Briefings in Bioinformatics, 22(6), 2021, 1–14

https://doi.org/10.1093/bib/bbab336
Problem Solving Protocol

DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences

Jian Zhang, Sina Ghadermarzi, Akila Katuwawala and Lukasz Kurgan

Corresponding authors: Jian Zhang, School of Computer and Information Technology, Xinyang Normal University, No.237, Nanhu Road, Xinyang 464000, Henan Province, P.R. China. E-mail: jianzhang@xynu.edu.cn and Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, Virginia 23284, USA. Tel.: (804) 827-3986. E-mail: lkurgan@vcu.edu

Abstract

Efforts to elucidate protein–DNA interactions at the molecular level rely in part on accurate predictions of DNA-binding residues in protein sequences. While there are over a dozen computational predictors of the DNA-binding residues, they are DNA-type agnostic and significantly cross-predict residues that interact with other ligands as DNA binding. We leverage a custom-designed machine learning architecture to introduce DNAgenie, first-of-its-kind predictor of residues that interact with A-DNA, B-DNA and single-stranded DNA. DNAgenie uses a comprehensive physiochemical profile extracted from an input protein sequence and implements a two-step refinement process to provide accurate predictions and to minimize the cross-predictions. Comparative tests on an independent test dataset demonstrate that DNAgenie outperforms the current methods that we adapt to predict residue-level interactions with the three DNA types. Further analysis finds that the use of the second (refinement) step leads to a substantial reduction in the cross predictions. Empirical tests show that DNAgenie's outputs that are converted to coarse-grained protein-level predictions compare favorably against recent tools that predict which DNA-binding proteins interact with double-stranded versus single-stranded DNAs. Moreover, predictions from the sequences of the whole human proteome reveal that the results produced by DNAgenie substantially overlap with the known DNA-binding proteins while also including promising leads for several hundred previously unknown putative DNA binders. These results suggest that DNAgenie is a valuable tool for the sequence-based characterization of protein functions. The DNAgenie's webserver is available at http://biomine.cs.vcu.edu/servers/DNAgenie/.

Key words: protein–DNA interactions; DNA-binding residues; A-DNA; B-DNA; single-stranded DNA; double-stranded DNA; prediction

Introduction

Protein–DNA interactions drive regulation of gene expression and DNA processing and repair [1, 2]. These interactions involve single-stranded DNA (ssDNA), double-stranded DNAs (dsDNA) [3] and a number of noncanonical DNA structures that include G-quadruplex [4–6], cruciform [7], i-motif [8], triplex [9] and hairpins [10], to name a few. The ssDNA-binding proteins are involved in DNA replication, recombination, and repair while the dsDNA-binding proteins play key roles in numerous cellular

Jian Zhang is an associate professor in the School of Computer and Information Technology at the Xinyang Normal University. His research expertise is in machine learning and structural bioinformatics. The website of his lab can be found at http://www.inforstation.com/biocomlab/.
Sina Ghadermarzi is a Ph.D. student in Computer Science at the Virginia Commonwealth University. His research focuses on the computational prediction of protein–drug and protein–ligand interactions and structural genomics.
Akila Katuwalala has recently graduated with Ph.D. in Computer Science from the Virginia Commonwealth University. He is a staff scientist with Adimab, LLC at Palo Alto in California. His research interests are in the computational prediction and characterization of intrinsic disorder and protein–ligand interactions.
Lukasz Kurgan is a fellow of AIMBE and the Robert J. Mattauch Endowed Professor of Computer Science at the Virginia Commonwealth University. His research work encompasses structural and functional characterization of proteins. He serves on the Editorial Board of *Bioinformatics* and as the Associate Editor-in-Chief of *Biomolecules*. Details about his research lab are at http://biomine.cs.vcu.edu/.
Submitted: 29 March 2021; **Received (in revised form):** 2 July 2021

processes that include DNA cleaving, chromosome packing and transcription [11, 12]. The dsDNA assumes several functionally different conformations with the B-DNA being the most abundant form and A-DNA and Z-DNA being the other common dsDNA subtypes [13, 14]. A-DNA and B-DNA are right-handed double helices that differ in the spatial arrangement of the base pairs, while Z-DNA is a left-handed duplex. Structural details of protein-DNA complexes are used to gain invaluable mechanistic insights into the corresponding protein functions [15–17], to characterize different modes of protein–DNA interactions [18] and to address a variety of other basic science and applied studies. The source data for these studies is generated by experimental techniques, such as X-ray crystallography and NMR. However, application of these techniques is relatively expensive and time-consuming. For instance, the structural genomics consortia reported solving about 13 500 structures over 15 years at the cost of 2 billion dollars, which converts to an average per-protein cost of $148 000 [19]. Consequently, cost- and time-efficient computational methods have been used to support and advance studies of protein–DNA interactions.

The computational methods that characterize these interactions on the protein side are categorized into two groups: structure-based versus sequence-based [20]. The former category identifies whether and how a given protein interacts with DNA by comparing the structure of a query protein to the structures of similar proteins that are in complex with DNA [21]. However, such structural information is available for a relatively small subset of DNA-binding proteins. A quick search of Protein Data Bank (PDB) reveals about 5600 protein-DNA complexes [22]. While homology modeling can be used to improve coverage by modeling proteins with unknown structure, this approach is computationally expensive, was estimated to cover only about 25% of proteins overall and 19% of eukaryotic proteins [23], and the resulting structural models may lack in quality for accurate prediction of interactions [24, 25].

The sequence-based predictors use protein sequence to provide either a coarse-grained level prediction of DNA-binding proteins (i.e., they identify DNA-binding proteins without details about the underlying interactions) or to predict DNA-binding residues (i.e., they identify the DNA-binding amino acids in the protein sequence) [20]. While they produce results at a lower resolution (protein-level and residue-level) compared to the atomic-level results generated by the structure-based methods, they can be applied to analyze any of the millions of proteins with the known sequences. The protein-level sequence-based predictors, which include DNA-Prot [26], DNAbinder [27], and StackDPPred [28], were summarized in a recent survey [20]. Here, we focus on the sequence-based methods that predict DNA-binding residues. They provide more details compared to the predictors of DNA-binding proteins while their results can be still used to identify DNA-binding proteins. Recent reviews [20, 29] and literature search reveal over a dozen residue-level sequence-based methods that include (chronologically) DBS-pred [30], DBS-PSSM [31], BindN [32], DNABindR [33], DP-Bind [34], DISIS [35], BindN-RF [36], DBindR [37], ProteDNA [38], NAPS [39], BindN+ [40], DNABR [41], TargetDNA [42], DRNApred [43], hybridNAP [20] and DNApred [44]. Predictions generated by these tools provide useful clues that support functional characterization of proteins. As an example, DRNApred [43] was recently used to characterize proteomes of coronaviruses [45] and Japanese encephalitis virus [46], to functionally characterize BEX3 [47] and σ^{E}TF [48] proteins, and to investigate interactome of ANKRD55 [49]. Although the residue-level methods generally produce accurate results [29], recent works reveal that they suffer a significant drawback. Namely, they incorrectly cross-predict residues that interact with other ligands (RNAs, proteins and small molecules) as DNA-binding [29, 50, 51]. For instance, depending on a method used, between 23 and 60% of the native RNA-binding residues were shown to be cross-predicted as DNA-binding [29]. In other words, these methods fail to reliably differentiate between interactions with DNA, RNA, proteins and small molecules. This is explained by the fact that they were trained using datasets consisting solely of DNA-binding proteins, whilst lacking proteins that interact with the other partners [29, 52].

The current residue-level methods provide DNA-type-agnostic predictions, i.e., their predictions do not differentiate between different types of DNAs. There are a handful of the protein-level predictors that tackle prediction of proteins that interact specifically with ssDNA and dsDNA [12, 53–56]. However, these methods predict the DNA type for the known DNA-binding proteins (i.e., they assume that the input protein binds DNA), do not differentiate between different types of dsDNA, and do not identify the DNA-binding residues. The current lack of the residue-level methods that address DNA-type-specific predictions is a substantial downside, given that the knowledge of the interacting DNA type provides useful functional clues [11, 12]. However, an accurate prediction of interactions with specific DNA types is rather challenging given that the existing methods struggle to differentiate between even more distinct partner type, such as DNA, RNA, protein and small molecules.

Motivated by the lack of suitable tools to solve these problems (i.e., DNA-agnostic prediction and cross-predictions), we introduce the first predictor of **DNA**-type-specific bindi**n**g residues i**n** prote**i**n sequ**e**nces, **DNAgenie**. We compile and share a new dataset that covers carefully curated A-DNA, B-DNA and ssDNA-interacting proteins as well as proteins that bind other partners to facilitate addressing the cross-predictions. DNAgenie combines a custom-designed machine learning architecture and a comprehensive physiochemical profile extracted from the input protein sequence to accurately predict A-DNA, B-DNA and ssDNA-binding residues. At the coarse-grained level, these predictions can be used to identify A-DNA-, D-DNA- and ssDNA-binding proteins, proteins that do not interact with DNA, as well as to differentiate between DNA types for the known DNA-binding proteins. Correspondingly, we compare the ability of DNAgenie to identify the DNA type for the DNA-binding proteins with the recently released best-performing tool [56]. Moreover, we use DNAgenie to produce and analyze putative A-DNA, B-DNA and ssDNA-binding proteins and residues in the human proteome.

Methods

Datasets

The training and benchmarking of DNAgenie's predictive model require a high-quality dataset of proteins that are experimentally annotated for interactions with a broad range of ligands. The annotations of the protein–DNA interactions serve as the ground truth to train and test predictive models. The annotations of the interactions with the other ligands (RNA, proteins and small molecules) are necessary to train a model that can accurately separate them from the protein–DNA interactions and to quantify the cross-predictions. We curated the data for the DNA-binding proteins from Protein Data Bank (PDB) [22]. First, we collected high-quality structures (resolution <3 Å) of proteins in complex with DNA. We ensure that the DNA chain is sufficiently long to determine DNA type, i.e., we reject complexes

where the DNA sequence <15 nucleotides long. Using geometry of the DNA molecule we identify 123 protein-ssDNA complexes, 185 protein-A-DNA complexes, and 954 protein-B-DNA complexes. Next, we map the PDB chains of these DNA-binding proteins into the corresponding full UniProt [57] sequences using SIFTS [58] to comprehensively annotate interactions with DNA and the other partners. Using the SIFTS's data, we identify other PDB chains that map to a given UniProt sequence. We combine the corresponding residue-level annotations of interactions extracted using the BioLip database [59] across these chains. This way we map experimental annotations of the binding residues from potentially multiple structures and from across multiple ligand types (including DNA-binding residues) onto a given UniProt sequence. This procedure, which was used in several recent studies [20, 52, 60], was shown to produce about 27% more complete coverage of the interactions compared to earlier works that use a single complex to annotate interactions [20].

