






Predicting bacterial transcription factor binding sites through machine learning and structural characterization based on DNA duplex stability

André Borges Farias ^{1,2,*}, Gustavo Sganzerla Martinez ³, Edgardo Galán-Vásquez ⁴, Marisa Fabiana Nicolás ¹, Ernesto Pérez-Rueda ^{2,*}

¹Laboratório Nacional de Computação Científica - LNCC, Avenida Getúlio Vargas, Petrópolis, Rio de Janeiro 25651075, Brazil

²Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica del Estado de Yucatán, Carretera Sierra Papacal, Mérida 97302, Yucatán, México

³Microbiology and Immunology, Dalhousie University, 5850 College Street, Halifax B3H 4H7, Nova Scotia, Canada

⁴Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, Circuito Escolar S/N, Mexico City 01000, México

*Corresponding authors. Ernesto Pérez-Rueda. E-mail: ernesto.perez@iimas.unam.mx; André Borges Farias. E-mail: bfarias.andre@gmail.com

Abstract

Transcriptional factors (TFs) in bacteria play a crucial role in gene regulation by binding to specific DNA sequences, thereby assisting in the activation or repression of genes. Despite their central role, deciphering shape recognition of bacterial TFs-DNA interactions remains an intricate challenge. A deeper understanding of DNA secondary structures could greatly enhance our knowledge of how TFs recognize and interact with DNA, thereby elucidating their biological function. In this study, we employed machine learning algorithms to predict transcription factor binding sites (TFBS) and classify them as directed-repeat (DR) or inverted-repeat (IR). To accomplish this, we divided the set of TFBS nucleotide sequences by size, ranging from 8 to 20 base pairs, and converted them into thermodynamic data known as DNA duplex stability (DDS). Our results demonstrate that the Random Forest algorithm accurately predicts TFBS with an average accuracy of over 82% and effectively distinguishes between IR and DR with an accuracy of 89%. Interestingly, upon converting the base pairs of several TFBS-IR into DDS values, we observed a symmetric profile typical of the palindromic structure associated with these architectures. This study presents a novel TFBS prediction model based on a DDS characteristic that may indicate how respective proteins interact with base pairs, thus providing insights into molecular mechanisms underlying bacterial TFs-DNA interaction.

Keywords: machine learning; transcription factor binding site; DNA duplex stability

Introduction

Transcription factors (TFs) constitute a class of critical proteins regulating many biological events. By discerning specific DNA-binding sites within the cellular milieu, TFs wield control over the gene expression patterns within organisms [1], impacting many intrinsic and extrinsic cellular processes [2, 3]. The DNA-binding sites of TFs are commonly represented as motifs through position weight matrices. Determining and characterizing how TFs recognize these motifs are crucial for understanding the regulatory functions of this class of proteins [4]. In bacteria, the most common DNA-binding structure identified so far is the helix-turn-helix (HTH) [5], allowing them to recognize DNA binding sites characterized by an inverted repeat (IR) architecture, resulting in palindromic nucleotide sequences. This arrangement, exemplified in *IcIR*, *TetR*, and *LacI*, typically involves TFs forming dimers or tetramers [6, 7]. These protein complexes interact with DNA using different HTH subunits, leading to a head-to-head configuration. In contrast, binding to sites with direct repeats (DR) architecture requires TFs to adopt a head-to-tail configuration

[8], as described for *BldC* [9], *Xis* [10], and *Atox1* [11], among others.

It is well-established that TFs recognize specific sequences upstream the transcription start site to regulate (activate or repress) gene expression [12]. However, the situation is further complicated by the ability of proteins to bind to DNA through various modes [13]. Understanding the role of DNA secondary structures can significantly enhance our knowledge about how this class of proteins recognizes and interacts with DNA, thereby elucidating their biological function. Techniques like one-hot encoding (OHE), where each unique character is assigned a distinct numerical ID, have been utilized in artificial intelligence models [14]. However, the numerical variables generated using these methods often lack direct biological significance, resulting in the limited interpretation of their biological role.

In the last decade, several databases have been developed to provide helpful information about DNA-binding motifs and genomic binding sites, such as *RegulonDB* [15], *JASPAR* [16], *CollecTF* [17], *TRANSFAC* [18], and *UniPROBE* [19]. Simultaneously,

Received: July 17, 2024. Revised: October 2, 2024. Accepted: November 1, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact journals.permissions@oup.com

diverse approaches have been developed and implemented for the prediction and analysis of transcription factor binding sites (TFBS), focusing on methods based on artificial intelligence [20]. In this regard, FCNsignal [21] employs fully convolutional neural networks (FCN) for predicting TF binding signals; whereas DeepBind [22], DESSO [23], and DeepSea [24] utilize convolutional neural networks (CNN). Additionally, DanQ [25] combines CNN with recurrent neural networks (RNN) to predict TFBS. While various methods based on deep learning have shown satisfactory results in predicting TFBS [26], models rooted in traditional machine learning (ML) approaches remain limited [27]. Implementing ML models enables accurate prediction and classification of TFBS and provides interpretability. The capacity to elucidate the outcomes of a model offers valuable insights into the structural attributes of protein-DNA binding regions, facilitating the delineation of the model's decision patterns. In essence, this entails comprehending the rationale behind its predictions or classifications derived from the input data, thus providing information about the essential characteristics the model is considering.

Considering the central role that TFs play in regulatory mechanisms, this study aims to utilize ML methodologies to develop a predictive model for identifying TF binding sites. Specifically, the model aims to differentiate between true TFs and non-TFs based on structural and thermodynamic parameters in order to identify the precise locations where regulatory proteins interact to positively or negatively modulate gene expression.

Additionally, we aim to investigate the structural variances within these regions, facilitating the development of a secondary ML model to differentiate whether a given region exhibits a direct or IR architecture. Furthermore, the robustness of our predictions is substantiated by structural model, highlighting the capability of our ML approach not only in identifying and classifying TFBSs but also points to a correlation between the features used to train the model with the specific mode of interaction employed by the respective TF.

Materials and methods

Dataset preparation and general procedure

This study acquired a dataset of 3098 TFBS sequences, which are recognized by 159 TFs from the bacterium *Escherichia coli* K12 from the RegulonDB database [15]. Furthermore, an additional set of 5452 bacterial TFBS sequences was obtained from the CollecTF database [17]. To ensure the integrity of the dataset, redundant sequences were identified and removed, totaling 559 and 3054 duplicate TFBS sequences from RegulonDB and CollecTF, respectively. Detailed information on sequence distribution by TF is provided in Supporting Information Tables S1 and S2.

Random sequences were generated for each TFBS to construct a negative dataset of sequences that did not represent TFBS while maintaining consistent length sizes and nucleotide composition. Given that a single TF can recognize multiple binding sites [28], resulting in TFBS of varying sizes, we employed the MEME program [29] for sequence alignment to ensure length uniformity. Subsequently, graphical representations of the sequences were generated using the WebLogo server [30].

Converting TFBS sequences into numerical features

The TFBS sequences, represented by nucleotides adenine (A), thymine (T), cytosine (C), and guanine (G), have been converted into numerical information using several parameters. These

parameters can be classified as either structural or thermodynamic. Structural parameters include twist [31], bend [31], major groove width [31], major groove depth [31], major groove size [32], and persistence length [33]. On the other hand, thermodynamic parameters encompass enthalpy [34], entropy [34], free energy [35], stacking energy [36], and DNA duplex stability (DDS [37]). In this study, we emphasize the DDS metric. AT nucleotides are bound together by two hydrogen bonds while GC consists of three. This physicochemical difference results in distinct free-energy profiles, and has been employed as good descriptors of interaction between proteins and DNA. This is the case of promoter sequences, which interact with RNA polymerase to initiate transcription. Promoter sequences have been described as per their levels of DDS, being distinguishable from other genomic regions across the three domains of life [38–41]. For each TFBS sequence, a DDS value was assigned to every dinucleotide in the sequences. For instance, AA, CG, and TC were converted to -1.00, -2.17, and -1.30, respectively. In addition, we display the code for converting all 16 possible combinations of dinucleotide pairs in the DDS values (see Table S3 in Supporting Information).

