Published online 28 January 2017

Nucleic Acids Research, 2017, Vol. 45, No. 10 ***e84***
doi: 10.1093/nar/gkx059

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 and
[2]Department of Computer Science, Virginia Commonwealth University, Richmond, 23284, USA

Received June 10, 2016; Revised January 19, 2017; Editorial Decision January 20, 2017; Accepted January 24, 2017

ABSTRACT

Protein-DNA and protein-RNA interactions are part of many diverse and essential cellular functions and yet most of them remain to be discovered and characterized. Recent research shows that sequence-based predictors of DNA-binding residues accurately find these residues but also cross-predict many RNA-binding residues as DNA-binding, and vice versa. Most of these methods are also relatively slow, prohibiting applications on the whole-genome scale. We describe a novel sequence-based method, DRNApred, which accurately and in high-throughput predicts and discriminates between DNA- and RNA-binding residues. DRNApred was designed using a new dataset with both DNA- and RNA-binding proteins, regression that penalizes cross-predictions, and a novel two-layered architecture. DRNApred outperforms state-of-the-art predictors of DNA- or RNA-binding residues on a benchmark test dataset by substantially reducing the cross predictions and predicting arguably higher quality false positives that are located nearby the native binding residues. Moreover, it also more accurately predicts the DNA- and RNA-binding proteins. Application on the human proteome confirms that DRNApred reduces the cross predictions among the native nucleic acid binders. Also, novel putative DNA/RNA-binding proteins that it predicts share similar subcellular locations and residue charge profiles with the known native binding proteins. Webserver of DRNApred is freely available at http://biomine.cs.vcu.edu/servers/DRNApred/.

INTRODUCTION

Interplay of proteins and the two types of nucleic acids, DNA and RNA, defines and regulates many cellular functions, such as DNA transcription, replication and repair (1,2), protein synthesis, regulation of gene expression, post-transcriptional modifications and posttranscriptional regulation (3–5). The protein–nucleic acids interactions are studied primarily using experimentally derived structures of the corresponding complexes. Unfortunately, use of the experimental methods is technically challenging and relatively expensive and thus only ∼6000 protein–nucleic acids complexes were characterized so far (source: Protein Data Bank (PDB) database (6) as of December 2016). However, the number of DNA-binding proteins is relatively substantial and was estimated to be on average close to 3% in eukaryotic organisms and 5% in animals (2). Similarly, the fraction of RNA-binding proteins was estimated to range between 2% and 8% of proteins in eukaryotic organisms (5). Simple math reveals that assuming the most conservative estimates we should know 2% of 43 million of known eukaryotic proteins (source: the NCBI's RefSeq database as of December 2016) = 860 thousand proteins that bind RNA and 3% of 43 million = 1.29 million proteins that bind DNA. The annotated DNA and RNA binding proteins can be obtained from several databases and datasets: RBPDB database (7), animalTFDB database (8,9), UniProt (10) database that includes annotations of nucleic acids binding via the gene ontology (GO) terms, and in datasets collected in recent studies (11,12). Using human proteome as an example, these databases collectively annotate 4.9% and 2.7% of the human proteins as DNA-binding and RNA-binding, respectively. These numbers are low given that the fraction of human transcription factors alone was approximated to be 7.9% (9) and the number of RNA-binding proteins was recently estimated to be at least 7.5% (13).

The substantial and growing gap between the number of known and the number of yet to be learned DNA and RNA binding proteins motivates the need to increase the pace of the characterization of protein–DNA and protein–RNA interactions. To this end, the experimental data are utilized to develop a number of time- and cost-efficient computational models for automated prediction of these interactions for the millions of the uncharacterized proteins.

*To whom correspondence should be addressed. Tel: +1 804 827 3986; Fax: +1 804 828 2771; Email: lkurgan@vcu.edu

These methods can be categorized into two types according to the input information that they use: structure-based versus sequence-based. The former methods predict the binding based on known protein structures, while the latter make the predictions solely from the protein sequences. The structure-based methods utilize information derived from the structure, typically based on shape and biophysical characteristics of the protein surface. However, structures are known for only ∼40 thousand distinct protein sequences (source: PDB database as of December 2016), limiting utility of the structure-based methods. Therefore, it is necessary to develop reliable computational methods to identify nucleic acids binding from the sequences. There are two types of sequence-based methods: those that predict DNA- or RNA- binding proteins versus DNA- or RNA-binding residues in protein sequences. The former type concerns a simple two-state prediction of whether a given protein sequence binds to DNA/RNA or not. The latter type goes much further by locating the binding residues (residues in contact with DNA/RNA) in the input protein sequences. Therefore, we focus on the computational prediction of DNA- and RNA-binding residues from protein chains. These methods can be used to find putative binding proteins in the vast sequence databases and to indicate putative sites of these interactions. A couple dozen of sequence-based methods that predict the DNA- or RNA-binding residues have been already published. They include DNA-binding predictors: DBS-Pred (14), DBS-PSSM (15), BindN (16), DNABindR (17,18), DP-Bind (19,20), DISIS (21), BindN+ (22), NAPS (23) and MetaDBSite (24); RNA-binding predictors: BindN (16), RNABindR (25–27), Pprint (28), RNAProB (29), PiRaNhA (30,31), BindN+ (22), NAPS (23), SPOT-seq (32) and Meta2 (33); and predictors of nucleic acid binding residues that do not discriminate between RNA and DNA: RBscore (34,35). Although the first two types of methods accurately predict their own type of nucleic acid binding (36,37), a recent study shows that they often fail to correctly recognize whether a given interaction is with RNA or DNA (36). More specifically, methods for the prediction of DNA-binding residues were found to cross-predict between 29% and 49% of RNA-binding residues as DNA-binding. Similarly, methods for the prediction of RNA-binding residues mispredict between 48% and 64% DNA-binding residues as RNA-binding. This observation was confirmed by another study that demonstrated that the current methods have AUC <0.5 (equivalent to a random prediction) for discriminating DNA- versus RNA-binding residues (37). A likely reason for this is that these methods were developed and tested using proteins that bind one type of nucleic acids (predictors of DNA binding were built using exclusively DNA-binding proteins) and thus they could not learn to discriminate between DNA and RNA. This is an important shortcoming because DNA and RNA binding residues carry out different cellular functions. Consequently, new methods that can accurately discriminate between the two nucleic acids are needed. Moreover, most of the existing methods require a substantial amount of runtime, which makes it very difficult to apply them on a large scale, say to analyze whole proteomes.

We describe a new method, DRNApred, that accurately and in high-throughput predicts protein–RNA and protein–DNA binding residues from protein sequences. Its ability to differentiate between these two types of nucleic acids stems from several novel design ideas that are empirically shown to boost predictive performance. Besides tests on a benchmark dataset, we also apply our runtime efficient method to predict the DNA- and RNA-binding proteins and residues in the human proteome to further validate predictive performance and annotate novel putative DNA- and RNA-binding proteins.

MATERIALS AND METHODS

Datasets

We expand a benchmark dataset of the DNA- and RNA-binding proteins from a recent comparative review in (36) with new data. These data are used to design datasets to empirically build and comparatively assess our method. The main advantage of the source dataset is the inclusion of a more complete annotations of nucleic acids binding residues when compared to the other, older datasets (15,22,24,25,27). This was accomplished by transferring annotations of binding residues between the same or virtually identical proteins in multiple complexes where they bind to potentially different fragments of DNA or RNA. As shown in (36), this increases the number of annotated binding residues by 14% and 10% for the DNA and RNA-binding residues, respectively. Previous datasets would only use annotations from one complex or use each complex independently. The annotation of binding residues follows standards in this field (36). A given residue is defined as binding if at least one of its atoms is closer than a cutoff distance from an atom of the DNA/RNA molecule. We use the 3.5 Å cutoff distance which was also the most often used in prior comparative studies (24,36,38,39) and when building prior predictors (36).

The original dataset was expanded by collecting 564 protein–DNA, 72 protein–RNA and 16 protein–DNA–RNA high resolution (better than 2.5Å) complexes that were released in PDB (6) after the original dataset was collected. The corresponding 892 DNA-binding and 145 RNA-binding chains were combined with the previous dataset to obtain total of 2827 DNA-binding and 1125 RNA-binding chains. Next, following the protocol in (36), we transferred binding annotations between proteins that share similar sequence and structure. We clustered proteins that share ≥80% sequence similarity and ≥0.5 TM scores and moved annotations between proteins in the same cluster. However, unlike in (36) where clustering was done separately for the DNA and RNA binding proteins, we clustered them together to further improve accuracy of annotations. We transferred DNA and RNA binding residues from all chains in the same cluster into a representative chain that has the largest number of binding residues. We also updated the deposition date of the representative chain to the most recent time among all chains in the same cluster. Following (37), we ensured that the proteins used to test and compare predictors are independent with the training proteins that are utilized to design the predictive model. This means that the test proteins are dissimilar in the sequence and deposited at a later time compared with the training proteins. This is also why we did not use test datasets from prior

studies that would inevitably share similarity to our training dataset. The datasets used by the predictors of DNA- and RNA-binding residues that are included in the comparative assessment were collected before November 2010. Correspondingly, the binding proteins released before November 2010 are assigned into the training dataset, and the remaining proteins are assigned into the test dataset. Moreover, we reduced the sequence similarity between training and test datasets. We filtered the test proteins by removing every sequence that shares >30% sequence similarity with any training sequence based on pairwise sequence similarity computed with the bl2seq program (40). Some of the existing predictors of DNA and RNA binding residues that we compare with could not complete predictions for proteins that are over 1000 residues long. Thus, we removed five and three such long proteins from the training and test datasets, respectively. Moreover, we also developed a version of the test dataset without the transfer of annotations of binding residues. The number of proteins and annotations of RNA and DNA binding residues in these datasets are summarized in Table 1. The differences in the number of binding residues between the two versions of the test datasets show that the transfer of annotations resulted in the enrichment by 17% in the DNA binding residues and by 18% in the RNA binding residues. Residues with missing coordinates in the training and test datasets (disordered residues for which we cannot annotate binding) are excluded from the evaluation. We balanced the training dataset by under-sampling the nonbinding residues since there are substantially more nonbinding residues than binding residues. Some of the non-binding residues bind to the other type of nucleic acid, i.e. DNA binding residues when predicting RNA binding and vice versa. These residues are important to study whether predictions discriminate between DNA and RNA binding. Thus, we keep all non-binding residues and under-sample 25% (15%) of the remaining nonbinding residues that do not bind to either DNA or RNA molecule. This way the number of the non-binding residues in the training dataset is about twice larger than the number of the DNA-binding (RNA-binding) residues.