We also curate a set of proteins that bind non-DNA partners. First, we select a clustered (to 30% sequence similarity) set of high-quality protein structure (resolution <3 Å) from PDB that do not interact with DNA. Like for the DNA-binding proteins, we map the PDB chains into the corresponding UniProt sequences with SIFTS. Next, we remove proteins that could bind DNA based on the information in UniProt, i.e., we eliminate proteins with 'transcription factor' and 'DNA binding' keywords and annotations. We comprehensively annotate residue-level interactions for the remaining proteins using the abovementioned approach and BioLip. We select a subset of these proteins at random to match the number of the annotated DNA-binding proteins. Finally, we cluster the resulting combined set of the DNA-binding and the non-DNA-binding proteins at 30% similarity using Blastclust to divide these data into training and test datasets. We place 70% of the resulting clusters into the training dataset, which we use to compute and optimize machine learning models, and the remaining 30% into the test dataset, which we use to empirically and comparatively evaluate the optimized models. This protocol ensures that the similarity between the training and test proteins is below 30%. We show a detailed breakdown of these datasets in Supplementary Table S1. These datasets are available at http://biomine.cs.vcu.edu/servers/DNAgenie/.

We use a recently released RNA-T benchmark dataset which includes 17 well-annotated RNA-binding proteins, utilizing the mapping protocol described above [20]. This dataset facilitates measurement of the cross-predictions among the RNA-binding residues. This is motivated by the fact that some of the RNA structures are similar to the A-DNA structure. RNA-T includes 409 RNA-binding residues and 5867 non-binding residues. The similarity of proteins in this dataset is below 30% when compared to the training dataset.

Assessment criteria

DNAgenie produces six predictions for every residue in the input protein sequence: three real-valued propensities that quantify likelihood that a given residue binds A-DNA, B-DNA and ssDNA; and three binary scores that categorize a given residue as A-DNA, B-DNA, ssDNA or non-DNA-binding. The binary predictions are produced by thresholding propensities, i.e., residues with propensities above a threshold are assumed binding, and otherwise they are assumed non-binding. Since the numbers of A-DNA-, B-DNA- and ssDNA-binding residues are much smaller compared to the non-binding residues (i.e., the data are highly imbalanced), some of the popular metrics, such as accuracy, should not be employed since their values are biased by the imbalance. We quantify the quality of the binary predictions using sensitivity (rate of correct predictions among the native binding residues) that is measured using thresholds that are set to maintain specific low values of false positive rate (FPR) at 5, 10 and 20%, which is equivalent to specificity of 95, 90 and 80%, respectively. This facilitates side-by-side comparisons between different predictors for each of the three thresholds. Moreover, we assess the cross-prediction using several metrics that were introduced in recent studies [20, 29, 52, 60, 61] including RatioCPR-D (ratio of the cross-prediction rate for DNA to sensitivity), RatioCPR-L (ratio of the cross-prediction rate for the other ligands to sensitivity), RatioOPR (ratio of the over prediction rate to sensitivity) using the binarity predictions that rely on the 5% FPR threshold (specificity = 0.95). The values of these ratios >1 suggest that a given predictor produces proportionally more correct than incorrect predictions, while ratios ≤1 mean that its outputs are at the level of a random predictor or worse. Inspired by related studies [20, 43, 51, 60–62], we assess the propensities with the commonly used AUC (area under the ROC curve), AULCratio, AUCPC-D, AUCPC-L and AUOPC. Larger values of the latter three measures (AUCPC-D, AUCPC-L and AUOPC) correspond to worse predictions, meaning predictions that are characterized by higher amounts of the cross/over-predictions. We provide detailed definitions of the above metrics in the Supplement.

DNAgenie model

DNAgenie employs a custom-designed two-layer architecture where the predictions generated by machine learning (ML) models in the first layer are refined in the second layer to reduce the cross-predictions (Figure 1).

The first layer includes four color-coded ML models that predict real-valued residue-level propensities for binding with A-DNA, B-DNA, ssDNA and with a collection of other partners that includes proteins, RNA and small molecules. The high predictive performance of this layer stems from the use of the comprehensive physiochemical profile that we produce from the input protein sequence. This profile covers a wide range of characteristics that are relevant to the protein–ligand interactions including relative solvent accessibility (RSA), intrinsic disorder and secondary structure that are predicted directly from the sequence, relative amino acid-level propensities (RAAP) that quantify tendency of amino acids to bind specific ligand types (RNA, DNA and proteins), evolutionary conservation (ECO), and biophysical properties including hydrophobicity, polarity and charge [20]. For instance, literature suggests that the DNA-binding residues are conserved, locate on the protein surface, and that certain amino acids are more likely to interact with DNA [20, 53, 63]. We predict RSA, intrinsic disorder and secondary structure directly from an input protein chain using ASAquick [64], IUPred2A [65] and the single-sequence version of PSIPRED [66], respectively. We select these methods based on their high predictive performance and low runtime. We develop the RAAP values for the A-DNA, B-DNA and ssDNA binding using a recently published approach [20]; we discuss these novel features in the Section A-DNA, B-DNA and ssDNA interaction indices. We generate the evolutionary conservation scores from the alignment profiles produced by fast HHblits tool [67]. We compute hydrophobicity, polarity and charge using indices from the AAindex database [68]. We process this comprehensive profile using a sliding-window approach to produce 423 features that quantify the respective characteristics individually and in combination with each other (e.g. we quantify the number of charged and conserved residues on the

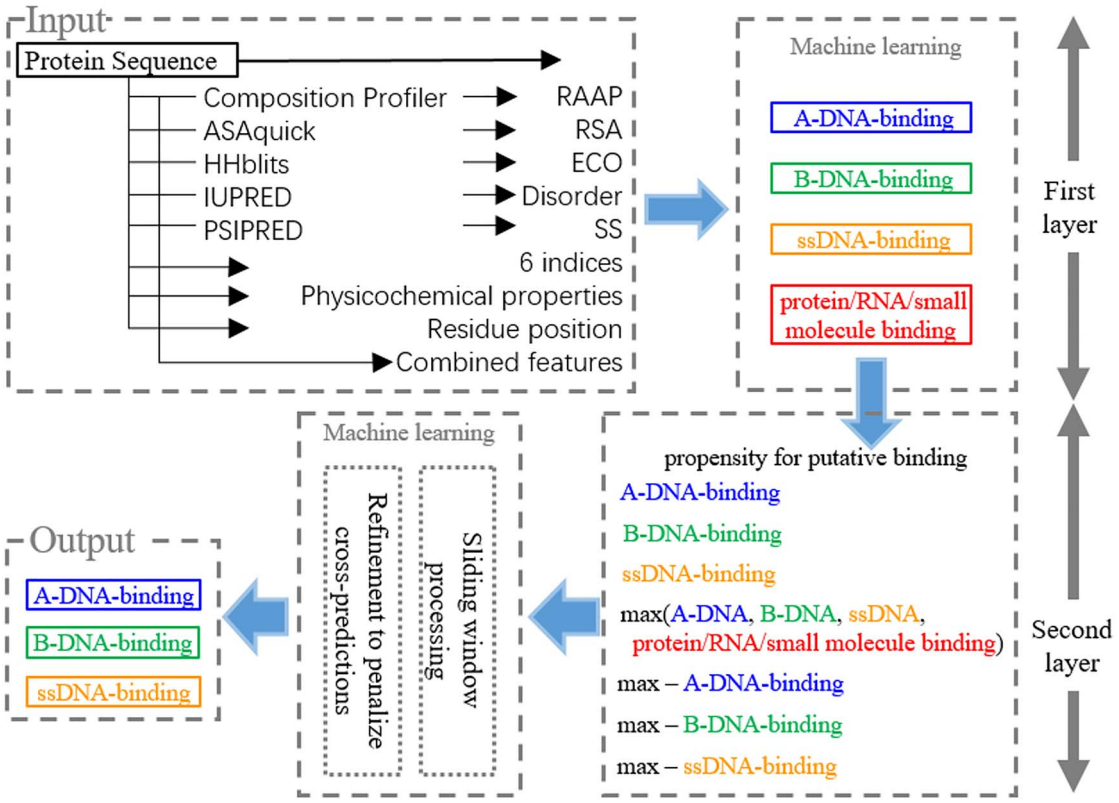

Figure 1. Architecture of DNAgenie. The input box denotes the physiochemical profile derived directly from the input protein sequence that covers relative amino acid propensities (RAAP) for binding, relative solvent accessibility (RSA), intrinsic disorder (Disorder), secondary structure (SS) and evolutionary conservation (ECO) and several other relevant biophysical properties. Seven machine learning models used by DNAgenie are denoted by color-coded boxes including the four models in the first layer that generate unrefined putative residue-level propensities for A-DNA, B-DNA, ssDNA and other (protein/RNA/small molecule) binding and the three models in the second layer that generate the refined putative residue-level propensities for A-DNA, B-DNA and ssDNA binding.

predicted surface). We detail these features in Supplementary Table S2.

The input for the second layer consists of the four predictions for the A-DNA, B-DNA, ssDNA and RNA/protein/small molecule binding passed from the first layer. Supplementary Table S3 describes features that are encoded from these four predictions, which we use as the input to the three color-coded ML models in the second layer. The second layer refines A-DNA, B-DNA and ssDNA predictions to minimize the cross-predictions. The cross predictions are reduced by comparing the unrefined putative residue-level propensities for the A-DNA, B-DNA and ssDNA interactions against each other and against the putative propensities for protein/RNA/small molecule binding. In other words, the refined residue-level propensities for A-DNA, B-DNA and ssDNA binding generated by DNAgenie can be seen as the cross-prediction reduced versions of the unrefined propensities generated in the first layer.

