Consensus-based prediction of RNA and DNA binding residues from protein sequences

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

[1]Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2V4 Canada
*To whom correspondence should be addressed. Email: lkurgan@ece.ualberta.ca

Abstract. Computational prediction of RNA- and DNA-binding residues from protein sequences offers a high-throughput and accurate solution to functionally annotate the avalanche of the protein sequence data. Although many predictors exist, the efforts to improve predictive performance with the use of consensus methods are so far limited. We explore and empirically compare a comprehensive set of different designs of consensuses, including simple approaches that combine binary predictions and more sophisticated machine learning models. We consider both DNA- and RNA-binding motivated by similarities in these interactions, which should lead to similar conclusions. We observe that the simple consensuses do not provide improved predictive performance when applied to sequences that share low similarity with the datasets used to build their input predictors. However, use of machine learning models, such as linear regression, Support Vector Machine and Naïve Bayes, results in the improved predictive performance when compared with the best individual predictors for the prediction of DNA- and RNA-binding residues.

1 Introduction

Interactions between proteins and DNA/RNA are at the heart of numerous cellular functions including regulation of gene expression, genome maintenance, recombination, replication and transcription, to name a few [1, 2]. The DNA-binding and RNA-binding proteins occupy a relatively large fraction of eukaryotic genomes, in the order of 3 to 5% [3] and 2 to 8% [1], respectively. However, only a small fraction of these interactions was annotated so far, primarily since the experimental methods that are used to determine the protein-DNA and protein-RNA interactions are technically challenging and relatively expensive. These methods are unable to keep pace with the rapid accumulation of the protein, DNA and RNA sequences; the current NCBI's RefSeq database [4] includes over 10 million DNA and RNA transcripts and about 52 million non-redundant proteins from over 51 thousand organisms. As a solution, the currently available experimental data are used to develop time- and cost-efficient computational tools that predict these interactions for the millions of the uncharacterized proteins.

Many computational predictors of the protein-DNA and protein-RNA interactions from the protein sequence and structure have been published and reviewed in the literature over the past several years [1, 5-12]. We focus on the prediction from protein chains since these methods can find the binding proteins and residues in the vast and rapidly growing sequence databases. Differences in the design and outcomes generated by various predictors can be exploited to build consensus-based predictors that take outputs generated by several individual predictors as the inputs. Research in related fields, such as sequence-based prediction of secondary structure and intrinsic disorder, shows that consensuses offer improved predictive performance when compared to the use of individual methods [13-19]. The differences in the design are also characteristic to the sequence-based prediction of DNA- and RNA-binding residues. The inputs to these methods, which represent information about each residue in the input protein sequence, differ in the scope and type of information used. The scope is defined based on the size of sequence segments centered on the predicted residues that are used to generate inputs, which varies widely between 3 and 41 residues [20, 21]. The considered types include various combinations of information about amino acid composition, physiochemical properties of the input amino acids, evolutionary profiles, sequence conservation, and structural characteristics that are predicted from the sequence, such as secondary structure and solvent accessibility. Past methods also utilized different types of predictive models, primarily generated by machine learning algorithms including neural network [20, 21], Support Vector Machine (SVM) [12, 22-24], Naïve Bayes [25], regression [26], decision tree [27], and random forest [28-30].

Consequently, a couple of studies investigated development of consensuses. Si *et al.* [31] developed MetaDBSite consensus that combines six DNA-binding predictors: DBS-pred [20], BindN [32], DP-Bind [26], DISIS [33], DNABindR [34], and BindN-RF [30] using SVM model. This consensus was shown to outperform each of the six predictors [31]. Similarly, Puton *et al.* [11] proposed Meta2 consensus that combines three RNA-binding predictors: PiRaNhA [35], Pprint [36], and BindN+ [22]. Although this approach merges the input predictions based on a simple weighted average, it still outperforms each of the three input predictors [11]. However, these two studies have drawbacks. First, some of the methods that they combine are no longer maintained and thus cannot be used. For instance, the current version of MetaDBSite combines only BindN and DP-Bind. Second, they did not compare and explore different ways to generate the consensuses but simply demonstrated that a given design is successful.

