Briefings in Bioinformatics, 2025, **26(1)**, bbaf016
https://doi.org/10.1093/bib/bbaf016
Review

Twenty years of advances in prediction of nucleic acid-binding residues in protein sequences

Sushmita Basu [1,*], Jing Yu [1], Daisuke Kihara [2,3], Lukasz Kurgan [1,*]

[1]Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Richmond, VA 23284, United States
[2]Department of Biological Sciences, Purdue University, 915 Mitch Daniels Boulevard, West Lafayette, IN 47907, United States
[3]Department of Computer Science, Purdue University, 305 N. University Street, West Lafayette, IN 47907, United States

*Corresponding authors. Sushmita Basu, Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA 23284, United States. E-mail: basudass@vcu.edu; Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA 23284, United States. E-mail: lkurgan@vcu.edu

Abstract

Computational prediction of nucleic acid-binding residues in protein sequences is an active field of research, with over 80 methods that were released in the past 2 decades. We identify and discuss 87 sequence-based predictors that include dozens of recently published methods that are surveyed for the first time. We overview historical progress and examine multiple practical issues that include availability and impact of predictors, key features of their predictive models, and important aspects related to their training and assessment. We observe that the past decade has brought increased use of deep neural networks and protein language models, which contributed to substantial gains in the predictive performance. We also highlight advancements in vital and challenging issues that include cross-predictions between deoxyribonucleic acid (DNA)-binding and ribonucleic acid (RNA)-binding residues and targeting the two distinct sources of binding annotations, structure-based versus intrinsic disorder-based. The methods trained on the structure-annotated interactions tend to perform poorly on the disorder-annotated binding and vice versa, with only a few methods that target and perform well across both annotation types. The cross-predictions are a significant problem, with some predictors of DNA-binding or RNA-binding residues indiscriminately predicting interactions with both nucleic acid types. Moreover, we show that methods with web servers are cited substantially more than tools without implementation or with no longer working implementations, motivating the development and long-term maintenance of the web servers. We close by discussing future research directions that aim to drive further progress in this area.

Keywords: protein–DNA interaction; protein–RNA interaction; nucleic acid-binding; DNA-binding residue; RNA-binding residue; intrinsic disorder; sequence-based prediction; machine learning; deep learning

Introduction

Interactions between biomolecules are major drivers of cellular processes. In particular, protein–nucleic acids interactions play crucial roles in a number of key cellular functions including deoxyribonucleic acid (DNA) replication, transcription, translation, and gene regulation [1–5]. Moreover, their misregulation is associated with several human diseases [6–8], providing further motivation to study these interactions. Several experimental techniques, such as pull-down assays, chromatin immunoprecipitation and CRISPR-Cas based approaches are used to study the protein–nucleic acids interactions [3, 9]. Additionally, atomic-level details that include information about the interacting residues and nucleotides can be obtained from experimentally solved structures of protein–nucleic acids complexes [10, 11]. However, obtaining structures of these complexes is relatively cost-intensive and challenging, especially when considering that nearly 250 million protein sequences are available as of September of 2024, with 90 million that were collected over the past 5 years [12]. To this point, computational methods that predict protein–nucleic acids interactions from sequences can be beneficial in bridging this large and growing function annotation gap. These computational tools are typically trained/generated from the limited amounts of the experimentally annotated data. The trained models can be used to produce predictions in a high-throughput manner for a large number of uncharacterized proteins, if their predictive performance is sufficiently good. The functional importance of protein–nucleic acids interactions combined with the availability of a sufficient amount of the corresponding experimental data for the model training and testing has motivated the development of several dozens of computational predictors of nucleic acid-binding residues in proteins.

These predictors can be divided into two groups, those that use the protein structure versus protein sequence as the input. The structure-based methods exploit the structural features, such as secondary structure, solvent accessibility, characteristics of spatial neighborhoods in the structure and shape complementarity, to derive predictive models [13–20]. Some of the popular and recent structure-based methods include (chronologically) aaRNA [17], NucleicNet [18], Graphbind [19], Geobind [20], and PNAbind

Received: October 3, 2024. **Revised:** December 24, 2024. **Accepted:** January 6, 2025

Table 1. Summary of surveys that discuss sequence-based predictors of the nucleic acid-binding residues. The articles are sorted chronologically with the most recent study at the top of the table. The 'Type of methods covered' column identifies whether a given survey covers 'Str' methods that were trained on the structure-annotated proteins and/or 'dis' methods that were trained on the disorder-annotated proteins. Bold font identifies data for this review

Reference (year published)	Target of assessment	Type of methods covered	No. of sequence-based predictors of DBRs/RBRs covered	No. of methods that were published after 2018 (year of the most recent method)	Types of training datasets discussed	Cross-prediction discussed
This study (NA)	**DNA and RNA**	**Str + Dis**	**87**	**34 (2024)**	**Yes**	**Yes**
[40] (2023)	DNA and RNA	Dis	3	2 (2022)	Yes	No
[30] (2022)	DNA and RNA	Str + Dis	13	3 (2021)	No	No
[34] (2020)	RNA	Str	28	2 (2019)	No	Yes
[31] (2019)	DNA	Str	8	0	No	No
[33] (2016)	DNA and RNA	Str	30	0	No	Yes
[37] (2016)	RNA	Str	12	0	No	No
[41] (2015)	DNA and RNA	Str	24	0	No	Yes
[29] (2015)	DNA	Str	7	0	No	No
[38] (2015)	RNA	Str	19	0	No	No
[35] (2013)	DNA	Str	13	0	No	No
[36] (2013)	DNA	Str	11	0	No	No
[28] (2013)	RNA	Str	10	0	No	Yes
[39] (2012)	RNA	Str	13	0	No	No

[21]. With the availability of a large number of structures predicted with AlphaFold2 [22], some of the recent methods, such as EquipNAS [23], GraphSite [24], and GraphNABP [25], utilize putative structures to train their models. In this survey, we focus the sequence-based predictors that use only the protein sequences as inputs to predict the DNA-binding residues (DBRs) and the ribonucleic acid (RNA)-binding residues (RBRs). Since the release of the first sequence-based predictors in 2004 [26, 27], many more methods were published [28–34]. We identified over two dozen sequence-based predictors that were released in the past 5 years, demonstrating that this is an active field of research. Availability of a large number of methods prompted the release of several surveys that we summarize in Table 1. These reviews typically enumerate the available predictors, list the corresponding references, summarize their predictive models, and provide guidance on how to select an appropriate method from a pool of multiple choices. They usually cover dozens of methods that include tools that predict DBRs [29, 31, 35, 36], RBRs [28, 34, 37–39] and both DBRs and RBRs [30, 33, 40, 41].

Most of the surveyed predictors are trained on training datasets using machine learning (ML) algorithms. Based on the source of the binding annotations in the training datasets, they are divided into two categories. The first includes data derived from the structures of protein–nucleic acids complexes, which we refer to as structure-annotated training data. The second category involves binding interfaces that are positioned in intrinsically disordered regions (IDRs), resulting in the disorder-annotated training data. IDRs lack stable structure under physiological conditions [42–45]. Protein sequences may have one or multiple IDRs, which in some cases may cover an entire sequence.

Several studies show that IDRs are functionally important and abundant in the nucleic acid-binding proteins [46–51]. Importantly, protein–nucleic acids interactions in IDRs differ in multiple ways from the interacting structured regions. The former typically have polymorphic conformations that are induced by interacting with different ligands [52, 53]. Moreover, they are enriched in disorder-promoting amino acids and form larger interfaces upon binding with partner molecules when compared with the interfaces in the structured regions [54–56]. These differences may impact ability of the corresponding methods to make accurate predictions, especially when they are applied to make predictions in IDRs while the underlying model was trained on the structure-annotated dataset, and vice versa. With that in mind, Table 1 reveals that virtually all past surveys focus on the methods that were trained on the structure-annotated proteins [28, 29, 31, 33–39, 41], with just two articles that consider tools that were trained from the disorder-annotated proteins [30, 40]. Moreover, one of these two articles covers just two predictors trained from the disorder-annotated data and discusses methods that targets interactions with other ligands, such as proteins and lipids [40], and the other mentions one disorder-trained predictor without discussing this aspect of the model training [30].

Another important aspect is cross-prediction between DBRs and RBRs. Research show that some of the current methods heavily cross-predict (mis-predict) DBRs as RBRs and vice versa [33, 41]. This issue was sporadically discussed in some of the surveys [28, 33, 34, 41]. The first independent survey covering the cross-prediction was authored by Zhao *et al.* [28] (Table 1), while the first prediction tool that assessed the cross-predictions was SPOT-Seq [57]. However, the cross-prediction analysis in both studies was limited only to the predictors of the RNA-binding proteins, which produce protein-level results that do not include prediction of RBRs. The most recent assessment, conducted in 2020, performed a residue-level analysis on the cross-prediction but it covered only predictors of RBRs [34]. We find that since the study by Yan *et al.* [33], none of the surveys in the past 8 years discussed the cross-prediction across predictors of DBRs and RBRs.

We provide a thorough and practical overview of this active field of research, substantially improving over the past surveys that are listed in Table 1. We cover the largest number of sequence-based nucleic acid-binding residue predictors, which include over two dozen new methods that were published in the past 5 years and which were not included in the previous review articles (Table 1). We examine how the underlying predictive models changed over time, investigate their availability and impact, and provide an expanded discussion of cross-predictions for the predictors of both DBRs and RBRs. Moreover, we categorize predictors on whether they are trained on the structure-annotated versus disorder-annotated proteins and discuss the corresponding implications. Altogether, we comprehensively review the 20-year long journey of the development of predictors of DBRs and RBRs.

Materials and methods

Selection of methods

We performed an extensive literature search to obtain the list of sequence-based predictors of nucleic acid-binding residues. We considered three main sources: (i) the past surveys that covered sequence-based nucleic acid-binding residue predictors [28–31, 33–39, 41]; (ii) studies that cite these surveys; and (iii) a manual search in PubMed using the following search keywords: [(Nucleic acid binding residue) OR (RNA binding residue) OR (DNA binding residue) OR (nucleotide binding residue)] AND [(prediction) OR (identification)], [(Nucleic acid binding site) OR (RNA binding site) OR (DNA binding site) OR (nucleotide binding site)] AND [(prediction) OR (identification)]. We further filtered and selected only those methods which are published in peer-reviewed Q1 journals. Using the above protocol, we identified 87 sequence-based nucleic acid-binding residue prediction methods, which more than doubles the number of such methods covered in each of the previous surveys (Table 1).

Predictive performance assessment

Predictive performance is an important aspect of surveying tools in this area. While every published method was evaluated and compared against a selection of other predictors, the list of the other methods is typically rather short. Summarizing results from across multiple sources is challenging since several factors must be considered to ensure that results of different methods can be directly compared. Specifically, the corresponding assessments must be performed on the same benchmark dataset, with same type of annotations of binding residues, and at least one common metric of predictive performance. Additionally, these studies often covered overlapping methods, making it difficult to boost coverage, particularly for more recently published tools. We considered these factors and were able to collect, report and discuss the predictive performance of 20 methods, with 13 of them published within the past 5 years.

We discuss the assessments of the DBR and the RBR predictors separately. The DNA-binding test dataset that we used to assess the DBR predictors was first reported by Patiyal *et al.* [58], and subsequently used in two recent studies [59, 60]. This dataset combines the DNA-binding datasets that were developed to assess two earlier predictors, hybridNAP [32] and ProNA2020 [61]. This test dataset contains 46 DNA-binding proteins with 965 DBRs and 9911 non-DBRs. The RNA-binding test dataset that we used to assess the RBR predictors was compiled to evaluate the Pprint2 method [62], and was also recently used in the article that introduces MucLiPred [60]. Similar to the DNA-binding dataset, it combines the datasets sourced from refs. [32] (hybridNAP) and [61] (ProNA2020). This test dataset contains 161 RNA-binding proteins with 6966 RBRs and 44,349 non-RBRs. The annotations of binding residues for the test proteins sourced from the hybridNAP article were obtained from the BioLip database [63, 64], which in turn was derived from PDB [65, 66]. The binding residue annotations for the test proteins from the ProNA2020 article were collected from the Protein–DNA Interface database [67] and the Protein-RNA Interface database [68], both of which also rely on the PDB-derived data.

