). Within each group, the retrieved TFBSs were aligned according to sequence similarity. Between groups, the alignments were determined according to the bridging structural alignments. We assessed the reliability of this approach by looking at TFBSs that have been found by differently seeded searches. Should the approach be valid, the sequence-based alignments of the same TFBS with different seeding TFBSs should be consistent with the structural alignments between the seeding TFBSs. To further verify the TFBS alignments, we paired each of the aligned TFBS sequences with the amino acid sequence of the corresponding HTH motif, and carried out direct coupling analysis (DCA) (47) between the amino acid and the nucleotide positions in the unified set of aligned amino acid and nucleotide sequences. The DCA results were correlated with inter-residue spatial distances in protein–DNA complexes.

Training a statistical energy function of jointed amino acid sequence and DNA sequence. We used the unified set of aligned amino acid sequences of the HTH motifs and the nucleotide sequences of the TFBSs to learn a statistical energy function $E(a_1, a_2, \ldots, a_{l_p}, b_1, b_2, \ldots, b_{l_d})$ which is related to the joint probability distribution $P(a_1, a_2, \ldots, a_{l_p}, b_1, b_2, \ldots, b_{l_d})$ of the amino acid and nucleotide sequences through the inversed Boltzmann relationship $E\ \ (a_1, a_2, \ldots, a_{l_p}, b_1, b_2, \ldots, b_{l_d}) = -\, ln\, P(a_1, a_2, \ldots, a_{l_p}, b_1, b_2, \ldots, b_{l_d})$, where $l_p =\ 14$ is the number of considered residues in the HTH motifs and $a_1, a_2, \ldots, a_{l_p}$are residue types, and $l_d = 8$ is the number of considered nucleotide in the DNA motifs and $b_1, b_2, \ldots, b_{l_d}$ are base types. More details are given in Supplementary Methods.

Z-scores and sequence profiles from statistical energies. For a TFR of a given HTH amino acid sequence, the statistical energy function was used to estimate the energy associated with an octamer DNA sequence motif, or to scan the GSS of the TFR to identify potential TFBSs. For the latter purpose, we scan each position of the GSS, considering it as a starting position of an octamer half site, followed by a gap segment and another complementary and reversed octamer half site. For each possible TFBS starting position, the gap width was allowed to vary systematically from –2 to 6 (a negative value corresponds to overlapping half sites). Each combination of the starting position with a gap width was considered as a candidate TFBS and its statistical energy computed accordingly. The energies of the two half sites are added together to give the overall energy of the position. The statistical energies of TFBSs for a given TFR were transformed into Z-scores by considering an average energy $\bar{e}$ and a standard variation σ, $Z = \frac{e - \bar{e}}{\sigma}$ When the statistical energy function was applied to scan the GSS of a given TFR, the average and standard deviations have been determined from energies computed by considering every position of the GSS as a possible TFBS starting position and calculating the respective statistical energy. When the energies of octamers were considered, the average energy and standard variation were determined from energies of all 65 536 possible octamers. The statistical energies of all 65 536 octamers can also be transformed into a sequence profile, by assuming the probability of each octamer to be proportional to $\exp(-e)$.

Experimental tests of some predicted TFBSs

From existing literatures, 20 TFRs whose TFBS in respective GSSs have been experimentally confirmed were collected and used to verify our prediction results. Besides these 20 TFRs, we carried out post-prediction validation experiments on 10 additional TFRs whose TFBS have not been reported before. These TFRs were selected so that their corresponding predictions were associated with varied significance as measured by P-values and/or Z-scores. The TFRs have been expressed from synthesized genes using *E. coli* host using standard techniques. Purified proteins were used for in vitro DNA binding assays by either electrophoretic mobility shift assay (EMSA) (48) or DNase I footprinting (49). Specifically, for 4 of the selected TFRs with best P-values smaller than 0.01, EMSA assay was carried out on only the highest-score candidate TFBS of each of them. For another 2 selected TFRs of best P-values of 0.21 and 0.26, respectively, EMSA assays were carried out on 6 highest P-value candidate TFBSs for each TFR. Evaluated by the statistical energy function, the 26 TFRs thus collected are associated with mostly low Z-scores. To purposefully include TFRs associated with higher Z-scores (namely, smaller absolute values), 4 additional TFRs associated with Z-scores between -2.5 to -2 have been selected. For each of these 4 TFRs, DNase I footprinting assays were carried out with DNA substrates containing 6 to 8 lowest Z-scores candidate TFBSs, disregarding their P-values.

To test whether or when the statistical energy function leads to reliable models about of the TFBS motifs, high-throughput Spec-Seq experiments (11,12) were carried out on 4 TFRs whose GSS TFBSs have been predicted with different Z-scores but all verified by experiments. In these experiments, the variability of the TFBS sequences of a TFR were explored in a single assay.

More experimental details are given in Supplementary Methods, Supplementary Tables S3, S4 and S7.

RESULTS AND DISCUSSION

Performance of the genome sequence-based predictions

Enrichment scores and P-values. Figure 2A shows the distribution of the actual normalized enrichment scores $e_N(x)$. For each TFR, only the maximum score has been counted. For comparison, the reference distribution of the controlling scores computed using randomly selected GSS sets is also shown. For a significant number of TFRs, the actual normalized enrichment scores are far above the controlling scores. The cumulated distribution of P-values computed according to the normalized scores and the reference distribution is shown in Figure 2B. For more than 54.8% TFRs, the P-values associated with the maximum scoring TFBSs are no more than 0.05, while for 28.7% of TFRs the P-values are no more than 0.01.

We note that many parameters for the genome sequence-based prediction have to be chosen empirically (some of them by trials and errors) in lack of a large amount of ground truth data against which our model could be optimized. For this reason, the normalization step and the conversion of the enrichment scores into P-values have been designed to minimize the impacts of these choices. Supplementary Figure S3A-C show data illustrating the robustness of the P-values with respect to some limited variations of the function mapping n_s (the number of substituted residues in DBD) to the enrichment weight w_i^p. In Supplementary Figure S3D, normalization factors computed with random palindrome sets containing different numbers of palindromes (30 000 versus 60 000) are compared, the results suggesting that a set of 60 000 palindromes is sufficiently large for the normalization factor to converge. Besides this, the data show that the absolute enrichment score depends strongly on the GC content of the query TFBS, suggesting the necessity for normalization.

We further notice that the actual TFBSs are predicted based not on the exact P-values, but on the ranking according to the estimated P-values. Some variations that have limited but still noticeable effects on the exact P-values (such as the mapping function 1 versus the mapping functions 2 and 3 in Supplementary Figure S3A) turned out to have little effects on the final top ranking TFBSs for most TFRs. This insensitivity added another layer of robustness around the genome-based prediction results with respect to parameter choices.

Experimental tests of predicted TFBSs. The experimental set used to verify the predictions includes 20 TFRs from the literatures and 10 TFRs tested by ourselves by either EMSA or DNase I footprinting assays. An example of the EMSA assay results is shown in Figure 3A, which clearly indicates that the TFR A0A0L8NT36 specifically binds to the predicted maximum scoring candidate TFBS A0A0L8NT36O but not to the TFBS P0ACT4O, which served as a negative control in this experiment. P0ACT4O was shown to bind to its own TFR in another EMSA assay (Supplementary Figure S4). An example of the DNase I footprinting results is shown in Figure 3B, which indicates that the TFR A6FXP3 also specifically binds to the two predicted TFBSs, protecting the sites from DNase I digestion. Results of the other DNA binding assay experiments are given in Supplementary Figure S5.