We consider five popular ML algorithms to train predictive model in both layers: logistic regression, weighted *k*-nearest-neighbor (*k*NN), Naïve Bayes, random forest and support vector machine. We motivate this selection by the successful use of these algorithms in related studies [20, 34, 43, 69–74]. However, in contrast to the past designs, we incorporate several innovative ideas including the use of the second/refinement layer that aims to reduce the cross-predictions, utilization of the RNA/protein/small molecule binding predictor in the first layer that facilitates the refinement, development and use of the novel RAAP for A-DNA, B-DNA and ssDNA binding, and consideration of a broad range of ML algorithms to develop the predictive model. We do not utilize the nowadays popular deep learning models since the amount of the training data is insufficient for their training (Supplementary Table S1), which would likely lead to overfitting. We use an empirical approach to adapt the predictive models trained with these five algorithms to the prediction of the specific DNA types and to maximize their predictive performance. We explore two-dimensional search space defined by empirical feature selection and selection of the ML algorithms. Feature selection aims to select a subset of the considered features that share low mutual correlation (i.e., which do not duplicate each other) and which are predictive for a specific DNA type. For each DNA type, we use wrapper feature selection with the best-first search [75] to select the best-performing subset of non-redundant features for each of the five ML algorithms. This allows us to adapt the same input profile to build accurate and selective models that predict residues that interact with A-DNA, B-DNA and ssDNA. The entire empirical design process relies exclusively on the 5-fold cross-validation tests on the training dataset. More precisely, we parametrize the models for first layer and the second layers inside the same cross-validation loop, ensuring that we do not overfit this dataset.

We compare results produced by the five ML algorithms on the training dataset in Supplementary Table S4. The results show that the support vector machine outperforms the other algorithms for the prediction of the residues that interact with each of the three DNA types. Support vector machine secures the highest AUC and AUCLratio scores coupled with the lowest/best

Table 1. Relative amino acid propensities (RAAP) for binding A-DNA, B-DNA and ssDNA

Amino Acid Type	Propensity for A-DNA binding	Propensity for B-DNA binding	Propensity for ssDNA binding
A	0.03	0.10	0.03
R	1.00	1.00	1.00
N	0.50	0.38	0.35
D	0.10	0.06	0.19
C	0.18	0.08	0.00
Q	0.35	0.36	0.14
E	0.09	0.03	0.06
G	0.22	0.22	0.34
H	0.56	0.56	0.50
I	0.14	0.09	0.22
L	0.06	0.00	0.10
K	0.76	0.76	0.56
M	0.15	0.18	0.15
F	0.19	0.19	0.38
P	0.00	0.15	0.08
S	0.34	0.34	0.18
T	0.36	0.50	0.36
W	0.62	0.35	0.62
Y	0.38	0.62	0.76
V	0.08	0.14	0.09

AUCPC-D, AUCPC-L and AUOPC values. This suggests that this ML model provides the most accurate predictions of the DNA-binding residues and the lowest rates of the cross-predictions. The support vector machine relies on the popular radial basis function kernel and we tune its hyperparameters C (complexity coefficient) and γ (width of the kernel function) using grid search where the parameter values are expressed as 2^x and $x = -10, -9, -8, \ldots, 10$. Consequently, we use the support vector machine models to implement DNAgenie.

These optimized models encapsulate relations between the selected physiochemical characteristics and DNA-binding, circumventing the need to use sequence alignment or homology. This means that DNAgenie can be used to predict virtually any protein sequence, irrespective of its similarity to other proteins, which we demonstrate empirically on the test dataset.

A-DNA, B-DNA and ssDNA interaction indices

We develop three new relative amino acid propensity (RAAP) indices which quantify likelihood that a given amino acid interacts with A-DNA, B-DNA and ssDNA. We follow a recent approach that has produced similar indices for binding to RNA, (type-agnostic) DNA and proteins [20]. First, we use Composition Profiler [76] to compute relative amino acid propensity for a specific DNA type by contrasting the corresponding set of DNA-binding residues against the non-DNA-binding residues collected from the training dataset. Next, we normalize these propensities across the three DNA types by first scaling them to the unit range and adjusting the scaled scores based on ranked averages across the DNA types. We list the resulting indices in Table 1.

We empirically test ability of these indices to identify residues that interact with A-DNA, B-DNA and ssDNA. For a set of training proteins that interact with a given DNA type, e.g., A-DNA-binding proteins, we compute differences of the A-DNA, B-DNA and ssDNA index values between the residues that interact with A-DNA and the remaining residues in the sequence. We compare these three differences and we mark a given protein as correctly predicted if the difference for the A-DNA index (the index for the selected DNA type) is higher than the difference for the other indices. This corresponds to a result where the A-DNA index is successful in marking the binding residues with the correct DNA type. We perform this test for the training proteins sets that bind A-DNA, B-DNA and ssDNA and summarize these results in Supplementary Table S5. The A-DNA index correctly predicts 70% of the A-DNA-binding proteins compared to only 23 and 7% of proteins that are incorrectly recognized as B-DNA and ssDNA; the corresponding rate of improvement over the second most common outcome is 70/23 = 3.04. Similarly, the B-DNA index correctly finds 54% of B-DNA-binding proteins, with the rate of improvement 54/31 = 1.74, while the ssDNA index marks 58% of the ssDNA proteins correctly, with the rate of improvement 58/24 = 2.42. These empirical results demonstrate that the three indices differentiate between the residues that interact with A-DNA, B-DNA and ssDNAs. We use these indices as one of the key innovative elements in the physiochemical profile utilized by DNAgenie and to adapt the current DNA type agnostic predictors of DNA-binding residues to predict interactions with A-DNA, B-DNA and ssDNA.

Results

Comparative assessment of the predictions of the A-DNA, B-DNA and ssDNA-binding residues

We benchmark predictions on the independent test dataset that covers the three types of DNA-binding residues, residues that bind the other ligand types and the non-binding residues. The assessment compares predictions against the native annotations of A-DNA, B-DNA, ssDNA, RNA, protein and small molecule binding. This allows us to evaluate the quality of the residue-level A-DNA, B-DNA and ssDNA-binding predictions and to assess the extent of cross-predictions of the DNA-binding among the residues that interact with the other partner molecules. The test dataset shares low (<30%) sequence similarity to the training data that was used in the cross-validation setting to design and optimize DNAgenie. The low similarity and the exclusion of the test set during the design

Table 2. Predictive performance of DNAgenie, the random baseline, and the DNA type-augmented predictions produced by the four state-of-the-art residue-level predictors of DNA-binding residues. We assess robustness of the predictive quality to different datasets by performing 10 tests on randomly selected 50% proteins from the test dataset. We report the corresponding averages and standard deviations. Statistical significance of differences in the predictive performance between DNAgenie and each of the other five predictors is quantified with the t-test for normal measurements as tested with the Anderson-Darling test; otherwise we use the Wilcoxon rank sum test. ++ and + mean that DNAgenie is significantly better at *P*-value <0.01 and *P*-value <0.05, respectively; = means that the difference is not significant (*P*-value ≥0.05). The sensitivities are reported at 5, 10 and 20% FPR. Bold font identifies the most accurate predictor for a given metric and DNA type

DNA type	Predictors	Sensitivity at 5% FPR	Sensitivity at 10% FPR	Sensitivity at 20% FPR	AUC	AULCratio
A-DNA	Random baseline	$0.050 \pm 0.005^{++}$	$0.110 \pm 0.008^{++}$	$0.209 \pm 0.018^{++}$	$0.514 \pm 0.006^{++}$	$0.951 \pm 0.109^{++}$
	TargetDNA	$0.268 \pm 0.016^{++}$	$0.411 \pm 0.028^{++}$	$0.622 \pm 0.020^{++}$	$0.774 \pm 0.016^{++}$	$5.839 \pm 0.653^{++}$
	HybridNAP	$0.185 \pm 0.025^{++}$	$0.322 \pm 0.017^{++}$	$0.493 \pm 0.019^{++}$	$0.702 \pm 0.014^{++}$	$4.084 \pm 0.850^{++}$
	BindN+	$0.196 \pm 0.035^{++}$	$0.361 \pm 0.029^{++}$	$0.527 \pm 0.034^{++}$	$0.722 \pm 0.014^{++}$	$4.633 \pm 0.995^{++}$
	DNApred	$0.310 \pm 0.021^{++}$	$0.464 \pm 0.026^{++}$	$0.650 \pm 0.019^{++}$	$0.789 \pm 0.017^{++}$	$7.196 \pm 0.801^{++}$
	DNAgenie	**0.483 ± 0.058**	**0.676 ± 0.026**	**0.831 ± 0.052**	**0.886 ± 0.037**	**11.896 ± 1.163**
B-DNA	Random baseline	$0.052 \pm 0.004^{++}$	$0.099 \pm 0.005^{++}$	$0.201 \pm 0.011^{++}$	$0.507 \pm 0.006^{++}$	$0.998 \pm 0.082^{++}$
	TargetDNA	$0.326 \pm 0.014^{++}$	$0.464 \pm 0.015^{++}$	$0.636 \pm 0.018^{++}$	$0.794 \pm 0.009^{++}$	$8.200 \pm 0.490^{++}$
	HybridNAP	$0.219 \pm 0.012^{++}$	$0.357 \pm 0.013^{++}$	$0.515 \pm 0.015^{++}$	$0.716 \pm 0.008^{++}$	$5.225 \pm 0.401^{++}$
	BindN+	$0.270 \pm 0.014^{++}$	$0.383 \pm 0.012^{++}$	$0.567 \pm 0.015^{++}$	$0.746 \pm 0.009^{++}$	$6.691 \pm 0.552^{++}$
	DNApred	$0.354 \pm 0.016^{++}$	$0.529 \pm 0.020^{++}$	$0.663 \pm 0.015^{++}$	$0.811 \pm 0.009^{++}$	$9.350 \pm 0.536^{++}$
	DNAgenie	**0.472 ± 0.044**	**0.644 ± 0.041**	**0.824 ± 0.037**	**0.884 ± 0.015**	**11.156 ± 1.161**
ssDNA	Random baseline	$0.047 \pm 0.004^{++}$	$0.108 \pm 0.023^{++}$	$0.219 \pm 0.026^{++}$	$0.502 \pm 0.005^{++}$	$0.933 \pm 0.093^{++}$
	TargetDNA	$0.193 \pm 0.016^{++}$	$0.351 \pm 0.052^{++}$	$0.560 \pm 0.039^{++}$	$0.757 \pm 0.025^{++}$	$3.999 \pm 0.563^{++}$
	HybridNAP	$0.142 \pm 0.014^{++}$	$0.245 \pm 0.034^{++}$	$0.454 \pm 0.039^{++}$	$0.683 \pm 0.018^{++}$	$2.729 \pm 0.505^{++}$
	BindN+	$0.153 \pm 0.025^{++}$	$0.281 \pm 0.051^{++}$	$0.491 \pm 0.044^{++}$	$0.709 \pm 0.025^{++}$	$3.346 \pm 0.425^{++}$
	DNApred	$0.213 \pm 0.039^{++}$	$0.412 \pm 0.051^{++}$	$0.576 \pm 0.045^{++}$	$0.774 \pm 0.027^{++}$	$4.816 \pm 0.723^{++}$
	DNAgenie	**0.487 ± 0.063**	**0.691 ± 0.088**	**0.850 ± 0.066**	**0.907 ± 0.018**	**16.581 ± 2.509**

process ensure that the measured performance reflects values that are expected when DNAgenie is applied on proteins for which sequence alignment or homology could not produce accurate results.

