A comprehensive overview of sequence-based protein binding residue predictions for structured and disordered regions

Amita Barik and Lukasz Kurgan

Department of Computer Science, Virginia Commonwealth University, Richmond, VA, U.S.A. 23284
lkurgan@vcu.edu

Knowledge of protein-protein interactions facilitates annotation of protein functions and drug development efforts. Computational prediction of protein-protein interactions is motivated by a growing gap between the number of known and the number of functionally annotated protein sequences. We focus on sequence-based predictors of protein-binding residues (PBRs). They rely on predictive models that were developed from training data where the native annotations of PBRs were collected either from structures of protein-protein complexes (structure-annotated predictors) or which are found in the intrinsically disordered/unstructured regions (disorder-annotated predictors). This is the first overview that considers both groups of methods. We survey a comprehensive set of 36 predictors including 19 structure-annotated and 17 disorder-annotated tools. The fact that six new methods were published in 2018 alone suggests that this is still an active research area. We discuss their availability, impact, predictive architectures and outputs. These predictors rely on different combinations of five main types of inputs and typically utilize machine learning-derived predictive models. Methods with webservers are the most highly cited, with the average citation count of 104, compared to the overall average of 74. We note lack of solutions that combine structure-annotated and disorder-annotated data to accurately predict PBRs in both structured and disordered regions.

1. Introduction

Proteins interact with nucleic acids, lipids, a variety of small ligands and proteins, prompting development of computational predictors of these interactions [1-10]. Understanding protein-protein interactions (PPIs) is important for a variety of applications, such as annotation of protein functions [11], drug discovery [12-14] study of disease mechanisms [15, 16], and the development of PPI networks [17, 18]. Several publicly available data repositories and resources that archive PPI data at the molecule (protein) and molecular (residue or atomic) levels have been developed. Examples includes the Mentha resource that integrates data about PPIs at the protein level [19], BioLip that annotates protein-binding residues [20] and Protein Data Bank (PDB) which provides access to detailed atomic-level structures of protein-protein complexes [21-23]. However, the scope of these resources is

very limited given the large number of already sequenced proteins, which has recently reached over 137 million (source: UniProt [24, 25] as of January 14, 2019). To compare, mentha offers access to only 741 thousand interactions (as of Jan 13, 2014), BioLip to 21.5 thousand residue-level interaction sites (as of Jan 4, 2019), and PDB to close to 27 thousand complexes (as of Jan 14, 2019). The very wide gap between the expected and annotated number of interactions motivates the development of fast and cost-effective computational predictors of PPIs [8, 26-32] which can be used to support more tedious, labor intensive and relatively expensive experimental techniques [33-35].

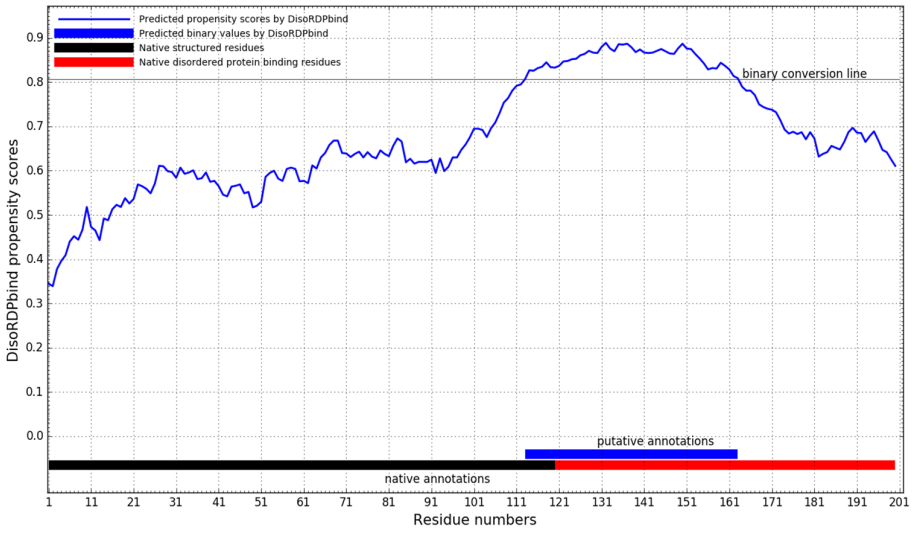

Figure 1. Prediction of PBRs for the nucleoplasmin protein from *Xenopus* (UniProt ID: P05221). Native annotation of PBRs, which were obtained from DisProt database (DisProt ID: DP00217), are shown with black and red horizontal bars at the bottom of the figure. The putative annotations generated with the DisoRDPbind method are shown with a thin blue line (for the putative numeric propensities) and with blue horizontal bars (for the putative binary predictions).

Numerous computational methods for the prediction of PPIs have been developed [8, 26-32]. They can be broadly classified into two categories based on the input: structure-based and sequence-based [8, 31]. The structure-based methods are limited in scope to a relatively small set of proteins that have structure and proteins for which structure can be accurately predicted. The sequence-based methods require only protein sequences as input and thus they can be applied to make predictions for any of the millions of the currently sequenced proteins. They predict PPIs at either the whole protein-level (whether or not a given protein

interacts with another protein) or at the residue-level (whether particular amino acids in the protein sequence interact with proteins). Prediction of interactions at the residue level arguably provides more insight/more detailed information and hence we focus on these methods. The residue-level methods predict protein binding residues (PBRs) either using a single protein sequence or a pair of sequences. This chapter surveys and describes a large collection of predictors that require a single protein sequence as input. These methods predict PBRs for every residue in the input protein chain. Discussion concerning a relatively small set of protein-pair-based methods [36-40] can be found in [8]. These methods find PBRs for a pair of protein sequences that are presumably interacting with each other.