In Supplementary Table S5, the DNA sequences of the experimentally verified true TFBSs and the highest P-value candidate TFBSs are given. The P-value ranks of the true TFBSs are also given. For 26 of the 30 TFRs, the true

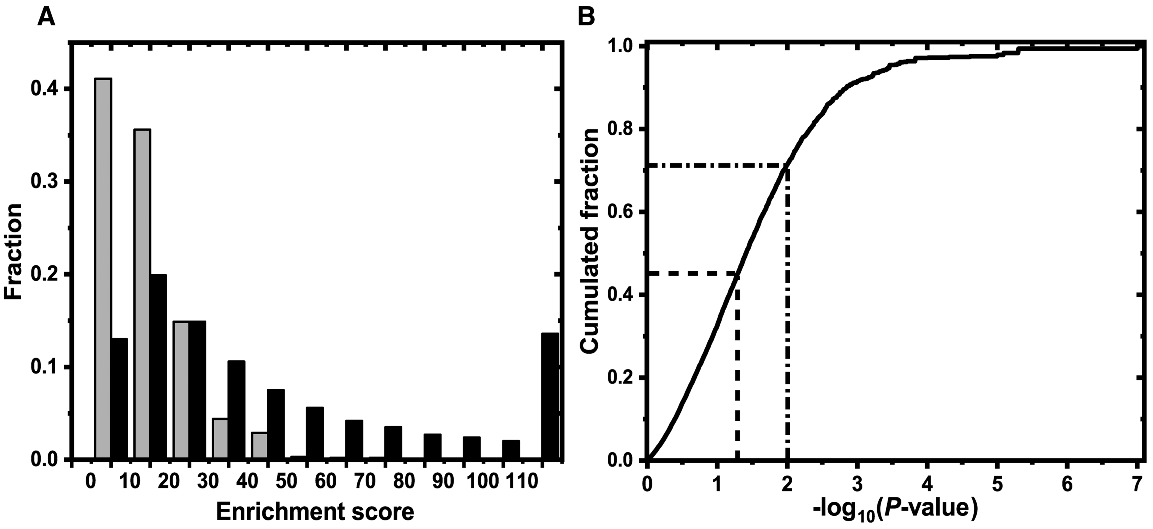


Figure 2. (**A**) Distribution of the actual normalized enrichment scores $e_N(x)$ (black bars) and the reference distribution of the controlling scores (grey bars). The right-most black bar corresponds to the sum of all entries with scores higher than 110. (**B**) The cumulated distribution of $-log_{10}$ (*P*-value) in TFR dataset. The dashed lines indicate the accumulated fraction at the *P*-value of 0.05 and the dash-dotted lines indicate the fraction at *P*-value 0.01.

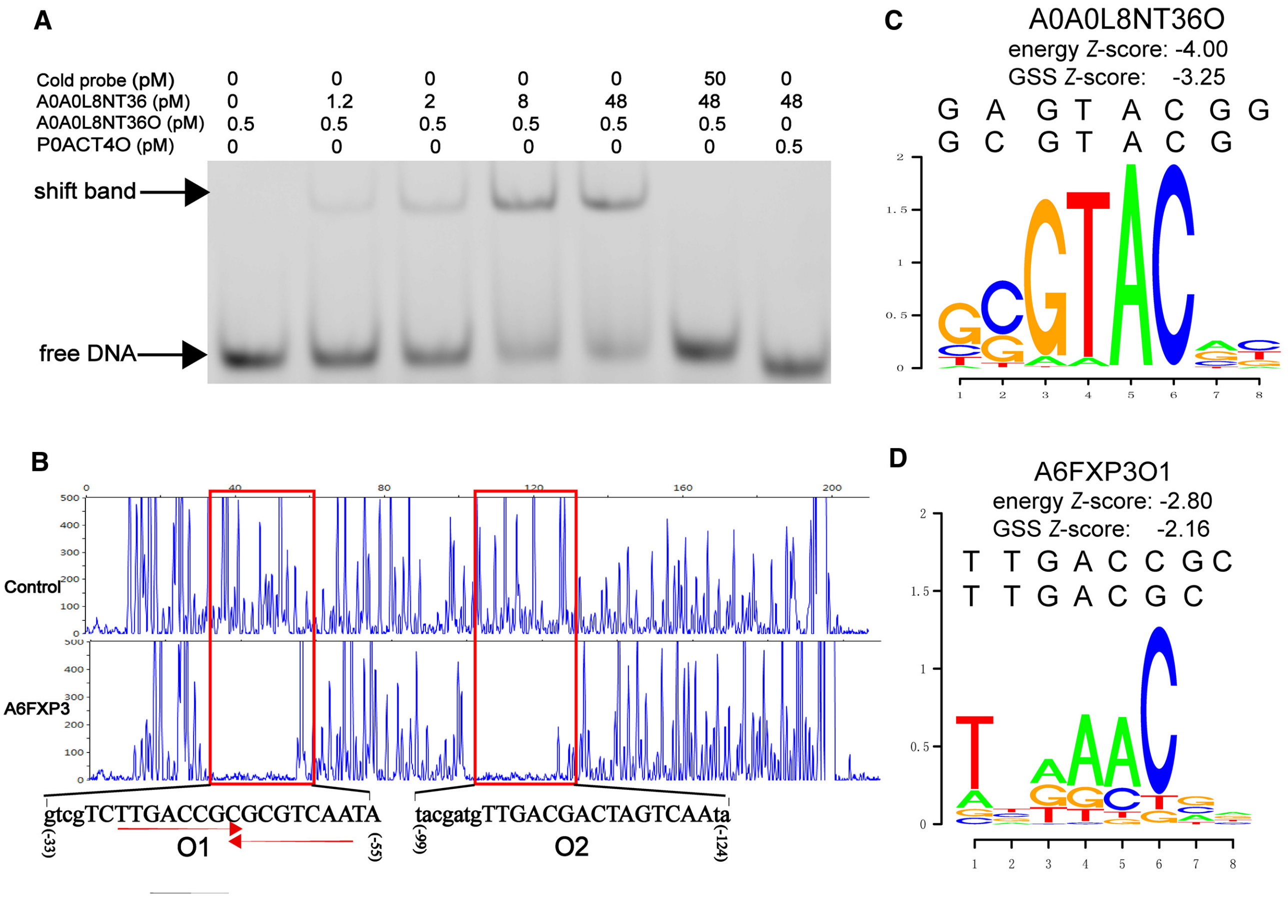


Figure 3. (**A**) The EMSA result for TFR A0A0L8NT36. The bound DNA and unbound DNA are indicated by 'shift band' and 'free DNA', respectively. A0A0L8NT36O is the predicted TFBS and the P0ACT4O a negative control. (**B**) The DNase I footprinting result of TFR A6FXP3. The DNA regions protected by binding with A6FXP3 are indicated by red boxes and the respective nucleotide sequences are shown below the boxes. Capital letters represent regions corresponding to predicted TFBSs. The red arrows below the sequence of TFBS O1 indicate the forward and backward octamer regions, which overlap by one nucleotide in this case. (**C**) and (**D**) show respective sequence logos of A0A0L8NT36 and A6FXP3, generated according to statistical energies computed for all possible nucleotide octamers. The actual nucleotide sequences of the forward and backward palindromic halves of the experimentally verified TFBSs are given with their *Z*-scores.

TFBSs are associated with either the highest *P*-values (21 TFRs) or the second highest *P*-values (5 TFRs). For the remaining 4 TFRs, the true TFBSs ranked between 6 and 34 according to the *P*-values.

P-values indicate the reliability of predictions. Figure 4A shows the *P*-values of the top ranking candidate TFBSs of the 30 TFRs and of the true TFBSs. The data clearly indicate that those top ranking candidate TFBSs associated with lower *P*-values are more likely to be the true TFBSs. According to the data, all predictions associated with *P*-values$\leq$0.01 can be safely accepted as reliable predictions, while there are some chances for top predictions associated with higher *P*-values to be false positives. Thus the *P*-value is a good indicator for reliability of the genome sequence-based predictions. We note that technically, the *P*-value representing statistical significance is a monotonic function of the normalized enrichment score. Thus the enrichment score normalization step in our workflow has been essential for obtaining this quantitative indicator.