Since DNAgenie is the first method that predicts A-DNA, B-DNA and ssDNA-binding residues, we compare it against a baseline implemented as a random-level predictor and the closest alternatives represented by a curated selection of state-of-the-art sequence-based predictors of DNA-binding residues: BindN+ [40], TargetDNA [42], hybridNAP [20] and DNApred [44]. These tools satisfy three selection criteria: availability as webservers or standalone software; fast predictions (under 10 minutes for an average size protein chain); and recent release, with the exception of the older and popular BindN+. We adapt their DNA type-agnostic predictions to cover the three DNA types by using A-DNA, B-DNA and ssDNA interaction indices that we devise using an approach described in a recent study [20]. Briefly, each index quantifies propensities of the 20 amino acids for interaction with a specific DNA type, reflecting compositional differences between the DNA-type-specific binding residues and residues that do not bind DNA. We multiply the original DNA type agnostic predictions by the indices to secure the three DNA-type-specific predictions for each of the four current predictors. This improves these predictions compared to using the original DNA type-agnostic prediction for each of the three DNA types. We demonstrate that empirically in the Section A-DNA, B-DNA and ssDNA interaction indices.

Table 2 quantifies predictive performance of DNAgenie and compares it with the baseline and the four alternatives. Results show that DNAgenie provides very accurate predictions across the three DNA types, with AUCs ranging between 0.88 (for B-DNA) and 0.91 (for ssDNA); Supplementary Figure S1A gives the corresponding ROC curves. The same is true based on the other metrics including AULCratio that quantifies the ratio of the measured AUC scores to the AUC scores of a random predictor for conservative predictions where the amount of the predicted DNA-binding residues does not exceed the amount of the native DNA-binding residues. Per this definition, the AULCratio values for the random baseline are around 1 while higher values denote more accurate results. DNAgenie secures AULCratio scores that span between 11.16 (for B-DNA) and 16.58 (for ssDNA), which corresponds to 1116 and 1658% improvement over the baseline, respectively. Moreover, sensitivity values of DNAgenie computed based on the conservative scenario with low 5% false positive rate (specificity = 95%) equal 48, 47 and 49% for the A-DNA, B-DNA and ssDNA binding. In other words, nearly half of the DNA-binding residues are correctly predicted at this low false-positive rate. The sensitivity values increase to the range between 0.64 to 0.69 when FPR is set to 10% and further increase to the range between 0.82 and 0.85 when FPR is set to 20%. The relation between sensitivity and specificity values is expressed by the ROC curves shown in Supplementary Figure S1A. We note that DNAgenie provides the best sensitivity values for the same specificity when compared to all other approaches, i.e., its ROC curves are above the other curves by a wide margin. Overall, Table 2 reveals the DNAgenie's results for the three DNA types that we quantify with multiple metrics are statistically significantly better than the baseline and the predictions generated by the four current augmented DNA type-agnostic predictors (*P*-value <0.01). The best of these methods, DNApred, obtains AUCs of 0.79, 0.81 and 0.77 for the prediction of the A-DNA, B-DNA and ssDNA interactions, respectively.

Analysis and evaluation of the cross-predictions

We empirically analyze the cross-predictions among the residues that interact with different DNA types and among the residues that bind non-DNA ligands (proteins, RNA and small molecules). Inspired by related works [29, 52], we quantify the cross-predictions among the DNA-binding residues with the

Table 3. Assessment of cross predictions generated by DNAgenie, the random baseline, and the DNA type-augmented predictions produced by the four state-of-the-art residue-level predictors of DNA-binding residues. Lower values of the AUCPCs and higher ratio values denote more accurate predictions (lower amount of cross-predictions). We assess robustness to different datasets by performing 10 tests on randomly selected 50% proteins from the test dataset. We report the corresponding averages and standard deviations. Statistical significance of differences in the predictive performance between DNAgenie and each of the other five predictors is quantified with the t-test for normal measurements as tested with the Anderson-Darling test; otherwise we use the Wilcoxon rank sum test. ++ and + mean that DNAgenie is significantly better at *P*-value <0.01 and *P*-value <0.05, respectively; = means that the difference is not significant (*P*-value ≥0.05). The binary assessments (CPR-D, CPR-L and OPR) are normalized between different predictors to maintain the same 5% FPR (specificity = 0.95). Bold font identifies the most accurate predictor for a given metric and DNA type

DNA type	Predictors	AUCPC-D	RatioCPR-D at 5% FPR	AUCPC-L	RatioCPR-L at 5% FPR	AUOPC	RatioOPR at 5% FPR
A-DNA	Random	$0.503 \pm 0.016^{++}$	$1.055 \pm 0.241^{++}$	$0.483 \pm 0.014^{++}$	$1.076 \pm 0.215^{++}$	$0.488 \pm 0.010^{++}$	$1.055 \pm 0.213^{++}$
	TargetDNA	$0.517 \pm 0.021^{++}$	$0.928 \pm 0.098^{++}$	$0.278 \pm 0.019^{++}$	$4.281 \pm 0.485^{++}$	$0.220 \pm 0.015^{++}$	$5.991 \pm 0.375^{++}$
	HybridNAP	$0.504 \pm 0.018^{++}$	$0.905 \pm 0.133^{++}$	$0.345 \pm 0.015^{++}$	$3.091 \pm 0.566^{++}$	$0.293 \pm 0.014^{++}$	$3.924 \pm 0.504^{++}$
	BindN+	$0.520 \pm 0.023^{++}$	$0.786 \pm 0.169^{++}$	$0.319 \pm 0.016^{++}$	$3.058 \pm 0.609^{++}$	$0.272 \pm 0.014^{++}$	$4.265 \pm 0.770^{++}$
	DNApred	$0.513 \pm 0.024^{++}$	$1.002 \pm 0.096^{++}$	$0.260 \pm 0.020^{++}$	$5.264 \pm 0.480^{+}$	$0.204 \pm 0.016^{++}$	$6.887 \pm 0.433^{++}$
	DNAgenie	**0.317 ± 0.051**	**2.007 ± 0.376**	**0.152 ± 0.037**	**7.295 ± 2.277**	**0.110 ± 0.025**	**10.584 ± 1.336**
B-DNA	Random	$0.492 \pm 0.015^{++}$	$0.944 \pm 0.119^{++}$	$0.504 \pm 0.007^{++}$	$1.106 \pm 0.107^{++}$	$0.493 \pm 0.006^{++}$	$1.021 \pm 0.076^{++}$
	TargetDNA	$0.454 \pm 0.014^{++}$	$1.246 \pm 0.077^{++}$	$0.248 \pm 0.014^{++}$	$4.797 \pm 0.698^{=}$	$0.203 \pm 0.009^{++}$	$6.785 \pm 0.334^{++}$
	HybridNAP	$0.472 \pm 0.011^{++}$	$1.210 \pm 0.126^{++}$	$0.319 \pm 0.018^{++}$	$3.531 \pm 0.548^{++}$	$0.282 \pm 0.008^{++}$	$4.469 \pm 0.257^{++}$
	BindN+	$0.457 \pm 0.018^{++}$	$1.438 \pm 0.251^{++}$	$0.281 \pm 0.016^{++}$	$3.996 \pm 0.595^{=}$	$0.251 \pm 0.009^{++}$	$5.520 \pm 0.281^{++}$
	DNApred	$0.455 \pm 0.016^{++}$	$1.150 \pm 0.056^{++}$	$0.230 \pm 0.014^{++}$	$5.786 \pm 0.805^{++}$	$0.187 \pm 0.009^{++}$	$7.373 \pm 0.364^{++}$
	DNAgenie	**0.291 ± 0.027**	**2.374 ± 0.550**	**0.172 ± 0.017**	**4.343 ± 0.567**	**0.102 ± 0.012**	**9.864 ± 0.925**
ssDNA	Random	$0.517 \pm 0.005^{++}$	$1.225 \pm 0.396^{++}$	$0.515 \pm 0.004^{++}$	$1.146 \pm 0.328^{++}$	$0.515 \pm 0.003^{++}$	$1.140 \pm 0.274^{++}$
	TargetDNA	$0.552 \pm 0.031^{++}$	$0.680 \pm 0.111^{++}$	$0.297 \pm 0.029^{++}$	$2.829 \pm 0.547^{++}$	$0.234 \pm 0.025^{++}$	$4.366 \pm 0.655^{++}$
	HybridNAP	$0.525 \pm 0.017^{++}$	$0.708 \pm 0.091^{++}$	$0.355 \pm 0.021^{++}$	$2.404 \pm 0.570^{++}$	$0.312 \pm 0.018^{++}$	$2.938 \pm 0.492^{++}$
	BindN+	$0.536 \pm 0.027^{++}$	$0.707 \pm 0.111^{++}$	$0.329 \pm 0.026^{++}$	$2.409 \pm 0.430^{++}$	$0.285 \pm 0.025^{++}$	$3.308 \pm 0.527^{++}$
	DNApred	$0.549 \pm 0.033^{++}$	$0.689 \pm 0.113^{++}$	$0.278 \pm 0.032^{++}$	$3.525 \pm 0.644^{++}$	$0.217 \pm 0.027^{++}$	$5.038 \pm 0.848^{++}$
	DNAgenie	**0.154 ± 0.025**	**4.872 ± 0.807**	**0.161 ± 0.033**	**4.689 ± 0.961**	**0.097 ± 0.022**	**10.064 ± 1.195**