The single-sequence predictors of PBRs take a protein sequence as their input and provide output in either binary format (each residue in the input protein sequence is classified as either PBR or non-PBR) or in both numeric (propensity score that quantifies likelihood that a given residue is protein-binding) and binary formats. Figure 1 shows example predictions of PBRs for the nucleoplasmin protein from *Xenopus* (UniProt ID: P05221) that were generated with the DisoRDPbind predictor [41, 42]. A thin blue line at the top of the figure shows the putative propensity scores that are generated for each residue in this protein sequence. Higher values of these scores correspond to a higher likelihood that a given residue binds proteins. These propensities are also used to generate the binary predictions. Residues with scores > threshold (0.807) are assumed to interact with proteins, while the remaining residues are assumed not to interact. This threshold was pre-optimized on a benchmark dataset to ensure that DisoRDPbind generates a low false positive rate of 10% (rate of incorrectly predicted PBRs). The binary predictions that stretch between positions 113 and 163 are represented by a thick blue horizontal line at the bottom of Figure 1. The native annotation of PBRs were obtained from the DisProt database [43, 44] and include intrinsically disordered protein-binding regions between positions 153 and 171 [45] and between positions 121 and 200 [46]. The predictions are in general in good agreement with the native PBRs. While DisoRDPbind fails to generate correct binary predictions at the C-terminus that is annotated as protein binding (positions 164 to 200), it still provides relatively high propensity scores for this region that suggests high likelihood of protein binding. This example not only explains format and interpretation of a sample prediction but it also shows that the putative propensities can be used to effectively supplement the binary predictions.

The single-sequence-based predictors of PBRs are being continually developed over the last couple of decades [8, 26-32]. We provide a comprehensive overview

of these predictors focusing on their availability, impact, predictive architectures, and outputs.

2. Computational prediction of protein binding residues from sequence

The single-sequence-based predictors of PBRs are typically derived with machine learning (ML) algorithms [47, 48]. These algorithms compute predictive models from a training dataset that is annotated with native PBRs. The ML algorithms optimize the architecture and parameters of the models such that the disagreements between their outputs and the native annotations in the training dataset is minimized. After the training is completed, the resulting models can be used to accurately predict PBRs in sequences of proteins that are not included in the training dataset [8].

The single-sequence-based predictors of PBRs are divided into two types based on the training datasets that they utilize: structure-annotated and disorder-annotated. The structure-annotated methods rely on training datasets where the annotations of PBRs are derived from structures of protein-protein complexes, typically collected from PDB [21-23]. The disorder-annotated predictors are optimized based on training datasets with intrinsically disordered protein-binding regions. The intrinsically disordered regions (IDRs) lack a stable three dimensional structure and they typically materialize as ensembles of multiple conformational states [49-56]. They are highly abundant in nature, particularly among eukaryotic organisms [57-60]. The structural plasticity of IDRs facilitates efficient interactions with a variety of different molecules [61-71], including proteins [72-76]. Many IDRs undergo disorder-to-structure transitions concomitant with their protein-binding activity [77-81], which means that some of these interactions are covered by the protein-protein complexes in PDB. Annotations of disordered PBRs can be collected from a variety of sources including PDB (based on regions with missing three dimensional structure) [82], DisProt [43, 44], and MobiDB [83, 84]. Moreover, both structure-annotated and disorder-annotated methods can be broadly classified into two categories based on the format of outputs that they generate. As discussed in Figure 1, a given predictor can produce either binary scores or both binary and propensity scores. Figure 2 illustrates a classification of current single-sequence-based predictors of PBRs on basis of the annotations and outputs. The efforts directed toward development of these predictors are well-balanced and include 19 structure-annotated methods and 17 disorder-annotated methods. However, while most of the disorder-annotated predictors provide

arguably more useful set of both output types, about half of the structure-annotated methods provide only the binary outputs.

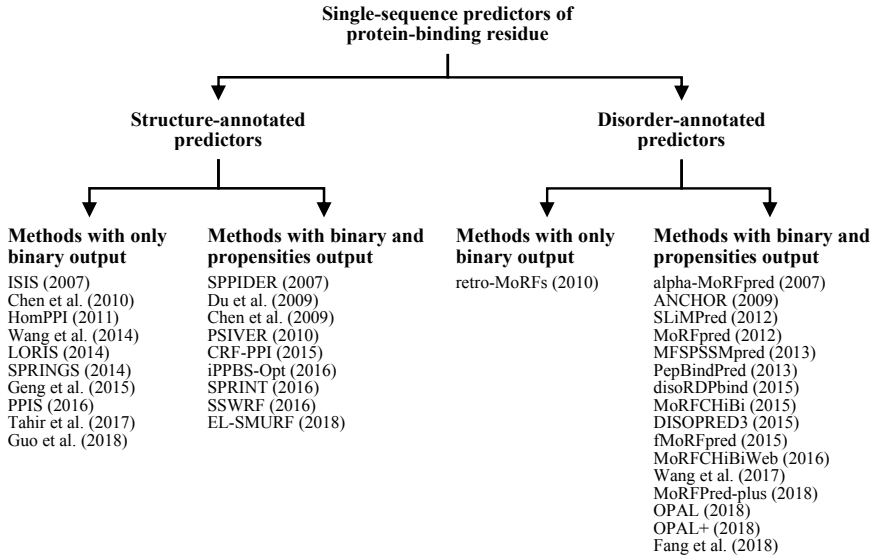

Figure 2. Classification of single-sequence-based predictors of PBRs based on the type of annotations used to generate the underlying predictive models and the outputs that they produce.

Several surveys of the single-sequence-based predictors of PBRs have been published [8, 26-32]. The main drawback of these articles is that they focus specifically on only the disorder-annotated [32] or structure-annotated [8, 26-31] methods. This is the first article that bridges that divide and discussed both types of computational predictors.

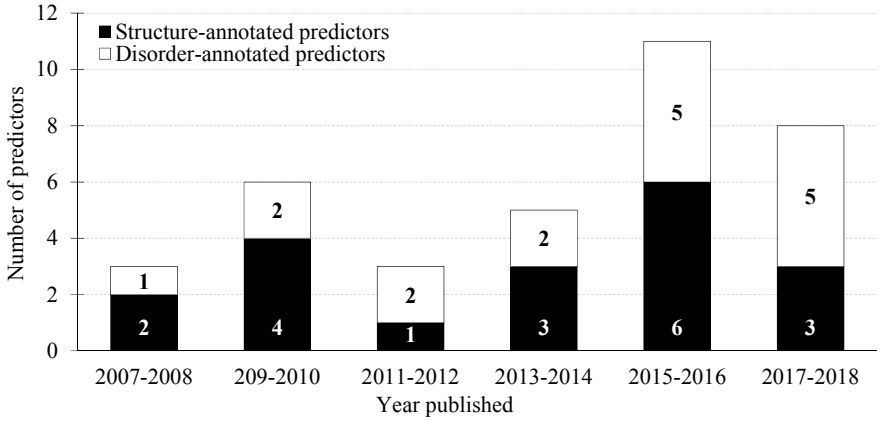

Figure 3. Historical timeline of the development of the single-sequence predictors of PBRs. The numbers inside the bars represent the number of predictors developed in the corresponding years.

