Briefings in Bioinformatics, 2017, 1250–1268

doi: 10.1093/bib/bbx168
Paper

Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains

Jian Zhang, Zhiqiang Ma and Lukasz Kurgan

Corresponding author: Lukasz Kurgan, Department of Computer Science, Virginia Commonwealth University, Richmond 23284, USA. Tel.: +1-804-827-3986; E-mail: lkurgan@vcu.edu

Abstract

Proteins interact with a variety of molecules including proteins and nucleic acids. We review a comprehensive collection of over 50 studies that analyze and/or predict these interactions. While majority of these studies address either solely protein–DNA or protein–RNA binding, only a few have a wider scope that covers both protein–protein and protein–nucleic acid binding. Our analysis reveals that binding residues are typically characterized with three hallmarks: relative solvent accessibility (RSA), evolutionary conservation and propensity of amino acids (AAs) for binding. Motivated by drawbacks of the prior studies, we perform a large-scale analysis to quantify and contrast the three hallmarks for residues that bind DNA-, RNA-, protein- and (for the first time) multi-ligand-binding residues that interact with DNA and proteins, and with RNA and proteins. Results generated on a well-annotated data set of over 23 000 proteins show that conservation of binding residues is higher for nucleic acid- than protein-binding residues. Multi-ligand-binding residues are more conserved and have higher RSA than single-ligand-binding residues. We empirically show that each hallmark discriminates between binding and non-binding residues, even predicted RSA, and that combining them improves discriminatory power for each of the five types of interactions. Linear scoring functions that combine these hallmarks offer good predictive performance of residue-level propensity for binding and provide intuitive interpretation of predictions. Better understanding of these residue-level interactions will facilitate development of methods that accurately predict binding in the exponentially growing databases of protein sequences.

Key words: protein–RNA interactions; protein–DNA interactions; protein–nucleic acid interactions; protein–protein interactions; DNA-binding residues; RNA-binding residues

Introduction

Proteins carry out their cellular functions by interacting with DNA [1], RNA [2], proteins [3] and a variety of other ligands [4–6]. Our understanding of these interactions and corresponding functional annotations of proteins are primarily derived from structural studies of protein–ligand complexes. These structural data are used to categorize these interactions, characterize the underlying physics and decipher patterns that define molecular recognition and specificity of interactions [7–11]. These limited in quantity data are also used to derive computational models that predict DNA, RNA and protein interactions from protein structures and sequences [12–15].

The characterization and prediction of these interactions can be done at a few different levels [16]. At the lowest

Jian Zhang is a Lecturer in the School of Computer and Information Technology at the Xinyang Normal University. His research interests are in machine learning and bioinformatics.

Zhiqiang Ma is a Professor of the College of Humanities and Sciences at the Northeast Normal University in Changchun. His research portfolio spans biometrics, bioinformatics and data mining.

Lukasz Kurgan is a Qimonda Endowed Professor of Computer Science at the Virginia Commonwealth University in Richmond. His research focuses on high-throughput structural and functional characterization of proteins and small RNAs. More details about his research group are at http://biomine.cs.vcu.edu/.

Submitted: 30 August 2017; **Received (in revised form):** 15 November 2017

resolution level, we ascertain these interactions at the whole-protein scale, e.g. predict whether a given protein binds RNA without further details about this interaction. At the medium resolution level, we additionally analyze and/or predict which residues in the protein sequence interact with the ligand. Finally, at the highest resolution level, we model or predict these interactions at the atomic scale using the three-dimensional (3D) structure of the protein and its ligand. The choice of the resolution often depends on the availability of the corresponding data. More specifically, whether only the sequence or both sequence and structure of the protein are available. The structure-based analysis and prediction of protein–ligand interactions are limited to a relatively small number of proteins that have 3D structures. As of August 2017, Protein Data Bank (PDB) [17, 18], the main database of protein structures, holds 123 000 structures that cover close to 42 000 proteins. Alternatively, a high-quality predicted structure can be used, but this inadvertently reduces quality of the analysis/prediction, and it still substantially limits the coverage. For instance, a recent study found that structures can be predicted for about 28% of human proteins [19]. Another larger-scale study that considered about 1500 organisms across the three kingdoms of life has estimated that the current structural coverage that includes known and predicted structures ranges between a few to 30% [20]. In contrast, sequence-based approaches require only the widely available protein sequences. Based on the UniProt resource [21, 22], the number of proteins with the known sequences currently stands at about 88 million as of August 2017. Moreover, proteomes (i.e. set of proteins thought to be expressed by organisms with genomes that were completely sequenced) are available for 99.5 000 of unique organisms. The sequence-based analysis is sometimes coupled with predicted or actual structural information, typically at a lower level, such as secondary structure or solvent accessibility, to improve the quality of the analysis and prediction.

We review a comprehensive set of recent articles that analyze and/or predict protein–protein, protein–DNA and protein–RNA interactions at all levels of resolution. Our first objective is to delineate a set of commonly analyzed and used hallmarks of these interactions, i.e. information concerning protein sequence and/or structure that can be used to identify interacting proteins and residues. Our second objective is to perform a comprehensive empirical analysis of these hallmarks. The novel and particularly unique aspect of this review and analysis is the wide-ranging scope. Unlike majority of previous studies, we analyze interactions with proteins, DNA and RNAs, instead of focusing on a single ligand. We are the first to analyze residues that bind to both RNA and proteins, and to both DNA and proteins, and to contrast and compare findings between the ligands. Moreover, we investigate whether the commonly used hallmarks are sufficient to predict binding residues in protein sequences, and we also develop and comparatively assess the first method that predicts residue-level propensity for protein–ligand interactions over the three types of ligands: DNA, RNAs and proteins.

Review of contributions that analyze and/or predict protein–DNA, protein–RNA and protein–protein binding

Based on manual analysis of PubMed queries, we found 52 studies, which were published over the past decade and that focus on the analysis and/or prediction of protein–RNA, protein–DNA and protein–protein interactions [23–74]. They are summarized in Table 1. Most of these studies (32 of 52) focus on the interactions with either DNA or RNA. More specifically, 12 works concern DNA binding, 20 RNA binding and 12 protein binding. In contrast, only five studies consider interactions with both type of nucleic acids, and even fewer (three works) focus on both protein–proteins and protein–nucleic acid interactions. Two of the latter three articles use information generated from protein sequences to build models to predict DNA-, RNA- and protein-binding residues [73, 72]. In the third article, the authors develop a predictor of protein-binding residues, and they analyze sequence- and structure-based hallmarks of protein-, DNA- and RNA-binding residues [74]. Moreover, none of these works consider investigating residues that interact with both nucleic acids and proteins partners. We also acknowledge methods that predict residues interacting with other types of ligands [15]. This family of predictors includes methods that predict binding to nucleotides [75, 76], vitamins, [77, 78], calcium [79], metals [80], as well as multiple types of small ligands [81, 82].

The 52 studies that focus on protein–RNA, protein–DNA and protein–protein interactions include 36 that develop predictors and 16 that analyze structural data, including 8 that use the analysis to derive a predictor (see the 'contributions' column in Table 1). Interestingly, most of these predictors are available as Web servers, which allow even novice users to conveniently perform predictions. Moreover, we observe a noticeable increase in the number of protein chains used for the analysis or the development of a given predictor with the time. The Person correlation coefficient between the year of publication and the average size of data sets used in a given year equals 0.6. This is expected as gradually more structural data are released in PDB.

We found that these 52 works consider and combine various types of information to characterize and/or predict the corresponding binding residues. The most frequently used information includes evolutionary conservation (ECO) (38 of 52), solvent accessibility (32 of 52) and propensity of amino acids (AAs) for binding (23 of 52 studies). The ECO is relevant, as binding residues are typically conserved across homologous protein sequences. The use of solvent accessibility is motivated by the fact that binding occurs on the protein surface. Finally, these studies also suggest that the type of the interacting AAs and their immediate neighbors in the protein sequence can be also used to determine propensity for binding. Each of the 52 studies uses at least one of these three hallmarks of binding residues and about half consider at least two hallmarks. Three articles cover the three hallmarks in the context of RNA-binding proteins [41, 42, 50] and one article for the protein binding [83]. A number of other factors that can be used to characterize binding residues were also considered. These include secondary structure [29, 31, 37, 41, 45, 56, 69, 73, 74], area of the binding sites [27, 35, 36, 56], regions of binding residues [27, 45], surface patches in the binding sites [26, 47, 49, 56, 61], shape of binding sites [29, 49, 56, 65] and their geometric similarity [25, 29, 60]. However, these factors were considered in a relatively small number of articles, and they primarily focus on the characteristics of the protein structure. This is in contrast to two of the main hallmarks, propensity of AAs for binding and ECO, which rely solely on the protein sequence.

While these studies contribute to improving our understanding of the determinants of binding in protein structures and sequences and provide community with putative annotations of binding, they also suffer a few shortcomings:

1. They use relative small data sets of chains that often cover fragments of complete protein sequences and that may

Table 1. Review of recent studies that focus on the analysis and/or prediction of protein–RNA, protein–DNA and protein–protein interactions

Type of interaction[1]	Reference	Year of publication	Number of chains used	Contributions[2]	Availability of prediction tools[3]	Considered hallmarks of binding residues[4]		
						AAP	ECO	RSA
D	[23]	2007	62	P	W		√	
D	[24]	2010	87	P	W		√	
D	[25]	2008	118	A		√		
D	[26]	2012	126	A and P		√	√	
D	[27]	2008	140	A and P			√	√
D	[28]	2013	206	P	W		√	√
D	[29]	2014	272	A and P		√	√	
D	[30]	2016	286	P	W	√	√	√
D	[31]	2012	337	P	W		√	
D	[32]	2014	435	P	W		√	
D	[33]	2016	584	P	W		√	√
D	[34]	2016	605	P	W		√	
R	[35]	2008	81	A		√		√
R	[36]	2012	81	P	W			√
R	[37]	2008	86	A and P	W	√	√	
R	[38]	2008	109	P	W		√	√
R	[39]	2014	116	P	W		√	√
R	[40]	2007	147	P	W	√	√	
R	[41]	2010	147	A and P		√	√	√
R	[42]	2011	160	P	W	√	√	√
R	[43]	2017	172	A and P	W	√	√	√
R	[44]	2010	205	P	S		√	√
R	[45]	2011	211	A				√
R	[46]	2008	302	P			√	
R	[47]	2010	316	P				√
R	[48]	2011	332	P	S		√	√
R	[49]	2015	344	P				√
R	[50]	2014	346	A and P	W	√	√	√
R	[51]	2016	443	P	S	√	√	
R	[52]	2011	569	P	W		√	
R	[53]	2014	952	P	W	√		
R	[54]	2011	4143	P	S	√		
P	[55]	2015	55	A		√	√	
P	[56]	2010	103	A		√		√
P	[57]	2013	122	P			√	√
P	[58]	2010	186	P	W		√	√
P	[59]	2014	186	P	W	√	√	√
P	[60]	2015	193	P	S		√	
P	[61]	2016	302	P				√
P	[62]	2015	422	P	S			√
P	[63]	2016	422	P	S		√	√
P	[64]	2016	422	P	S		√	√
P	[65]	2011	1496	A		√		√
P	[66]	2016	1905	P	S		√	√
D, R	[67]	2007	147	P		√		√
D, R	[68]	2014	149	A			√	√
D, R	[69]	2014	1017	P	W		√	
D, R	[70]	2015	1950	P	W	√	√	√
D, R	[71]	2017	4604	P	W	√	√	√
D, R, P	[72]	2015	86	A and P	W		√	√
D, R, P	[73]	2015	315	P	W	√		
D, R, P	[74]	2011	446	A		√	√	
D, R, P, DP, RP	This work	N/A	23458	A and P	W	√	√	√

Note: Studies are sorted by the size of the data set they used and grouped by the type(s) of interactions that they cover.
[1]D, R, P, DP and RP stand for DNA binding, RNA binding, protein binding, DNA and protein binding, and RNA and protein binding, respectively.
[2]A and P stand for analysis and prediction, respectively.
[3]W and S correspond to the availability of the Web server and source code, respectively.
[4]AAP, ECO and RSA refer to the three main hallmarks of binding: AA type preferences, ECO and RSA, respectively.

provide incomplete annotations of binding. Works that analyze DNA binding, RNA binding and protein binding use only up to 605, 4143 and 4604 protein sequences in complex with DNA, RNA and proteins, respectively. They were collected from PDB [17], and thus, they may not provide complete coverage of the underlying protein sequence. Moreover, these data sets include chains that typically have limited pairwise sequence similarity that were derived by selecting one chain from a cluster of similar chains. While this provides a desired uniform sampling of the sequence space, it also results in under-annotating binding residues. In other words, when the same chain is found in multiple complexes, only one of them is used to annotate binding, while in fact this chain could bind to ligands of the same type (e.g. different fragments of RNA) localized in different binding sites across these complexes. Thus, the annotations of binding should be transferred between these complexes to ensure that they are more complete.