Machine learning models

A comprehensive screening for the classification task (TFBS or non-TFBS) was conducted employing various classifiers: Random Forest (RF), stochastic gradient descent (SGD), support vector classifier (SVC), linear support vector classifier (LinearSVC), AdaBoost, Gradient Boosting, XGBoost and Decision Tree (DT) classifiers. All algorithms, except for XGBoost, were implemented using scikit-learn version 0.24.2 [42]. To ensure consistency, we maintained the hyperparameters at their default values across all datasets. Each model underwent validation through an 80-20 train-test split under a 10-fold cross validation process to ensure every data point is covered both in the train and test steps. The models were evaluated based on their accuracy (ACC), precision (Prec), recall (Rec), F1-score (F1), and area under the ROC curve (AUC). Finally, the best performing algorithm, i.e. highest accuracy, precision, recall, F-1 score, and AUC, was selected.

In addition, a second model was developed to predict the category of TFBS, distinguishing between DR and IR sequences. However, due to the substantial disparity in data volume between both classes, creating an imbalanced dataset, we split the classification task into two subtasks: DR versus non-DR, and IR versus non-IR. Among the dataset of 3204 sequences, only 316 were categorized as TFBS-DR, while 738 sequences from the TFBS-IR class were selected. The negative dataset for these two subtasks were random sequences. We then trained a model to distinguish between TFBS-DR and TFBS-IR sequences. For this purpose, we utilized 316 TFBS-DR sequences and 738 TFBS-IR sequences. We also utilized SHapley Additive exPlanations (SHAP v. 0.42.1) [43] to map the decision patterns of our models. This approach facilitated the identification of nucleotide positions that significantly contribute to differentiating a TFBS-DR from a TFBS-IR.

The methodology developed in this work is illustrated in Fig. 1, depicting the sequential steps undertaken to achieve our research objectives. The dataset containing the train/test data and sequences used in this work is publicly available at <https://github.com/farias-ab/TFBS-Prediction.git>.

Results and discussion

TFBS are well represented by DDS

Initially, we carried out a study to evaluate the impact of converting TFBS nucleotide sequences into numerical values

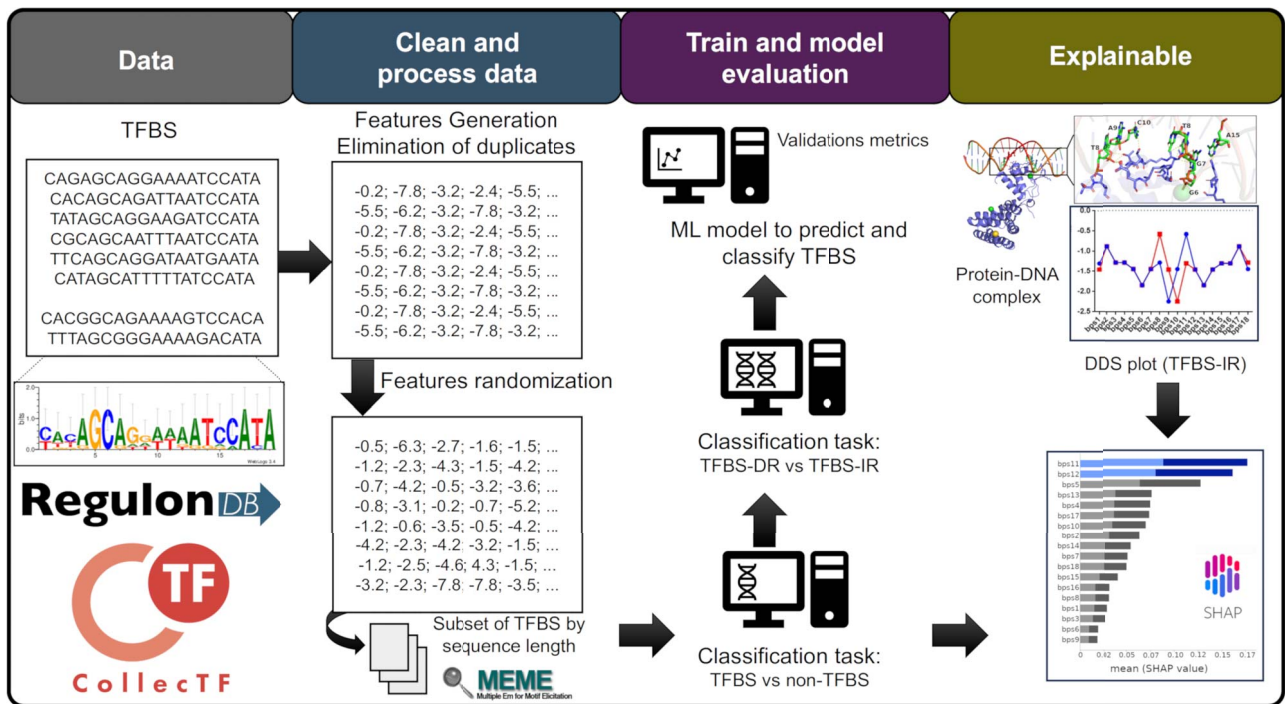


Figure 1. Schematic illustrating the workflow for acquiring, training, validating, and interpreting the predictive model designed to TFBS.

related to thermodynamic or structural data. In this preliminary assessment, we focused on sequences of *Escherichia coli* K12 obtained from the RegulonDB database. Thus, we partitioned our dataset based on the length of nucleotide base pairs (bp) to ensure uniform dimensions across the data. More than 92% of the dataset consists of TFBS with size ranging between 8 and 20 bp (see [Supplementary Fig. S1](#)). After partitioning our dataset, TFBS sequences spanning 12 to 15 bp were converted into both structural and thermodynamic variables for training the ML model, employing the RF algorithm. Validation metrics for all parameters utilized are depicted in [Fig. 2](#).

We observe that thermodynamic features, particularly DDS and enthalpy, appear to more accurately represent TFBS, with average accuracy values of 0.75 and 0.76, respectively. In contrast, structural features exhibit slightly lower validation metrics, especially size and length, where the average accuracy values were only 0.65 and 0.69, respectively. However, width presented an average accuracy value of 0.75, similar to DDS, indicating that it is the best structural descriptor for TFBS. It is important to highlight that DDS has been successfully used to predict DNA binding process in prokaryotic cells (archaeal and bacterial) [38, 44], suggesting that they are reliable descriptors to represent protein-DNA binding events to be exploited by ML models.

In the initial screening to determine the type of descriptor to be used in the work, we observed that the overall quality of the model was intermediate, with accuracy ranging around 70–75%. This indicated the need to enhance our performance in predicting TFBS. To address this, we expanded our dataset by incorporating TFBS sequences from other organisms sourced from the CollecTF database, thereby enriching our training set with additional data. [Table 1](#) presents the influence of each database on the model's predictability. The nucleotide sequences with length sizes ranging from 13 to 16 bp were separated into RegulonDB data and CollecTF data, as illustrated in the [Fig. S4](#). The model performance improves significantly when utilizing the CollecTF dataset, particularly for the 13 bp and 16 bp models. This improvement is likely

attributable to the larger volume of data provided by CollecTF, which enhances the model's ability to generalize across diverse TFBS. Recent work has evaluated the impact of data quantity on the behavior of ML models [45–47], highlighting how the size and quality of training datasets can significantly influence the predictability and generalization capabilities of these models. In addition, Bailly et al. demonstrated that, in the scenarios studied, ML models were less influenced by dataset size and consistently outperformed deep learning models [48].

Hence, we opted to select DDS as a feature for training ML models, we investigated whether the profile of DDS values from a TFBS is different from a randomly generated sequence (see [Fig. S2](#) in Supporting Information). We noticed significant differences between the TFBS sequences compared with random sequences, except for the subset with 9 bp.

[Figure 3](#) presents the individual (gray line) and the average (black line) of DDS profile of eight TFBS. The logo plot reveal notable symmetrical patterns in specific TFBS, particularly the IR motifs. This symmetry is evident in the average value profile of the 57 Cra TFBS sequences, where a distinct valley is discernible in the central region, marked by the conserved dinucleotide CG ([Fig. 3A](#)). Similar symmetrical signals are evident in the average value profile of the 28 GalR TFBS sequences, characterized by one valley composed of conserved dinucleotide CG ([Fig. 3B](#)). Moreover, the average value profile of the 30 GalS TFBS sequences reveals one valley consisting of GC ([Fig. 3C](#)). Additionally, we observed in these three cases, two well-conserved peaks composed by AA and TT dinucleotide, positioned in proximity to this valley. On the other hand, the 19 sequences of GntR TFBS ([Fig. 3D](#)) exhibit distinct profiles, lacking a well-conserved valley composed of GC. Instead of AA and TT, as observed previously, two highly conserved peaks composed of TA and TA are evident at positions 5-6 and 11-12.