2.1. *Overview of sequence based predictors of protein binding residues*

Well over 30 single-sequence predictors of PBRs have been developed. Figure 3 depicts a historical timeline of these efforts. The structure-annotated methods are represented using black bars (total of 19) while disordered-annotated methods are shown with white bars (total of 17). The first two methods, the structure-annotated ISIS [85] and the disorder-annotated alpha-MoRFpred [86, 87] were released in 2007. We observe that while initially the structure-annotated predictors were dominant (6 structure-annotated vs. 3 disorder-annotated methods were published between 2007 and 2010), recent years have observed a substantial shift towards the development of the disorder-annotated methods (10 disorder-annotated vs. 9 structure-annotated methods were released between 2015 and 2018). The fact that 2018 alone has seen six new methods [88-93] suggests that this is still an active research area.

Table 1 provides references and information about the year of publication, availability and citation counts for the 36 predictors. We provide URLs (Uniform Resource Locators) for the implementations that are currently (as of January 7, 2019) available for 24 of the 36 predictors. The implementations can be provided in two ways: as webservers and/or standalone code. The webservers are comparatively easier to use and primarily target less computer savvy users who want to perform *ad hoc* predictions. The computations are performed on the server

side, and the end user only need access to Internet and a web browser to run the predictions. The predicted results are returned to the users typically via email or/and the web browser. Only 4 (21%) out of the 19 structure-annotated predictors provide webservers while 11 (65%) out of 17 of the disorder-annotated methods have webserver facilities. The source code is helpful for users who want to run the predictions on their hardware, which could be because they need to predict a large dataset of proteins or embed a given predictor into a larger bioinformatics pipeline. About 37% (7 out of 19) of the structure-annotated methods and 41% (7 out 17) of the disorder-annotated methods provide this option. Finally, five disorder-annotated predictors including ANCHOR [94-96], MFSPSSM [97], MoRFChiBi [98], DISOPRED3 [99], and MoRFCHiBiWeb [100] as available as both webserver and source code. Moreover, 12 predictors were either never made available to the community (10 methods) or the support was discontinued after they were made available at the time of publication. The latter include the predictor by Chen et al. [101] and SSWRF [102].

Table 1 also quantifies citations, arguably one of the most important aspects of impact generated by these predictors. We note that we use only one reference, the one with the highest citation counts, for the few methods that were published in multiple articles to avoid duplicate counting when measuring these values. We quantify the total and the annual number of citations collected from Google Scholar. Overall, the 36 predictors were cited 2649 times, with an impressive average of 74 citations per method and median of 31 citations per method. The average and median increase to 104 and 54 for the 15 methods that offer webservers. The annual citation counts are more suitable for direct comparisons between methods and they reveal that the most-cited methods include DISOPRED3 [99], iPPBS-Opt [103], ANCHOR [94-96], alpha-MoRFpred [86, 87] and MoRFpred [104, 105], all of which secure over 30 citations annually. Interestingly, the availability of these predictors is directly correlated with their citations. Predictors that are not available and those that are available as only standalone code have the average annual citations of 10.3 and 5.2, respectively. The methods that are available as only webservers secure on average 17.3 citations per year while the five predictors that are provided as both standalone code and webserver have accumulated an average of 27.8 citations annually. This agrees with common sense since computational methods that are not made available or that have to be installed locally are less likely to be utilized by the end users, and thus less likely to be cited.

Table 1. Overview of the single-sequence-based predictors of PBRs. The methods divided into two groups: 19 structure-annotated and 17 disorder-annotated, and sorted by the publication year in the ascending order within each group. The 'Type' column indicates whether a given method is available as the online webserver (WS) and/or standalone source code (SC); NA means that neither webserver nor source code is available. The 'URL' column gives the page where the method can be found as of January 7, 2019, where NLA means that the method is 'no longer available' while the published article claims that it was originally available. The 'Total citations' column gives the number of citations collected from Google Scholar on January 7, 2019. To avoid duplicate counting of citations for methods that are published in multiple articles, we use the one with the highest number of citations. The 'annual citations' column gives an average number of citations per year since a given method was published.

Type	Method	Ref.	Year published	Availability		Citations	
				Type	URL	Total	Annual
Structure-annotated predictors	ISIS	[85]	2007	NA	NA	244	22
	SPPIDER	[106]	2007	WS	http://sppider.cchmc.org/	272	25
	Du et al.	[107]	2009	NA	NA	11	1
	Chen et al.	[101]	2009	SC	NLA	126	14
	PSIVER	[108]	2010	WS	http://mizuguchilab.org/PSIVER/	114	14
	Chen et al.	[109]	2010	SC	http://mail.ustc.edu.cn/~bigeagle/BMCBioinfo2010/index.htm	39	5
	HomPPI	[110]	2011	WS	http://ailab1.ist.psu.edu/PSHOMPPIv1.2/index.html	55	8
	Wang et al.	[111]	2014	NA	NA	50	13
	LORIS	[112]	2014	SC	https://sites.google.com/site/sukantamondal/software	24	6
	SPRINGS	[113]	2014	SC	https://sites.google.com/site/predppis/	12	3
	CRF-PPI	[114]	2015	SC	http://csbio.njust.edu.cn/bioinf/CRF-PPI	6	2
	Geng et al.	[115]	2015	NA	NA	10	3
	iPPBS-Opt	[103]	2016	WS	http://www.jci-bioinfo.cn/iPPBS-Opt	94	47
	PPIS	[116]	2016	SC	http://csbio.njust.edu.cn/bioinf/PPIS	13	7
	SPRINT	[117]	2016	SC	http://sparks-lab.org/server/SPRINT/	32	16
	SSWRF	[102]	2016	SC	NLA	30	15
	Tahir et al	[118]	2017	NA	NA	4	4
	Guo et al	[91]	2018	NA	NA	6	6
	EL-SMURF	[92]	2018	SC	http://github.com/QUST-AIBBDRC/EL-SMURF/	0	0
Disorder-annotated predictors	alpha-MoRFpred	[86, 87]	2007	NA	NA	445	40
	ANCHOR	[94-96]	2009	WS + SC	http://anchor.enzim.hu	386	43
	retro-MoRFs	[119]	2010	NA	NA	27	3
	SLiMPred	[120]	2012	WS	http://bioware.ucd.ie/~compass/biowareweb//Server_pages/slimpred.php	54	9
	MoRFpred	[104, 105]	2012	WS	http://biomine.cs.vcu.edu/servers/MoRFpred/	192	32
	MFSPSSMpred	[97]	2013	WS + SC	http://webapp.yama.info.waseda.ac.jp/fang/MoRFs.php	33	7
	PepBindPred	[121]	2013	WS	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/pepbindpred.php	17	3
	DisoRDPbind	[41, 42]	2015	WS	http://biomine.cs.vcu.edu/servers/DisoRDPbind/	44	15
	MoRFCHiBi	[98]	2015	WS + SC	https://morf.msl.ubc.ca/index.xhtml	35	12
	DISOPRED3	[99]	2015	WS + SC	http://bioinf.cs.ucl.ac.uk/disopred	199	66
	fMoRFpred	[81]	2015	WS	http://biomine.cs.vcu.edu/servers/fMoRFpred/	35	12
	MoRFCHiBiWeb	[100]	2016	WS + SC	http://morf.chibi.ubc.ca:8080/mcw/index.xhtml	22	11
	Wang et al.	[122]	2017	NA	NA	2	2
	MoRFPred-plus	[93]	2018	SC	https://github.com/roneshsharma/MoRFpred-plus/wiki/MoRFpred-plus	8	8
	OPAL	[88]	2018	WS	http://www.alok-ai-lab.com/tools/opal/	8	8
	OPAL+	[89]	2018	SC	https://github.com/roneshsharma/OPAL-plus/wiki/OPAL-plus-Download	0	0
	Fang et al.	[90]	2018	NA	NA	0	0