area under the cross-prediction curve for DNA (AUCPC-D) and cross-prediction rate for DNA (CPR-D). These two measures quantify the extent to which a given type of DNA binding is predicted among the residues that bind the other two types of DNA. Lower values of area correspond to fewer cross-predictions. To ease interpretation, the cross-prediction rate is computed as a ratio (RatioCPR-D) between the number of the cross predictions and the sensitivity (the number of correct predictions for a given DNA type). This way random-level predictions have ratio = 1, with higher values denoting the rate of improvement over the random baseline. We also assess the cross-predictions among the residues that interact with the other ligands (proteins, RNA and small molecules) based on two metrics, AUCPC-L and RatioCPR-L. Finally, we evaluate the rate of incorrect predictions among the non-binding residues with AUOPC (area under the over-prediction rate curve) and RatioOPR (ratio of over-prediction rate among the non-binding residues and sensitivity). We define these metrics in the Section Assessment criteria.

Table 3 reveals that DNAgenie produces minimal amounts of cross-predictions across the three DNA types. On average, over the three DNA types, DNAgenie secures RatioCPR-D = 3.08, RatioCPR-L = 5.44 and RatioOPR = 10.17. The fact that RatioCPR-D < RatioCPR-L means that the amount of the cross-predictions between the three DNA types is larger compared to the cross-predictions of DNA binding among residues that interact with the other molecules. This is expected given that the different DNA types are much more similar to each other compared to the similarity between the DNA and the other molecules (proteins, RNA and small molecules). However, even for the most challenging case, the corresponding RatioCPR-D shows that DNAgenie is 308% better than the baseline. To compare, the best alternative, DNApred, substantially cross-predicts between the three DNA types (average RatioCPR-D = 0.95, which is at the level of the baseline) and produces more cross predictions among the residues that bind the other partners (average RatioCPR-L = 4.86). The cross-prediction curves shown in the Supplementary Figure S1B (for DNA-binding residues), S1C (for the other ligand types) and S1D (for the non-binding residues) reveal a large margin of improvement between the DNAgenie's curves and the curves of the other methods. These plots demonstrate that the improvements in the DNAgenie's cross-prediction rates are consistent over the entire range of the sensitivity values. Moreover, Table 3 shows that the corresponding AUCPC-D, AUCPC-L and AUOPC scores produced by DNAgenie are significantly better when compared with the other methods and for each DNA type (*P*-value <0.01).

Empirical analysis also reveals that the major reason for the low amounts of the cross-predictions is the use of the second (refinement) layer in the DNAgenie's model. While the first layer's models achieve the average AUCPC-D (area under the curve that quantifies rate of the incorrect predictions of DNA binding across DNA types) = 0.29, this area shrinks to 0.25 (16% improvement) after the refinement in the second layer. When broken by the DNA type, AUCPC-D decreases from 0.35 to 0.32 for A-DNA (*P*-value = 0.15), from 0.33 to 0.29 for B-DNA (*p*-value <0.05) and from 0.19 to 0.15 for ssDNA (*P*-value <0.05). This means that the use of the second layer provides consistent improvements, over the three DNA types, which in case of B-DNA and ssDNA are also statistically significant. By contrast, DNApred yields the average AUCPC-D = 0.51.

Assessment of the cross-predictions in RNA-binding proteins

We analyze the cross-predictions among the RNA-binding proteins since the A-form of DNA is structurally similar to some

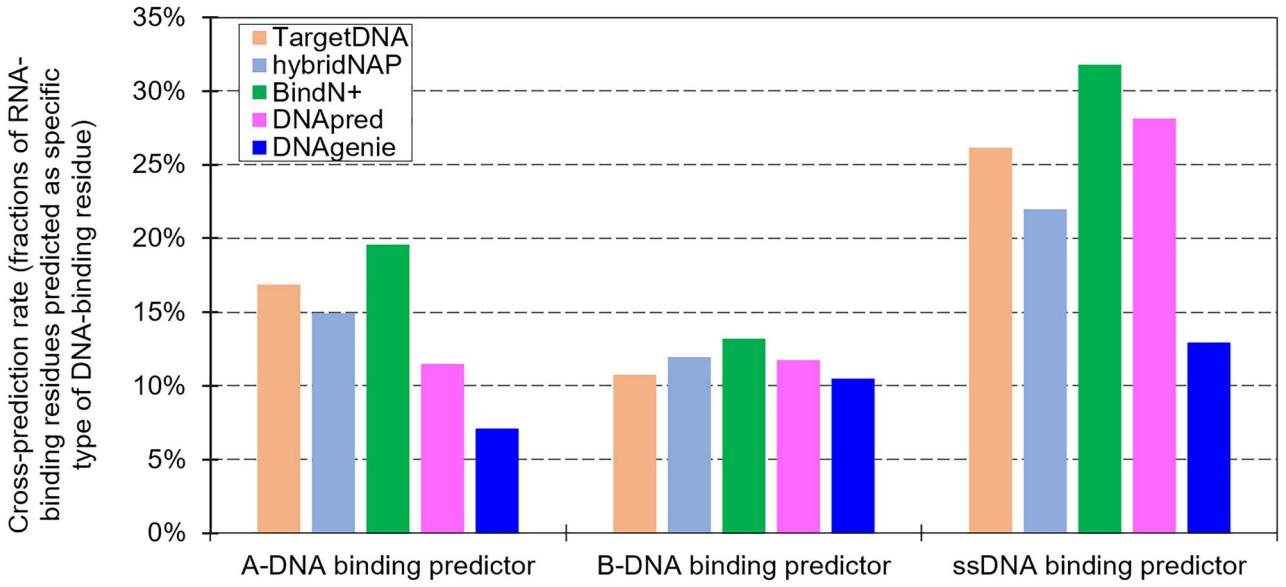

Figure 2. Comparison of the cross-predictions rates among the RNA binding residues (i.e., fraction of the RNA-binding residues predicted as A-DNA, B-DNA and ssDNA-binding residues) in the RNA-T dataset. Similar to Table 2, the predictions are normalized between different predictors to maintain the same 5% FPR (specificity = 0.95) on the test dataset.

of the RNA structures. Moreover, recent work demonstrates that predictors of DNA-binding residues cross-predict over 20% of RNA-binding residues as DNA-binding [43]. Figure 2 illustrates the cross-prediction rates on the recently published RNA-T benchmark dataset, i.e., the rate of the prediction of A-DNA, B-DNA and ssDNA residues among the native RNA-binding residues. When considering predictions of the A-DNA-binding residues, TargetDNA, hybridNAP and BindN+ cross-predicts over 15% of the RNA-binding residues as A-DNA-binding residues while DNApred and DNAgenie yield the lowest/best rates at about 11.5 and 7.1%, respectively. On average, across the three DNA types, DNAgenie obtains the lowest cross prediction rate at 10.2%, compared to 16.3% for hybridNAP, 17.1% for DNApred, 17.9% for TargetDNA and 21.5% for BindN+. These are relatively low rates given that DNAgenie secures the average sensitivity of 48.1% (4.7 times higher compared to the cross-prediction rate) for the DNA-binding residues (Table 2). This test demonstrates that DNAgenie accurately differentiates between DNA-binding and RNA-binding residues.

Comparative assessment of the predictions of dsDNA and ssDNA-binding proteins

Several methods are available for the coarse-grained prediction that identifies whether a given DNA-binding protein interacts with ssDNA or dsDNA [12, 53–56]. We apply the residue-level predictions of A-DNA-, B-DNA- and ssDNA-binding residues generated by DNAgenie to differentiate between the ssDNA and dsDNA partners for the DNA-binding proteins in the test dataset. We compute the propensity for the ssDNA binding at the protein level by calculating the average of the residue-level propensities for the predicted ssDNA-binding residues. Similarly, we use one minus the average of the propensities for the A-DNA- and B-DNA-binding residues to quantify the protein-level propensity for the dsDNA binding. We compare these results with the most recent protein-level predictor by Sharma and colleagues that was shown to outperform the older tools [56]. We use the author-provided implementation of this tool to collect the protein-level propensities for the ssDNA and dsDNA binding. We report results for the two best performing ensembles of three machine learning models: ensemble 1 that relies on the majority-based prediction and ensemble 2 that select the model with the highest propensity [56]. For the ensemble 1, we use the propensity that is calculated as the average of the propensities of the 2 or 3 models that are in the majority, while for ensemble 2, we use the propensity of the selected model. Table 4 summarizes the results. We setup the binary predictions to the same 5, 10 and 20% FPRs, which allows us to directly compare the corresponding sensitivity values across different methods.

DNAgenie offers better coarse-grained predictions of the ssDNA- and dsDNA-binding proteins. The sensitivity computed at the 5% FPR is higher by 18% and AUC improves from 0.786 to 0.863 when compared with the best current tool. Supplementary Figure S2 gives the corresponding ROC curves. We note that DNAgenie additionally provides accurate predictions of the DNA-binding residues that are categorized by DNA type into A-DNA, B-DNA and ssDNA. In contrast, the other methods do not predict the DNA-binding residues and do not differentiate between different dsDNA types.