It is noteworthy mentioning that all TFBS depicted in [Fig. 3A-D](#) belong to the same family. These observations raise the possibility that these symmetrical patterns observed in DDS plot might

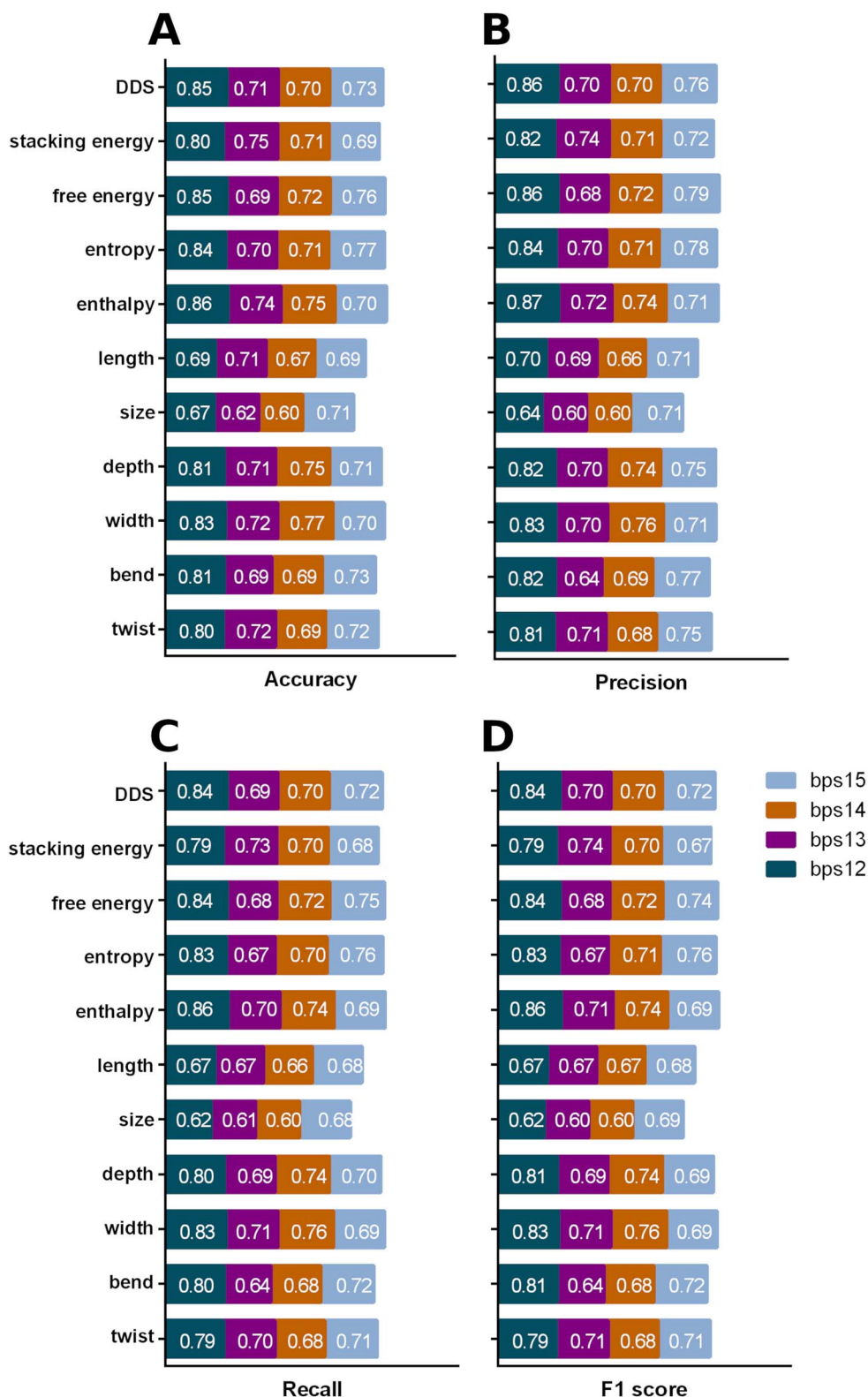


Figure 2. **Evaluation of structural and thermodynamic features of TFBS.** Genomic sequences were converted into features for training an ML model using the Random Forest algorithm. Performance metrics include accuracy (A), precision (B), recall (C), and F1 score (D).

correlate with the mechanisms of TFBS-IR recognize their TFs [49]. In contrast, a distinct profile was observed concerning the TFBS-DR (Fig. 3E-H), lacking the pronounced and poorly preserved valleys observed in IR motifs. Fig. 3E illustrates the average value profile of the 10 AgaR TFBS sequences, revealing two valleys at

the ends formed by the moderately conserved CG dinucleotides (positions 2-3 and 12-13). For the OmpR TFBS, no valleys were discerned. However, the average value profile of the 25 sequences exhibited two peaks comprising moderate (values around 1 bit) conserved dinucleotides, TT and AA (Fig. 3F). Notably, the average

Table 1. Comparison of RF model performance using RegulonDB and CollecTF datasets. Performance metrics including accuracy (ACC), precision (Prec), recall (REC), F1-score, confusion matrices, and area under the ROC curve (AUC) are provided for each dataset, highlighting the differences in classification results when using RegulonDB versus CollecTF across different sequence lengths.

	Model 13bp		Model 14bp		Model 15bp		Model 16bp	
	RegulonDB	CollecTF	RegulonDB	CollecTF	RegulonDB	CollecTF	RegulonDB	CollecTF
Dataset TFBS, nonTFBS	631, 631	1366, 1336	411, 412	434, 440	258, 259	556, 556	173, 173	112, 118
ACC	0.7747	0.9760	0.8121	0.9086	0.8750	0.9417	0.7857	0.9348
Prec	0.7728	0.9760	0.8133	0.9093	0.8720	0.9413	0.7891	0.9361
REC	0.7732	0.9758	0.8119	0.9087	0.8766	0.9437	0.7903	0.9265
F1-score	0.7730	0.9759	0.8119	0.9085	0.87356	0.9415	0.7857	0.9308
Confusion matrix	[[87 28] [29 109]]	[[250 7] [6 278]]	[[70 13] [18 64]]	[[78 10] [6 81]]	[[40 5] [8 51]]	[[100 3] [10 110]]	[[27 5] [10 28]]	[[27 1] [2 16]]
AUC	0.8615	0.9973	0.8909	0.9818	0.9465	0.9876	0.8602	0.9454