Table 2. Predictive architecture and outputs generated by the single-sequence-based predictors of PBRs. The methods divided into two groups: 19 structure-annotated and 17 disorder-annotated, and sorted by the publication year in the ascending order within each group. The 'Input' sub-columns include information extracted directly from the amino acid sequence (AAS); evolutionary information (EVO); and structural properties predicted from the sequence that include putative relative solvent accessibility (pRSA), putative secondary structure (pSS) and putative intrinsic disorder (pDIS). The √ means that a given input type is utilized as one of the inputs while blank cell indicates that it is not considered. The 'Predictive model' column categorizes the models into two groups: those generated with machine learning (ML) algorithms and those that rely on a scoring function (SF) generated either by an empirical formula or using an alignment score. The ML learning models include neural network (NN), K-nearest neighbor (KNN), probabilistic neural network (PNN), support vector machine (SVM), random forest (RF), naïve Bayes (NB), regularized logistic function (RLF), and LR (logistic regression). The √ in the 'Outputs' column means that a given type of output is generated by the predictor.

Type	Method	Ref.	Inputs					Predictive model	Outputs	
			AAS	EVO	pRSA	pSS	pDIS		Binary	Propensity
Structure-annotated predictors	ISIS	[85]		√	√	√		ML (NN)	√	
	SPPIDER	[106]	√	√	√			ML (KNN)	√	√
	Du et al.	[107]	√	√	√			ML (SVM)	√	√
	Chen et al.	[101]	√	√	√			ML (RF)	√	√
	PSIVER	[108]		√	√			ML (NB)	√	√
	Chen et al.	[109]	√	√	√			ML (SVM)	√	
	HomPPI	[110]		√	√		√	SF	√	
	Wang et al.	[111]		√	√			ML (SVM)	√	
	LORIS	[112]	√	√	√	√		ML (RLF)	√	
	SPRINGS	[113]	√	√	√	√		ML (NN)	√	
	CRF-PPI	[114]	√	√	√			ML (RF)	√	√
	Geng et al.	[115]		√	√			ML (NB)	√	
	iPPBS-Opt	[103]	√		√			ML (KNN)	√	√
	PPIS	[116]	√	√	√			ML (RF)	√	
	SPRINT	[117]	√	√	√	√		ML (SVM)	√	√
	SSWRF	[102]	√	√	√			ML (SVM, RF)	√	√
	Tahir et al.	[118]	√	√				ML (KNN, PNN, SVM)	√	
	Guo et al	[91]	√	√	√			ML (SVM)	√	
	EL-SMURF	[92]	√	√	√			ML (RF)	√	√
Disorder-annotated predictors	alpha-MoRFpred	[86, 87]	√			√	√	ML (NN)	√	√
	ANCHOR	[94-96]	√					SF	√	√
	retro-MoRFs	[119]	√				√	SF	√	
	SLiMPred	[120]	√		√	√	√	ML (NN)	√	√
	MoRFpred	[104, 105]	√	√	√		√	ML (SVM)	√	√
	MFSPSSMpred	[97]		√				ML (SVM)	√	√
	PepBindPred	[121]				√	√	ML (NN)	√	√
	DisoRDPbind	[41, 42]	√			√	√	ML (LR)	√	√
	MoRFCHiBi	[98]	√					ML (SVM)	√	√
	DISOPRED3	[99]	√	√				ML (SVM)	√	√
	fMoRFpred	[81]	√			√	√	ML (SVM)	√	√
	MoRFCHiBiWeb	[100]	√	√			√	ML (NB)	√	√
	Wang et al.	[122]	√	√	√	√		ML (SVM	√	√
	MoRFPred-plus	[93]	√	√				ML (SVM)	√	√
	OPAL	[88]	√		√		√	ML (SVM)	√	√
	OPAL+	[89]	√		√	√		ML (SVM)	√	√
	Fang et al.	[90]		√				ML (SVM)	√	√

2.2. *Architectures of the predictors of protein binding residues*

Table 2 provides details about the architectures of the 36 single-sequence predictors of PBRs. We deconstruct the architectures into three major parts: inputs, predictive models and outputs, and we discuss each of these individually.