2. Most of these studies (44 of 52) focus on one type of ligand and none consider residues that bind to multiple ligand types, including residues that bind to both DNA and proteins and to both RNA and proteins, which we name multi-ligand-binding residues. Consequently, the hallmarks that characterize binding were rarely compared between RNA, DNA and protein binding and were never quantified and compared with the multi-ligand-binding residues.
3. Most of the predictors of DNA, RNA and proteins binding use difficult to comprehend inputs and intricate black-box models to produce predictions. Thus, factors that contribute to their predictions are unclear diminishing ability of the users to interpret the predictions, e.g. to understand why a given residue is predicted as binding to RNA.

Motivated by this analysis, we perform first-of-its-kind comprehensive comparative analysis of the main hallmarks of protein, DNA and RNA binding. We develop a large data set of complete protein sequences that includes annotations transferred from identical protein fragments in different complexes. We use these data to analyze and compare the three hallmarks between residues that bind RNA, DNA, proteins, both RNA and proteins and both DNA and proteins. Additionally, we also empirically test whether use of just these three hallmarks is sufficient to predict binding residues in protein sequences. We also develop, empirically test and deploy as a Web server scoring functions that use hallmarks computed/predicted from a protein sequence to predict residue-level propensity for DNA, RNA and protein binding. We empirically compare these scoring functions against existing methods for the sequence-based prediction of DNA-, RNA- and protein-binding residues.

Setup of the empirical analysis

Data sets

We collect proteins, which were solved structurally in complex from the BioLiP database in October 2015 [84]. BioLip stores high-quality semi-manually curated annotations of biologically relevant protein–ligand interactions extracted from PDB. It annotates a given residue as binding if the distance between an atom of this residue and an atom of the ligand $<0.5+$ sum of the Van der Waal's radii of the two atoms [84]. BioLip includes 5913 DNA-binding, 20 731 RNA-binding, 163 589 protein-binding and 112 797 ligand-binding chains, some of which are redundant. In contrast to the other studies that typically consider one complex per protein and which may cover a fragment of the complete protein sequence, we use complete proteins that combine annotations from potentially multiple complexes. We map BioLiP sequences into UniProt [85] with SIFTS [86] to ensure that we work with complete protein sequences and to transfer annotations from multiple BioLiP/PDB protein chains that are linked to the same UniProt protein. Next, we remove UniProt IDs that correspond to protein fragments and combine annotations of binding residues from all PDB structures that are mapped to the same protein (UniProt ID). This way we annotate 27% more binding residues when compared with the best case scenario of how prior works annotate binding residues, i.e. when chains with the highest number of binding residues are used to cover the complete protein sequence. This enrichment is accomplished by transferring binding residues from BioLiP/PDB chains that cover the same fragment of the UniProt sequence. Supplementary Figure S1 summarizes the enrichment in the annotations of binding residues for different types of ligands. The final data set includes 23 458 proteins (817 DNA-binding, 1040 RNA-binding, 17 594 protein-binding and 14 327 ligand-binding proteins) that were extracted from 303 030 chains from BioLip and annotated using 294 447 structures of PDB chains (Table 2). This means that besides providing a more complete annotation of binding residues, our data set is at least five times bigger than any of the data sets used in the prior studies (Table 1); this includes the data sets that we used in our recent articles that address prediction of protein-binding residues [87] and assessment and development of predictors of DNA- and RNA-binding residues [71, 88]. We consider binding residues that interact with DNA, RNA, proteins, DNA and proteins, RNA and proteins and nonbinding residues, which lack annotations of binding including binding to small ligands. This data set is used to benchmark the hallmarks of DNA, RNA and protein binding and to empirically compare predictive performance of different combinations of these hallmarks. The comparison is performed based on the 3-fold cross-validation test on this benchmark data set. In this test, the proteins are divided at random into three equally sized and disjoint subsets of the benchmark data set. Models generated using proteins from two subsets are tested on the third subset, and this is repeated three times to use every subset once as the test subset.

We also compare predictive performance of these hallmarks with the results obtained by current methods that predict DNA-, RNA- and protein-binding residues from protein sequences. This assessment is based on test data sets collected from recent relevant studies. They include the DNA_T and RNA_T data sets that were published as a part of a comparative review of methods that predict DNA- and RNA-binding residues [88] (Table 2). Similar to our data set, these test data sets incorporate the transferred annotations. We also use the test data set that was introduced in [58] and which was used to assess several predictors of protein-binding residues [58, 83, 89, 90]. We enrich the original test data set with the transferred annotations using the abovementioned protocol, and we name the resulting data set protein_T (Table 2). Moreover, we ensure that the proteins from our large benchmark data set that are used to train our model share low similarity with the proteins from these three test data sets. We use BLASTCLUST (using default settings including coverage -L 0.9) to remove proteins from the benchmark data set that were annotated using at least one PDB chain that shares 30% or higher similarity with the proteins from a given test data set. Only the remaining proteins are used to compute our models. We emphasize that the similarity is measured against PDB chains that were used to annotate a given sequence collected from UniProt, resulting in a stricter filter of similarity. We name

Table 2. Summary of the benchmark and test data sets used

Type of data set	Data sets	Number of proteins	Number of PDB chains	Number of binding residues	Number of nonbinding residues	Total number of residues
Benchmark	All	23 458	294 447	1 070 757	5 070 108	6 140 865
	DNA binding	817	19 152	19 987	153 475	199 881
	RNA binding	1040	45 337	38 899	158 506	223 814
	Protein binding	17 594	269 705	791 918	3 444 379	4 405 232
	Small ligand binding	14 327	240 335	267 385	3 665 994	4 370 210
	DNA and protein binding	386	13 703	2507	63 956	93 995
	RNA and protein binding	686	39 412	5846	76 183	125 751
Test	DNA_T	47	138	875	8231	9106
	RNA_T	17	46	409	5448	5857
	Protein_T	72	764	2452	15 688	18 140

Note: The sum of the number of binding residues is larger than the total listed in the 'total' column, as some residues bind to multiple ligands. The 'number of PDB chains' column gives the number of structures of chains from PDB that were used to annotate binding for a given data set. The three test data sets were collected from [58, 95].

the corresponding data sets training30_DNA, training30_RNA and training30_protein. As the training data sets include proteins that share <30% similarity to the proteins in the three corresponding test data sets, the annotations transferred into the proteins in the test data sets could not originate from any of the training proteins. The benchmark data sets, including a mapping of these proteins into the corresponding PDB chains and the native annotations of binding residues; the training30_DNA, training30_RNA and training30_protein training data sets; and the DNA_T, RNA_T and protein_T test data sets, are all available at http://biomine.cs.vcu.edu/servers/hybridNAP/.

Computation of the three hallmarks of binding residues

We analyze the three most commonly considered hallmarks of binding: (i) propensity for binding of AAs in the sequence; (ii) relative solvent accessibility (RSA) of residues in the structure; and (iii) ECO of residues in the sequences. They are analyzed for each ligand type including the two types of multi-ligand binding.

The AA propensity for binding is quantified as Relative difference in abundance of a given AA type (RAA) between binding residues and the corresponding nonbinding residues located on the protein surface. We focus on the surface to eliminate a confounding factor related to a bias in composition of AAs in the protein core that typically is not involved in binding. RAA is defined as the difference between fractions of a given AA type among binding residues and among the surface nonbinding residues divided by the fraction among the latter residues. Positive (negative) values denote enrichment (depletion) among binding residues compared with the nonbinding residues. We compute the relative differences using Composition Profiler [91]. We score these propensities by considering the binding residues in the sequence. We use a weighted average, where a weight = 0.5 was assigned to the residue that is scored and 0.25 to each of its neighbors.

The area of solvent accessibility (ASA) for each residue in a protein was derived from the protein structure, based on the corresponding PDB file, using the DSSP program [92]. Then, RSA was calculated by dividing the native ASA by empirically derived maximal value of ASA for a given AA type obtained from [93]; these values are also used to annotate surface residues to compute RAA. Moreover, to consider proteins that do not have structures and for which the native RSA cannot be calculated, we use putative ASA generated with ASAquick [94] from the sequences, and the corresponding putative RSA values. We empirically analyze whether these predictions can be used to substitute native solvent accessibility for the purpose of prediction the binding residues solely from the protein sequence.

To generate ECO scores, we compute multiple sequence alignment (MSA) by running HHBlits [95] against the redundancy reduced UniProt20 database ver. 2015_06 using the default parameters. MSA is used to generate $n \times 20$ matrix of position-specific frequencies p_c where $c = 1, 2, \ldots 20$ represents the 20 AA types, and n is the length of the protein chain. Based on [96], we calculate ECO scores from this matrix as follows:

$$ECO = \frac{\log \sum_{i=1}^{i=20} p_c^2(i)/p_0(i)}{\log \sum_{i=1}^{i=20} p_c(i)/p_0(i)},$$

where i is position of a residue in the sequence, and $p_0(i)$ is the BLOSUM62 background distribution for ith position [97]. We use the hidden Markov model-based scores rather than the position specific scoring matrix (PSSM)-based scores, as they are faster to compute and were shown to provide a better measure of ECO [50].

Selection of physicochemical properties of AAs that explain their propensity for binding

We investigate whether propensities of AAs to bind a given ligand and are related to their physicochemical properties. We use the AAindex database [98] that includes a comprehensive set of 544 indices that quantify physicochemical properties of AAs. First, the 20 AAs were split into two groups: those with large relative difference (enrichment or depletion) between binding and the surface nonbinding residues and the remaining AAs. Details how the corresponding cutoffs were established are given in Supplementary Figure S2. Next, we assess whether the difference in the values of the AAindex between the two sets of residues is statistically significant. We evaluate each index individually, and we assert that a given AAindex potentially explains propensity for binding if the difference is significant. We use the Shapiro–Wilk test to check whether the data are normal, and next we test significance with the Student's t-test for normal data and with the Wilcoxon signed-rank test otherwise. For the indices with the statistically significant difference

(P-value < 0.05), we calculate a correlation between their values and the relative difference values. We use the point-biserial correlation coefficient to quantify correlation between binary and continuous/binary measurements, Spearman's rank correlation coefficient if at least one measurement is discrete and Pearson correlation coefficient (PCC) when both measurements are continuous. One, most appropriate, correlation coefficient is computed for each significant index. We pick the AA index with the highest correlation coefficient to point to the corresponding physicochemical property that best explains propensity for binding to a given ligand type.

Statistical analysis of differences of hallmark values

We compare RAA, RSA and ECO values between residues that bind different types of ligands and nonbinding residues based on their cumulative distributions. We use the Kolmogorov–Smirnov test, a nonparametric test, which quantifies a distance between empirical cumulative distribution functions, to assess whether pairs of cumulative distributions for the same hallmark are statistically different. We assume that two cumulative distributions are significantly different when P-value < 0.001.

Sequence-based prediction of binding residues

Besides evaluating whether each of the three hallmarks can be used to differentiate between binding and nonbinding residues, we investigate whether combining them together would result in a stronger discriminatory power that would be sufficient to accurately identify binding residues. We use a simple linear combination of the three hallmarks:

$$y = w_1 + w_2 \times \mathrm{RAA} + w_3 \times \mathrm{RSA} + w_4 \times \mathrm{ECO},$$

where y is the estimated propensity for binding, and w_1, w_2, w_3 and w_4 are coefficients that are determined empirically from a training data set (training fold in the cross-validation) to minimize error between y and the native annotation of binding. In particular, we use the least squares algorithm [99] implemented in Matlab R2017a to estimate these coefficients. We develop five such functions for DNA binding, RNA binding, protein binding, DNA and protein binding and RNA and protein binding. They are unlikely to overfit our benchmark data sets with over 6 million residues, as they use only four parameters. We perform cross-validation on the benchmark data sets to investigate whether these linear functions could be used to accurately predict propensity for binding for proteins that were not used to determine the four coefficients. Moreover, we also compare scores computed with the linear functions build on our training data sets with current methods that predict DNA-, RNA- and protein-binding residues on the corresponding test data sets.

Evaluation of the sequence-based prediction of binding residues

We empirically evaluate whether the propensity for binding computed based on values of a given hallmark or outputs generated by the scoring functions can accurately discriminate binding from nonbinding residues. We compare the propensity scores with the native annotation of binding by applying commonly used AUC [area under the receiver operator characteristic (ROC) curve] to quantify the predictive performance [38, 41, 42, 46, 73, 88]. The ROC curve is a relation between TPrate (sensitivity) and FPrate (1-specificity) over the entire range of propensity values. $\mathrm{TPrate} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and $\mathrm{FPrate} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$, where TP and TN denote the numbers of correctly predicted native binding and native nonbinding residues, respectively, FN is the number of incorrectly predicted binding residues (native binding residues predicted as nonbinding) and FP is the number of incorrectly predicted nonbinding residues (native nonbinding residues predicted as binding). Inspired by [100], we also introduce two additional measures: AULC (area under the ROC curve for low range of the FPrate values) and AULCratio (ratio of achieved AULC relative to AULC of a method that generates predictions at random). Motivated by the fact that binding residues are infrequent compared with the nonbinding residues, AULC focuses on predictions where the number of false positives is equal or smaller than the number of native binding residues. In other words, AULC is the area under the ROC curve for the low FPrate values that correspond to the FP values ranging between $\mathrm{FP} = 0$ and $\mathrm{FP} = \mathrm{TP} + \mathrm{FN}$ (the number of native positives). This range corresponds to the predictions, where the number of predicted binding residues does not exceed the number of native binding residues. AULCratio > 1 indicates that the corresponding predictor is better than a method that generates random predictions.

