OXFORD

Briefings in Bioinformatics, 17(1), 2016, 88–105

doi: 10.1093/bib/bbv023
Advance Access Publication Date: 1 May 2015
Paper

A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues

Jing Yan, Stefanie Friedrich and Lukasz Kurgan

Corresponding author. Lukasz Kurgan, University of Alberta, 9107 116 Street, Edmonton, Alberta, Canada T6G 2V4. Tel: +1-780-492-5488; Fax: +1-780-492-1811; Email: lkurgan@ece.ualberta.ca

Abstract

Motivated by the pressing need to characterize protein–DNA and protein–RNA interactions on large scale, we review a comprehensive set of 30 computational methods for high-throughput prediction of RNA- or DNA-binding residues from protein sequences. We summarize these predictors from several significant perspectives including their design, outputs and availability. We perform empirical assessment of methods that offer web servers using a new benchmark data set characterized by a more complete annotation that includes binding residues transferred from the same or similar proteins. We show that predictors of DNA-binding (RNA-binding) residues offer relatively strong predictive performance but they are unable to properly separate DNA- from RNA-binding residues. We design and empirically assess several types of consensuses and demonstrate that machine learning (ML)-based approaches provide improved predictive performance when compared with the individual predictors of DNA-binding residues or RNA-binding residues. We also formulate and execute first-of-its-kind study that targets combined prediction of DNA- and RNA-binding residues. We design and test three types of consensuses for this prediction and conclude that this novel approach that relies on ML design provides better predictive quality than individual predictors when tested on prediction of DNA- and RNA-binding residues individually. It also substantially improves discrimination between these two types of nucleic acids. Our results suggest that development of a new generation of predictors would benefit from using training data sets that combine both RNA- and DNA-binding proteins, designing new inputs that specifically target either DNA- or RNA-binding residues and pursuing combined prediction of DNA- and RNA-binding residues.

Key words: DNA-binding proteins; transcription factors; RNA-binding proteins; protein–DNA binding; protein–RNA binding; protein–nucleic acids binding

Introduction

Interplay of proteins, DNA and RNA defines and regulates many cellular-level activities. DNA-binding proteins are driving regulation of gene expression and DNA transcription, replication and repair [1, 2]. The RNA-binding proteins that interact with several types of RNAs, such as mRNA, tRNA and rRNA, are involved in a variety of cellular functions including protein synthesis, regulation of gene expression, posttranscriptional modifications and posttranscriptional regulation [3–5]. The number of DNA-binding proteins in a genome is relatively substantial and was estimated to be on average close to 3% of a eukaryotic genome and 5% of an animal genome, which translates to about 800 proteins per animal genome [2]. Similarly, the fraction of RNA-binding proteins was estimated to range between 2 and 8% of eukaryotic genomes [5]. Experimental determination of protein–DNA and protein–RNA interactions is technically challenging and relatively expensive and thus only a small fraction of these interactions was characterized so far. With the recent rapid accumulation of the protein, DNA and RNA data, i.e. as of November 2014, NCBI's RefSeq database [6] includes >9 million of DNA and RNA transcripts and about 47 million nonredundant

Jing Yan is a Ph.D. candidate at the University of Alberta. Her research concentrates on the development and applications of high-throughput structural bioinformatics methods.
Stefanie Friedrich is Master's student at the Stockholm University and pre-PhD student at Science for Life Laboratory in Solna, Sweden. Her research interests are in the areas of machine learning, pattern detection and data mining methods applied to bioinformatics.
Lukasz Kurgan is a Professor at the University of Alberta. His research group focuses on high-throughput structural and functional characterization of proteins and small RNAs.

Submitted: 19 December 2014; **Received (in revised form):** 16 March 2015

proteins from 49 000 organisms (source: http://www.ncbi.nlm.nih.gov/refseq/), there is a pressing need to increase the pace of the characterization of protein–DNA and protein–RNA inter actions. To this end, the experimental data are being used to de velop time- and cost-efficient computational models that could be used to perform automated prediction of these interactions for the millions of the uncharacterized proteins.

A number of models that predict the protein–DNA/–RNA interactions from the protein sequence and structure have been published and reviewed in the literature over the past several years [7–10]. Some efforts have been also recently made to predict protein-binding nucleotides in the RNA sequences using similar types of methodologies as for the prediction of the nucleotide-binding residues in proteins [11, 12]. Our focus is on the computational prediction of DNA- and RNA-binding residues from protein chains. These methods can be used to find the binding proteins in the vast sequence databases and to indicate sites of these interactions. Table 1 summarizes recent comparative reviews of the predictors of DNA-binding residues [13, 14] and RNA-binding residues [7, 15, 16]. These comparative analyses provide useful clues about the predictive performance of these predictors and help the end users to select a suitable method from among many available choices. However, these reviews and the corresponding predictive models focus solely on the prediction of interactions with just one of the two nucleic acids types. They do not consider how well they separate between DNA and RNA interactions. In other words, they do not test specificity of these predictions when applying a method for the prediction of DNA-binding residues to predict RNA-binding residues and vice versa. This is an important oversight given similarity between DNA and RNA molecules and the fact that the end users would not want them to be confused. Another drawback of the prior comparative reviews is that they consider data sets with incomplete annotations of binding residues. This is because the annotations are based on a single structure of protein–DNA or protein–RNA complex, which could be partial if only a fragment of DNA or RNA is considered in a given complex or if the same protein is involved in other binding events.

Our review addressed these two drawbacks and offers a considerably expanded scope (Table 1). We review a more complete set of sequence-based predictors of DNA- and RNA-binding residues, including newer methods. We discuss how they define binding residues, overview their predictive models and summarize their outputs, which was neglected in the prior studies. We perform comparative evaluation of predictive quality of methods that are conveniently available as web servers and which are runtime-efficient. This assessment is based on three new benchmark data sets that consider DNA-binding proteins, RNA-binding proteins and for the first time a combined set of DNA- and RNA-binding proteins. The data sets use protein–DNA and protein–RNA complexes released after September 2010 (date when the newest data set used by the published methods was collected) where the proteins have low similarity to the proteins in the complexes that were released before that date. This way the benchmark data are fair between predictors, and the corresponding predictive performance concerns novel (in regards to the interactions with DNA and RNA) proteins. Importantly, our evaluation uses two commonly considered definitions of binding residues (based on cutoff distances of 3.5 Å and 5 Å between atoms of protein and the nucleic acids) and considers a more complete set of binding annotations, which are transferred between multiple complexes with the same or similar protein.

Table 1. Summary and comparison of recent reviews concerning prediction of DNA- and RNA-binding residues from protein sequences

Review article (year published)	Scope of descriptive component					Scope of empirical component						
	Coverage	# methods	Year published of newest method	Defines binding	Discusses outputs of methods	Benchmark data set used				Considers cross prediction (RNA-binding on DNA-binding proteins and vice versa)	Evaluates consensus methods	Considers combined DNA- and RNA-binding prediction
						Year collected	# proteins	Cutoff(s) to define binding	Complete annotations of binding			
This review	DNA and RNA	30 (14 + 16)	2012	yes	yes	2013	531	3.5; 5	yes	yes	yes	yes
[7] (2013)	RNA	10	2011	no	no	2012	106	undefined	no	no	no	no
[14] (2013)	DNA	11	2011	no	no	2012	301	3.5	no	no	no	no
[16] (2012)	RNA	13	2011	yes	no	2010	198	5	no	no	no	no
[15] (2012)	RNA	7	2011	no	no	2011	44	3.5	no	no	yes	no
[13] (2011)	DNA	6	2009	yes	no	2010	232	3.5; 4; 4.5; 5; 5.5; 6	no	no	yes	no

Moreover, we design and assess several types of consensuses including a logic-based approach, majority vote and an extension of the majority vote in the form of a machine learning (ML)-based consensus that uses regression. We show that logic-based and majority vote consensuses do not offer improvements when tested on our challenging (low similarity) benchmark set, while the more sophisticated ML-based consensus provides improved predictive quality. We also propose a new approach to combine predictions of DNA- and RNA-binding residues in which the four possible outcomes include 'DNA&RNA-binding', 'DNA-binding', 'RNA-binding' and 'non-binding'. We develop a first-of-its-kind ML-based prototype predictor of these four outcomes. This method achieves relatively good predictive performance and outperforms a naive approach that directly merges predictions of DNA-binding and RNA-binding residues. In particular it reduces the rate of mispredictions between DNA and RNA binding residues. We also empirically demonstrate that the ML consensuses provide better predictions for longer binding regions and that they predict relatively few binding residues in proteins that do not interact with DNA or RNA.

Materials and methods

Benchmark data sets

Similar to other studies [7, 13–16], the benchmark data sets were extracted from structures of protein–DNA and protein–RNA complexes collected from the Protein Data Bank (PDB)[17]; these data were obtained in September 2013. The definitions of the binding residues differ between studies, with the most prevalent approach based on a cutoff distance, i.e. a given residue is considered as binding if at least one of its side chain or backbone atoms is closer than the cutoff from an atom of the RNA/DNA molecule [18]. Table 1 ['cutoff(s) to define binding' column] and Table 2 ('cutoff' column) reveal that the prior comparative reviews and 29 of 30 predictors of binding residues use this definition, although the cutoff values used vary considerably. We apply two frequently used values at 3.5 Å and 5 Å to accommodate for these differences. We note that the 3.5 Å is used most often when designing the prediction methods (13 of 30 methods in Table 2), while both of the considered thresholds were used in the prior reviews (Table 1). We collected total of 1082 high-quality X-ray structures (resolution better than 2.5 Å) of protein–DNA complexes, 271 protein–RNA complexes and 4 complexes that include both DNA and RNA. These complexes are split into chains and the chains that have no binding residues or are shorter than 30 amino acids in length are removed. As a result, we obtained 1935 (1939) DNA-binding chains and 981 (985) RNA-binding chains for distance cutoff of 3.5 Å (5 Å).

Motivated by a recent work that evaluated predictive quality of methods that find small ligand binding pockets on the protein surface [51], we improve the annotations of binding residues by transferring these annotations between similar proteins. This similarity stems from the fact that the structures of protein–DNA and protein–RNA complexes could concern paralogs, similar or the same proteins in different organisms, and structures of the same proteins solved at different resolutions or with different co-factors. Using the procedure introduced in [51], we first find proteins that are similar in their structure and sequence. The structural similarity is expressed with the template modeling (TM) score [52]. The similarity in the sequence is measured with the sequence identity expressed as a fraction of aligned residues over the length of the shorter sequence, where the alignment is calculated using bl2seq [53] with default parameters; we only consider the aligned proteins for which *e*-value <0.001. The two similarity scores are used to perform clustering of protein chains where two chains are assigned into the same cluster if their TM score >0.5 and the sequence identity $>80\%$ [51]. The chains in the same cluster are assumed to be sufficiently similar and are represented by one chain with the largest number of binding residues. The annotations of binding residues of the remaining chains in the same cluster are transferred into this chain. This is done based on the alignment with bl2seq (*e*-value <0.001) where annotations are transferred for positions that are matched in the alignment. As a result of the transfer, the number of annotated DNA-binding residues was enlarged by 13.7 and 7.9% for the annotations based on the 3.5 Å and 5 Å threshold, respectively. Similarly, the number of RNA-binding residues increased by 9.7 and 4.7%, respectively.