The five commonly used elements of the input are computed from the amino acid sequence (AAS), evolutionary information (EVO), and from three relevant predicted structural properties: putative relative solvent accessibility (pRSA), putative secondary structure (pSS) and putative intrinsic disorder (pDIS). AAS-based input typically quantifies amino acid composition, residue-level physiochemical properties, and/or position of amino acids in the sequence. EVO is usually calculated from the position specific scoring matrix generated from the input protein chain with the PSI-BLAST algorithm [123]. While AA and EVO are calculated directly from the protein sequence, the other three input elements (pRSA, pSS, and pDIS) are predicted from the sequence with bioinformatics tools. RSA is a measure of residue-level solvent exposure which is calculated by dividing the predicted solvent accessible surface of a given residue in the input protein sequence by the maximum possible solvent accessible surface area of the same amino acid type. Virtually all structure-annotated methods, except for [118], use pRSA, compared to only 5 out of 17 disorder-annotated tools. This is expected since disordered protein-binding regions do not have a well-defined surface when compared to the structured protein-protein interfaces. The pSS is generated with one of the popular tools [124-127] that include PSIPRED [128], PROFphd [129], and Distill [130]. The pDIS is clearly relevant for the disorder-annotated predictors and there are many accurate algorithms that can be used to provide this input [131-136]. Correspondingly, majority of the disorder-annotated predictors (9 out of 17) use this input, compared to only one structure-annotated method. Overall, we observe that none of the methods uses all five types of inputs, while majority of the methods (22 out of 36) use between 3 and 4 input types.

Over 90% of methods (33 out of 36), which exclude just one structure-annotated and two disorder-annotated predictors, apply machine learning (ML) algorithms to generate predictive models. Some of the frequently used ML models are support vector machines (used by 17 methods), neural networks (6 predictors), random forest (5 methods), and K-nearest neighbors (3 methods) and Naïve Bayes (3 methods). The three non-ML predictors rely on relatively simple predictive models in the form of an empirical formula or sequence alignment [94-96, 110, 119].

The outputs of the predictors, which we explain in the 'Introduction' section, take two forms: binary values and propensity scores. While all predictors discussed here generate binary outputs, some of them do not provide the propensities. To be more precise, all but one disorder-annotated predictor provide both binary and propensity scores, while only 47% of the structure-annotated methods (9 out of 19 methods) generate both outputs. This lack of propensities is a drawback since these values can be used to provide useful context for the binary predictions. For instance, propensities can be used to find missing putative PBRs (i.e, residue that are not predicted as PBRs in binary but which have relatively high propensities, which is illustrated in Figure 1) or to find lower quality predictions of PBRs (i.e., residues predicted as PBRs in binary but with relatively low propensity scores).

Overall, our analysis reveals that the single-sequence predictors of PBRs are characterized by a wide range of architectures that rely on five major types of inputs and that use a variety of ML-derived predictive models. The fact that none of the current predictors applies all five input types opens an opportunity to develop more accurate models that would take advantage of all types of inputs.

3. Summary and recommendations

The low coverage and importance of the current annotations of protein-protein interactions have stimulated the development of numerous single-sequence predictors of PBRs. We discuss two categories of these computational predictors that were trained using structure-annotated vs. disorder-annotated datasets. We show that recent development efforts have shifted towards the disorder-annotated predictors, with four methods that were released in 2018 alone. Most of these methods are made available to the community as either source code or webserver, with only a few that offer both options. Empirical analysis reveals that methods that include webservers are cited at much higher rates, with an average of 104 citations. We also summarize the predictive architecture of the 36 predictors of PBRs. They rely on various combinations of five main types of inputs and use of a wide range of primarily ML-derived predictive models.

Although most of the predictors are made available to the community as webservers and/or standalone programs, 1/3 (12 out of 36) were not released or their availability was discontinued. They are unlikely to be ever used and consequently they are cited substantially less often. We believe that methods that are not made available should not be published and that the subsequent support

should be guaranteed by the authors as part of the publication process. These standards are already imposed in some of the lead publication venues, such as *Nucleic Acids Research* where webserver articles are expected to be functional and maintained for at least five years after publication. We also stress the importance of the propensity outputs that provide a useful and effective context to the binary outputs, which we demonstrate in Figure 1. The designers of these methods should strive to provide the propensities, which unfortunately is not the case for about half of the structure-annotated methods.

A perhaps surprising observation is the separation between the disorder-annotated and the structure-annotated methods. There is not a single method that combines both types of training data to provide a more complete solution capable of predicting PBRs in both structured and disordered regions. Such cross-over solutions should be developed in the near future.

A recent comparative review reveals that the structure-annotated predictors generate moderately accurate predictions of PBRs [8]. The binary outputs of the best structure-annotated predictors offer accuracies at about 80% (the high value is driven the large majority of easy to predict non-binding residues) and modest levels of correlation between the predicted and native PBRs, with the Matthews correlation coefficient of 0.21. The putative propensities that they produce are characterized by moderate AUC values in the 0.65 to 0.69 range, where the overall AUC values range between 0.5 and 1. We also emphasize a recent empirical observation that these methods heavily cross-predict residues that bind other ligands (RNA, DNA and a variety of small ligands) as PBRs. As many as 20 to 40% of the residues that interact with these other ligands are predicted as PBRs, when compared to similar levels of sensitivity (rate of correct predictions for native PBRs) [8]. This means that the current methods predict PBRs at similar rates among the native PBRs as among the residues that bind the other ligands. This likely stems from the fact that training datasets of these methods are solely use proteins with PBRs, lacking a sufficient population of proteins that interact with other ligands. This results in an inability to train the predictive models that could differentiate between PBRs and other types of ligand-binding residues. Thus, we propose that more accurate solutions that use better training datasets and that can more specifically target PBRs should be developed.

Acknowledgment

This research was supported in part by the by the National Science Foundation (grant 1617369) and the Robert J. Mattauch Endowment funds to L.K.