We also assess binarized scores that indicate whether a given residue binds. This assessment is based on commonly used $\mathrm{sensitivity} = \mathrm{TPrate} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and $\mathrm{specificity} = 1 - \mathrm{FPrate} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$. Sensitivity quantifies fraction of correctly predicted binding residues among the native binding residues. Specificity measures fraction of correctly predicted nonbinding residues among the native nonbinding residues. To ensure that results of multiple methods can be compared side by side, we compute the sensitivity at a predefined $\mathrm{specificity} = 0.95$. This value of specificity corresponds to the $\mathrm{FPrate} = 0.05$. Finally, we also report values of the Matthews correlation coefficient (MCC):

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

to assess the binarized predictions. $\mathrm{MCC} = 0$ denotes a random prediction, while $\mathrm{MCC} = 1$ corresponds to a perfect prediction.

Comparative empirical analysis of main hallmarks of protein–DNA, protein–RNA and protein–protein binding

Analysis of propensity of AAs for RNA, DNA and protein binding

The propensity is determined with RAA that quantifies relative difference in abundance of a given AA type between binding residues and surface nonbinding residues; positive/negative values correspond to enrichment/depletion. RAA values for residues that bind DNA-, RNA-, proteins- and the multi-ligand-binding residues are given in Table 3. While majority of AAs have low absolute RAA values, several are preferentially enriched or depleted with the relative differences as high as over 200% (2.05 enrichment for Arg in residues that bind both DNA and proteins) and as low as −69% (0.69 depletion for Glu in DNA binding).

The DNA- and RNA-binding residues are enriched in positively charged Lys, Arg and His and depleted in negatively charged Asp and Glu (Table 3). This pattern is because of stabilizing ionic interactions between the positively charged residues and phosphate group of DNA and RNA [9, 11, 50]. The other strong pattern is the enrichment of the aromatic residues (Phe,

Table 3. Propensity of AAs for binding to DNA, RNA, proteins and multi-ligand binding to both DNA and proteins and both RNA and proteins when compared with the surface nonbinding residues

AA	Binding to a single ligand			Multi-ligand binding	
	DNA	RNA	Protein	DNA and protein	RNA and protein
Ala	−0.40	−0.25	−0.08	**−0.54**	−0.40
Arg	**1.40**	**1.33**	0.12	**2.06**	**1.87**
Asn	0.09	−0.10	−0.15	−0.09	−0.22
Asp	**−0.63**	**−0.62**	−0.33	**−0.67**	**−0.64**
Cys	0.06	0.17	**0.76**	0.10	−0.03
Gln	−0.05	−0.17	−0.11	−0.00	−0.17
Glu	**−0.70**	**−0.64**	−0.34	**−0.58**	**−0.61**
Gly	−0.17	0.04	−0.25	−0.41	−0.26
His	**0.52**	**0.54**	0.18	**0.51**	**0.64**
Ile	−0.09	0.11	**0.71**	−0.09	0.16
Leu	−0.40	−0.18	**0.61**	−0.35	−0.10
Lys	**0.49**	**0.51**	−0.38	0.36	**0.55**
Met	**0.40**	**0.55**	**0.92**	0.36	**0.55**
Phe	**0.51**	0.41	**1.18**	**0.64**	**0.48**
Pro	**−0.46**	−0.26	−0.17	−0.44	−0.23
Ser	0.23	−0.14	−0.13	0.04	−0.40
Thr	0.12	−0.10	−0.07	0.21	−0.17
Trp	**1.14**	0.30	**0.95**	**0.95**	**0.77**
Tyr	**0.92**	0.33	**0.71**	**0.71**	**0.67**
Val	−0.27	0.06	0.37	−0.24	−0.10

Note: Positive (negative) value corresponds to enrichment (depletion) of a given AA type in binding residues compared with its occurrence in all residues from the data set. AAs characterized by large values of enrichment (>0.45) and depletion (< −0.45) are shown in bold; details on these thresholds are given in Supplementary Figure S2. The analysis was performed on the benchmark data sets.

Trp and Tyr) that was shown to be related to the π–π stacking interactions in complexes with nucleic acids [101, 102]. RAA values for DNA- and RNA-binding residues are highly correlated, with PCC = 0.91, and the difference between these values lacks statistical significance (Supplementary Table S1). Consequently, this hallmark is unlikely to accurately differentiate between binding to RNA and DNA.

Analysis of results for the protein-binding residues points to the enrichment in several AAs (Table 3), which is in agreement with the literature. Enrichment in Cys was previously observed and was linked to coupling of Cys residues in the protein–protein interfaces [103]. Similarly, enrichment in aromatic residues (Phe, Trp and Tyr) was found in homodimeric protein complexes [104], and preferred coupling between aromatic residues was observed in protein–protein-binding interfaces [105]. The latter study also noted depletion of the charged–charged pairs, particularly for the residues of opposing charge, which could explain the negative values of RAA for Asp, Glu and Lys. Preference of Met, Phe and Trp to be involved in protein–protein binding was shown in [104, 106]. The enrichment of Ile and Leu was pointed to correlate with the increased propensity for formation of energetic protein–protein interaction hot spots [106, 107]. Interestingly, the RAA values for the protein binding are modestly similar to the corresponding RAA values for DNA binding (PCC = 0.43; *P*-value = 0.36) and RNA binding (PCC = 0.42; *P*-value = 0.11) (Supplementary Table S1). This means that RAA can be potentially used to separate protein- and nucleic acid-binding residues.

The RAA values for the multi-ligand-binding residues follow the values for the corresponding DNA- and RNA-binding residues (Table 3). Supplementary Table S1 reveals that similarity is high between the RAA values for the DNA binding and each multi-ligand binding, and between the RAA values for the RNA binding and each multi-ligand binding (PCC ≥ 0.90; *P*-value ≥ 0.87). Analogously high similarity is true between the binding to both DNA and proteins and to both RNA and proteins (PCC = 0.95; *P*-value = 0.97). Interestingly, similarity between the RAA values of the protein binding and the two types of the multi-ligand binding is only modest (PCC ≤ 0.44; *P*-value ≤ 0.19) (Supplementary Table S1). This means that the enrichment in AAs for the multi-ligand-binding residues is driven by the binding to the nucleic acids. We also note the particularly high levels of enrichment of Arg in the multi-ligand-binding residues for both RNA and DNA.

Supplementary Figure S3 contrasts the RAA values with a recently developed set of propensities for the DNA, RNA and protein binding [74]. To the best of our knowledge, there are no existing propensity scales for the multi-ligand-binding residues. The other propensities were also derived empirically from protein–protein and protein–nucleic acid complexes. However, they are based on a smaller set of 153 protein–protein, 81 protein–RNA and 212 protein–DNA complexes versus 17 594, 1040 and 817 that we use, respectively. Moreover, our data set uses arguably more complete set of binding residues that incorporates annotations transferred from multiple protein chains/complexes that are linked to the same UniProt protein. One immediately apparent difference is that the RAA values are both positive (for AA that are enriched among the binding residues) and negative (for the depleted residues), while the other propensities are strictly positive where low (high) values correspond to depletion (enrichment). We argue that the RAA values are easier to interpret, as the value of 0 is a clear breaking point. Moreover, the RAA values and the other three sets of propensities are correlated. PCC between RAA and the propensities from [74] equals 0.77 for the DNA binding, 0.71 for the RNA binding and 0.60 for the protein binding. The lower correlation for the protein binding could be attributed to the fact that the other AA index was derived using solely heterodimers, while we use both homo and heterodimers. Among the AAs with the five lowest and the five highest propensities, seven are in common between the two indices for the RNA binding and five are in common for the DNA and the protein binding.

Using an approach described above, we find statistically significant physicochemical properties of AAs that are the most correlated with the RAA values. These are polarizability (AAindex CHAM820101) for residues that bind DNA and that bind both DNA and proteins, charge (AAindex KLEP840101) for residues that bind RNA and that bind both RNA and proteins and polarity (AAindex RADA880108) for the protein binding. We visualize these selected properties in Supplementary Figure S4. Given the high correlation of their RAA values, we combine charge and polarizability to analyze DNA binding, RNA binding and the two types of multi-ligand-binding residues. The scatter plots demonstrate that these properties separate majority of AAs with high absolute values of RAA (red markers) from the AAs with low absolute values (green markers). We note that the importance of polar residues for DNA and RNA binding was shown in [11, 108] and for protein–protein binding in [106, 107].

Overall, our analysis in the context of the DNA, RNA and protein binding is in agreement with the literature, while we are the first to contrast side by side the propensity of AAs for binding DNA, RNA and proteins and to provide insights for the multi-ligand-binding residues.

Analysis of the three hallmarks of RNA, DNA and protein binding

We analyze differences in the values of the three hallmarks between residues that bind various ligands, residues that do not bind ligands (nonbinding residues) and all residues. The source data that includes values of the three hallmarks are available at http://biomine.cs.vcu.edu/servers/hybridNAP/.

Supplementary Figure S5 shows distributions of the RAA values. The RNA- and DNA-binding residues (Supplementary Figure S5A and B) have substantially higher values of RAA compared with the nonbinding and all residues, while the distributions for the other types of ligand are similar to the distributions for the nonbinding and all residues. This means that RAA values can identify the DNA- and DNA-binding residues in a given protein sequence but not the protein-binding residues.

Supplementary Figure S6 compares distributions of the RSA values. As expected, residues that bind ligands have higher RSA values than the nonbinding residues and the set of all residues (Supplementary Figure S6A–E). Interestingly, the two types of the multi-ligand-binding residues have higher RSA values than the values for DNA-, RNA- and protein-binding residues (Supplementary Figure S6F). This means that residues that bind both proteins and nucleic acids are overall more exposed to solvent. About 12% of all residues and 14% of nonbinding residues have RSA = 0, which means that they are buried in the core of the protein. As expected, the corresponding fractions of buried residues among the binding residues are nearly 0. Bars in Supplementary Figure S6 that represent relative ratio of the fractions of binding and nonbinding residues reveal that a much larger proportion of binding residues is found among residues with RSA > 0.5. Moreover, majority of RNA-, DNA- and protein-binding residues have RSA values ranging between 20 and 70% (Supplementary Figure S6F). The multi-ligand-binding residues are substantially overrepresented among residues with high solvent accessibility (tall white bars on the right-hand side of Supplementary Figure S6D and E) and majority of them attain RSA values in the 30–80% range.

Supplementary Figure S7 compares distributions of the ECO values. Residues that bind RNA-, DNA- and the multi-ligand-binding residues are more conserved then the nonbinding residues (Supplementary Figure S7A and B, D and E). About 60% of these binding residues have ECO > 0.5, while only about 30% of the nonbinding residues have such high conservation scores. Moreover, residues that bind DNA and that bind DNA and proteins have the largest proportion of highly conserved residues (white bars in Supplementary Figure S7A and D for ECO > 0.9). In contrast, conservation of the protein-binding residues is on par with the nonbinding residues (Supplementary Figure S7C). Our results agree with literature. DNA-binding residues were shown to be more conserved than other surface residues [109], while the sequence conservation of the protein-binding residues was found to be similar to the conservation of surface residues [110].

Figure 1 shows cumulative distributions of the values of RAA, RSA and ECO, which are used to analyze statistical significance of differences in the values of these hallmarks between the seven sets of residues (RNA binding, DNA binding, protein binding, RNA and protein binding and DNA and protein binding, nonbinding and all). Distributions for each set of residues are given in Supplementary Figures S8–S10. Table 4 summarizes the analysis of significance, and Supplementary Table S2 provides the corresponding statistics. Figure 1A reveals that RAA values separate the various sets of residues into two clusters: DNA- and RNA-binding residues that have similar and high relative AA propensities for binding; the remaining sets that share lower values of RAA. Correspondingly, Table 4 demonstrates that the RAA values are significantly different (P-value < 0.001) between RNA-binding residues and the other residue sets and between DNA-binding residues and the other residue sets, but not between RNA- and DNA-binding residues and not between the other sets of residues. Figure 1B suggests that use of RSA results in three clusters: all and nonbinding residues; residues that bind to RNA, DNA and proteins; and the two sets of the multi-ligand-binding residues. Table 4 confirms that RSA of all five sets of binding residues is significantly larger than RSA of nonbinding and all residues (P-value < 0.001). RSA values of the multi-ligand-binding residues are also significantly larger when compared with the RNA-binding, DNA-binding and protein-binding residues (P-value < 0.001), while they are not significantly different between the three latter sets of binding residues. Figure 1C shows that values of ECO provide separation into two clusters: all, nonbinding and protein-binding residues versus DNA-, RNA- and multi-ligand-binding residues. Consequently, the conservation scores are not significantly different between protein binding, nonbinding and all residues, and between DNA-, RNA- and multi-ligand-binding residues (Table 4). However, they are significantly different between all residues/nonbinding residues/protein-binding residues and the DNA-binding/RNA-binding/multi-binding residues (P-value < 0.001). Finally, Table 4 reports that none of the considered hallmarks is statistically different between DNA- and RNA-binding residues and between the two types of the multi-ligand-binding residues. Separation of these residues would require other information.

