Supporting Materials for

DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues

Jing Yan[1] and Lukasz Kurgan[2,*]

[1] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2V4, Canada

[2] Department of Computer Science, Virginia Commonwealth University, Richmond, 23284, U.S.A.

* To whom correspondence should be addressed. Tel: +1 804 827 3986; Email: lkurgan@vcu.edu

Collection of the dataset of non-binding human proteins

We developed a dataset of human proteins that are unlikely to bind DNA and RNA. Similar to the protocol defined in ref. (1), we selected proteins from the UniProt's complete human proteome. This dataset includes proteins that satisfy the following stricter, when compared to ref. (1), seven conditions:

1. Only reviewed entries in UniProt, i.e., these proteins have underwent manual evaluation.
2. Subcellular location does not include nucleus, chromosome, or nucleoplasm.
3. Functional annotations expressed with the gene ontology (GO) terms do not include DNA, RNA, nucleotide, nucleic acid, DNA binding, RNA binding, or nucleotide binding.
4. Protein names do not contain DNA, RNA, nucleic acid, nucleotide, or ribosomal.
5. Functions annotated in UniProt do not include DNA binding, RNA binding, nucleic acid binding, or nucleotide binding function.
6. UniProt records do not have any of the following keywords: DNA, RNA, nucleic acid, nucleotide, ribosomal, ribosome, ribosomal protein, or chromosome.
7. No annotations of interactions with DNA, RNA, or nucleotide.

Using these criteria, we collected a set of 5996 human proteins that are unlikely to bind DNA and RNA. Based on the protocol in ref. (1), we further filtered these proteins by removing the sequences that share ≥30% sequence similarity with each other or with any sequence in the training dataset. This reduces the redundancy within the dataset and also reduces a likelihood that these proteins bind to DNA or RNA given that proteins in the training dataset bind to these nucleic acids. We selected at random 82 proteins from the resulting dataset matching the size of the test dataset to reduce the computational cost of evaluation of multiple considered methods on this non-binding dataset, particularly given the high runtime of some of the existing predictors.

Evaluation criteria

The binary predictions (binding vs. nonbinding) were assessed and compared between different methods using the following three measures:

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

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

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

where TP is the number of true positives (correctly predicted binding residues), FN is the number of false negatives (incorrectly predicted binding residues), FP is the number of false positives (incorrectly predicted nonbinding residues) and TN is the number of true negatives (correctly predicted nonbinding residues).

Selection of amino acid (AA) indices

We collected AA indices from the AAindex database (2). Some of these indices are redundant with each other (they quantify similar properties) and some may not be relevant to the prediction of the RNA and DNA binding. Therefore, we empirically selected a subset of non-redundant and predictive AA indices. Specifically, we removed the indices that are incomplete (with missing values) and those that are not predictive (lack correlations with the prediction outcomes) or redundant (have high mutual correlation with other indices). We computed the point-biserial correlation (PBC) of each index with the DNA-binding annotations and RNA-binding annotations in the training dataset to quantify whether these indices are predictive. Indices with PBC <0.05, which indicates that they offer low predictive power, were removed. Next, we removed the redundant indices among the remaining indices. The indices were sorted based on PBC values in the descending order. We started from the top ranked index, and we added the next ranked index into the pool of retained indices only if its Pearson correlation (PCC) with each of the indices in the pool is ≤ 0.9. As a result, we selected 164 indices that are predictive and non-redundant for the prediction of the DNA-binding residues and 105 indices for the prediction of the RNA-binding residues.

Feature selection and parameterization of predictive models in the first layer

To implement the second step of the first layer, we empirically selected a subset of non-redundant and predictive features. These features are arguably useful to discriminate between binding residues and non-binding residues. There are two types of non-binding residues: the non-binding residues that do not bind to either DNA or RNA and the non-binding residues that do not bind the target type of nucleic acid but bind the other type. For example, when predicting DNA-binding residues, the non-DNA-binding residues include residues that bind RNA and that do not bind either DNA or RNA. We selected features that differentiate between binding and non-binding residues and that also minimize the number of DNA-binding residues that are confused for RNA-binding and vice versa. To do this we assigned weights to the residues in our training dataset. By default, the residues have a weight of 1. We assign weight >1 to the residues that could be cross predicted. This includes the RNA-binding residues for the dataset we used to develop DNA-binding prediction method. It also includes the DNA-binding residues for the dataset we used to develop RNA-binding prediction method. Next, we passed the weight values along with the value of the features into the algorithm that computes the logistic regression model. When building the model, the magnitudes of the prediction errors for the instances (residues) with weight >1 are correspondingly increased when compared to the magnitude of the prediction errors for the instances with weight of 1. This way the regression minimizes mis-predictions of residues with weights > 1. We selected the best weight value by considering values ranging from 1 to 4 with step of 0.2. For each of the considered weight value, we empirically selected a subset of predictive and non-redundant features from the original set of considered features using a two-step feature selection. We performed the selection exclusively using the training dataset with the 5-fold cross validation protocol. We divided the training proteins into 5 folds such that protein chains in a given test fold are dissimilar to the training sequences (sequences in the training folds). This simulates the tests on the test dataset. We clustered the chains in the training dataset using CD-HIT at 30% sequence identity. We assigned the proteins that are clustered in the same group to the same cross-validation fold. In the first step of feature selection, we applied a wrapper-based approach to rank the features. For each feature, we calculated its predictive performance (measured by AULC) when used as an input to univariate logistic

regression model based on the 5-fold cross validation on the training dataset. In the second step, we executed the best first search-based feature selection using the wrapper with logistic regression model to select a subset of predictive and non-redundant features. Starting with the top ranked feature, we accepted the next best-ranked feature into a selected set of features only if the addition of this feature improves AULC by at least 0.0001 based on the 5-fold cross validation on the training dataset when compared with the feature set before this addition. To compare, the AULC of a random predictor = 0.003. We went through the sorted list of features once to select the subset of features. Depending on the weight values (we repeat the selection for each considered value of weight), we selected between 28 (23) and 41 (31) features for the prediction of DNA (RNA)-binding residues. Using the weight = 1.8 (3.6) as an example, Supporting Figure S1 shows the improvement of AULC by gradually (one by one) adding the 41 (31) selected features into the feature subset along the feature selection process. We observed a steady increase in the predictive performance as additional features are added into the set of selected features.