The original redundant data sets were reduced after the clustering to the nonredundant data set of 356 (359) DNA-binding proteins, and 175 (175) RNA-binding proteins based on the cutoff at 3.5 Å (5 Å). We split them into training and test proteins based on their release date. We observe that the data sets used by the considered predictors of DNA- and RNA-binding residues were collected before September 2010. Correspondingly, the binding proteins released before September 2010 constitute the training set, which we use to select and compute consensuses. The proteins released after September 2010 are less likely to be used to train the published methods. Furthermore, we reduce this set of proteins by excluding those that are similar to the training proteins. Using CD-HIT [54], we clustered all training and test protein chains sharing $\geq 30\%$ sequence similarity and we removed all test proteins that are in the same cluster with any of the training chains. The remaining test proteins share $<30\%$ sequence similarity with training proteins. Based on the cutoff at 3.5 Å (5 Å), our training data set contains 293 (295) DNA-binding proteins and 149 (149) RNA-binding proteins. The test proteins were used to establish the following data sets: 'DNA_T' test data set that includes 47 (48) DNA-binding proteins, 'RNA_T' test data set that contains 17 (17) RNA-binding proteins, and the combined 'COMB_T' test data set that has 64 (65) nucleic acid-binding proteins when applying the cutoff at 3.5 Å (5 Å); 'T' denotes the fact that the annotations were transferred. We also define the corresponding three test data sets where the annotations were not transferred: 'DNA_NT', 'RNA_NT' and 'COMB_NT'. The full set of both training and test chains that includes annotations with and without transfer and for both cutoff is available in the Supplementary data; these data can be used to derive the six data sets, which would facilitate future comparative studies.

We also developed a data set that includes proteins that are unlikely to interact with DNA/RNA to investigate whether binding residues would be still predicted in these chains. We consider human proteins from the complete human proteome collected from the UniProt [55]. We included proteins that satisfy the following constrains: not located in nucleus; not annotated as DNA-binding, RNA-binding and nucleic acid-binding in Gene Ontology [56]; not having DNA, RNA, nucleic acid or nucleotide terms in protein name; not having the DNA, RNA, nucleic acid or nucleotide keyword. We clustered the resulting set of 12 361 human proteins using CD-HIT with 30% sequence similarity and selected one chain from each cluster to remove similarity that could bias evaluation. Next, we removed chains that share $\geq 30\%$ sequence similarity with any chains in our training set to further reduce possibility that these chains bind

Table 2. Summary of predictors of DNA and RNA binding residues

Method	Year	Ref[a]	Cut-off	Considered types of input features[b]											Prediction model[c]	Web server[d]		
				AC	PP	PA	PS	SA	PSSM	Max Hom	Wild Span	StL	SeL	WS		URL	Output	
																	bin	pr
Predictors of DNA binding residues																		
DBS-pred	2004	[18]	3.5	✓										3	NN	www.abren.net/dbs-pred/	✓	
DBS-PSSM	2005	[19]	3.5						✓					5	NN	dbspssm.netasa.org	✓	✓
BindN	2006	[20]	3.5		✓									11	SVM	bioinfo.ggc.org/bindn/	✓	✓
Ho *et al.*	2007	[21]	3.5						✓					7	SVM	N/A		
DP-Bind	2006, 2007	[22, 23]	4.5						✓					7	SVM, KLR, PLR	lcg.rit.albany.edu/dp-bind	✓	✓
DISIS	2007	[24]	6.0			✓	✓	✓		✓				3, 9	NN & SVM	cubic.bioc.columbia.edu/services/disis[g]		
DNABindR	2006, 2008	[25, 26]	1.0[e]	✓										21	Naïve Bayes	turing.cs.iastate.edu/PredDNA/index.html [g]		
BindN-RF	2009	[27]	3.5		✓			✓	✓					11	RF	bioinfo.ggc.org/bindn-rf/ [g]		
DBindR	2009	[28]	3.5	✓			✓		✓					11	RF	www.cbi.seu.edu.cn/DBindR/DBindR.htm [g]		
DBD-Threader	2009	[29]	4.5									✓		N/A	Template-based	cssb.biology.gatech.edu/skolnick/ webservice/DBD-Threader/index.html	✓	
ProteDNA	2009	[30]	4.5[f]						✓				✓	11	SVM	protedna.csie.ntu.edu.tw/method.php	✓	
BindN+	2010	[31]	3.5		✓			✓	✓					11	SVM	bioinfo.ggc.org/bindn+/	✓	✓
NAPS	2010	[32]	4.5	✓	✓				✓					7	C4.5	proteomics.bioengr.uic.edu/NAPS/ [g]		
DNABR	2012	[33]	3.5	✓	✓				✓					9	RF	www.cbi.seu.edu.cn/DNABR/ [g]	✓	✓
Predictors of RNA binding residues																		
Jeong *et al.*	2004	[34]	6.0	✓			✓							41	NN	N/A		
Jeong *et al.*	2006	[35]	6.0						✓					15	NN	N/A		
BindN	2006	[20]	3.5		✓									11	SVM	bioinfo.ggc.org/bindn/	✓	✓
PRINTR	2008	[36]	ENTANGLE				✓		✓					15	SVM	210.42.106.80/printr/ [g]		
RISP	2008	[37]	3.5						✓					7	SVM	grc.seu.edu.cn/RISP [g]		
Pprint	2008	[38]	6.0						✓					11,13,15	SVM	www.imtech.res.in/raghava/pprint/	✓	✓
RNAProB	2008	[39]	6, 5,3.5						✓					25	SVM	N/A		
BindN+	2010	[31]	3.5		✓			✓	✓					11	SVM	bioinfo.ggc.org/bindn+/	✓	✓
PiRaNhA	2009, 2010	[40, 41]	3.9		✓	✓			✓					23	SVM	www.bioinformatics.sussex.ac.uk/PIRANHA [g]		
NAPS	2010	[32]	4.5	✓	✓				✓					7	C4.5	proteomics.bioengr.uic.edu/NAPS [g]		
ProteRNA	2010	[42]	5.0				✓		✓		✓			23	SVM	N/A		
RBRpred	2010	[43]	6.0	✓		✓	✓		✓					15	SVM	N/A		
Wang *et al.*	2011	[44]	6.0		✓	✓			✓					15	SVM	N/A		
PRBR	2011	[45]	3.5		✓		✓		✓					11	RF	www.cbi.seu.edu.cn/PRBR/ [g]	✓	✓
SPOT-Seq	2011	[46]	4.5									✓		N/A	Template-based	sparks.informatics.iupui.edu	✓	
RNABindR	2006, 2007, 2012	[16, 47, 48]	5.0						✓					25	SVM	einstein.cs.iastate.edu/RNABindR/	✓	✓

Methods used in the empirical assessment are shown in bold.

[a] Ref – reference.

[b] AC – amino acid composition; PP – physiochemical properties of amino acids; PA – predicted solvent accessibility (ASA); PS – predicted secondary structure; SA – sequence alignment; PSSM – position-specific scoring matrix; MaxHom – MaxHom algorithm [49]; WildSpan – WildSpan algorithm [50]; StL –template library of structures; SeL –template library of sequences; WS – window size.

[c] NN – neural network; SVM – support vector machine; KLR – kernel logistic regression; PLR – penalized logistic regression; RF – random forest; C4.5 – decision tree.

[d] bin – outputs binary prediction; pr – outputs numeric propensity score.

[e] An amino acid is a DNA-binding residue if its ASA computed in the protein-DNA complex using NACCESS is smaller than its ASA in the unbounded protein by at least 1Å^2.

[f] A residue is regarded as involved in sequence-specific binding with the DNA if one or more heavy atoms in its side chain fall within 4.5 Å from the nucleobases of the DNA.

[g] Web server was not available as of December 2013 when the predictions were collected.

to RNA or DNA. To do that we clustered the human proteins with proteins from the training data set using CD-HIT at 30% similarity and we kept only the 6559 human proteins that were not located in the same cluster with any chains from our training set. To have the size similar to the size of our test data sets and to accommodate for the computational cost of predictions, we selected at random 50 proteins from this set of human proteins to form the data set of the nonbinding proteins.

Selection of methods included in the empirical assessment

The empirical assessment includes sequence-based methods for the prediction of DNA- and RNA-binding residues that were selected from the comprehensive list of 30 methods shown in Table 2. We selected nine predictors that were available as web servers as of December 2013 when the predictions were collected and which are runtime-efficient, i.e. they predict an average size protein sequence with 200 residues in under 10 min; this assures that we cover methods that are convenient to use for the end users. We use the most recent versions of methods that have multiple versions. We include four predictors of DNA-binding residues, DBS-PSSM [19], DP-Bind [22, 23], ProteDNA [30] and BindN+ [31], and three for the predictions of RNA-binding residues, Pprint [38], BindN+ [31] and RNABindR [16, 47, 48]. DP-Bind implements a family of methods that includes three ML models, support vector machine (SVM), kernel logistic regression (KLR) and penalized logistic regression, and two types of consensuses of these models [23]. We consider the default KLR classifier-based model, DP-Bind(klr), and the default majority-vote-based consensus, DP-Bind(maj). ProteDNA offers predictions in two modes, one with high-precision and another balanced; we use the latter version, ProteDNA(B), that provides a better balance between sensitivity and specificity [30]. We also consider two recent consensus-based approaches, which combine predictions of multiple methods: MetaDBSite [14] for the DNA-binding and the consensus by Puton et al. [15] for the RNA-binding. In total we examine 10 predictors, including three consensus-based approaches, which cover a comprehensive range of designs. These methods include a variety of predictive algorithms (Table 2), such as neural networks, SVMs, regression, Bayesian classifiers and consensuses, and they make use of several different types of inputs, such as evolutionary profiles, sequence alignment, composition of amino acids and physiochemical properties of amino acids.

From the list of recent methods we exclude DBS-pred [18] and BindN [20], which were superseded by DBS-PSSM and BindN+, respectively; DBD-Threader [29] and SPOT-Seq [46] that rely on libraries of structures of protein–DNA and protein–RNA complex and took excessive amount of time to run; and several methods that do not offer a web server including the predictor by Ho et al. [21], by Jeong et al. [34, 35], RNAProB [39], ProteRNA [42], RBRpred [43], and method by Wang et al [44]. We also could not consider DISIS [24], DNABindR [25, 26], BindN-RF [27], DBindR [28], NAPS [32], DNABR [33], PRINTR [36], RISP [37], PiRaNhA [40, 41] and PRBR [45] because their web servers were either no longer maintained or unavailable at the time of our experiment.