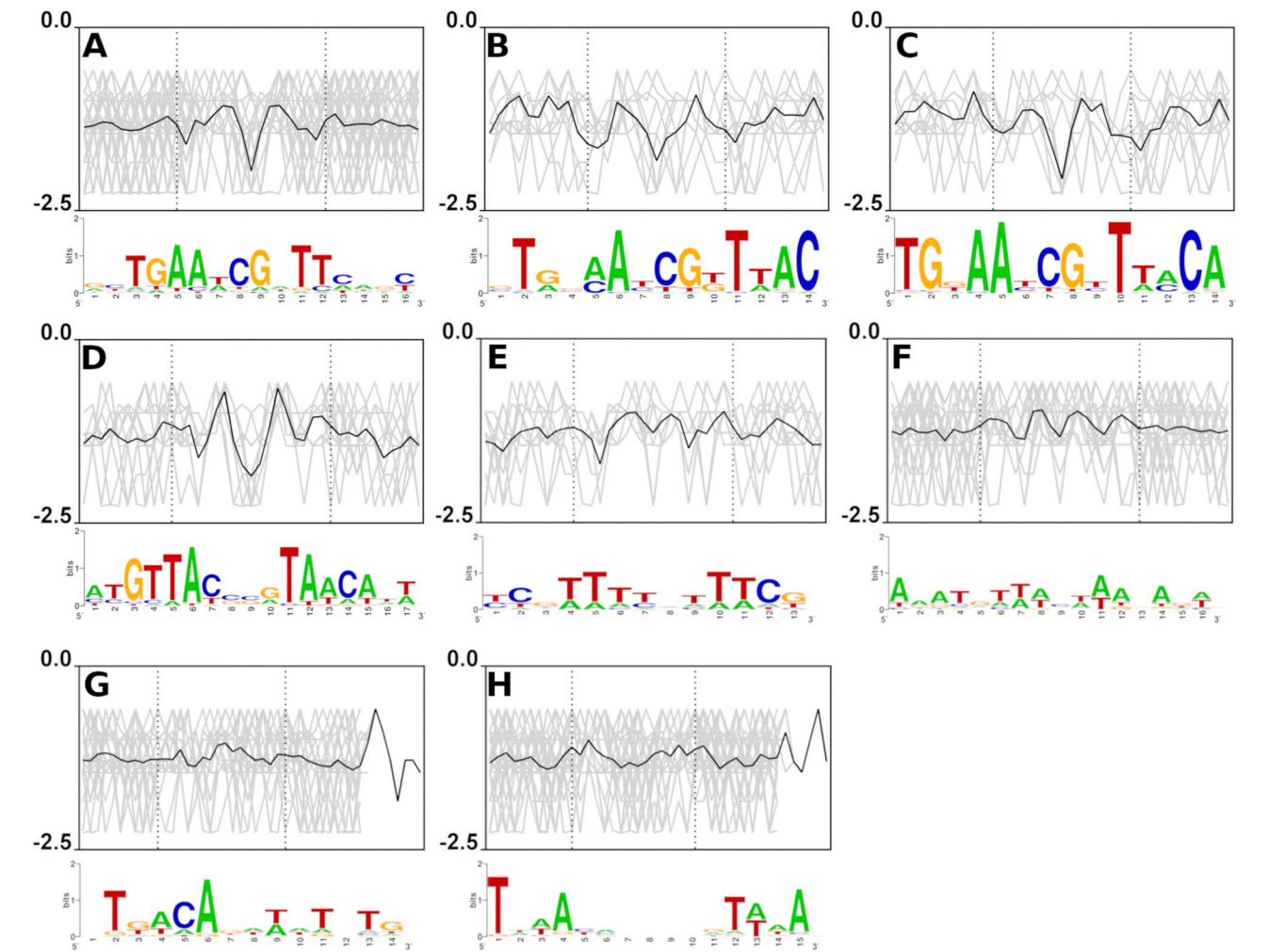


Figure 3. Comparison of DDS profiles between TFBS-IR and TFBS-DR for TFs Cra (A), GalR (B), GalS (C), GntR (D), AgaR (E), OmpR (F), PhoB (G), and PhoP (H). Black dots indicate the TFBS positions. DDS values for each TFBS sequence are depicted in gray, with the average value shown by black lines. The sequence representations below each graph were generated from the TFBS region (between the dotted lines) using the WebLogo server.

value profile of the 32 PhoB TFBS (Fig. 3G) and 40 PhoP TFBS sequences did not reveal any well-conserved nucleotides, peaks, or valleys (Fig. 3H).

The translation of nucleotide sequences into thermodynamic values, specifically DDS, presents an unexplored domain in the context of ML models for predicting TFBS. To our knowledge, there are no previous studies where DDS has been applied to binding site prediction. Given the absence of prior research utilizing

parameters like DDS, we acknowledge the potential significance of these thermodynamic features in our analysis.

Random Forest accurately predict TFBS

After initial screening to establish the most appropriate descriptor to convert the sequences into input data for the model, we conducted an assessment of various ML algorithms to evaluate their performance in classifying sequences as TFBS or non-TFBS.

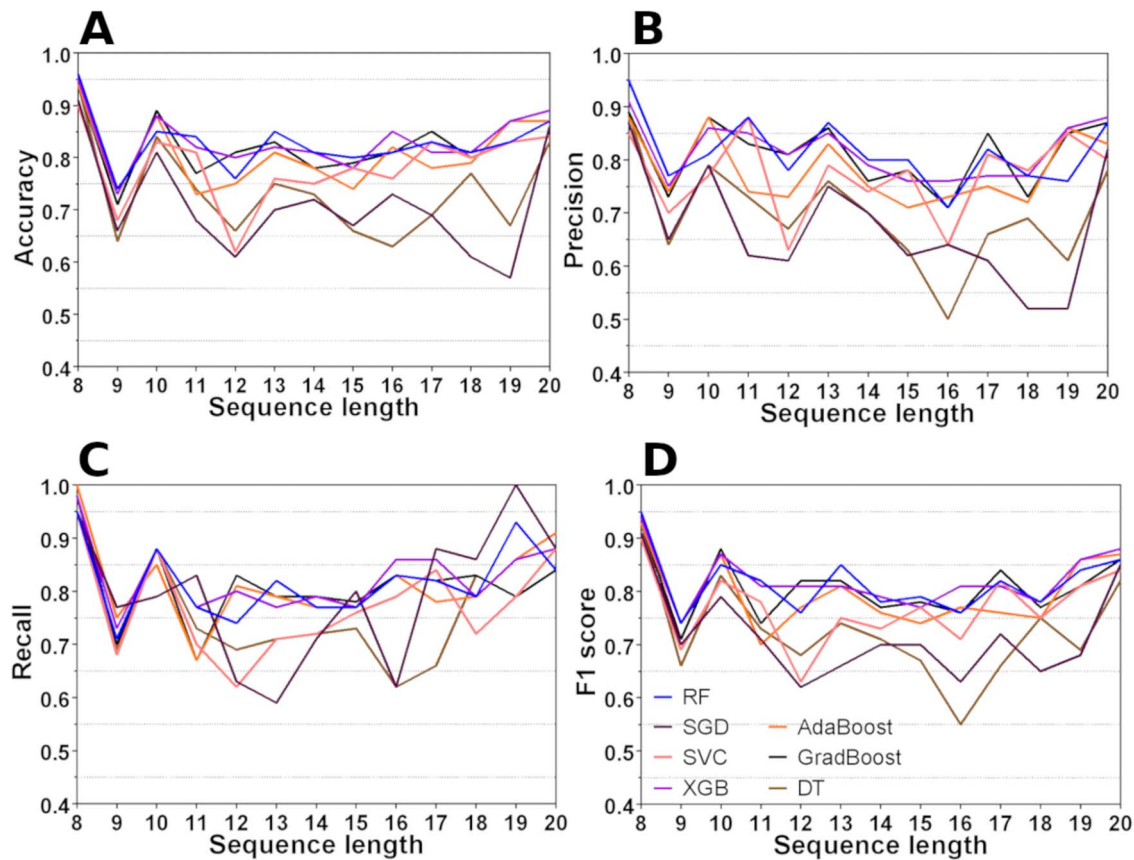


Figure 4. Comparative analysis of TFBS prediction algorithms into different sequence sizes. The performance of each algorithm is shown by lines: Random Forest (blue line), SGD (dark purple line), SVC (pink line), Adaboost (orange line), Gradboost (black line), DT (brown line), and XGB (purple line).

The negative set, comprising sequences that do not correspond to TFBS, was generated by randomly sorting TFBS sequences. However, it is well-known that one TF can recognize various DNA binding sites [2, 50], and these TFBS may not have the same number of nucleotides, leading to variations in their sizes. The majority of TFBS sequences presents in our dataset comprises 8 to 20 bp, encompassing 92.93% of our dataset. Conversely, shorter (4 to 7 bp) and longer (21 to 163 bp) TFBS constitute only 0.61 and 6.46% of the dataset, respectively. Figure 4 presents the validation results for each ML model (for more details see Table S4 in Supporting Information).

Figure 4A illustrates the accuracy values achieved by various algorithms across diverse datasets. Notably, the RF, XGBoost, and Gradient Boosting models exhibited higher accuracy compared to the others. Ensemble methods consistently exhibit superior performance across a range of prediction scenarios within biological contexts [51]. The ensemble methods mentioned achieved an optimal balance between bias and variance, resulting in more robust and generalizable models compared to simpler models like SGD and SVC. Additionally, ensemble models are capable of capturing non-linear relationships and complex interdependencies in data, rendering them better suited for addressing complex prediction problems compared to DT and linear models such as SGD and SVC. Precision, recall, and F1-score, as illustrated in Fig. 4B–D, further emphasize the effectiveness of our ensemble models in predicting TFBS. Among the models, the RF model stands out, consistently achieving an AUC above 80% across all datasets (Table 2), covering TFBS sequences ranging from 8 to 20 bp. These results suggest that our model has a high predictive accuracy in

Table 2. AUC obtained for the ML models. RF, Random Forest; SGD, stochastic gradient descent; SVC, support vector classifier; GXB, gradient boost; Ada, Adaboost; Grad, Gradboost; DT, Decision Tree.

bp	RF	SGD	SVC	XGB	Ada	Grad	DT
8	0.98	0.96	0.96	0.97	0.97	0.97	0.94
9	0.80	0.74	0.77	0.79	0.80	0.80	0.64
10	0.94	0.86	0.91	0.95	0.93	0.97	0.84
11	0.86	0.79	0.81	0.87	0.79	0.84	0.74
12	0.87	0.65	0.70	0.89	0.86	0.88	0.66
13	0.91	0.79	0.83	0.91	0.90	0.91	0.75
14	0.88	0.79	0.85	0.90	0.88	0.90	0.73
15	0.89	0.79	0.85	0.87	0.83	0.88	0.67
16	0.89	0.80	0.83	0.87	0.87	0.88	0.63
17	0.90	0.81	0.88	0.91	0.88	0.89	0.68
18	0.89	0.81	0.86	0.91	0.87	0.92	0.78
19	0.96	0.66	0.94	0.94	0.96	0.92	0.67
20	0.95	0.92	0.94	0.94	0.94	0.95	0.83

distinguishing TFBS from random sequences, highlighting its potential as an important tool for identifying new bacterial TFBS.