The 20 predictors of RBRs and DBRs output numeric propensity scores for the corresponding type of binding for each amino acid in the input protein sequence. These propensities are used to generate binary state predictions (binding versus non-binding) using a threshold, where residues with propensities > threshold are assumed to bind, and the remaining residues are assumed not to bind. Based on their frequent use in the source evaluation articles [32, 58–62] and related comparative surveys [34, 38, 41, 69–73], we used the area under the receiver operating characteristic curve (AUC) to evaluate the predicted propensity scores and the Matthews Correlation Coefficient (MCC) to evaluate the predictions of binary states:

$$\mathrm{MCC} = (\mathrm{TN} * \mathrm{TP} - \mathrm{FN} * \mathrm{FP}) / \sqrt{[(\mathrm{TP}+\mathrm{FP})\,(\mathrm{TP}+\mathrm{FN})\,(\mathrm{TN}+\mathrm{FP})\,(\mathrm{TN}+\mathrm{FN})]}$$

where TP, TN, FN, and FP are the numbers of true positives (correctly predicted binding residues), true negatives (correctly predicted non-binding residues), false negatives (binding residues incorrectly predicted as non-binding), and false positives (non-binding residues incorrectly predicted as binding), respectively. MCC values range between −1 and 1, where 0 denotes predictions at random levels and a higher positive score indicates stronger correlation between predictions and the native annotations of binding. The AUC is the area under the curve defined by the true positive rates (TPR) versus false positive rates (FPR) computed over the thresholds equal to all unique predicted propensities:

$$\mathrm{TPR} = \mathrm{TP} / (\mathrm{TP}+\mathrm{FN})$$

$$\mathrm{FPR} = \mathrm{FP} / (\mathrm{FP}+\mathrm{TN})$$

While AUC theoretically ranges from 0 to 1, the AUC of a random predictor is ~0.5 and higher values that are > 0.5 suggest stronger predictive performance. We collected the AUC and MCC scores for the predictors of DBRs from refs. [58–60, 74, 75] and for the predictors of RBRs from refs. [60, 62].

Results

We summarized details concerning the comprehensive list of the 87 predictors, such as their name, prediction target (DBRs, RBRs, or both), type of predictive models and training datasets, outputs and consideration given to cross-predictions, in Table 2. We covered 36 predictors of DBRs, 29 predictors of RBRs, and 22 predictors that target both DBRs and RBRs. Some methods are designed for specific types of nucleic acids and proteins. Specifically, SDCpred [76] predicts residues that interact with the mono- and dinucleotide-specific DNAs; SRCpred [77] targets predictions of the dinucleotide-specific RNA binding; DNAgenie [78] predicts A-DNA, B-DNA and single-stranded DNA binding residues; and TSNAPred [79] predicts residues that interact with the A-DNA, B-DNA, single stranded DNA, mRNA, tRNA, and rRNA. Moreover, EPDRNA [80] generates predictions of nucleic acid-binding residues in proteins associated with human diseases including cancer, cardiovascular and neurodegenerative diseases.

We analyzed the 87 methods from multiple complementary perspectives including a historical overview, their availability and impact, predictive performance, the consideration of the

Table 2. Sequence-based predictors of DBRs and RBRs. The methods are arranged in the chronological order. The ‘target’ column shows types of predicted binding residues: DNA-binding (D), RNA-binding (R), and DNA- and RNA-binding (DR). The ‘predictive architecture’ column covers neural network (NN), support vector machine (SVM), naïve Bayes (NB), logistic regression (LR), random forest (RF), template-based prediction (TB), decision tree (DT), linear regression (LR), and long short-term memory (LSTM) NN; deep-learning models are marked with [D]; we also name the protein language model (PLM) that a given method uses, if any, inside the round brackets. The ‘training dataset’ column differentiates between methods trained from the structure-annotated interactions (S), disorder-annotated interactions (D), and both types of annotations (S + D). The ‘output’ column includes propensity scores (P), binary state (B), and both (B + P). The ‘cross-prediction’ column shows whether the cross-prediction was not mentioned (NM), discussed but not considered (Dis), or corrected/considered (Corr) when designing a given method.

Ref	Year published	Name	Target (D, R, DR)	Predictive architecture (PLM types used for feature extraction)	Training dataset (S, D, S + D)	Output (B, P, B + P)	Cross-prediction (NM/Dis/-Corr)
[26]	2004	DBS-pred	D	NN	S	P	NM
[27]	2004	Jeong *et al.*	R	NN	S	B	NM
[81]	2005	DBS-PSSM	D	NN	S	B + P	NM
[82]	2006	BindN	DR	SVM	S	B + P	NM
[83]	2006	DNABindR	D	NB	S	P	NM
[84]	2006	Jeong *et al.*	R	NN	S	B	NM
[85, 86]	2006	DP-Bind	D	LR, SVM	S	B + P	NM
[87]	2007	DISIS	D	NN, SVM	S	B	NM
[88]	2007	Ho *et al.*	D	SVM	S	B	NM
[89, 90]	2006	RNABindR	R	NB	S	B	NM
[91]	2008	Pprint	R	SVM	S	B + P	NM
[92]	2008	PRINTR	R	SVM	S	B	NM
[93]	2008	RISP	R	SVM	S	B + P	NM
[94]	2008	RNAProB	R	SVM	S	B	NM
[95]	2009	BindN-RF	D	RF	S	B + P	NM
[96]	2009	DBD-Threader	D	TB	S	B + P	NM
[97]	2009	DBindR	D	RF	S	B + P	NM
[98]	2009	ProteDNA	D	SVM	S	B	NM
[76]	2009	SDCpred	D	NN	S	P	NM
[99, 100]	2009	PiRaNhA	R	SVM	S	B + P	NM
[101]	2010	BindN+	DR	SVM	S	B + P	NM
[102]	2010	NAPS	DR	DT	S	B + P	Dis
[103]	2010	PRNA	R	RF	S	B + P	NM
[104]	2010	ProteRNA	R	SVM	S	B	NM
[105]	2010	RBRpred	R	SVM	S	B	NM
[106]	2010	RNA	R	LR	S	B	NM
[107]	2011	MetaDBSite	D	NB, NN, LR, RF, SVM	S	B	NM
[108]	2011	PRBR	R	RF	S	B + P	NM
[77]	2011	SRCpred	R	NN	S	P	NM
[109]	2011	Choi and Han	R	SVM	S	B	NM
[110]	2011	Wang *et al.*	R	SVM	S	B	NM
[57]	2011	SPOT-Seq	R	TB	S	B	NM
[111]	2012	DNABR	D	RF	S	B + P	NM
[112]	2012	meta2	R	SVM	S	P	NM
[113]	2013	TargetS	D	SVM	S	P	NM
[114]	2014	Pan *et al.*	R	RF	S	B	NM
[115]	2014	SPOT-Seq-DNA	D	TB	S	B	NM
[116]	2014	SPOT-Seq-RNA	R	TB	S	B	NM
[117]	2014	RNABindRplus	R	SVM, LR	S	B + P	NM
[17]	2014	aaRNA	R	NN	S	B + P	NM
[118]	2015	RBRIdent	R	RF	S	B + P	NM
[119]	2015	SNBRFinder	DR	SVM, TB	S	B + P	NM
[120, 121]	2015	DisoRDPbind	DR	LR	D	B + P	NM
[122]	2015	RBScore-SVM	R	SVM	S	B	NM
[123]	2016	Dang *et al*	D	RF	S	P	NM
[124]	2016	DQPred-DBR	D	SVM	S	B + P	NM
[125]	2016	FastRNABindR	R	RF, SVM	S	P	NM
[126]	2016	TargetDNA	D	SVM	S	B + P	NM
[127]	2017	DRNApred	DR	LR	S	B + P	Corr
[128]	2017	PRODNA	D	Sparse Representation	S	P	NM
[129]	2017	EL_PSSM-RT	D	RF, SVM	S	B	NM
[130]	2018	PDRLGB	D	Light Gradient Boosted DT	S	B	NM

(Continued)

Table 2. Continued

Ref	Year published	Name	Target (D, R, DR)	Predictive architecture (PLM types used for feature extraction)	Training dataset (S, D, S + D)	Output (B, P, B + P)	Cross-prediction (NM/Dis/-Corr)
[131]	2018	funDNApred	D	Fuzzy Cognitive Map	S	P	NM
[132]	2019	DNAPred	D	SVM	S	B + P	NM
[32]	2019	hybridNAP	DR	LR	S	B + P	NM
[133]	2019	NucBind	DR	SVM, TB	S	B + P	Dis
[134]	2019	PSPrint-seq	R	RF	S	B	NM
[135]	2019	iProDNA-CapsNet	D	[D] Convolutional NN, Feed forward NN	S	B + P	NM
[61]	2020	ProNA2020	DR	NN, SVM (ProtVec)	S	B + P	NM
[136]	2020	EL_LSTM	D	[D] LSTM NN	S	B	NM
[137]	2021	SPDH	D	SVM	S	P	NM
[78]	2021	DNAgenie	D	LR, k-nearest neighbor, NB, RF, SVM	S	B + P	Corr
[138]	2021	NCBRPred	DR	[D] Recurrent NN	S	B + P	Corr
[139]	2021	bindEmbed21	DR	[D] Convolutional NN (ProtT5)	S	B + P	NM
[140]	2022	MTDsite	DR	[D] Bidirectional-LSTM NN	S	B	Dis
[141]	2022	DeepDISOBind	DR	[D] Convolutional NN	D	B + P	Corr
[142]	2022	PredDBR	D	[D] Convolutional NN, NN	S	B	NM
[58]	2022	DBPred	D	[D] Convolutional NN	S	B	NM
[79]	2022	TSNAPred	DR	[D] CapsNet, Light Gradient Boosted DT, NN	S	B + P	Dis
[143]	2022	iDRNA-ITF	DR	[D] NN, Bidirectional gated recurrent NN (CAN-NER)	S	B + P	Corr
[62]	2023	Pprint2	R	[D] Convolutional NN	S	B	NM
[144]	2023	Guan *et al.*	D	[D] Transformer, Convolutional NN	S	P	NM
[145]	2023	proRBR	R	RF	S	B + P	NM
[146]	2023	HybridRNAbind	R	[D] Convolutional NN, Recurrent NN, RF	S + D	B + P	Corr
[147]	2023	GLMSite	DR	[D] Graph NN (ProtTrans and ESMFold)	S	B + P	NM
[59]	2024	CLAPE-DB	D	[D] Convolutional NN (ProtBERT)	S	B + P	NM
[148]	2024	HybridDBRpred	D	[D] Transformer, NN	S + D	B + P	Corr
[80]	2024	EPDRNA	DR	RF, k-nearest neighbor, LR, XGBoost	S	B + P	NM
[149]	2024	DRBpred	DR	Light Gradient Boosted DT	S	B + P	NM
[60]	2024	MucLiPred	DR	[D] NN (BERT)	S	B + P	NM
[150]	2024	ULDNA	D	[D] LSTM NN (ESM2 and ProtTrans)	S	B + P	NM
[151]	2024	SOFB	DR	[D] Convolutional NN, Bidirectional-LSTM NN (NA Bert, ProtT5)	S	B	NM
[152]	2024	GPSFun	DR	[D] Graph NN (ProtT5-XL-U50)	S	B + P	NM

(*Continued*)

Table 2. Continued

Ref	Year published	Name	Target (D, R, DR)	Predictive architecture (PLM types used for feature extraction)	Training dataset (S, D, S + D)	Output (B, P, B + P)	Cross-prediction (NM/Dis/-Corr)
[153]	2024	GPSite	DR	[D] Graph NN (ProtTrans and ESMFold)	S	B + P	NM
[75]	2024	PDNApred	D	[D] Convolutional NN, Bidirectional Gated Recurrent Unit NN (ESM2 and ProtT5)	S	B + P	NM
[74]	2024	DIRP	D	[D] Convolutional NN (ESM2 and ProtTrans)	S	B + P	NM
[154]	2024	DeepDBS	D	[D] LSTM NN, RF	S	B + P	NM

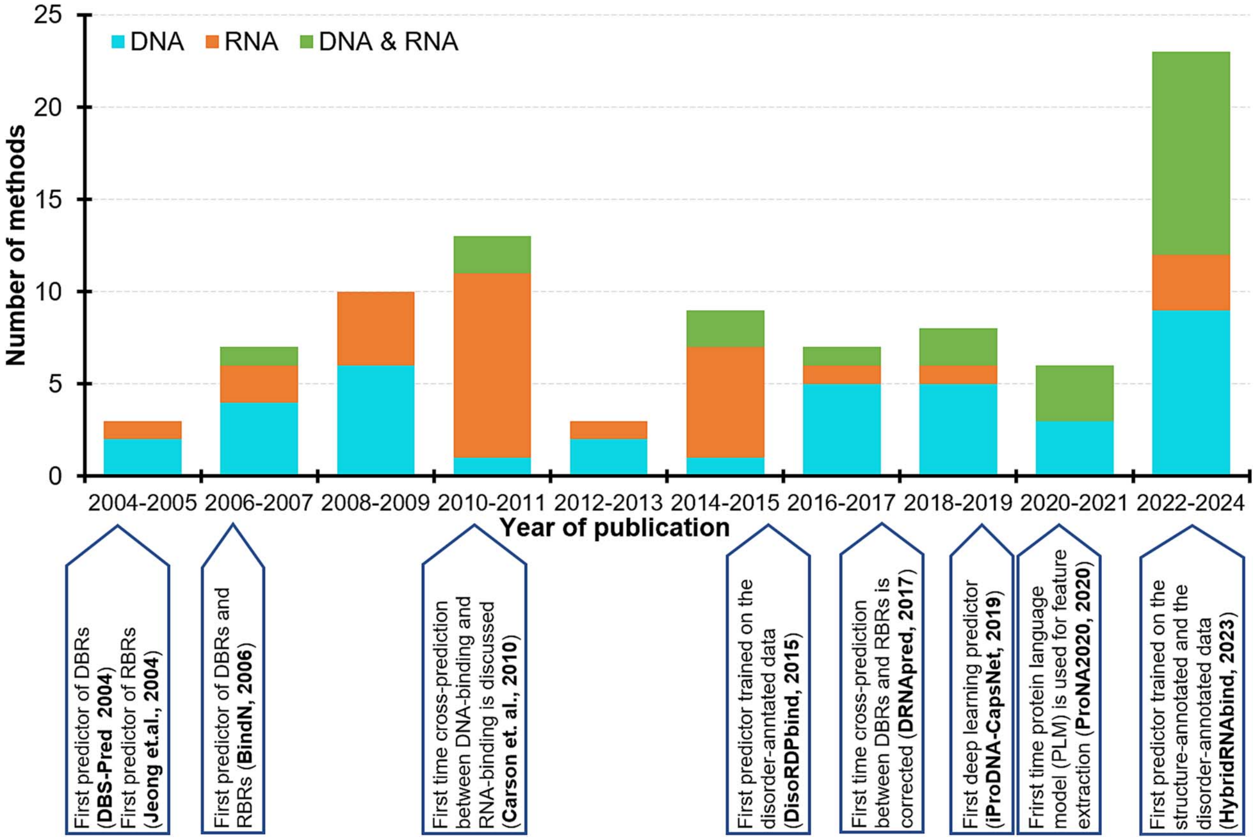

Figure 1. Timeline of the release of the 87 sequence-based nucleic acid binding residue predictors. The color-coded bars represent methods that target prediction of DBRs (blue), RBRs (orange), and both DBRs and RBRs (green). The major milestones are shown at the bottom in the blue-bordered boxes.

structure-based versus disorder-based annotations of binding in their training datasets, and cross-predictions between DBRs and RBRs.