References

1. Ding, X.M., et al., *Computational prediction of DNA-protein interactions: a review.* Curr Comput Aided Drug Des, 2010. **6**(3): p. 197-206.
2. Chen, K. and L. Kurgan, *Investigation of atomic level patterns in protein--small ligand interactions.* PLoS One, 2009. **4**(2): p. e4473.
3. Sudha, G., R. Nussinov, and N. Srinivasan, *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): p. 141-50.
4. Fornes, O., et al., *On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions.* Adv Protein Chem Struct Biol, 2014. **94**: p. 77-120.
5. Barik, A., et al., *Molecular architecture of protein-RNA recognition sites.* J Biomol Struct Dyn, 2015. **33**(12): p. 2738-51.
6. Chowdhury, S., J. Zhang, and L. Kurgan, *In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome.* Proteomics, 2018: p. e1800064.
7. Zhao, H., Y. Yang, and Y. Zhou, *Prediction of RNA binding proteins comes of age from low resolution to high resolution.* Mol Biosyst, 2013. **9**(10): p. 2417-25.
8. Zhang, J. and L. Kurgan, *Review and comparative assessment of sequence-based predictors of protein-binding residues.* Brief Bioinform, 2018. **19**(5): p. 821-837.
9. Hu, G., et al., *Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions.* Structure, 2012. **20**(11): p. 1815-22.
10. Chen, K., et al., *A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds.* Structure, 2011. **19**(5): p. 613-21.
11. Orii, N. and M.K. Ganapathiraju, *Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function.* PLoS One, 2012. **7**(11): p. e49029.
12. Sperandio, O., *Editorial: Toward the design of drugs on protein-protein interactions.* Curr Pharm Des, 2012. **18**(30): p. 4585.
13. Petta, I., et al., *Modulation of Protein-Protein Interactions for the Development of Novel Therapeutics.* Mol Ther, 2016. **24**(4): p. 707-18.
14. Athanasios, A., et al., *Protein-Protein Interaction (PPI) Network: Recent Advances in Drug Discovery.* Curr Drug Metab, 2017. **18**(1): p. 5-10.

15. Kuzmanov, U. and A. Emili, *Protein-protein interaction networks: probing disease mechanisms using model systems.* Genome Med, 2013. **5**(4): p. 37.
16. Nibbe, R.K., et al., *Protein-protein interaction networks and subnetworks in the biology of disease.* Wiley Interdiscip Rev Syst Biol Med, 2011. **3**(3): p. 357-67.
17. De Las Rivas, J. and C. Fontanillo, *Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell.* Brief Funct Genomics, 2012. **11**(6): p. 489-96.
18. Hao, T., et al., *Reconstruction and Application of Protein-Protein Interaction Network.* Int J Mol Sci, 2016. **17**(6).
19. Calderone, A., L. Castagnoli, and G. Cesareni, *mentha: a resource for browsing integrated protein-interaction networks.* Nat Methods, 2013. **10**(8): p. 690-1.
20. Yang, J., A. Roy, and Y. Zhang, *BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1096-103.
21. Burley, S.K., *PDB40: The Protein Data Bank celebrates its 40th birthday.* Biopolymers, 2013. **99**(3): p. 165-9.
22. Burley, S.K., et al., *Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive.* Methods Mol Biol, 2017. **1607**: p. 627-641.
23. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
24. The UniProt, C., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.
25. UniProt, C., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(Database issue): p. D204-12.
26. Ezkurdia, I., et al., *Progress and challenges in predicting protein-protein interaction sites.* Brief Bioinform, 2009. **10**(3): p. 233-46.
27. Fernandez-Recio, J., *Prediction of protein binding sites and hot spots.* Wiley Interdisciplinary Reviews-Computational Molecular Science, 2011. **1**(5): p. 680-698.
28. Aumentado-Armstrong, T.T., B. Istrate, and R.A. Murgita, *Algorithmic approaches to protein-protein interaction site prediction.* Algorithms Mol Biol, 2015. **10**: p. 7.
29. Xue, L.C., et al., *Computational prediction of protein interfaces: A review of data driven methods.* FEBS Lett, 2015. **589**(23): p. 3516-26.
30. Esmaielbeiki, R., et al., *Progress and challenges in predicting protein interfaces.* Brief Bioinform, 2016. **17**(1): p. 117-31.
31. Maheshwari, S. and M. Brylinski, *Predicting protein interface residues using easily accessible on-line resources.* Brief Bioinform, 2015. **16**(6): p. 1025-34.
32. Meng, F., V.N. Uversky, and L. Kurgan, *Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions.* Cell Mol Life Sci, 2017. **74**(17): p. 3069-3090.

33. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration.* Nature Biotechnology, 1999. **17**(10): p. 1030-1032.
34. Sobott, F. and C.V. Robinson, *Protein complexes gain momentum.* Curr Opin Struct Biol, 2002. **12**(6): p. 729-34.
35. Yates, J.R., 3rd, *Mass spectrometry. From genomics to proteomics.* Trends Genet, 2000. **16**(1): p. 5-8.
36. Pitre, S., et al., *PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.* BMC Bioinformatics, 2006. **7**: p. 365.
37. Shi, M.G., et al., *Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset.* Amino Acids, 2010. **38**(3): p. 891-899.
38. Chang, D.T.H., Y.T. Syu, and P.C. Lin, *Predicting the protein-protein interactions using primary structures with predicted protein surface.* Bmc Bioinformatics, 2010. **11**.
39. Amos-Binks, A., et al., *Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences.* Bmc Bioinformatics, 2011. **12**.
40. Xia, B., et al., *PETs: A Stable and Accurate Predictor of Protein-Protein Interacting Sites Based on Extremely-Randomized Trees.* Ieee Transactions on Nanobioscience, 2015. **14**(8): p. 882-893.
41. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind.* Methods Mol Biol, 2017. **1484**: p. 187-203.
42. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Res, 2015. **43**(18): p. e121.
43. Vucetic, S., et al., *DisProt: a database of protein disorder.* Bioinformatics, 2005. **21**(1): p. 137-40.
44. Piovesan, D., et al., *DisProt 7.0: a major update of the database of disordered proteins.* Nucleic Acids Res, 2016. **D1**: p. D219-D227.
45. Conti, E. and J. Kuriyan, *Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha.* Structure, 2000. **8**(3): p. 329-38.
46. Hierro, A., et al., *Structural and functional properties of Escherichia coli-derived nucleoplasmin. A comparative study of recombinant and natural proteins.* Eur J Biochem, 2001. **268**(6): p. 1739-48.
47. Larranaga, P., et al., *Machine learning in bioinformatics.* Brief Bioinform, 2006. **7**(1): p. 86-112.
48. Kurgan, L. and Y. Zhou, *Machine learning models in protein bioinformatics.* Curr Protein Pept Sci, 2011. **12**(6): p. 455.
49. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J Mol Biol, 1999. **293**(2): p. 321-331.