Evaluation measures and protocols

The considered predictors of DNA- and RNA-binding residues output either only the binary prediction (binding versus nonbinding) or binary prediction combined with a real-valued score that quantifies propensity for binding. We evaluate both types of outputs [57], and we exclude residues with missing atomic coordinates in the source structure files (i.e. disordered residues), as we could not compute their annotations of binding. The binary predictions are assesses based on

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}}$$

where TP is the number of true-positive results (correctly predicted binding residues), FN is the number of false-negative results (incorrectly predicted binding residues), FP is the number of false-positive results (incorrectly predicted nonbinding residues) and TN is the number of true-negative results (correctly predicted nonbinding residues).

The predicted propensities are evaluated using receiver operating curve (ROC), which is a plot of false-positive rate (FPR), which equals 1 − Specificity, against the true-positive rate (TPR), which is the same as Sensitivity. These two rates are computed by binarizing the propensities using thresholds. Similar to other studies we report the area under the ROC curve (AUC).

Moreover, we introduce a new measure, called ‘Ratio’, which quantifies the amount of cross-prediction between DNA- and RNA-binding residues. Ratio is defined as the fraction of native DNA-binding residues that are predicted as RNA-binding, and the fraction of native RNA-binding residues that are predicted as DNA-binding.

We also evaluate statistical significance of the differences in predictive quality between the best-performing method and each of the other considered methods. For a given data set, we randomly select 70% of proteins to calculate the corresponding Matthews correlation coefficient (MCC) and AUC values. This is repeated 10 times, and we compare the corresponding 10 paired results. Given that the measurements are normal, as tested using the Anderson–Darling test [58] with 0.05 significance, we use the paired *t*-test to investigate significance; otherwise we use the Wilcoxon rank sum test. The difference between a given pair of predictors is assumed statistically significant if *P*-value <0.05.

Prediction of DNA and RNA binding

DNA and RNA binding amino acids share similar biochemical properties, e.g. they are positively charged and have strong propensity to interact with the negatively charged phosphate backbone of DNA or RNA [22, 47]. However, the corresponding interactions carry out different cellular function and thus a given predictor should be able to separate DNA-binding residues from the RNA-binding residues. This is perhaps as crucial as the ability to separate RNA-/DNA-binding residues from the nonbinding residues. Here, we consider two types of predictions: with two (binary) and four outcomes. The former represents the prediction of a residue as binding versus nonbinding, for binding to either DNA or RNA. We are the first to consider the prediction with four outcomes: DNA&RNA-binding residue that binds to both DNA and RNA, DNA-binding residue (which does not bind to RNA), RNA-binding residue (which does not bind to DNA) and nonbinding residue. The results of the two-outcome-based predictions of the DNA binding and of the RNA binding can be combined to obtain the four outcomes as shown in Table 3.

Table 3. The conversion of the prediction of DNA-binding residues and the prediction of RNA-binding residues into the combined prediction of the DNA- and RNA-binding residues

Outcome		Two outcome predictions of RNA binding	
		RNA-binding	Nonbinding
Two outcome predictions of DNA binding	DNAbinding	DNA&RNA-binding	DNA-binding only
	Nonbinding	RNA-binding only	Nonbinding

Design and assessment of consensus predictors

A consensus method combines the predictions from several individual predictors of the DNA-binding residues or the RNA-binding residues. Prior works demonstrate that use of a consensus could provide improved predictive performance. Si *et al.* [13] developed a consensus MetaDBSite that integrates six predictors of DNA-binding residues. They have shown that MetaDBSite outperforms each of the six individual methods. Similarly, for the prediction of RNA-binding residues, Puton *et al.* [15] developed consensus of three predictors that provides improved predictive quality when compared with these methods.

We consider a comprehensive range of designs of consensuses and empirically assess their predictive performance. We are the first to investigate logic-based consensuses, which are selected as the best-performing (according to the MCC score on the training data set) combination of k methods, $k = 1, 2, \ldots, N$ where N is the total number of predictors of RNA- or DNA-binding residues that we consider in our empirical assessment. The predictions of the k methods are combined using two approaches, based on logical OR and logical AND operators. Specifically, the AND-based consensus assumes that a given residue is predicted as binding only if all k methods predict it as binding; otherwise this residue is predicted as nonbinding. The OR-based approach predicts a given residues as binding if any of the k methods predict it as binding. We also considered a majority-vote-based consensus predictor. This consensus predicts a residue as binding only if over half of the input methods predict so. This design generates the number of predicted binding residues that is lower than a consensus that uses only the logical OR and higher than if only the logical AND is used given that the same input predictors are used. The above two types of consensuses are simple to implement by an end user and do not involve any parameterization, which reduces risk of over fitting into a given benchmark data set.

We also extend these relatively simple consensuses to a more sophisticated ML consensus using linear logistic regression model. This model implements weighted average of the input predictions and uses both the binary predictions and the propensity scores generated by the individual DNA-binding or RNA-binding predictors. We generate the regression model on the training data set and assess its predictions on a given test data set. As the number of the nonbinding residues is substantially larger that the number of the binding residues in our training set, we under-sampled the nonbinding residues. For each training chain, we randomly sampled without replacement 25% of the nonbinding residues, and as a result, their number is about twice larger than the number of binding residues. The propensity scores generated by the regression model are binarized using the cutoff that corresponds to the maximal values of MCC on the training data set.

We also implement and empirically test first-of-its-kind method for the predictions of DNA- and RNA-binding residues based on the four outcomes. We considered three different approaches: 'single consensus', 'multiple consensus' and 'machine learning consensus'. The 'single consensus' combines outputs generated by a single DNA-binding predictor and a single RNA-binding predictor. We use the best-performing, according to the MCC score on the training data set, predictors and apply the rules summarized in Table 3 to merge their predictions. As consensuses of RNA-binding (DNA-binding) predictors outperform individual predictors on the training data set, the 'multiple consensus' approach extends the single consensus by integrating results of multiple predictors of RNA-binding residues or multiple predictors of DNA-binding residues. In other words, this approach combines outputs generated by a consensus of DNA-binding predictors and a consensus of RNA-binding predictors. We examine the combination of the two logic-based consensuses (multiple consensus logic) and two majority-vote-based consensuses (multiple consensus majority vote). Also, we combine the DNA-binding residue predictions with the RNA-binding residue predictions predicted by the corresponding two ML consensus predictors (multiple consensus ML). We also design and test a novel consensus that combines predictions generated by all considered predictors of DNA-binding and RNA-binding residues using the logistic regression model (DNA and RNA ML consensus). This is a single regression model rather than a combination of two regression models that is implemented in the multiple consensus ML. All these consensuses were build using only the training data set, i.e. the specific combinations of methods used in the multiple consensuses were selected based on maximizing the MCC value on the training data set, and the regression model for the DNA and RNA ML consensus was also generated on the training data set.

Results

Sequence-based predictors of RNA- and DNA-binding residues

Table 2 summarizes 14 sequence-based methods for the prediction of DNA-binding residues and 16 for the RNA-binding residues. Perhaps their most striking characteristic is that these predictors define binding residues in different ways. Virtually all predictors, except for DNABindR [25, 28] and PRINTR [39], define a given residue as binding if at least one of its atoms is closer than a cutoff distance from an atom of the RNA/DNA molecule. However, the cutoff values vary widely between 3.5 Å and 6 Å. The most commonly used value is 3.5 Å. Similarly, prior comparative reviews [7, 13–16] also often consider value of 3.5 Å and 5.0 Å, which is why we apply these two cutoffs.

The predictive models can be divided into two types: 'sequence-only' models that perform predictions using solely the sequence and sequence-derived one-dimensional descriptors [59], such as secondary structure and solvent accessibility; and 'template-based' models that rely on a library of structural templates. The latter group of methods uses the input sequence to find a structure in complex with DNA or RNA that has similar sequence, and they use this structure to perform predictions. The two 'template-based' approaches, DBD-Threader [50] for the prediction of DNA-binding residues and SPOT-Seq [46] for the RNA-binding residues provide accurate predictions but they also require relatively long runtime; our tests using the web servers show runtime values up to several hours per protein for DBD-Threader and 20 min to a few hours for SPOT-Seq. Interestingly, SPOT-Seq was shown to discriminate between RNA- and DNA-binding proteins [46], while we investigate whether this could be also accomplished with the sequence-only models.

The predictive strategy used by the 'sequence-only' methods consists of two steps. First, each residue in the input protein sequence is encoded into a vector of numerical features. Next, these features are used as inputs to a predictive model that outputs a binary value (binding versus nonbinding) and, for some methods, also a numeric score that quantifies propensity for the binding (Table 2). The information used to compute features for a given residue is collected from a window of residues that are adjacent to this residue in the sequence. The sizes of this window vary widely between methods, ranging from 3 (one residue on each side of the predicted residue) to 41; the most frequently used value is 11 (Table 2). The sequence-only predictors use a variety of designs that vary both on the information that is used to generate the features and the predictive models used. The input features include information derived directly from the protein sequence including amino acid composition (identity), and physiochemical properties of the input amino acids, such as pKa value of side chains, hydrophobicity, molecular mass and charge. Some features are also computed from one-dimensional structural characteristics that are predicted from the sequence, such as secondary structure and solvent accessibility. The most common input is based on the results of multiple sequence alignment of the input chain into a large sets of protein sequences (such as the *nr* database), primarily in the form of the evolutionary profile quantified with the position-specific scoring matrix (PSSM). This is related to the fact that PSSM can be used to quantify conservation of residues and the binding residues were shown to be conserved in the sequence [49, 60, 61]. Two predictors substitute PSSM with another way to find conserved residues. ProteRNA method [42] uses the WildSpan algorithm [62] while DISIS [24] uses MaxHom [63] algorithm. The predictive models are exclusively implemented based on a variety of ML algorithms including neural networks, SVMs, Naïve Bayes and decision trees. The SVM is used most often, which is motivated by empirical results that demonstrate that this type of model usually provides strong predictive performance [15, 20]. However, we note that different methods were trained and tested on different data sets, which vary in terms of their release date, size, resolution of structures used to generate annotation of binding, sequence similarity within the data set and definition of binding annotation. Moreover, they were evaluated using different protocols (e.g. using test sets and a variety of cross-validation types) and the predictive performance was assessed using different measures. Therefore, we could not use the results reported in the original articles to directly compare predictive quality of these methods. Our tests of methods that offer web server indicate that DBS-pred [18] and BindN [20] are among the fastest methods that complete the prediction of DNA-, RNA-binding residues for an average-sized protein with about 200 amino acids in <1 s.

Recent studies also investigated development of consensus approaches. Si *et al.* [13] have implemented a consensus method MetaDBSite that integrates predictions from six DNA-binding predictors: DBS-pred [18], BindN [20], DP-Bind [23], DISIS [24], DNABindR [25] and BindN-RF [27]. The results of these predictors are combined using the SVM model, and the resulting consensus was shown to outperform each individual predictor. Similarly, Puton *et al.* [15] assessed predictive quality of seven sequence-based methods for prediction of RNA-binding residues and developed a consensus that combines predictions from the top three predictors: PiRaNhA [41], Pprint [38] and BindN+ [31]. The outputs of these methods were merged together using weighted average where the weights are the AUC values on their benchmark data set. Again, empirical results have shown that their consensus outperforms the results generated by each of the three single predictors. These results motivated us to further investigate feasibility of building accurate consensus-based approaches.