Historical overview

Figure 1 illustrates the timeline of the release of the 87 predictors, where we highlighted eight major milestones. The first methods that predict exclusively DBRs (DBS-pred) or RBRs (predictor by Jeong et al.) were published in 2004 [26, 27]. Since then, at least one method was released each year. DBS-pred [26] is the first predictor of DBRs that was released as a web server. This contribution introduced a simple predictive architecture in the form of a shallow feed forward neural network (NN) and defined how to collect datasets with annotations of DBRs, which were used to train and test this predictive model. At the same time, Jeong et al. published two methods, one in 2004 [27] and another in 2006 [84], but neither predictor was released for public use (no code and no web server). RNABindR that was published in 2007 [89], is the first predictor of RBRs that is available as a web server.

As another milestone (Fig. 1), BindN that was published in 2006 [82], is the first method that predicts both DBRs and RBRs. Except for this aspect, BindN arguably did not contribute to moving other aspects of this field forward as it utilizes a relatively simple design, which consists of a classical support vector machine (SVM) model that uses just three biochemical features of amino acids as the input (pK_a, hydrophobicity, and molecular mass). This simplicity motivated the subsequent release of an improved BindN+ version in 2010 [101], which was designed to utilize more sophisticated inputs that rely on evolutionary information. By 2009, 20 methods were released and yet the issue of the cross-predictions between DBRs and RBRs did not surface. Authors of NAPS [102], a tool that predicts DBRs and RBRs, were the first to briefly discuss this concern but they did not address it in their design. It took several more years until 2016 when the cross-prediction was quantified and compared across predictors in the comparative study by Yan *et al.* [33]. This study motivated the development and release in 2017 of DRNApred [127], which was designed to accurately predict and discriminate between DBRs and RBRs (i.e. minimize the cross-predictions). The main innovation that was introduced in DRNApred is a two-layered architecture, where predictions of DBRs and RBRs produced by the first layer are input together into the second predictive layer that refines them to minimize cross-predictions. Importantly, until 2015 all the methods were developed using the structure-annotated training data. As a major milestone (Fig. 1), DisoRDPbind that was published in 2015 [120], is the first tool that was designed using the disorder-annotated training and test datasets, extending the DBR/RBR annotation protocols that were introduced in mid 2000s.

The predictive architectures of the methods, see Table 2, are predominantly based on ML algorithms, with a just few exceptions where template-based approaches are used [57, 96, 115, 116]. For example, SNBRFinder [119] and NucBind [133] utilize a template-based approach to predict protein structure from the input sequence, which is followed by the application of ML-generated models, particularly SVM, that identify putative DBRs and RBRs from the predicted structures. Significant majority of the predictors of DBRs and RBRs rely on shallow ML algorithms. Many different shallow ML algorithms were tried including the most widely used SVM, which was applied by itself in 19 predictors [82, 88, 91–94, 98, 99, 101, 104, 105, 109, 110, 113, 122, 124, 126, 132, 137] and in combination with some other algorithms in additional 9 predictors [61, 78, 85, 87, 107, 117, 119, 129, 133]. The next two popular shallow ML algorithms are random forest (RF), which was applied in 14 methods [78, 80, 95, 97, 103, 108, 111, 114, 118, 123, 125, 129, 145, 146], and shallow NN that were used in 10 methods [17, 26, 27, 61, 76, 77, 81, 84, 87, 107]. Some of the other shallow ML algorithms include logistic regression (LR) [78, 80, 85, 107, 117, 120, 121, 127], naïve Bayes (NB) [78, 83, 89, 107], linear regression [32, 106] and decision trees (DT) [61, 102, 130, 149]. Interestingly, we observed a substantial increase in the use of deep NNs (DNNs) over the past 5 years, where 24 of the 34 predictors rely on the DNN models. Released in 2019, iProDNA-CapsNet [135] is the first method that applied the DNN model, marking another major milestone (Fig. 1). The main innovation behind this predictor was the formulation and training of the predictive model that involves two two-dimensional convolutional layers connected to a fully connected feed forward layer. However, iProDNA-CapsNet relies on rather generic inputs, in the form of evolutionary information that was previously utilized to design several past predictors. Analysis of Table 2 reveals that the convolutional NNs are the most widely used architecture of DNNs among the predictors of RBRs and DBRs [58, 59, 62, 74, 75, 135, 139, 141, 142, 144, 146, 151], although several other architectures that include unidirectional and bidirectional recurrent networks, transformers, and graph networks were also used. In addition to the development of the DNN models, we identified a recent trend of using pre-trained protein language models (PLMs) as feature/input extraction tools [59–61, 74, 75, 139, 147, 150–153]. Simply put, PLMs process an input protein sequence in a way similar to processing a sentence in human language, where functional motifs and domains of the protein act as words in the sentence [155], producing a vector of numerical features for each amino acid. PLMs have been used to solve several bioinformatics problems and literature shows that their use tends to lead to improvements in the predictive performance of models [156, 157]. In the context of the prediction of nucleic acid-binding residues, the ProNA2020 method in 2020 [61] was the first to use PLM called ProtVec [158] for the feature/input extraction, denoting another milestone (Fig. 1). Henceforth, eleven other predictors including bindEmbed21 (PLM: ProtTrans-ProtT5 [159]) [139], iDRNA-ITF (PLM: CANNER [160]) [143], CLAPE-DB (PLM: ProtTrans-ProtBERT [159]) [59], MucLiPred (PLM: ProtTrans-ProtBERT [159]) [60], GLMSite (PLM: ProtTrans [159], ESMFold [161]) [147], ULDNA (PLM: ESM2 [161], ESM-MSA [162] and ProtTrans [159]) [150], SOFB (PLM: ProtTrans-ProtT5 [159]) [151], GPSFun (PLM: ProtTrans- ProtT5-XL-U50 [159]) [152], GPSite (PLM: ProtTrans [159], ESMFold [161]) [153], PDNApred (PLM: ProtTrans-ProtT5 [159] and ESM2 [161]) [75] and DIRP (PLM: ProtTrans [159] and ESM2 [161]) [74] use PLMs for the feature extraction. The latest milestone was the development and release of HybridRNAbind in 2023 [146], which is the first tool that was trained using both structure- and disorder-annotated training data, bridging the two annotations types. Additionally, this RBR predictor also minimizes cross-prediction between RBRs and DBRs [146]. Soon after, HybridDBRpred, which targets prediction of DBRs and similarly combines the structure- and disorder-annotated training data, was published [148].

The above historical overview (Fig. 1) suggests that the timeline can be divided into two distinct decades, each defined by a number of unique characteristics. The first decade spans the period between 2004 and 2014, and the second decade includes years from 2014 onwards. The major focus during the first decade was on developing predictors that target either DBRs or RBRs, resulting in 16 DBR predictors, 17 RBR predictors and just three tools that predict both DBRs and RBRs. Moreover, these methods rely on relatively simple shallow ML algorithms and they were trained exclusively on the structure-annotated proteins. The second decade is a more dynamic period where five major milestones took place (Fig. 1). Although many predictors of either DBRs or RBRs were still developed, 19 methods that target both types of nucleic acid-binding residues were released in this period including 6 out of the 12 methods published in 2024 (Table 1). Furthermore, we observed a big shift in the choice of predictive architectures that increasingly included deep ML models and modern PLMs for the feature/input extraction, with the underlying objective to improve predictive performance. The second decade also brought consideration to the cross-predictions between DBRs and RBRs and predicting both structure- and disorder-annotated interactions. The former likely stems from the increased focus on predicting both types of interactions, which inevitably brings the matter of evaluating whether these predictions overlap. Altogether, the second decade featured development of more sophisticated predictive models that attempted to address the challenging issues affecting quality (cross-predictions) and scope (disorder- and structure-annotated) of DBRs and RBRs predictions.

Availability and impact

An important aspect of computational predictors is their accessibility to users. Supplementary Table S1 provides details concerning the mode of availability (a web server (WS), a standalone code (SC), both or neither, as declared in the corresponding publication) and their current availability (whether or not the declared mode is currently available). We verified the availability of these implementations in November 2024 when we collected these data by checking the web links from the reference articles. This led to three outcomes: available (we list the corresponding links in Supplementary Table S1), not working (links in the original reference did not work as of November 2024), and never available (authors did not make their tools available in the original reference).

The two modes of availability, WS and SC, differ in multiple aspects. Users can access WSs via a web browser and these predictions are computed on the server side, typically without installing any software on the user's side. This makes a WS an arguably easier to use option, however, each prediction request is usually limited to a single protein or a small batch of proteins and the runtime might be affected by the ongoing server load. On the other hand, SC has to be downloaded and installed by the user and the computations are performed on the user's hardware. SC can be challenging to install, requiring users to install one or more third-party applications and have specific hardware and/or software infrastructure. However, SC offers certain advantages when compared to WS, including ability to be embed into other bioinformatics pipelines and to generate predictions at a large scale. We found that 75 out of the 87 listed methods provided at least one mode of availability at the time of their publication (Supplementary Table S1). Out of these 75 methods, 56 (77%) were originally released as WSs, either solely or along with SC, making WSs the most common mode of availability.

Though SC was provided for the first time in 2010 by the authors of PRNA [103], this option was rather uncommon until recently. Out of the 28 methods that were published with SC, 21 were released in the past 5 years. Moreover, 9 of these 21 methods were originally published with both WS and SC, which arguably broadens the utility of the methods when compared to tools that offer one mode of availability. However, as of November 2024 only 35 out of the 75 originally available methods have a working WS or SC; the links for the other 40 no longer work. Among the 35 currently available methods, 12 are only WS, 13 are only SC, and 10 are both WS and SC. The overall availability rate for the sequence-based predictors of RBRs and DBRs is at 40% (35 out of 87), which is same as the 40% rate for the predictors of protein-binding residues [71] and lower than the recently reported 71% rate for the predictors of the disordered binding residues [40].

We also analyzed the impact or popularity of the sequence-based predictors of RBRs and DBRs based on their citations, which we collected from the Google Scholar in November 2024 (Supplementary Table S1). While we collected the total number of citations for each method, we relied on the corresponding annual citation rates (i.e. total citations divided by the number of years since publication) which are more appropriate to compare between methods. We also excluded the methods which are published from 2023 onwards, since they are too new to reliably measure their citations.

We found that methods that did not offer either mode of availability (i.e. not made available) are cited substantially less (median annual citations = 3) compared to the tools which originally had at least one mode of availability (median annual citations = 8). These median annual citation counts are much larger for the methods that have currently working WS and/or SC (median annual citations = 13). Among these working tools, the tools with only working WS receive the most citations (median annual citations = 15), closely followed by the methods with both WS and SC (median annual citations = 14). Moreover, methods that are available solely as SC are comparatively less cited (median annual citations = 5). We hypothesize that the higher citations for the methods with working WSs are because this mode of availability is accessible to a much broader group of users, including those who have limited technical expertise and computational resources. On the other hand, methods that were not made available secure relatively poor citation numbers, which suggests that availability strongly affects the rate of use (citations). Moreover, we also observed that methods which were originally available but which currently do not work obtain much fewer citations when contrasted against tools with working WS and/or SC (median annual citations = 8 versus 13, respectively). This implies that availability of methods should be maintained after their release, or otherwise their impact is much diminished.

Lastly, we briefly comment on a few most impactful/cited methods. BindN [82], the first tool that predicts both DNA- and RNA-binding residues, has the highest total citations of 495 (Supplementary Table S1). Since then, this tool has undergone two upgrades [95, 101], with the most recent version, BindN+ [101], released in 2010, which received over 200 citations to date. DBS-pred [26], predictor of DBRs, is the only other tool that has total citations at over 400. Among predictors of RBRs, Pprint [91] with an overall citations of 322, is the most cited. In fact, Pprint along with DP-Bind [85] (overall citations of 270) are the only two methods that have maintained their implementations for over 10 years, which likely contributed to their high citation counts. There are 21 methods which have been cited >100 times and hybridNAP [32], which was published in 2019, is the most recent method to collect such high number of citations.