To this end, we explore and empirically compare different ways to generate consensuses and we apply only the currently available and well-maintained input predictors. We investigate the use of simple consensuses and more sophisticated machine learning models. We consider the prediction of both the DNA-binding and the RNA-binding motivated by similarities in the main characteristics of these interactions, e.g., these binding residues in the protein are positively charged and have strong propensity to interact with the negatively charged phosphate backbone of DNA or RNA [37, 38]. In other words, we expect similar conclusions for both types of binding.

2 Materials and Methods

2.1 Selection of methods included in the consensus

We selected eight out of about 30 methods that are available for the prediction of DNA- and RNA-binding residues. These are all methods that were available as reliably working (i.e., able to predict large dataset of proteins) webservers as of December 2013 (when we collected the data) which are characterized by relatively low runtime (i.e., they predict an average sized protein chains with 200 residues in under 10 minutes). We applied the most recent versions of predictors that have multiple versions. The eight methods include five predictors of DNA-binding residues: DBS-PSSM [39], two version of DP-Bind [26, 37], ProteDNA [24], and BindN+ [22]; and three predictors of the RNA-binding residues: Pprint [36], BindN+ [22], and RNABindR [12, 23, 38]. For the DP-Bind, we use two "default" versions based on the kernel logistic regression (KLR) classifier, DP-Bind(klr), and an ensemble of three classifiers, DP-Bind(maj). For the ProteDNA which offers two modes, we use the balanced version, ProteDNA(B), that provides a better balance between sensitivity and specificity [24].

2.2 Datasets

The datasets were collected from the protein-DNA and protein-RNA complexes deposited in the Protein Data Bank (PDB)[40] as of September 2013. We annotated the binding residues utilizing the most prevalent approach based on the cut-off distance at 3.5Å, i.e., a given residue is considered as binding if at least one of its side chain or backbone atoms is closer than 3.5Å from an atom of the RNA/DNA molecule [20]. We collected all 1935 DNA-binding chains and 981 RNA-binding chains which have high-quality X-ray structures, i.e., resolution better than 2.5Å. Next, we improved the annotations of the binding residues by transferring these annotations between homologous proteins using procedure introduced in ref. [41]. Consequently, the number of annotated DNA- and RNA-binding residues was enlarged by 13.7% and 9.7%, respectively. The original redundant datasets were reduced to the non-redundant set 531 DNA- and RNA-binding chains. We divided this dataset into two subsets, the TRAINING and TEST datasets. The former dataset is used to design our consensuses and includes 445 chains that were deposited into PDB before September 2010, the date when the most recent dataset used to build the considered eight predictors was collected. The latter dataset includes newer depositions to assure that we test on independent data were not used to design the considered predictors. Moreover, the original dataset was clustered at 30% similarity using CD-HIT [42] and we removed from the TEST dataset all proteins that ended up in clusters that included any of the proteins from the TRAINING set. This way the final version of the TEST dataset includes 65 chains that share low, <30%, similarity with the chains that are used to design our consensuses and that were used to design the input methods.

2.3 Evaluation

The predictors of DNA- and RNA-binding residues output either only the binary prediction (binding vs. non-binding) or binary prediction together with a real-valued propensity for binding. We evaluate both outputs and exclude residues with missing atomic coordinates in the source structure files (i.e., disordered residues) since we could not complete their annotation of binding. The binary predictions are assessed using accuracy = (TP+TN)/(TP+TN+FP+FN), sensitivity = TP/(TP+FN), specificity =

TN/(FP+TN), and MCC = (TP×TN-FN×FP)/√[(TP+FN)×(TP+FP)×(TN+FP)×(TN+FN)], where TP is the number of true positives (correctly predicted binding residues), FN is the number of false negatives (incorrectly predicted binding residues), FP is the number of false positives (incorrectly predicted non-binding residues), and TN is the number of true negatives (correctly predicted non-binding residues). We primarily rely on the MCC given the unbalanced nature of our datasets, i.e., the number of binding residues is lower than the number of non-binding residues. The propensities are evaluated using Receiver Operating Curve (ROC), which is a plot of false positive rate (FPR = 1 – specificity), against the true positive rate (TPR = sensitivity). These two rates are computed by binarizing the propensities using thresholds and we report the area under the ROC curve (AUC).