Assessment of the predictive performance of the sequence-based prediction of DNA-binding and RNA-binding residues

We perform empirical assessment of the 10 selected computationally efficient sequence-only predictors that are available as web servers on the test data sets. The evaluation uses DNA_T (with DNA-binding proteins only), RNA_T (with RNA-binding proteins only) and COMB_T (with DNA- and RNA-binding proteins) benchmark test data sets where annotations were transferred between similar proteins, which results in a more complete set of annotations. We also compare these results with the results based on the original DNA_NT, RNA_NT and COMB_NT test data sets without the transfer. We consider two definitions of binding residues, using the cutoffs at 3.5 Å and 5 Å.

Predictive performance on the data sets with DNA-binding or RNA-binding proteins

Table 4 reveals that predictive performance of the individual predictors of DNA-binding residues [DBS-PSSM, DP-Bind(maj), DP-Bind(klr) and BindN+] on the DNA_T data set is relatively similar, with MCC values ranging between 0.293 and 0.307 (0.304 and 0.343), and AUC ranging between 0.795 and 0.797 (0.769 and 0.778) for the cutoff at 3.5 Å (5 Å). The only exception is the ProteDNA method that is characterized by lower predictive quality on this test data set. A likely explanation is the fact that this method was designed to find binding residues specifically in the transcription factors, which are a subset of our data set of DNA-binding proteins. This is corroborated by the relatively low value of sensitivity that was obtained by this predictor. Interestingly, the MetaDBSite consensus is also underperforming when compared with the results reported by the authors [13]. The reason is that four methods that this consensus was originally designed to use are no longer maintained. Consequently, instead of combining results of six predictors the current version of MetaDBSite is a simple ensemble of BindN and DP-Bind based on the logical AND, i.e. a given residue is predicted as DNA-binding if both methods predict it as such. Prediction qualities on the data set where binding residues are defined based on the 5 Å threshold are characterized by a decrease in sensitivity compared with the threshold of 3.5 Å. This means that the considered predictors do not predict the residues located between 3.5 and 5 Å as well as those that are closer than 3.5 Å. This could be because most of these methods (BindN+, DBS-PSSM and MetaDBSite) were trained using the 3.5 Å cutoff, while the remaining two methods (DP-Bind and ProteDNA) also use a lower cutoff at 4.5 Å.

Table 4. Results of empirical assessment of predictors of the DNA- or RNA-binding residues on the DNA_T or RNA_T data sets, respectively

Method	Binding residues defined based on 3.5 Å threshold						Binding residues defined based on 5 Å threshold					
	Sensitivity	Specificity	MCC	Significance	AUC	Significance	Sensitivity	Specificity	MCC	Significance	AUC	Significance
DNA binding on DNA_T data set												
Machine learning consensus	0.478	0.916	0.354		0.831		0.521	0.877	0.374		0.807	
Majority vote (BindN+(DNA), DBS_PSSM, ProteDNA(B)	0.447	0.907	0.314	+			0.366	0.923	0.322	+		
DBS-PSSM	0.721	0.753	0.307	+	0.796	+	0.656	0.773	0.343	+	0.772	+
Logic consensus (BindN+ AND DBS-PSSM)	0.424	0.912	0.305	+			0.360	0.924	0.317	+		
DP-Bind(maj)	0.598	0.823	0.301	+			0.539	0.844	0.338	+		
DP-Bind(klr)	0.590	0.824	0.297	+	0.795	+	0.525	0.844	0.328	+	0.778	+
BindN+	0.482	0.879	0.293	+	0.797	+	0.410	0.892	0.304	+	0.769	+
MetaDBSite consensus	0.325	0.935	0.267	+			0.263	0.942	0.260	+		
ProteDNA(B)	0.093	0.982	0.142	+			0.038	0.998	0.152	+		
RNA binding on RNA_T data set												
Machine learning consensus	0.242	0.962	0.234		0.755		0.273	0.934	0.229		0.726	
BindN+	0.399	0.891	0.219	=	0.738	+	0.341	0.896	0.222	=	0.704	+
Majority vote (BindN+(RNA), RNAbindR, Pprint)	0.457	0.854	0.212	+			0.407	0.862	0.226	=		
Meta2 consensus	0.526	0.812	0.211	+			0.463	0.816	0.214	+		
Logic consensus (BindN+ AND RNABindR AND Pprint)	0.244	0.950	0.203	+			0.190	0.951	0.181	+		
RNABindR	0.575	0.739	0.178	+	0.724	+	0.546	0.751	0.207	+	0.711	+
Pprint	0.433	0.796	0.141	+	0.681	+	0.381	0.799	0.136	+	0.648	+

Significance of the difference in MCC and AUC values between the best-performing method and other methods on a given data set was assessed based on 10 tests that use 70% of randomly chosen proteins; + (=) in the Sig column denotes that the difference was (was not) significant at *P*-value <0.05. AUC values could not be computed for DP-Bind(maj), MetaDBSite, ProteDNA(B), Meta2 and the four new consensuses, as these methods provide only the binary predictions. Methods are sorted by the MCC value.

We designed two types of consensuses: logic-based consensus, which combines individual predictors of DNA-binding residues (BindN+, DBS-PSSM, DP-Bind and ProteDNA) using all permutations of logical OR and logical AND operators; and majority vote consensus, which combines them using a majority voting rule (see Materials and Methods section for details). The top three logic-based and majority vote consensuses that obtain highest MCCs on the training data set are compared with the best-performing individual predictor, BindN+, in Supplementary Figures S1A and S1B, respectively. The best logic-based consensus combines BindN+ and DBS-PSSM using logical AND, which means that a given residue is predicted as binding only if both methods predict it as binding. The best majority vote consensus combines BindN+, DBS-PSSM and ProteDNA, which means that a given residue is predicted as binding only if at least two of these methods predict it as binding. The results show that although the logic-based and ma jority vote consensuses improve the prediction performance on the training data set, they do not deliver these improvements on the test data set. The logic-based approach provides similar (2.6% worse) MCC when compared with the best-performing individual predictor, DBS-PSSM, for the binding threshold equal 3.5Å (5Å) on the test data set. Although majority vote consensus slightly im proves MCC by 0.7% for the binding threshold at 3.5 Å, its MCC drops by 2.1% for the threshold at 5 Å. The reason for the lack of im provement is that the test data set is dissimilar to the training data set (<30% sequence similarity) and such simple combinations of in dividual predictors did not translate well between these two data sets. Motivated by this, we extended these designs of the consen suses into a more advanced ML consensus that applies linear logis tic regression (see Materials and Methods for details). The ML con sensus outperforms all single predictors by at least 4.7% (3.1%) in MCC, 3.4% (3.0%) in AUC for the cutoff at 3.5 Å (5 Å), and these differ ences are statistically significant (Table 4).

Analysis of the results concerning prediction of the RNA-binding residues leads to similar observations (Table 4). Predictive performance of the three considered predictors (BindN+, RNABindR and Pprint) vary between 0.141 and 0.219 (0.136 and 0.222) in MCC, and between 0.681 and 0.738 (0.648 and 0.711) in AUC on the RNA_T test data set based on the cutoff at 3.5 Å (5 Å). The Meta2 consensus is not performing as well as previously reported [15]. This is because some of the methods Meta2 was originally designed to combine are no longer available. The logic-based consensus, which outperforms other considered consensuses on the training data set (Supplementary Figure S1A), integrates predictions from BindN+, RNABindR and Pprint using logical AND. The majority vote consensus also com bines these three individual predictors (Supplementary Figure S1B). Similar to the results for the DNA-binding, these two types of simple consensuses do not perform well on the test data set. They only achieve equivalent or slightly worse MCC compared with the best-performing predictor, BindN+, on this data set. However, the ML-based consensus outperforms all the individ ual predictors. More specifically, its MCC is higher by at least 1.5% (0.7%), AUC by at least 1.8% (1.5%) and specificity by at least 7.2% (3.8%) for the threshold of 3.5 Å (5 Å).

Overall, we conclude that methods for the prediction of DNA-binding (RNA-binding) residues are characterized by relatively good predictive performance measured by their values of MCC and AUC when tested on the dissimilar (in the sequence) proteins that bind DNA (RNA). Their AUC is at about 0.8 (0.7), and their predictions have modest correlation with the native annotations at about 0.3 (0.2). They generally have relatively high specificity coupled with modest sensitivity, which means that they predict a subset of native binding residues with high predictive quality while missing the remaining binding residues. Our analysis reveals that a simple consensus based on majority vote or logic does not improve the predictive performance when applied to predict proteins that are dissimilar to the proteins that were used to develop this consensus. At the same time, a more sophisticated logistic regression-based consensus outperforms all individual methods in the prediction of DNA-binding and RNA-binding residues, even for the dissimilar chains. We also note that predictive quality (see AUC and sensitivity values in Table 4) on the data set where binding residues are annotated based on the larger cutoff at 5 Å are consistently slightly worse than for the cutoff at 3.5 Å. This suggests that residues that are closer to RNA or DNA are easier to discriminate from the nonbinding residues.

Predictive performance on the data set with DNA- and RNA-binding proteins

We are the first to comprehensively assess predictive performance of the considered predictors on the COMB_T data set that combines DNA- and RNA-binding proteins, see Table 5. We observe a drop in MCC when compared with the results in Table 4. This is a universal pattern, irrespective of whether we assess predictors of DNA- or RNA-binding residues, and it reveals that these methods confuse the two types of binding residues. Sensitivity stays the same, as the annotation of the binding residues does not change compared with when we consider prediction of DNA- or RNA-binding residues; we just introduce additional nonbinding residues.

Considering individual predictors of DNA-binding residues, the MCC on the COMB_T data set (Table 5) is lower by 3.8–5.5% (3.8-6.4%) when compared with the results on the DNA_T data set (Table 4) for the cutoff at 3.5 Å (5 Å). The only exception is ProteDNA, which has low sensitivity and MCC and which predicts a relatively small number of residues that selectively bind transcription factors. Ratio, which quantifies fraction of RNA-binding residues that are predicted to be DNA-binding, reveals that at least 28.9% (25.0%) and as many as 48.7% (42.9%) of the RNA-binding residues are mispredicted at the cut-off of 3.5 Å (5 Å). Although the majority vote and logic consensuses do not offer improved MCC when compared with their input methods on this test data set, their Ratio is reduced to 23.2%. This means that the individual predictors do not agree on the misprediction of the RNA-binding residues as DNA-binding for a substantial number of cases, i.e. they mispredict different residues. The ML-based consensus that we designed again outperforms all other predictors on this data set. It secures the highest MCC equal 0.311 (0.326) and also the highest AUC of 0.841 (0.81), and these improvements are statistically significant (Table 5). Figure 1A (Figure 1B) shows the ROCs for the ML consensus and all the individual predictors that generate the propensity scores on the COMB_T data sets at 3.5 Å (5 Å) binding cutoff. Notably, the TPR of our ML consensus is higher than the TPR of any individual predictors for almost the entire range of FPR values. However, this consensus still has a problem of substantial levels of mispredictions between DNA and RNA binding residues, which is demonstrated by the moderate values of Ratio (Table 5). We solve this problem by proposing a new design of the ML consensus that combines prediction of both DNA- and RNA-binding residues.