We also highlight a few recently published applications of these highly cited tools in biological contexts to further substantiate their impact. Multiple predictors including BindN [82], BindN+ [101] and metaDBsite [107] were used in tandem to characterize the DSrC protein from sulfur oxidizing bacterium *Allochromatium vinosum* [163]. Similarly, DRNApred [127], Pprint [91] and RNAbind-Plus [117] were applied to study the GRP20 protein in the context of flower development [164], while Pprint and hybridNAP [32] were applied to investigate roles of the SIX1 protein in the iron metabolism associated with progression of endometrial cancer [165]. These predictors were also used individually, with an example of DRNApred [127] that was recently applied to characterize proteins encoded by a viral mycobacteriophage gene [166] and CRESS DNA viruses [167]. Methods that have low runtimes were used to analyze entire proteomes. For instance, DisoRDPbind [120] was used to investigate length variation of short tandem repeats in *A. thaliana* [168], abundance and function of intrinsic disorder in the polyomavirus [169], and functions of the RNA-binding proteins in human [48]. DisoRDPbind together with DRNAPred and Pprint were applied to analyze the COVID-19 proteome [170]. Moreover, these tools were used to generate predictions at the scale of multiple proteomes and these data are conveniently available to the users via specialized databases. For instance, GPSite's predictions for the entire Swiss-Prot database are available in the GPSiteDB database [153], while DisoRDPbind's predictions for 273 reference proteomes, which cover the popular and model organisms, can be conveniently obtained from the DescribePROT database [171, 172].

To summarize, our analysis revealed a significant amount of interest in the area of the nucleic acid binding residue prediction

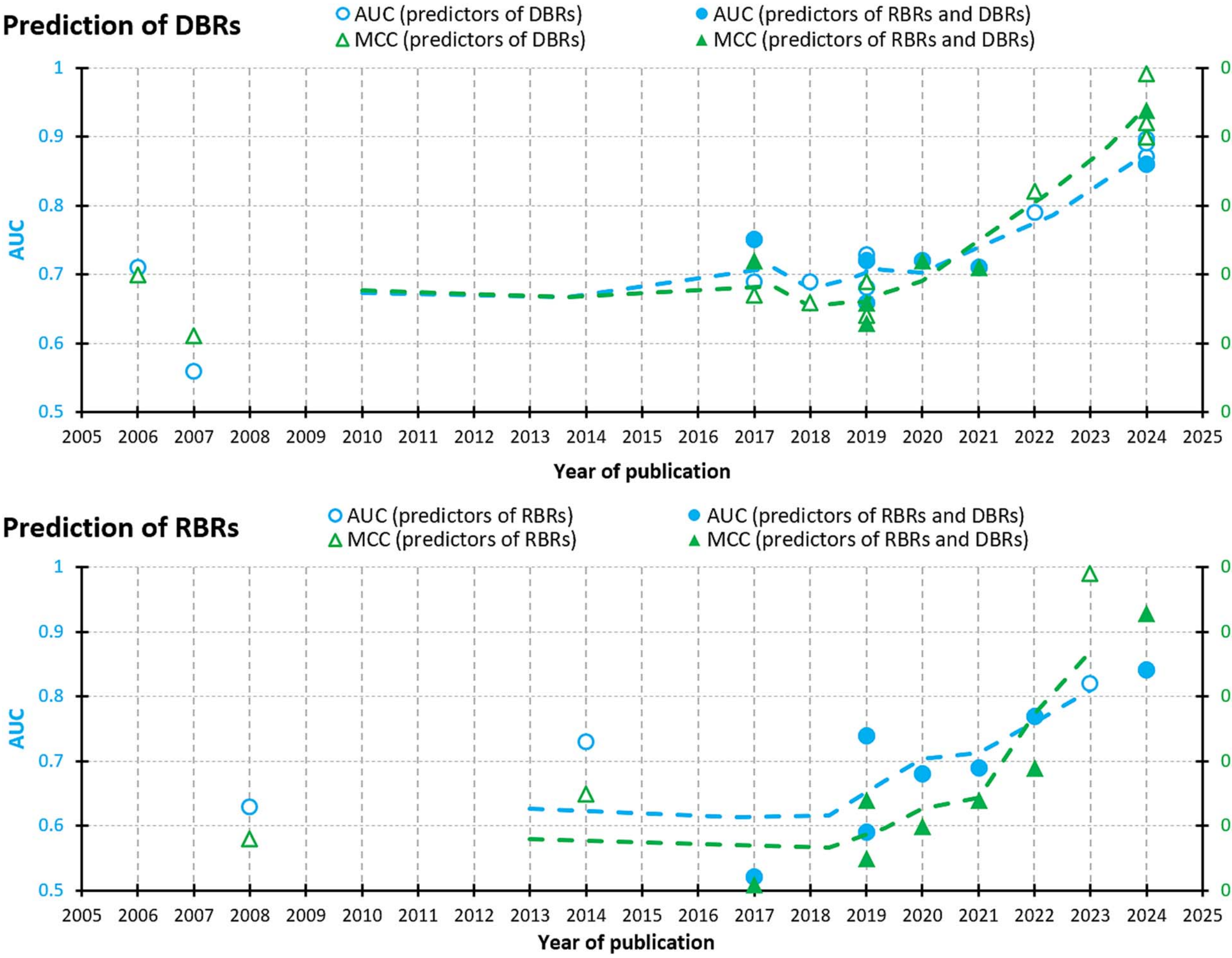

Figure 2. Relation between the publication year and predictive performance for the corresponding methods that was measured on the same benchmark dataset of 46 DNA-binding proteins that was introduced and applied in refs. [58–60, 74, 75] (top panel) and the same benchmark dataset of the 161 RNA-binding proteins that was used in refs. [60, 62] (bottom panel). Hollow markers denote methods that predict DBRs (top panel) or RBRs (bottom panel) while solid markers are for predictors of DRBs and RBRs (both panels). The primary/left y-axis quantifies the AUC values (blue markers) and the secondary/right y-axis gives the MCC values (green markers). The color-coded dashed lines are the moving averages of the corresponding metrics calculated over three consecutive methods based on the publication years. The numerical values of AUCs and MCCs are given in the Supplementary Tables S2 (results for the top panel) and S3 (results for the bottom panel).

(i.e. the corresponding predictors were collectively cited over 6600 times and 21 of them were cited over 100 times), where methods with working WSs are the most cited/popular. These observations are in good agreement with a recent analysis for a broader collection of predictors of protein structure and function [173].

Predictive performance

We summarized predictive performance of the current predictors and analyzed how it evolved over time in Fig. 2. We compared performance of 16 predictors of DBRs and 10 predictors of RBRs. We used the same test datasets and performance metric for each of the two collections of methods to ensure that results can be directly compared across the corresponding tools. We relied on popular and recently developed test datasets that were introduced in refs. [58, 62]. The dataset for the assessment of the DBR predictors includes 46 DNA-binding proteins and was used in refs. [58–60, 74, 75], while the other dataset covers 161 RNA-binding proteins and was used in refs. [60, 62]. Figure 2 reports AUC that evaluates quality of the putative propensity scores and MCC that quantifies quality of the binary state predictions. Further details can be found in the Materials and Methods section.

Figure 2 visualizes relations between predictive performance of the 20 predictors and their corresponding publication time. The performance of both DBR and RBR predictors varies widely between modest levels (AUC of ∼0.6; MCC of ∼0.2) and high levels (AUC > 0.80; MCC > 0.4), with an overall trend of improving with passing years. We computed moving average-based trends by averaging AUC and MCC scores over a window of three chronologically consecutive methods (dashed lines in Fig. 2). From these trends, we found that earlier methods secure similar and relatively modest performance with an average AUC <0.7 and MCC <0.2. The performance trends upwards starting around 2020, with the recent methods having AUC well >0.8 and MCC of ∼0.5. The best-performing predictors of DBRs are DIPR (2024) with AUC of 0.89 and MCC of 0.49 and PDNApred (2024) with AUC of 0.90 and MCC of 0.49; Supplementary Table S2. The best predictors of RBRs are MucLiPred (2024) with AUC of 0.84 and MCC of 0.43 and Pprint2 (2023) which secures AUC of 0.82 and MCC of 0.49; Supplementary Table S3. This demonstrates that the most accurate predictions of RBRs and DRBs are produced with similar levels of performance, with predictors of DBRs performing slightly better.

We also analyzed relation between the AUC and MCC scores and found that these scores are inconsistent for some methods, i.e. they should roughly follow a linear relation while some results deviate from this trend (Supplementary Fig. S1). MCC is a threshold-dependent measure that is derived from the predicted propensities for binding while AUC directly evaluates the propensities without the use of a threshold. The articles that presented these results used a 'default' threshold of 0.5 to generate the binary predictions for the calculation of MCC. Setting the same threshold value fails to account for the differences in the ranges and distributions of the propensities produced by different methods, leading to MCC values that are computed at different rates of positive (binding) to negative (non-binding) predictions. This could be the reason for the observed inconsistencies. A better strategy is to apply thresholds that standardize the predictions of different methods to a consistent prediction rate (say 5% or 10% FPR). However, the overall correlations between MCC and AUC metrics are high, with the Pearson's correlation coefficient of 0.95 and 0.87 for the DBR and the RBR predictors, respectively.

As we discuss in the Historical Overview section, recent methods often rely on the DL-based models when compared with older methods that primarily use shallow ML-based algorithms (Table 2). Correspondingly, we investigated whether the recent improvements in the predictive importance could be attributed to the use of the more sophisticated predictive models. We limited this analysis to the predictors that were published since 2019 when the first DL-based predictor was released. Using the corresponding results from Supplementary Tables S2 and S3, we found that the DL-based DBR predictors have substantially higher predictive performance than the shallow ML-based predictors that were published in the same period of time, with median AUC = 0.86 versus 0.72 and median MCC = 0.40 versus 0.18. Similar observations are true for the RBR predictors, where the DL-based methods secure median AUC = 0.80 versus 0.68 for the other group of predictors, and MCC = 0.31 versus 0.10, respectively. Recent studies in related areas including protein function prediction [174, 175] and intrinsic disorder prediction [176] similarly showed that DL-based predictive models substantially outperform the shallow ML-based models. Our analysis reveals that the same is true for the prediction of the nucleic acid binding residues.

We also investigated whether similar patterns of differences can be observed when comparing recently released (since 2019) predictors that use PLMs for the feature/input extraction versus those that do not utilize PLMs. Based on the data in Supplementary Tables S2 and S3, we found that predictors of DBRs that use PLMs secure median AUC = 0.87 and median MCC = 0.42 versus median AUC = 0.72 and median MCC = 0.18 for the methods that do not use PLMs. For the predictors of RBRs, we similarly observed that the PLM-utilizing methods generate on average more accurate predictions than the other methods, with median AUC = 0.77 versus 0.72 and median MCC = 0.19 versus 0.14, respectively. These results suggest that the application of PLMs produces improvements in the predictive performance for the predictors that cover both types of binding residues. Our observation is also supported by an empirical analysis that demonstrated that the use of the ProtT5 PLM produces higher levels of accuracy than the use of the popular multiple sequence alignment-based inputs for equivalent models that predict DBRs and RBRs [139].

However, we note a limitation of our analysis. The predictive performance can be impacted by sequence similarity between the training and test proteins. In principle, predictors should be tested on the test proteins that share low similarity (typically <30%) with the training proteins. Several of the recent predictors in the above analysis were tested under this low similarity regime [58–60, 62, 74, 75], including DBpred, MucLiPred, CLAPE-DB, Pprint2, PDNApred, and DIRP. The training datasets of the other methods may share higher levels of similarity, and thus their reported predictive performance might be inflated. However, this does not affect our observations since the methods for which the performance was tested on the low similarity test proteins are the most accurate. Additionally, the two test datasets are structure-annotated, which means that this assessment does not reflect the performance on the disorder-annotated protein–nucleic acids interactions. We discuss this issue in a following section.

Structure-annotated and disorder-annotated training datasets

We divided the 87 predictors into two groups based on their training data: trained from the structure-annotated interactions (structure-trained) versus trained from the disorder-annotated interactions (disorder-trained). The 'Training dataset' column in Table 2 shows the type of interaction annotations used for training each method, where 'S' represents structure-annotated datasets and 'D' represents disorder-annotated datasets. The binding annotations for these two types of training datasets were obtained from two distinct sources. The structure-annotated binding residues were collected from the structures of the protein–nucleic acids complexes which are extracted directly from the PDB database [65, 66] or indirectly from the PDB-derived BioLip database [63, 64]. These annotations rely on certain criteria, such as distance between interacting atoms and number of atoms interacting per residue, to identify binding versus non-binding residues. On the other hand, the disorder-annotated interactions were derived from the DisProt database [177, 178], the largest repository of experimentally validated intrinsically disordered proteins. These annotations concern binding sequence regions rather than specific binding residues, as is the case for the structure-based annotations. In other words, disorder-annotated training datasets include disordered regions that are involved in binding, assuming (imprecisely) that all residues in the corresponding region are binding. The lack of the more precise annotation of the corresponding binding residues in the disordered regions is a result of an inherent difficulty in capturing these details without the structure. This limitation was discussed in the context of the recent community assessments of predictions of disordered binding regions [70, 179].