2.4 Considered consensus designs

We consider a comprehensive set of simple consensuses designed as the best performing (i.e., securing the highest MCC score on the TRAINING dataset) combinations of *k* methods, *k* = 2, …, *N* where *N* is the number of considered predictors of RNA- or DNA-binding residues. The binary predictions of the *k* methods are combined using logical OR and logical AND operators. The latter design assumes that a given residues is predicted as binding only if all *k* methods predict it as binding; otherwise this residue is predicted as non-binding. The former design predicts a given residues as binding if any of the *k* methods predict it as binding. We used these two operators individually and mixed them together. For instance, given *N* = 3 for the prediction of RNA-binding residues, we explore designs that include “1 AND 2”, “1 AND 2 AND 3”, “1 AND 3”, “1 OR 3”, “1 OR 2 OR 3”, “(1 AND 3) OR 2”, “1 AND (2 OR 3)”, etc. In total, we considered 10 and 116 designs for the prediction of RNA-binding residues (*N* = 3) and DNA-binding residues (*N* = 5), respectively. We select one, best-performing consensus (i.e., consensus that provides the highest value of MCC on the TRAINING dataset) for the prediction of DNA-binding residues and for the prediction of RNA-binding residues.

We also utilize more sophisticated designs where the predictions for a given residue, including both binary values and propensities, from the *N* methods are combined using predictive models generated by five different popular types of machine learning algorithms. We include the linear logistic regression (LLR), C4.5 decision tree (C4.5), *k*-nearest neighbor (kNN), SVM, and Naïve Bayes (NB) using the implementations from the WEKA platform [43]. Each of these classifiers was parameterized based on five-fold cross validation on the TRAINING dataset. We use grid search to select parameters that provide the maximal value of MCC. For LLR, we adjust the number of boosting iterations *n* = {0, 1, …, 10}; for C4.5 we parameterize confidence factor *c* = {0.05, 0.1,…, 0.5} and minimal number of instances per leaf node *m* = {1, 2, …, 5} that are used for pruning; for kNN we optimize number of neighbors *k* = {1, 2, …, 30}; for SVM we use the Gaussian kernel and find the best values of complexity parameter C = {2^{-3}, 2^{-1}, …, 2^{3}} and width of the kernel gamma = {2^{-2}, 2^{0}, …, 2^{8}}. Since all these consensuses generate real-values propensity as the output, we binarize it to obtain the binary prediction (binding vs. non-binding) by selecting a threshold that gives maximal value of MCC on the TRAINING dataset.

	Method	Accuracy	Sensitivity	Specificity	MCC	Sig	AUC	Sig
DNA-binding	ML consensus LLR	0.857	0.594	0.873	**0.304**		**0.839**	
	ML consensus C4.5	0.889	0.485	0.915	0.301	=	0.789	+
	ML consensus kNN	0.810	0.682	0.818	0.287	+	0.826	+
	ML consensus SVM	0.823	0.648	0.834	0.286	+	0.742	+
	ML consensus NB	0.805	0.664	0.814	0.273	+	0.829	+
	Simple consensus	0.890	0.424	0.919	0.267	+		
	DBS-PSSM	*0.771*	*0.721*	*0.774*	*0.266*	+	*0.810*	+
	BindN+	*0.865*	*0.482*	*0.888*	*0.256*	+	*0.806*	+
	DP-Bind(maj)	*0.810*	*0.598*	*0.823*	*0.247*	+		
	DP-Bind(klr)	*0.814*	*0.590*	*0.828*	*0.246*	+	*0.794*	+
	MetaDBSite consensus	0.898	0.325	0.933	0.221	+		
	ProteDNA(B)	*0.937*	*0.093*	*0.990*	*0.158*	+		
RNA-binding	ML consensus LLR	0.920	0.257	0.939	**0.128**		**0.731**	
	ML consensus SVM	0.919	0.249	0.938	0.123	+	0.618	+
	ML consensus NB	0.931	0.215	0.952	0.121	=	0.727	+
	Meta2 consensus	0.768	0.526	0.774	0.116	+		
	ML consensus kNN	0.927	0.218	0.947	0.115	+	0.711	+
	BindN+	*0.841*	*0.399*	*0.854*	*0.114*	+	*0.706*	+
	Simple consensus	0.915	0.244	0.933	0.113	+		
	RNABindR	*0.714*	*0.575*	*0.718*	*0.105*	+	*0.712*	+
	ML consensus C4.5	0.942	0.154	0.965	0.100	+	0.610	+
	Pprint	*0.773*	*0.433*	*0.782*	*0.084*	+	*0.667*	+