Similarly, assessment of the predictors of the RNA-binding residues on the COMB_T data set demonstrates that the results are worse when compared with the results on the RNA_T data set. Specifically, MCC is lower by 5.7–10.5% (5.7–11%); AUC by 1.2–3.1% (1.7–3.6%); and specificity by 1.4–3.7% (1.7–4%) based on

Table 5. Results of empirical assessment of predictors of the DNA- or RNA-binding residues on the COMB_T data set

Type of binding	Method	Binding residues defined based on 3.5 Å threshold							Binding residues defined based on 5 Å threshold						
		Sensitivity	Specificity	Ratio	MCC	Significance	AUC	Significance	Sensitivity	Specificity	Ratio	MCC	Significance	AUC	Significance
DNA binding	Machine learning consensus	0.478	0.922	0.267	0.311		0.841		0.521	0.881	0.310	0.326		0.810	
	Majority vote (BindN+(DNA), DBS_PSSM, ProteDNA(B)	0.447	0.916	0.232	0.277	+			0.366	0.925	0.196	0.282	+		
	Logic consensus (BindN+ AND DBS-PSSM)	0.424	0.919	0.232	0.267	+			0.360	0.925	0.196	0.278	+		
	DBS-PSSM	0.721	0.774	0.487	0.266	+	0.810	+	0.656	0.785	0.429	0.298	+	0.784	+
	BindN+	0.482	0.888	0.289	0.256	+	0.806	+	0.410	0.896	0.250	0.266	+	0.773	+
	DP-Bind(maj)	0.598	0.823	0.467	0.247	+			0.539	0.834	0.421	0.275	+		
	DP-Bind(klr)	0.590	0.828	0.445	0.246	+	0.794	+	0.525	0.839	0.404	0.270	+	0.770	+
	MetaDBSite consensus	0.325	0.933	0.230	0.221	+			0.263	0.938	0.185	0.216	+		
	ProteDNA(B)	0.093	0.990	0.000	0.158	+			0.038	0.999	0.000	0.160	+		
RNA binding	Machine learning consensus	0.242	0.945	0.240	0.128		0.730		0.273	0.905	0.330	0.120		0.699	
	Majority vote (BindN+(RNA), RNAbindR, Pprint)	0.457	0.821	0.551	0.116	+			0.407	0.823	0.499	0.120	=		
	Meta2 consensus	0.526	0.774	0.616	0.116	+			0.463	0.778	0.550	0.116	=		
	BindN+	0.399	0.854	0.498	0.114	+	0.706	+	0.341	0.856	0.427	0.111	+	0.668	+
	Logic consensus (BindN+ AND RNABindR AND Pprint)	0.244	0.933	0.279	0.113	+			0.190	0.933	0.234	0.097	+		
	RNABindR	0.575	0.718	0.643	0.105	+	0.712	+	0.546	0.722	0.601	0.120	=	0.694	=
	Pprint	0.433	0.782	0.478	0.084	+	0.667	+	0.381	0.782	0.454	0.079	+	0.629	+

Significance of the difference in MCC and AUC values between the best-performing method and other methods on a given data set was assessed based on 10 tests that use 70% of randomly chosen proteins; + (=) in the Sig column denotes that the difference was (was not) significant at *P*-value <0.05. AUC values could not be computed for DP-Bind(maj), MetaDBSite, ProteDNA(B), Meta2 and the four new consensuses, as these methods provide only the binary predictions. Methods are sorted by the MCC value.

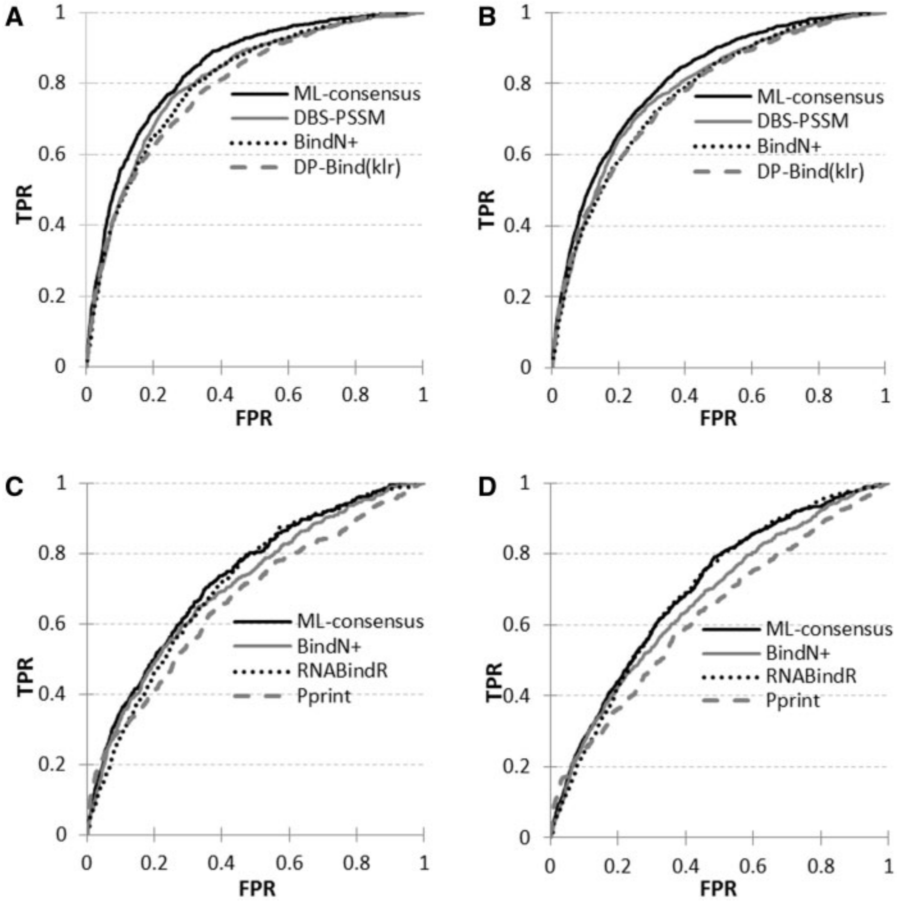

Figure 1. The ROCs for the ML consensuses and the individual predictors of DNA- and RNA-binding residues. Panels **A** and **B** compare the DNA-binding predictors on the COMB_T data set with the binding thresholds of 3.5A and 5A, respectively; Panels **C** and **D** compare the RNA-binding predictors on the COMB_T data set with the binding threshold of 3.5 Å and 5 Å, respectively.

the cutoff at 3.5 Å (5 Å). Most importantly, the identical sensitivity coupled with the lower specificity indicates that predictors of RNA-binding residues mispredict the DNA-binding residues as RNA-binding, which is further confirmed by the values of Ratio. Ratio tells that these methods mispredict between 47.8 and 64.3% (42.7 and 60.1%) DNA-binding residues as RNA-binding depending on the cutoff value. Although the logic-based and majority vote consensuses do not improve predictive performance when compared with their input predictors on this data set, the former consensus provides relatively low values of Ratio. However, the ML-based consensus outperforms all the individual predictors. It secures the highest MCC, AUC and specificity and the lowest (best) Ratio. The ROCs of this consensus and all the individual predictors are shown in Figure 1C and D. Overall, the ML consensus achieves the best performance when considering the entire range of FPR values. Pprint performs well at low FPR (<0.05) value, while its TPR drops substantially for higher values of FPR. RNABindR curve overlaps with our ML consensus curve at larger values of FPR, but this method has lower PRF when FPR <0.45 and <0.25 for the binding cutoffs at 3.5 Å and 5 Å, respectively. The only weakness of the ML consensus is still relatively high values of Ratio, in spite of the fact that they are lower than for the other methods.

We further investigate the mentioned above misprediction by assessing methods that target RNA-binding (DNA-binding) on the data sets with the DNA-binding proteins (RNA-binding proteins); see Supplementary Table S1. The considered predictors obtain relatively low specificity between about 0.7 and 0.9 (ideally specificity should be 1) and high Ratio between about 0.25 and 0.7 (ideally Ratio should be 0), except for ProteDNA that predicts only a small subset of DNA-binding residues. We note that the logic and ML consensuses overall secure lower val ues of Ratio when compared with the individual predictors. These results confirm our observations based on the results in Table 5.

The results on the COMB_T, DNA_T and RNA_T data sets (Tables 4 and 5) indicate that current methods that predict DNA-binding or RNA-binding residues are characterized by good predictive performance. However, although these predictors perform well on their own type of binding, they also overpredict the other type of binding residues, i.e. predictors of RNA-binding (DNA-binding) residues also predict a large number of DNA-binding (RNA-binding) residues as RNA-binding (DNA-binding). This means that they tend to predict nucleic acids-binding residues rather than more specific DNA- or RNA-binding residues.

Predictive performance on the data sets with and without transfer of annotations

We assess the predictors of DNA-binding or RNA-binding residues on the data sets where we did not transfer the binding residues between the similar proteins: DNA_NT, RNA_NT and COMB_NT. In these data sets, we used binding residues

Table 6. Results of empirical assessment of consensus-based methods on the COMB_T data set when considering prediction of combined DNA- and RNA-binding residues and individual prediction of DNA- or RNA-binding residues

Method	Prediction of DNA and RNA binding				Prediction of DNA or RNA binding	
	DNA&RNA	DNA	RNA	non-DNA & non-RNA	DNA versus non-DNA	RNA versus non-RNA
Sensitivity						
Single consensus	N/A	0.101	0.164	0.839	0.482	0.399
Multiple consensus logic	N/A	0.207	0.103	0.899	0.424	0.244
Multiple consensus majority vote	N/A	0.085	0.259	0.821	0.447	0.457
Multiple consensus machine learning	N/A	0.261	0.078	0.914	0.478	0.242
RNA and DNA machine learning consensus	N/A	0.392	0.125	0.929	0.392	0.125
Specificity						
Single consensus	0.908	0.962	0.942	0.552	0.888	0.854
Multiple consensus logic	0.957	0.951	0.974	0.438	0.919	0.933
Multiple consensus majority vote	0.922	0.976	0.895	0.590	0.916	0.821
Multiple consensus machine learning	0.955	0.956	0.986	0.451	0.922	0.945
RNA and DNA machine learning consensus	1.000	0.941	0.981	0.409	0.941	0.981
MCC						
Single consensus	N/A	0.074	0.072	0.277	0.256	0.114
Multiple consensus logic	N/A	0.159	0.076	0.281	0.267	0.113
Multiple consensus majority vote	N/A	0.086	0.081	0.280	0.277	0.116
Multiple consensus machine learning	N/A	0.220	0.084	0.318	0.311	0.128
RNA and DNA machine learning consensus	N/A	0.290	0.118	0.315	0.290	0.118
Ratio						
Single consensus	N/A	N/A	N/A	N/A	0.289	0.498
Multiple consensus logic	N/A	N/A	N/A	N/A	0.232	0.279
Multiple consensus majority vote	N/A	N/A	N/A	N/A	0.232	0.551
Multiple consensus machine learning	N/A	N/A	N/A	N/A	0.267	0.240
RNA and DNA machine learning consensus	N/A	N/A	N/A	N/A	0.183	0.064