Using structural data to obtain an extended and unified set of aligned HTHs and TFBSs

Structure-based alignments of TFBSs are feasible. The results shown in Figure 5A and B suggest that TFR TFBSs of highly dissimilar nucleotide sequences can be aligned based on the structures of TFR-DNA complexes. The high structural similarity between the TFRs' HTH motifs allowed them to be aligned with each other easily. After applying the geometric transformations that align the HTH motifs to the DNA structures, 4-bp core DNA segments in half palindromes can be unambiguously aligned with each other (Figure 5A), the mutual root mean square deviations (RMSD) of DNA backbone atom positions being no more than 2.3 Å and mostly below 1.5 Å for the 16 complexes considered here. The TFBS sequences for these TFRs with experimentally determined complex structures can thus be aligned faithfully according to the structure-based alignments of the core DNA segments.

Sequence-based alignments of dissimilar TFBSs are not reliable. Despite the high structure similarity, the amino acid sequences of the HTH motifs of TFRs included for structural alignments have been kept low (the pairwise sequence identities between the HTH motifs <50% except for one pair (P0C093 and Q9KVD2, whose HTH motif sequence identity is 89%), so that the TFRs recognize TFBSs of distinguished nucleotide sequences. Because of the lack of sequence similarity, sequence-based alignments of the TFBSs of these TFRs generated erroneous results as judged by the structure-based alignments (Figure 5B).

Obtaining an extended and unified set of jointly aligned HTH and TFBS sequences. The structure-based alignment, while being accurate, can cover only a handful of TFRs. To obtain extended groups of jointly aligned TFRs and TFBSs, each of the 16 structurally aligned TFBSs have been used to seed iterative searches of TFBSs of similar sequences within the genome sequence-based prediction results that were of *P*-values $\leq$0.01. In the meantime, the amino acid sequences of the HTH motifs of the corresponding TFRs have also been aligned. This led to extended groups of jointed and simultaneously aligned sequences of HTHs and TFBSs. Finally, the groups generated from different seeding TFBSs were combined to obtain a single unified set, the alignment between the TFBS sequences in different groups bridged by the structure-based alignments between the seeding TFBSs. In total, the unified set contained 6932 unique HTH motifs and TFBSs. Supplementary Table S6 lists the sizes of the groups and the number of overlapping TFBSs between groups. Among the 16 groups, 11 contained >100 members, with varied number of between-group overlapping TFBSs. Supplementary Table S6 also gives the numbers of overlapping TFBSs for which their sequence alignments with the seeding TFBSs of the different groups were consistent with the structural-based alignments between the seeding TFBSs. The data indicate a significant overall fraction (86%) of consistent overlapping TFBSs, suggesting that a majority of entries in the extended unified set contained correctly aligned TFBSs. Another piece of evidence in support of the overall correctness of the unified joint alignment has come from direct coupling analysis (DCA) (47) between the amino acid residue types at positions of the HTH motif and the nucleotide types at positions of the core DNA segments. In Figure 5C, the value of direct information (DI) which indicates the extent of coupled residue type substitutions in the joint protein-DNA sequence set are plotted against the inter-residue spatial distances computed from structures of protein–DNA complexes. The results show that the distances between amino acid residue-nucleotide pairs of larger values of DI are all around or below 5 Å.

Performance of the statistical energy function learned from the unified set of HTH and TFBS sequences

Tests against the genome sequence-based predictions. The trained statistical energy function allows the prediction of TFBS nucleotide sequences from a given amino acid sequence of the HTH motif. We used two sets of data to test the accuracy of such predictions. The first set comprised 1055 randomly selected entries from the training data. The second set does not contain any entries from the training data. It comprised 1000 randomly selected TFRs with sequence identities between any two HTH motifs being <85% and TFBSs predicted by the genome sequence-based method being associated with *P*-values $\leq$0.001. In addition, the TFBSs were required to be longer than 12 bp and of GC contents between 25% and 75%. For each TFR in the tests sets, two ways of using the energy function to predict TFBS have been considered. The first is to find the lowest energy DNA octamers based only on the energy function. The second is to use the energy function to scan the TFR's GSS and identify sites of lowest *Z*-scores. The TFBSs predicted by the genome sequence-based method with small *P*-values served as references to judge the TFBS predicted using the statistical energy function. To consider *Z*-score-dependent accuracies, the prediction results for each test sets have been evenly divided into five groups associated with different *Z*-score ranges. Within each group, the distributions of the number of identical nucleotides between the TFBSs pre-