Table 1. Results of empirical assessment of predictors of the DNA- or RNA-binding residues on the TEST dataset. Significance of the difference in MCC and AUC values between the best performing method and other methods for a given binding type was assessed based on 10 tests that utilize 70% of randomly chosen proteins; if the measurements are normal, as tested using the Anderson–Darling test at 0.05 significance, we use the paired *t*-test; otherwise we use the Wilcoxon rank sum test; + (=) in the Sig column denotes that the difference was (was not) significant at *p*-value <0.05. AUC values could not be computed for DP-Bind(maj), MetaDBSite, ProteDNA(B), Meta2, and the two simple consensuses since these methods provide only the binary predictions. The highest MCC and AUC values for each type of binding are given in **bold** font. Individual predictors are denoted with *italics*.

3 Results and discussion

The predictive performance of the considered individual methods, the best performing simple consensus and the considered five machine learning consensuses on the TEST dataset for the prediction of the DNA-binding and the RNA-binding residues is summarized in Table 1. The methods are sorted by their MCC values. We also include results for the two published consensuses: MetaDBSite [31] for the DNA-binding and the Meta2 consensus by Puton et al. [11] for the RNA-binding; their predictions were collected using the corresponding webservers.

The selected simple consensuses, which are characterized by the best predictive performance on the TRAINING dataset, include the AND-based combinations: "BindN+ AND DBS-PSSM" for the prediction of DNA-binding residues, and "BindN+ AND RNABindR AND Pprint" for the RNA-binding residues. Although these consensuses provide improvements in predictive quality when compared with the individual predictors on the TRAINING dataset (increase in MCC by 0.01 and 0.04 for the DNA-and RNA-binding, respectively), Table 1 reveals that this did not translate into the TEST dataset. The simple consensuses obtain the same predictive performance as the best individual method, MCC of 0.267 vs. 0.266 of the best individual method DBS-PSSM for the DNA-binding and 0.113 vs. 0.114 of the best BindN+ for the RNA-binding. The reason is that TEST dataset shares low sequence similarity with the TRAINING set. This results in differences in predictions of individual methods between the two datasets that negatively affect accuracy of the simple designs of the consensus. In fact, the simple consensuses that obtain the best results on the TEST dataset for the DNA-binding "BindN+ OR DBS-PSSM AND DP-Bind(klr) OR ProteDNA(B)" and for the RNA-binding "BindN+ AND RNABindR" secure higher MCCs that equal 0.291 and 0.118, respectively, on that dataset. We conclude that the consensuses that rely on the simple design that combines binary predictions are unlikely to provide improved predictive performance when applied to sequences that share low similarity with the datasets used to build their input predictors.

Fig. 1. ROC generated on the TEST dataset for the best performing ML consensus and the considered individual predictors that generate real-values propensity scores for prediction of DNA-binding and RNA-binding residues.