Assessment of model generalizability through external validation

To further assess the predictive power and applicability of our models, we performed external validation using bacterial TFBS sequences obtained from the Prodoric database [52]. Given that our models are sequence-length dependent, we stratified the Prodoric dataset by sequence length, selecting TFBS with length

Table 3. External validation results using the Prodoric database with RF models for TFBS and nonTFBS across 13, 14, and 15 bp sequences.

		Prec	REC	F1-score	Dataset
Model 13 bp	TFBS	0.94	0.73	0.82	369
	nonTFBS	0.53	0.88	0.66	129
Model 14 bp	TFBS	0.88	0.60	0.71	275
	nonTFBS	0.45	0.80	0.58	113
Model 15 bp	TFBS	0.90	0.59	0.72	441
	nonTFBS	0.40	0.80	0.53	148

sizes of 13, 14, and 15 bp. We ensured that none of the sequences in the external validation set overlapped with those used in the training phase, maintaining the independence of the validation data. For external validation, we used 369, 275, and 441 TFBS with 13, 14, and 15 bp, respectively. To generate a negative dataset (nonTFBS), we created random nucleotide sequences matching the length of the TFBS sequences.

The external validation results are presented in Table 3. Notably, the 13 bp model exhibited the highest performance, achieving a precision of 0.94, recall of 0.73, and an F1-score of 0.82 for TFBS predictions. This strong performance indicates that the 13 bp model is highly reliable in identifying true TFBS, with a favorable balance between precision and recall. In contrast, the performance metrics for nonTFBS predictions were lower, with a precision of 0.53, recall of 0.88, and an F1-score of 0.66. This suggests that the model struggles with nonTFBS predictions, likely due to the randomness of the negative dataset, which may lack discernible patterns for the model to generalize effectively.

The 14 and 15 bp models showed a decline in predictive performance compared to the 13 bp model. For the 14 bp model, the precision, recall, and F1-score for TFBS were 0.88, 0.60, and 0.71, respectively. Similarly, the 15 bp model achieved a precision of 0.90, recall of 0.59, and an F1-score of 0.72. Furthermore, the nonTFBS predictions for both the 14 and 15 bp models exhibited further declines in accuracy, with precision values of 0.45 and 0.40, and F1-scores of 0.58 and 0.53, respectively. These results are consistent with the hypothesis that randomly generated nonTFBS sequences, lacking biological relevance, present a challenge for the model's generalization capabilities.

We compared our model with a recent study, who employed ML algorithms such as RF and XGBoost for TFBS prediction using *S. cerevisiae* PBM data [27]. However, due to the lack of publicly available models and datasets from their work, we focused on comparing the performance metrics reported in their paper with those achieved in our models (Table S5). Our results demonstrate that both our RF and XGBoost models, particularly those trained on sequences of 11 and 13 bp lengths, outperform the models by previous study in terms of recall while achieving comparable precision.

These findings confirm the robustness of our models in identifying TFBS across various sequence lengths. However, the variability in nonTFBS predictions highlights an area for future improvement. Refining the generation of nonTFBS sequences to include more biologically relevant negative examples, rather than relying on purely random sequences, may improve the model's ability to distinguish between TFBS and nonTFBS regions with greater accuracy.

Model interpretability using DNA duplex stability

SHAP has emerged as a powerful tool for understanding the contributions of individual features to ML models. By decomposing

the model's predictions into contributions from each input feature, SHAP enables us to quantify the impact of specific bp on the model's decision-making process. We consider that the search for the interpretability of the DDS values used as input data for the model can help in understanding the functional role of specific motifs related to protein-DNA interaction. DNA sequences are characterized by complex dependencies between adjacent bp and higher-order sequence motifs [53]. These dependencies arise from various factors, including the structural properties of DNA, such as base stacking and hydrogen bonding, as well as the functional constraints imposed by biological processes such as transcription, replication, and DNA-protein interactions. Figure 5 presents the results of the SHAP analysis, illustrating the influence of individual bp on TFBS prediction.

The importance analysis of the model reveals a significant pattern. The most influential features for TFBS classification, frequently involve adjacent bp, as observed for bp4 and bp5 (Fig. 5A), bp11 and bp12 (Fig. 5B), and bp8 and bp7 (Fig. 5C). This proximity suggests a potential functional significance wherein specific regions of DNA might play a pivotal role in molecular recognition and interaction with proteins. This region's importance may vary depending on the characteristics of associated TFs families, considering factors like amino acid conservation and the structural arrangement of the DNA binding domain (DBD) motif. For instance, in our model for predicting 8-bp TFBS (Fig. 5D), particular attention is drawn to the central region, comprising bp4 and bp6. This focus on the central region underscores its potential role as a key determinant in TFBS classification, suggesting that interactions within this segment may hold crucial information for understanding TF binding specificity and regulatory function.

These results reveal that the use of DDS values to represent TFBS can provide a valuable tool for elucidating the intricate dynamics of protein-DNA interactions. FadR, a well-studied member of the GntR family of TFs, serves as an illustrative example in this regard (Fig. 6).

We obtained the structure of the crystallized protein-DNA complex (Fig. 6A) from the Protein Data Bank, under PDB code id 1H9T [54]. By converting TFBS sequences into DDS values (Fig. 6B), we were able to capture a phenomenon related to the symmetry of TFBS, linked by a symmetric and opposite region forming between the bp of position 8-11, similar to a palindrome typical of several TFBS (see Fig. S3 in Supporting Information). We also observe two well-conserved regions, comprising the bp of position 2-7 and 12-17. The feature importance analysis (Fig. 6C) underscores the significance of conserved bp in predicting TFBS, with bp11 being the exception. Notably, bp7, comprised of nucleotides G7 and T8, stands out due to its crucial interactions with amino acids R49, R45, and S7 in both the A and B chains of the FadR dimer. Another noteworthy feature highlighted by the SHAP analysis is bp10, where a significant interaction between the C/G10 nucleotide and the T46 amino acid of both protein chains was observed. Korostelev *et al.* noted that conserved positions likely contribute to initial DNA binding, while correlated positions fine-tune interactions with specific sites [55]. Furthermore, Yeo *et al.* performed structural studies on *Bacillus halodurans* FadR [56], highlighting critical interactions between bp equivalent to bp6-7 and bp9-10 in Fig. 6B.

Although further studies need to be conducted to explore additional families of transcriptional factors, a notable correlation emerges between DDS values and experimentally observed DNA-protein contacts. Korostelev *et al.* used several crystal structures of related TFs in the DNA-bound form and demonstrated a significant correlation between specific pairs and contacting