To sum up, we characterize the three hallmarks for the DNA, RNA and protein binding, and we are the first to contrast them side by side, compute and contrast them for multi-ligand-binding residues and assess significance of their differences. Among our novel findings, we empirically demonstrate that the multi-ligand-binding residues are significantly more solvent exposed than the residues that bind to DNA, RNA and proteins and significantly more evolutionary conserved when compared with the protein-binding residues.

Prediction of protein–DNA-, protein–RNA- and protein–protein-binding residues using the three hallmarks of binding

Combining hallmarks leads to improved and accurate discrimination of binding residues

We evaluate and visualize relation between propensity for binding and the values each of the three hallmarks and their combinations. We combine hallmarks using linear regression, which results in the following five linear functions:

$$y_{\text{DNA\_binding}} = -0.4139 + 0.4133 \times \text{RAA} + 0.2097 \times \text{RSA} + 0.3115 \times \text{ECO}$$

$$y_{\text{RNA\_binding}} = -0.4661 + 0.4218 \times \text{RAA} + 0.2788 \times \text{RSA} + 0.3842 \times \text{ECO}$$

$$y_{\text{protein\_binding}} = -0.4932 + 0.4024 \times \text{RAA} + 0.5596 \times \text{RSA} + 0.3018 \times \text{ECO}$$

$$y_{\text{DNA\_and\_protein\_binding}} = -0.5217 + 0.2648 \times \text{RAA} + 0.4539 \times \text{RSA} + 0.4527 \times \text{ECO}$$

$$y_{\text{RNA\_and\_protein\_binding}} = -0.5066 + 0.2436 \times \text{RAA} + 0.4642 \times \text{RSA} + 0.3891 \times \text{ECO}$$

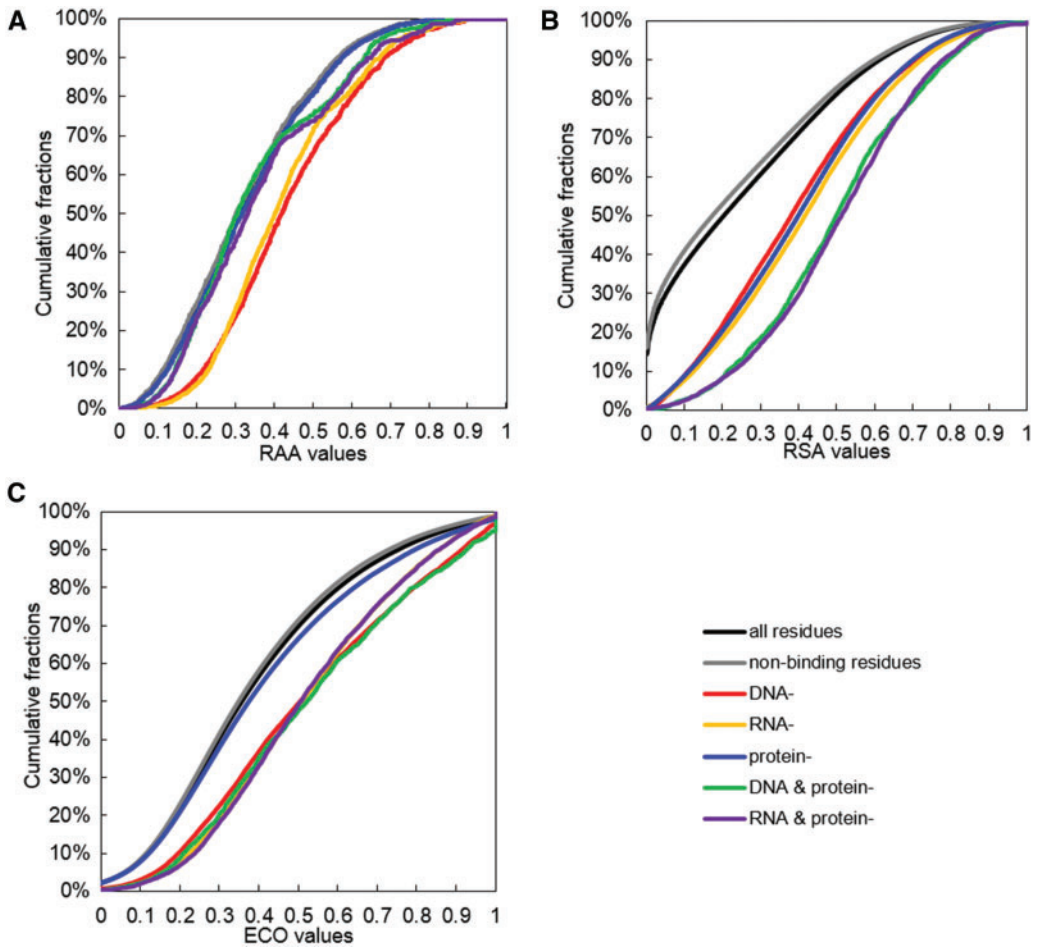

Figure 1. Cumulative distributions of the values of RAA (**A**), RSA (**B**) and ECP (**C**) on the corresponding benchmark data sets. The y-axis shows the fraction of the values below the number shown on the x-axis. Cumulative distributions for individual types of ligands are provided in the Supplementary Figures S8 (for Figure 1A), S9 (for Figure 1B) and S10 (for Figure 1C) .

Table 4. Hallmarks of binding that offer statistically significant differences (P-value < 0.001) for a given combination of residue sets defined in the table head and the first row

	All residues	Nonbinding	DNA binding	RNA binding	Protein binding	DNA and protein binding	RNA and protein binding
All residues	N/A		RAA RSA ECO	RAA RSA ECO	RSA	RSA ECO	RSA ECO
Nonbinding		N/A	RAA RSA ECO	RAA RSA ECO	RSA	RSA ECO	RSA ECO
DNA binding			N/A		RAA ECO	RAA RSA	RAA RSA
RNA binding				N/A	RAA ECO	RAA RSA	RAA RSA
Protein binding					N/A	RSA ECO	RSA ECO
DNA and protein binding						N/A	
RNA and protein binding							N/A

Note: For instance, 'RSA ECO' in the right top corner means that differences in RSA and ECO between RNA and protein-binding residues (column) are significant when compared with all residues (row). The analysis was performed on the benchmark data sets.

All coefficients are positive. This means that higher values of RAA, RSA and ECO are associated with higher propensity for binding. Differences in values of coefficients between functions reflect relative predictive value of the corresponding hallmarks. The highest coefficients for the RNA and DNA binding are for RAA, while for the protein, DNA and protein and RNA and protein binding the highest are for RSA. This agrees with Figure 1 where these hallmarks are shown to provide the best separation between the nonbinding and the corresponding binding residues.

Supplementary Table S3 compares predictive quality of the propensities for binding computed using RAA, RSA, ECO and their regression-based combinations when compared with the native annotations of binding; we report results from cross-validation on the benchmark data set and results where modeling was performed on the whole benchmark data set. The

corresponding putative propensities for binding on the benchmark data set are available at http://biomine.cs.vcu.edu/servers/hybridNAP/. RAA provides strong predictive quality for the DNA, RNA and multi-ligand-binding residues ($AULC_{ratio} > 2.9$) and performs poorly for protein binding ($AULC_{ratio} < 1.8$). The sensitivity (TPrate) for the prediction of DNA, RNA and both multi-ligand-binding residues is substantially higher (between 0.13 and 0.17) compared with the FPrate that is set to equal 0.05. RSA provides relatively accurate predictions for both types of multi-ligand-binding residues ($AULC_{ratio} > 3$, and TPrate > 0.14 at FPrate = 0.05), but its predictive quality for the RNA-, DNA- and protein-binding residues is lower ($AULC_{ratio} < 2.7$). ECO is shown to be useful to find the DNA-binding and DNA- and protein-binding residues ($AULC_{ratio} > 3$, and TPrate > 0.14 at FPrate = 0.05). Overall, residues that bind nucleic acids are easier to detect with the three hallmarks, while protein-binding residues are more elusive. However, when the hallmarks are combined together protein-binding residues can be predicted with reasonably high levels of predictive quality ($AULC_{ratio} = 3.83$, and TPrate = 0.18 at FPrate = 0.05). Overall, we found that combining two or more hallmarks using our simple linear functions results in a large increase of predictive performance for each type of binding when compared with the use of individual markers of binding. The DNA- and RNA-binding residues and the multi-ligand-binding residues can be predicted with $AULC_{ratio} > 5.8$, AUC between 0.75 and 0.82 and TPrate between 0.25 and 0.36 at FPrate = 0.05, depending on the ligand type (Supplementary Table S3). The results of experiments based on cross-validation are virtually identical with the modeling using the whole benchmark data set, which means that as expected our linear functions do not overtif the benchmark data sets. This is because our models rely on just four parameters (coefficients in the linear function). To sum up, our empirical results demonstrate that while some of the hallmarks do not provide sufficient discriminatory power to accurately predict some types of binding residues (e.g. RAA to predict protein-binding residues), a simple linear combination of these hallmarks provides accurate predictions across the five types of binding residues.

Figure 2 visualizes relation between values of the three hallmarks and native annotations of DNA binding expressed with ratio of fraction of the binding residues to fraction of nonbinding residues for a given range of RAA, RSA and ECO values. Results for the other types of binding are given in Supplementary Figure S11. In the 3D space defined by the values of RAA, RSA and ECO, outputs of the linear functions for $y > 0$ and $y < 0$ correspond to two subspaces, where the residues are predicted to bind and not to bind, respectively. The plane shown in Figure 2 is a boundary between these two subspaces and is defined by the values of the three markers where $y_{DNA\_binding}=0$. The ball-shaped points that are color coded to represent ratios of binding versus nonbinding residues [darker green (red) denotes higher ratio of binding (nonbinding) residues] reveal that the native annotations of binding agree with the placement of the plane that is based on propensity for binding predicted from the three hallmarks. The green (red) points are primarily above (below) the plane, which means that the three hallmarks can be used to separate binding from nonbinding residues. Importantly, Figure 2 also illustrates how combining these hallmarks results in the improved separation of the two types of residues. For instance, residues with RAA <0.5 would be categorized as nonbinding if only this marker would be used; this stems from the one-dimensional (1D) plot shown in front of the RAA axis. However, some of these residues are likely to bind DNA,

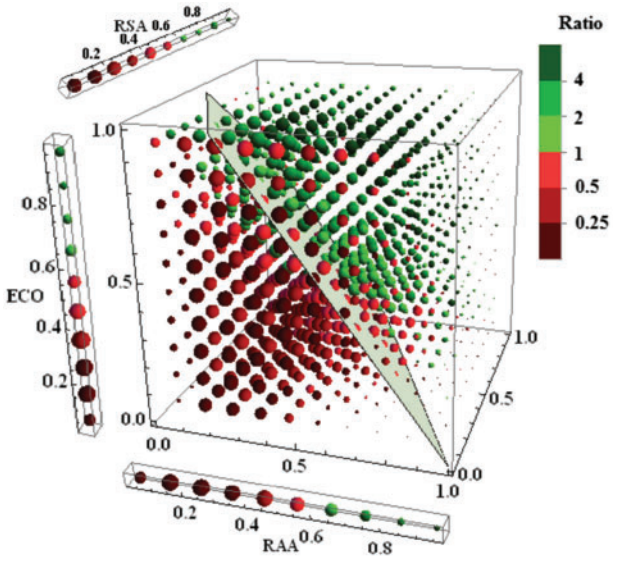

Figure 2. Likelihood of binding residues in 3D space defined by the three hallmarks of binding and in the three 1D spaces defined by each hallmark (1D plots at the edges of the 3D plot). Likelihood is estimated with ratio of fraction of binding residues to fraction of nonbinding residues for a given range of RAA, RSA and ECO values. Darker green (red) ball-shaped points denote higher ratio of binding (nonbinding) residues; their size corresponds to number of residues. Plane separates the 3D space into two subspaces for $y > 0$ and $y < 0$ that correspond to values of hallmarks for which residues are predicted to bind and not to bind, respectively. Code to recreate this figure as an interactive plot is available at http://biomine.cs.vcu.edu/servers/hybridNAP/ under the Datasets and Supporting Information link.

in particular these with high conservation and RNA values (top, left and far corner of the 3D plot), and they can be found using the 3D model.