There are no DNA&RNA binding residues in this data set and thus we cannot compute sensitivity and MCC for this outcome. Values of Ratio cannot be computed for the combined prediction of RNA and DNA binding. The 'multiple consensus logic' uses the two best-performing logic-based consensuses that we built for the prediction of DNA-binding residues and RNA-binding residues, respectively; 'multiple consensus majority vote' combines the two best-performing majority-vote-based consensuses for the prediction of DNA- and RNA-binding residues, respectively; 'multiple consensus machine learning' is the combination of the two machine learning consensus for the prediction of DNA- and RNA-binding residues, respectively; and 'DNA and RNA machine learning consensus' combines predictions generated by all considered predictors of DNA-binding and RNA-binding residues using logistic regression model.

annotated based on a single complex, which was also practiced in the prior studies. The results that are summarized in the Supplementary Tables S2 and S3 are analogous to the results when using the more complete set of annotations that include transfer from the similar proteins. That is, relative (with each other) predictive quality of individual methods is similar and their overall predictive performance is good. Similarly, the logic and majority vote consensuses do not provide improved predictive performance while the ML consensus improves over the individual predictors on each of the three data sets, i.e. this consensus has higher MCC and AUC values. However, when directly comparing Table 4 with the corresponding Supplementary Table S2 or Table 5 with the corresponding Supplementary Table S3, we observe that use of the transferred (more complete) annotations results in a decrease in AUC and sensitivity. Specifically, for the predictors of DNA-binding residues, the decrease ranges between 0.4% (0.5%) and 1.3% (1.1%) in AUC, and between 0.6% (0.2%) and 2.9% (1.6%) in sensitivity for the threshold used to define binding residues at 3.5 Å (5 Å). Similarly, for the methods that predict RNA-binding residues, the AUC is lower by up to 0.8% (0.8%) and sensitivity by up to 0.5% (0.6%). These results suggest that the binding residues that were transferred from similar proteins are more difficult to predict for the current methods when compared with the remaining binding residues. This could be explained by the fact that these methods were designed (trained) on data set without the transfer of annotations. Such data sets would include false negatives, which are binding residues that were annotated as nonbinding since their annotations were not transferred from a similar protein. Apparently mislabeling these residues in the training data sets prevents these methods from correctly identifying them as binding on our test data sets. Consequently, we believe that a new generation of predictors should use training data sets with the transferred annotations.

Assessment of the predictive performance of the sequence-based combined prediction of DNA- and RNA-binding residues

The published predictors were designed specifically to target either protein–DNA or protein–RNA interactions. We empirically demonstrate that these methods cross-predict into the other nucleic acid type, i.e. methods that predict DNA-binding also mispredict RNA-binding residues and vice versa. One way to potentially alleviate this drawback is to redefine these two prediction tasks as a single prediction with four outcomes: DNA&RNA-binding, DNA-binding, RNA-binding and nonbinding residue. We are the first to design such predictors and comprehensively assess their predictive performance. Our design integrates multiple predictors of DNA- and RNA-binding residues based on three types of consensuses: single consensus, multiple

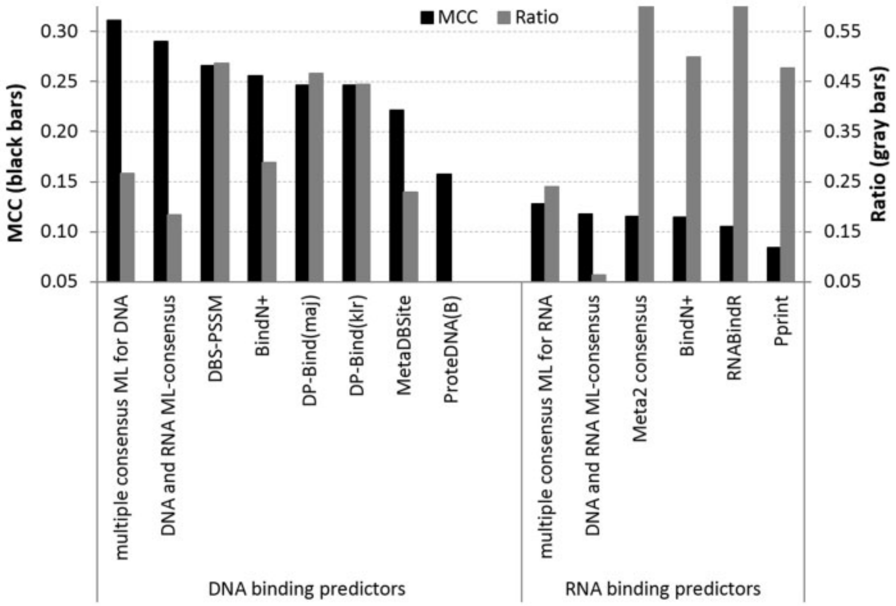

Figure 2. Comparison between the DNA and RNA ML consensus that targets combined prediction of DNA- and RNA-binding residues and the considered predictors of DNA- or RNA-binding residues. The predictors of DNA- or RNA-binding residues include the two ML-based DNA- or RNA- binding consensuses. The evaluation considers prediction of DNA-binding residues (left side of the figure) and prediction of RNA-binding residues (right side of the figure) on the COMB_T test data set with the binding residues annotated using cut-off at 3.5 Å.

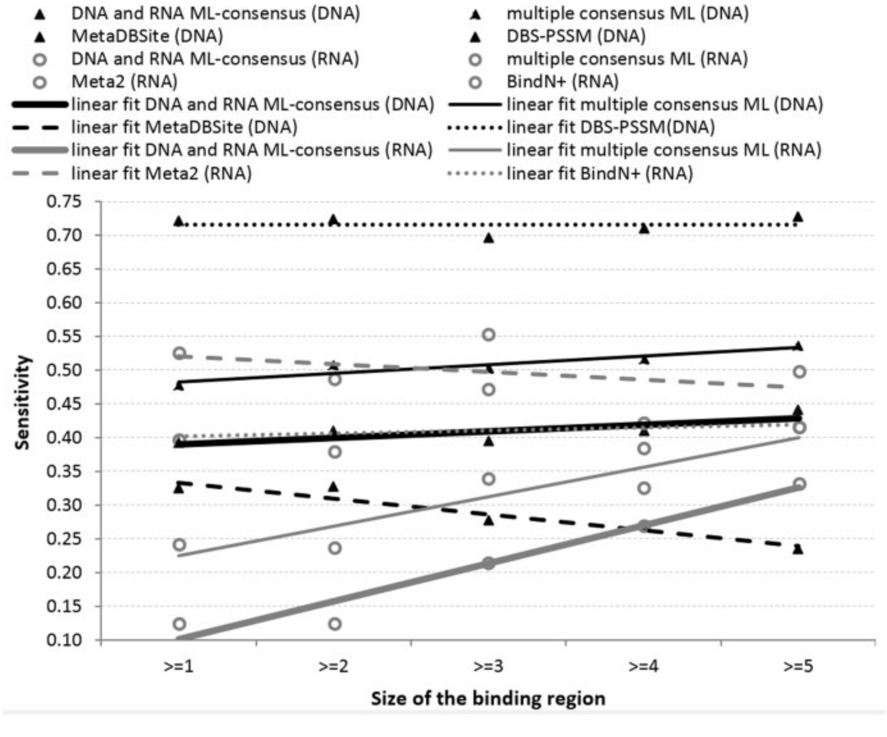

Figure 3. Sensitivity in the function of the minimal size of binding regions for the ML-based consensus predictors, the best individual predictors and the existing consensus predictor for the prediction of DNA- or RNA-binding residues. We consider only the binding residues that are in regions of the size larger or equal to the value shown on the x-axis; the other binding residues were removed from the assessment. The linear fit into the sensitivity values is shown using lines. The black and gray lines are for the DNA-binding and RNA-binding predictors, respectively. The line type denotes a given predictor.

consensus and ML consensus. The single consensus combines the best-performing (i.e. providing the highest MCC) on the training data set predictor of DNA-binding residues, BindN+ (DNA version), with the best-performing predictor of RNA-binding residues, BindN+ (RNA version). The multiple consensus approach combines multiple predictors of DNA-binding residues and RNA-binding residues. We consider three designs of the multiple consensuses: multiple consensus logic, multiple