Table 2 shows that large majority of methods, 83 out of 87, was developed using solely the structure-annotated training datasets. On the other hand, DisoRDPbind [120, 121] and DeepDISOBind [141] are the two methods that are exclusively trained on the disorder-annotated data (Table 2). Both methods provide prediction of DBRs and RBRs alongside prediction of protein-binding residues. The rather low number of tools trained from the disorder-annotated interactions can be attributed to the fact that the corresponding data was released relatively recently, in early 2010s [180–182]. Interestingly, two recent studies reported that the structure-trained methods perform well on the structure-annotated proteins, whereas they secure poor to modest performance on the disorder-annotated proteins and vice versa [146, 148]. For example, when considering the predictors of RBRs [146], the structure-trained MTDsite performs the best on the structure-annotated interactions with AUC = 0.76, whereas, its AUC drops to 0.60 for the disorder-annotated interactions [146]. Similarly, the disorder-trained DeepDISOBind performs best on disorder-annotated interactions with AUC = 0.72, whereas, it secures a much lower AUC of 0.64 for the structure-annotated interactions [146]. Similar observations were published for the predictors of DBRs [148]. The fact that the ground truth annotations of binding for the disorder-annotated versus structure-annotated datasets are different (residues versus regions) and come from different source databases may explain the dichotomy in the method development efforts, i.e. methods are typically designed using either structure-annotated or disorder-annotated training sets and consequently they do not work equally well across the two annotation types. Similar observation of the dichotomy of structure-trained versus disorder-trained predictions was reported for the sequence-based predictors of the protein-binding residues [183].

To this end, two recently published methods, HybridRNAbind [146] and HybridDBRpred [148], were designed to address this dichotomy by targeting both types of annotations, i.e. they are trained and tested on datasets composed of both types of annotations. HybridRNAbind that predicts RBRs performs relatively well with the AUC of 0.76 on the structure-annotated interactions and AUC of 0.72 for the disorder-annotated interactions [146]. Similarly, HybridDBRpred that targets prediction of DBRs secures AUCs of 0.83 and 0.77 for the structure-annotated and the disorder-annotated interactions, respectively [148]. The development of these two methods shows that it is possible to build predictors that work well across the two annotation types, and suggests that these efforts should continue.

Cross-prediction between DBRs and RBRs

DNA and RNA share relatively high levels of similarity in their physicochemical nature, as both are made up of a monomeric unit having a nitrogenous base and a sugar-phosphate group. Given their resemblance at the molecular level, it is reasonable to expect that predictors of nucleic aci- binding residues may face difficulties to accurately discriminate between DBRs and RBRs. Besides accurately predicting putative binding residues, these methods also should be free from the cross-predictions where DBRs are confused for RBRs and vice versa. High levels of cross-predictions would mean that the corresponding methods predict residues that interact with nucleic acids in the type agnostic manner.

Two early surveys conducted empirical assessments of cross-predictions for several predictors of DBRs and RBRs, covering methods that were published before 2014 [33, 41]. These studies reported similar findings suggesting that none of these older methods accurately discriminates between DBRs and RBRs. Miao *et al.* found that several accurate predictors of RBRs also obtain high AUC scores when tested on predicting DBRs in the DNA-binding proteins, which implies that these methods predict binding residues irrespective of whether they bind DNA or RNA [41]. While they also show that some methods, such as PRNA, RNABindRPlus, RBScore-SVM, discriminate between DBRs and RBRs, the predictive performance of these tools on the RNA-binding datasets is low, with AUCs ~0.5 [41]. The study by Yan *et al.* quantified cross-predictions by measuring the fraction of DBRs (or RBRs) that are mis-predicted as RBRs (or DBRs) [33]. They found that among the RBR predictors, RNABindR has the highest cross-prediction rate, incorrectly classifying >60% of DBRs as RBRs while BindN+ has the lowest rate, at ~45%. For the DBR predictors, DBS-PSSM generates the most cross-predictions, with 44% RBRs predicted as DBRs. A positive exception is ProteDNA, predictor of DBRs, which accurately distinguishes DBRs from RBRs but has low sensitivity for DBRs, since it was designed to specifically predict DBRs in transcription factors [33]. These assessments emphasize the need to develop new predictors that minimize the cross-predictions between DNA and RNA. They motivated the release of DRNApred [127], which is the first method that was specifically designed to restrict cross-predictions. Consequently, a subsequent survey of the RBR predictors shows that DRNApred performs the best in terms of the cross-predictions [34]. However, these comparative studies are relatively old and do not cover recently published predictors [33, 34, 41]. We note that authors of several newer methods, such as NCBRpred [138], iDRNA-ITF [143], DNAgenie [78], DeepDISOBind [141], MTDsite [140], HybridDBRpred [148] and HybridRNAbind [146], evaluated cross-predictions and designed their models to minimize them. However, these studies considered/evaluated relatively few methods and used different datasets and metrics, constraining comparative analysis to a rather limited number of recent tools. Moreover, many recently published methods that include seven methods from 2024 and several that predict both DBRs and RBRs [60–62, 74, 75, 139, 147, 149, 151–154] overlooked this crucial aspect. To this end, we note that reports of predictive performance that do not account for the cross-predictions should be interpreted with caution.

Discussion

Sequence-based prediction of nucleic acid-binding residues is a mature and an active research area. We identified nearly 90 predictors that were published over the last two decades, including 29 that were published over the past 5 years and 12 that were released in 2024. We discussed multiple practical characteristics of these methods including their availability and impact, key features of their predictive models, and major aspects related to their training and assessment. We observed that the last decade produced a noteworthy progress in terms of improvements in the predictive quality, use of sophisticated predictive models and PLMs, and advancements on multiple vital and challenging issues, such as targeting of the two distinct annotation types (structure-based versus disorder-based) and cross-predictions.

Our analysis of the availability reveals that predictors that have a web server or a standalone code are cited substantially more than tools without implementations. Methods with the arguably easier to use web servers attract more citations when compared to the tools with the code. Furthermore, we found that methods that have implementations that are working and are maintained over the long term (in particular working and maintained web server) secure much higher citation counts compared to tools with originally available implementations that currently do not work. In the context of predictive models, the recently released predictors increasingly rely on modern deep network architectures. Our analysis revealed that these models outperform the previously-dominant shallow machine-learning algorithms, which is in line with results of similar analyses in related area of protein structure and function prediction [174–176]. We also stressed the impact of training on two distinct types of datasets: structure- versus disorder-annotated. Majority of predictors were trained on the structure-annotated interactions and they perform poorly when tested on the disorder-annotated interactions. Similarly, the predictors trained on the disorder-annotated datasets perform rather poorly on the structure-annotated interactions. The underlying dichotomy of the predictive models and their inability to crossover to the other type of annotations mirrors the prediction of protein-binding residues [183]. This motivated the development of the HybridDBRpred and HybridRNAbind methods that perform relatively well for both structure and disorder-annotated interactions. However, some protein–nucleic acids interactions are driven by the assembly of protein chains, such as the ones found in the ribosomal complexes [184, 185]. The sequence-based methods that are trained on monomeric protein chains may underperform for these proteins, particularly when compared with the structure-based methods that use the corresponding complexes for the training. Finally, we indicated that cross-predictions between DBRs and RBRs is a crucial aspect of empirical assessments of predictive performance, and yet this aspect was not evaluated for many of the recently published tools. Consequently, while some of the recently released predictors were shown to produce accurate results (AUC > 0.8 and MCC > 0.4; Supplementary Tables S2 and S3), users should also quantify and analyze their cross-prediction rates to formulate a more holistic picture of their performance. In general, users should avoid predictors with high cross-prediction rates since this means that their predictions are nucleic acids type agnostic. The binding residues that are predicted by the accurate sequence-based tools can be used to support subsequent modelling of the protein–nucleic acids interactions. In particular, tools that specialize in modelling and predicting binding specificity for the protein–nucleic acids complexes [18, 186–190] would likely benefit from the knowledge where a given DNA or RNA binds on the protein surface. While some of these tools target specific types of proteins, such as transcription factors [186, 188, 190], other tools can be applied to a more generic class of nucleic acid-binding proteins [18, 187, 189].

Our analysis motivates several considerations for future work. Many of the recently released predictors of RBRs and DBRs utilize PLMs to derive inputs to the predictive models, with ProtT5 and ProtBERT from the ProtTrans project [159] being the common selections. Our empirical analysis suggests that the use of PLMs leads to substantial improvements in predictive performance for predictors of both DBRs and RBRs, which is based on their overall higher AUC and MCC values when compared to the predictors that do not utilize PLMs. However, PLMs used by the current predictors were produced using generic collections of protein sequences while several PLMs that were designed for specific types/classes of proteins were released in recent years. For example, IDP-BERT was designed to capture characteristics of intrinsically disordered proteins [191]; ProGen was built from sequences of five families of lysozymes [192]; and IgLM was trained using antibody sequences [193]. We believe that similar efforts geared towards developing and using PLMs that target nucleic acid-binding proteins should drive further improvements in accuracy for the predictors of DBRs and RBRs. As a first step in this direction, the authors of the SOFB predictor adopted the generic ProtT5 PLM to make it more suitable for the recognition of nucleic acid-binding residues, and named this model NABert [151]. Moreover, structural and functional aspects of proteins are typically conserved in their amino acid sequences. Consequently, evolutionary profiles generated using protein sequence databases are commonly used to predict nucleic acids binding residues [62, 74, 78, 81, 86, 88, 91, 92, 94, 101, 129, 133, 153] and in related areas, such as the prediction of secondary structures [194–196], and intrinsic disorder [40, 197–199]. A few studies have pointed to the impact of the quality of the evolutionary profile on the predictive performance, which in turns stem from the size and quality of the underlying sequence alignment databases [81, 200]. These considerations offer additional opportunities to improve predictive performance of future RBRs and DBRs predictors.

Another important consideration concerns the ability of current and future predictors to accurately discriminate between DBRs and RBRs. The authors of future methods should measure and comparatively assess cross-predictions and design their models to minimize them. While in same aspects DNA and RNA are relatively similar, they are distinct in the structures of their binding interfaces [201, 202]. The π-stacking interactions between the aromatic amino acids and the nucleobases or sugar moieties of DNA and RNA play vital role in protein–nucleic acids recognition [203]. Previous studies highlighted differences between stacking interactions with DNA and RNA in terms of their rate of occurrences and preferences of amino acids with respective nucleobases [204, 205]. These fine details, which are typically extracted from 3D structures of protein–nucleic acids complexes, could be perhaps approximated from protein sequences or sequence-predicted protein structures, providing a way to improve the ability to distinguish between DBRs and RBRs. Moreover, future comparative assessments should cover the cross-prediction aspect to reveal which current methods accurately identify DNA versus RNA-binding residues and which are nucleic acids type agnostic. The last such study was published in 2020 [34] and is relative outdated given the large number of methods that were released subsequently. Furthermore, these assessments should cover cross-predictions between DBR and RBRs and also between DBRs/RBRs and residues that interact with other types of ligands, such as proteins, peptides and small molecules. Similar evaluations were done recently in the context of developing predictors of the protein-binding residues [206].

We also emphasize the substantial impact of ensuring sustained/long-term availability of web servers for the predictors of RBRs and DBRs. The current 40% availability rate should be improved to match levels in other areas, such as the 71% rate for the intrinsic disorder predictors [40]. As shown in a recent study [173] and our analysis, this is likely to increase their scientific impact that is indirectly measured by their citation rates. This can be accomplished by requiring the commitment to support web servers for an extended period of time at the point of publication, which would benefit both the developers and users.

Lastly, the structures of the protein–nucleic acids complexes are useful to investigate atomic level details of these interactions. They can be predicted when the native structures are unavailable, which is relatively common. Many such predictors are available including several docking-based tools [207–210]. The recently released Flex-LZerD that considers flexibility of the protein upon docking to nucleic acids [211], partly addresses predictions for the disordered binding regions. The release of AlphaFold3 that predicts structures of protein-ligand complexes, where ligands include proteins, nucleic acids, small molecules and ions [212], is also notable. However, AlphaFold3 authors note that their model generates '*spurious structural order (hallucinations) in disordered regions*' [212], which is a major drawback in the context of prediction of nucleic acid-binding residues that frequently reside in the disordered regions [46, 47, 49–51]. Moreover, these methods can be applied only when the structure of the nucleic acid is known, in contrast to the tools that we review which make predictions solely from the protein sequences.

Key Points

- Eighty-seven predictors of nucleic acid-binding residues in protein sequences were developed in the last two decades
- Machine learning is the primary approach to develop these predictors

- Recent use of deep learning and protein language models resulted in substantial gains in predictive performance
- Cross-predictions between RNA-binding and DNA-binding residues are a significant challenge
- Predictors with working web servers enjoy high citation rates, motivating development and long-term maintenance of web servers

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare no competing interests.