Supporting Figure S2 compares the predictive quality of the logistic regression models trained by using different weight values and the corresponding selected feature subsets. The predictive quality is measured by AULRC on the training dataset based on the 5-fold cross validation. This measure quantifies the amount of cross prediction between RNA and DNA binding residues. We selected the weight value of 1.8 (3.6) with the corresponding subset of 41 (31) features that secures the best predictive quality (lowest AULRC). We utilized these parameters to implement the predictor of the DNA (RNA)-binding residues. We combined the selected features with the 30 features that compose the evolutionary profile. This led to an additional improvement in predictive quality, as shown in Supporting Figure S1. We input the resulting 71 (61) features into logistic regression to build the two prediction models, one for the prediction of DNA-binding and one for RNA-binding.

Feature selection in the second layer

We selected a subset of predictive and non-redundant features using the same feature selection procedure as for the first layer. We first ranked features based on the predictive quality (measured by AULC) of the corresponding univariate logistic regression models on the 5-fold cross validation on the training set. Then starting from the top ranked features, we accepted the next ranked feature into our feature set based on two conditions. First, if the AULC value is not worse by more than 0.0001 compared to the prediction obtained with the model from the first layer. Second, if the AULRC is better by (drops by) at least 0.001 when compared to the prediction using feature set before the addition. We ended up selecting 3 and 3 features for the prediction of DNA and RNA-binding residues, respectively, which are input into the corresponding logistic regression model to generate the final predictions.

Application and assessment on the human proteome

We applied our DRNApred and BindN+ to perform large-scale predictions on the complete human proteome. We assessed their predictive performance by quantifying the extent of their cross-predictions on the known binding proteins from the human proteome and we investigated whether novel binding proteins predicted with DRNApred are likely to be correctly predicted.

Assessment of predictive performance

To evaluate predictions of the DNA/RNA-binding proteins on the entire human proteome, we calculated the extent of the cross predictions of the native DNA and RNA binding proteins. In other words, we evaluated whether a given method specifically predicts only the desired one target type of binding proteins without

confusing DNA-binding and the RNA-binding proteins. We calculated the ratio of the fraction of correct predictions to the fraction of the incorrect cross predictions among the known binding proteins. For example, for the prediction of the DNA-binding proteins, we calculated the ratio of the fraction of the correctly predicted known DNA-binding proteins to the fraction of the predicted DNA-binding proteins among the known RNA-binding proteins. This ratio quantifies the ability of a given method to predict the correct type of binding proteins while maintaining low rate of mis-predictions of the incorrect type of binding proteins. A random predictor would attain the ratio = 1. This means that its fraction of correct predictions in the correct type of nucleic acid is equal to the fraction of incorrect predictions in the other type of the nucleic acid. The ratio > 1 indicates better than random prediction, with a higher number corresponding to a more accurate method.

Moreover, we also compared the cross predictions at the residue level. That is, we assessed whether the considered methods specifically predict the target type of binding residues without confusing the DNA-binding and RNA-binding residues. The proteins in the human proteome are annotated per sequence. That means that we have the information whether a given protein binds to DNA or RNA, but not which amino acids in that protein bind to DNA or RNA. Thus, we perform the tests at the residue level indirectly by investigating whether the predicted binding residues are located in the target type of binding proteins. We calculated the ratio of the fraction of predicted binding residues in the correct type of binding proteins to the fraction of the predicted binding residues in the incorrect type of binding proteins using the set of known binding proteins. For example, for the prediction of the DNA-binding residues, we calculated the ratio of the fraction of the predicted DNA-binding residues in the known DNA-binding proteins to the fraction of the predicted DNA-binding residues in the known RNA-binding proteins. A random predictor would secure ratio = 1. The ratio > 1 indicates that a given method is better than random and higher values correspond to a stronger predictive performance.

Validation of novel DNA and RNA binding proteins and residues predicted by DRNApred

We analyzed the novel DNA and RNA-binding proteins and residues predicted by DRNApred. These residues and proteins do not overlap with the known DNA and RNA-binding proteins.

First, we investigated and compared the annotations of the subcellular localization between the novel and known binding proteins. A pattern of similar localization would suggest that the novel binding proteins are in fact correctly predicted. The subcellular location is annotated based on the GO Cellular Components (GO-CC) terms collected from the UniProt resource for the human proteome. We used all proteins for which this information is complete. Our goal was to find out whether the GO-CC terms associated with the known binding proteins are similar to the GO-CC terms of the novel putative binding proteins. First, for each GO-CC term we calculated its fraction of occurrence (defined as number of occurrences divided by the number of proteins) among the known binding proteins. We also calculated the fraction of occurrence of this GO-CC term in the whole human proteome to establish a point of reference. The GO-CC term is assumed to be enriched in the known binding proteins if the fraction of occurrence in these proteins is much higher (at least 100% increase) than the fraction in the whole proteome. Next, we investigated whether each enriched GO-CC term is also enriched in the novel putative binding proteins. We calculated the fraction of their occurrence among the novel predicted binding proteins, and compared them with the corresponding points of reference (fractions in the whole proteome excluding the known binding proteins). We considered a given GO-CC term as enriched in the novel putative binding proteins if its fraction of occurrence in these proteins is much higher (by at least 100%) compared to the reference. We hypothesized that the putative novel DNA

and RNA binding proteins are correctly predicted if their enriched GO-CC terms cover most of the GO-CC terms enriched in the known binding proteins.