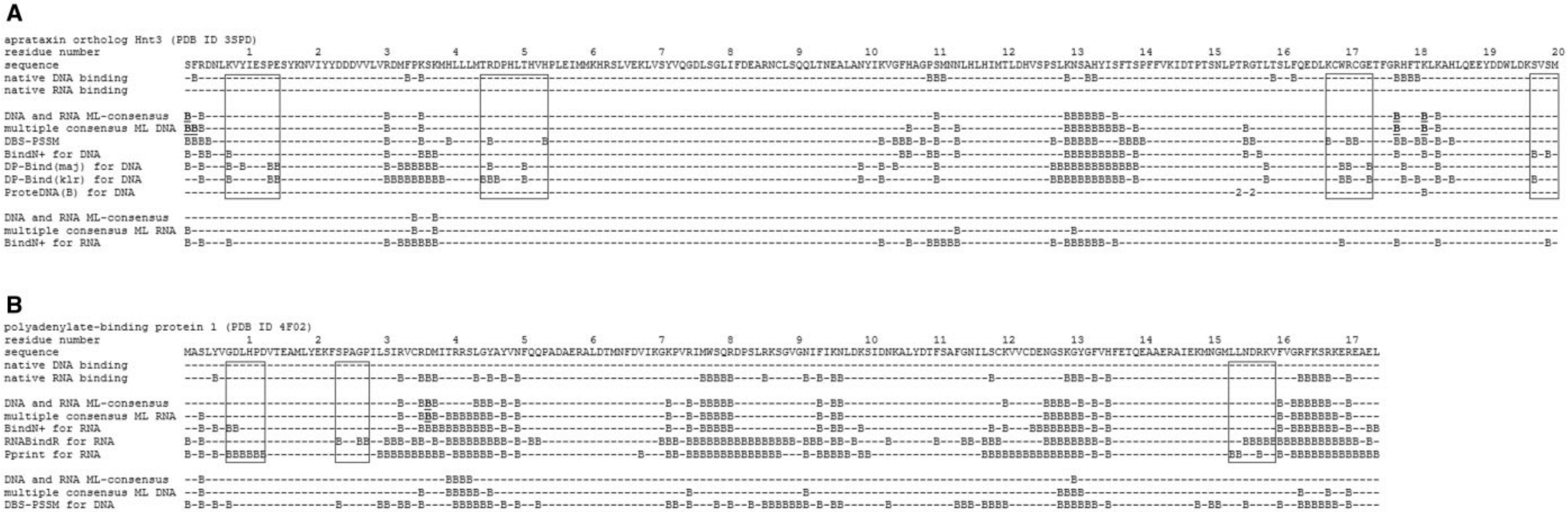
A
apraxatin ortholog Hnt3 (PDB ID 3SPD)
residue number
sequence
SFRDNLKVYIESPESYKNVIYYDDDVVLVRDMFPKSKMHLLLMTRDPHLTHVHPLEIMMKHRSLVEKLVSYVQGDLSGLIFDEARNCLSQQLTNEALANYIKVGFHAGPSMNNLHLHIMTLDHVSPSLKNSAHYISFTSPFFVKIDTPTSNLPTRGTLTSLFQEDLKCWRCGETFGRHFTKLKAHLQEEYDDWLDKSVSM
native DNA binding
native RNA binding
DNA and RNA ML-consensus
multiple consensus ML DNA
DBS-PSSM
BindN+ for DNA
DP-Bind(maj) for DNA
DP-Bind(klr) for DNA
ProteDNA(B) for DNA
DNA and RNA ML-consensus
multiple consensus ML RNA
BindN+ for RNA
B
polyadenylate-binding protein 1 (PDB ID 4F02)
residue number
sequence
MASLYVGDLHPDVTEAMLYEKFSPAGPILSIRVCRDMITRRSLGYAYVNFQQPADAERALDTMNFDVIKGKPVRIMWSQRDPSLRKSGVGNIFIKNLDKSIDNKALYDTFSAFGNILSCKVVCDENGSKGYGFVHFETQEAAERAIEKMNGMLLNDRKVFVGRFKSRKEREAEL
native DNA binding
native RNA binding
DNA and RNA ML-consensus
multiple consensus ML RNA
BindN+ for RNA
RNABindR for RNA
Pprint for RNA
DNA and RNA ML-consensus
multiple consensus ML DNA
DBS-PSSM for DNA

Figure 4. Two case studies that illustrate the working of the ML consensuses. Panel **A** concerns the DNA-binding apratаxin ortholog Hnt3 (PDB ID: 3SPD), and Panel **B** shows the RNA-binding polyadenylate-binding protein 1 (PDB ID: 4F02). 'B' denotes binding residues and '-' denotes the nonbinding residues. Boxes and bold denote results that are discussed in the text. Disordered regions in these two proteins (regions with no coordinates) are omitted. Panel **A** includes the following lines (from top to bottom): residue number, sequence, native annotation of DNA binding residues, native annotation of RNA-binding residues, blank line, predictions from the DNA and RNA ML model for the DNA binding residues, ML model for DNA-binding, DBS-PSSM, BindN+(DNA), DP-Bind(maj), DP-Bind(klr), ProteDNA(B), blank line, predictions from the DNA and RNA ML model for the RNA binding residues, ML model for the RNA-binding residues and the best individual RNA-binding residue predictor BindN+(RNA). Panel **B** includes the following lines (from top to bottom): residue number, sequence, native annotation of DNA binding residues, native annotation of RNA-binding residues, blank line, predictions from the DNA and RNA ML model for the RNA binding residues, ML model for the RNA-binding residues, BindN+, RNABindR, Pprint, blank line, predictions from the DNA and RNA ML model for the DNA binding residues, ML model for the DNA-binding residues and the best individual DNA-binding residue predictor DBS-PSSM.

consensus majority vote and multiple consensus ML. Moreover, the DNA and RNA ML consensus combines predictions generated by all considered predictors of DNA-binding and RNA-binding residues using the logistic regression model (see 'Design and assessment of consensus predictors' section for details). We assess these methods on the COMB_T test data set that includes annotations that were transferred from similar proteins and where binding residues were defined with the cutoff at 3.5 Å (Table 6). This data set shares low, <30%, sequence similarity with the training data set that was used to develop consensuses. There are no DNA&RNA-binding residues in this data set, so we cannot compute sensitivity and MCC for this outcome.

All multiple consensuses outperform the single consensus in MCC for the combined prediction of DNA and RNA binding (Table 6). Moreover, the single consensus substantially overpredicts the RNA&DNA outcome with the corresponding specificity at 0.908. The multiple consensuses reduce this overprediction obtaining specificities between 0.922 and 0.957. The result of this overprediction for both single and multiple consensuses is the relatively low sensitivity for the prediction of DNA binding and the prediction of RNA binding, i.e. many of the RNA or DNA binding residues are predicted to bind both RNA and DNA. However, the RNA and DNA ML consensus, which is inherently designed to predict the four outcomes, correctly does not predict the DNA&RNA binding residues (specificity = 1) and secures high values of specificity and MCC. Its MCC is higher by 7 and 3.4% for the prediction of DNA-binding residues and RNA-binding residues, respectively, when compared with the best multiple consensus. This result demonstrates that the RNA and DNA ML consensus provides improved predictive performance when compared with the other consensuses.

We applied the considered consensuses to predict DNA-binding residues and RNA-binding residues separately (the two right-most columns in Table 6). The predictions of the consensuses that consider four outcomes are converted into prediction of DNA-binding residues as follows: 'DNA&RNA-binding' and 'DNA-binding' are assigned as 'DNA-binding'; 'RNA-binding' and 'non-binding' are assigned as 'non-binding'. For the prediction of RNA-binding residues, the conversion assumes 'RNA-binding' for the 'DNA&RNA-binding' and 'RNA-binding' predictions, and 'non-binding' for the 'DNA-binding' and 'non-binding' predictions. Table 6 shows that the two ML consensuses outperform the other types of consensuses having higher values of MCC and specificity. The main observation is that the RNA and DNA ML consensus offers substantially reduced values of Ratio, at 0.183 and 0.064 for the DNA and for the RNA binding, respectively, compared with the second best Ratios of 0.232 and 0.240. This means that this novel type of consensus generates predictions with lower rate of mispredictions between DNA- and RNA-binding residues.

We compare results generated by the two ML consensuses for the prediction of DNA-binding residues with the considered predictors of DNA-binding, see Figure 2. The DNA and RNA ML consensus obtains MCC of 0.290, which is lower than MCC of 0.311 of the multiple consensus ML for the prediction of DNA-binding residues (black bars in Figure 2). However, the former consensus has by far the lowest values of Ratio at only 0.183 (gray bars in Figure 2), except for the ProteDNA that predicts a small subset of DNA-binding residues and has the lowest MCC. Similar conclusions are true when considering prediction of the RNA-binding residues (Figure 2). The DNA and RNA ML consensus secures MCC of 0.118, which is lower compared with the best MCC of 0.128 obtained by the multiple consensus ML. It also boasts the lowest value of Ratio at 0.064 compared with the second lowest value at 0.240. Most importantly, the novel DNA and RNA ML consensus improves over all individual predictors having higher MCC while providing much lower Ratio for prediction of the RNA and the DNA binding residues (Figure 2). These results suggest that the development of consensuses for the combined prediction of DNA- and RNA-binding residues could offer a viable solution to generate high-quality prediction of DNA- or RNA-binding residues where the cross-predictions are substantially reduced.

Assessment of the predictive performance on binding regions

We investigate the predictive quality of our consensus predictors and all considered individual predictors on the predictions of DNA- or RNA-binding regions, defined as a stretch of consecutive binding residues. Figure 3 analyzes relation between sensitivity (fraction of correct predictions among the native binding residues) and the minimal length of the binding regions. We observe an increase in the sensitivity with the length of the binding regions for the DNA and RNA ML consensus and the multiple consensus ML for the prediction of DNA binding residues and prediction of RNA binding residues (solid lines in Figure 3). On the other hand, the current consensuses, MetaDBSite and Meta2, are characterized by lower sensitivity for the longer binding regions (dashed lines in Figure 3). The best-performing individual predictors, DBS-PSSM for the DNA binding and BindN+ for the RNA binding (Table 5) offer the same levels of sensitivity irrespective of the length of the binding regions (dotted lines in Figure 3). Although DBS-PSSM has the highest sensitivity, this method also has lowest specificity and MCC that is lower than MCCs of the considered consensuses (Table 5), which means that it overpredicts binding residues. All together, we conclude that the ML consensuses work especially well for the longer binding regions.

Assessment of the predictive performance on proteins that do not interact with DNA and RNA

We test the considered predictors on 50 human proteins that do not interact with either DNA or RNA molecules to estimate their specificity, which ideally should equal 1. Except for ProteDNA, which only predicts a small subset of DNA-binding residues, the considered individual DNA-binding predictors have specificity between 0.78 and 0.87. The multiple consensus ML for the prediction of DNA-binding residues and the MetaDBSite consensus have higher specificities at 0.92 and 0.93, respectively. The highest specificity at 0.95 is achieved by the DNA and RNA ML consensus. Similar results are observed for the RNA binding predictors. The three individual RNA-binding predictors and the Meta2 predictor obtain specificity ranging between 0.78 and 0.86. The multiple consensus ML has specificity of 0.97 while the RNA- and DNA-ML consensus secures the highest specificity of 0.99. Overall, the ML consensuses, and in particular the novel design that combines prediction of RNA and DNA binding residues, offer reduced levels of FP predictions.

Case studies

We illustrate predictions of the most successful in our tests ML consensuses and all considered individual predictors of DNA- and RNA-binding residues on two proteins selected from the test data set. The overall predictive performance measured with MCC for the consensuses on these two proteins is similar to the value on the whole test data set. Figure 4A compares predictions for the DNA-binding aprataxin ortholog Hnt3 (PDB ID: 3SPD). We observe that virtually all binding regions (except for the residues near position 160) were captured by most predictors. Both ML consensuses for the prediction of DNA-binding residues filter FP predictions (nonbinding residues predicted as binding) at both termini (shown using boxes in Figure 4A). These boxed regions are relatively far away from the native binding regions. Moreover, they annotate a few binding residues that were predicted by a subset of individual predictors (shown in bold and underline in Figure 4A) which are either correct predictions or immediately adjacent to the native binding residues. The RNA and DNA ML consensus reduces some of the FP generated by the multiple consensus ML, particularly near position 135. The best performing in our tests individual method that predicts RNA-binding residues (last line in Figure 4A) generate FP that generally line up with the location of the DNA binding residues. However, the ML consensuses, in particular the novel DNA and RNA ML consensus, substantially reduces these mispredictions. Similar observations are true for the predictions for the RNA-binding polyadenylate-binding protein 1 (PDB ID: 4F02) shown in Figure 4B. The two ML consensuses filter out FP generated by the individual predictors of RNA binding residues in the boxed regions that are relatively far from the native binding regions. They also correctly locate binding residue at position 36 that was missed by one of the individual RNA-binding predictors. Moreover, the best performing in our tests predictor of the DNA binding residues incorrectly predicts relatively many DNA binding residues (last line in Figure 4B) which again align with the native RNA binding residues. The ML approaches for the prediction of DNA binding residues reduce the number of these mispredictions by a large factor.