We extracted the complete human proteome (69 178 human proteins) from the UniProt database to apply and evaluate our method on the proteome scale. We annotated a comprehensive set of the native RNA and DNA binding proteins in this proteome using the databases utilized in (41): UniProt (10), RBPDB (7), animalTFDB (8,9) and two recently curated datasets (11,12). By combining information from these five resources we annotated 3360 DNA-binding proteins (4.9% of human proteome) and 1855 RNA-binding proteins (2.7% of the human proteome). These two datasets are available at http://biomine.cs.vcu.edu/servers/DRNApred/.

We also collected a set of proteins that are unlikely to bind either DNA or RNA. These proteins are used to investigate whether the considered methods predict binding residues in these proteins. We applied protocol from (36) to collect such human proteins based on their subcellular localization, names, functional annotations and keywords using reviewed entries in UniProt. This dataset includes 82 dissimilar (<30% pairwise similarity) non-binding proteins to match the size of the test dataset. Details are given in the Supporting Materials.

Evaluation criteria

Evaluation of predictive quality is performed for the two types of predictions: binary prediction (binding versus non-binding residues) and real-valued propensities. The real-valued propensities quantify the propensity that a given residue binds a given type of nucleic acid. We exclude residues with missing atomic coordinates in the source structure files (i.e. disordered residues) since their annotations of binding could not be computed. The binary predictions were assessed using sensitivity, specificity and Matthews correlation coefficient (MCC). These three measures were used in similar works that addressed prediction of DNA or RNA binding residues (17,22,24,27,33,38,42,43); definitions are provided in the Supporting Materials.

The predicted propensities are evaluated using the receiver operating curves (ROCs). ROC is a plot of false-positive rate against the true-positive rate computed by binarizing the propensities using thresholds. FPR = 1 – specificity and quantifies fraction of non-binding residues incorrectly predicted as binding (false positives). TPR is the same as sensitivity and quantifies fraction of correctly predicted binding residues (true positives). We report the area under the ROC curve (AUC), the same as in a number of related studies (22,27,33,44). The fraction of the DNA- (RNA-) binding residues is 8.2% (4.8%) in our training dataset. Thus, even a small FPR = 0.2 corresponds to the prediction where the binding residues are over-predicted by 2.5 (4) times compared to their native number. Therefore, we focus our assessment of the predictive performance on the part of the ROC where number of FPs is no bigger than the number of actual positives (native binding residues). This corresponds to the lower (left side) part of ROC where FPR ≤8.2% (5.4%) for DNA and ≤4.8% (4.5%) for RNA on the training (test) dataset. Consequently, we also report the area under this lower part of curve (AULC). We could not compute the measures introduced in (37), which include weighted arithmetic mean of AUC and mean of AUC. These are computed as average of per protein scores that are impossible to compute for half of our test dataset. This is because we include both RNA and DNA binding proteins and our point is to test all predictors on the complete test dataset with both types of proteins. This way when testing RNA binding predictors half of the test proteins does not include RNA binding residues and thus computation of these types of AUC values is not possible; the same is true for the DNA binding predictors.

We evaluate the extent to which different methods cross-predict between the two types of nucleic acid binding residues using the ratio measure and the ratio curve that were introduced in (36). 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. The ratio curve is a plot of ratio against the TPR, which is calculated by binarizing the propensities using thresholds. We report the area under the ratio curve (AURC). Given the low numbers of binding

Table 1. Summary of datasets

Dataset name	Number of proteins	Number and fraction of DNA binding residues	Number and fraction of RNA binding residues
Training	488	7823 (7.6%)	4699 (4.6%)
Test with transferred annotations	82	968 (5.1%)	808 (4.2%)
Test without transferred annotations	82	827 (4.3%)	682 (3.6%)

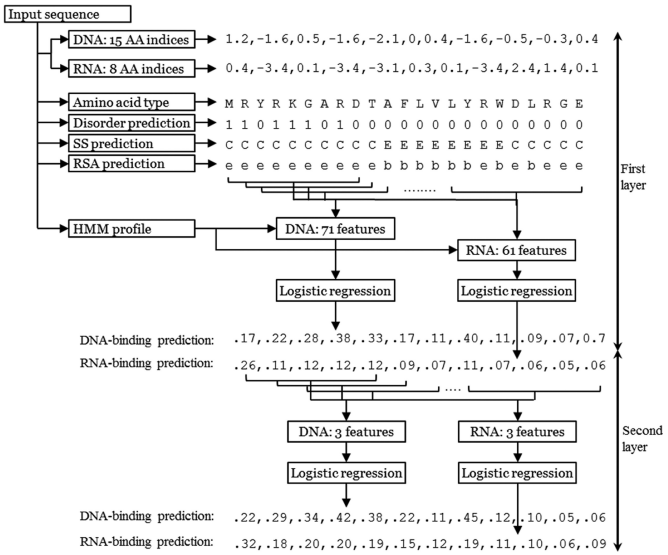

Figure 1. Architecture of DRNApred predictor.

residues we also quantify the area under the lower part of the curve where TPR $\leq$0.5 (AULRC). The larger values of TPR are less interesting because such predictions result in prohibitively high FPRs.

Similar to (36,41), we evaluate significance of the differences in predictive quality between the best-performing prediction method and each of the other considered methods on the test dataset. The underlying goal is to investigate if the results on the test dataset are not biased by a subset of test proteins by measuring whether the predictive quality is consistent over different subsets of proteins. We randomly select 70% of test proteins to calculate the corresponding MCC and AUC values. This is repeated 10 times and we compare the corresponding 10 paired results. If the measurements are normal, as tested using the Anderson–Darling test (45) with 0.05 significance, then we apply the paired *t*-test to investigate significance. Otherwise, we utilize the Wilcoxon rank sum test. The difference between a given pair of predictors is assumed statistically significant if *P*-value $<$0.05.

Architecture of the DRNApred predictor

DRNApred predicts DNA- and RNA-binding residue using a two-layer design, see Figure 1. The first predictive layer includes three steps. In the first step, a variety of physicochemical and biochemical properties together with hidden Markov model (HMM) based evolutionary profile and putative intrinsic disorder, secondary structure and solvent accessibility are calculated and predicted from the input protein sequence. We considered 15 and 8 physicochemical and biochemical properties for the prediction of DNA and RNA binding, respectively. In the second step, these inputs are processed using a sliding window to generate a set of empirically selected 71 (for DNA binding) and 61 (for RNA binding) features. In the third step, these feature sets are input into logistic regression models that generate predictions of the DNA- and RNA-binding residues. Use of the logistic regression is motivated by several factors: (i) this model has been used in a related study (41); (ii) simplicity of this linear model reduces likelihood of overfitting the training dataset compared to models that use a larger number of parameters; (iii) this model is fast to generate on the training dataset which is important given a relatively large size of our dataset and (iv) Predictions with this model are also fast, which is crucial for a runtime efficient (high-throughput) predictor.

The second predictive layer uses the predictions generated in the first layer as the input, see Figure 1. It aims to improve over the predictions from the first layer by considering information about the putative propensities for RNA and DNA binding for the residues that are adjacent in the sequence. The corresponding neighborhoods in the sequence are defined using sliding windows (Figure 1). Intuitively, amino acids surrounded by a large number of residues that have high putative propensities for DNA binding are more likely to bind DNA compared to amino acids surrounded primarily by residues that have lower predicted propensities for DNA binding. Moreover, amino acids neighboring many residues that have high propensities for both types of binding are likely to bind both RNA and DNA. However, amino acids surrounded by many residues with high propensities for DNA binding and a few residues that have modest propensities for RNA binding are more likely to bind only DNA. We use the predictions of the DNA-binding and RNA-binding residues generated in the first layer to compute two sets of three empirically selected numerical features that discern such patterns. One set of features is for the prediction of DNA binding and the other set for the prediction of RNA binding. These feature sets are fed into two corresponding logistic regression models that re-predict the DNA-binding and RNA-binding residues.

DRNApred is designed to accurately, specifically, and in high-throughput fashion predict DNA-binding and RNA-binding residues from the protein sequence. In particular, we focused on reducing the cross predictions between DNA and RNA binding residues. We have utilized three novel strategies to address this aim. First, our model is designed on a new training dataset with both DNA-binding and RNA-binding proteins compared to the prior methods used either DNA or RNA binding proteins. Second, we explicitly penalize the cross predictions in the regression model in the

first layer. Third, the second layer is designed to further reduce the cross predictions.

Design of the first predictive layer

We applied a shotgun approach by generating a large variety of structural and physicochemical properties of the input sequence and encoding them into a large number of numerical features. Moreover, we computed these features utilizing sliding windows of different sizes. Next, we empirically selected a smaller subset of predictive and non-redundant features is from this large set of considered features.

In the first step of the first predictive layer (Figure 1), we consider a comprehensive set of properties of the input sequence including amino acid (AA) type, information derived from putative intrinsic disorder, secondary structure (SS) and solvent accessibility (SA), AA indices that quantify physicochemical properties of residues in the input protein sequence, and evolutionary profile of that sequence. These properties have already been used in previous studies that focused on the prediction of DNA or RNA binding (36). We predicted intrinsic disorder with IUPred (46) and Espritz (47) methods. SS was predicted with the fast version of PSIPRED that does not use sequence alignment (48) and SA was predicted with PROFphd (49), NETASA (50) and RVP-net (51) methods. These predictions were performed using runtime-efficient predictors to ensure that DRNApred is computationally efficient. We collected AA indices from the AAindex database (52). We considered 164 (DNA) and 105 (RNA) non-redundant and relevant to our prediction indices that we empirically selected from the original list of over 500 indices; details are given in the Supporting Materials. The evolutionary profiles were generated using HHblits (53) with the default parameter settings and the *nr* database. The profiles are in the form of $N * 30$ matrix, where N is the length of the input protein sequence. For each position n_i in the input sequence, $i = 1,2,\ldots N$, these profiles consist of 30 scores. These scores include 20 values that represent observed frequencies of the 20 AA types in homologous proteins, seven transition frequency scores that quantify probabilities to observe a match, insertion or deletion after this position and the three local diversity values that quantify the diversity of the aligned sequences in a region around the position n_i.