Second, we analyzed and compared the residue level predictions between the novel and known binding proteins. We compared the levels of the positively charged residues (Arginine and Lysine) between the binding and nonbinding residues. This is motivated by the observation that DNA and RNA binding residues are positively charged in order to bind to the negatively charged phosphate backbone of the DNA or RNA molecule. We expected and empirically confirmed that the fractions of the putative positively charged binding residues among the known DNA and RNA binding proteins are substantially higher than the fractions among the putative non-binding residues in these proteins and among the residues in the human proteins. The fractions are defined as a number of positively charged putative binding residues divided by the number of putative binding residues. We also computed the same fractions among the novel putative DNA and RNA binding proteins. We hypothesized that the putative novel DNA and RNA binding residues are likely predicted correctly if their fractions are much higher than the fractions for the non-binding residues in these proteins and for the residues in the human proteins. At the same time, these fractions should be comparable to the corresponding fractions for the known binding proteins.

REFERENCES

1. Yan, J., Friedrich, S. and Kurgan, L. (2016) A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform*, **17**, 88-105.
2. Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. and Kanehisa, M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic acids research*, **36**, D202-D205.
3. Rao, H., Zhu, F., Yang, G., Li, Z. and Chen, Y. (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic acids research*, **39**, W385-W390.

Supporting Table S1. Description of features that were considered in the design of the DRNApred method. Exposed residues are determined using the prediction from the PROFphd method. For the aggregated feature that were computed for the exposed residues, we calculate the average value in two ways: sum of the information for the exposed residues divided by the number of the exposed residues in the window, and sum of the information for the exposed residues divided by the size of the window. The standard deviation is only calculated in the first case. Since the AA indices utilized in the DNA-binding and RNA-binding predictions are different, the corresponding aggregated features are different. Consequently, the number of features for the DNA-binding prediction is shown first and is followed by the number of features for the prediction of RNA-binding that is given inside brackets.

Feature type	**Input type**		**Description**	**Window size *w***	**number of features**
Per residue	Amino acid type		20 dimensional binary vector to encode the amino acid type	*w*=3	60
	Disorder, SS, RSA		We include probability and binary values from 9 methods (5 disorder + 3 RSA + SS).	*w*=3	90
	HMM profile		20 amino acid emission frequencies + 7 transition frequencies +3 local diversities	N/A	30
Aggregated	All residues	Amino acid type	Amino acid composition (20 values for each window size)	*w*={9,11,13,15,17,19, 21, protein length}	160
			Amino acids are divided into 3 groups based on their properties (e.g. charge, hydrophobicity, etc.) (3). We calculate the composition/transition/distribution of the amino acids in each group.	*w*={9,11,13,15,17,19, 21, protein length}	1176
		Disorder, SS, RSA	Content of binary predictions over the window of size w	*w*={9,11,13,15,17,19, 21, protein length}	328
			Average value and standard deviation of the probability predictions over the window of size *w*		
		AA indices	Average value and standard deviation of AA indices over the window of size *w*	*w*={9,21}	656(420)
	Exposed residues	Amino acid type	Amino acid composition of the exposed residues	*w*={9,11,13,15,17,19, 21, protein length}	320
			Composition of the exposed amino acids in each group	*w*={9,11,13,15,17,19, 21, protein length}	336
		Disorder, SS, RSA	Content of binary predictions of the exposed residues over the window of size w	*w*={9,11,13,15,17,19, 21, protein length}	440
			Average value and standard deviation of the probability predictions of the exposed residues over the window of size *w*		
		AA indices	Average value and standard deviation of AA indices of the exposed residues over the window of size *w*	*w*={9,21}	984(630)

Supporting Table S2. Comparison of predictive performance of DRNApred and the other considered predictors of the DNA-binding residues and predictors of the RNA-binding residues on the test datasets. The results on the complete test dataset that includes transferred annotations of binding residues are shown at the top of the table. The results on a version of the test dataset without the transfer of annotations are given at the bottom. Sensitivity, MCC and ratio are calculated from the binary predictions which are converted from the putative real-valued propensities. The conversion applies thresholds for which the number of predicted binding residues equal to the number of native binding residues in the test dataset; this results in the same predicted positive rate PPR = (TP+FP)/(TP+FP+TN+FN) = 0.05 for all methods. Significance of the difference in MCC, ratio, AULC and AULRC values between the best performing method and other methods was assessed based on 10 repetitions that utilize 70% of randomly chosen from the test dataset proteins; + (=) in the Sig column denotes that the difference was (was not) significant at *p*-value <0.05. Methods are sorted by their AULRC value.

Type of dataset	Type of binding	Methods	Sensitivity	MCC	Sig	ratio	Sig	AUC	AULC	Sig	AURC	AULRC	Sig
With transferred annotations	DNA binding	DRNApred	0.25	0.21		0.06		0.77	0.010		0.26	0.039	
		BindN+	0.22	0.18	+	0.13	+	0.79	0.008	+	0.35	0.069	+
		DP-Bind(svm)	0.24	0.20	=	0.14	+	0.75	0.009	+	0.43	0.087	+
		DP-Bind(klr)	0.24	0.20	=	0.15	+	0.76	0.009	+	0.43	0.087	+
		DP-Bind(plr)	0.22	0.18	+	0.16	+	0.74	0.008	+	0.44	0.093	+
		DBS-PSSM	0.21	0.17	+	0.18	+	0.77	0.008	+	0.41	0.095	+
	RNA binding	DRNApred	0.16	0.12		0.02		0.67	0.005		0.25	0.029	
		Pprint	0.15	0.11	=	0.10	+	0.66	0.005	=	0.51	0.121	+
		RNABindR	0.14	0.10	+	0.16	+	0.73	0.004	+	0.51	0.135	+
		BindN+	0.12	0.08	+	0.20	+	0.67	0.003	+	0.63	0.195	+
Without transferred annotations	DNA binding	DRNApred	0.23	0.20		0.06		0.78	0.008		0.28	0.040	
		BindN+	0.20	0.16	+	0.13	+	0.80	0.006	+	0.36	0.072	+
		DP-Bind(svm)	0.23	0.19	=	0.14	+	0.76	0.007	=	0.44	0.087	+
		DP-Bind(klr)	0.23	0.20	=	0.15	+	0.77	0.007	=	0.43	0.088	+
		DP-Bind(plr)	0.21	0.18	=	0.15	+	0.76	0.007	+	0.45	0.092	+
		DBS-PSSM	0.20	0.16	+	0.17	+	0.79	0.006	+	0.41	0.098	+
	RNA binding	DRNApred	0.14	0.11		0.01		0.65	0.003		0.26	0.032	
		Pprint	0.15	0.12	=	0.09	+	0.67	0.004	=	0.50	0.117	+
		RNABindR	0.14	0.11	=	0.15	+	0.74	0.003	=	0.51	0.135	+
		BindN+	0.11	0.08	+	0.17	+	0.68	0.002	+	0.63	0.191	+