Case study

We showcase blind/*de novo* prediction produced by DNAgenie on one of the test proteins, human DNA methyltransferase 3A (DNMT3A). This protein shares low 4.9% similarity with the training proteins, i.e., the maximal pairwise similarity across all training proteins measured with BLAST is 4.9% [77, 78]. We emphasize that this case study is meant to illustrate DNAgenie's predictions and compare them side-by-side with the other predictors. Recently released structural details of the interaction of DNMT3A with B-DNA serve as the ground truth to assess these predictions [79]. DNAgenie's predictive quality, expressed with AUC, is similar to the average AUCs on the test dataset, representing an average/typical case. Figure 3 illustrates the 3D structure of the complex with B-DNA with the color-coded annotations of DNAgenie's predictions. The correct predictions of

Table 4. Predictive performance of DNAgenie and the ensembles 1 and 2 proposed by Sharma and colleagues for the prediction of ssDNA and dsDNA partners of the DNA-binding proteins in the test dataset. The binary assessment with sensitivity is normalized between different predictors to maintain the same 5, 10 and 20% FPR. Bold font identifies the most accurate predictor for a given metric

Protein-level predictor of ssDNA and dsDNA binding	Sensitivity at 5% FPR	Sensitivity at 10% FPR	Sensitivity at 20% FPR	AUC
Ensemble 1	0.17	0.34	0.68	0.786
Ensemble 2	0.24	0.26	0.42	0.739
DNAgenie	**0.42**	**0.44**	**0.73**	**0.863**

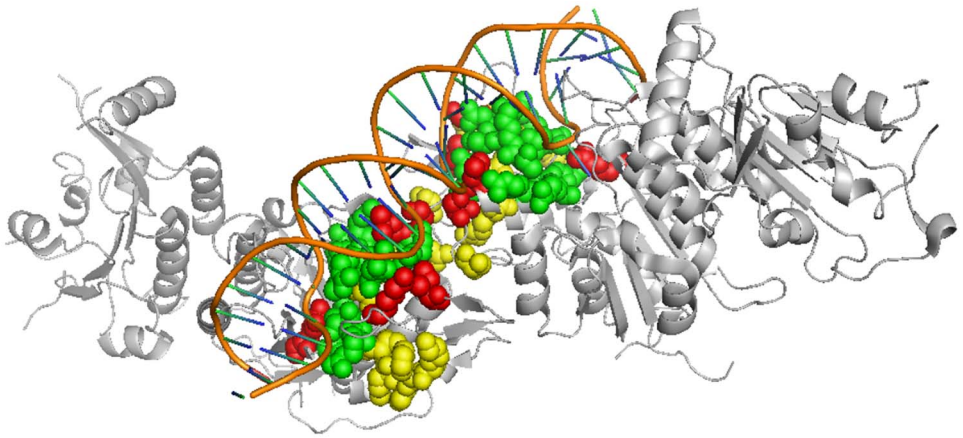

Figure 3. Structure of the human DNA methyltransferase 3A (DNMT3A) in complex with B-DNA (PDB ID: 5YX2 chain D). The protein structure is shown with the gray cartoon representation. DNA-structure is represented by the orange double helix. The residues represented using color-coded balls identify DNAgenie's predictions where green are true positives (correctly predicted B-DNA-binding residues), red are false negatives (native B-DNA-binding residues incorrectly predicted as non-B-DNA-binding residues), and yellow are false positives (native non-B-DNA-binding residues incorrectly predicted as B-DNA-binding residues).

the B-DNA-binding residues shown in green (true positives) extend along the double helix. The false positives (native non-B-DNA-binding residues incorrectly predicted as B-DNA-binding residues) marked in yellow are located nearby the interaction site. We argue that they provide useful clues, especially since the native annotations of binding residues rely on a somehow arbitrary distance-based definition, i.e., residue is defined as binding if the distance between an atom of this residue and a DNA atom is <0.5 Å $+$ the sum of the Van der Waal's radii of the two atoms [59]. The yellow residues could be marked as green if the 0.5 Å factor would increase. Supplementary Figure S3 provides side-by-side comparison of the B-DNA-binding predictions produced by DNAgenie and the DNA type-agnostic predictions generated by the four selected sequence-based predictors of DNA-binding residues. DNAgenie successfully identifies 12 of the 23 native B-DNA-binding residues, with the false positives clustered in a close proximity of the B-DNA-binding residues. The second-best DNApred correctly finds six B-DNA-binding residues, but with many false positives scattered along the sequence at positions far from the native B-DNA-binding residues. The other three tools face similar problems, with predictions distributed along the entire protein chain. This example illustrates that DNAgenie produces on average more true positives than the alternative tools, which is evident based on its high sensitivity values in Table 2, and also generates arguably more useful false positives that are localized nearby the true positives.

Prediction and analysis of A-DNA, B-DNA and ssDNA-binding residues and proteins in the human proteome

We apply DNAgenie to make predictions for the 20 350 proteins from the UniProt's reference human proteome [57]. We evaluate veracity of these predictions by computing their overlap with the list of the currently known DNA-binding proteins. First, we collect the DNA-binding proteins from ENPD, the largest database of the nucleic acid-binding proteins [80]. Given natural variations in protein sequences, we annotate human proteins from the UniProt's reference proteome as DNA-binding if they share over 90% similarity (quantified with BLAST) with any of the human DNA-binding proteins from ENPD. This results in a list of 2062 experimentally annotated DNA-binding proteins. Second, we independently use Pfam domains [81] to annotate DNA-binding proteins. We manually analyze Pfam domains in the human proteome and find 672 domains that interact with DNA. We identify 2218 human proteins that have at least one of these domains. Third, we combine the 2062 DNA-binding proteins from ENPD and the 2218 proteins that have Pfam's DNA-binding domains to establish the final set of 2763 verified DNA-binding proteins. Next, we use the putative A-DNA, B-DNA and ssDNA-binding residues generated by DNAgenie at the low 5% FPR to identify putative DNA-binding proteins. We calibrate this residue-level to protein-level prediction conversion to generate the number of putative DNA-binding proteins that is similar to the number of the verified DNA-binding proteins. We apply two conditions to define a given protein as DNA-binding. First, the fraction of putative DNA-binding residues must be higher than 10% to reduce likelihood of including spurious predictions. Second, the protein must include at least one long segment of DNA-binding residues (equivalent of a DNA-binding domain) which is composed of at least 90% of residues predicted as A-DNA-, B-DNA- or ssDNA-binding residues within a window of 15 consecutive residues. This approach generates 2778 putative DNA-binding proteins, which constitute 13.6% of the human proteome and include 1201 A-DNA-binding, 1404 B-DNA-binding, and 713 ssDNA-binding proteins.

We compare the 2778 putative DNA-binding proteins against the 2763 known DNA-binding proteins. DNAgenie predicts 529 (25.7%) of the 2062 ENPD-annotated proteins and 737 (26.5%) of the complete set of the 2763 verified DNA-binding proteins. The amount of the overlap is driven in part by the use of the low 5% FPR-based predictions, which limits their sensitivity to about 48%, as we show in Table 2. We assess statistical significance of the overlap between the predicted and the verified DNA-binding proteins by comparing the predictions with a randomized baseline. We compute overlap between a randomly selected set of 2778 human protein and the 2763 verified DNA-binding proteins to implement the baseline, and we repeat this sampling 1000 times to establish confidence intervals. The corresponding average and standard deviation for the overlap of the baseline are 13.8% $\pm$ 0.6%, with the maximum of 15.5%. The 26.5% overlap generated by DNAgenie is about two times larger than the average of the baseline and this difference is statistically significant (P-value <0.01).

We also analyze novel putative DNA-binding proteins produced by DNAgenie to investigate whether they share certain characteristics (i.e., subcellular localization and Pfam domains) that are associated with the known DNA-binding proteins. If true, that would suggest that these predictions provide informative leads to identify novel DNA-binding proteins. Correspondingly, we further analyze the 2778–737 = 2041 novel putative DNA-binding proteins produced by DNAgenie. First, we investigate whether they share subcellular location annotations that are characteristic for the verified DNA-binding proteins. Using the GO-slim analysis of the cellular component annotations in PANTHER [82], we find 42 cellular components that are statistically enriched among the verified DNA-binding proteins, when compared to the reference human proteome (P-value <0.05 using the Fisher's test with the discovery rate correction, fold enrichment >2, and 15 or more occurrences per annotation to ensure robustness of the statistics). We repeat this analysis for the novel putative DNA-binding proteins and identify 27 significantly enriched cellular components, out of which 48.1% (13 annotations) are in common with the components enriched for verified DNA binders. For context, no significantly enriched cellular component are produced when we run the same analysis for a random set of 2763 human proteins, which is equivalent to the size of the collection of the verified DNA-binding proteins. Figure 4 summarizes the cellular components that are enriched in the novel putative DNA-binding proteins. As a couple of highlights, they are found among ribosomal and mitochondrial proteins, which agrees with literature [83, 84]. Second, we examine Pfam domains present in the novel putative DNA-binding proteins. We find that 10.5% of them include at least one domain that suggests binding to DNA, such as BEX, CENP, Cyclin, MBD_C, NPIP and several types of the zinc finger domains. Overall, we discover that 35.5% of these novel DNA binders include at a minimum one of the relevant Pfam domains or is annotated with a cellular component term that is associated with the verified DNA-binding proteins. These results suggest that at least some of the novel predictions constitute promising leads to identify previously unknown DNA-binding proteins.

We list the protein IDs of the 2778 DNA-binding proteins predicted by DNAgenie in the Supplementary Table S6. We also share more detailed information at http://biomine.cs.vcu.edu/servers/DNAgenie/. The latter data includes UniProt accession numbers, sequences, predictions of the A-DNA, B-DNA and ssDNA-binding residues, markers for inclusion in ENPD, and listing of relevant Pfam domains that are categorized as either DNA-binding or likely to interact with DNA.

DNAgenie webserver

DNAgenie is publicly available as a webserver at http://biomine.cs.vcu.edu/servers/DNAgenie/; we also offer a mirror site at http://www.inforstation.com/webservers/DNAgenie/. With the user's convenience in mind, the webserver performs calculations on the server side and we allow batch predictions for up to five proteins in a single request. We encourage users to contact the authors directly in case if large-scale predictions are needed. The only required input are the FASTA-formatted protein sequences. The server outputs numeric propensities for the A-DNA, B-DNA and ssDNA-binding and the three corresponding binaries predictions (binding versus non-binding) for each amino acid in the input protein chain(s). The results are available in two convenient ways: as an HTML page-formatted report and a parseable csv file. Users have an option to provide email address where the links to the results are sent upon completion of the predictions.