We process the abovementioned properties in the second step of the first predictive layer using sliding windows to generate a large set of numerical features (Figure 1). Sliding windows are centered on the predicted residue a_i to accommodate for information carried by the adjacent residues. For each property, we consider two types of features:

- Per residue features that are computed for each residue in the window. We apply a sliding window of size 3 to include the information about a_i and its two immediate neighbors. The corresponding features are $[V_{i-1}, V_i, V_{i+1}]$, where V is a feature vector that includes the AA type and predicted disorder, SS and SA. We use default values for the neighbors of residues at the either termini of the sequence that are missing.
- Aggregate feature that are computed by combining information coming from multiple residues in the window. This information includes AA types, values of selected AA indices, and predicted disorder, SS and SA. We combine these values over the whole sliding window. Moreover, we also filter the positions in the window using the SA predictions to combine values only for the solvent exposed residues in the window. This is motivated by the fact that binding residues are typically located on the protein surface. We vary the window size from 9 to 21 with a step of 2. We also compute the same aggregated values for the entire protein chain. These aggregate features are inspired by recent works (41,54–56).

Detailed description of the features is given in Supporting Table S1. In total, we considered 4580 features for the prediction of the DNA-binding residues, and 3990 features for the prediction of the RNA-binding residues.

Next, we empirically selected a subset of non-redundant and predictive features that can discriminate between DNA-binding, RNA-binding and non-binding residues. We introduced weights for the residues in the training dataset to improve discrimination between DNA and RNA binding residues. We set weights to values >1 for resides that could be cross-predicted (RNA-binding residues in the dataset to develop DNA-binding prediction method and vice versa) and to 1 for the remaining residues. These weights are passed into the regression model to amplify the errors associated with the cross-predictions. We used the training dataset to optimize the values of the weights and to perform feature selection. We implemented feature selection with a wrapper-based approach and the best first search. We performed the selection for each of considered 16 values of weights (between 1 and 4 with step of 0.2). We selected weight values and features that leads to the best predictive performance measured with AULC (to choose features) and AULRC (to select the weight) based on five-fold cross validation on the training dataset. Consequently, we selected the weights = 1.8 and 3.6 with the corresponding sets of 71 and 61 features for the prediction of the DNA-binding and RNA-binding residues, respectively. Detailed description of the feature selection and selection of the weights is given in the Supporting Materials and Supporting Figures S1 and S2.

Design of the second predictive layer

We used the predictions of the DNA and RNA binding residues generated by the two models from the first predictive later to more accurately re-predict propensities for DNA and RNA binding. The design of the second layer includes two steps. In the first step, we generated a set of per residue features and aggregate features from the two predictions using a sliding window. For the per residue features, we set the window size to 3 to include the predictions of the DNA-binding and RNA-binding for the two immediate neighbors of the predicted residue. The aggregate features include the content of predicted DNA-binding and RNA-binding residues and averages and standard deviations of the predicted propensities for DNA-binding and RNA-binding in the windows. We used ten window sizes between 3 and 21 with a step of 2. We also calculated the same aggregate values for the whole sequence. This totals

to 122 features. In the second step, we empirically selected a subset of predictive and non-redundant features using the same feature selection procedure and cross-validation on the training dataset that we used in the first predictive layer. Consequently, we selected two sets of three features, one for the prediction of DNA-binding residues and the other for the prediction of the RNA-binding residue. More details on the feature selection are given in the Supporting Materials. Each feature set is input into the corresponding logistic regression model to generate the final predictions.

RESULTS AND DISCUSSION

Improvement in predictive performance due to the use of novel design strategies

We used three novel design ideas to reduce the cross-predictions between the two types of nucleic acid binding residues: training dataset that includes DNA-binding and RNA-binding proteins; weights to compute regression models in the first predictive layer; and the second predictive layer. We compared the results obtained by a complete predictive model with the results obtained when designing the model without the use of these strategies to quantify their impact on the predictive performance. We considered and compared the following four scenarios:

1) The predictor developed on the training dataset with just one target type of nucleic acid binding proteins (referred to as only DNA (RNA) binding data).
2) The predictor trained on the combined training dataset of both DNA-binding and RNA-binding proteins (referred to as combined data).
3) The predictor designed on the combined training dataset and using the weights to minimize the cross predictions (referred to as combined data with penalty).
4) The complete predictor implemented using two predictive layers based on the combined training dataset and weights (referred to as second layer).

We evaluated the predictive performance for these four scenarios on the test dataset, see Figure 2. For the DNA-binding models, the predictive quality measured by AUC and AULC is very similar across the four scenarios. However, the amount of cross predictions quantified by AURC and AULRC decrease dramatically as we improve our design by adding additional strategies. We reduced the cross prediction measured with AURC (AULRC) by 10% (25%), while maintaining similar overall predictive quality measured with AUC and AULC when comparing the 'only DNA-binding data' and 'combined data' scenarios. The model based on the 'combined data with penalty' scenario that uses weights further decreases AURC and AULRC by 12% and 20%, respectively, while maintaining equivalent values of AUC and AULC. The last 'second layer' scenario provides the best predictive performance by again decreasing the amount of cross predictions. This is evidenced by lower values of AURC and AULRC and similar values of AUC and AULC when compared to the 'combined data with penalty' scenario. The same observations are true for the models that predict RNA-binding residues. The model based on the 'second layer' scenario maintains the overall predictive quality measured with AUC and AULC and substantially reduces the cross prediction measured with AURC and AULRC when compared to the other three scenarios. Consequently, we applied the two models that use all four strategies in the DRNApred method.

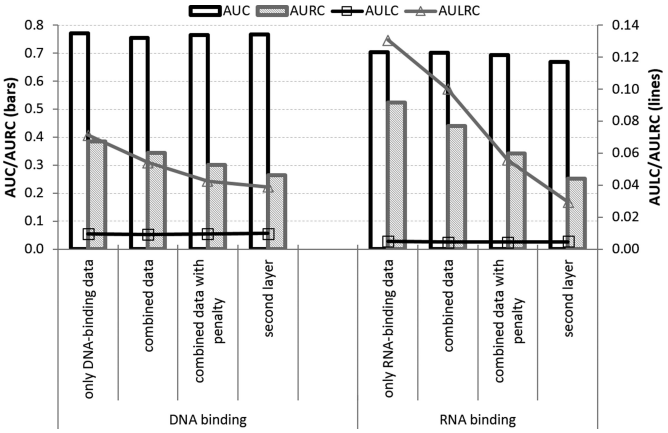

Figure 2. Comparison of predictive performance on the test dataset using different designs of the models for the prediction of DNA-binding (RNA-binding) residues. Bars for the AUC and AURC are quantified with the *y*-axis on the left side while lines for the AULC and AULRC are quantified with the scale on the *y*-axis on the right. The 'only DNA-binding data' ('only RNA-binding data') scenario is the model for the prediction of DNA-binding (RNA-binding) residues designed on the training dataset with just DNA-binding (RNA-binding) proteins; the 'combined data' scenario is for the model built on the combined training dataset with both DNA-binding and RNA-binding proteins; the 'combine data with penalty' scenario is for the model that uses combined training dataset and weights that are used to penalize the cross predictions; the 'second layer' scenario considers model that extends the 'combine data with penalty' scenario with the second layer.

Comparative assessment of predictive performance for the prediction of the DNA binding residues and RNA binding residues

We assessed DRNApred for the prediction of DNA-binding residues and the prediction of RNA-binding residues on the test dataset with the complete (transferred) annotations of binding. We compared these results with the results generated by existing methods that predict DNA or RNA binding residues. We included five predictors of DNA-binding residues and three methods that predict RNA-binding residues that were selected for empirical assessment in a recent comparative review (36). The criteria used to select these methods in that article were availability of web-servers and short runtime. More specifically, the selected methods predict an average size protein sequence with 200 residues in under 10 min.

DRNApred that was evaluated for the prediction of DNA binding residues secures similar overall predictive quality quantified with AUC and AULC values when compared to the representative predictors of DNA binding residues. These results are summarized in the top portion of the Supporting Table S2. AULC is the area under the ROC curve where FPR $<$ 5.4%. Since there are 5.4% of native DNA binding residues in the test dataset the corresponding part of the ROC curve covers predictions where the num-

ber of non-binding residues incorrectly predicted as binding (false positives) is smaller than the number of native binding residues (positives). In other words, this is where the predictor does not overpredict the binding residues. Although DRNApred's AUC is lower than that of the best method BindN+ and comparable to the other considered methods, DRNApred's AULC is the highest and significantly better than the AULC values of all other methods. The corresponding ROC curves are shown in Supporting Figure S3A (complete curve) and S3B (curve used to compute AULC) in the Supporting Materials. The curves of BindN+ and DBS-PSSM are better than the DRNApred's curve when FPR values are high and worse for the arguably more practical range with the lower values of FPR<5.4% (Supporting Figure S3B in the Supporting Materials). Importantly, the fraction of correctly predicted binding residues (TPR) of DRNApred is about six times higher than its FPR at FPR = 5.4%. Close to 30% of the native DNA-binding residues can be found at this low FPR. This means that DRNApred correctly locates a large fraction of native binding residues when mis-predicting a relatively low fraction of the native non-binding residues. We binarize the propensities generated by the considered methods to classify each residue as binding (propensity > threshold) and non-binding (propensity ≤ threshold). The threshold is determined to ensure that the number of predicted binding residues equals to the number of native binding residues in the test dataset. These binarized predictions are assessed with sensitivity and MCC; specificity is virtually identical for different methods given how the threshold was selected. DRNApred offers slightly higher sensitivity and comparable MCC when compared to the other considered predictors of DNA binding residues (see the top portion of the Supporting Table S2). Although DRNApred's overall predictive performance for the prediction of the DNA binding residues is similar to the other methods, our predictor significantly reduces the cross predictions between DNA and RNA binding residues. This is measured with AURC (area under the ratio curve) and AULRC (area under the lower range of the ratio curve where TPR<0.5). DRNApred obtains the lowest AURC and AULRC values which are lower by (0.35 – 0.26)/0.35 = 26% and (0.069 – 0.039)/0.069 = 43%, respectively, compared to the second best BindN+ predictor. Figure 3A which plots of the values of ratio against TPR, further validates this conclusion. It shows that DRNApred is substantially better than the other methods because it achieves the lowest ratio over the entire range of TPR values. For instance, at TPR = 0.5 DRNApred's ratio = 0.21, which means that when correctly predicting 50% of the DNA binding residues 21% of the RNA binding residues were also predicted as DNA binding. To compare, the other methods obtain much higher ratios = 0.41, 0.36 and 0.30 for DP-Bind, DBS-PSSM and BindN+, respectively, at the same TPR = 0.5. Comparison of the ratio when the number of predicted binding residues equals to the number of native binding residues in the test dataset (see the top portion of the Supporting Table S2) similarly shows that our method significantly reduces the cross predictions. DRNApred obtains the lowest ratio value which is lower by (0.13 – 0.06)/0.13 = 54% compared to the second best BindN+.