Supporting Table S3. Comparison of predictive performance of DRNApred and the selected best-performing other approach based on combination of DB-Bind(svm) and Pprint for the prediction of residues that interact with both DNA and RNA. For each method, sensitivity, MCC and ratio are calculated from the binary predictions which are converted from the putative real-valued propensities. The conversion applies thresholds for which the number of predicted residues that bind both DNA and RNA binding equals to the number of native residues that bind both RNA and DNA. Significance of the difference in MCC, ratio, AULC, and AULRC values between DRNApred and DB-Bind(svm)+Pprint was assessed based on 10 repetitions that utilize 70% of randomly chosen test proteins; + (=) in the Sig column denotes that the difference was (was not) significant at p-value <0.05. Methods are sorted by their AULRC values.

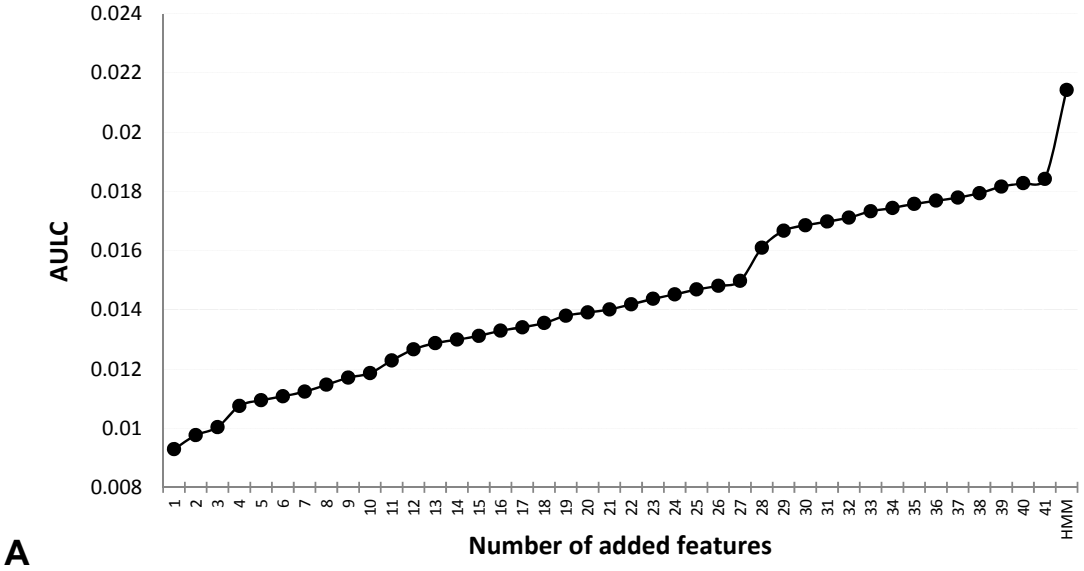
Methods	Sensitivity	MCC	Sig	ratio	Sig	AUC	AULC	Sig	AURC	AULRC	Sig
DRNApred	0.12	0.11		0.05		0.77	0.0010		0.35	0.059	
DP-Bind(svm) + Pprint	0.04	0.03	+	0.06	+	0.86	0.0003	+	0.37	0.086	+

Supporting Table S4. Comparison of predictive performance of DRNApred and the selected best-performing methods for the prediction of RNA-binding residues and the prediction of DNA-binding residues on proteins that interact with RNA/DNA hybrids. Sensitivity, MCC and ratio are calculated from the binary predictions that are converted from the putative real-valued propensities. The conversion applies thresholds for which the number of predicted residues that bind DNA (RNA) equals to the number of native residues that bind DNA (RNA). Statistical significance of the differences could not be estimated given the small size of these datasets. Methods are sorted by their AULRC values.

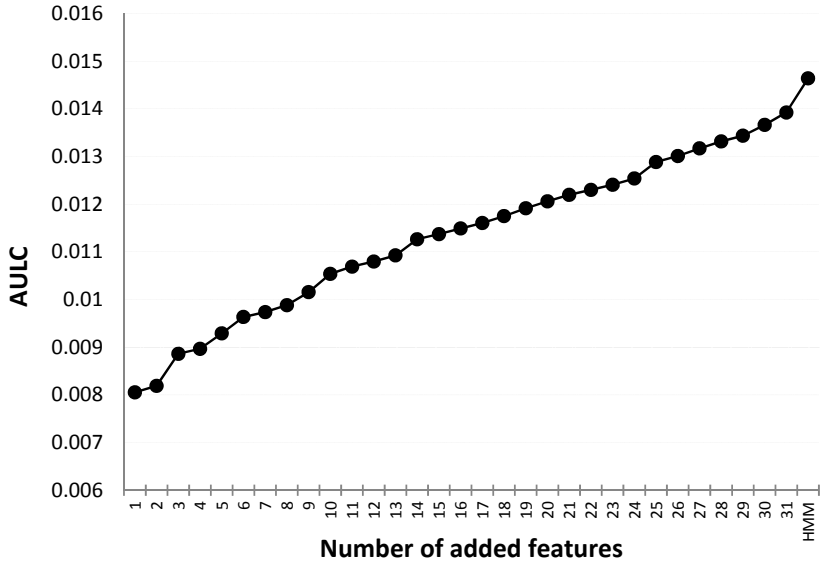
	Methods	Sensitivity	MCC	ratio	AUC	AULC	AURC	AULRC
DNA binding	DRNApred	0.39	0.33	0.57	0.78	0.012	0.61	0.147
	DP-Bind(svm)	0.46	0.40	0.70	0.82	0.008	0.74	0.258
RNA binding	DRNApred	0.18	0.10	0.00	0.74	0.003	0.26	0.038
	Pprint	0.09	0.07	0.09	0.79	0.004	0.39	0.112

Supporting Table S5. Comparison of predictive performance of DRNApred and the other considered methods for the prediction of DNA and RNA-binding proteins on the test dataset. Prediction of all methods are set such that their FPR = 0.05. Significance of the difference in MCC and AUC values between the best performing method and other methods was assessed based on 10 repetitions that utilize 70% of randomly chosen from the test dataset proteins; + (=) in the Sig column denotes that the difference was (was not) significant at p-value <0.05. Methods are sorted by their AUC value.