Hallmarks computed from sequence provide accurate prediction of binding residues

While RAA and ECO are computed from the protein sequences, RSA is obtained from the structures, thus constraining results of our analysis to the structurally solved proteins. Given the rapid growth of the protein sequence space and the fact that majority of proteins lack structural coverage [20, 111], we analyze hallmarks computed solely from the sequences. To this end, we use putative RSA values predicted from the sequences with ASAquick [94]. This predictor was empirically assessed to generate putative RSA with mean absolute error (MAE) of 11% and PCC with the native RSA of 0.65 ± 0.1 using an independent (from the data used to develop this method) benchmark data set. Evaluation of predictions from this method on our benchmark data set shows that it is MAE = 11.4% and PCC = 0.64 ± 0.05. This slightly lower than originally estimated predictive quality suggests that predictions generated by ASAquick do not overfit our data set.

Figure 3 quantifies differences in the predictive performance when the native RSA is replaced by the putative RSA (source data are available in Supplementary Tables S3 and S4). The corresponding values of the putative RSA and propensities for binding computed using our regressions that use the putative values of RSA on the benchmark data set are available at http://biomine.cs.vcu.edu/servers/hybridNAP/. Supplementary Figure S12 shows side by side the ROC curves that correspond to models that use native versus putative RSA values. As expected, the

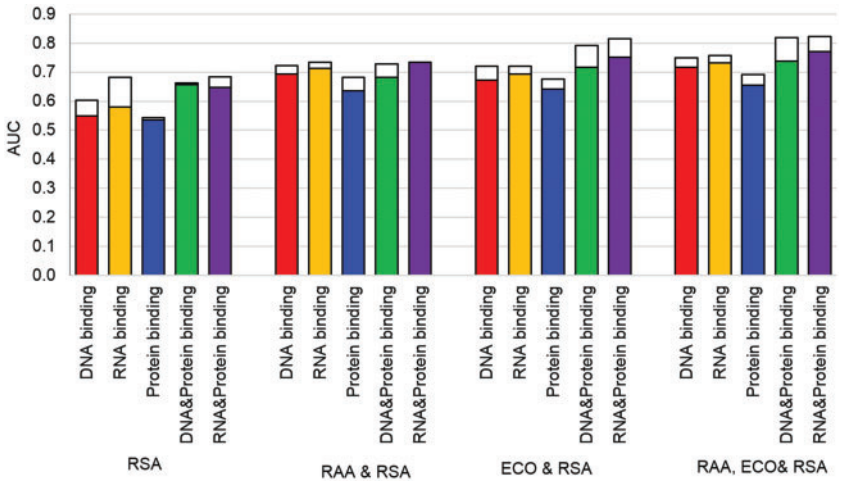

Figure 3. The loss of predictive performance measured with AUC when using putative RSA predicted from the sequence instead of the native RSA computed from the structure. The results include the use of putative RSA alone and the three regression models that combine putative RSA with the other two hallmarks of binding. Solid and color coded by the type of binding bars show AUC when using the putative RSA, while hollow bars when using the native RSA. The analysis was performed based on the cross-validation on the benchmark data sets. Source values are available in Supplementary Tables S3 and S4.

use of putative values results in a reduction of predictive quality. Our results reveal that the magnitude of this reduction is modest. Importantly, combining putative RSA with the other two hallmarks leads to models that provide accurate predictions of binding residues. The DNA-, RNA- and multi-ligand-binding residues are predicted with $AULC_{ratio} > 5$, $AUC > 0.72$ and $TPrate > 0.22$ at $FPrate = 0.05$. The prediction of the protein-binding residues is less accurate (similar to when using the native RSA) but still provides useful clues to find these residues in the input protein chain ($AULC_{ratio} = 3.3$, $AUC = 0.66$ and $TPrate = 0.15$ at $FPrate = 0.05$). Our results agree with the fact that these hallmarks are often used to build sequence-based predictors of protein- and nucleic acid-binding residues (Table 1). The results also reveal that accurate prediction of the five types of binding residues solely from the protein sequence is possible with the use of the three hallmarks.

Motivated by the work in [112], we empirically investigate whether and how the predictive performance varies when applying the best regression models that rely on the three hallmarks and putative RSA to predict proteins that carry out specific functions. We annotate molecular functions for proteins in the benchmark data set using gene ontology (GO) terms that we collect from the UniProt resource. To ensure that the sample size is sufficiently large to obtain robust estimates, we evaluate predictive performance for the molecular functions that include at least 30 proteins. Supplementary Table S5 shows AUC and $AULC_{ratio}$ values for the resulting 8, 8, 31, 6 and 7 functions for the proteins involved in DNA, RNA, protein, DNA and protein as well as RNA and protein binding, respectively. Both AUC and $AULC_{ratio}$ values are significantly better than random ($AUC > 0.5$ and $AULC_{ratio} > 1$) for all considered functions. $AULC_{ratio}$ exceed 3 for the functions of proteins that interact with DNA, RNA and with both DNA and proteins, and is no smaller than 1.99 for the protein- and RNA and protein-binding proteins. Binding residues for DNA repressors, DNA activators, hydrolases and transferases are predicted with $AULC_{ratio}$ surpassing 5.

HybridNAP predictor of binding residues

The five scoring functions that are based on the three hallmarks of binding are deployed as a Web server called hybridNAP (hybrid prediction of Nucleic Acids and Protein binding). A flowchart that explains how predictions are performed is visualized in Figure 4. The Web server is freely available at http://biomine.cs.vcu.edu/servers/hybridNAP/. It accepts queries with up to 10 FASTA-formatted protein sequences and provides real-valued propensities for RNA, DNA, protein, RNA and protein as well as DNA and protein binding for each residue in the submitted sequences. This is the first method that provides predictions of multi-ligand-binding residues. HybridNAP is also the first to simultaneously provide predictions of DNA-, RNA- and protein-binding residues. Results are stored in a parsable text file, which is archived for at least 1 month on the server and which can be accessed via URL provided in the browser window and sent to a user-provided email address.

We empirically compare hybridNAP with a selection of representative methods for the prediction of DNA-, RNA- and protein-binding residues on the corresponding three test data set: DNA_T, RNA_T and protein_T. We could not identify any methods that predict multi-ligand-binding residues. As hybridNAP is available as a Web server, we select methods that are also available as Web servers, and we use these Web servers to collect their predictions. We include the methods that were assessed in the recent comparative review of predictors of DNA- and RNA-binding residues, where the same selection criteria were applied [88]. They include BindN+ [24], DBS-PSSM [113] and DP-Bind(klr) [23] for the prediction of DNA-binding residues and BindN+ [24], RNABindR [40] and Pprint [37] for the prediction of RNA-binding residues. We also compare the predictions of protein-binding generated by hybridNAP with two representative methods that predict protein-binding residues: SPRINGS [83] and PSIVER [58]. The hybridNAP models were generated using the training30_DNA, training30_RNA and training30_protein data sets that include proteins for which the PDB chains used to annotate them share <30% similarity with the DNA_T, RNA_T and protein_T test data sets, respectively. The three training data sets and the predictions from all considered methods, including hybridNAP, on the three test data sets are available at http://biomine.cs.vcu.edu/servers/hybridNAP/.

The results in Figure 5A and Supplementary Table S6 show that hybridNAP secures $AUC = 0.69$, $TPrate = 0.17$ at $FPrate = 0.05$, $MCC = 0.19$ and $AULC_{ratio} = 3.4$ for the prediction of

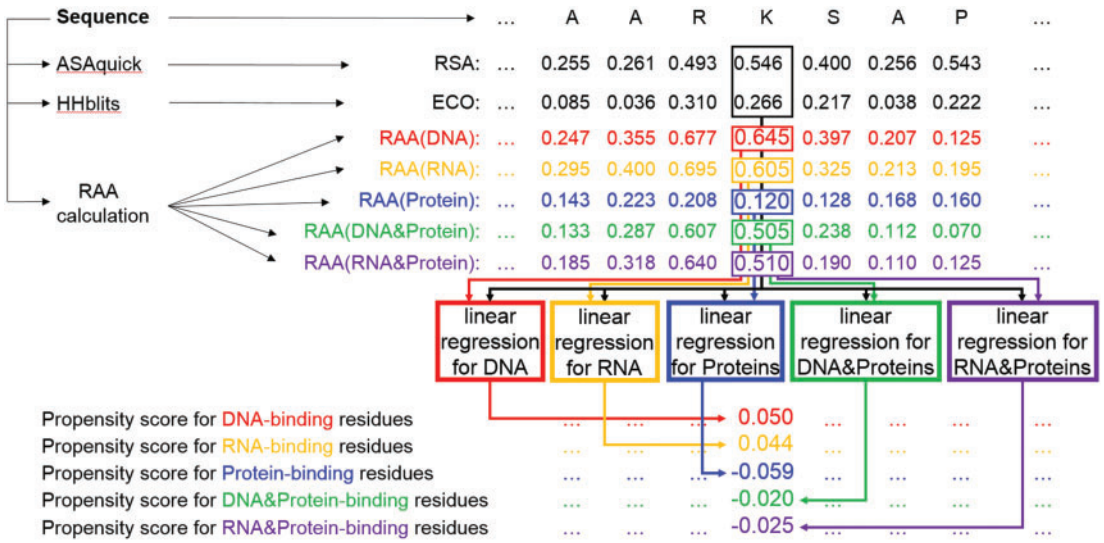

Figure 4. A flowchart that summarizes the predictions with the hybridNAP method. We color coded the five predictive models that that are integrated into hybridNAP that focus on the prediction of DNA binding (red), RNA binding (yellow), protein binding (blue), DNA and protein binding (green) and RNA and protein binding (violet).

DNA-binding residues on the DNA_T data set. These results are lower when compared with the other methods. DP-Bind, DBS-PSSM and BindN+ obtain AUCs at about 0.8, TPrates between 0.28 and 0.30 at FPrate = 0.05 and MCCs at about 0.3. However, their $AULC_{ratios}$ are more similar to hybridNAP. This is also evident in Figure 5A, where the ROC curve of hybridNAP is relatively close to the curves for the other three methods for the low FPR values. This part of the curve is arguably more practical given that this is where the number of false positives (nonbinding residues predicted as DNA binding) is relatively low and is kept below the number of native DNA-binding residues. Figure 5B and Supplementary Table S6 provide results for the prediction of RNA-binding residues on the RNA_T data set. HybridNAP's results are similar to Pprint and somewhat lower than the results of BindN+ and RNABindR. HybridNAP's AUC = 0.67 and MCC = 0.15 compared with AUCs and MCCs of the other methods that range between 0.68 and 0.74 and between 0.19 and 0.23, respectively. Similarly, hybridNAP's $AULC_{ratio} = 4.2$ as opposed to the corresponding values for the other methods that are between 4.4 and 6.1. Finally, Figure 5C and Supplementary Table S6 summarize results for the prediction of protein-binding residues on the protein_T data set. The predictive quality of hybridNAP is similar to PSIVER and lower than SPRINGS. The three methods provide modest levels of predictive performance with AUCs between 0.59 and 0.63, MCCs between 0.1 and 0.14 and $AULC_{ratio}$ between 1.6 and 2.3.

We observe that the results of hybridNAP on the three test data sets (Supplementary Table S6) are lower than the results on the benchmark data set (Supplementary Table S4), particularly for the RNA and protein binding. To compare, AUC = 0.69 versus 0.72 for the DNA binding, 0.67 versus 0.73 for the RNA binding and 0.59 versus 0.66 for the protein binding. We also compared results of the representative predictors of DNA-binding residues (DP-Bind that secures largest values of AUC and $AULC_{ratio}$ in Supplementary Table S6), RNA-binding residues (RNABindR that secures similar predictive performance to the best performing BindN+ that was unavailable at the time of these tests) and protein binding (SPRINGS that secures largest values of AUC and $AULC_{ratio}$ in Supplementary Table S6) between the test data sets and the corresponding benchmark data sets (last row in the Supplementary Table S4). Similarly as for hybridNAP, these results on the test data sets are lower when compared with the results on the benchmark data sets for the RNA and protein binding, i.e. AUC = 0.72 versus 0.80 for RNABindR and AUC = 0.61 versus 0.63 for SPRINGS. The reason is that these test data sets include only the proteins that bind a given ligand type, e.g. the RNA_T test data set consists solely of the RNA-binding proteins. These data sets do not include relatively easier to predict nonbinding residues in the proteins that do not bind this ligand. This is in contrast to the benchmark data set that incorporates a much broader population of proteins which interact with a variety of ligands.

Overall, our empirical analysis demonstrates that hybridNAP offers modestly accurate results. Although they are not as good as the results produced by some of the current predictors, hybridNAP's predictions are sufficiently accurate to offer practical insights. This is especially evident, given the relatively high $AULC_{ratio}$ values for the prediction of DNA- and RNA-binding residues. HybridNAP's TPrates = 0.17 for the DNA binding, = 0.18 for the RNA binding and = 0.1 for the protein binding (Supplementary Table S6) are between 2 and 3.6 times higher than the corresponding FPrate = 0.05. Moreover, Figure 5 reveals that HybridNAP secures TPrate = 0.3 at the FPrate = 0.11 for the DNA binding, = 0.12 for the RNA binding and = 0.18 for the protein binding. This means that these results are substantially better than random. The lower predictive performance of hybridNAP when compared with the other methods is not surprising, as the other methods use sophisticated predictive models (neural networks and support vector machines) and a large number of predictive inputs [58, 83, 88]. This is in contrast to hybridNAP that uses a simple regression with just the three inputs. Importantly, our intention is not to provide the most accurate method but to demonstrate that the three hallmarks computed from the protein sequence are sufficient to provide practical levels predictive performance. Moreover, hybridNAP is the only method that simultaneously provides predictions of DNA-, RNA- and protein-binding residues. This is substantially more convenient and runtime efficient than having to use multiple predictors. Importantly, our Web server is also the first to predict multi-ligand-binding residues. Finally, our methods and the reported predictive performance can be also used as a benchmark in future works that will develop more advanced and more accurate predictors.