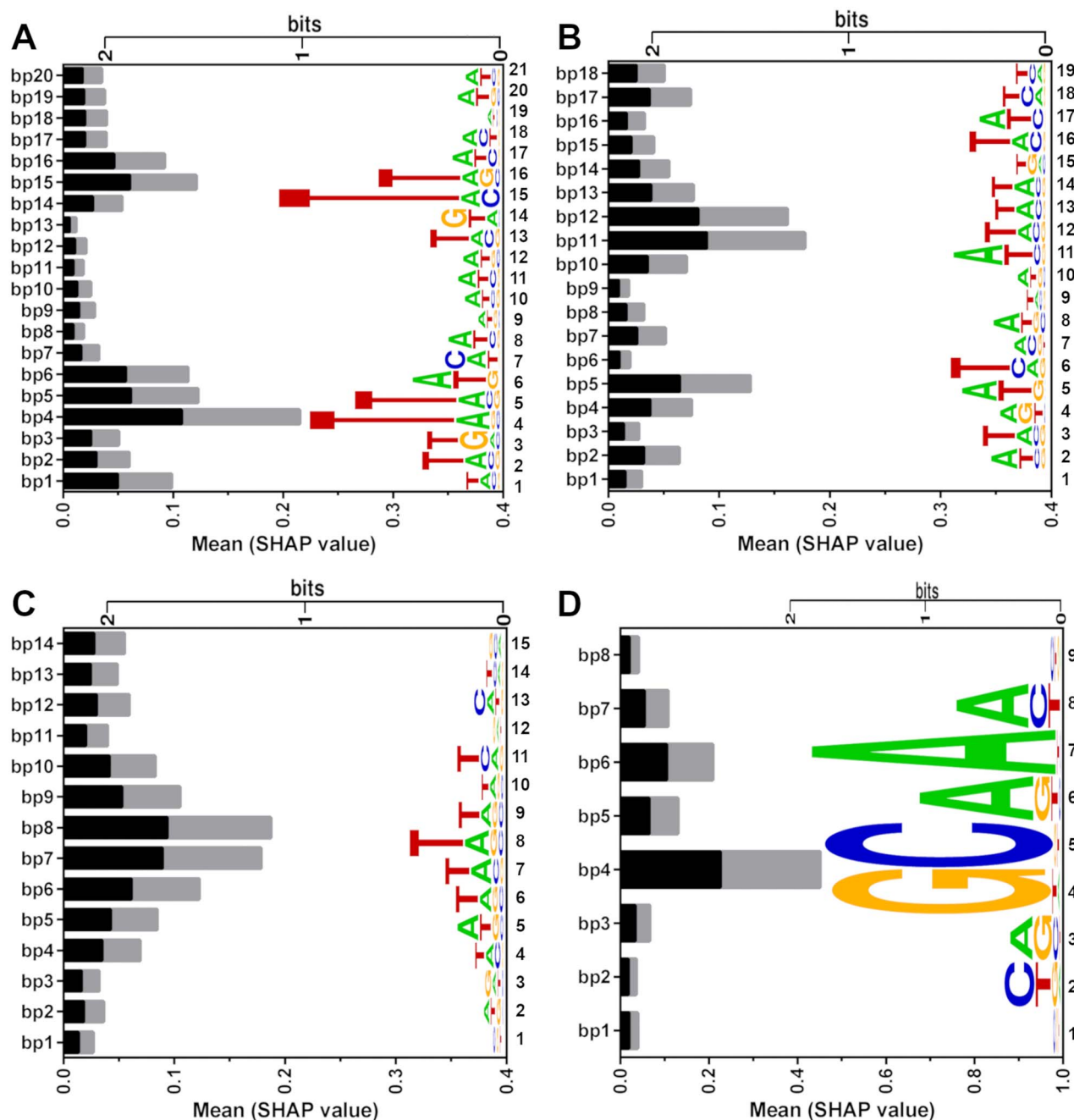


Figure 5. SHAP analysis (stacked bar chart) of DDS values alongside sequence LOGO plots for four different sequence lengths: A (20 bp), B (18 bp), C (14 bp), and D (8 bp). The SHAP analysis highlights the impact of the bp of TFBS (black) and nonTFBS (gray) on the model's prediction, while the sequence logos illustrate conserved nucleotide residues within these regions.

positions. They suggested that such correlations, when combined with sequence data, could become powerful tools for studying the evolution of TF families and the coevolution of TF-DNA interactions [55]. This observation underscores the promise of DDS-based approaches in predicting TFBS and characterizing protein-DNA interactions. Several studies have proposed that the conservation of bp within a motif correlates significantly with the number of contacts they establish with the bound TF [55, 57, 58]. Consequently, structural analyses of TFBS have been employed to predict amino acid—base contacts for TFs, offering valuable insights into protein-DNA interactions that warrant further experimental validation [59–63].

The composition of nucleotides can differentiate a TFBS-DR from a TFBS-IR

After successfully predicting TFBS, we developed a model capable of distinguishing TFBS-DR from TFBS-IR, highlighting the implications of this distinction for understanding gene regulatory dynamics. By elucidating the nucleotide-level characteristics that define these regulatory elements, we aim to shed light on the nuanced interplay between sequence structure and function in the context of transcriptional regulation. To achieve this, we employed the RF algorithm, chosen for its superior performance in TFBS classification, particularly with a subset of sequences consisting of 15 bp (see Table S6). We further subdivided the classification

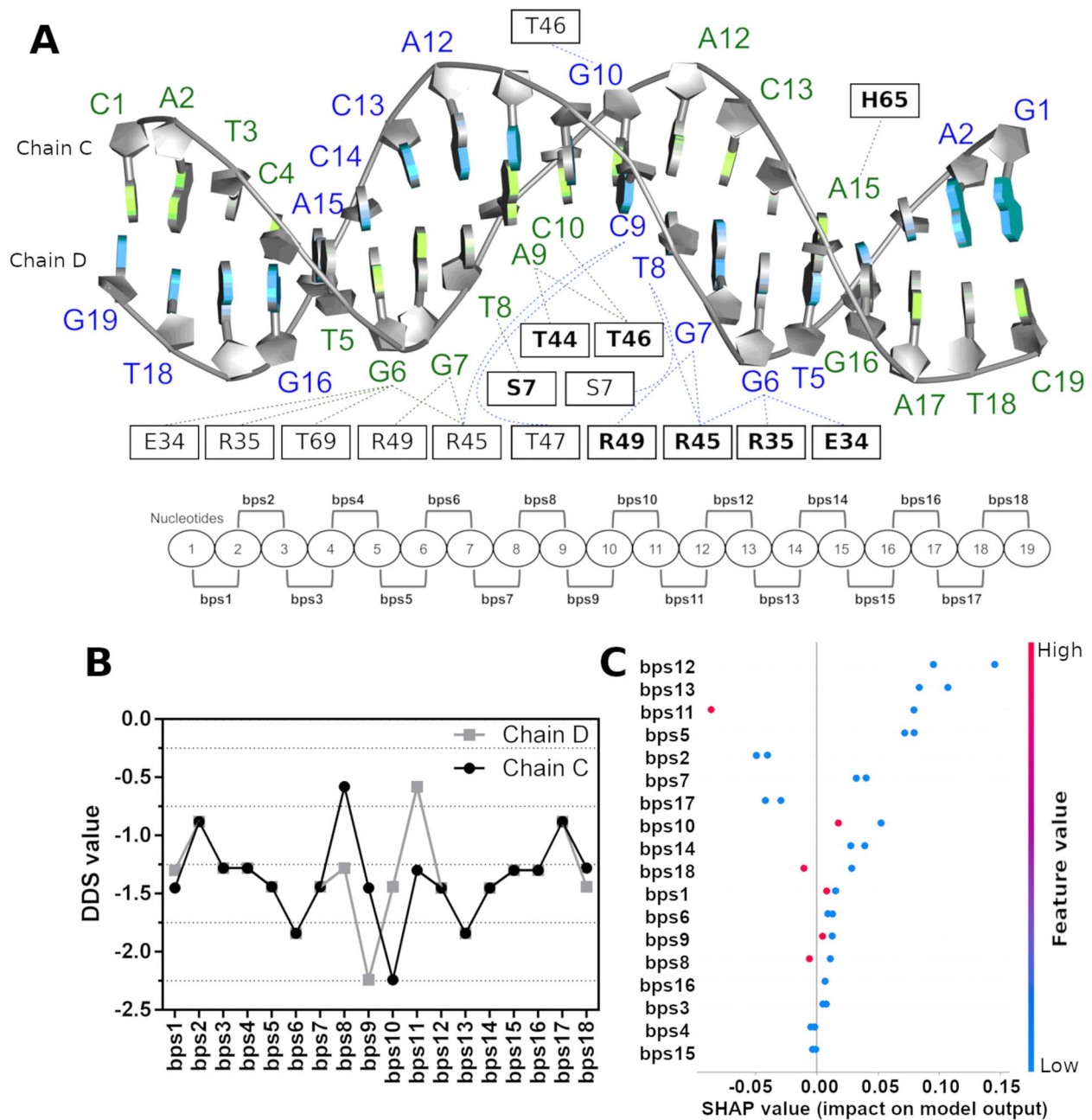


Figure 6. **Correlation between the structure of the TFBS-FadR complex and DDS values.** (A) Depiction of intermolecular interactions between TFBS and FadR observed in the crystallographic structure (PDB code: 1H9T). The dotted lines indicate the hydrogen bond interactions as visualized using the PyMOL software. (B) Conversion of TFBS, chains C and D, into DDS values. (C) Assessment of the impact of nucleotide bp from chains C and D on the TFBS prediction model.

task into three distinct categories: DR versus non-DR, IR versus non-IR, and DR versus IR. This segmentation allowed us to assess whether the data quantity imbalance between the TFBS-IR (738 sequences) and TFBS-DR (316 sequences) classes could lead to overfitting (Table 4).