We observed similar results for the prediction of RNA-binding residues (Figure 3B and the top portion of the Supporting Table S2 in the Supporting Materials). DRNApred offers equivalent predictive quality measured with AUC, AULC values when compared with the other predictors. Most importantly, our method produces significantly lower amounts of cross predictions that are quantified with AURC and AULRC values. The DRNApred's AURC and AULRC are substantially lower by (0.51 – 0.25)/0.51 = 51% and (0.121 – 0.029)/0.121 = 76%, respectively, compared to the second best Pprint methods. ROC curves show that RNABindR is better than other predictors when FPR is relatively high (Supporting Figure S3C in the Supporting Materials), but it is outperformed by DRNApred when FPR<4.5% (Supporting Figure S3D in the Supporting Materials), i.e. when the number of non-binding residues incorrectly predicted as binding (false positives) is lower than the number of native RNA binding residues. Ratio curve in Figure 3B further confirms the conclusion that our method significantly reduces the cross-predictions. This is true over the whole range of the TPR values. For example, at the TPR = 0.5 the ratios equal 0.20, 0.52, 0.54 and 0.72 for DRNApred, RNABindR, Pprint and BindN+, respectively. Comparison when setting all methods to generate the number of predicted binding residues equal to the number of native binding residues reveals that DRNApred provides slightly higher sensitivity and MCC and a much smaller ratio (see the top portion of the Supporting Table S2). The ratio of our predictor is lower by (0.1 – 0.02)/0.1 = 80% when compared to the second best Pprint method.

We also tested all predictors on the test dataset that does not include the transferred annotations. In other words, each protein is annotated using a single protein-DNA or protein-RNA complex. The results from DRNApred are slightly worse on this dataset compared with the dataset with the transferred annotations. This can be observed by comparing the corresponding top and bottom portions of the Supporting Table S2. The AULCs are 0.010 (dataset with transferred annotations) versus 0.008 (dataset without transferred annotations) for DNA binding and 0.005 versus 0.003 for RNA binding. Similarly, AULRCs are 0.039 versus 0.040 for DNA binding and 0.029 versus 0.032 for RNA binding; we note that higher value of AULRC indicates more cross-predictions. The assessment of binary predictions is also consistent with MCCs of 0.21 versus 0.20 and sensitivity of 0.25 versus 0.23 for DNA binding and MCCs of 0.12 versus 0.11 and sensitivity of 0.16 versus 0.14 for RNA binding. However, the same is true for the other predictors. For instance, we noted a slight but consistent across all predictors reduction in sensitivity and AULC values when using the dataset without the transfer of annotations. The decrease in sensitivity suggests that these methods successfully predict the transferred DNA and RNA binding residues. These predictions are at the rates that are the same or slightly higher than for the binding residues in the dataset without the transfer. This supports a hypothesis that transferred annotations share similar characteristics with the annotations in the dataset without the transfer. Importantly, DRNApred maintains significantly lower cross-prediction rates measured with ratio, AURC and AULRC values when compared to the other predictors on both types

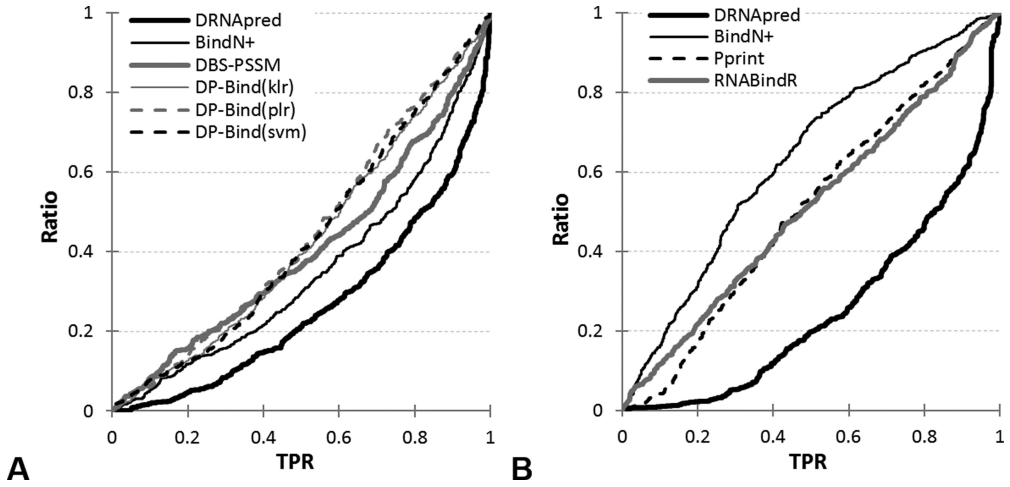

Figure 3. Comparison of the ratio curves for DRNApred and the considered predictors of the DNA and RNA binding residues on the test dataset. The ratio curve is the plot of Ratio against TPR values (fractions of correctly predicted binding residues). The curve that is closer to the *x*-axis for the same TPR corresponds to better predictions, i.e. lower amount of cross-predictions between RNA and DNA binding residues. Panel A is for the prediction of DNA-binding residues, and Panel B for the prediction of RNA-binding residues.

of nucleic acids binding on this test dataset (see the bottom portion of the Supporting Table S2). Its overall predictive performance measured with MCC, AUC and AULC is on par with the other predictors which is consistent with the results on the test dataset with the transferred annotations.

To sum up the results on both versions of the test dataset, DRNApred substantially reduces the cross predictions between DNA-binding and RNA-binding residues while maintaining similar overall predictive quality when compared to the existing methods. We analyzed this further in the next section. Moreover, our predictor correctly predicts the largest number of DNA-binding or RNA-binding residues when the number of predicted binding residues is reasonably low and not larger than the number of native binding residues.

We assessed the considered methods on the dataset composed on proteins that are unlikely to bind nucleic acids. We quantified the predictive quality with the FPR (fraction of non-binding residues predicted as binding) because there are no binding residues in this dataset. The results reveal that all methods obtain comparable and low FPR values that range between 2% and 5% (2–4%) for the prediction of DNA (RNA)-binding residues. Among the predictors of the DNA-binding residues, BindN+, DP-Bind(svm), DP-Bind(klr) and DBS-PSSM secure FPR = 3%, DP-Bind(plr) has FPR = 4%, and DRNApred obtains FPR = 5%. Considering the predictors of the RNA-binding residues, DRNApred and BindN+ generate predictions characterized by FPR = 2% while Pprint has FPR = 4%.

Finally, we assessed whether our predictive model and other predictors would stay current in a near future. We tested all methods using progressively newer depositions. Our test dataset includes proteins that were released between 2010 and 2015. We divided them into three similarly sized subsets with proteins deposited before 2012 (22 chains), in 2012 (31 chains) and after 2012 (29 chains). Supporting Figure S4 shows that none of the methods shows either decreasing or increasing trend of their predictive performance in correlation with the release dates. This observation applies to both the evaluation of the predictive accuracy with AULC (Supporting Figure S4A) and ratio of cross-predictions (Supporting Figure S4B). This suggest that DRNApred and other predictors should maintain similar levels of predictive performance in the near future.

Comparative assessment of predictive performance for the prediction of proteins that bind different types of RNAs

A few studies have shown that predictive performance for the prediction of RNA binding residues varies across different types of RNAs (57,58). Moreover, some of these RNAs are more similar to DNA, for instance in terms of the electrostatics of their interactions with proteins (57). This could affect the amount of their cross-predictions with the DNA binding residues. To study this, we evaluated predictions on the test dataset with the transferred annotations that includes only proteins that interact with a specific type of RNAs and the DNA binding proteins. We considered four categories of RNAs that had sufficient number of binding residues in the test dataset: mRNAs, dsRNAs, tRNAs and rRNAs. We found 8, 5, 4 and 2 proteins and 162, 91, 138, 66 residues that bind mRNAs, double stranded RNAs, tRNAs and rRNAs. Supporting Figure S5A shows the rate of cross predictions of the predictors of DNA binding residues for the four considered types of RNAs and compares it with the overall rate over all RNA-binding residues in our test dataset. We found that DRNApred offers the lowest ratio of the cross-predictions across the four types of RNAs. In other words, it predicts fewest RNA binding residues as DNA binding. Moreover, the ratios for the three versions of DP-Bind are consistently higher for the rRNAs and dsRNAs when compared to mRNAs and tRNAs. This observation agrees with the results in (57) where proteins that bind rRNAs and DNA were shown to have a more similar charge profile when compared to the similarity between DNA, tRNA and mRNA binding proteins. Supporting Figure S5B compares predictive quality measured with

AULC for the four predictors of RNA binding residues. DRNApred, RNABindR and BindN+ have similar AULC values for mRNAs, tRNAs and dsRNAs. Pprint provides higher predictive quality for tRNAs and dsRNAs while DRNApred is substantially better for rRNAs. The highest AULC values across most methods are for prediction of rRNAs and the lowest for mRNAs. This is in agreement with (57) where prediction of rRNA binding and mRNA binding from protein structures was shown to have highest and lowest quality, respectively. Supporting Figure S5C focuses on the AULRC values that quantify how the ratio of cross-predictions changes across different rates of correct predictions of RNA binding residues (TPR values). Given that lower AULRC values indicate lower cross-prediction rates, DRNApred is shown to predict the fewest DNA binding residues as RNA binding when predicting each of the corresponding types of RNAs. In the nutshell, this analysis shows that overall predictive performance of DRNApred is on par with the other predictors of RNA binding residues across the four types of RNAs while it provides a substantial reduction in the rate of cross-predictions.