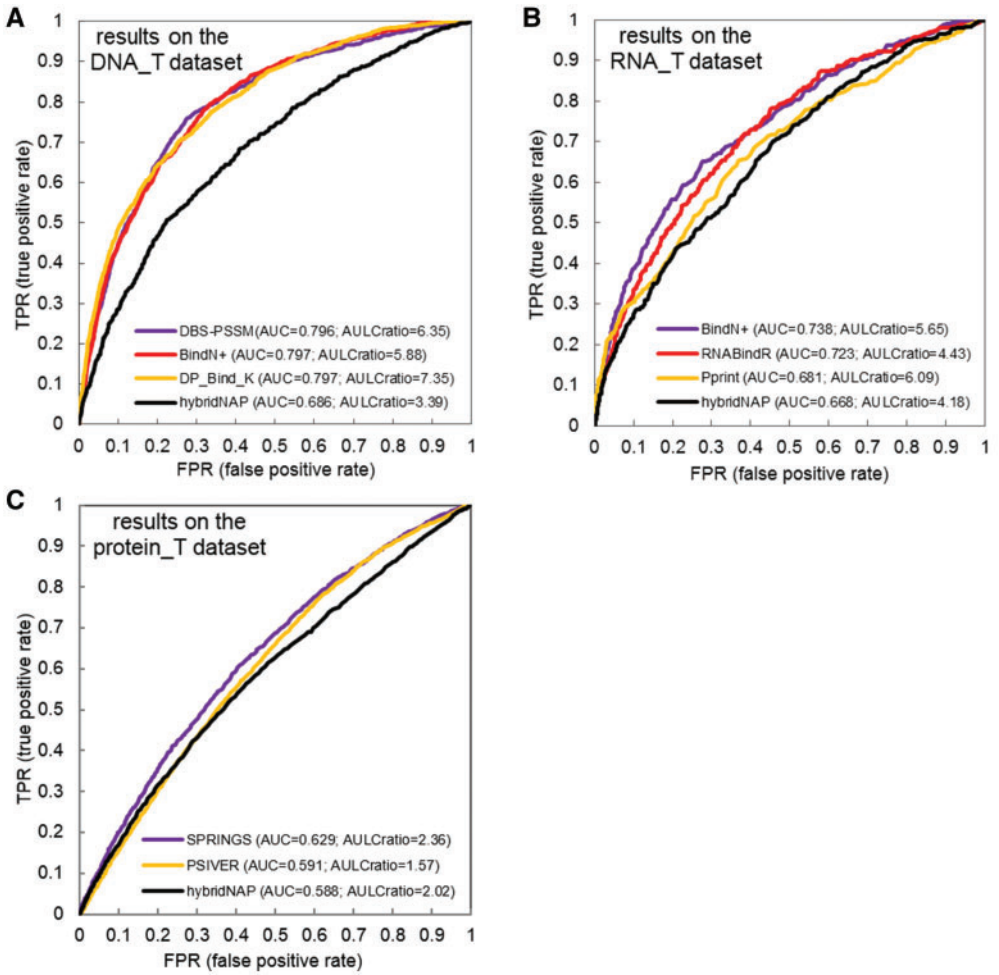

Figure 5. ROC curves for hybridNAP and the existing predictors of DNA-binding residues on the DNA_T test data set (47 proteins) (**A**), predictors of RNA-binding residues on the RNA_T test data set (17 proteins) (**B**) and predictors of protein-binding residues on the protein_T test data set (72 proteins) (**C**). The corresponding values of AUC and AULCratio are given in the figure legends.

Two characteristics that can potentially affect predictive quality of a given model that is measured on a test data set include similarity between proteins in the training and test data sets and the size of the training data set. The predictors that we consider in our empirical evaluation use training proteins that share <30% similarity with the proteins in the test data sets [58, 83, 88]. This is in agreement with the similarity for our training data sets. Next, we empirically investigate whether the size of the data sets used to train our regression-based models impacts the predictive performance. We reduce the size of our training data set to be the same as the size of the smallest data sets that were used to train the corresponding predictors of DNA-, RNA- and protein-binding residues. We select at random 62 proteins from training30_DNA to train our model for the prediction of DNA-binding residues, as the training data sets of BindN+, DBS-PSSM and DP-Bind are the same and include 62 proteins (we call this data set training30small_DNA). Similarly, we select at random 86 proteins from training30_RNA to train our model for the prediction of RNA-binding residues, as the training data sets of BindN+, RNABindR and Pprint have 86, 147 and 174 proteins, respectively (training30small_RNA data set). Finally, we randomly pick 186 proteins from training30_protein to build the predictor of protein-binding residues because both PSIVER and SPRINGS use training data set with this number of proteins (training30small_protein data set). We randomize the selection of the training proteins 10 times (i.e. we create 10 training30small_DNA, 10 training30small_RNA and 10 training30small_protein data sets) to build 10 corresponding predictors. These data sets are available at http://biomine.cs.vcu.edu/servers/hybridNAP/. Supplementary Table S6 reports the averages and SDs of the corresponding ten tests for the prediction of DNA-, RNA- and protein-binding residues for hybridNAP. The results demonstrate that the predictive quality of hybridNAP is similar irrespective of the sizes of the training data sets. To compare, AUC values ($AULC_{ratio}$ values) when using the complete versus small training data set are 0.686 versus 0.685 (3.39 versus 3.31) for DNA binding, 0.668 versus 0.668 (4.18 versus 4.19) for RNA binding and 0.588 versus 0.588 (2.02 versus 2.00) for protein binding. Similarly, the MCC values are virtually identical when comparing the two sets of results side by side. This analysis suggests that the predictive performance of hybridNAP can be compared side by side with the results of the other methods on the test data sets that we use.

We also study whether the measured predictive quality is affected by the imbalanced nature of the test data sets in which the majority of residues are nonbinding. We undersample the native nonbinding residues at random to match their number to the number of native nonbinding residues. The randomized

undersampling is repeated 10 times to create 10 balanced test data sets. Supplementary Table S7 reports averages and SDs of the corresponding 10 tests on the balanced DNA_T, RNA_T and protein_T test data sets. The results are similar to the results on the original test data sets (Supplementary Table S6) for all considered predictors including hybridNAP. For instance, AUCs of hybridNAP for the prediction of DNA-binding residues on the balanced versus imbalanced data set are 0.686 versus 0.686; for the prediction of RNA binding, they are 0.668 versus 0.672; and for the prediction of protein binding, they are 0.588 versus 0.589. Similarly, the corresponding AUCs of BindN+ are 0.794 versus 0.797 for the prediction of DNA-binding residues, and 0.731 versus 0.738 for the prediction of RNA-binding residues. One exception is the MCC values that are higher for the balanced test data set. For instance, hybridNAP's MCCs are 0.31 for the balanced DNA-binding test data set versus 0.19 for the original DNA-binding test data set, 0.26 versus 0.15 for the RNA binding and 0.14 versus 0.11 for the protein binding. We observe similar increases for the other methods. These differences stem from the fact that the predictions on the original data sets include a much larger number of false positives compared with the predictions on the balanced data set, while the number of true positives remains the same. We argue that the higher MCCs on the balanced data set could be misleading, as these predictors will be ultimately applied on full protein chains that are imbalanced. This analysis reveals that the majority of the used measures of the predictive quality can be used to accurately assess results on the original and balanced test data sets.

We also compare hybridNAP with regressions that use of one or two hallmarks of binding on the same three test data sets (Supplementary Table S8). This comparison reveals that use of the three hallmarks provides improved predictive performance when contrasted with the use of two or a single hallmark. The $AULC_{ratio}$ (MCC) improves from 2.96 to 3.39 (from 0.16 to 0.19) for the prediction of DNA-binding residues, from 2.44 to 4.18 (from 0.12 to 0.15) for the RNA-binding residues and from 1.37 to 2.02 (from 0.07 to 0.11) for the protein-binding residues, when comparing the best single hallmark with use of hybridNAP. Similarly, when FPrate = 0.05, the TPrate (sensitivity) goes up from 0.135 to 0.167, from 0.131 to 0.176 and from 0.066 to 0.096 for the prediction of DNA-, RNA- and protein-binding residues, respectively (Supplementary Table S8). This demonstrates that the three hallmarks provide complementary information for the sequence-based prediction of the nucleic acid-binding residues.

To summarize, our analysis reveals that combining the three hallmarks provides accurate prediction of DNA- and RNA-binding residues. We also note that hybridNAP is the first approach that provides prediction of multi-ligand-binding residues. Moreover, its predictions can be easily linked to the underlying hallmarks of binding, providing easy-to-understand interpretation of the results. The latter means that the hybridNAP's users would not only learn whether a given residue is likely to bind DNA, RNA and/or proteins, but (s)he would also learn about underlying factors that contribute to this prediction, such as specific levels of conservation, RSA and/or RAA. We explore this further in the next section.

Use of unexplored hallmarks leads to improved predictive performance for the current predictors

Many of the current predictors do not use some of the three hallmarks. Among the predictors that we compared with, BindN+, DBS-PSSM and DP-Bind do not use RSA and RAA values, and RNABindR does not take advantage of any of the hallmarks, while Pprint and PSIVER do not use RSA and RAA, respectively. We empirically investigate whether addition of the missing hallmarks would improve their predictive performance. We follow two rules of thumb to augment the original propensities output by these methods. First, as binding is unlikely for buried residues, we set the new propensity = (original_propensity + RSA)/2 for the residues with the putative RSA ≤ 0.1. This effectively lowers the resulting propensity for these residues, given how low the RSA values are. The second rule relies on an assertion that residues that are highly conserved or have high RAA value are more likely to bind, and thus, their propensity should be amplified. Thus, among the AAs on the surface (with putative RSA > 0.1), we increase the binding propensity for the residues that have ECO or RAA values in the top 5%. More specifically, we use ECO and RAA values that are normalized to the unit range and compute the new propensity = (original_propensity + ECO or RAA)/2. Figure 6 compares predictive performance measured with AUC and $AULC_{ratio}$ between the original predictors (black bars) with the predictors that were augmented using these two rules (gray bars). Figure 6A shows that the use of the two simple rules improves the values of $AULC_{ratio}$ for all considered methods. This means that the hallmarks can be used to more accurately find binding residues among the predictions with high propensities (when FPrate values are low). Figure 6B shows that the overall AUC values also improve by a small margin across all considered predictors. We hypothesize that this margin could be further increased if the missing hallmarks were used as inputs to optimize the predictive models instead of being used to postprocess the predictions.

Case studies

We visualize predictions from hybridNAP for two proteins, one that binds DNA and one that binds RNA and proteins. The AUCs of the corresponding hybridNAP's predictions are similar to the average values based on the cross-validation. We also contrast hybridNAP's predictions with the predictions from BindN+.