Funding

This work was supported in part by the National Science Foundation (grants DBI 2146027 and IIS 2125218) and the Robert J. Mattauch Endowment funds to L.K. D.K. acknowledges support from National Science Foundation (DBI 2146026, IIS 2211598, DMS 2151678) and National Institutes of Health (R01GM133840).

References

1. Djebali S, Davis CA, Merkel A. *et al.* Landscape of transcription in human cells. *Nature* 2012;**489**:101–8. https://doi.org/10.1038/nature11233.
2. Evande R, Rana A, Biswas-Fiss EE. *et al.* Protein–DNA interactions regulate human papillomavirus DNA replication, transcription, and oncogenesis. *Int J Mol Sci* 2023;**24**. https://doi.org/10.3390/ijms24108493.
3. 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**:3018–30. https://doi.org/10.1021/acs.jproteome.1c00074.
4. Oyejobi GK, Yan X, Sliz P. *et al.* Regulating protein–RNA interactions: advances in targeting the LIN28/Let-7 pathway. *Int J Mol Sci* 2024;**25**. https://doi.org/10.3390/ijms25073585.
5. Peng Z, Oldfield CJ, Xue B. *et al.* A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol Life Sci* 2014;**71**:1477–504. https://doi.org/10.1007/s00018-013-1446-6.
6. Cookson MR. RNA-binding proteins implicated in neurodegenerative diseases. *Wiley Interdiscip Rev RNA* 2017;**8**. https://doi.org/10.1002/wrna.1397.
7. Bonczek O, Wang L, Gnanasundram SV. *et al.* DNA and RNA binding proteins: from motifs to roles in cancer. *Int J Mol Sci* 2022;**23**. https://doi.org/10.3390/ijms23169329.
8. Tao YN, Zhang Q, Wang H. *et al.* Alternative splicing and related RNA binding proteins in human health and disease. *Signal Transduct Target Ther* 2024;**9**. https://doi.org/10.1038/s41392-024-01734-2.
9. Valuchova S, Fulnecek J, Petrov AP. *et al.* A rapid method for detecting protein–nucleic acid interactions by protein induced fluorescence enhancement. *Sci Rep* 2016;**6**:6.
10. Chen S, Yan K, Liu B. PDB-BRE: a ligand–protein interaction binding residue extractor based on protein data Bank. *Proteins* 2024;**92**:145–53. https://doi.org/10.1002/prot.26596.
11. Burley SK, Bhikadiya C, Bi C. *et al.* RCSB protein data bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 2021;**49**:D437–51. https://doi.org/10.1093/nar/gkaa1038.
12. UniProt C. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31.
13. Zhu XL, Ericksen SS, Mitchell JC. DBSI: DNA-binding site identifier. *Nucleic Acids Res* 2013;**41**. https://doi.org/10.1093/nar/gkt617.
14. Tsuchiya Y, Kinoshita K, Nakamura H. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins* 2004;**55**:885–94. https://doi.org/10.1002/prot.20111.
15. Tjong H, Zhou HX. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res* 2007;**35**:1465–77. https://doi.org/10.1093/nar/gkm008.
16. Zhou WQ, Yan H. A discriminatory function for prediction of protein–DNA interactions based on alpha shape modeling. *Bioinformatics* 2010;**26**:2541–8. https://doi.org/10.1093/bioinformatics/btq478.
17. Li S, Yamashita K, Amada KM. *et al.* Quantifying sequence and structural features of protein–RNA interactions. *Nucleic Acids Res* 2014;**42**:10086–98. https://doi.org/10.1093/nar/gku681.
18. Lam JH, Li Y, Zhu L. *et al.* A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat Commun* 2019;**10**:10. https://doi.org/10.1038/s41467-019-12920-0.
19. Xia Y, Xia CQ, Pan X. *et al.* GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res* 2021;**49**. https://doi.org/10.1093/nar/gkab044.
20. Li PP, Liu ZP. GeoBind: segmentation of nucleic acid binding interface on protein surface with geometric deep learning. *Nucleic Acids Res* 2023;**51**. https://doi.org/10.1093/nar/gkad288.
21. Sagendorf JM, Mitra R, Huang J. *et al.* Structure-based prediction of protein–nucleic acid binding using graph neural networks. *Biophys Rev* 2024;**16**:297–314. https://doi.org/10.1007/s12551-024-01201-w.
22. Varadi M, Anyango S, Deshpande M. *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44. https://doi.org/10.1093/nar/gkab1061.
23. Roche R, Moussad B, Shuvo MH. *et al.* EquiPNAS: improved protein–nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *Nucleic Acids Res* 2024;**52**. https://doi.org/10.1093/nar/gkae039.
24. Shi WT, Singha M, Pu L. *et al.* GraphSite: ligand binding site classification with deep graph learning. *Biomolecules* 2022;**12**. https://doi.org/10.3390/biom12081053.
25. Li X, Mei L, Ding X. *et al.* GraphNABP: identifying nucleic acid-binding proteins with protein graphs and protein language models. *Int J Biol Macromol* 2024;**280**. https://doi.org/10.1016/j.ijbiomac.2024.136356.
26. 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**:477–86. https://doi.org/10.1093/bioinformatics/btg432.

27. Jeong E, Chung IF, Miyano S. A neural network method for identification of RNA-interacting residues in protein. *Genome Inform* 2004;**15**:105–16.
28. Zhao HY, Yang YD, Zhou YQ. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol Biosyst* 2013;**9**:2417–25. https://doi.org/10.1039/c3mb70167k.
29. Si JN, Zhao R, Wu RL. An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci* 2015;**16**:5194–215. https://doi.org/10.3390/ijms16035194.
30. Cui FF, Zhang Z, Cao C. *et al.* Protein–DNA/RNA interactions: machine intelligence tools and approaches in the era of artificial intelligence and big data. *Proteomics* 2022;**22**. https://doi.org/10.1002/pmic.202100197.
31. Emamjomeh A, Choobineh D, Hajieghrari B. *et al.* DNA–protein interaction: identification, prediction and data analysis. *Mol Biol Rep* 2019;**46**:3571–96. https://doi.org/10.1007/s11033-019-04763-1.
32. 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**:1250–68. https://doi.org/10.1093/bib/bbx168.
33. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* 2016;**17**:88–105. https://doi.org/10.1093/bib/bbv023.
34. 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**. https://doi.org/10.3390/ijms21186879.
35. Gromiha MM, Nagarajan R. Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein–DNA complexes. *Adv Protein Chem Struct Biol* 2013;**91**:65–99. https://doi.org/10.1016/B978-0-12-411637-5.00003-2.
36. Nagarajan R, Ahmad S, Gromiha MM. Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res* 2013;**41**:7606–14. https://doi.org/10.1093/nar/gkt544.
37. Liu ZP, Chen LN. Prediction and dissection of protein–RNA interactions by molecular descriptors. *Curr Top Med Chem* 2016;**16**:604–15. https://doi.org/10.2174/1568026615666150819110703.
38. Si JN, Cui J, Cheng J. *et al.* Computational prediction of RNA-binding proteins and binding sites. *Int J Mol Sci* 2015;**16**:26303–17. https://doi.org/10.3390/ijms161125952.
39. 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.
40. Basu S, Kihara D, Kurgan L. Computational prediction of disordered binding regions. *Comput Struct Biotechnol J* 2023;**21**:1487–97. https://doi.org/10.1016/j.csbj.2023.02.018.
41. Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol* 2015;**11**. https://doi.org/10.1371/journal.pcbi.1004639.
42. Dunker AK, Lawson JD, Brown CJ. *et al.* Intrinsically disordered protein. *J Mol Graph Model* 2001;**19**:26–59. https://doi.org/10.1016/S1093-3263(00)00138-8.
43. Tompa P, Fuxreiter M, Oldfield CJ. *et al.* Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* 2009;**31**:328–35. https://doi.org/10.1002/bies.200800151.
44. Lieutaud P, Ferron F, Uversky AV. *et al.* How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disord Proteins* 2016;**4**:e1259708. https://doi.org/10.1080/21690707.2016.1259708.
45. Habchi J, Tompa P, Longhi S. *et al.* Introducing protein intrinsic disorder. *Chem Rev* 2014;**114**:6561–88. https://doi.org/10.1021/cr400514h.
46. Basu S, Bahadur RP. A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell Mol Life Sci* 2016;**73**:4075–84. https://doi.org/10.1007/s00018-016-2283-1.
47. Wang C, Uversky VN, Kurgan L. Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from eukaryota, bacteria, and archaea. *Proteomics* 2016;**16**:1486–98. https://doi.org/10.1002/pmic.201500177.
48. Zhao B, Katuwawala A, Oldfield CJ. *et al.* Intrinsic disorder in human RNA-binding proteins. *J Mol Biol* 2021;**433**. https://doi.org/10.1016/j.jmb.2021.167229.
49. Varadi M, Zsolyomi F, Guharoy M. *et al.* Functional advantages of conserved intrinsic disorder in RNA-binding proteins. *PloS One* 2015;**10**. https://doi.org/10.1371/journal.pone.0139731.
50. Dyson J. Role of intrinsic disorder in protein-protein and protein–nucleic acid interactions. *FEBS Journal* 2012;**279**:9–9.
51. Munshi S, Gopi S, Asampille G. *et al.* Tunable order-disorder continuum in protein–DNA interactions. *Nucleic Acids Res* 2018;**46**:8700–9. https://doi.org/10.1093/nar/gky732.
52. Hsu WL, Oldfield CJ, Xue B. *et al.* Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci* 2013;**22**:258–73. https://doi.org/10.1002/pro.2207.
53. Fung HYJ, Birol M, Rhoades E. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Curr Opin Struct Biol* 2018;**49**:36–43. https://doi.org/10.1016/j.sbi.2017.12.007.
54. Williams RM, Obradovi Z, Mathura V. *et al.* The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput* 2001;89–100. https://doi.org/10.1142/9789814447362_0010.
55. Wu ZH, Hu G, Yang J. *et al.* In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces. *FEBS Lett* 2015;**589**:2561–9. https://doi.org/10.1016/j.febslet.2015.08.014.
56. Zhao B, Kurgan L. Compositional bias of intrinsically disordered proteins and regions and their predictions. *Biomolecules* 2022;**12**. https://doi.org/10.3390/biom12070888.
57. Zhao HY, Yang YD, Zhou YQ. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol* 2011;**8**:988–96. https://doi.org/10.4161/rna.8.6.17813.
58. Patiyal S, Dhall A, Raghava GPS. A deep learning-based method for the prediction of DNA interacting residues in a protein. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbac322.
59. Liu YF, Tian BX. Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning. *Brief Bioinform* 2024;**25**. https://doi.org/10.1093/bib/bbad488.
60. Zhang JS, Wang RH, Wei LY. MucLiPred: multi-level contrastive learning for predicting nucleic acid binding residues of proteins. *J Chem Inf Model* 2024;**64**:1050–65. https://doi.org/10.1021/acs.jcim.3c01471.
61. Qiu J, Bernhofer M, Heinzinger M. *et al.* ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J Mol Biol* 2020;**432**:2428–43. https://doi.org/10.1016/j.jmb.2020.02.026.

62. Patiyal S, Dhall A, Bajaj K. *et al.* Prediction of RNA-interacting residues in a protein using CNN and evolutionary profile. *Brief Bioinform* 2023;**24**. https://doi.org/10.1093/bib/bbac538.

63. Yang JY, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013;**41**:D1096–103. https://doi.org/10.1093/nar/gks966.

64. Zhang CX, Zhang X, Freddolino PL. *et al.* BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Res* 2024;**52**:D404–12. https://doi.org/10.1093/nar/gkad630.

65. Berman HM, Westbrook J, Feng Z. *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42. https://doi.org/10.1093/nar/28.1.235.

66. Burley SK, Bhikadiya C, Bi C. *et al.* RCSB protein data bank (RCSB.Org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res* 2023;**51**:D488–508. https://doi.org/10.1093/nar/gkac1077.

67. Norambuena T, Melo F. *The protein–DNA interface database*. BMC *Bioinformatics* 2010;**11**:11. https://doi.org/10.1186/1471-2105-11-262.

68. Lewis BA, Walia RR, Terribilini M. *et al.* PRIDB: a protein–RNA interface database. *Nucleic Acids Res* 2011;**39**:D277–82. https://doi.org/10.1093/nar/gkq1108.

69. Katuwawala A, Peng Z, Yang J. *et al.* Computational prediction of MoRFs, short disorder-to-order transitioning protein binding regions. *Comput Struct Biotechnol J* 2019;**17**:454–62. https://doi.org/10.1016/j.csbj.2019.03.013.

70. Del Conte A, Mehdiabadi M, Bouhraoua A. *et al.* Critical assessment of protein intrinsic disorder prediction (CAID)—results of round 2. *Proteins* 2023. https://doi.org/10.1002/prot.26582.

71. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinform* 2018;**19**:821–37. https://doi.org/10.1093/bib/bbx022.

72. Wang C, Kurgan L. Review and comparative assessment of similarity-based methods for prediction of drug–protein interactions in the druggable human proteome. *Brief Bioinform* 2019;**20**:2066–87. https://doi.org/10.1093/bib/bby069.