The findings reveal that distinguishing a TFBS-DR from a random sequence yields an accuracy of 0.858, whereas the accuracy significantly increases to 0.966 when classifying a TFBS-IR from nonTFBS-IR sequence. This disparity can be attributed to the prevalence of symmetry of IR in relation to DR, thus leading to a more characteristic and distinguishable signal from random sequences. Additionally, the model developed for TFBS-DR versus TFBS-IR classification achieved an accuracy of 0.891 and a

precision of 0.855. Notably, the generated models demonstrate robustness against overfitting, as indicated by the results of the K-fold cross-validation test, which reveal no discernible pattern of overfitting.

Conclusion

In recent years, the exponential growth of biological data available in databases has paved the way for the development of efficient and robust ML-based models. Among these, models for predicting TFBS hold significant relevance for understanding gene regulation. While existing models often rely on the OHE strategy for translating nucleotide sequences, which lacks inherent biological

Table 4. Validation of the ML model for classifying TFBS-DR and TFBS-IR sequences.

	DR vs nonDR	IR vs nonIR	IR vs DR
Dataset	316 DR 316 non-DR	738 IR 738 non-IR	738 IR 316 DR
ACC	0.858	0.966	0.891
Prec	0.855	0.966	0.855
Rec	0.855	0.967	0.881
F1-Score	0.855	0.966	0.866
K-fold	[0.890, 0.812,	[0.959, 0.973,	[0.896, 0.839,
cross	0.936, 0.888,	0.986, 0.973,	0.821, 0.877,
validation	0.905, 0.968,	0.966, 0.993,	0.876, 0.895,
	0.905, 0.936,	0.959, 0.979,	0.895, 0.828,
	0.936, 0.905]	0.979, 0.986]	0.857, 0.866]

or physical meaning, our work aimed to explore alternative methods. In this study, we investigated several thermodynamic and structural parameters to convert TFBS into meaningful features for training ML models. Our approach not only aimed at accurate prediction but also sought to provide insights into the structural characterization of protein-DNA complexes.

The findings demonstrated in this work point that DDS effectively represented the TFBS dataset, revealing characteristic symmetry patterns reminiscent of palindromic TFBS. By subdividing the dataset based on the number of nucleotides, we ensured uniform evaluation by various ML algorithms. Using the RF algorithm, we achieved an average accuracy of over 82% in distinguishing TFBS from random sequences. Furthermore, we developed a model capable of differentiating TFBS into IR and DR with an accuracy of 89%. The model trained on sequences with a length of 13 bp demonstrated the best performance and is therefore the most recommended for TFBS prediction. However, since TFBS can vary in size and given the specific challenges associated with different contexts, it may be necessary to conduct predictions using models designed for various bp lengths.

These results underscore the potential of converting nucleotide sequences into DDS values, which may provide valuable insights into how bacterial proteins recognize DNA structures. Ongoing studies are investigating how this approach can aid in structural characterization. Preliminary findings suggest a promising avenue for developing models that enhance our understanding of protein-DNA interaction sites.

Key Points

- We employed ML algorithms to predict TFBS and classify them as DR or IR.
- We divided the set of TFBS nucleotide sequences by size, ranging from 8 to 20 bp, and converted them into thermodynamic data known as DDS.
- We demonstrate that the RF algorithm accurately predicts TFBS and effectively distinguishes between IR and DR.
- Interestingly, upon converting the bp of several TFBS-IR into DDS values, we observed a symmetric profile typical of the palindromic structure associated with these architectures.
- We show a novel TFBS prediction model based on a DDS characteristic that may indicate how respective proteins interact with bp.

Supplementary data

Supplementary data are available at Briefings in Bioinformatics online and publicly available at <https://github.com/farias-ab/TFBS-Prediction.git>.

Conflict of interest: None declared.

Funding

This work was supported by PAPIIT-DGAPA UNAM grant IN220523 (Ernesto Perez-Rueda) and IA207423 (Edgardo Galan-Vasquez), and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES/UNAM, COOPBRAS, n. 05/2019, 88887.368759/2019-00 (MFN) - Finance Code 001. A.B.F. was supported by a fellowship from CNPq (process no. 420622/2023-3). M.F.N. was supported by a fellowship from CNPq (process no. 305895/2022-2) and FAPERJ-CNE (process no. E-26/200.555/2023). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Lambert SA, Jolma A, Campitelli LF. et al. The human transcription factors. *Cell* 2018;**172**:650–65. <https://doi.org/10.1016/j.cell.2018.01.029>.
2. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* 2012;**13**:613–26. <https://doi.org/10.1038/nrg3207>.
3. Heltberg ML, Krishna S, Jensen MH. On chaotic dynamics in transcription factors and the associated effects in differential gene regulation. *Nat Commun* 2019;**10**:71.
4. Inukai S, Kock KH, Bulyk ML. Transcription factor—DNA binding: beyond binding site motifs. *Curr Opin Genet Dev* 2017;**43**:110–9. <https://doi.org/10.1016/j.gde.2017.02.007>.
5. Perez-Rueda E. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res* 2000;**28**:1838–47. <https://doi.org/10.1093/nar/28.8.1838>.
6. Fan L, Wang T, Hua C. et al. A compendium of DNA-binding specificities of transcription factors in *Pseudomonas syringae*. *Nat Commun* 2020;**11**:4947.
7. Swint-Kruse L, Elam CR, Lin JW. et al. Plasticity of quaternary structure: twenty-two ways to form a LacI dimer. *Protein Sci* 2001;**10**:262–76. <https://doi.org/10.1110/ps.35801>.
8. Fernandez-Lopez R, Ruiz R, delÂ Campo. et al. Structural basis of direct and inverted DNA sequence repeat recognition by helix—Turn—Helix transcription factors. *Nucleic Acids Res* 2022;**50**:11938–47. <https://doi.org/10.1093/nar/gkac1024>.
9. Schumacher MA, Den Hengst CD, Bush MJ. et al. The MerR-like protein BldC binds DNA direct repeats as cooperative multimers to regulate *Streptomyces* development. *Nat Commun* 2018;**9**:1139.
10. Abbani MA, Papagiannis CV, Sam MD. et al. Structure of the cooperative Xis—DNA complex reveals a micronucleoprotein filament that regulates phage lambda intasome assembly. *Proc Natl Acad Sci* 2007;**104**:2109–14. <https://doi.org/10.1073/pnas.0607820104>.
11. Muller PAJ, Klomp LWJ. ATOX1: a novel copper-responsive transcription factor in mammals? *Int J Biochem Cell Biol* 2009;**41**:1233–6. <https://doi.org/10.1016/j.biocel.2008.08.001>.
12. Weidemüller P, Kholmatov M, Petsalaki E. et al. Transcription factors: bridge between cell signaling and gene regulation. *Proteomics* 2021;**21**:2000034.