Binding type	Methods	TPR at FPR = 0.05	MCC at FPR = 0.05	Sig	AUC	Sig
DNA	DRNApred	0.27	0.26		0.68	
	DP-Bind(svm)	0.06	0.00	+	0.54	+
	DP-Bind(klr)	0.04	-0.05	+	0.53	+
	BindN+	0.12	0.10	+	0.52	+
	DP-Bind(plr)	0.02	-0.10	+	0.45	+
	DBS-PSSM	0.00	-0.19	+	0.44	+
RNA	DRNApred	0.30	0.32		0.65	
	Pprint	0.24	0.26	+	0.63	=
	RNABindR	0.12	0.11	+	0.59	+
	BindN+	0.00	-0.16	+	0.45	+

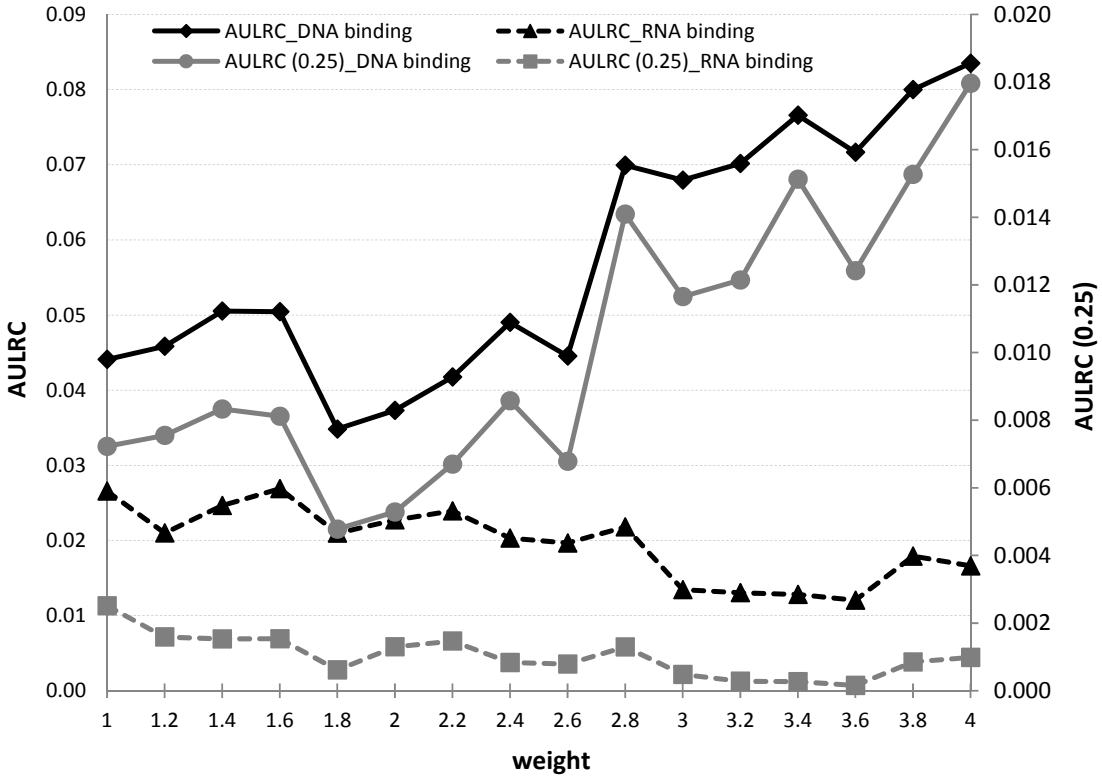


A

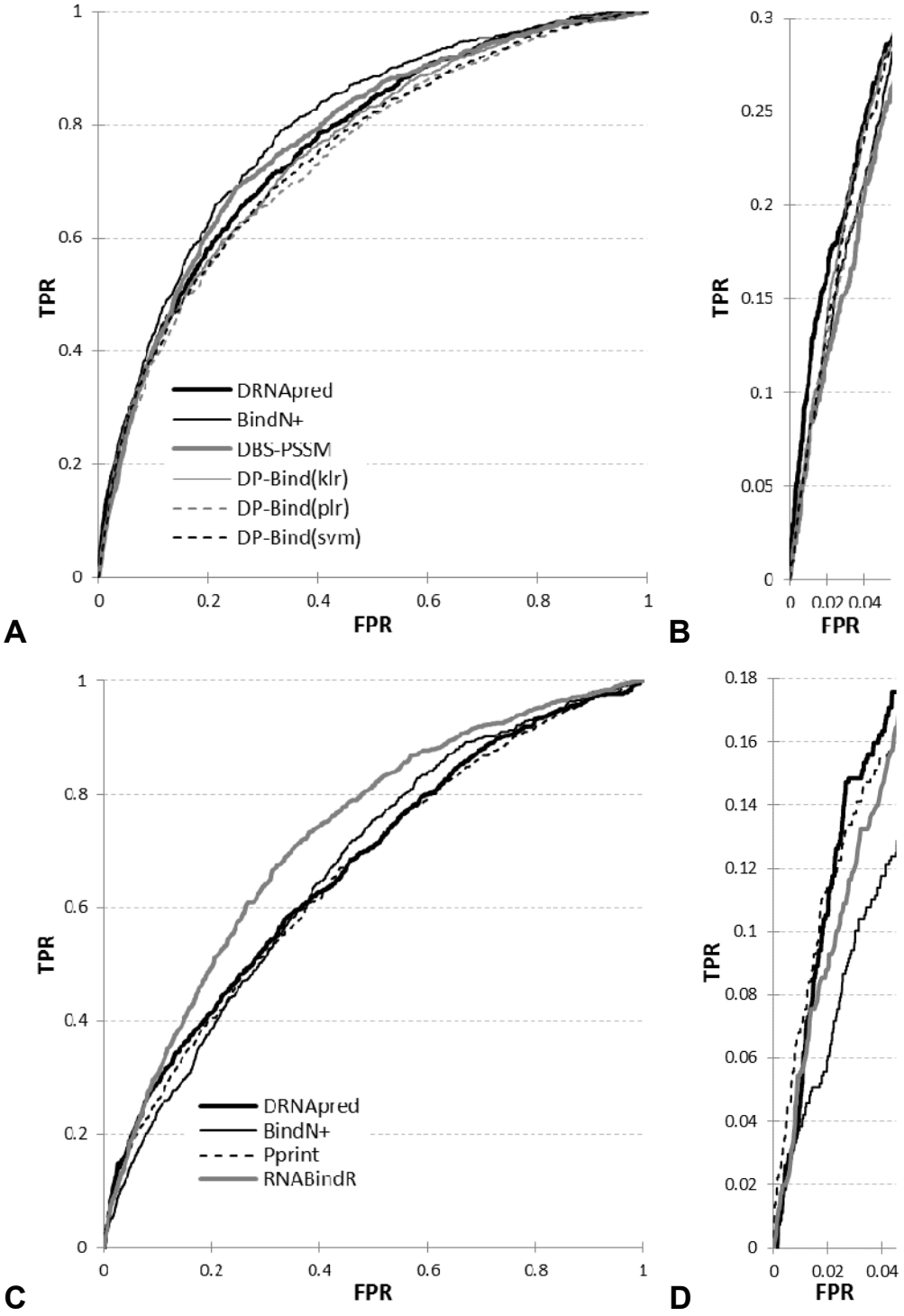


B

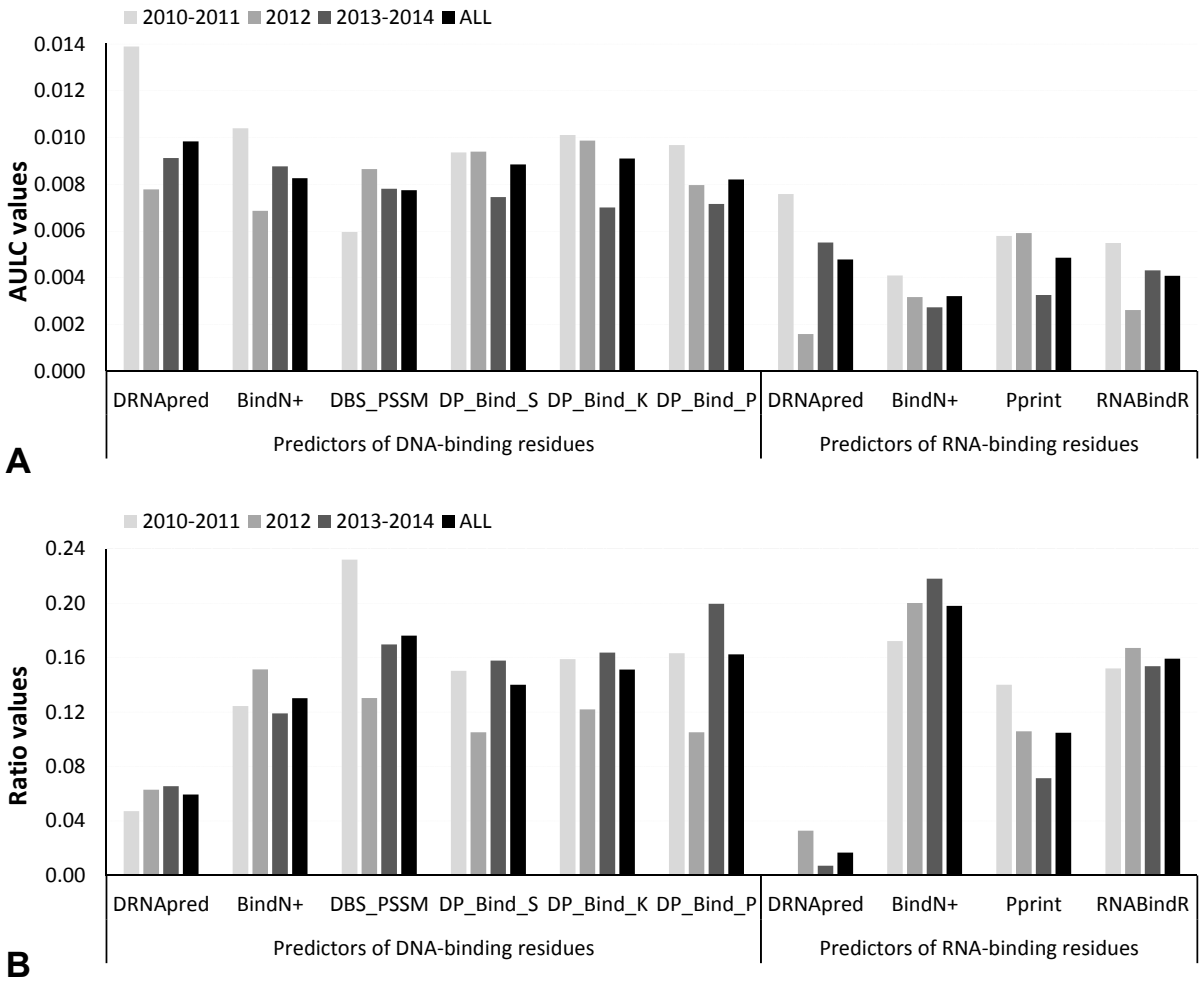
Supporting Figure S1. Improvement in the value of AULC through the feature selection based on 5-fold cross validation on the training dataset. Panel A is for the prediction of DNA-binding residues with the weight value = 1.8. Panel B is for the prediction of RNA-binding residues with the weight value = 3.6. X-axis is the number of features added through the best first search in the feature selection. Last index on the x-axis 'HMM' represents addition of the entire HMM profile that includes 30 features.