73. Katuwawala A, Oldfield CJ, Kurgan L. Accuracy of protein-level disorder predictions. *Brief Bioinform* 2020;**21**:1509–22. https://doi.org/10.1093/bib/bbz100.

74. Liu YC, Lin YJ, Chang YY. *et al.* Deciphering the language of protein–DNA interactions: a deep learning approach combining contextual embeddings and multi-scale sequence modeling. *J Mol Biol* 2024;**436**. https://doi.org/10.1016/j.jmb.2024.168769.

75. Zhang LR, Liu TG. PDNAPred: interpretable prediction of protein–DNA binding sites based on pre-trained protein language models. *Int J Biol Macromol* 2024;**281**. https://doi.org/10.1016/j.ijbiomac.2024.136432.

76. Andrabi M, Mizuguchi K, Sarai A. *et al.* Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. *BMC Struct Biol* 2009;**9**:30.

77. Fernandez M, Kumagai Y, Standley DM. *et al.* Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics* 2011;**12 Suppl 13**:S5. https://doi.org/10.1186/1471-2105-12-S13-S5.

78. Zhang J, Ghadermarzi S, Katuwawala A. *et al.* DNAgenie: accurate prediction of DNA-type-specific binding residues in protein sequences. *Brief Bioinform* 2021;**22**. https://doi.org/10.1093/bib/bbab336.

79. Nie WJ, Deng L. TSNAPred: predicting type-specific nucleic acid binding residues via an ensemble approach. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbac244.

80. Sun CZ, Feng YE. EPDRNA: a model for identifying DNA–RNA binding sites in disease-related proteins. *Protein J* 2024;**43**:513–21. https://doi.org/10.1007/s10930-024-10183-3.

81. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005;**6**:6. https://doi.org/10.1186/1471-2105-6-33.

82. 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**:W243–8. https://doi.org/10.1093/nar/gkl298.

83. Yan CH, Terribilini M, Wu F. *et al.* Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics* 2006;**7**:7. https://doi.org/10.1186/1471-2105-7-262.

84. Jeong EN, Miyano S. A weighted profile based method for protein–RNA interacting residue prediction. *Transactions on Computational Systems Biology IV* 2006;**3939**:123–39. https://doi.org/10.1007/11732488_11.

85. Hwang S, Gou ZK, Kuznetsov IB. DP-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* 2007;**23**:634–6. https://doi.org/10.1093/bioinformatics/btl672.

86. Kuznetsov IB, Gou Z, Li R. *et al.* Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins* 2006;**64**:19–27. https://doi.org/10.1002/prot.20977.

87. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. *Bioinformatics* 2007;**23**:I347–53. https://doi.org/10.1093/bioinformatics/btm174.

88. Ho SY, Yu FC, Chang CY. *et al.* Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. *Biosystems* 2007;**90**:234–41. https://doi.org/10.1016/j.biosystems.2006.08.007.

89. 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**:W578–84. https://doi.org/10.1093/nar/gkm294.

90. Terribilini M, Lee JH, Yan C. *et al.* Prediction of RNA binding sites in proteins from amino acid sequence. *RNA* 2006;**12**:1450–62. https://doi.org/10.1261/rna.2197306.

91. Kumar M, Gromiha AM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 2008;**71**:189–94. https://doi.org/10.1002/prot.21677.

92. Wang Y, Xue Z, Shen G. *et al.* PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids* 2008;**35**:295–302. https://doi.org/10.1007/s00726-007-0634-9.

93. Tong J, Jiang P, Lu ZH. RISP: a web-based server for prediction of RNA-binding sites in proteins. *Comput Methods Programs Biomed* 2008;**90**:148–53. https://doi.org/10.1016/j.cmpb.2007.12.003.

94. Cheng CW, Su ECY, Hwang JK. *et al.* Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics* 2008;**9**:9.

95. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* 2009;**10 Suppl 1**:S1. https://doi.org/10.1186/1471-2164-10-S1-S1.

96. Gao M, Skolnick J. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. *PLoS Comput Biol* 2009;**5**. https://doi.org/10.1371/journal.pcbi.1000567.

97. 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**:30–5. https://doi.org/10.1093/bioinformatics/btn583.
98. Chu WY, 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**:W396–401. https://doi.org/10.1093/nar/gkp449.
99. Murakami Y, Spriggs RV, Nakamura H. *et al.* PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res* 2010;**38**:W412–6. https://doi.org/10.1093/nar/gkq474.
100. Spriggs RV, Murakami Y, Nakamura H. *et al.* Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics* 2009;**25**:1492–7. https://doi.org/10.1093/bioinformatics/btp257.
101. 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 Suppl 1**:S3. https://doi.org/10.1186/1752-0509-4-S1-S3.
102. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res* 2010;**38**:W431–5. https://doi.org/10.1093/nar/gkq361.
103. Liu ZP, Wu LY, Wang Y. *et al.* Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics* 2010;**26**:1616–22. https://doi.org/10.1093/bioinformatics/btq253.
104. Huang YF, Chiu LY, Huang CC. *et al.* Predicting RNA-binding residues from evolutionary information and sequence conservation. *BMC Genomics* 2010;**11**:S2. https://doi.org/10.1186/1471-2164-11-S4-S2.
105. Zhang T, Zhang H, Chen K. *et al.* Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr Protein Pept Sci* 2010;**11**:609–28. https://doi.org/10.2174/138920310794109193.
106. Li Q, Cao Z, Liu H. Improve the prediction of RNA-binding residues using structural neighbours. *Protein Pept Lett* 2010;**17**: 287–96. https://doi.org/10.2174/092986610790780279.
107. Si J, Zhang Z, Lin B. *et al.* MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst Biol* 2011;**5 Suppl 1**:S7. https://doi.org/10.1186/1752-0509-5-S1-S7.
108. Ma X, Guo J, Wu J. *et al.* Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins* 2011;**79**: 1230–9. https://doi.org/10.1002/prot.22958.
109. Choi S, Han K. Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics* 2011;**12 Suppl 13**:S7. https://doi.org/10.1186/1471-2105-12-S13-S7.
110. Wang CC, Fang Y, Xiao J. *et al.* Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids* 2011;**40**:239–48. https://doi.org/10.1007/s00726-010-0639-7.
111. 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**: 1766–75. https://doi.org/10.1109/TCBB.2012.106.
112. Puton T, Kozlowski L, Tuszynska I. *et al.* Computational methods for prediction of protein-RNA interactions. *J Struct Biol* 2012;**179**:261–8. https://doi.org/10.1016/j.jsb.2011.10.001.
113. Yu DJ, Hu J, Yang J. *et al.* Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**:994–1008. https://doi.org/10.1109/TCBB.2013.104.
114. Pan XY, Zhu L, Fan YX. *et al.* Predicting protein–RNA interaction amino acids using random forest based on submodularity subset selection. *Comput Biol Chem* 2014;**53**:324–30. https://doi.org/10.1016/j.compbiolchem.2014.11.002.
115. Zhao HY, 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**. https://doi.org/10.1371/journal.pone.0096694.
116. Yang, Y.D., Zhao H., Wang J., Zhou Y., *SPOT-Seq-RNA: predicting protein–RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. Protein Structure Prediction*, 3rd Edition, 2014. **1137**: p. 119–30 https://doi.org/10.1007/978-1-4939-0366-5_9.
117. Walia RR, Xue LC, Wilkins K. *et al.* RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PloS One* 2014;**9**. https://doi.org/10.1371/journal.pone.0097725.
118. Xiong DP, Zeng JY, Gong HP. RBRIdent: an algorithm for improved identification of RNA-binding residues in proteins from primary sequences. *Proteins* 2015;**83**:1068–77. https://doi.org/10.1002/prot.24806.
119. Yang XX, Wang J, Sun J. *et al.* SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PloS One* 2015;**10**. https://doi.org/10.1371/journal.pone.0133260.
120. Peng ZL, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* 2015;**43**. https://doi.org/10.1093/nar/gkv585.
121. Peng ZL, Wang C, Uversky VN. *et al.* Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol Biol* 2017;**1484**:187–203. https://doi.org/10.1007/978-1-4939-6406-2_14.
122. Miao ZC, Westhof E. Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res* 2015;**43**:5340–51. https://doi.org/10.1093/nar/gkv446.
123. Dang TKL, Meckbach C, Tacke R. *et al.* A novel sequence-based feature for the identification of DNA-binding sites in proteins using Jensen–Shannon divergence. *Entropy* 2016;**18**. https://doi.org/10.3390/e18100379.
124. Chai H, Zhang J, Yang G. *et al.* An evolution-based DNA-binding residue predictor using a dynamic query-driven learning scheme. *Mol Biosyst* 2016;**12**:3643–50. https://doi.org/10.1039/C6MB00626D.
125. El-Manzalawy Y, Abbas M, Malluhi Q. *et al.* FastRNABindR: fast and accurate prediction of protein–RNA interface residues. *PloS One* 2016;**11**. https://doi.org/10.1371/journal.pone.0160395.
126. 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* 2017;**14**:1389–98. https://doi.org/10.1109/TCBB.2016.2616469.
127. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**. https://doi.org/10.1093/nar/gkx059.
128. Shen C, Ding Y, Tang J. *et al.* Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information. *Molecules* 2017;**22**. https://doi.org/10.3390/molecules22122079.

129. Zhou JY, Lu Q, Xu R. *et al.* EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation. *BMC Bioinformatics* 2017;**18**:18. https://doi.org/10.1186/s12859-017-1792-8.

130. Deng L, Pan J, Xu X. *et al.* PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine. *BMC Bioinformatics* 2018;**19**:19. https://doi.org/10.1186/s12859-018-2527-1.

131. Amirkhani A, Kolahdoozi M, Wang C. *et al.* Prediction of DNA-binding residues in local segments of protein sequences with fuzzy cognitive maps. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**:1372–82. https://doi.org/10.1109/TCBB.2018.2890261.

132. Zhu YH, 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**:3057–71. https://doi.org/10.1021/acs.jcim.8b00749.

133. 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**:930–6. https://doi.org/10.1093/bioinformatics/bty756.

134. Jung Y, El-Manzalawy Y, Dobbs D. *et al.* Partner-specific prediction of RNA-binding residues in proteins: a critical assessment. *Proteins* 2019;**87**:198–211. https://doi.org/10.1002/prot.25639.

135. Nguyen BP, Nguyen QH, Doan-Ngoc GN. *et al.* iProDNA-CapsNet: identifying protein–DNA binding residues using capsule neural networks. *BMC Bioinformatics* 2019;**20**. https://doi.org/10.1186/s12859-019-2604-0.

136. Zhou JY, Lu Q, Xu R. *et al.* EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning. *IEEE/ACM Trans Comput Biol Bioinform* 2020;**17**:124–35. https://doi.org/10.1109/TCBB.2018.2858806.

137. Yao LS, Wang HD, Bin YN. Predicting hot spot residues at protein–DNA binding interfaces based on sequence information. *Interdiscip Sci* 2021;**13**:1–11. https://doi.org/10.1007/s12539-020-00399-z.

138. Zhang J, Chen QC, Liu B. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Brief Bioinform* 2021;**22**. https://doi.org/10.1093/bib/bbaa397.

139. Littmann M, Heinzinger M, Dallago C. *et al.* Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci Rep* 2021;**11**. https://doi.org/10.1038/s41598-021-03431-4.

140. Sun Z, Zheng S, Zhao H. *et al.* To improve prediction of binding residues with DNA, RNA, carbohydrate, and peptide via multi-task deep neural networks. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:3735–43. https://doi.org/10.1109/TCBB.2021.3118916.

141. Zhang FH, Zhao B, Shi W. *et al.* DeepDISOBind: accurate prediction of RNA-, DNA- and protein-binding intrinsically disordered residues with deep multi-task learning. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbab521.

142. Hu J, Bai YS, Zheng LL. *et al.* Protein–DNA binding residue prediction via bagging strategy and sequence-based cube-format feature. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:3635–45. https://doi.org/10.1109/TCBB.2021.3123828.

143. Wang N, Yan K, Zhang J. *et al.* iDRNA-ITF: identifying DNA-and RNA-binding residues in proteins based on induction and transfer framework. *Brief Bioinform* 2022;**23**. https://doi.org/10.1093/bib/bbac236.

144. Guan S, Zou Q, Wu H. *et al.* Protein–DNA binding residues prediction using a deep learning model with hierarchical feature extraction. *IEEE/ACM Trans Comput Biol Bioinform* 2023;**20**:2619–28. https://doi.org/10.1109/TCBB.2022.3190933.

145. Agarwal A, Kant S, Bahadur RP. Efficient mapping of RNA-binding residues in RNA-binding proteins using local sequence features of binding site residues in protein-RNA complexes. *Proteins* 2023;**91**:1361–79. https://doi.org/10.1002/prot.26528.

146. Zhang F, Li M, Zhang J. *et al.* HybridRNAbind: prediction of RNA interacting residues across structure-annotated and disorder-annotated proteins. *Nucleic Acids Res* 2023;**51**:e25. https://doi.org/10.1093/nar/gkac1253.

147. Song YD, Yuan Q, Zhao H. *et al.* Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures. *Brief Bioinform* 2023;**24**. https://doi.org/10.1093/bib/bbad360.