13. Siggers T, Gordán R. Protein—DNA binding: complexities and multi-protein codes. *Nucleic Acids Res* 2014;**42**:2099–111. <https://doi.org/10.1093/nar/gkt1112>.
14. Joiret M, Leclercq M, Lambrechts G. et al. Cracking the genetic code with neural networks. *Front Artif Intell* 2023;**6**:1128153.
15. Tierrafría VH, Rioualen C, Salgado H. et al. RegulonDB 11.0: comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12. *Microb Genom* 2022;**8**. <https://doi.org/10.1099/mgen.0.000833>.
16. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2022;**50**:D165–73. <https://doi.org/10.1093/nar/gkab1113>.
17. Kiliç S, White ER, Sagitova DM. et al. CollecTF: a database of experimentally validated transcription factor-binding sites in bacteria. *Nucleic Acids Res* 2014;**42**:D156–60. <https://doi.org/10.1093/nar/gkt1123>.
18. Wingender E. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996;**24**:238–41. <https://doi.org/10.1093/nar/24.1.238>.
19. Hume MA, Barrera LA, Gisselbrecht SS. et al. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein—DNA interactions. *Nucleic Acids Res* 2015;**43**:D117–22. <https://doi.org/10.1093/nar/gku1045>.
20. Martinez GS, Perez-Rueda E, Kumar A. et al. Explainable artificial intelligence as a reliable annotator of archaeal promoter regions. *Sci Rep* 2023;**13**:1763.
21. Zhang Q, He Y, Wang S. et al. Base-resolution prediction of transcription factor binding signals by a deep learning framework. *PLoS Comput Biol* 2022;**18**:e1009941. <https://doi.org/10.1371/journal.pcbi.1009941>.
22. Alipanahi B, Delong A, Weirauch MT. et al. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8. <https://doi.org/10.1038/nbt.3300>.
23. Yang J, Ma A, Hoppe AD. et al. Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. *Nucleic Acids Res* 2019;**47**:7809–24. <https://doi.org/10.1093/nar/gkz672>.
24. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning—based sequence model. *Nat Methods* 2015;**12**:931–4. <https://doi.org/10.1038/nmeth.3547>.
25. Quang D, Xie X, DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* 2016;**44**:e107–7. <https://doi.org/10.1093/nar/gkw226>.
26. Zeng Y, Gong M, Lin M. et al. A review about transcription factor binding sites prediction based on deep learning. *IEEE Access* 2020;**8**:219256–74. <https://doi.org/10.1109/ACCESS.2020.3042903>.
27. Yaman OU, Çalik P. MachineTFBS: motif-based method to predict transcription factor binding sites with first-best models from machine learning library. *Biochem Eng J* 2023;**198**:108990. <https://doi.org/10.1016/j.bej.2023.108990>.
28. Canals A, Pieretti S, Muriel-Masanes M. et al. ToxR activates the *vibrio cholerae* virulence genes by tethering DNA to the membrane through versatile binding to multiple sites. *Proc Natl Acad Sci* 2023;**120**:e2304378120.
29. Bailey TL, Johnson J, Grant CE. et al. The MEME suite. *Nucleic Acids Res* 2015;**43**:W39–49. <https://doi.org/10.1093/nar/gkv416>.
30. Crooks GE, Hon G, Chandonia J-M. et al. WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90. <https://doi.org/10.1101/gr.849004>.
31. Karas H, Knüppel R, Schulz W. et al. Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Bioinformatics* 1996;**12**:441–6. <https://doi.org/10.1093/bioinformatics/12.5.441>.
32. Gorin AA, Zhurkin VB, Wilma K. B-DNA twisting correlates with base-pair morphology. *J Mol Biol* 1995;**247**:34–48. <https://doi.org/10.1006/jmbi.1994.0120>.
33. Hogan ME, Austin RH. Importance of DNA stiffness in protein—DNA binding specificity. *Nature* 1987;**329**:263–6. <https://doi.org/10.1038/329263a0>.
34. Sugimoto N, Nakano S-I, Yoneyama M. et al. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res* 1996;**24**:4501–5. <https://doi.org/10.1093/nar/24.22.4501>.
35. Breslauer KJ, Frank R, Blöcker H. et al. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci* 1986;**83**:3746–50. <https://doi.org/10.1073/pnas.83.11.3746>.
36. Pérez A, Noy A, Filip Lankas F. et al. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res* 2004;**32**:6144–51. <https://doi.org/10.1093/nar/gkh954>.
37. SantaLucia J, Hicks D. The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 2004;**33**:415–40. <https://doi.org/10.1146/annurev.biophys.32.110601.141800>.
38. Martinez GS, Sarkar S, Kumar A. et al. Characterization of promoters in archaeal genomes based on DNA structural parameters. *MicrobiologyOpen* 2021;**10**:e1230.
39. Martinez GS, Perez-Rueda E, Kumar A. et al. CDBProm: the comprehensive directory of bacterial promoters. *NAR Genom Bioinform* 2024;**6**:lqae018. [eprint: https://academic.oup.com/nargab/article-pdf/6/1/lqae018/56727634/lqae018.pdf](https://academic.oup.com/nargab/article-pdf/6/1/lqae018/56727634/lqae018.pdf).
40. Bansal M, Kumar A, Yella VR. Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. *Curr Opin Struct Biol* 2014;**25**:77–85. <https://doi.org/10.1016/j.sbi.2014.01.007>.
41. Yella VR, Bansal M. DNA structural features of eukaryotic TATA-containing and TATA-less promoters. *FEBS Open Bio* 2017;**7**:324–34. <https://doi.org/10.1002/2211-5463.12166>.
42. Pedregosa F, Varoquaux G, Gramfort A. et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
43. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S. et al. (eds.), *Advances in Neural Information Processing Systems*, Vol. **30**, Red Hook, NY, USA: Curran Associates, Inc., 2017.
44. Martinez GS, Pérez-Rueda E, Sarkar S. et al. Machine learning and statistics shape a novel path in archaeal promoter annotation. *BMC Bioinformatics* 2022;**23**:171.
45. Schwabe D, Becker K, Seyferth M. et al. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *npj Digit. Med* 2024;**7**:203.
46. Burgoon LD, Kluxen FM, Hüser A. et al. The database makes the poison: how the selection of datasets in QSAR models impacts toxicant prediction of higher tier endpoints. *Regul Toxicol Pharmacol* 2024;**151**:105663. <https://doi.org/10.1016/j.yrtph.2024.105663>.
47. Dimitsaki S, Gavrilidis GI, Dimitriadis VK. et al. Benchmarking of machine learning classifiers on plasma proteomic for COVID-19 severity prediction through interpretable artificial intelligence. *Artif Intell Med* 2023;**137**:102490. <https://doi.org/10.1016/j.artmed.2023.102490>.
48. Bailly A, Blanc C, Francis É. et al. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Comput Methods Programs Biomed* 2022;**213**:106504. <https://doi.org/10.1016/j.cmpb.2021.106504>.

49. Coons LA, Burkholder AB, Hewitt SC. et al. Decoding the inversion symmetry underlying transcription factor DNA-binding specificity and functionality in the genome. *iScience* 2019;**15**:552–91. <https://doi.org/10.1016/j.isci.2019.04.006>.
50. Chen Y, Lin Y-C-D, Luo Y. et al. Quantitative model for genome-wide cyclic AMP receptor protein binding site identification and characteristic analysis. *Brief Bioinform* 2023;**24**:bbad138.
51. Boulesteix A-L, Janitza S, Kruppa J. et al. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Min Knowl Discov* 2012;**2**:493–507. <https://doi.org/10.1002/widm.1072>.
52. Dudek C-A, Jahn D. PRODORIC: state-of-the-art database of prokaryotic gene regulation. *Nucleic Acids Res* 2022;**50**:D295–302. <https://doi.org/10.1093/nar/gkab1110>.
53. Stormo GD, Fields DS. Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem Sci* 1998;**23**: 109–13. [https://doi.org/10.1016/S0968-0004\(98\)01187-6](https://doi.org/10.1016/S0968-0004(98)01187-6).
54. Van Aalten DMF. The structural basis of acyl coenzyme A-dependent regulation of the transcription factor FadR. *EMBO J* 2001;**20**:2041–50. <https://doi.org/10.1093/emboj/20.8.2041>.
55. Korostelev YD, Zharov IA, Mironov AA. et al. Identification of position-specific correlations between DNA-binding domains and their binding sites. Application to the MerR family of transcription factors. *PLoS One* 2016;**11**:e0162681. <https://doi.org/10.1371/journal.pone.0162681>.
56. Yeo HK, Park YW, Lee JY. Structural basis of operator sites recognition and effector binding in the TetR family transcription regulator FadR. *Nucleic Acids Res* 2017;**45**:4244–54. <https://doi.org/10.1093/nar/gkx009>.
57. Morozov AV, Siggia ED. Connecting protein structure with predictions of regulatory sites. *Proc Natl Acad Sci* 2007;**104**:7068–73. <https://doi.org/10.1073/pnas.0701356104>.
58. Mirny LA. Structural analysis of conserved base pairs in protein–DNA complexes. *Nucleic Acids Res* 2002;**30**:1704–11.
59. Suvorova IA, Korostelev YD, Gelfand MS. GntR family of bacterial transcription factors and their DNA binding motifs: structure, positioning and co-evolution. *PLOS ONE* 2015;**10**:e0132618. <https://doi.org/10.1371/journal.pone.0132618>.
60. Mahony S, Auron PE, Benos PV. Inferring protein–DNA dependencies using motif alignments and mutual information. *Bioinformatics* 2007;**23**:i297–304. <https://doi.org/10.1093/bioinformatics/btm215>.
61. Camas FM, Alm EJ, Poyatos JF. Local gene regulation details a recognition code within the LacI transcriptional factor family. *PLoS Comput Biol* 2010;**6**:e1000989. <https://doi.org/10.1371/journal.pcbi.1000989>.
62. Desai TA, Rodionov DA, Gelfand MS. et al. Engineering transcription factors with novel DNA-binding specificity using comparative genomics. *Nucleic Acids Res* 2009;**37**:2493–503. <https://doi.org/10.1093/nar/gkp079>.
63. Luscombe NM, Thornton JM. Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* 2002;**320**:991–1009. [https://doi.org/10.1016/S0022-2836\(02\)00571-5](https://doi.org/10.1016/S0022-2836(02)00571-5).