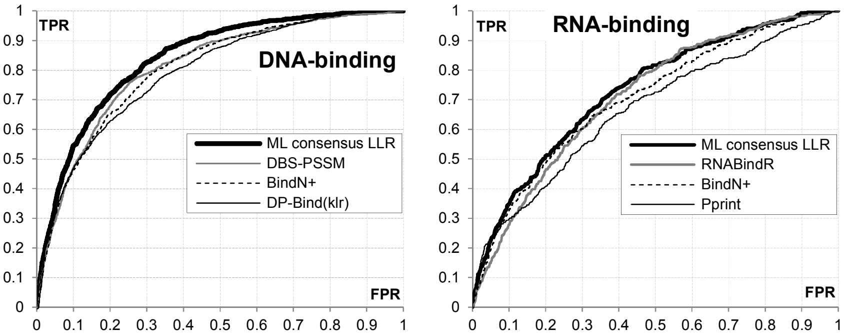

Table 1 demonstrates that consensuses based on certain machine learning models offer improved predictive performance when compared with the best individual predictors. In particular, the linear regression (LLR model) secures the highest MCC and AUC values for prediction of both RNA- and DNA-binding residues, and these values are significantly higher than the values offered by the individual predictors. The ROCs of the LLR consensus and the corresponding individual predictors that generate real-values propensities are compared in Figure 1. These curves reveal that this consensus outperforms the other methods for virtually entire range of the FPR values, except for the low FPR<0.04 for the RNA-binding where Pprint offers slightly higher TPR values. Two other machine learning models, SVM and NB, also offer improvements for the prediction of RNA- and DNA-binding residues. The other two models, C4.5 and kNN, provide improvements for the prediction of DNA-binding residues but not for the prediction of the RNA-binding residues. To sum up, we observe that consensuses that rely on certain more sophisticated models provide improved predictive performance, even when tested using chains that share low sequence similarity with proteins that were used to build their input predictors.

Fig. 2. Correlation between pairs of individual predictors (narrow bars) and the best performing ML consensus and each individual predictor (wide bars) for the prediction of DNA-binding residues (black bars) and RNA-binding residues (gray bars).

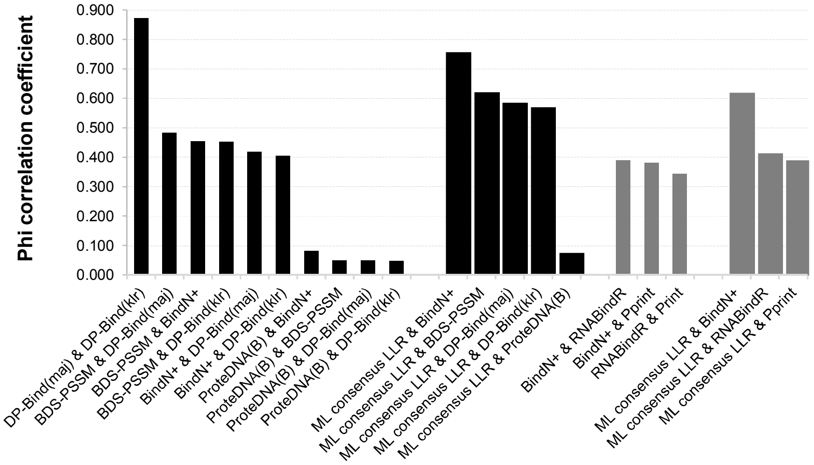

Figure 2 provides insights that may explain why consensuses are successful. It gives values of the Phi correlation coefficient (PhiCC), which is an equivalent of the Pearson correlation coefficient for a pair of binary variables, between the binary predictions of all pairs of the individual methods (thin bars) and between the binary predictions of our LLR consensus and each individual predictor. Except for the pair of DP-Bind(maj) and DP-Bind(klr) methods that share very similar design [26, 37] and consequently secure high correlation close to 0.9, the predictions of the other methods are only modestly correlated with the PhiCC values < 0.5 for the DNA-binding and < 0.4 for the RNA-binding. This could be explained by substantial differences in the design of these methods. For instance, BindN+ uses information concerning physiochemical properties of the input amino acids, sequence alignment, evolutionary profiles, and the SVM model. DP-Bind uses regression model and inputs that solely rely on the evolutionary profiles. DBS-PSSM also uses the evolutionary profiles but with the neural network model. RNABindR applies SVM model and the evolutionary profiles. We also note the low correlations for any pair of methods for the prediction of DNA-binding residues that includes ProteDNA(B). This method predicts a subset of DNA-binding residues that bind transcription factors, which is why it secures low sensitivity (Table 1) and has low correlations. The modest levels of correlations between individual predictors are exploited by the consensus. In other words, since all individual predictors offer relatively good predictive performance and their predictions are substantially different (modestly correlated), these predictions must complement each other. A similar observation was made in the context of the sequence-based prediction of intrinsic disorder [16]. The wide bars in Figure 2 suggest that the LLR-based consensus has higher correlation with the individual methods, >0.57 for the prediction of DNA-binding residues and >0.39 for the prediction of RNA-binding residues (except for ProteDNA(B) which under-predicts the binding residues). This combined with the fact that our consensus obtains higher predictive performance means that it effectively takes advantage of this complementarity between the input predictors.