Comparative assessment of predictive performance for the prediction of proteins that bind both DNA and RNA, and RNA/DNA hybrids

We tested whether the reduced amount of cross-predictions affects DRNApred's ability to predict residues that bind both RNA and DNA, and binding residues in proteins that interact with RNA/DNA hybrids. We compared the predictions from DRNApred with the predictions generated by combining the current predictors of DNA and RNA binding residues. We selected DP-bind(svm) for the prediction of DNA binding residues and Pprint for the prediction of RNA binding residues because these two methods have secured the highest AULC values (Supporting Table S2).

To evaluate prediction of residues that bind both RNA and DNA, we developed a dataset of proteins that bind RNA and DNA by following the same protocol as we used for the test dataset. We collected 90 proteins chains from PDB that are in complex with both types of nucleic acids. Next, we transferred annotations of DNA and RNA binding residues between the proteins that are in the same clusters defined by high sequence similarity ($\geq$80%) and high structure similarity ($\geq$0.5 TM scores). Each of the eight resulting clusters was represented by a chain that has the largest number of binding residues. Next, we eliminated proteins that are over 1000 residues long that could not be predicted with DP-bind and proteins that do not have residues that bind both RNA and DNA. Supporting Table S3 summarizes assessment of the predictions of residues that bind RNA and DNA for the remaining four proteins (PDB IDs: 1ZBI, 1MSW, 2QKB, 4H8K). The results reveal that DRNApred secures stronger predictive performance than the current methods. It has significantly higher AULC values which means that is correctly predicts more residues that bind RNA and DNA at low values of FPR when compared with the DP-bind(svm)+Pprint method. DRNApred also has significantly lower value of the area under the ratio curve, AULRC. This reveals that our predictor makes fewer incorrect predictions such that the native residues that bind either DNA or RNA are predicted as residues that bind both RNA and DNA. We also assessed binary predictions, i.e. a given residue binds RNA and RNA versus it does not bind RNA and DNA. The same as on the main test dataset, the threshold to obtain binary predictions from the putative propensities is set to ensure that the number of predicted DNA and RNA binding residues equals to the number of native DNA and RNA binding residues. The DRNApred's sensitivity = 12% and its ratio which measures the fraction of residues that bind either RNA or DNA that were predicted as binding both RNA and DNA is 5%. These values are better than the sensitivity = 4% and ratio = 6% obtained with the DP-bind(svm)+Pprint approach (Supporting Table S3). Overall, the empirical results demonstrate that DRNApred relatively accurately predicts residues that bind both types of nucleic acids.

To assess predictions for proteins that interact with RNA/DNA hybrids, we collected 21 proteins chains from 17 complex with the hybrids that were available in PDB. Next, we clustered them into groups of chains that share high sequence similarity ($\geq$80%) and high structure similarity ($\geq$0.5 TM scores) to remove redundancy and to transfer annotations of DNA and RNA binding residues. Among the eight resulting clusters, six did not have DNA/RNA binding residues based on our definition of binding. The remaining two proteins (PDB IDs: 2HVR and 2Q2T) include 11 RNA-binding and 52 DNA-binding residues, which we annotated based on the corresponding base. Supporting Table S4 compares the results from DRNApred and two methods that secured the highest AULC values on the test dataset: DP-bind(svm) for the prediction of DNA-binding residues and Pprint for the prediction of RNA-binding residues. The results demonstrate that DRNApred obtains lower values of AURC and AULRC for the predictions of both RNA and DNA binding residues. This means that it cross-predicts fewer residues than the other methods. DRNApred and the other methods have comparable values of AULC and AUC. We emphasize similarity in the AULC values which quantify ability to correctly identify RNA and DNA binding residues for low values of FPR. We also evaluated binary predictions that were defined in the same way as on the main test dataset. Namely, we set the cut-offs to define binary predictions from the putative propensities to ensure that the number of predicted DNA (RNA) binding residues equals to the number of native DNA (RNA) binding residues. For the prediction of DNA binding residues, the DRNApred's sensitivity and MCC are slightly lower than DP-bind(svm)'s values, but the ratio of our predictor is better (Supporting Table S4). For the prediction of RNA binding residues, DRNApred boasts better values of sensitivity, MCC and ratio when compared with Pprint.

To sum up, the results obtained for proteins that bind both RNA and DNA and for proteins that interact with DNA/RNA hybrids agree with the evaluation on the test dataset. Altogether they suggest that DRNApred accurately predicts DNA binding residues and RNA binding residues while reducing cross-predictions between these two nucleic acids.

Analysis of the predicted binding residues

We observed that the native RNA and DNA binding residues tend to cluster together in the protein sequences. This is because close proximity in the sequence usually implies proximity in the corresponding structure. This in turn is relevant since regions on the protein surface that interact with the nucleic acids tend to be relatively large due to the large size of the RNA and DNA molecules. Moreover, the annotations of the binding residues suffer inaccuracies given how they are defined. The use of a distance between atoms in protein and nucleic acids results in a somehow arbitrary inclusion or exclusion of binding residues that are close to the cut-off value that is used to define binding. This means that some of the non-binding residues adjacent to the annotated binding residues could be in fact involved in binding. Altogether, these observation points to a conjecture that residues that are in close proximity in the sequence to the annotated binding residues are more likely to in fact bind DNA/RNA compared to residues that are far away. In other words, false positives localized close to the native binding residues are more desirable (more likely to be true positives) compared to the false positives that are farther away from the binding residues.

We investigated binding residues predicted by different methods to compare how close they are from the native binding residues. We quantified the distance either as a number of positions in the sequences or the Euclidian distance measured between alpha-carbon atoms of the corresponding residues in the protein structures. For each method, we counted the number of correctly predicted binding residues, i.e. the residues with distance = 0 from the native binding residues. We also counted the incorrectly predicted binding residues that are a specific number of residues or a specific distance in structure measured in Å away from the nearest native binding residue. The corresponding fractions of these predicted binding residues out of the total number of the predicted binding residues are plotted in Figure 4. We argue that DRNApred predicts higher quality false positives compared to the other considered methods since its predictions are located closer to the native binding residues. This is true for the prediction of both DNA and RNA binding residues irrespective of how the distance is measured. Our empirical results reveal that 46% (Figure 4A) and 51% (Figure 4C) of residues predicted by DRNApred as RNA and DNA binding, respectively, are just up to five positions in the sequence away from the nearest native binding residues. To compare, the second best method generates 33% (Figure 4A) and 44% (Figure 4C) of its predicted RNA and DNA binding residues at that distance from the nearest native binding residues. Similarly, 41% (Figure 4B) and 48% (Figure 4D) of residues predicted by DRNApred to bind DNA and RNA, respectively, are no farther than 8 Å from a native binding residue. The corresponding fractions for the second best method are substantially lower and equal 32% (Figure 4B) and 41% (Figure 4D).

The observation that DRNApred correctly predicts more binding residues at the corresponding distance is consistent with its higher MCC and sensitivity values (Supporting Table S2). Moreover, as the distance increases the DRNApred's curve saturates faster and reaches a much higher value compared to the curves from the other methods. This means that our model cross predicts much fewer residues than the other methods. The fraction of the putative binding residues predicted in the incorrect type of binding proteins can be read from the gap between the value of 1 and the value of the fraction at the far end of a given curve. Specifically, DRNApred mis-predicts 20% of DNA-binding residues in the RNA-binding proteins (Figure 4C and D) and 18% RNA-binding residues in the DNA-binding proteins (Figure 4A and B). To compare, the corresponding values are 35% (Figure 4C and D) and 44% Figure 4A and B) for the second best BindN+ and Pprint methods, respectively. Overall, DRNApred correctly finds more binding residues and captures more putative binding residues that are likely to bind to DNA (RNA) although they lack such annotation in the test dataset. Importantly, our model generates substantially fewer strong mis-predictions that are defined as the putative RNA binding residues identified in the DNA binding proteins and the putative DNA binding residues found in the RNA binding proteins.

We also evaluated how predictive quality measured with MCC and TPR would change if the predicted binding residues which are 0, $\leq$1, $\leq$2 and $\leq$3 residues away in the sequence from the nearest native binding residue would be considered as correctly predicted (Supporting Figure S6). We have done this analysis using the distance in the sequence because these predictions are performed in the sequence without the knowledge of the protein structure. We argue that the corresponding false positives that we reconsider as true positives could be in fact interacting with the nucleic acids or be useful to identify the nearby binding residues. As expected, both MCC and TPR for all considered methods improve as we included additional true positives. Interestingly, inclusion of just the adjacent positions (distance = 1 on the x-axis) results in a substantial increase in DRNApred's TPR by ~8% for both DNA binding (from 25% to 33%) and RNA binding (from 16% to 24%) compared to when only the native binding residues are considered. DRNApred's MCC also registers a large increase from 0.21 to 0.31 for the DNA binding and from 0.12 to 0.22 for RNA binding. At distance = 3, our method achieves the TPR = 0.38 (0.31) and MCC = 0.39 (0.31) for the prediction of the DNA (RNA)-binding residues. Moreover, DRNApred secures the largest increases in both MCC and TPR when compared to the other methods. This again demonstrates that our predictor is better at finding desirable, high quality false positives that could be in fact relevant to the nucleic acid binding.