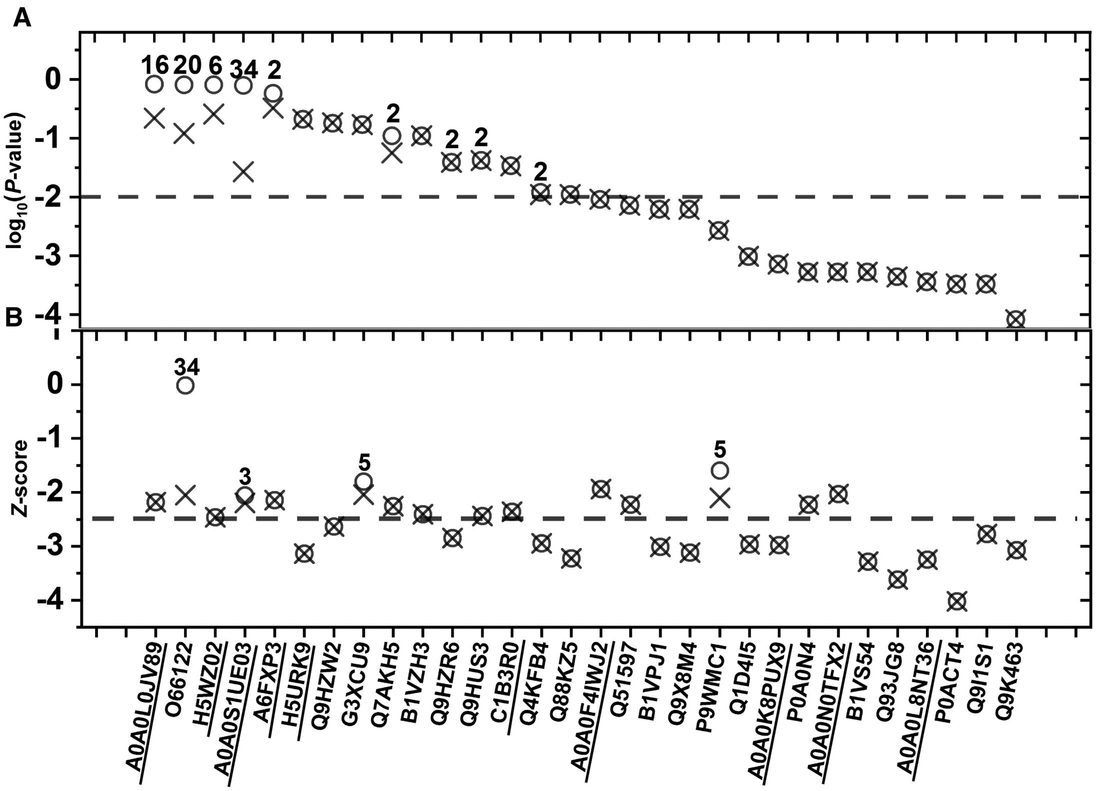


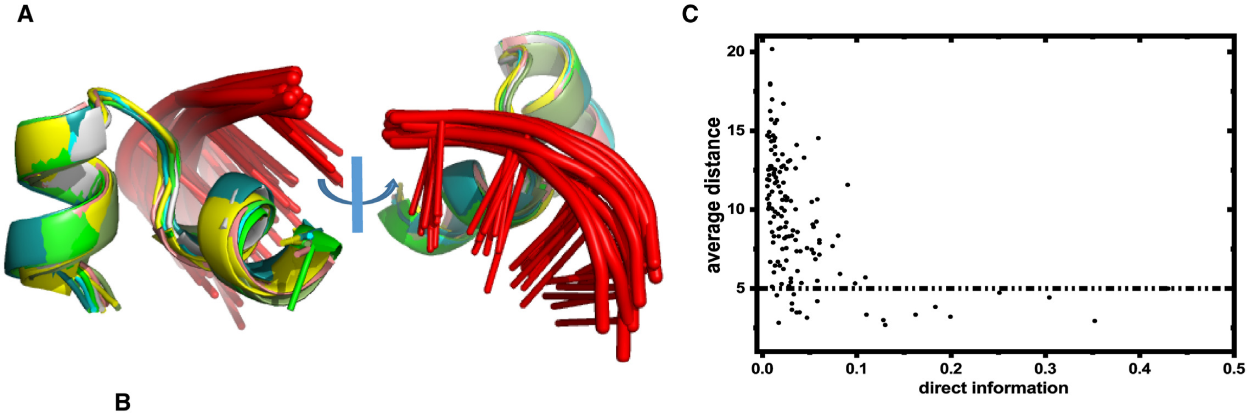
Figure 4. (**A**) The *P*-values of the top-tanking TFBSs (crosses) predicted by the genome sequence-based method and of the experimentally verified true TFBSs (circles) for 30 TFRs. The dashed horizontal line separates *P*-values above and below 0.01. (**B**) The *Z*-scores of the top ranking TFBSs predicted by the statistical energy method (crosses) and of the experimentally verified true TFBSs (circles). The dashed line separates *Z*-scores below and above –2.5. In both (A) and (B), the two types of symbols falling on top of each other indicates that the top-ranking prediction is a true TFBS. Along the horizontal axis, the TFRs are identified by their UniProt IDs, TFRs experimentally investigated in the current study underlined.

dicted by the statistical energy function methods and the TFBSs predicted by the genome sequence-based method for the same TFRs have been estimated.

The results for predictions based on the statistical energies of DNA octamers are given in Figure 6A (first test set) and Figure 6B (second test set). The agreements between the energy-based and the genome sequence-based predictions increase as the *Z*-scores decrease. For the first set containing training data, the fractions of predictions with no more than 2 non-identical nucleotides accounted for 41%, 52%, 66%, 63% and 75% of data for the five *Z*-score ranges between –∞ and 0 separated by values of –3.04, –3.21, –3.41 and –3.69, respectively, each range containing equal amount of data. For the second test set not containing any training data, the respective fractions are 21%, 37%, 39%, 45% and 63%. If for each TFR we consider any of the top 5 lowest energy non-redundant (sequences differ by two or more nucleotides) octamer sequences to be possible predictions, the respective fractions increased to 89%, 94%, 93%, 95% and 98% for the first test set and to 71%, 74%, 74%, 73% and 82% for the second test set. Thus, the prediction of DNA octamers without referring to any DNA sequences in the GSSs show reasonable success rates even on non-training data.

When the statistical energy function was applied to scan the GSS sequences, much higher success rates were achieved than considering all possible octamer sequences. For the first test set composed of training data, the fractions of predictions with >80% sequence identity accounted for 44%, 65%, 79%, 85% and 90% for respective *Z*-score ranges separated by values of –2.39, –2.71, –2.99 and –3.33. For the second test set containing no training data, the respective fractions are 31%, 48%, 72%, 84% and 86%. If any of the top 5 predictions are considered to be possible, the respective fractions for the first test data set are 63%, 90%, 97%, 98%, 99%, while those for the second test data set are 68%, 85%, 95%, 97% and 98%. Thus for a TFR in the lowest three *Z*-score ranges, the chance for the top 5 predictions to contain a presumably correct TFBS is above 95% (Supplementary Figure S3E shows the distribution of the total number of candidate TFBS sites per GSS segment, which varies around 52 between 40 to 65).

Ranks of the experimentally verified true TFBSs by the statistical energy function. Figure 4B shows the GSS *Z*-scores of the true TFBSs of the 30 TFRs and the *Z*-scores of the top ranking candidate TFBSs for the same TFRs. The *Z*-scores have been computed by using the statistical energy function to scan the respective GSSs. For 26 of the TFRs, the true TFBSs correspond to the lowest *Z*-score ones. For the remaining four TFRs, the true TFBS ranks are 3, 5, 5 and 34, respectively.



B

P0A0N4	-----TATAGACCGATCGATCGGTCTATA	TATAGACCGATCGATCGGTCTATA
P0ACT4	----------------CCTATCAATGATAGA------	---CCTATCAATGATAGA---------------
Q9AMH9	--------TTCGTACGGCGTACGAA--------	-TTCGTACGGCGTACGAA-----------
Q9KVD2	TTACGTGAGTACTCACGTAA------------	---TTACGTGAGTACTCACGTAA----

Figure 5. (**A**) Structure alignments of HTH motifs and 4-bp core DNA segments in protein-DNA complexes. Only the HTH motifs have been fitted for superimposing structures. For clarity, only 8 complexes (PDB IDs: 1jt0, 1pqi, 3zql, 4gct, 5dy0, 5k7z, 6c31, 4i6z) have been included. (**B**) Sequence-based and structure-based alignments between the nuleotide sequences of the TFBSs of 4 TFRs (PDB IDs: 1jt0, 1pqi, 3zql, 4gct). The first column gives the UniProt IDs of the TFRs, the second column gives the alignment maximizing the nuleotide type identity at aligned positions, and the third column gives the alignment generated from the structure alignment of the TFR-DNA complexes. The four bases colored in red cooresponding to the 4-bp core segments shown in (A). (**C**) The correlation between values of direct information computed from the aligned DBD and TFBS sequences and the inter-residue spatial distances computed from the structures of protein–DNA complexes. In this plot, each point corresponds to a pair of one amino acid position of the HTH motif and one nucleotide position of the octamer TFBS half site. The direct information values have been obtained through Directed Coupling Analysis of the unified set of aligned HTH and TFBS sequences. The spatial distances have been calculated as averages of distances between backbone atom pairs in four protein–DNA complexes (PDB IDs: 1jt0, 1pqi, 3zql, 4gct). If both the sequence alignment and the structural models are correct, high direct information between two residues would imply direct interactions (and thus short distances) between the residues in the 3D structure. The plot shows that all of the amino acid-nucleotide residue pairs associated with high direct information are in short spatial distances from each other.

GSS Z-scores indicate reliability of statistical energy function-derived TFBS motifs. In theory, the statistical energy function can be used not only to rank the candidate TFBS sites in the upstream GSS sequence profiles, but also to model sequence variations of the TFBS. In practice, whether a statistical energy-derived TFBS model (such as the sequence profile) for a particular TFR is reliable depends on the available data. More specifically, it depends on whether the training data have unbiasedly covered the respective regions in the joint sequence space of the DBD and the TFBS. The answer to this question is unknown *a priori*. On the other hand, the following reasoning suggests that the GSS *Z*-score (the *Z*-score of the lowest energy GSS sites among all candidate palindromic sites in the upstream GSS) can be used as an empirical indicator of the validity or reliability of the statistical energy function-derived TFBS model for a particular TFR. First, applying the statistical energy function to a TFR leads to two *Z*-scores. One is the octamer energy *Z*-score (or energy *Z*-score for short), which is the *Z*-score of the lowest energy 8-nucleotide core fragment among all possible 8-nucleotide fragments. The other is the GSS *Z*-score. The two *Z*-scores have different meanings. The energy *Z*-score reflects the extent to which the TFBS sequence is restrained by the statistical energy model. The lower the energy *Z*-score, the less variable or more restrained the TFBS sequences are. The GSS *Z*-score reflects how good the statistical energy restraints are satisfied by an actual upstream (candidate) TFBS sequence. While a low energy *Z*-score alone cannot guarantee that the TFBS model or profile produced by the statistical energy function is unbiased or reliable, a sufficiently low GSS *Z*-score unambiguously indicates that an upstream TFBS actually exists in consistence with the sequence model derived from the statistical energy function. This latter fact can be considered as strong evidence to support the applicability of the statistical energy function to the corresponding TFR. In other words, low GSS *Z*-score effectively indicates high reliability of the TFBS motif model derived from the statistical energy function.