Finally, we analyze predictive performance of the two existing consensuses. The MCC of MetaDBSite is relatively low and lower than MCC of some of the considered individual predictors (Table 1). The reason is that this approach is currently implemented a simple consensus "BindN AND DP-Bind" since the other four predictors that it was originally designed to include are no longer available. The Meta2 consensus for the prediction of RNA-binding residues outperforms its input predictors Pprint and BindN+ (Table 1). This consensus is based on a weighted average, which is more complex than our simple consensus designs, but is less sophisticated than our machine learning designs. Correspondingly, Meta2 provides lower predictive performance than our consensuses based on LLR, SVM and NB models.

To summarize, our empirical study suggests that sequence-based prediction of RNA- and DNA-binding residues would benefit from the use of machine learning consensuses. Such consensuses exploit complementarity between individual predictors to generate predictions with significantly higher predictive quality when compared with the individual predictors, even for the chains characterized by low sequence similarity with the proteins used to develop these predictors.

4 References

1. Re, A., et al., *RNA-protein interactions: an overview.* Methods Mol Biol, 2014. **1097**: p. 491-521.
2. Dey, B., et al., *DNA-protein interactions: methods for detection and analysis.* Mol Cell Biochem, 2012. **365**(1-2): p. 279-99.
3. Charoensawan, V., D. Wilson, and S.A. Teichmann, *Genomic repertoires of DNA-binding transcription factors across the tree of life.* Nucleic Acids Res, 2010. **38**(21): p. 7364-77.
4. Pruitt, K.D., et al., *RefSeq: an update on mammalian reference sequences.* Nucleic Acids Res, 2014. **42**(Database issue): p. D756-63.
5. Zhao, H., Y. Yang, and Y. Zhou, *Prediction of RNA binding proteins comes of age from low resolution to high resolution.* Mol Biosyst, 2013. **9**(10): p. 2417-25.
6. Fornes, O., et al., *On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions.* Adv Protein Chem Struct Biol, 2014. **94**: p. 77-120.
7. Kauffman, C. and G. Karypis, *Computational tools for protein-DNA interactions.* Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 2012. **2**(1): p. 14-28.