50. Uversky, V.N., J.R. Gillespie, and A.L. Fink, *Why are "natively unfolded" proteins unstructured under physiologic conditions?* Proteins, 2000. **41**(3): p. 415-427.
51. Dunker, A.K., et al., *Intrinsically disordered protein.* J Mol Graph Model, 2001. **19**(1): p. 26-59.
52. Uversky, V.N. and A.K. Dunker, *Understanding protein non-folding.* Biochim Biophys Acta, 2010. **1804**(6): p. 1231-64.
53. Habchi, J., et al., *Introducing protein intrinsic disorder.* Chem Rev, 2014. **114**(13): p. 6561-88.
54. Uversky, V.N., *Introduction to intrinsically disordered proteins (IDPs).* Chem Rev, 2014. **114**(13): p. 6557-60.
55. van der Lee, R., et al., *Classification of Intrinsically Disordered Regions and Proteins.* Chemical Reviews, 2014. **114**(13): p. 6589-6631.
56. Lieutaud, P., et al., *How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe.* Intrinsically Disord Proteins, 2016. **4**(1): p. e1259708.
57. Xue, B., A.K. Dunker, and V.N. Uversky, *Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.* J Biomol Struct Dyn, 2012. **30**(2): p. 137-49.
58. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
59. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cell Mol Life Sci, 2015. **72**(1): p. 137-51.
60. Peng, Z., M.J. Mizianty, and L. Kurgan, *Genome-scale prediction of proteins with long intrinsically disordered regions.* Proteins, 2014. **82**(1): p. 145-58.
61. Dyson, H.J., *Roles of intrinsic disorder in protein-nucleic acid interactions.* Mol Biosyst, 2012. **8**(1): p. 97-104.
62. Varadi, M., et al., *Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins.* PLoS One, 2015. **10**(10): p. e0139731.
63. Meng, F., et al., *Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments.* Int J Mol Sci, 2015. **17**(1).
64. Peng, Z., et al., *More than just tails: intrinsic disorder in histone proteins.* Mol Biosyst, 2012. **8**(7): p. 1886-901.
65. Xue, B., et al., *Protein intrinsic disorder as a flexible armor and a weapon of HIV-1.* Cell Mol Life Sci, 2012. **69**(8): p. 1211-59.
66. Fan, X., et al., *The intrinsic disorder status of the human hepatitis C virus proteome.* Mol Biosyst, 2014. **10**(6): p. 1345-63.
67. Xue, B., et al., *Structural disorder in viral proteins.* Chem Rev, 2014. **114**(13): p. 6880-911.

68. Peng, Z., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome.* Cell Mol Life Sci, 2014. **71**(8): p. 1477-504.
69. Wu, Z., et al., *In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces.* FEBS Lett, 2015. **589**(19 Pt A): p. 2561-9.
70. Na, I., et al., *Autophagy-related intrinsically disordered proteins in intra-nuclear compartments.* Mol Biosyst, 2016. **12**(9): p. 2798-817.
71. Peng, Z., et al., *Resilience of death: intrinsic disorder in proteins involved in the programmed cell death.* Cell Death Differ, 2013. **20**(9): p. 1257-67.
72. Fuxreiter, M., et al., *Disordered proteinaceous machines.* Chem Rev, 2014. **114**(13): p. 6806-43.
73. Tompa, P., et al., *Close encounters of the third kind: disordered domains and the interactions of proteins.* Bioessays, 2009. **31**(3): p. 328-35.
74. Patil, A. and H. Nakamura, *Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks.* FEBS Lett, 2006. **580**(8): p. 2041-5.
75. Hu, G., et al., *Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions.* Int J Mol Sci, 2017. **18**(12).
76. Xue, B., et al., *Stochastic machines as a colocalization mechanism for scaffold protein function.* FEBS Lett, 2013. **587**(11): p. 1587-91.
77. Dyson, H.J. and P.E. Wright, *Coupling of folding and binding for unstructured proteins.* Curr Opin Struct Biol, 2002. **12**(1): p. 54-60.
78. Fuxreiter, M., et al., *Disordered proteinaceous machines.* Chem Rev, 2014. **114**(13): p. 6806-43.
79. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* J Mol Biol, 2006. **362**(5): p. 1043-59.
80. Vacic, V., et al., *Characterization of molecular recognition features, MoRFs, and their binding partners.* J Proteome Res, 2007. **6**(6): p. 2351-66.
81. Yan, J., et al., *Molecular recognition features (MoRFs) in three domains of life.* Mol Biosyst, 2016. **12**(3): p. 697-710.
82. DeForte, S. and V.N. Uversky, *Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree.* Protein Sci, 2016. **25**(3): p. 676-88.
83. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D471-D476.
84. Di Domenico, T., et al., *MobiDB: a comprehensive database of intrinsic protein disorder annotations.* Bioinformatics, 2012. **28**(15): p. 2080-2081.
85. Ofran, Y. and B. Rost, *ISIS: interaction sites identified from sequence.* Bioinformatics, 2007. **23**(2): p. e13-6.
86. Oldfield, C.J., et al., *Comparing and combining predictors of mostly disordered proteins.* Biochemistry, 2005. **44**(6): p. 1989-2000.

87. Cheng, Y., et al., *Mining alpha-helix-forming molecular recognition features with cross species sequence alignments.* Biochemistry, 2007. **46**(47): p. 13468-77.
88. Sharma, R., et al., *OPAL: prediction of MoRF regions in intrinsically disordered protein sequences.* Bioinformatics, 2018. **34**(11): p. 1850-1858.
89. Sharma, R., et al., *OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences.* Proteomics, 2018: p. e1800058.
90. Fang, C., et al., *Identifying MoRFs in Disordered Proteins Using Enlarged Conserved Features*, in *Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology*. 2018, ACM: Chengdu, China. p. 50-54.
91. Guo, H., et al., *Predicting protein-protein interaction sites using modified support vector machine.* International Journal of Machine Learning and Cybernetics, 2018. **9**(3): p. 393-398.
92. Wang, X., et al., *Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique.* Bioinformatics, 2018.
93. Sharma, R., et al., *MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles.* Journal of Theoretical Biology, 2018. **437**: p. 9-16.
94. Meszaros, B., I. Simon, and Z. Dosztanyi, *Prediction of protein binding regions in disordered proteins.* PLoS Comput Biol, 2009. **5**(5): p. e1000376.
95. Dosztanyi, Z., B. Meszaros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins.* Bioinformatics, 2009. **25**(20): p. 2745-6.
96. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.
97. Fang, C., et al., *MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation.* BMC Bioinformatics, 2013. **14**: p. 300.
98. Malhis, N. and J. Gsponer, *Computational identification of MoRFs in protein sequences.* Bioinformatics, 2015. **31**(11): p. 1738-44.
99. Jones, D.T. and D. Cozzetto, *DISOPRED3: precise disordered region predictions with annotated protein-binding activity.* Bioinformatics, 2015. **31**(6): p. 857-63.
100. Malhis, N., M. Jacobson, and J. Gsponer, *MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences.* Nucleic Acids Res, 2016.
101. Chen, X.W. and J.C. Jeong, *Sequence-based prediction of protein interaction sites with an integrative method.* Bioinformatics, 2009. **25**(5): p. 585-91.