The above reasoning is supported by results in Figure 3C, D and Supplementary Figure S6, in which a number of sequence profiles derived from the statistical energy are shown together with the experimentally verified TFBS sequences in GSSs. Among these Figures, Supplementary Figure S6F shows the results for TFR O66122, the only TFR (among the 30 TFRs) for which the predicted TFBS failed in experimentally tests. In general, the results indicate qualitative correlation between the GSS *Z*-scores and the agreements of the actual TFBS sequences to the statistical energy-derived sequence profiles.

Such a correlation in all TFRs is illustrated in Figure 7A, in which the similarity between the actual upstream TFBSs and the statistical energy-derived TFBS profiles is plotted against the GSS *Z*-score. The similarity has been measured as the number of identical nucleotide residues averaged over a small GSS *Z*-score range around a given *Z*-score value.

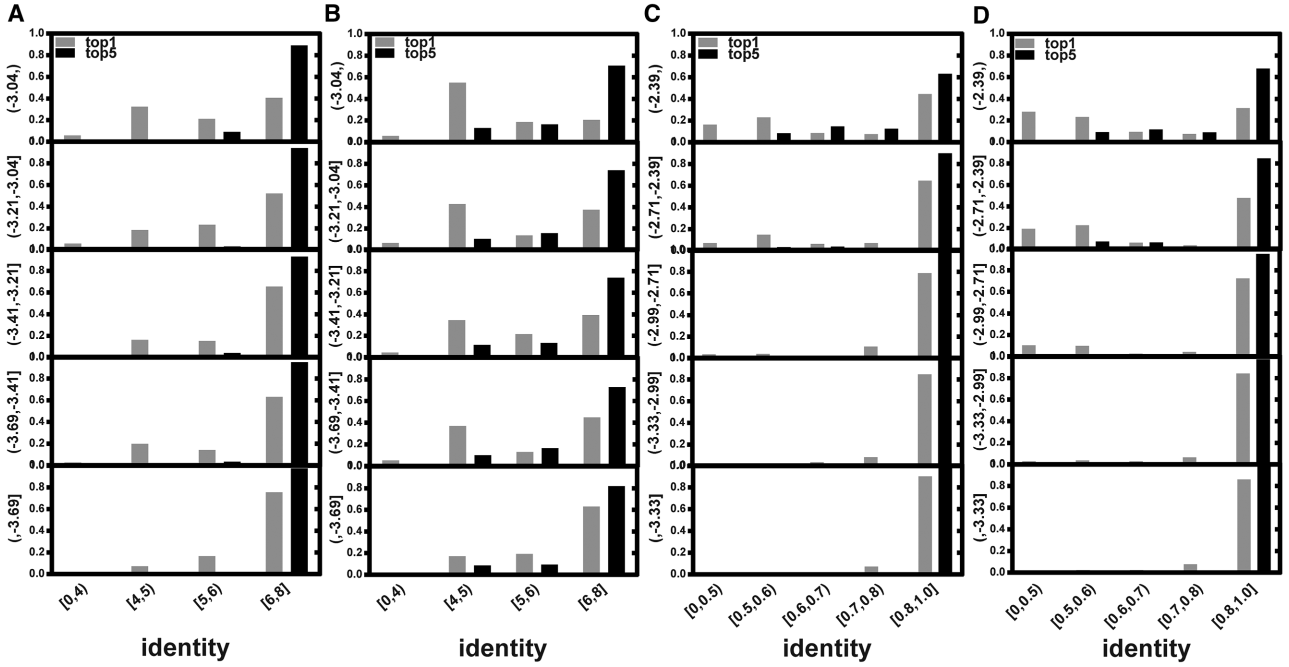


Figure 6. Agreements between the TFBSs predicted by the statistical energy function method and those by the genome sequence-based method. Two test sets of TFRs with TFBSs predicted by the genome sequence-based method with high significance have been considered. The first set of TFRs have been randomly selected from the training data for the statistical energy function (panels **A** and **C**) and the second set included non-training TFRs predicted by the genome sequence-based method with P-values ≤ 0.001 (panels **B** and **D**). Data in panels A and B correspond to predictions made based on energies computed for DNA octamers, and those in panels C and D correspond to predictions made by using the energy function to scan the genome sequence segments. The horizontal axes indicate the number (A and B) or fraction (C and D) of identical nucleotides between the TFBSs predicted by the statistical energy function and those predicted by the genome sequence-based method. The vertical axes indicate fractions of TFR entries. The test data have been evenly divided into five groups covering different Z-score ranges and the fractions within each group have been determined separately, shown in one subpanel for one Z-score range. The Z-score ranges are indicated on the right sides of the panels A and C, respectively.

At the GSS Z-score around –3.0, the average number of identical nucleotide residues is about 8 (or four identical nucleotide residues per half palindrome).

The Spec-Seq experiments provided data to verify the statistical energy-derived TFBS profiles for specific TFRs. In these experiments, probing DNA libraries have been designed by randomizing the experimentally verified upstream TFBS sequences at 4 (or occasionally 3) nucleotide positions, with the positions varied in different libraries to cover all positions of interest (see Figure 7 and Supplementary Figures S7 and S8 for the libraries used in the current work). If the chosen upstream TFBS sequence is indeed close to the consensus binding sequence of the corresponding TFR, different library members should be able to bind to the TFR in a conserved mode, allowing sequence variability at individual nucleotide positions to be mapped out by analyzing the sequencing results of bound and unbound fractions of the libraries. For one TFR(A0A0L8NT36), whose TFBS was predicted with a low GSS Z-score of –3.24, the Spec-Seq profile agrees excellently with both the statistical energy-derived motif and the upstream TFBS sequence (Figure 7B and Supplementary Figure S7A). For another TFR(A0A0F4IWJ2) associated with a GSS Z-score of –1.9, there are also obvious similarities between the Spec-Seq profiles, the statistical energy-derived motif, and the upstream TFBS sequence, the latter two exhibiting not outright but still substantial extent of similarity (see Figure 7C and Supplementary Figure S7B). For each of the remaining two TFRs (A0A0S1UE03 and A0A0N0TFX2) analyzed by Spec-Seq, standard analyses of the corresponding Spec-Seq data did not lead to consistent profiles between the forward and backward halves of the palindrome (Supplementary Figure S8), while neither of the experimental data-based forward and backward profiles agree with the upstream TFBS sequence or the statistical energy-derived TFBS motif (Supplementary Figure S8). This is probably because for these TFRs, the upstream TFBS sequences deviate too much from the (unknown) consensus TFBS sequences, leading the perturbed DNA sequences in the Spec-Seq libraries to not retain the same relative binding orientations and position shifts, prohibiting consistent and meaningful analysis of the Spec-Seq data. In summary, the Spec-Seq results in Figure 7 and Supplementary Figure S7 confirm our reasoning that the GSS Z-score (or the agreement between the upstream TFBS and the statistical energy-derived TFBS profile) can effectively indicate the applicability of the statistical energy function for TFBS motif prediction.