Summary

Prediction of the DNA-binding residues in protein sequence is a difficult problem. The current DNA type-agnostic solutions lack in two aspects: the ability to differentiate DNA-binding residues from the residues that interact with other partners (i.e., they cross-predict residues that interact with RNAs, proteins and small molecules as DNA binding); and the ability to predict interactions with specific DNA types. DNAgenie provides the first and accurate solution to both challenges, as we demonstrate through extensive comparative empirical tests and application to the human proteome. Importantly, DNAgenie does not rely on sequence similarity or homology, which means that it provides accurate results for virtually any protein sequence. This is evident based on the results on the test dataset, which simulates a scenario where DNAgenie is used to predict sequences that share low similarity (<30%) with its training proteins.

There are multiple factors that explain high-quality of the results produced by DNAgenie. First, we utilize the training dataset that covers proteins that interact with DNA, RNA, proteins and small molecules, allowing our machine learning models to successfully learn to differentiate between different ligand types. This is in contrast to the prior tools that were trained using datasets composed solely of the DNA-binding proteins [29, 52]. Second, we represent the input protein sequence using a broad physiochemical profile that covers a comprehensive collection of relevant sequence-derived structural, evolutionary, biophysical and biochemical information. Third, we use the two-layered topology (Figure 1) where we apply the custom-designed second layer to refine the initial predictions generated by the first stage, with the objective to minimize the cross-predictions. The use of the comprehensive profile leads to a very accurate prediction of the A-DNA, B-DNA and ssDNA binding by the machine learning models from the first layer. On average, over the three DNA types, the first-layer models secure AUC = 0.890 and sensitivity = 0.47 at the low 5% FPR. After the refinement in the second layer, DNAgenie's models generate nearly identical average AUC = 0.893 and sensitivity = 0.48 (at 5% FPR). Importantly, this refinement leads to the statistically significant reduction in the cross-predictions, which we discuss in the Section Analysis and evaluation of the cross-predictions. More specifically, the average AUCPC-D of the

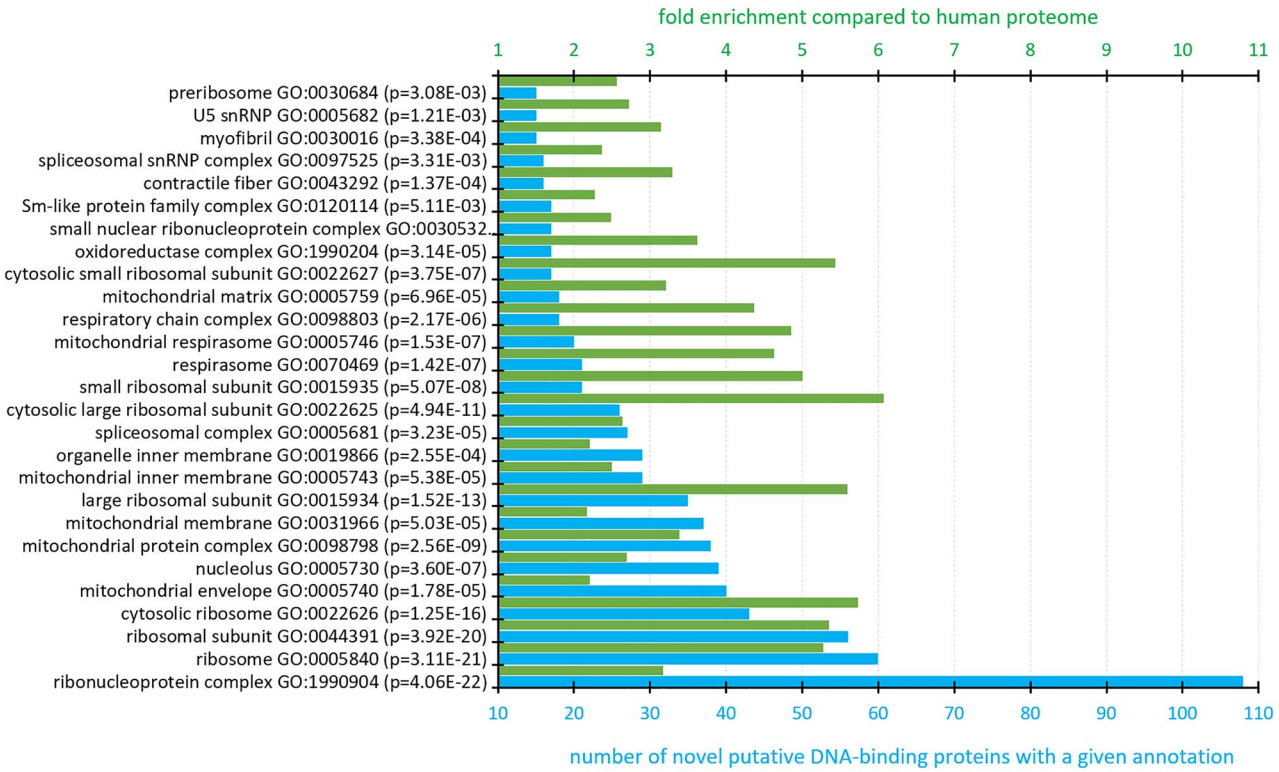

Figure 4. Cellular components that are significantly enriched among the novel putative DNA-binding proteins produced by DNAgenie. * identifies annotations that are in common with the components that are significantly enriched in the verified DNA-binding proteins. The plot is sorted by the number of proteins that have a given enriched annotation (blue bars). Analysis was performed with PANTHER where *p*-values were computed using the Fisher's test with the discovery rate correction, minimal fold enrichment is set to 2, and annotations with 15 or more occurrences are used to ensure robustness of the statistics.

first layer's models of 0.29 is improved to 0.25 (16% improvement) after the refinement. To compare, DNApred, the best current method (as shown in by its authors [44] and in our Table 2), obtains the average AUC = 0.79, the average sensitivity = 0.29 (at 5% FPR), and the average AUCPC-D = 0.51. The improvements offered by the DNAgenie over DNApred and other considered methods are statistically significant for each of the three DNA types (*P*-values <0.05). Moreover, we note that lower predictive performance of DNApred and other methods that we compare with can be explained by the fact that they were not originally designed to predict A-DNA-, B-DNA- and ss-DNA-binding residues.

We empirically demonstrate that DNAgenie's predictions on the human proteome significantly overlap with the known DNA-binding proteins while also covering several hundred novel putative DNA-binding proteins. Utilizing two sources of independent data, Pfam domains and GO annotations, we argue that some of these novel putative DNA binders are likely to interact with DNA because they harbor domains that suggest DNA binding and since they share subcellular locations that are enriched for the currently known DNA-binding proteins. The putative DNA-binding proteins should be investigated experimentally to either confirm or refute the predictions. Nowadays, a wide array of methods that include functional proteomics experiments and mass spectrometry can be used for that purpose. Some of the popular approaches include affinity purification [85], chromatin immunoprecipitation (ChIP) [86, 87] electrophoresis mobility shift assay (EMSA), and more recently CRISPR (regularly clustered interspaced palindromic repeats)-based approaches [88].

Altogether, the strong predictive performance on the test dataset coupled with the accurate predictions on the human proteins, which include numerous promising leads for novel DNA-binding proteins, demonstrate that DNAgenie is a valuable tool for computational, sequence-based characterization of protein functions. DNAgenie's webserver, training and test datasets, and predictions and annotations for the human proteins are available at http://biomine.cs.vcu.edu/servers/DNAgenie/.

We envision extending DNAgenie in two directions. The current version is limited to the three major DNA types: A-DNA, B-DNA and ssDNA. This is due to the lack of a sufficient amount of experimental data for the interactions with the other DNA types. While the current amount of data is insufficient to accurately train the predictive model and to build an adequately large and dissimilar test dataset, this is likely to change in a near future. The other interesting extension is to consider taxonomic differences in the protein–DNA interactions. DNAgenie is designed to make predictions across all domains of life, as it was trained on the dataset that covers eukaryotic (36%), bacterial (43%), archaeal (5%) and viral (16%) proteins. However, we anticipate that models optimized specifically for eukaryotic versus prokaryotic proteins would be different and would provide more accurate results when used for the taxonomically compatible proteins.

Key Points

- Sixteen methods are available for the prediction of the DNA-binding residues in protein sequences
- Current predictors of the DNA-binding residues are DNA type agnostic and significantly cross-predict residues that interact with other ligands
- DNAgenie is the first sequence-based predictor of amino acids that interact with A-DNA, B-DNA and single-stranded DNA
- DNAgenie offers accurate predictions of the DNA-binding residues, low cross-predictions rates and high-quality coarse-grained predictions of ssDNA- and dsDNA-binding proteins
- DNAgenie identifies promising leads for previously unknown DNA-binding proteins in the human proteome

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

The Robert J. Mattauch Endowment funds to L.K., and by the National Natural Science Foundation of China (grant 61802329); the Innovation Team Support Plan of University Science and Technology of Henan Province (grant 19IRTSTHN014); the Nanhu Scholars Program for Young Scholars of the Xinyang Normal University to J.Z.