8. Liu, L.A. and P. Bradley, *Atomistic modeling of protein-DNA interaction specificity: progress and applications*. Curr Opin Struct Biol, 2012. **22**(4): p. 397-405.
9. Gromiha, M.M. and R. Nagarajan, *Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein-DNA complexes.* Adv Protein Chem Struct Biol, 2013. **91**: p. 65-99.
10. Ding, X.M., et al., *Computational prediction of DNA-protein interactions: a review*. Curr Comput Aided Drug Des, 2010. **6**(3): p. 197-206.
11. Puton, T., et al., *Computational methods for prediction of protein-RNA interactions*. J Struct Biol, 2012. **179**(3): p. 261-8.
12. Walia, R.R., et al., *Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art.* Bmc Bioinformatics, 2012. **13**.
13. Yan, J., M. Marcus, and L. Kurgan, *Comprehensively designed consensus of standalone secondary structure predictors improves Q3 by over 3%.* J Biomol Struct Dyn, 2014. **32**(1): p. 36-51.
14. Zhang, H., et al., *Critical assessment of high-throughput standalone methods for secondary structure prediction*. Brief Bioinform, 2011. **12**(6): p. 672-88.
15. Fan, X. and L. Kurgan, *Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus*. J Biomol Struct Dyn, 2014. **32**(3): p. 448-64.
16. Peng, Z. and L. Kurgan, *On the complementarity of the consensus-based disorder prediction.* Pac Symp Biocomput, 2012: p. 176-87.
17. Kozlowski, L.P. and J.M. Bujnicki, *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins*. BMC Bioinformatics, 2012. **13**: p. 111.
18. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder*. Bioinformatics, 2015. **31**(2): p. 201-8.
19. Albrecht, M., et al., *Simple consensus procedures are effective and sufficient in secondary structure prediction*. Protein Eng, 2003. **16**(7): p. 459-62.
20. Ahmad, S., M.M. Gromiha, and A. Sarai, *Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information.* Bioinformatics, 2004. **20**(4): p. 477-86.
21. Jeong, E., I.F. Chung, and S. Miyano, *A neural network method for identification of RNA-interacting residues in protein.* Genome Inform, 2004. **15**(1): p. 105-16.
22. Wang, L., et al., *BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features.* BMC Syst Biol, 2010. **4 Suppl 1**: p. S3.
23. Terribilini, M., et al., *RNABindR: a server for analyzing and predicting RNA-binding sites in proteins*. Nucleic Acids Research, 2007. **35**: p. W578-W584.
24. Chu, W.Y., et al., *ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors*. Nucleic Acids Research, 2009. **37**: p. W396-W401.
25. Lee, J.H., et al., *Striking similarities in diverse telomerase proteins revealed by combining structure prediction and machine learning approaches*. Pac Symp Biocomput, 2008: p. 501-12.
26. Hwang, S., Z.K. Gou, and I.B. Kuznetsov, *DP-Bind: a Web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins*. Bioinformatics, 2007. **23**(5): p. 634-636.
27. Carson, M.B., R. Langlois, and H. Lu, *NAPS: a residue-level nucleic acid-binding prediction server*. Nucleic Acids Research, 2010. **38**: p. W431-W435.
28. Ma, X., et al., *Sequence-Based Prediction of DNA-Binding Residues in Proteins with Conservation and Correlation Information*. Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2012. **9**(6): p. 1766-1775.

29. Ma, X., et al., *Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature.* Proteins-Structure Function and Bioinformatics, 2011. **79**(4): p. 1230-1239.
30. Wang, L.J., M.Q. Yang, and J.Y. Yang, *Prediction of DNA-binding residues from protein sequence information using random forests*. Bmc Genomics, 2009. **10**.
31. Si, J., et al., *MetaDBSite: a meta approach to improve protein DNA-binding sites prediction.* BMC Syst Biol, 2011. **5 Suppl 1**: p. S7.
32. Wang, L.J. and S.J. Brown, *BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences.* Nucleic Acids Research, 2006. **34**: p. W243-W248.
33. Ofran, Y., V. Mysore, and B. Rost, *Prediction of DNA-binding residues from sequence.* Bioinformatics, 2007. **23**(13): p. I347-I353.
34. Yan, C.H., et al., *Predicting DNA-binding sites of proteins from amino acid sequence*. Bmc Bioinformatics, 2006. **7**.
35. Murakami, Y., et al., *PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences*. Nucleic Acids Research, 2010. **38**: p. W412-W416.
36. Kumar, M., A.M. Gromiha, and G.P.S. Raghava, *Prediction of RNA binding sites in a protein using SVM and PSSM profile.* Proteins-Structure Function and Bioinformatics, 2008. **71**(1): p. 189-194.
37. Kuznetsov, I.B., et al., *Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins.* Proteins-Structure Function and Bioinformatics, 2006. **64**(1): p. 19-27.
38. Terribilini, M., et al., *Prediction of RNA binding sites in proteins from amino acid sequence.* Rna-a Publication of the Rna Society, 2006. **12**(8): p. 1450-1462.
39. Ahmad, S. and A. Sarai, *PSSM-based prediction of DNA binding sites in proteins.* BMC Bioinformatics, 2005. **6**: p. 33.
40. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
41. Chen, K., et al., *A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds.* Structure, 2011. **19**(5): p. 613-21.
42. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-2.
43. Frank, E., et al., *Weka-A Machine Learning Workbench for Data Mining.* Data Mining and Knowledge Discovery Handbook, Second Edition, 2010: p. 1269-1277.