The above proposition given, Figure 7D shows the Z-score distribution of all TFRs on a 2D plane spanned by the energy Z-score and the GSS Z-score. From this distribution, we can estimate that >10 000 TFRs are associated with energy Z-scores < –3.5 and GSS Z-scores < –3.0 (there are 1 980 TFRs if we require the energy Z-score < –4.0), suggesting that for these TFRs, the statistical energy function may produce meaningful TFBS sequence models. In

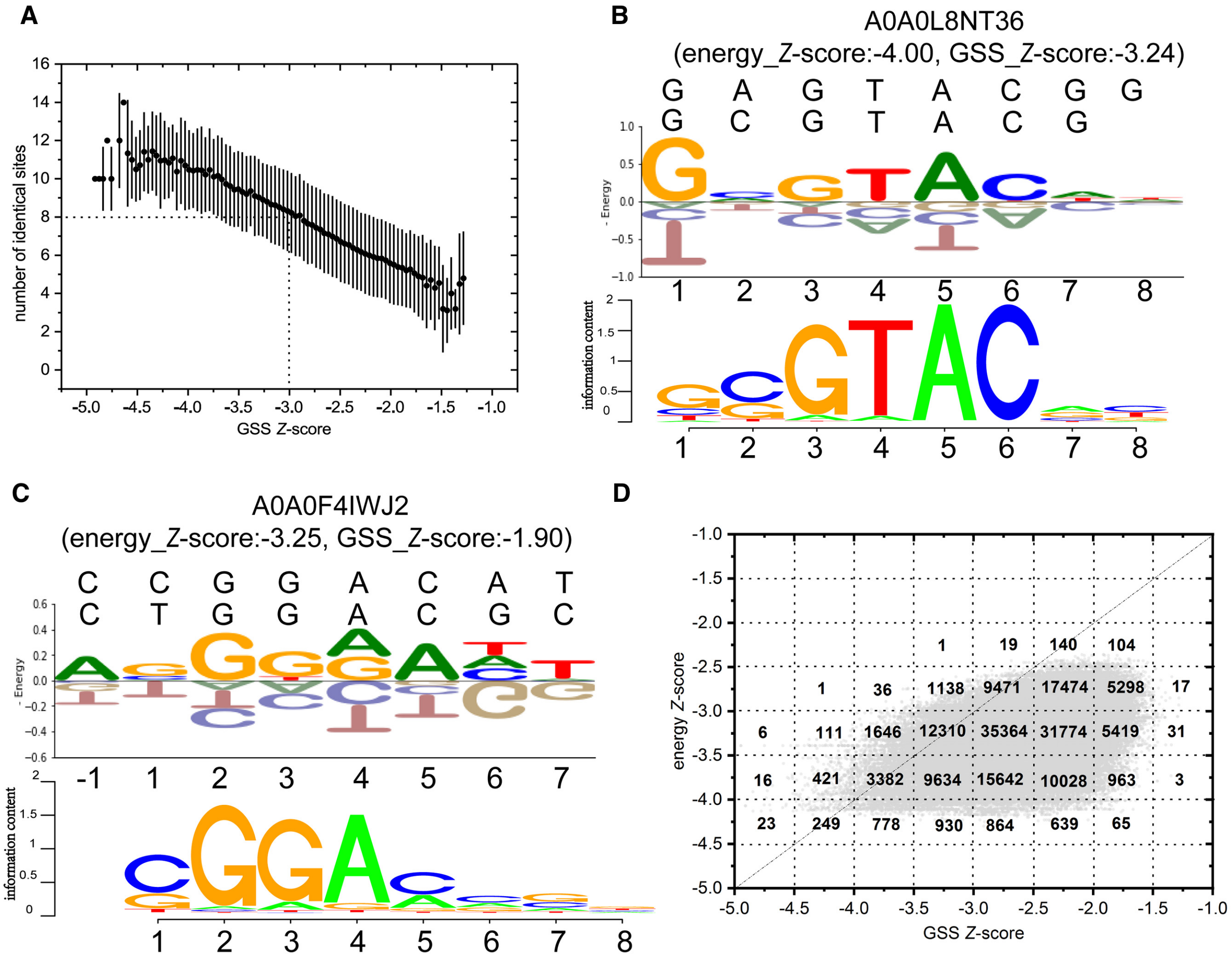


Figure 7. (**A**) The correlation of GSS Z-score and the similarity between the actual upstream TFBS and the statistical energy-derived TFBS motif. The similarity has been measured as the number of identical nucleotide residues averaged over a small GSS Z-score range (±0.04) around a given Z-score value. In (**B**) and (**C**), the TFBS sequence patterns derived from the Spec-Seq data (upper panels) are compared with statistical energy function-derived sequence motifs (lower panels) for the TFRS A0A0L8NT36 and A0A0F4IWJ2, respectively. The energy logos from the Spec-Seq data have been computed by enforcing symmetry of the forward and backward half palindromes (described as the second approach in Supplementary Method). The nucleotide positions in the core octamer are numbered from 1 to 8 to be consistent with results in Supplementary Figures S7 and S8. (**D**) A scattering plot showing the distribution of all TFRs on a 2D plane spanned by the energy Z-score and the GSS Z-score. Non-empty cells are labelled by the numbers of TFRs that fall into corresponding cells.

applications investigating specific TFRs, such a model may be applied to scan genome sequences to propose new TFBSs located far from the TFR genes. As examples, Supplementary Table S7 shows likely TFBSs discovered at different genome locations for three TFRs associated with low Z-scores.

Combining predictions based on genome-sequences and on statistical energies. Results in Figure 4A and B suggested that for seven TFRs for which the genome sequence-based method did not produce reliable predictions, the statistical energy method still predicted the true TFBSs as of the lowest Z-scores. These include the TFRs A0A0L0JV89 (P-value = 0.22), H5WZ02 (P-value = 0.26), A6FXP3 (P-value = 0.32), Q7AKH5 (P-value = 0.06), Q9HZR6 (P-value = 0.04), Q9HUS3 (P-value = 0.04) and Q4BFK4 (P-value = 0.011). Only for two TFRs, G3XCU9 (P-value = 0.17) and P9WMC1(P-value = 0.003), the genome sequence-based method but not the statistical energy function predicted the true TFBSs as the top results. The scattering plot of the smallest P-values versus the lowest Z-scores for 12 000 TFRs is shown in Figure 8A, which indicates that the two indicators are to a very large extent uncorrelated. These results suggest that the energy function generated by statistical learning significantly complements the genome sequence-based approach. For a particular TFR, more robust final results may be obtained by considering the top ranking candidates identified by the two methods simultaneously, the best predictions according to one method cross-checked by considering the significance indicators of the other method. As examples, Figure 8 shows scattering plots of P-values versus Z-scores for top predictions by both methods for 5 TFRs. For 4 of them, the true TFBSs are included in the top 10 results of both methods. For two TFRs, the top predictions of both methods agree with each other and they turn out to be the true TFBSs (Figure 8B and C).