148. Zhang J, Basu S, Kurgan L. HybridDBRpred: improved sequence-based prediction of DNA-binding amino acids using annotations from structured complexes and disordered proteins. *Nucleic Acids Res* 2024;**52**:e10. https://doi.org/10.1093/nar/gkad1131.

149. Kabir MWU, Alawad DM, Pokhrel P. *et al.* DRBpred: a sequence-based machine learning method to effectively predict DNA-and RNA-binding residues. *Comput Biol Med* 2024;**170**:108081. https://doi.org/10.1016/j.compbiomed.2024.108081.

150. Zhu YH, Liu Z, Liu Y. *et al.* ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction. *Brief Bioinform* 2024;**25**. https://doi.org/10.1093/bib/bbae040.

151. Zhang B, Hou Z, Yang Y. *et al.* SOFB is a comprehensive ensemble deep learning approach for elucidating and characterizing protein–nucleic-acid-binding residues. *Commun Biol* 2024;**7**:679. https://doi.org/10.1038/s42003-024-06332-0.

152. Yuan QM, Tian C, Song Y. *et al.* GPSFun: geometry-aware protein sequence function predictions with language models. *Nucleic Acids Res* 2024;**52**:W248–55. https://doi.org/10.1093/nar/gkae381.

153. Yuan QM, Tian C, Yang YD. Genome-scale annotation of protein binding sites via language model and geometric deep learning. *Elife* 2024;**13**:13. https://doi.org/10.7554/eLife.93695.3.

154. Daanial Khan Y, Alkhalifah T, Alturise F. *et al.* DeepDBS: identification of DNA-binding sites in protein sequences by using deep representations and random forest. *Methods* 2024;**231**:26–36. https://doi.org/10.1016/j.ymeth.2024.09.004.

155. Yu LJ, Tanwar DK, Penha EDS. *et al.* Grammar of protein domain architectures. *Proc Natl Acad Sci U S A* 2019;**116**:3636–45. https://doi.org/10.1073/pnas.1814684116.

156. Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* 2021;**19**:1750–8. https://doi.org/10.1016/j.csbj.2021.03.022.

157. Zhang S, Fan R, Liu Y. *et al.* Applications of transformer-based language models in bioinformatics: a survey. *Neuro-Oncology Advances* 2023;**5**. https://doi.org/10.1093/bioadv/vbad001.

158. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS One* 2015;**10**. https://doi.org/10.1371/journal.pone.0141287.

159. Elnaggar A, Heinzinger M, Dallago C. *et al.* ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.

160. Zhu Y, Wang G. CAN-NER: convolutional attention network for Chinese named entity recognition. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl Hlt 2019)* 2019;**Vol. 1**:3384–93.

161. Lin ZM, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574.

162. Rao RM. *et al. MSA Transformer*, in *Proceedings of the 38th International Conference on Machine Learning*. In: Marina M, Tong Z, (eds.), *PMLR: Proceedings of Machine Learning Research*, 2021, 8844–56.

163. Ghosh S, Bagchi A. Structural study to analyze the DNA-binding properties of DsrC protein from the dsr operon of sulfur-oxidizing bacterium Allochromatium vinosum. *J Mol Model* 2019;**25**. https://doi.org/10.1007/s00894-019-3945-3.

164. Wang J, Ma X, Hu Y. *et al.* Regulation of micro- and small-exon retention and other splicing processes by GRP20 for flower development. *Nat Plants* 2024;**10**:66–85. https://doi.org/10.1038/s41477-023-01605-8.

165. Zhang Z, Li B, Wang Z. *et al.* Novel LncRNA LINC02936 suppresses ferroptosis and promotes tumor progression by interacting with SIX1/CP Axis in endometrial cancer. *Int J Biol Sci* 2024;**20**:1356–74. https://doi.org/10.7150/ijbs.86256.

166. Chong, Qui E, Habtehyimer F, Germroth A. *et al.* Mycobacteriophage Alexphander gene 94 encodes an essential dsDNA-binding protein during lytic infection. *Int J Mol Sci* 2024;**25**. https://doi.org/10.3390/ijms25137466.

167. Yadhav Y, Selvaraj K, Ramasamy S. *et al.* Computational studies on rep and capsid proteins of CRESS DNA viruses. *Virus* 2024;**35**: 17–26. https://doi.org/10.1007/s13337-024-00858-x.

168. Reinar WB, Greulich A, Stø IM. *et al.* Adaptive protein evolution through length variation of short tandem repeats in. *Sci Adv* 2023;**9**. https://doi.org/10.1126/sciadv.add6960.

169. Lanclos N, Radulovic P, Bland J. *et al.* Implications of intrinsic disorder and functional proteomics in the merkel cell polyomavirus life cycle. *J Cell Biochem* 2023;**125**. https://doi.org/10.1002/jcb.30485.

170. 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* 2020.

171. Basu S, Zhao B, Biró B. *et al.* DescribePROT in 2023: more, higher-quality and experimental annotations and improved data download options. *Nucleic Acids Res* 2024;**52**:D426–33. https://doi.org/10.1093/nar/gkad985.

172. Zhao B, Katuwawala A, Oldfield CJ. *et al.* DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Res* 2021;**49**:D298–308. https://doi.org/10.1093/nar/gkaa931.

173. Song J, Kurgan L. Availability of web servers significantly boosts citations rates of bioinformatics methods for protein function and disorder prediction. *Bioinform Adv* 2023;**3**:vbad184. https://doi.org/10.1093/bioadv/vbad184.

174. Bileschi ML, Bryant D, Sanderson T. *et al.* Using deep learning to annotate the protein universe. *Nat Biotechnol* 2022;**40**:932–7. https://doi.org/10.1038/s41587-021-01179-w.

175. Boadu F, Lee AHY, Cheng JL. Deep learning methods for protein function prediction. *Proteomics* 2024. https://doi.org/10.1002/pmic.202300471.

176. Zhao B, Kurgan L. Deep learning in prediction of intrinsic disorder in proteins. *Comput Struct Biotechnol J* 2022;**20**:1286–94. https://doi.org/10.1016/j.csbj.2022.03.003.

177. Vucetic S, Obradovic Z, Vacic V. *et al.* DisProt: a database of protein disorder. *Bioinformatics* 2005;**21**:137–40. https://doi.org/10.1093/bioinformatics/bth476.

178. Aspromonte MC. *et al.* DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res* 2023.

179. Necci M, Piovesan D., CAID Predictors *et al.* Critical assessment of protein intrinsic disorder prediction. *Nat Methods* 2021;**18**: 472–81. https://doi.org/10.1038/s41592-021-01117-3.

180. Piovesan D, Tabaro F, Mičetić I. *et al.* DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res* 2016;**D1**:D219–27. https://doi.org/10.1093/nar/gkw1056.

181. Katuwawala A, Ghadermarzi S, Kurgan L. Computational prediction of functions of intrinsically disordered regions. *Prog Mol Biol Transl Sci* 2019;**166**:341–69. https://doi.org/10.1016/bs.pmbts.2019.04.006.

182. Aspromonte MC, Nugnes MV, Quaglia F. *et al.* DisProt in 2024: improving function annotation of intrinsically disordered proteins. *Nucleic Acids Res* 2024;**52**:D434–41. https://doi.org/10.1093/nar/gkad928.

183. 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**:4729–38. https://doi.org/10.1093/bioinformatics/btaa573.

184. Hsieh J, Andrews AJ, Fierke CA. Roles of protein subunits in RNA-protein complexes: lessons from ribonuclease P. *Biopolymers* 2004;**73**:79–89. https://doi.org/10.1002/bip.10521.

185. Madru C, Lebaron S, Blaud M. *et al.* Chaperoning 5S RNA assembly. *Genes Dev* 2015;**29**:1432–46. https://doi.org/10.1101/gad.260349.115.

186. Wetzel JL, Zhang KQ, Singh M. Learning probabilistic protein–DNA recognition codes from DNA-binding specificities using structural mappings. *Genome Res* 2022;**32**:1776–86. https://doi.org/10.1101/gr.276606.122.

187. Mitra R, Li J, Sagendorf JM. *et al.* Geometric deep learning of protein–DNA binding specificity. *Nat Methods* 2024;**21**:1674–83. https://doi.org/10.1038/s41592-024-02372-w.

188. Christensen RG, Enuameh MS, Noyes MB. *et al.* Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics* 2012;**28**:I84–9. https://doi.org/10.1093/bioinformatics/bts202.

189. 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.

190. Yanover C, Bradley P. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for CH zinc fingers. *Nucleic Acids Res* 2011;**39**:4564–76. https://doi.org/10.1093/nar/gkr048.

191. Pang YH, Liu B. IDP-LM: prediction of protein intrinsic disorder and disorder functions based on language models. *PLoS Comput Biol* 2023;**19**. https://doi.org/10.1371/journal.pcbi.1011657.

192. Madani A, Krause B, Greene ER. *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* 2023;**41**:1099–106. https://doi.org/10.1038/s41587-022-01618-2.

193. Shuai RW, Ruffolo JA, Gray JJ. IgLM: infilling language modeling for antibody sequence design. *Cell Systems* 2023;**14**:979–989.e4. https://doi.org/10.1016/j.cels.2023.10.001.

194. Dong B, Liu Z, Xu D. *et al.* Impact of multi-factor features on protein secondary structure prediction. *Biomolecules* 2024;**14**. https://doi.org/10.3390/biom14091155.

195. Zhang, J., Qian J., Zou Q., Zhou F., Kurgan L., Recent advances in computational prediction of secondary and supersecondary structures from protein sequences, in *Protein Supersecondary Structures: Methods and Protocols*, A.E. Kister, Editor. 2025, Springer US: New York, NY. p. 1–19 https://doi.org/10.1007/978-1-0716-4213-9_1.

196. Jiang Q, Jin X, Lee SJ. *et al.* Protein secondary structure prediction: a survey of the state of the art. *J Mol Graph Model* 2017;**76**:379–402. https://doi.org/10.1016/j.jmgm.2017.07.015.
197. Mirabello C, Wallner B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. *PloS One* 2019;**14**. https://doi.org/10.1371/journal.pone.0220182.
198. Zhao B, Kurgan L. Surveying over 100 predictors of intrinsic disorder in proteins. *Expert Rev Proteomics* 2021;**18**:1019–29. https://doi.org/10.1080/14789450.2021.2018304.
199. Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform* 2019;**20**:330–46. https://doi.org/10.1093/bib/bbx126.
200. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. *Proteins* 2002;**46**:197–205. https://doi.org/10.1002/prot.10029.
201. Kulandaisamy A, Srivastava A, Nagarajan R. *et al.* Dissecting and analyzing key residues in protein–DNA complexes. *J Mol Recognit* 2018;**31**. https://doi.org/10.1002/jmr.2692.
202. Barik A, C N, Pilla SP. *et al.* Molecular architecture of protein-RNA recognition sites. *J Biomol Struct Dyn* 2015;**33**:2738–51. https://doi.org/10.1080/07391102.2015.1004652.
203. Baker CM, Grant GH. Role of aromatic amino acids in protein–nucleic acid recognition. *Biopolymers* 2007;**85**:456–70. https://doi.org/10.1002/bip.20682.
204. Wilson KA, Holland DJ, Wetmore SD. Topology of RNA–protein nucleobase-amino acid π-π interactions and comparison to analogous DNA-protein π-π contacts. *RNA* 2016;**22**:696–708. https://doi.org/10.1261/rna.054924.115.
205. Wilson KA, Kung RW, D'souza S. *et al.* Anatomy of noncovalent interactions between the nucleobases or ribose and π-containing amino acids in RNA–protein complexes. *Nucleic Acids Res* 2021;**49**:2213–25. https://doi.org/10.1093/nar/gkab008.
206. Zhang F, Shi W, Zhang J. *et al.* PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection. *Bioinformatics* 2020;**36**:i735–44. https://doi.org/10.1093/bioinformatics/btaa806.
207. Tuszynska I, Magnus M, Jonak K. *et al.* NPDock: a web server for protein–nucleic acid docking. *Nucleic Acids Res* 2015;**43**:W425–30. https://doi.org/10.1093/nar/gkv493.
208. Tuszynska I, Bujnicki JM. DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC Bioinformatics* 2011;**12**:348. https://doi.org/10.1186/1471-2105-12-348.
209. Yan Y, Zhang D, Zhou P. *et al.* HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* 2017;**45**:W365–73. https://doi.org/10.1093/nar/gkx407.
210. Zheng J, Hong X, Xie J. *et al.* P3DOCK: a protein–RNA docking webserver based on template-based and template-free docking. *Bioinformatics* 2020;**36**:96–103. https://doi.org/10.1093/bioinformatics/btz478.
211. Christoffer C, Kihara D. Modeling protein–nucleic acid complexes with extremely large conformational changes using flex-LZerD. *Proteomics* 2023;**23**:e2200322. https://doi.org/10.1002/pmic.202200322.
212. Abramson J, Adler J, Dunger J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**. https://doi.org/10.1038/s41586-024-07487-w.

Briefings in Bioinformatics, 2025, **26(1)**, bbaf016
https://doi.org/10.1093/bib/bbaf016
Review