References

1. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res* 2010;**38**(21):7364–77.
2. Stormo GD, Zhao Y. Determining the specificity of protein–DNA interactions. *Nat Rev Genet* 2010;**11**(11):751–60.
3. Xie Z, Hu S, Qian J, *et al*. Systematic characterization of protein-DNA interactions. *Cell Mol Life Sci* 2011;**68**(10):1657–68.
4. Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* 2015;**43**(18):8627–37.
5. Mishra SK, Tawani A, Mishra A, *et al*. G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep* 2016;**6**(1):38144.
6. Brázda V, Hároníková L, Liao J, *et al*. DNA and RNA Quadruplex-binding proteins. *Int J Mol Sci* 2014;**15**(10):17493–517.
7. Brázda V, Laister RC, Jagelská EB, *et al*. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol Biol* 2011;**12**(1):33.
8. Zeraati M, Langley DB, Schofield P, *et al*. I-motif DNA structures are formed in the nuclei of human cells. *Nat Chem* 2018;**10**(6):631–7.
9. Chan PP, Glazer PM. Triplex DNA: fundamentals, advances, and potential applications for gene therapy. *J Mol Med (Berl)* 1997;**75**(4):267–82.
10. Chou SH, Chin KH, Wang AH. Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res* 2003;**31**(10):2461–74.
11. Marceau AH. Functions of single-strand DNA-binding proteins in DNA replication, recombination, and repair. *Methods Mol Biol* 2012;**922**:1–21.
12. Wang W, Liu J, Zhou X. Identification of single-stranded and double-stranded DNA binding proteins based on protein structure. *BMC Bioinformatics* 2014;**15**(Suppl 12):S4.
13. Ghosh A, Bansal M. A glossary of DNA structures from A to Z. *Acta Crystallogr D Biol Crystallogr* 2003;**59**(Pt 4):620–6.
14. Potaman VN, Sinden RR. *DNA, in DNA Conformation and Transcription*. Boston, MA: Springer USA, 2005, 3–17.
15. Wagner FR, Dienemann C, Wang H, *et al*. Structure of SWI/SNF chromatin remodeller RSC bound to a nucleosome. *Nature* 2020;**579**(7799):448–51.
16. Jiang J, Chan H, Cash DD, *et al*. Structure of Tetrahymena telomerase reveals previously unknown subunits, functions, and interactions. *Science* 2015;**350**(6260):aab4070.
17. Yang H, Jeffrey PD, Miller J, *et al*. BRCA2 function in DNA binding and recombination from a BRCA2-DSS1-ssDNA structure. *Science* 2002;**297**(5588):1837–48.
18. Murphy FVT, Churchill ME. Nonsequence-specific DNA recognition: a structural perspective. *Structure* 2000;**8**(4):R83–9.
19. Grabowski M, Niedzialkowska E, Zimmerman MD, *et al*. The impact of structural genomics: the first quindecennial. *J Struct Funct Genomics* 2016;**17**(1):1–16.
20. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief Bioinform* 2019;**20**(4):1250–68.
21. Zhao H, Wang J, Zhou Y, *et al*. Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome. *PLoS One* 2014;**9**(5):e96694.
22. wwPDB consortium. Protein data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* 2019;**47**(D1):D520–8.
23. Mizianty MJ, Fan X, Yan J, *et al*. Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr D Biol Crystallogr* 2014;**70**(11):2781–93.
24. Si J, Zhao R, Wu R. An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci* 2015;**16**(3):5194–215.
25. Maheshwari S, Brylinski M. Predicting protein interface residues using easily accessible on-line resources. *Brief Bioinform* 2015;**16**(6):1025–34.
26. Kumar KK, Pugalenthi G, Suganthan PN. DNA-Prot: identification of DNA binding proteins from protein sequence information using random Forest. *J Biomol Struct Dyn* 2009;**26**(6):679–86.
27. Kumar M, Gromiha MM, Raghava GP. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* 2007;**8**:463.
28. Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics* 2019;**35**(3):433–41.
29. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* 2016;**17**(1):88–105.
30. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 2004;**20**(4):477–86.
31. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC bioinformatics* 2005;**6**(1):33.

32. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* 2006;**34**(suppl_2):W243–8.
33. Yan C, Terribilini M, Wu F, *et al*. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC bioinformatics* 2006;**7**(1):262.
34. Hwang S, Gou Z, Kuznetsov IB. DP-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* 2007;**23**(5):634–6.
35. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. *Bioinformatics* 2007;**23**(13):i347–53.
36. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* 2009;**10**(S1):S1.
37. Wu J, Liu H, Duan X, *et al*. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* 2009;**25**(1):30–5.
38. Chu W-Y, Huang YF, Huang CC, *et al*. ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res* 2009;**37**(suppl_2):W396–401.
39. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res* 2010;**38**(suppl_2):W431–5.
40. Wang L, Huang C, Yang MQ, *et al*. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010;**4**(S1):S3.
41. Ma X, Guo J, Liu HD, *et al*. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**(6):1766–75.
42. Hu J, Li Y, Zhang M, *et al*. Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**14**(6):1389–98.
43. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**(10): e84.
44. Zhu Y-H, Hu J, Song XN, *et al*. DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. *J Chem Inf Model* 2019;**59**(6):3057–71.
45. Giri R, Bhardwaj T, Shegane M, *et al*. Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses. *Cell Mol Life Sci* 2021;**78**:1655–1688.
46. Bhardwaj T, Saumya KU, Kumar P, *et al*. Japanese encephalitis virus-exploring the dark proteome and disorder-function paradigm. *FEBS J* 2020;**287**(17):3751–76.
47. do Amaral MJ, Araujo TS, Díaz NC, *et al*. Phase separation and disorder-to-order transition of human brain expressed X-linked 3 (hBEX3) in the presence of small fragments of tRNA. *J Mol Biol* 2020;**432**(7):2319–48.
48. Lim YY, Lim TS, Choong YS. Structural approaches for the DNA binding motifs prediction in bacillus thuringiensis sigma-E transcription factor (sigma(E)TF). *J Mol Model* 2019;**25**(10):301.
49. Ugidos N, Mena J, Baquero S, *et al*. Interactome of the autoimmune risk protein ANKRD55. *Front Immunol* 2019;**10**:2067.
50. Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol* 2015;**11**(12):e1004639.
51. Su H, Liu M, Sun S, *et al*. Improving the prediction of protein-nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* 2019;**35**(6):930–6.
52. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinform* 2018;**19**(5):821–37.
53. Wang W, Sun L, Zhang S, *et al*. Analysis and prediction of single-stranded and double-stranded DNA binding proteins based on protein sequences. *BMC bioinformatics* 2017;**18**(1):300.
54. Ali F, Arif M, Khan ZU, *et al*. SDBP-Pred: prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM. *Anal Biochem* 2020;**589**:113494.
55. Tan C, Wang T, Yang W, *et al*. PredPSD: a gradient tree boosting approach for single-stranded and double-stranded DNA binding protein prediction. *Molecules* 2020;**25**(1):98.
56. Sharma R, Kumar S, Tsunoda T, *et al*. Single-stranded and double-stranded DNA-binding protein prediction using HMM profiles. *Anal Biochem* 2021;**612**:113954.
57. UniProt C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**(D1):D506–15.
58. Dana JM, Gutmanas A, Tyagi N, *et al*. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res* 2019;**47**(D1):D482–9.
59. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res* 2012;**41**(D1):D1096–103.
60. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 2019;**35**(14):i343–53.
61. Zhang J, Ghadermarzi S, Kurgan L. Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes versus disordered proteins. *Bioinformatics* 2020;**36**(18):4729–38.
62. Wang K, Hu G, Wu Z, *et al*. Comprehensive survey and comparative assessment of RNA-binding residue predictions with analysis by RNA type. *Int J Mol Sci* 2020;**21**(18):6879.
63. Gromiha MM, Saranya N, Selvaraj S, *et al*. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome Science* 2011;**9**(Suppl 1):S13.
64. Faraggi E, Zhou YQ, Kloczkowski A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins* 2014;**82**(11):3170–6.
65. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* 2018;**46**(W1):W329–37.
66. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. *Bioinformatics* 2000;**16**(4):404–5.
67. Remmert M, Biegert A, Hauser A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**(2):173–5.
68. Kawashima S, Pokarowski P, Pokarowska M, *et al*. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**(Database issue):D202–5.
69. Lou W, Wang X, Chen F, *et al*. Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random Forest and Gaussian Naïve Bayes. *PLoS One* 2014;**9**(1):e86703.

70. Cai Y, He JF, Li XL, *et al*. A novel computational approach to predict transcription factor DNA binding preference. *J Proteome Res* 2009;**8**(2):999–1003.
71. Qian Z, Lu L, Liu XJ, *et al*. An approach to predict transcription factor DNA binding site specificity based upon gene and transcription factor functional categorization. *Bioinformatics* 2007;**23**(18):2449–54.
72. Li Q, Cao Z, Liu H. Improve the prediction of RNA-binding residues using structural neighbours. *Protein Pept Lett* 2010;**17**(3):287–96.
73. Walia RR, Caragea C, Lewis BA, *et al*. Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics* 2012;**13**(1):89.
74. Terribilini M, Sander JD, Lee JH, *et al*. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res* 2007;**35**(Web Server):W578–84.
75. Kohavi R, John GH. Wrappers for feature subset selection. *Artificial Intelligence* 1997;**97**(1–2):273–324.
76. Vacic V, Uversky VN, Dunker AK, *et al*. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC bioinformatics* 2007;**8**(1):211.
77. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**(17):3389–402.
78. Hu G, Kurgan L. Sequence similarity searching. *Curr Protoc Protein Sci* 2019;**95**(1):e71.
79. Zhang ZM, Lu R, Wang P, *et al*. Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature* 2018;**554**(7692):387–91.
80. Tak Leung RW, Jiang X, Chu KH, *et al*. ENPD-A database of eukaryotic nucleic acid binding proteins: linking gene regulations to proteins. *Nucleic Acids Res* 2019;**47**(D1): D322–9.
81. el-Gebali S, Mistry J, Bateman A, *et al*. The Pfam protein families database in 2019. *Nucleic Acids Res* 2019;**47**(D1):D427–32.
82. Mi H, Muruganujan A, Huang X, *et al*. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* 2019;**14**(3):703–21.
83. Avliyakulov NK, Lukes J, Ray DS. Mitochondrial histone-like DNA-binding proteins are essential for normal cell growth and mitochondrial function in Crithidia fasciculata. *Eukaryot Cell* 2004;**3**(2):518–26.
84. de S, Varsally W, Falciani F, *et al*. Ribosomal proteins' association with transcription sites peaks at tRNA genes in Schizosaccharomyces pombe. *RNA* 2011;**17**(9):1713–26.
85. Yang TT, Chow CW. Elucidating protein: DNA complex by oligonucleotide DNA affinity purification. *Methods Mol Biol* 2012;**809**:75–84.
86. Ma T, Ye Z, Wang L. Genome wide approaches to identify protein-DNA interactions. *Curr Med Chem* 2019;**26**(42):7641–54.
87. Massie CE, Mills IG. Mapping protein-DNA interactions using ChIP-sequencing. *Methods Mol Biol* 2012;**809**:157–73.
88. Cozzolino F, Iacobucci I, Monaco V, *et al*. Protein-DNA/RNA interactions: an overview of investigation methods in the -omics era. *J Proteome Res* 2021;**20**(6):3018–30.