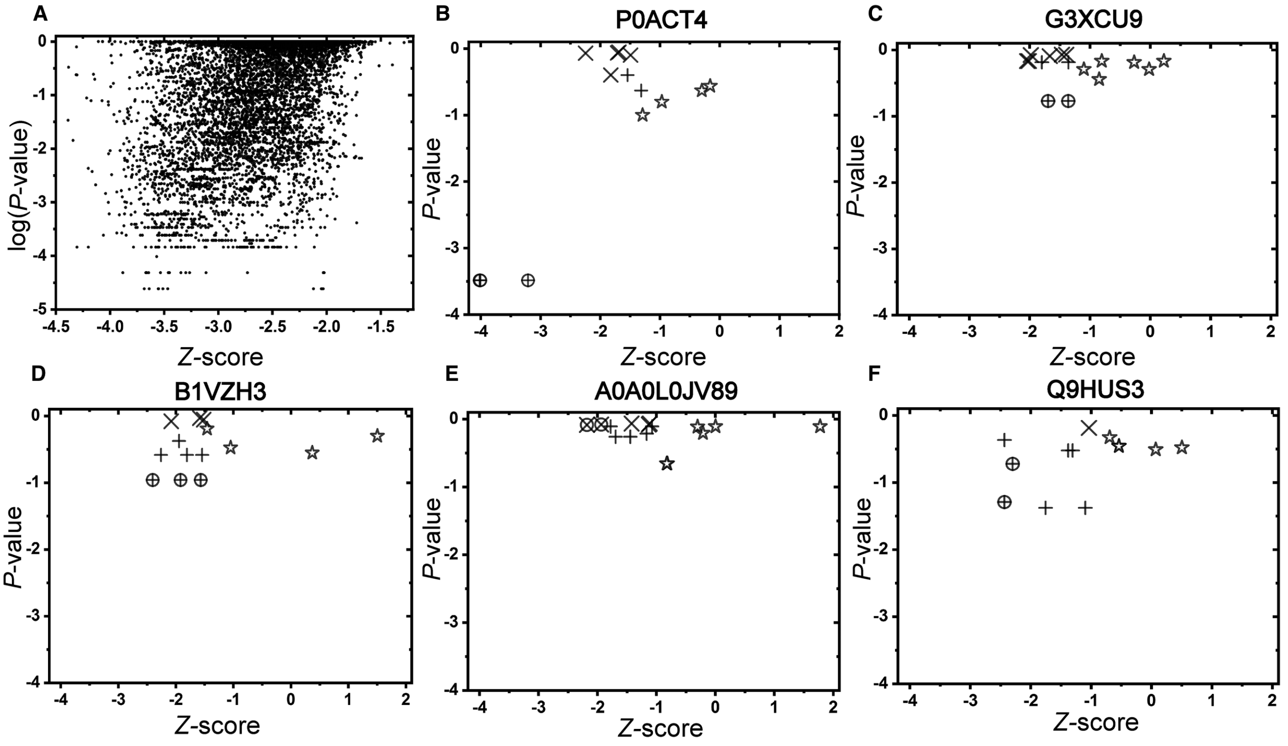


Figure 8. (**A**) The scattering plot of best Z-score (horizontal axis) and log_{10} (P-value) (vertical axis) of prediction results for 12 000 randomly selected TFRs. (**B**)-(**F**) The scattering plots of Z-score (horizontal axis) and log_{10} (P-value) (vertical axis) of top-ranking predictions for five TFRs. The UniProt IDs of the TFRs are indicated above respective plots. The tilted crosses (◊) represent TFBSs only in the top 10 of the statistical energy function predictions, stars (☆) represent TFBSs only in the top 10 of the genome sequence-based predictions, and the crosses (+) represent TFBS in the top 10 sets of both methods. Points corresponding to experimentally verified true TFBSs are enclosed by open cycles (◯).

For the TFRs Q9HUS3 and G3XCU9, the true TFBSs are of either the lowest Z-scores or the lowest P-values, all in the top 10 results of both methods (Figure 8D and E). For the TFR A0A0L0JV89, no prediction of significant P-values was made and the true TFBSs are those of the lowest Z-scores (Figure 8F). Overall, statistical learning significantly extended the range of TFRs for which reliable TFBSs can be predicted. According to the results in Figure 6D, which shows that for the Z-score range between –2.7 and –2.4, the fractions of successful predictions on non-training data are 86% considering only the top 1 candidate and 98% considering the top 5 candidates, we may choose a threshold Z-score of –2.5 to separate between 'reliable' and 'unreliable' predictions. Then the statistical energy-based predictions for 48.3% of all the TFRs in the TFR dataset can be considered as reliable. If this set is combined with those satisfying the P-value$\leq$0.01 condition in genome sequence-based predictions, the overall coverage increases to 59.6%, which more than doubled the coverage of the genome sequence-based method alone.

CONCLUSIONS

In summary, we have developed an approach combining genome sequence data and protein-DNA complex structure data to systematically predict DNA binding sites of transcription factors of TetR family. First, we have developed a computational workflow to streamline predictions based on genome sequences, quantitative P-values proposed to represent statistical significance of results. The definition of a quantitative P-value allowed predicted candidates to be ranked and reliability of results assessed. This has allowed us to filter out those highly reliable prediction results to construct a large training set of TFBSs and TFRs covering diverse sequence space for subsequent statistical learning. By incorporating structural information from a handful of protein-DNA complexes of diverged sequences, we were able to construct a unified set of more than 6 000 aligned TFBS sequences each jointed with the amino acid sequence of its respective recognition protein motif. This large set enabled the training of a statistical energy function, which represents the joint distribution of the amino acid sequences of TFRs (more specifically, the TFRs' DNA-recognizing HTH motifs) and the nucleotide sequences of the TFBSs. Benchmarking against experimental results have validated the use of P-values in the genome sequence-based method and of Z-scores in the statistical energy-based method to rank candidates and to indicate reliability of predictions.

Combining the genome sequence-based and the statistical energy-based prediction results leads to more robust predictions on individual TFRs. The joint application of both methods more than doubles the number of TFRs with predictable TFBSs. Besides ranking given TFBSs, the statistical energy function may also be applied to model the variable TFBS sequences of given TFRs. The applicability can be judged empirically by the GSS Z-score. If considered applicable, the statistical energy can be employed to scan entire genome sequences to discover new TFBS. Besides TFBS discovery, data shown in Figure 6A and B suggest that the statistical energy function may also be applied to design new core DNA octamer motifs for TFRs of given amino acid sequences, although the success rates may be lower, but can still be acceptable (63% for top one design and 82% in top 5 designs) for sufficiently low energy *Z*-score (-3.7) results (Figure 6B). In the future, the statistical energy function can be further improved when more structures of TFR-DNA complexes become available, allowing the training data to be extended to cover wider regions in the sequence space.

Compared with most previous computational approaches to TFBS prediction, our approach integrated genome sequence and structural data to produce predictions for as many TFRs as possible. As a result, the TFR-TFBS interactions predictable by our method significantly extend currently known specific interactions between TFs and DNA. We have used each of the 1 000 TFRs contained in the second test set constructed from the genome sequence-based predictions to query the FootprintDB database (50). For only 84 (371) TFRs, results of *E*-values smaller than $10^{-10}(10^{-5})$ were returned. Among the returned TFBS DNA sequences, only 11 (16) exhibited sequence identities above 80% with the genome sequence method-predicted top results (P-value ≤ 0.001). Thus the newly predicted TFR-TFBS interactions can be applied to improve the construction of gene regulatory networks. To facilitate this, a public webserver has been provided for others to retrieve our prediction results or to make predictions on new TFRs with our models (http://biocomp.ustc.edu.cn/servers/tfbs-predict.php). For interested users, a downloadable package provided at the same address contains source codes with installation guide and input/output examples, intermediate data for predictions on the experimentally verified TFR sets, and a collection of predicted TFBSs of significant *P*-values or *Z*-scores.

Our study on the TetR family can be considered as an example that shows the substantial benefits of integrating sequence and structural data in data-driven models predicting specific protein-DNA interactions. This framework of data integration can be extended to other prokaryotic TF families for which sufficient structural data are available. A preliminary survey of TF-DNA complex structures in PDB suggested that for 13 of the 15 TF families considered to be the most important prokaryotic regulators by Ramos *et al*. (51), more than one protein-DNA complex structures of different members of the same families have been reported. Besides the TetR family, the other families with more than 5 non-redundant protein-DNA complex structures are the MarR, GntR, Crp, DeoR, LysR, LacI and MerR families, which respectively have structures available for 14, 11, 10, 9, 9, 7 and 6 family members. Just like the TFRs, the structures of the DNA binding domains and the bound DNA fragments of different members of the same family can be well-aligned simultaneously. Thus the structure alignment-based approach may be extended to these other TF families, especially those with more structural data, with the caveat that family-specific re-parameterization of the method may be needed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr. Banghui Liu and Chen Chen for help with the experiments.

FUNDING

National Key R&D Program of China [2018YFA0900703 to H.L., 2018YFA0901600 to Q.C.]; National Natural Science Foundation of China [31971175, U1732156 to Q.C.]; Youth Innovation Promotion Association, Chinese Academy of Sciences [2017494 to Q.C.]. Funding for open access charge: University of Science and Technology of China.
Conflict of interest statement. None declared.