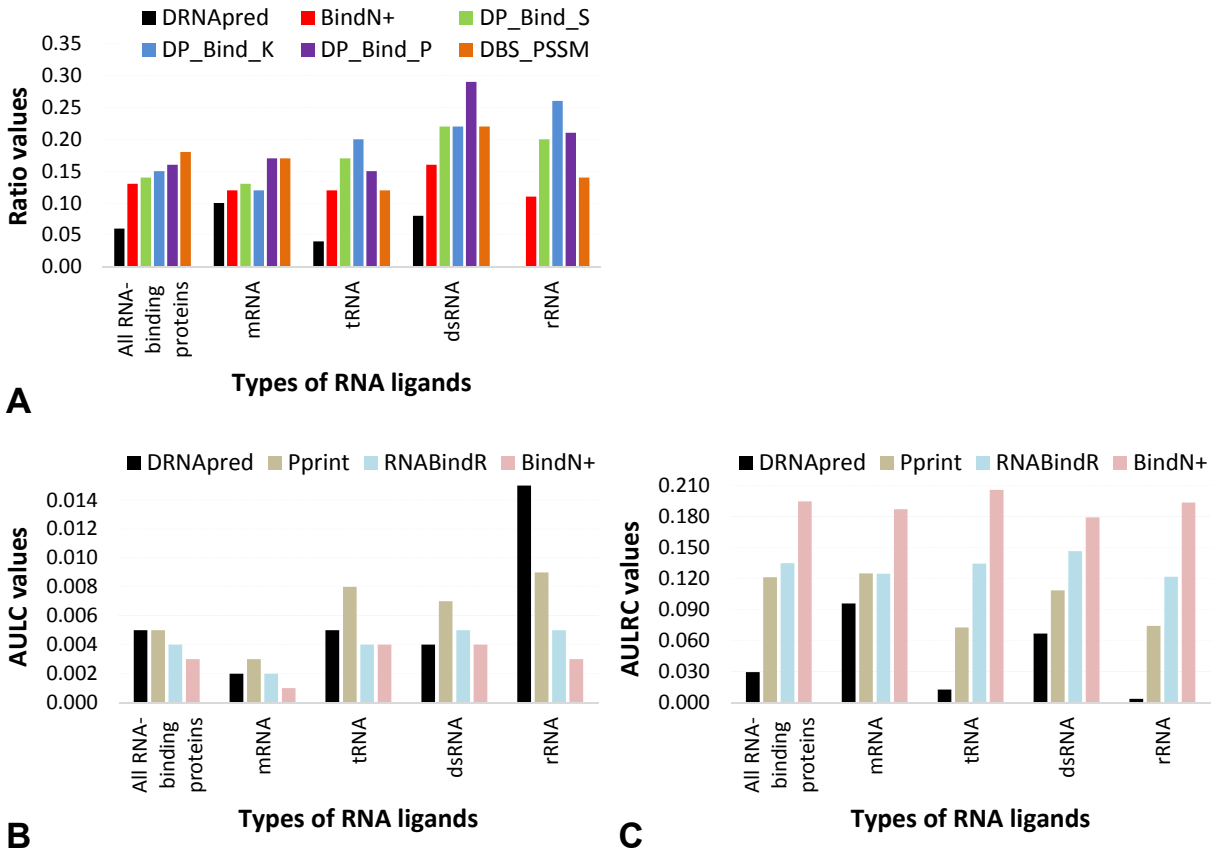
Supporting Figure S2. Predictive performance measured by AULRC on the training dataset based on 5-fold cross validation for the models that use different weights. AULRC (0.25) is equivalent to AULRC but using a smaller cutoff on TPR at 0.25. Lines represent the results for the prediction of DNA-binding residues, and dotted lines indicate the results for the prediction of RNA-binding residues.



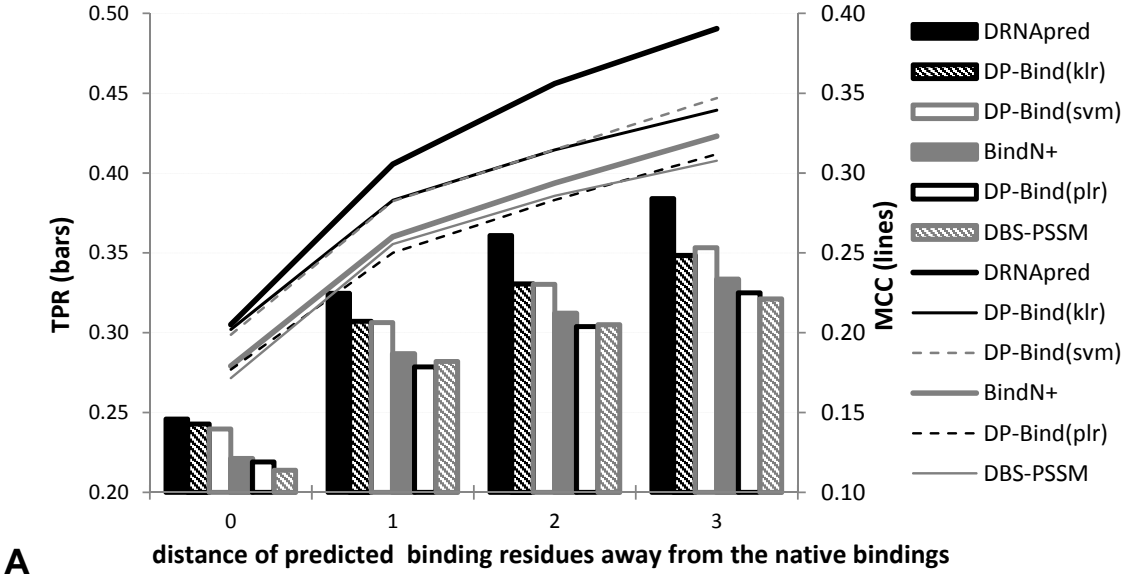
Supporting Figure S3. Comparison of ROCs of DRNApred and the other considered predictors of the DNA and RNA binding residues on the test dataset. Panels A and B are for the DNA binding and panels C and D for the RNA binding. Panels B and D focus on the ROC curve where FPR<5.4% for the DNA binding and <4.5% for the RNA binding; AULC is calculated as the area under that part of the ROC curve.