Comparative assessment of predictive performance for the prediction of the DNA/RNA-binding proteins

We tested performance of DRNApred and the other predictors of the DNA and RNA-binding residues for the predictions of the DNA and RNA-binding proteins on the test dataset. A protein is annotated as binding to DNA (RNA) if at least one residue in this protein is annotated as binding to DNA (RNA). A protein is assumed to be predicted as binding to DNA (RNA) if the number of predicted DNA (RNA)-binding residues in this protein is larger than

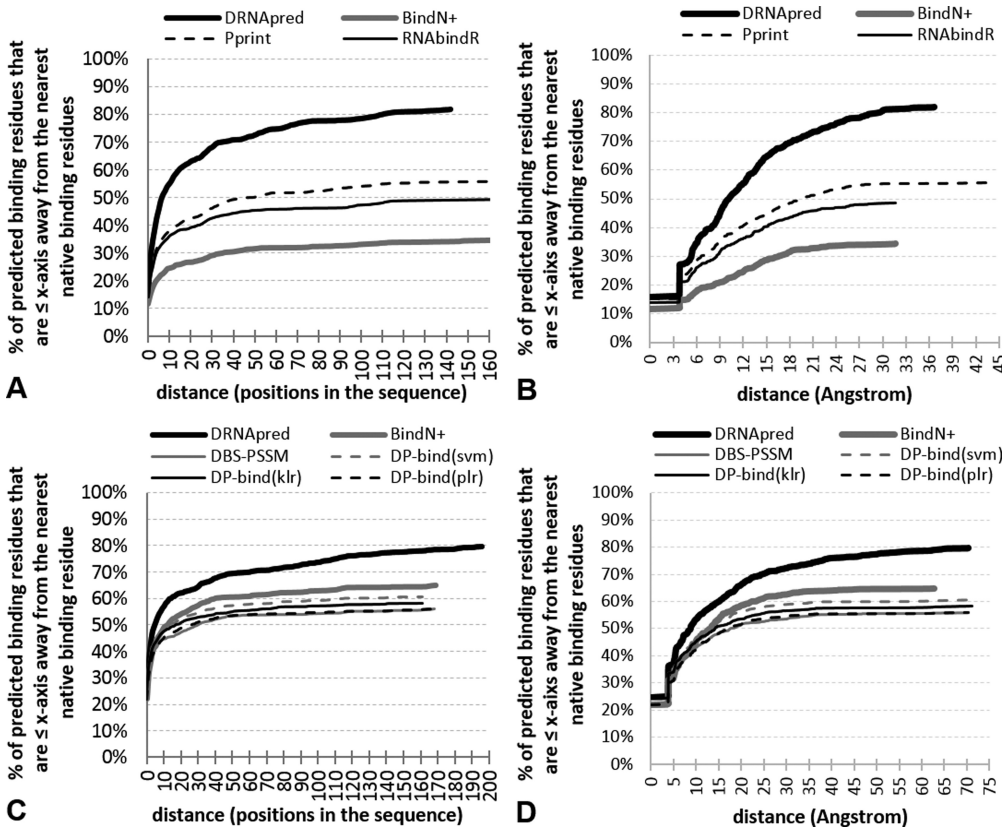

Figure 4. Summary of the distance between the predicted binding residues and the nearest native binding residues. The distance is measures either as the number of positions in the sequence (panels **A** and **C**) or the Euclidian distance between alpha-carbon atoms of the two residues in the protein structure (panels **B** and **D**). Panels A and B summarizes results for the prediction of the DNA-binding residues and panels C and D for the prediction of the RNA-binding residues. The summary is quantified with fractions of putative binding residues that are ≤ distance shown on the x-axis away from the nearest native binding residue. The fraction is defined as the count of residues up to a given distance away divided by the total number of the putative binding residues. The curves do not reach the fraction of 1 because the remaining residues are predicted in proteins that do not have the corresponding native binding residues (the distance to the nearest native binding residue is undefined). These are putative RNA binding residues that are predicted in the DNA binding proteins and vice versa.

a small threshold. This is to accommodate for the predicted false positives. The threshold is set so the FPR of a given method on the test set equals 5%. DRNApred outperforms the other methods by a wide margin (Supporting Table S5). DRNApred's MCC is statistically significantly better than MCCs of the other methods. The TPR of our predictor is 5 and 6 times higher than the corresponding FPR for the DNA and RNA binding, respectively, and is also much higher than the TPR values of the other predictors.

Besides evaluating predictions at the low FPR, we varied the thresholds (the minimal number of the predicted binding residues that corresponds to prediction of a binding protein) using the complete range. We plotted relation between the corresponding TPR and FPR values (ROC curve) in Figure 5. The plot shows that DRNApred improves over the other methods for small and modest values of FPR. Predictions when TPR values are high are arguably less interesting since they would lead to a substantial overprediction of the DNA or RNA binding proteins. By tuning the threshold, DRNApred achieves maximal MCC = 0.31 and 0.36 for the prediction of the DNA and RNA-binding proteins, respectively, compared to the second best method DP-Bind(svm) with MCC = 0.23 and RNABindR with MCC = 0.28. The main reason why the other methods offer lower predictive quality is that they cross-predict between DNA and RNA binding residues. In other words, their correct predictions of DNA binding proteins are coupled with the incorrect predictions of RNA binding proteins as DNA binding, resulting in high FPRs and low AUC and MCC values.

Comparative evaluation of runtime

Runtime is a key factor that determines whether a given predictor can be applied in a high-throughput manner to annotate a large collection of proteins. The considered predictors, except for DRNApred, utilize evolutionary profile derived with PSI-BLAST as one of their inputs. The calculation of the profile is the main computational cost of these methods. We approximated a lower bound of their runtime by the time to run PSI-BLAST. Based on the database and the number of iterations that each of these methods used to run PSI-BLAST, we divided them into two groups. The first group includes DP-Bind, DBS-PSSM, Pprint and RNABindR that use the *nr* database with at least three iterations of PSI-BLAST (referred to as *PSIBlast on nr*). The second group includes BindN+ that uses much smaller UniProt database with three iterations of PSI-BLAST. Fig-

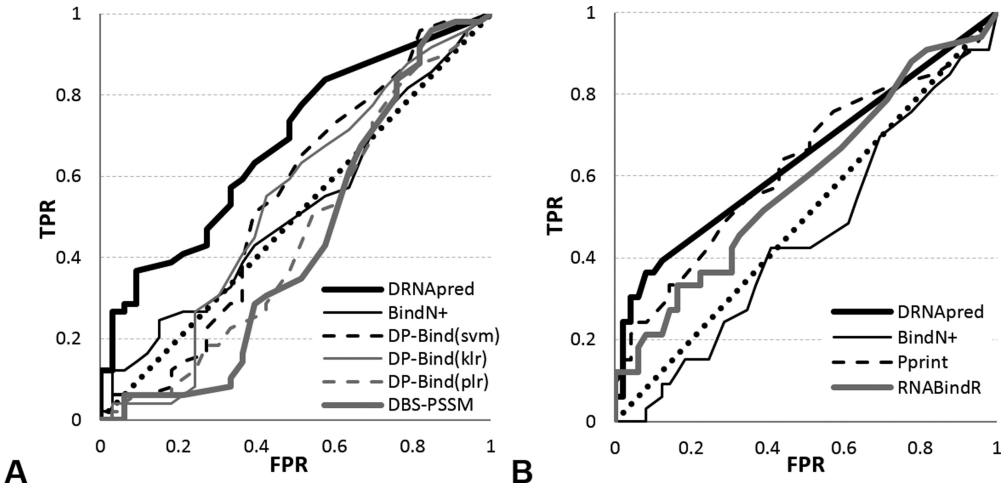

Figure 5. Comparison of ROCs for DRNApred and the other predictors for the prediction of DNA and RNA-binding proteins on the test dataset. Panel **A** is for the prediction of the DNA-binding proteins and Panel **B** is for the prediction of the RNA-binding proteins. The dotted black diagonal line represents a random prediction.

ure 6 compares the runtime of DRNApred and the other methods based on predictions on the test dataset using the same hardware (i7-CPU and 23GB RAM). Although the absolute value of the runtime depends on computer hardware used, we focused on relative differences which are hardware independent. DRNApred is at least 3 orders of magnitude faster than the other methods that utilize PSI_BLAST against the *nr* database. DRNApred's runtime is comparable to the runtime of BindN+. Both methods predict an average size protein in ∼15 s using a modern desktop computer.

We interpolated the measured runtime using second degree polynomials that provide a relatively accurate fit (see lines in Figure 6). We used these polynomials to estimate and compare the total runtime to predict the complete human proteome, the largest proteome with ∼70 thousand proteins. DRNApred and BindN+ are estimated to take about 48 and 21 days, respectively, using a single i7-CPU. The other methods will take substantially more time, 1475 days that correspond to over 4 years. The actual DRNApred's runtime on the human proteome was 55 days, which is close to the estimate. However, we performed calculations using eight processors in parallel, each dedicated to a different subset of proteins, which reduced the runtime to 7 days. These results suggest that DRNApred is sufficiently fast to perform genome-wide predictions using a desktop computer.

In the nutshell, the runtime of DRNApred is relatively low and comparable to the fastest current method. Thus, besides offering substantially better predictive performance DRNApred can be used to perform large scale predictions.

Assessment of predictive performance on the known DNA and RNA binding proteins in the human proteome

We applied DRNApred and BindN+ (the other runtime-efficient method) to predict the RNA and DNA binding residues and the RNA and DNA binding proteins in the human proteome. We used the predicted DNA and RNA binding residues generated by both methods to define putative DNA and RNA binding proteins, respectively. The predicted binding proteins have the number of the corresponding predicted binding residues higher than a threshold = 5%. This cut-off corresponds to the FPR of the prediction of the binding proteins on the test dataset. This is to accommodate for spurious predictions that are associated with the false positive predictions inherent in the outputs of these predictive models. We assessed the predictive performance by measuring whether these methods specifically predict only the target type of binding proteins/residues among the known binding proteins in the human proteome. In other words, we evaluated whether their predictions of the DNA-binding residues target primarily the known DNA-binding proteins and how many of these predictions are for the known RNA-binding proteins, and vice versa. We also assessed whether novel binding proteins predicted with DRNApred are likely to be correctly predicted. These are predicted binding proteins that not overlap with the native binding proteins. Technical details of this assessment are explained in the Supporting Materials.