REFERENCES

1. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.
2. Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res.*, **34**, D74–D81.
3. Vaquerizas,J.M., Teichmann,S.A. and Luscombe,N.M. (2012) How do you find transcription factors? Computational approaches to compile and annotate repertoires of regulators for any genome. *Methods Mol. Biol. (Clifton, N.J.)*, **786**, 3–19.
4. Banf,M. and Rhee,S.Y. (2017) Computational inference of gene regulatory networks: approaches, limitations and opportunities. *Biochim. Biophys. Acta*, **1860**, 41–52.
5. Mercatelli,D., Scalambra,L., Triboli,L., Ray,F. and Giorgi,F.M. (2020) Gene regulatory network inference resources: a practical overview. *Biochim. Biophys. Acta*, **1863**, 194430–194451.
6. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet*, **11**, 751–760.
7. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegromontero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I. and Cook,K.B. (2014) Determination and Inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
8. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B. and Bussemaker,H.J. (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
9. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J. and Sillanpää,M.J. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
10. Meng,X. and Wolfe,S.A. (2006) Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat. Protoc.*, **1**, 30–45.
11. Stormo,G.D., Zuo,Z. and Chang,Y.K. (2014) Spec-seq: determining protein-DNA-binding specificity by sequencing. *Brief. Funct. Genomics*, **14**, 30–38.
12. Zuo,Z., Chang,Y. and Stormo,G.D. (2015) A quantitative understanding of lac repressor's binding specificity and flexibility. *Quant. Biol.*, **3**, 69–80.
13. Zuo,Z. and Stormo,G.D. (2014) High-Resolution specificity from DNA sequencing highlights alternative modes of lac repressor binding. *Genetics*, **198**, 1329–1343.
14. McCue,L.A., Thompson,W., Carmack,C.S., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
15. Liu,B., Zhang,H., Zhou,C., Li,G., Fennell,A., Wang,G., Kang,Y., Liu,Q. and Ma,Q. (2016) An integrative and applicable phylogenetic

footprinting framework for *cis*-regulatory motifs identification in prokaryotic genomes. *BMC Genomics*, **17**, 578–590.

16. Katara,P., Grover,A. and Sharma,V. (2012) Phylogenetic footprinting: a boost for microbial regulatory genomics. *Protoplasma*, **249**, 901–907.
17. Laing,E., Sidhu,K. and Hubbard,S.J. (2008) Predicted transcription factor binding sites as predictors of operons in *Escherichia coli* and *Streptomyces coelicolor*. *BMC Genomics*, **9**, 79–84.
18. Pavesi,G., Mereghetti,P., Mauri,G. and Pesole,G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.*, **32**, W199–W203.
19. Li,G., Liu,B., Ma,Q. and Xu,Y. (2011) A new framework for identifying *cis*-regulatory motifs in prokaryotes. *Nucleic Acids Res.*, **39**, e42.
20. Li,G., Liu,B. and Xu,Y. (2010) Accurate recognition of *cis*-regulatory motifs with the correct lengths in prokaryotic genomes. *Nucleic Acids Res.*, **38**, e12.
21. Yan,B., Methe,B.A., Lovley,D.R. and Krushkal,J. (2004) Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family *Geobacteraceae*. *J. Theor. Biol.*, **230**, 133–144.
22. Francke,C., Kerkhoven,R., Wels,M. and Siezen,R.J. (2008) A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. *BMC Genomics*, **9**, 145–164.
23. Yan,B., Lovley,D.R. and Krushkal,J. (2007) Genome-wide similarity search for transcription factors and their binding sites in a metal-reducing prokaryote *Geobacter sulfurreducens*. *Biosystems*, **90**, 421–441.
24. Yu,Z., Reichheld,S.E., Savchenko,A., Parkinson,J. and Davidson,A.R. (2010) A comprehensive analysis of structural and sequence conservation in the TetR family transcriptional regulators. *J. Mol. Biol.*, **400**, 847–864.
25. Wels,M., Francke,C., Kerkhoven,R., Kleerebezem,M. and Siezen,R.J. (2006) Predicting *cis*-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res.*, **34**, 1947–1958.
26. Liu,B., Yang,J., Li,Y., McDermaid,A. and Ma,Q. (2018) An algorithmic perspective of de novo *cis*-regulatory motif finding based on ChIP-seq data. *Brief. Bioinform.*, **19**, 1069–1081.
27. Zambelli,F., Pesole,G. and Pavesi,G. (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res.*, **37**, W247–W252.
28. Liu,L.A. and Bradley,P. (2012) Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Curr. Opin. Struct. Biol.*, **22**, 397–405.
29. Liu,Z., Guo,J.T., Li,T. and Xu,Y. (2008) Structure-based prediction of transcription factor binding sites using a protein-DNA docking approach. *Proteins*, **72**, 1114–1124.
30. Siggers,T.W. and Honig,B. (2007) Structure-based prediction of C_2H_2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.*, **35**, 1085–1097.
31. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
32. Pujato,M., Kieken,F., Skiles,A.A., Tapinos,N. and Fiser,A. (2014) Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic Acids Res.*, **42**, 13500–13512.
33. Alibes,A., Nadra,A.D., De Masi,F., Bulyk,M.L., Serrano,L. and Stricher,F. (2010) Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. *Nucleic Acids Res.*, **38**, 7422–7431.
34. Pareja,E., Pareja-Tobes,P., Manrique,M., Pareja-Tobes,E., Bonal,J. and Tobes,R. (2006) ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms. *BMC Microbiol.*, **6**, 29–39.
35. Cuthbertson,L. and Nodwell,J.R. (2013) The TetR family of regulators. *Microbiol. Mol. Biol. Rev.*, **77**, 440–475.
36. Croxatto,A., Chalker,V.J., Lauritz,J., Jass,J., Hardman,A., Williams,P., Camara,M. and Milton,D.L. (2002) VanT, a homologue of *Vibrio harveyi* LuxR, regulates serine, metalloprotease, pigment, and biofilm production in *Vibrio anguillarum*. *J. Bacteriol.*, **184**, 1617–1629.
37. Maity,T.S., Close,D.W., Valdez,Y.E., Nowaklovato,K.L., Martiarbona,R., Nguyen,T.T., Unkefer,P.J., Honggeller,E., Bradbury,A. and Dunbar,J. (2012) Discovery of DNA operators for TetR and MarR family transcription factors from *Burkholderia xenovorans*. *Microbiology*, **158**, 571–582.
38. Persikov,A.V. and Singh,M. (2013) De novo prediction of DNA-binding specificities for Cys_2His_2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 97–108.
39. Persikov,A.V., Singh,M. and Osada,R. (2008) Predicting DNA recognition by Cys_2His_2 zinc finger proteins. *Bioinformatics*, **25**, 22–29.
40. Khamis,A.M., Motwalli,O., Oliva,R., Jankovic,B.R., Medvedeva,Y.A., Ashoor,H., Essack,M., Gao,X. and Bajic,V.B. (2018) A novel method for improved accuracy of transcription factor binding site prediction. *Nucleic Acids Res.*, **46**, e72.
41. Li,Y., Chen,C., Kaye,A.M. and Wasserman,W.W. (2015) The identification of *cis*-regulatory elements: a review from a machine learning perspective. *Biosystems*, **138**, 6–17.
42. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–840.
43. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
44. Ahn,S.K., Cuthbertson,L. and Nodwell,J.R. (2012) Genome context as a predictive tool for identifying regulatory targets of the TetR family transcriptional regulators. *PLoS One*, **7**, e50562
45. Pearson,C.E., Zorbas,H., Price,G.B. and Zannis-Hadjopoulos,M. (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J. Cell. Biochem.*, 63, 1–22.
46. Burley,S.K., Berman,H.M., Bhikadiya,C., Bi,C., Chen,L., Di Costanzo,L., Christie,C., Dalenberg,K., Duarte,J.M. and Dutta,S. (2018) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
47. Morcos,F., Pagnani,A., Lunt,B., Bertolino,A., Marks,D.S., Sander,C., Zecchina,R., Onuchic,J.N., Hwa,T. and Weigt,M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, E1293–E1301.
48. Hellman,L.M. and Fried,M. (2007) Electrophoretic Mobility Shift Assay (EMSA) for DETECTING Protein-Nucleic acid interactions. *Nat. Protoc.*, **2**, 1849–1861.
49. Carey,M., Peterson,C.L. and Smale,S.T. (2013) DNase I Footprinting. *CSH Protoc.*, **2013**, 469-471.
50. Sebastian,A. and Contrerasmoreira,B. (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, **30**, 258–265.
51. Ramos,J.L., Martínez-Bueno,M., Molina-Henares,A.J., Terán,W., Watanabe,K., Zhang,X., Gallegos,M.T., Brennan,R. and Tobes,R. (2005) The TetR family of transcriptional repressors. *Microbiol. Mol. Biol. Rev.*, **69**, 326–356.