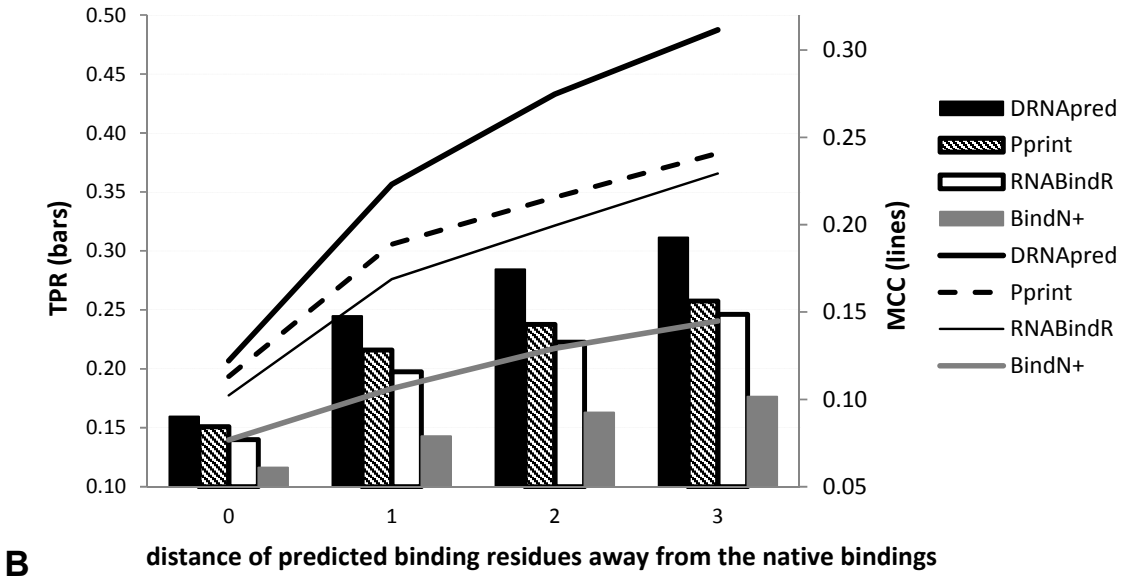
Supporting Figure S4. Predictive performance of DRNApred and other predictors of DNA and RNA binding residues on test proteins sorted according to their release date. Different shades of bars denote the release dates where darker shades correspond more recently released proteins and black denotes the entire test dataset. Panel A shows AULC values while panel B gives ratio values.



Supporting Figure S5. Predictive performance for proteins from the test dataset that bind different types of RNAs including mRNA, tRNA, dsRNA and rRNA. Panel A shows values of ratio for the predictors of DNA binding residues that quantifies fraction of the RNA-binding residues predicted as DNA binding. Panels B and C give AULC and AULRC values for the predictors of RNA binding residues.

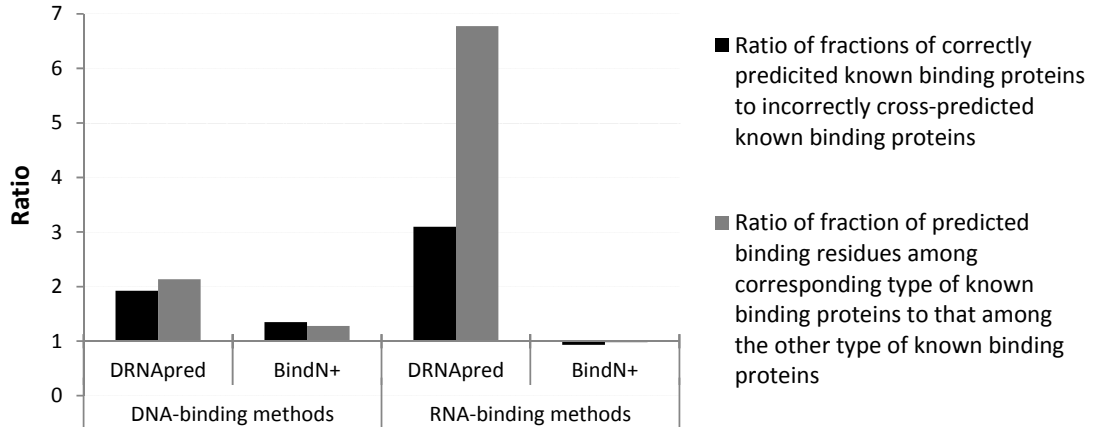


A



B

Supporting Figure S6. Comparison of MCC and TPR values for DRNApred and other considered predictors of DNA and RNA binding residues when reconsidering putative binding residues that are close to native binding residues as correctly predicted (true positives). The predicted binding residues that are no farther than 0, 1, 2, and 3 positions (*x*-axis) in the sequence from the closest native binding residue are considered as correct predictions. TPR values are shown using bars and the y-axis on the left. MCC values are shown using lines and the *y*-axis on the right. Panel A is for the predictors of the DNA-binding residues while Panel B is for the predictors of the RNA-binding residues.



Supporting Figure S7. Predictive performance of DRNApred and BindN+ for the prediction of binding proteins and residues in the known binding proteins from the human proteome. The y-axis shows ratio between the fraction of predictions on the correct type of known binding proteins and the fraction of predictions on the cross predicted type of known binding proteins. Random predictor would return ratio =1 and higher ratio indicates a smaller amount of cross predictions. Black (gray) bars summarize comparison for the prediction of the binding proteins (residues).