Assessment of predictive performance

We quantified the amount of cross predictions of the DNA and RNA binding proteins among the known human DNA-binding and RNA-binding proteins. In particular, we calculated and compared the ratio of the fraction of the correctly predicted known binding proteins to incorrectly cross predicted known binding proteins. The results are shown using black bars in Supporting Figure S7. Both DRNApred and BindN+ generate better than random results for the prediction of the DNA-binding proteins, i.e. their ratio values > 1. BindN+ obtains ratio = 1.3, which means that it predicts 1.3 times higher fraction of DNA-binding proteins in the known DNA-binding proteins than in the known RNA-binding proteins. DRNApred outperforms BindN+ by securing ratio = 1.9, which corresponds to (1.9 – 1.3)/1.3 = 46% improvement. Moreover, BindN+ secures ratio ≈ 1 for the prediction of the RNA-binding

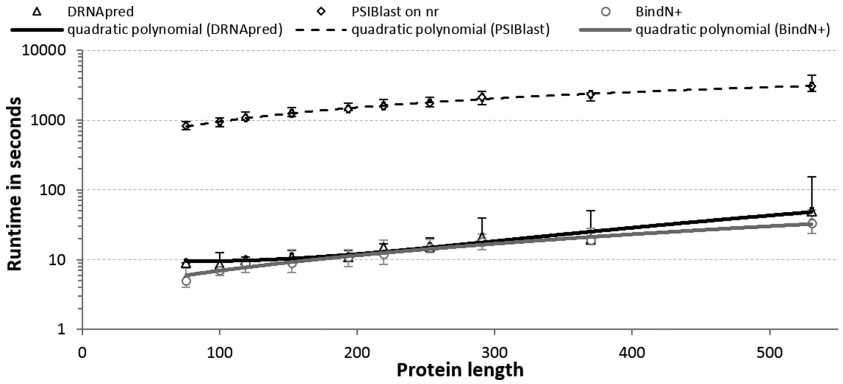

Figure 6. Comparison of runtime in the function of protein length for DRNApred and the other predictors of the DNA and RNA binding residues on the test dataset. The *y*-axis is the runtime in seconds shown using base 10 logarithmic scale. The *x*-axis is the protein length. We sorted proteins by their sequence length and divided them into 10 equally sized sets that include proteins with increasing size. The plot reports the median runtime (markers) and the 25th and 75th centiles (error bars) against the median protein length for each of the 10 protein sets. The measurements were made using a modern desktop computer with i7-CPU and 23GB RAM. Lines show quadratic polynomial fit into the measured data.

proteins. This reveals that this method substantially cross-predicts the DNA-binding proteins as RNA binding. The predictions of the RNA-binding proteins by DRNApred are substantially better, with the ratio = 3.1, an improvement by 310%. This means that DRNApred predicts three times more correct RNA binding proteins compared to the incorrectly cross predicted DNA binding proteins. Overall, these results demonstrate that DRNApred provides specific predictions of the DNA binding and the RNA binding proteins.

We also assessed the predictive quality of DRNApred and BindN+ by comparing their cross predictions of the predicted binding residues in the sets of known DNA-binding and RNA-binding proteins. We calculated the ratio of the fraction of the predicted binding residues among the correct type of known binding proteins to the fraction of the cross predicted putative binding residues in the other type of known binding proteins. The results are shown using grey bars in Supporting Figure S7. DRNApred achieves ratio of 2.1 for the prediction of the DNA-binding residues. This means that it predicts over two times higher fraction of DNA-binding residues in the known DNA-binding proteins than in the known RNA-binding proteins. To compare, BindN+ obtains ratio = 1.3, which suggests that it cross predicts a more substantial number of DNA binding residues. DRNApred also outperforms BindN+ when considering the prediction of RNA-binding residues. BindN+ obtains a ratio at ∼1 indicating that it predicts similar fraction of RNA binding residues in both known DNA-binding and RNA-binding proteins. DRNApred secures a high ratio = 6.8. The observation that DRNApred accurately and specifically predicts each type of the nucleic acid binding on the human proteome is consistent with the conclusions that we have reached on the test dataset.

Evaluation of novel putative RNA and DNA binding proteins

We investigated the degree of an overlap in subcellular locations between the putative novel binding proteins and the known binding proteins. We annotated the locations based on the gene ontology cellular component (GO-CC) terms. We created a list of the GO-CC terms that are substantially enriched in the known binding proteins, by at least three folds, when compared to their abundance in the whole proteome. These terms are significantly associated with the DNA binding and the RNA binding proteins. Next, we calculated the fraction of these terms that are substantially enriched, by at least 100%, in the novel binding proteins. A high fraction indicates that both known and novel putative RNA and DNA binding proteins share similar subcellular locations. Results are shown in Figure 7. The *x*-axis shows the minimal level of enrichments of the considered GO-CC terms in the known binding proteins. The numbers of these terms, which are shown above the bars, are fairly high indicating that they can be used to pinpoint the subcellular locations of the native binders. As the required enrichment of the GO-CC in the known binders grows from at least 3- to 9-fold so does the fraction of these terms that are also significantly enriched in the novel putative binders. These fractions start at 64% and 86% for the DNA and RNA binding proteins, respectively, when considering the over 100 terms that are enriched by at least 3-fold in the native binders. Given that we use 42 and 95 terms that are enriched by at least 9-fold in the DNA and RNA binding proteins, respectively, 100% and 93% of them are also enriched in the novel putative binders. This revels that virtually all of the subcellular locations that are significantly associated with the native RNA and DNA binding proteins are also significantly enriched in the novel RNA and DNA binding proteins that were predicted by DRNApred. In other words, the locations of the putative and native RNA and DNA proteins are in agreement, suggesting that the novel binding proteins are possibly predicted correctly.

We analyzed whether the predicted binding residues in the novel binding proteins are similar to the binding residues in the native binding proteins. Since one of the hallmarks of the DNA and RNA binding is inclusion of charged residues, we compared the fractions of the positively charged residues among the predicted binding and nonbinding residues in these proteins with the fractions in the known binding proteins and in the whole proteome. Results are summarized in Figure 8. Overall, about 11%

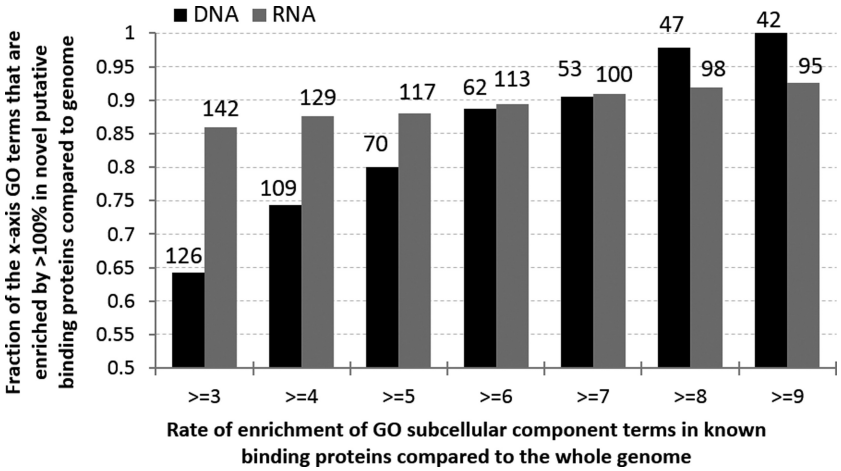

Figure 7. Fraction of the gene ontology cellular component (GO-CC) terms associated with the known binding proteins that are also enriched by at least 100% in novel putative binding proteins. The enrichment in the GO-CC terms is computed against their abundance in the proteome. The *x*-axis shows the minimal level of enrichments of the GO-CC terms in the known binding proteins, the corresponding numbers of significantly enriched terms are shown above the bars. Grey (black) bars summarize results for the RNA (DNA) binding proteins.

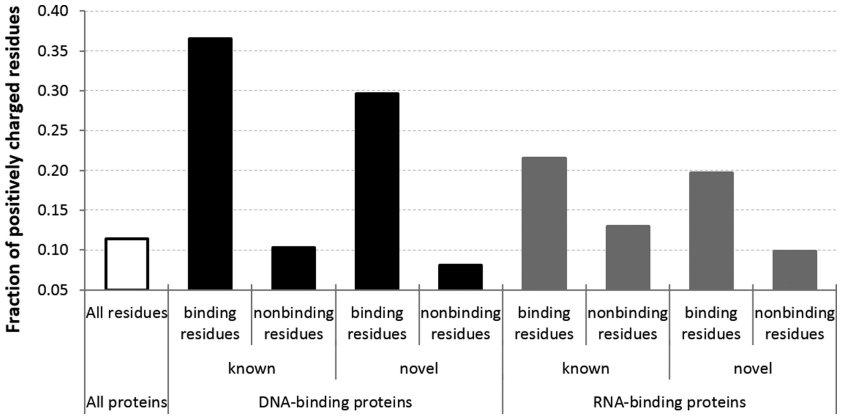

Figure 8. Fraction of the positively charged residues among the binding and nonbinding residues in the known and novel binding proteins and among the residues in the entire human proteome. Grey (black) bars summarize results for the RNA (DNA) binding proteins. The hollow bar shows the results for the human proteome.

of residues in the human proteome are positively charged. There are 3.2 and 2.6 times (1.9 and 1.7 times) more positively charged residues among the predicted DNA-binding residues (RNA-binding residues) in the known and novel putative DNA-binding proteins (RNA-binding proteins), respectively, when compared to the human proteome. While this is expected for the native binders, the similar levels of the enrichment in the putative novel binders suggest that they are likely correctly identified by DRNApred. Moreover, the fraction of the positively charged residues among the putative nonbinding residues in both known and putative DNA and RNA binding proteins is similar to the level of the positively charged residues in the proteome. The differences in the levels of the positively charged residues between the putative binding and nonbinding residues support our claim that the putative binding residues generated by DRNApred are likely to bind the two nucleic acids. This observation is consistent for both native and novel putative DNA and RNA binding proteins.

CONCLUSIONS

Many methods for accurate prediction of the DNA- and RNA-binding residues from the protein sequence have been published. However, we and others have shown that these methods cross-predict a substantial number of the nucleic acid binding residues (DNA-binding residues are predicted as RNA-binding as vice versa). Most of these methods also require a relatively high amount of runtime. We introduced a new predictor, DRNApred, which accurately discriminates between DNA-binding and RNA-binding residues and proteins. DRNApred requires a low amount of runtime. It predicts an average sized protein with 200 residues in 15 s on a modern desktop computer, and thus it can be applied on a whole proteome scale.