102. Wei, Z.S., et al., *Protein-protein interaction sites prediction by ensembling SVM and sample-weighted random forests.* Neurocomputing, 2016. **193**: p. 201-212.
103. Jia, J.H., et al., *iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets.* Molecules, 2016. **21**(1).
104. Disfani, F.M., et al., *MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins.* Bioinformatics, 2012. **28**(12): p. i75-83.
105. Oldfield, C.J., V.N. Uversky, and L. Kurgan, *Predicting Functions of Disordered Proteins with MoRFpred.* Methods Mol Biol, 2018. **1851**.
106. Porollo, A. and J. Meller, *Prediction-based fingerprints of protein-protein interactions.* Proteins-Structure Function and Bioinformatics, 2007. **66**(3): p. 630-645.
107. Du, X.Q., J.X. Cheng, and J. Song, *Improved Prediction of Protein Binding Sites from Sequences Using Genetic Algorithm.* Protein Journal, 2009. **28**(6): p. 273-280.
108. Murakami, Y. and K. Mizuguchi, *Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites.* Bioinformatics, 2010. **26**(15): p. 1841-8.
109. Chen, P. and J. Li, *Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information.* BMC Bioinformatics, 2010. **11**: p. 402.
110. Xue, L.C., D. Dobbs, and V. Honavar, *HomPPI: a class of sequence homology based protein-protein interface prediction methods.* BMC Bioinformatics, 2011. **12**: p. 244.
111. Wang, D.A., R. Wang, and H. Yan, *Fast prediction of protein-protein interaction sites based on Extreme Learning Machines.* Neurocomputing, 2014. **128**: p. 258-266.
112. Dhole, K., et al., *Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier.* Journal of Theoretical Biology, 2014. **348**: p. 47-54.
113. Singh, G., et al., *SPRINGS: Prediction of Protein-Protein Interaction Sites Using Artificial Neural Networks.* J Proteomics Computational Biol, 2014. **1**.
114. Wei, Z.S., et al., *A Cascade Random Forests Algorithm for Predicting Protein-Protein Interaction Sites.* IEEE Trans Nanobioscience, 2015. **14**(7): p. 746-60.
115. Geng, H.J., et al., *Prediction of Protein-Protein Interaction Sites Based on Naive Bayes Classifier.* Biochemistry Research International, 2015.
116. Liu, G.H., H.B. Shen, and D.J. Yu, *Prediction of Protein-Protein Interaction Sites with Machine-Learning-Based Data-Cleaning and Post-Filtering Procedures.* Journal of Membrane Biology, 2016. **249**(1-2): p. 141-153.

117. Taherzadeh, G., et al., *Sequence-Based Prediction of Protein-Peptide Binding Sites Using Support Vector Machine.* Journal of Computational Chemistry, 2016. **37**(13): p. 1223-1229.
118. Tahir, M. and M. Hayat, *Machine learning based identification of protein-protein interactions using derived features of physiochemical properties and evolutionary profiles.* Artificial Intelligence in Medicine, 2017. **78**: p. 61-71.
119. Xue, B., A.K. Dunker, and V.N. Uversky, *Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction.* International Journal of Molecular Sciences, 2010. **11**(10): p. 3725-3747.
120. Mooney, C., et al., *Prediction of Short Linear Protein Binding Regions.* Journal of Molecular Biology, 2012. **415**(1): p. 193-204.
121. Khan, W., et al., *Predicting Binding within Disordered Protein Regions to Structurally Characterised Peptide-Binding Domains.* Plos One, 2013. **8**(9).
122. Wang, Y., et al., *A sequence-based computational method for prediction of MoRFs.* Rsc Advances, 2017. **7**(31): p. 18937-18945.
123. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
124. Zhang, H., et al., *Critical assessment of high-throughput standalone methods for secondary structure prediction.* Brief Bioinform, 2011. **12**(6): p. 672-88.
125. Meng, F. and L. Kurgan, *Computational Prediction of Protein Secondary Structure from Sequence.* Curr Protoc Protein Sci, 2016. **86**: p. 2 3 1-2 3 10.
126. Chen, K. and L. Kurgan, *Computational prediction of secondary and supersecondary structures.* Methods Mol Biol, 2013. **932**: p. 63-86.
127. Meng, F., V. Uversky, and L. Kurgan, *Computational Prediction of Intrinsic Disorder in Proteins.* Curr Protoc Protein Sci, 2017. **88**: p. 2 16 1-2 16 14.
128. Buchan, D.W., et al., *Scalable web services for the PSIPRED Protein Analysis Workbench.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W349-57.
129. Rost, B., *Prediction In 1D: secondary structure, membrane helices, and accessibility In Structural Bioinformatics*, in *Structural Bioinformatics*, P. Bourne and H. Weissig, Editors. 2002, Wiley New Jersey, USA. p. 559-588.
130. Bau, D., et al., *Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins.* BMC Bioinformatics, 2006. **7**: p. 402.
131. Hu, G., et al., *Quality Assessment for the Putative Intrinsic Disorder in Proteins.* Bioinformatics, 2018.
132. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder.* Bioinformatics, 2015. **31**(2): p. 201-8.

133. Peng, Z.L. and L. Kurgan, *Comprehensive comparative assessment of in-silico predictors of disordered regions.* Curr Protein Pept Sci, 2012. **13**(1): p. 6-18.
134. Peng, Z. and L. Kurgan, *On the complementarity of the consensus-based disorder prediction.* Pac Symp Biocomput, 2012: p. 176-87.
135. Deng, X., J. Eickholt, and J. Cheng, *A comprehensive overview of computational protein disorder prediction methods.* Mol Biosyst, 2012. **8**(1): p. 114-21.
136. Dosztanyi, Z., B. Meszaros, and I. Simon, *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins.* Briefings in Bioinformatics, 2010. **11**(2): p. 225-243.