The native DNA-binding residues are at the N terminus of the transcriptional repressor protein (Figure 7A; hollow boxes below the horizontal axis+). The hybridNAP's scoring function finds majority of these binding residues (black crosses), and its predictions generally agree with the outputs of BindN+ (blue crosses). Driven by the simplicity of its model, a substantial benefit of hybridNAP is availability of values of the three hallmarks that suggest the underlying factors that explain the prediction. Generally, we note that if at least one hallmark has high values (highlighted with solid green markers), the corresponding residue is likely to bind. The end users of the hybridNAP's Web server can take advantage of the same annotations of high (green) and low (red) values of hallmarks that are available among Web server's outputs generated online. Predictions of RNA- and protein-binding residues for the ribosomal protein S1A are shown in Figure 7B. The native binding residues are distributed oven the entire chain and include six multi-ligand-binding residues that bind both RNA and proteins (solid black boxes below the horizontal axis). BindN+ and hybridNAP are relatively successful in finding the native RNA-binding residues (hollow boxes below the horizontal axis); the former method finds fewer binding residues but at a lower false-positive rate. The hybridNAP method predicts protein-binding residues in three of four clusters of native protein-binding residues (solid gray boxes below the horizontal axis). It also correctly predicts two residues that bind both RNA and proteins, for two more its

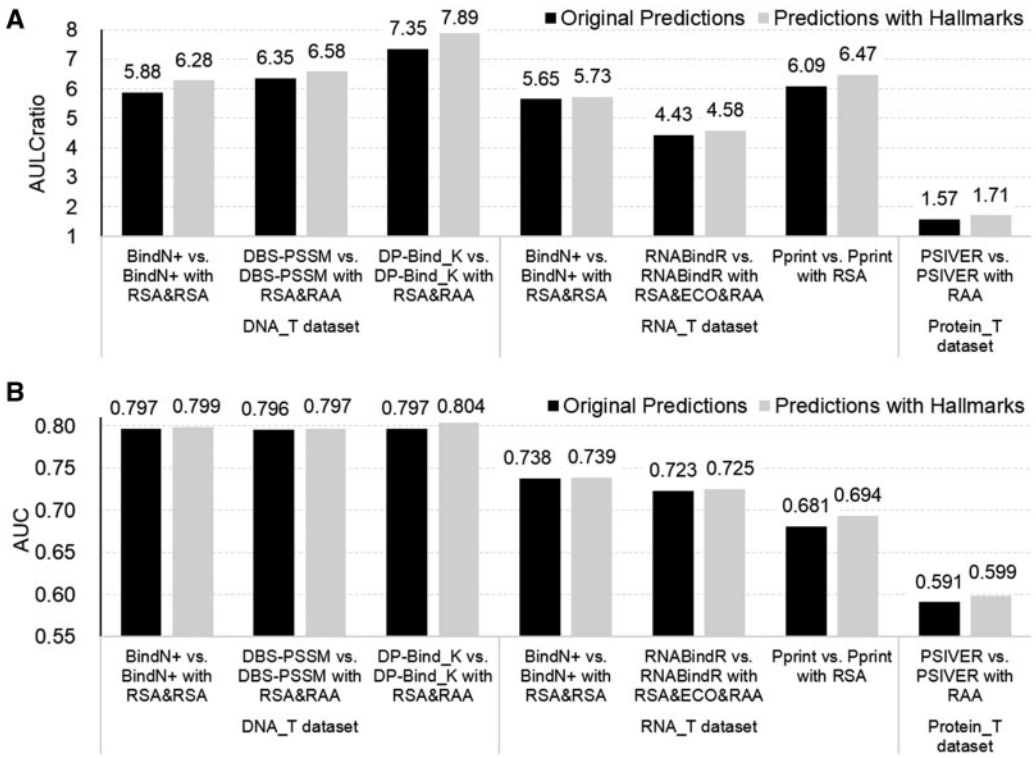

Figure 6. Comparison of the predictive performance of the existing predictors of DNA- or RNA- or protein-binding residues (black bars) with these methods augmented with the use of hallmarks (gray bars) on the DNA_T/RNA_T/protein_T test data sets. (**A** and **B**) show the values of AULCratio and AUC, respectively. The outputs of predictors are improved with the hallmarks that they do not use. BindN+, DBS-PSSM and DP-Bind are augmented with the use RSA and RAA values; RNABindR is improved with the use of RSA, ECO and RAA; Pprint and PSIVER are enhanced with the use of RSA and RAA, respectively.

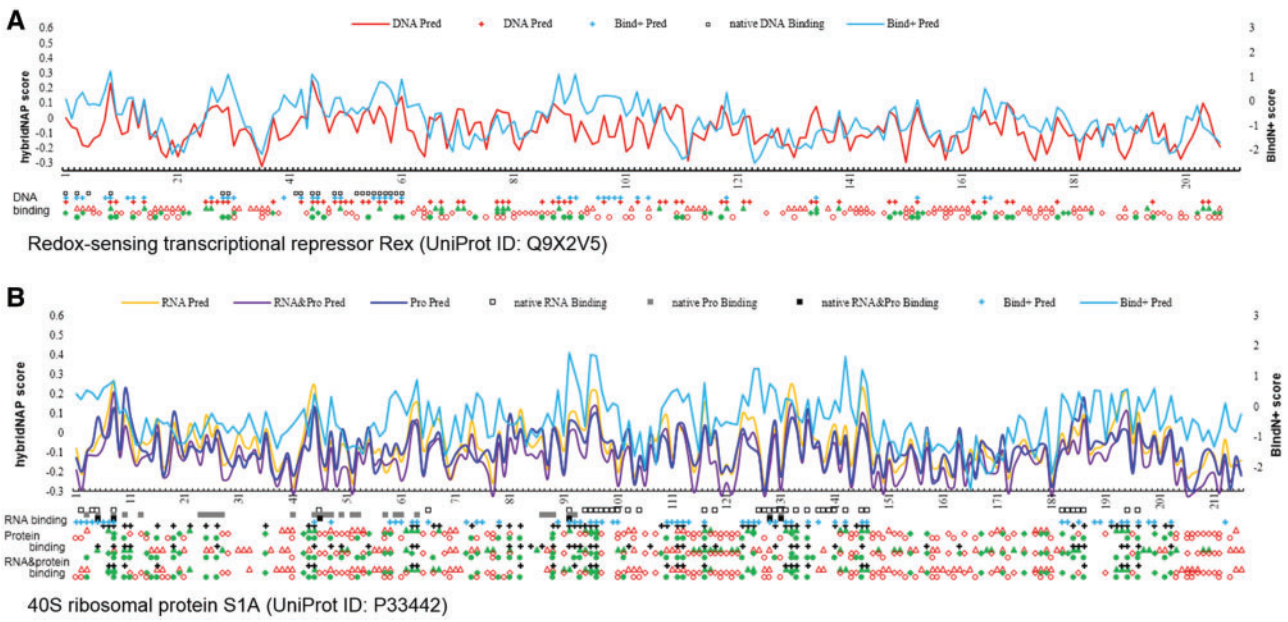

Figure 7. Case studies that illustrate scores generated by hybridNAP and compare them with the native annotations of binding and predictions from BindN+. (**A**) shows results for the DNA-binding Redox-sensing transcriptional repressor Rex protein (UniProt ID: Q9X2V5). (**B**) is for the RNA- and protein-binding 40S ribosomal protein S1A (Uniprot ID: P33442). The x-axis represents the protein sequence. The top part of the plot shows the scores generated by hybridNAP (red, yellow, dark blue and purple lines for the prediction of DNA binding, RNA binding, protein binding and RNA and protein binding) and by BindN+ (light blue line). The markers underneath the x-axis line provide annotations of native and predictive binding residues. In (**A**), one cluster of three lines of markers shows results for DNA binding. In (**B**), there are three clusters of three lines for the RNA, protein and RNA and protein binding. Squares denote the native binding residues where gray squares with black borders are for DNA binding, hollow squares for RNA binding, solid gray for protein binding and solid black for residues that bind both RNA and proteins. Crosses denote predicted binding residues, in black for hybridNAP and in blue for BindN+. Triangles, diamonds and circles represent values of the three hallmarks: RAA, RSA and ECO, respectively. Solid green markers indicate high values of these hallmarks, and while hollow red markers represent low values.

predictions are just one residue away from the native annotations, and for the last two its predictions are two and three residues away. Overall, hybridNAP's predictions are on par with BindN+ for the prediction of RNA- and DNA-binding residues. However, hybridNAP also predicts protein-binding residues as well as multi-ligand-binding residues, and has the advantage of informing end users of the factor(s) that drive its predictions.

Conclusions

We substantially expand our recent related studies that concern prediction of protein–nucleic acid interactions [71, 88] or protein–protein interactions [87], use smaller data sets and focus on the assessment of predictions instead of characterizing binding residues. Here, we review a comprehensive set of over 50 recent studies that predict and/or characterize protein–nucleic acids and protein–protein interactions with the goals of analyzing main hallmarks of binding using a large data set and building a first-of-its-kind method that predicts RNA-, DNA- and protein- binding residues.

We found that most of the 50 surveyed studies concern either protein–DNA or protein–RNA binding, while only a few consider both protein–protein and protein–nucleic acid binding. While different works may contemplate different characteristics that can be derived from protein sequences and structures to investigate and predict binding residues, they generally agree on three main hallmarks of binding: ECO, RSA and propensity of AAs for binding (RAA). We also found that these articles share a few deficiencies. They use relative small data sets with incomplete annotations of binding, typically focus on one type of ligand and rarely compare between RNA-, DNA- and protein-binding, and never studied the multi-ligand-binding residues.

Motivated by the conclusions from the review, we present a comprehensive and large-scale comparative analysis of propensity of residues to bind proteins, RNAs, DNA, proteins and RNAs and proteins and DNA in the context of their RSA, ECO and RAA values. We use substantially larger data sets with markedly more complete annotations of binding when compared with the prior studies. Our analysis suggests that propensities of AAs for binding depend on the ligand type, and in case of the multi-ligand-binding residues, they are driven by the binding to the nucleic acids. We found that this hallmark is not suitable to differentiate between binding to RNA and DNA but can be successfully used to separate protein-binding and nucleic acid-binding residues. The residues that interact with nucleic acids are significantly more conserved in the sequence when compared with the protein-binding and nonbinding residues. Moreover, residues that interact with proteins have conservation values that are similar to the nonbinding residues. While all binding residues are generally biased to be more solvent exposed, we found that this bias is stronger for the multi-ligand-binding residues. Interestingly, we also discovered that none of the three hallmarks is statistically different between DNA- and RNA-binding residues and between the two types of the multi-ligand-binding residues. We empirically show that merging the information coming from the three hallmarks leads to improved ability to predict binding residues in protein sequences. We also demonstrate that predictive performance of these predictions is sufficient to relatively accurately find DNA-, RNA-, protein-, RNA- and protein- and DNA- and protein-binding residues in protein sequences and structures.

We combine the three hallmarks, using predicted from sequence RSA, to develop sequence-based hybridNAP predictor of propensities for protein, DNA, RNA and multi-ligand binding. This method is freely available at http://biomine.cs.vcu.edu/servers/hybridNAP/. We empirically show that hybridNAP provides modest predictive performance compared with the current sequence-based predictors of DNA- and RNA-binding residues. Although it is outperformed by some of the predictors, particularly for the prediction of DNA-binding residues, hybridNAP has several unique advantages. This is the first method that concurrently provides predictions of DNA-, RNA- and protein-binding residues in contrast to the other methods that predict binding to a single type of ligand (DBS-PSSM, DP-Bind, RNABindR, Pprint, PSIVER, SPRINGS and SSWRF) and to DNA and RNA (BindN+). Moreover, hybridNAP is the first method that provides predictions of the multi-ligand-binding residues, including residues that bind both DNA and proteins, and both RNA and proteins. It also offers arguably easy to understand interpretation of the putative propensities for binding, which are derived based on the values of the three hallmarks. The hybridNAP's Web page provides access to the data sets, values of the three hallmarks of binding and predictions from all considered methods on the data sets used in this project. These resources can be used in future studies to benchmark and develop more accurate predictors.

Key Points

- We review over 50 studies that focus on the analysis and/or prediction of protein–RNA, protein–DNA and protein–protein interactions.
- Three main hallmarks of DNA-, RNA- and protein-binding residues include relative solvent accessibility, evolutionary conservation and propensity of AAs for binding.
- Residues that bind both nucleic acids and proteins are more conserved and have higher solvent accessibility than residues that bind one type of ligands.
- Merging the information coming from the three hallmarks leads to improved ability to predict binding residues in protein sequences when compared with using a single hallmark.
- Linear combinations of the three hallmarks can be used to accurately predict DNA-, RNA- protein-, protein- and DNA- as well as protein- and RNA-binding residues in protein sequences.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported in part by Scholarship funded by China Scholarship Council to J.Z. and by the Qimonda Endowed Chair position to L.K.