We designed DRNApred by considering a comprehensive set of features extracted from a diverse set of sources of sequence-derived information extracted from a dataset with both DNA-binding and RNA-binding proteins. This information includes amino acid types, physicochemical properties of amino acids, evolutionary profiles, and putative

intrinsic disorder, secondary structure, solvent accessibility. We performed empirical selection of a subset of predictive and non-redundant features from this large set of considered features. We also implemented a weight-based mechanism and incorporated the second predictive layer to reduce the cross-predictions. We empirically demonstrated that these novel design strategies substantially reduce the amount of cross predictions. Moreover, we comparatively tested DRNApred on the test dataset for the prediction of the DNA and RNA-binding residues and proteins. We showed that DRNApred substantially reduces the cross predictions when compared to several existing methods. We empirically demonstrated that in spite of the reduced cross-prediction our predictive model relatively accurately finds residues that bind both RNA and DNA as well as residues that interact with the RNA/DNA hybrids. Our empirical analysis also includes assessment of the predictive performance on different types of RNAs. Importantly, our empirical analysis revealed that our predictor finds arguably higher quality false positives that are located nearby the native binding residues. It predicts substantially fewer DNA binding residues in the RNA binding proteins and vice versa when compared with the considered current predictors. Furthermore, we compared predictive performance for the prediction of the DNA-binding and RNA-binding proteins. We showed that DRNApred secures the highest AUCs and outperforms the other methods by correctly predicting more DNA- and RNA-binding proteins at the same false positive rate. Our empirical tests also demonstrated that DRNApred is computationally efficient. It is at least 3 orders of magnitude faster than majority of the other methods, excluding BindN+. We showed that DRNApred and BindN+ have similar runtime profiles, which means that these two methods can be used to perform genome-wide predictions on a desktop computer. However, our tests indicated that DRNApred provides better predictive performance and the lowest levels of cross predictions.

A substantial number of the DNA and RNA binding proteins are yet to be discovered in the human proteome. We applied our runtime-efficient DRNApred method to perform large-scale prediction and assessment of the DNA and RNA binding proteins and binding residues in the human proteome. We compared predictive quality between DRNApred and BindN+, in particular focusing on the cross prediction between RNA and DNA binding among the known binders. We showed that DRNApred substantially reduces the cross predictions at both residue and protein levels when compared to BindN+. We also analyzed whether the putative novel binding proteins generated by DRNApred possess certain hallmarks of the native binding proteins. We showed that subcellular locations and content of positively charged residues among their binding residues are similar between novel and native binders. This provides support to the claim that DRNApred can be used to discover novel DNA and RNA binding proteins in human.

The DRNApred's webserver is freely available at http://biomine.cs.vcu.edu/servers/DRNApred/. It accepts queries that consists of up to 100 protein FASTA-formatted protein sequences and provides predictions of both RNA and DNA binding residues. Results are stored in a parsable text file that is archived on the server for at least 1 month. User can download the file from URL address that is provided in the browser window upon completion of the prediction. Results are also send to a user-provided email address. The webpage also includes Supporting Materials, training and test datasets, and the datasets with the native DNA- and RNA-binding proteins in human.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Chen Wang for his contribution to setup the webserver and to perform assessment based on the protein structures.

FUNDING

Natural Sciences and Engineering Research Council of Canada [Discovery grant no. 298328] (in part); Qimonda Research Chair (to L.K.). Funding for open access charge: Qimonda Research Chair (to L.K.).
Conflict of interest statement. None declared.

REFERENCES

1. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, doi:10.1186/gb-2000-1-1-reviews001.
2. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
3. Re,A., Joshi,T., Kulberkyte,E., Morris,Q. and Workman,C.T. (2014) RNA–protein interactions: an overview. *RNA Sequence Struct. Funct.: Comput. Bioinformatic Methods*, 491–521.
4. Noller,H.F. (2005) RNA structure: reading the ribosome. *Science*, **309**, 1508–1514.
5. Glisovic,T., Bachorik,J.L., Yong,J. and Dreyfuss,G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Berglund,A.-C., Sjölund,E., Östlund,G. and Sonnhammer,E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36**, D263–D266.
8. Zhang,H.-M., Chen,H., Liu,W., Liu,H., Gong,J., Wang,H. and Guo,A.-Y. (2012) AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.*, **40**, D144–D149.
9. Zhang,H.-M., Liu,T., Liu,C.-J., Song,S., Zhang,X., Liu,W., Jia,H., Xue,Y. and Guo,A.-Y. (2015) AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.*, **43**, D76–D81.
10. Consortium,U. (2014) Activities at the universal protein resource (UniProt). *Nucleic Acids Res.*, **42**, 7486.
11. Baltz,A.G., Munschauer,M., Schwanhäusser,B., Vasile,A., Murakawa,Y., Schueler,M., Youngs,N., Penfold-Brown,D., Drew,K. and Milek,M. (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.
12. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M. and Wei,G. (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
13. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.

14. Ahmad,S., Gromiha,M.M. and Sarai,A. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.
15. Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33.
16. Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.
17. Yan,C., Terribilini,M., Wu,F., Jernigan,R.L., Dobbs,D. and Honavar,V. (2006) Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, **7**, 1.
18. Lee,J.-h., Hamilton,M., Gleeson,C., Caragea,C., Zaback,P., Sander,J.D., Li,X., Wu,F., Terribilini,M. and Honavar,V. (2008) *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. NIH Public Access, p. 501.
19. Hwang,S., Gou,Z. and Kuznetsov,I.B. (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, **23**, 634–636.
20. Kuznetsov,I.B., Gou,Z., Li,R. and Hwang,S. (2006) Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins: Struct. Funct. Bioinformatics*, **64**, 19–27.
21. Ofran,Y., Mysore,V. and Rost,B. (2007) Prediction of DNA-binding residues from sequence. *Bioinformatics*, **23**, i347–i353.
22. Wang,L., Huang,C., Yang,M.Q. and Yang,J.Y. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, **4**, 1.
23. Carson,M.B., Langlois,R. and Lu,H. (2010) NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res.*, **38**, W431–W435.
24. Si,J., Zhang,Z., Lin,B., Schroeder,M. and Huang,B. (2011) MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. *BMC Syst. Biol.*, **5**, S7.
25. Terribilini,M., Lee,J.-H., Yan,C., Jernigan,R.L., Honavar,V. and Dobbs,D. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.
26. Terribilini,M., Sander,J.D., Lee,J.-H., Zaback,P., Jernigan,R.L., Honavar,V. and Dobbs,D. (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.
27. Walia,R.R., Caragea,C., Lewis,B.A., Towfic,F., Terribilini,M., El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics*, **13**, 1.
28. Kumar,M., Gromiha,M.M. and Raghava,G. (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins: Struct. Funct. Bioinformatics*, **71**, 189–194.
29. Cheng,C.-W., Su,E.C., Hwang,J.-K., Sung,T.-Y. and Hsu,W.-L. (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, **9**, S6.
30. Spriggs,R.V., Murakami,Y., Nakamura,H. and Jones,S. (2009) Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics*, **25**, 1492–1497.
31. Murakami,Y., Spriggs,R.V., Nakamura,H. and Jones,S. (2010) PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res.*, **38**, W412–W416.
32. Zhao,H., Yang,Y. and Zhou,Y. (2011) Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol.*, **8**, 988–996.
33. Puton,T., Kozlowski,L., Tuszynska,I., Rother,K. and Bujnicki,J.M. (2012) Computational methods for prediction of protein–RNA interactions. *J. Struct. Biol.*, **179**, 261–268.
34. Miao,Z. and Westhof,E. (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res.*, **43**, 5340–5351.
35. Miao,Z. and Westhof,E. (2016) RBscore&NBench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database. *Nucleic Acids Res.*
36. 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.
37. Miao,Z. and Westhof,E. (2015) A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput. Biol.*, **11**, e1004639.
38. Nagarajan,R., Ahmad,S. and Gromiha,M.M. (2013) Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res.*, **41**, 7606–7614.
39. Puton,T., Kozlowski,L., Tuszynska,I., Rother,K. and Bujnicki,J.M. (2012) Computational methods for prediction of protein-RNA interactions. *J. Struct. Biol.*, **179**, 261–268.
40. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, **32**, W20–W25.
41. Peng,Z. and Kurgan,L. (2015) High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res.*, **43**, e121.
42. Zhao,H., Yang,Y. and Zhou,Y. (2013) Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol. bioSystems*, **9**, 2417–2425.
43. Tong,J., Jiang,P. and Lu,Z.-h. (2008) RISP: a web-based server for prediction of RNA-binding sites in proteins. *Comput. Methods Programs Biomed.*, **90**, 148–153.
44. Ma,X., Guo,J., Liu,H.-D., Xie,J.-M. and Sun,X. (2012) Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *Computat. Biol. Bioinformatics, IEEE/ACM Trans.*, **9**, 1766–1775.
45. Anderson,T.W. and Darling,D.A. (1952) Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *Ann. Math. Stat.*, 193–212.
46. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
47. Walsh,I., Martin,A.J., Di Domenico,T. and Tosatto,S.C. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
48. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
49. Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Bioinformatics*, **20**, 216–226.
50. Ahmad,S. and Gromiha,M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, **18**, 819–824.
51. Ahmad,S., Gromiha,M.M. and Sarai,A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**, 1849–1851.
52. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.
53. Remmert,M., Biegert,A., Hauser,A. and Söding,J. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
54. Disfani,F.M., Hsu,W.L., Mizianty,M.J., Oldfield,C.J., Xue,B., Dunker,A.K., Uversky,V.N. and Kurgan,L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, **28**, i75–i83.
55. Mizianty,M.J. and Kurgan,L. (2011) Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics*, **27**, i24–i33.
56. Maheshwari,S. and Brylinski,M. (2015) Predicting protein interface residues using easily accessible on-line resources. *Brief. Bioinform.*, **16**, 1025–1034.
57. Ahmad,S. and Sarai,A. (2011) Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. *BMC Struct. Biol.*, **11**, 8.
58. Fernandez,M., Kumagai,Y., Standley,D.M., Sarai,A., Mizuguchi,K. and Ahmad,S. (2011) Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics*, **12**(Suppl. 13), S5.