Overall, the case studies demonstrate that the ML consensuses successfully reduce some of the FP generated by the individual predictors and correctly predict binding residues even if some of the individual predictors do not. The novel DNA and RNA ML consensus further reduces some of the FP generated by the multiple consensus ML.

Conclusions

High-throughput identification of nucleic acid-protein interactions is critical to improve our understanding of macromolecular functions and biophysical mechanisms of gene regulation. We performed a comprehensive review of 30 sequence-based predictors of DNA- or RNA-binding residues. Although these methods vary in their design, they commonly use evolutionary information and sliding windows to encode inputs and SVM as the predictive model. This suggests that the binding residues tend to appear in conserved sequence segments. The input features used to predict DNA-binding residues overlap with the inputs used by the predictors of RNA-binding residues, which is not surprising given the chemical similarity between DNA and RNA. Our empirical assessment of DNA-binding (RNA-binding) predictors on the DNA-binding (RNA-binding) proteins that have working web servers reveals that they are characterized by acceptable levels of predictive performance. They have AUCs at about 0.7–0.8 and MCCs between 0.1 and 0.3 when measured on a hard data set of proteins characterized by low sequence similarity to the proteins used to design these methods. However, when tested on the test data set that include both RNA- and DNA-binding proteins, we found that these predictors are guilty of substantial amounts of cross-prediction, i.e. they predict RNA-binding residues as DNA-binding and vice versa. In other words, they are unable to properly separate DNA from RNA binding residues. This is likely the results of use of similar input features and the fact that these methods were trained based on data sets that use either only DNA-binding or only RNA-binding proteins. A unique characteristic of our empirical assessment is the fact that we used multiple similar proteins to annotate binding residues, thus providing a more complete annotation. We found that these binding residues transferred from similar proteins, when compared with previous assessments that used only a single structure, are more challenging to predict by the current methods. We speculate that this is because these 'additional' binding residues are mislabeled in the training data set of these methods.

Motivated by the prior success in building consensus-based predictors [13, 15], we designed and empirically tested a simple logic-based consensuses based on combinations of logical OR and logical AND operators, a majority vote consensus, and a more sophisticated ML consensus. We show that the logic and majority-vote-based consensuses do not offer improvements when tested on the hard test data set. However, the ML consensuses provide improved predictive performance when compared with the individual methods for the prediction of DNA-binding residues and for the prediction of RNA-binding residues on the same hard test set. We also performed first-of-its-kind study concerning combined prediction of DNA- and RNA-binding residues. We designed three types of consensuses to address this prediction, including a ML-based approach. The ML consensus offers strong predictive performance in the combined prediction and, most importantly, also for the prediction of DNA-binding or RNA-binding residues individually. We empirically show that this consensus provides higher values of MCC compared with the best-performing individual predictors while it also substantially reduces the cross-prediction. We also assessed how the consensuses and individual predictors perform on longer binding regions and show that the ML consensuses perform better for longer binding regions. When tested on the nonbinding proteins, once again the ML consensuses secure the lowest levels of FP and the highest specificity. Finally, we illustrate these empirical results using two case studies. They demonstrate that the ML consensuses filter out false predictions of the binding residues generated by individual predictors that are located relatively far from the native binding residues.

Finally, our results prompt several recommendations. First, we found that many of the original web servers are either no longer maintained or only transiently online. Besides this being inconvenient for the end users, it also negatively affects consensuses that rely on the web server calculations. We recommend that new consensuses should be built using stand-alone, local implementation of the input predictors and that the standards in supporting web servers should be improved. Although our consensuses rely on the web servers, as some of the input predictors do not offer stand-alone versions, we use methods that were available over extended period of time; we performed predictions with these methods between December 2013 and early 2015. Second, new generations of DNA-binding (RNA-binding) specific predictors are needed. Such methods would not only separate the DNA-binding (RNA-binding) residues from the nonbinding residues but also from the RNA-binding (DNA-binding) residues. This could be accomplished via a number of avenues including (1) building training data sets that combine both RNA- and DNA-binding proteins; (2) design of new inputs that are predictive specifically for either DNA- or RNA-binding residues; (3) by considering building methods that address combined prediction of DNA- and RNA-binding residues; and (4) by combining the different types of inputs and features used by the current predictors of DNA-binding and RNA-binding residues (Table 2). The latter would lead to faster runtime compared with a consensus that combines outputs of these predictors. Third, we advocate that the currently predominant annotation of the binding residues should be improved by transfer from similar proteins, instead of using individual complexes. This would improve completeness of the annotations and would lead to the development of more accurate predictors.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Our detailed analysis of a comprehensive set of 30 sequence-based predictors of DNA- or RNA-binding residues covers their design, outputs and availability.
- Modern predictors that offer web servers are characterized by good overall predictive performance but they cannot discriminate between DNA- and RNA-binding residues.
- Consensus-based methods based on ML provide improved predictive performance when compared with individual prediction methods.
- Annotation of DNA- or RNA-binding residues should combine information from the corresponding complexes that involve the same or similar proteins.
- New prediction methods that better discriminate between DNA- and RNA-binding residues should be built by using training data sets that combine both RNA- and DNA-binding proteins, designing new inputs that specifically target either DNA- or RNA-binding residues, and by pursuing combined prediction of DNA- and RNA-binding residues.

Funding

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada [Discovery grant no. 298328 to L.K].

References

1. Luscombe NM, Austin SE, Berman HM, *et al*. An overview of the structures of protein-DNA complexes. *Genome Biol* 2000;**1**: REVIEWS001.
2. Charoensawan V, Wilson D, Teichmann SA. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res* 2010;**38**:7364–77.
3. Re A, Joshi T, Kulberkyte E, *et al*. RNA-protein interactions: an overview. *Methods Mol Biol* 2014;**1097**:491–521.
4. Noller HF. RNA structure: reading the ribosome. *Science* 2005; **309**:1508–14.
5. Glisovic T, Bachorik JL, Yong J, *et al*. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett* 2008;**582**: 1977–86.
6. Pruitt KD, Brown GR, Hiatt SM, *et al*. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;**42**: D756–63.
7. Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol Biosyst* 2013;**9**:2417–25.
8. Fornes O, Garcia-Garcia J, Bonet J, *et al*. On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. *Adv Protein Chem Struct Biol* 2014;**94**:77–120.
9. Kauffman C, Karypis G. Computational tools for protein-DNA interactions. *Wiley Interdiscip Rev Data Min Knowl Discov* 2012; **2**:14–28.
10. Liu LA, Bradley P. Atomistic modeling of protein-DNA interaction specificity: progress and applications. *Curr Opin Struct Biol* 2012;**22**:397–405.
11. Choi S, Han K. Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Comput Biol Med* 2013;**43**:1687–97.
12. Panwar B, Raghava GP. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. *Genomics* 2015;**105**:197–203.
13. 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.
14. 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.
15. Puton T, Kozlowski L, Tuszynska I, *et al*. Computational methods for prediction of protein-RNA interactions. *J Struct Biol* 2012;**179**:261–8.
16. 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**:89.
17. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
18. 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.
19. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005;**6**:33.
20. Wang LJ, 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.
21. 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.
22. Kuznetsov IB, Gou ZK, Li R, *et al*. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins Struct Funct Bioinform* 2006;**64**:19–27.
23. 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.
24. Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. *Bioinformatics* 2007;**23**:I347–53.
25. Yan CH, Terribilini M, Wu FH, *et al*. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics* 2006;**7**:262.
26. Lee JH, Hamilton M, Gleeson C, *et al*. Striking similarities in diverse telomerase proteins revealed by combining structure prediction and machine learning approaches. *Pac Symp Biocomput* **2008**:501–12.
27. Wang LJ, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics* 2009;**10**:S1.

28. Wu JS, Liu HD, Duan XY, *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.
29. 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**:e1000567.
30. 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.
31. 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.
32. Carson MB, Langlois R, Lu H. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res* 2010;**38**: W431–5.
33. 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.
34. Jeong E, Chung IF, Miyano S. A neural network method for identification of RNA-interacting residues in protein. *Genome Inform* 2004;**15**:105–16.
35. Jeong EN, Miyano S. A weighted profile based method for protein-RNA interacting residue prediction. *Trans Comput Syst Biol Iv* 2006;**3939**:123–39.
36. 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.
37. Tong J, Jiang P, Lu ZH. RISP: A web-based server for prediction of RNA-binding sites in proteins. *Comput Methods Program Biomed* 2008;**90**:148–53.
38. Kumar M, Gromiha AM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins Struct Funct Bioinform* 2008;**71**:189–94.
39. 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**:S6.
40. Spriggs RV, Murakami Y, Nakamura H, *et al*. Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics* 2009;**25**:1492–7.
41. 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**:W412–16.
42. Huang YF, Chiu LY, Huang CC, *et al*. Predicting RNA-binding residues from evolutionary information and sequence conservation. *BMC Genomics* 2010;**11**:S2.
43. 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 Peptide Sci* 2010;**11**:609–28.
44. Wang CC, Fang YP, Xiao JM, *et al*. Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids* 2011;**40**:239–48.
45. Ma X, Guo J, Wu JS, *et al*. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins Struct Func Bioinform* 2011;**79**:1230–9.
46. 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 Biology* 2011;**8**:988–96.
47. Terribilini M, Lee JH, Yan CH, *et al*. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA Pub RNA Soc* 2006;**12**:1450–62.
48. 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.
49. Pupko T, Bell RE, Mayrose I, *et al*. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 2002;**18** (Suppl 1):S71–7.
50. 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**:e1000567.
51. Chen K, Mizianty MJ, Gao J, *et al*. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure* 2011;**19**:613–21.
52. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;**33**: 2302–9.
53. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
54. Huang Y, Niu B, Gao Y, *et al*. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
55. UniProt C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
56. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
57. Baldi P, Brunak S, Chauvin Y, *et al*. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000;**16**:412–24.
58. Anderson TW, Darling DA. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann Math Stat* 1952;**23**:193–212.
59. Kurgan L, Disfani FM. Structural protein descriptors in 1-dimension and their sequence-based predictions. *Curr Protein Pept Sci* 2011;**12**:470–89.
60. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 1996;**257**:342–58.
61. Zvelebil MJ, Barton GJ, Taylor WR, *et al*. Prediction of protein secondary structure and active-sites using the alignment of homologous sequences. *J Mol Biol* 1987;**195**:957–61.
62. Hsu CM, Chen CY, Hsu CC, *et al*. Efficient discovery of structural motifs from protein sequences with combination of flexible intra- and inter-block gap constraints. *Adv Knowl Discov Data Mining Proc* 2006;**3918**:530–9.
63. Schneider R, Sander C. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res* 1996;**24**:201–5.