References

1. Siggers T, Gordan R. Protein-DNA binding: complexities and multi-protein codes. *Nucleic Acids Res* 2014;**42**(4):2099–111.
2. Cook KB, Hughes TR, Morris QD. High-throughput characterization of protein-RNA interactions. *Brief Funct Genomics* 2015;**14**(1):74–89.
3. Sudha G, Nussinov R, Srinivasan N. An overview of recent advances in structural bioinformatics of protein-protein

interactions and a guide to their principles. *Prog Biophys Mol Biol* 2014;**116**(2–3):141–50.
4. Chen K, Kurgan L. Investigation of atomic level patterns in protein–small ligand interactions. *PLoS One* 2009;**4**(2): e4473.
5. Dudev T, Lim C. Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. *Chem Rev* 2014;**114**(1):538–56.
6. Peng T, Yuan X, Hang HC. Turning the spotlight on protein-lipid interactions in cells. *Curr Opin Chem Biol* 2014;**21**:144–53.
7. Gallina AM, Bork P, Bordo D. Structural analysis of protein-ligand interactions: the binding of endogenous compounds and of synthetic drugs. *J Mol Recognit* 2014;**27**(2):65–72.
8. Nagarajan R, Chothani S, Ramakrishnan C, *et al*. Structure based approach for understanding organism specific recognition of protein-RNA complexes. *Biol Direct* 2015;**10**(1):8.
9. Ellis JJ, Broom M, Jones S. Protein-RNA interactions: structural analysis and functional classes. *Proteins* 2006;**66**(4): 903–11.
10. Prabakaran P, Siebers JG, Ahmad S, *et al*. Classification of protein-DNA complexes based on structural descriptors. *Structure* 2006;**14**(9):1355–67.
11. Lejeune D, Delsaux N, Charloteaux B, *et al*. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* 2005;**61**(2):258–71.
12. Ehrenberger T, Cantley LC, Yaffe MB. Computational prediction of protein-protein interactions. *Methods Mol Biol* 2015; **1278**:57–75.
13. Si J, Zhao R, Wu R. An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci* 2015;**16**(3):5194–215.
14. Puton T, Kozlowski L, Tuszynska I, *et al*. Computational methods for prediction of protein-RNA interactions. *J Struct Biol* 2012;**179**(3):261–8.
15. Roche D, Brackenridge DA, McGuffin LJ. Proteins and their interacting partners: an introduction to protein-ligand binding site prediction methods. *Int J Mol Sci* 2015;**16**(12):29829–42.
16. Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol Biosyst* 2013;**9**(10):2417–25.
17. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**(1):235–42.
18. Berman HM, Kleywegt GJ, Nakamura H, *et al*. The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* 2012;**20**(3):391–6.
19. Zhang QC, Petrey D, Deng L, *et al*. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* 2012;**490**(7421):556–60.
20. Mizianty MJ, Fan X, Yan J, *et al*. Covering complete proteomes with X-ray structures: a current snapshot. *Acta Crystallogr D Biol Crystallogr* 2014;**70**(11):2781–93.
21. Apweiler R, Bairoch A, Wu CH, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**(90001): D115–19.
22. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69.
23. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics* 2007;**23**(5):634–6.
24. Wang L, Huang C, Yang MQ, *et al*. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol* 2010;**4(Suppl 1)**:S3.
25. Sathyapriya R, Vijayabaskar MS, Vishveshwara S, Nussinov R. Insights into protein-DNA interactions through structure network analysis. *PLoS Comput Biol* 2008;**4**(9):e1000170.
26. Dey S, Pal A, Guharoy M, *et al*. Characterization and prediction of the binding site in DNA-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters. *Nucleic Acids Res* 2012;**40**(15):7150–61.
27. Ahmad S, Keskin O, Sarai A, *et al*. Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res* 2008;**36**(18):5922–32.
28. Liu R, Hu J. DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins* 2013; **81**(11):1885–99.
29. Wang W, Liu J, Xiong Y, *et al*. Analysis and classification of DNA-binding sites in single-stranded and double-stranded DNA-binding proteins using protein information. *IET Syst Biol* 2014;**8**(4):176–83.
30. Zhou J, Xu R, He Y, *et al*. PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. *Sci Rep* 2016;**6**:27653.
31. Ma X, Guo J, Liu HD, *et al*. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**(6):1766–75.
32. Zhao H, Wang J, Zhou Y, *et al*. Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome. *PLoS One* 2014;**9**(5): e96694.
33. Hu J, Li Y, Zhang M, *et al*. Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs. *IEEE/ACM Trans Comput Biol Bioinform* 2017. doi:10.1109/TCBB.2016.2616469.
34. Dang TKL, Meckbach C, Tacke R, *et al*. A novel sequence-based feature for the identification of DNA-binding sites in proteins using Jensen–Shannon divergence. *Entropy* 2016; **18**(10):379.
35. Bahadur RP, Zacharias M, Janin J. Dissecting protein-RNA recognition sites. *Nucleic Acids Res* 2008;**36**(8):2705–16.
36. Barik A, Mishra A, Bahadur RP. PRince: a web server for structural and physicochemical analysis of protein-RNA interface. *Nucleic Acids Res* 2012;**40**(Web Server Issue): W440–4.
37. Kumar M, Gromiha MM, Raghava GP. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* 2008;**71**(1):189–94.
38. Wang Y, Xue Z, Shen G, *et al*. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids* 2008;**35**(2):295–302.
39. Chen YC, Sargsyan K, Wright JD, *et al*. Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic Acids Res* 2014;**42**(3):e15.
40. Terribilini M, Sander JD, Lee JH, *et al*. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res* 2007;**35(Web Server)**:W578–84.
41. Zhang T, Zhang H, Chen K, *et al*. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr Protein Pept Sci* 2010;**11**(7):609–28.
42. Fernandez M, Kumagai Y, Standley DM, *et al*. Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics* 2011;**12(Suppl 13)**:S5.
43. Luo J, Liu L, Venkateswaran S, *et al*. RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites. *Sci Rep* 2017;**7**(1):614.

44. Liu ZP, Wu LY, Wang Y, *et al*. Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics* 2010;**26**(13):1616–22.
45. Gupta A, Gribskov M. The role of RNA sequence and structure in RNA–protein interactions. *J Mol Biol* 2011;**409**(4): 574–87.
46. Cheng CW, Su EC, Hwang JK, *et al*. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics* 2008;**9(Suppl 12)**:S6.
47. Perez-Cano L, Fernandez-Recio J. Optimal Protein-RNA Area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins* 2010;**78**:25–35.
48. Wang CC, Fang Y, Xiao J, *et al*. Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids* 2011;**40**(1):239–48.
49. Ren H, Shen Y. RNA-binding residues prediction using structural features. *BMC Bioinformatics* 2015;**16**(1):249.
50. Li S, Yamashita K, Amada KM, *et al*. Quantifying sequence and structural features of protein-RNA interactions. *Nucleic Acids Res* 2014;**42**(15):10086–98.
51. Sun M, Wang X, Zou C, *et al*. Accurate prediction of RNA-binding protein residues with two discriminative structural descriptors. *BMC Bioinformatics* 2016;**17**(1):231.
52. Walia RR, Xue LC, Wilkins K, *et al*. RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One* 2014;**9**(5):e97725.
53. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* 2011;**12**(1):489.
54. Choi S, Han K. Prediction of RNA-binding amino acids from protein and RNA sequences. *BMC Bioinformatics* 2011; **12(Suppl 13)**:S7.
55. Sudha G, Singh P, Swapna LS, *et al*. Weak conservation of structural features in the interfaces of homologous transient protein-protein complexes. *Protein Sci* 2015;**24**(11):1856–73.
56. London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide-protein binding strategies. *Structure* 2010;**18**(2):188–99.
57. Asadabadi EB, Abdolmaleki P. Predictions of protein-protein interfaces within membrane protein complexes. *Avicenna J Med Biotechnol* 2013;**5**:148–57.
58. Murakami Y, Mizuguchi K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* 2010;**26**(15): 1841–8.
59. Singh G, Dhole K, Pai PP, *et al*. SPRINGS: prediction of protein-protein interaction sites using artificial neural networks. *PeerJ PrePrints* 2014; **2**:e266v2.
60. Laine E, Carbone A. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein protein interactions. *PLoS Comput Biol* 2015;**11**(12):e1004580.
61. Hwang H, Petrey D, Honig B. A hybrid method for protein-protein interface prediction. *Protein Sci* 2016;**25**(1):159–65.
62. Maheshwari S, Brylinski M. Prediction of protein-protein interaction sites from weakly homologous template structures using meta-threading and machine learning. *J Mol Recognit* 2015;**28**(1):35–48.
63. Liu GH, Shen HB, Yu DJ. Prediction of protein-protein interaction sites with machine-learning-based data-cleaning and post-filtering procedures. *J Membr Biol* 2016;**249**(1–2): 141–53.
64. Wei ZS, Han K, Yang JY, *et al*. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. *Neurocomputing* 2016;**193**:201–12.
65. Baussand J, Camproux AC. Deciphering the shape and deformation of secondary structures through local conformation analysis. *BMC Struct Biol* 2011;**11**(1):9.
66. Maheshwari S, Brylinski M. Template-based identification of protein-protein interfaces using eFindSitePPI. *Methods* 2016; **93**:64–71.
67. Baker CM, Grant GH. Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers* 2007;**85**(5–6): 456–70.
68. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. *Nat Rev Mol Cell Biol* 2014;**15**(11):749–60.
69. Hu J, He X, Yu DJ, *et al*. A new supervised over-sampling algorithm with application to protein-nucleotide binding residue prediction. *PLoS One* 2014;**9**(9):e107676.
70. Yang X, Wang J, Sun J, *et al*. SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PLoS One* 2015;**10**(7):e0133260.
71. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. *Nucleic Acids Res* 2017;**45**:e84.
72. Munteanu CR, Pimenta AC, Fernandez-Lozano C, *et al*. Solvent accessible surface area-based hot-spot detection methods for protein–protein and protein–nucleic acid interfaces. *J Chem Inf Model* 2015;**55**(5):1077–86.
73. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* 2015;**43**(18):e121.
74. Gromiha MM, Saranya N, Selvaraj S, *et al*. Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes. *Proteome Sci* 2011;**9(Suppl 1)**:S13.
75. Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* 2012;**28**(3): 331–41.
76. Yu DJ, Hu J, Huang Y, *et al*. TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble. *J Comput Chem* 2013;**34**(11):974–85.
77. Yu DJ, Hu J, Yan H, *et al*. Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace SVMs ensemble. *BMC Bioinformatics* 2014;**15**(1):297.
78. Panwar B, Gupta S, Raghava GPS. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC Bioinformatics* 2013;**14**(1):44.
79. Horst JA, Samudrala R. A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognit Lett* 2010;**31**(14):2103–12.
80. Passerini A, Lippi M, Frasconi P. Predicting metal-binding sites from protein sequence. *IEEE/ACM Trans Comput Biol Bioinform* 2012;**9**(1):203–13.
81. Yu DJ, Hu J, Yang J, *et al*. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**(4):994–1008.
82. Yu DJ, Hu J, Li QM, *et al*. Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction. *IEEE Trans Nanobioscience* 2015;**14**: 45–58.

83. Singh G, Dhole K, Pai PP, *et al*. SPRINGS: prediction of protein-protein interaction sites using artificial neural networks. *J Proteomics Comput Biol* 2014;**1**:7.
84. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* 2013;**41**:D1096–103.
85. UniProt C. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
86. Velankar S, Dana JM, Jacobsen J. SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res* 2013;**41**:D483–9.
87. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief Bioinform* 2017, in press.
88. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform* 2016;**17**(1):88–105.
89. Huang J, Deng R, Wang J. metaPIS: a sequence-based meta-server for protein interaction site prediction. *Protein Pept Lett* 2013;**20**:218–30.
90. Zhu Y, Zhou WQ, Dai DQ, *et al*. Identification of DNA-binding and protein-binding proteins using enhanced graph wavelet features. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**10**(4): 1017–31.
91. Vacic V, Uversky VN, Dunker AK, *et al*. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* 2007;**8**(1): 211.
92. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**(12):2577–637.
93. Tien MZ, Meyer AG, Sydykova DK, *et al*. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One* 2013; **8**(11):e80635.
94. Faraggi E, Zhou YQ, Kloczkowski A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins* 2014;**82**(11):3170–6.
95. Remmert M, Biegert A, Hauser A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
96. Fischer J, Mayer C, Soding J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* 2008;**24**(5):613–20.
97. Dou Y, Zheng X, Yang J, *et al*. Prediction of catalytic residues based on an overlapping amino acid classification. *Amino Acids* 2010;**39**(5):1353–61.
98. Kawashima S, Pokarowski P, Pokarowska M, *et al*. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**:D202–5.
99. Holland PW, Welsch RE. Robust regression using iteratively reweighted least-squares. *Commun Stat Theory Methods* 1977; **6**(9):813–27.
100. Meng F, Kurgan L. DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* 2016;**32**(12):i341–50.
101. Duh Y, Hsiao YY, Li CL, *et al*. Aromatic residues in RNase T stack with nucleobases to guide the sequence-specific recognition and cleavage of nucleic acids. *Protein Sci* 2015; **24**(12):1934–41.
102. Wilson KA, Kellie JL, Wetmore SD. DNA-protein pi-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res* 2014;**42**(10):6726–41.
103. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. *J Mol Biol* 2003;**325**(2):377–87.
104. Brinda KV, Kannan N, Vishveshwara S. Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Eng* 2002;**15**(4):265–77.
105. Halperin I, Wolfson H, Nussinov R. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure* 2004;**12**:1027–38.
106. Ma BY, Elkayam T, Wolfson H, *et al*. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci USA* 2003;**100**(10):5772–7.
107. Hu ZJ, Ma BY, Wolfson H, *et al*. Conservation of polar residues as hot spots at protein interfaces. *Proteins* 2000;**39**(4):331–42.
108. Gromiha MM, Fukui K. Scoring function based approach for locating binding sites and understanding recognition mechanism of protein-DNA complexes. *J Chem Inf Model* 2011; **51**(3):721–9.
109. Luscombe NM, Thornton JM. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* 2002;**320**(5):991–1009.
110. Caffrey DR, Somaroo S, Hughes JD, *et al*. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 2004;**13**(1):190–202.
111. Khafizov K, Madrid-Aliste C, Almo SC, *et al*. Trends in structural coverage of the protein universe and the impact of the protein structure initiative. *Proc Natl Acad Sci USA* 2014; **111**(10):3733–8.
112. Nagarajan R, Ahmad S, Gromiha MM. Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res* 2013;**41**(16):7606–14.
113. Ahmad S, Sarai A. PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 2005;**6**:33.