Computational and Structural Biotechnology Journal 17 (2019) 454–462

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Computational Prediction of MoRFs, Short Disorder-to-order Transitioning Protein Binding Regions

Akila Katuwawala [a], Zhenling Peng [b], Jianyi Yang [c], Lukasz Kurgan [a,*]

[a] Department of Computer Science, Virginia Commonwealth University, USA
[b] Center for Applied Mathematics, Tianjin University, Tianjin, China
[c] School of Mathematical Sciences, Nankai University, Tianjin, China

ARTICLE INFO

Article history:
Received 6 February 2019
Received in revised form 22 March 2019
Accepted 23 March 2019
Available online 26 March 2019

Keywords:
Intrinsic disorder
Intrinsically disordered regions
Molecular recognition features
Disordered protein binding
Short linear motifs
Semi-disorder
Protein-protein interactions

ABSTRACT

Molecular recognition features (MoRFs) are short protein-binding regions that undergo disorder-to-order transitions (induced folding) upon binding protein partners. These regions are abundant in nature and can be predicted from protein sequences based on their distinctive sequence signatures. This first-of-its-kind survey covers 14 MoRF predictors and six related methods for the prediction of short protein-binding linear motifs, disordered protein-binding regions and semi-disordered regions. We show that the development of MoRF predictors has accelerated in the recent years. These predictors depend on machine learning-derived models that were generated using training datasets where MoRFs are annotated using putative disorder. Our analysis reveals that they generate accurate predictions. We identified eight methods that offer area under the ROC curve (AUC) ≥ 0.7 on experimentally-validated test datasets. We show that modern MoRF predictors accurately find experimentally annotated MoRFs even though they were trained using the putative disorder annotations. They are relatively highly-cited, particularly the methods available as webservers that on average secure three times more citations than methods without this option. MoRF predictions contribute to the experimental discovery of protein-protein interactions, annotation of protein functions and computational analysis of a variety of proteomes, protein families, and pathways. We outline future development and application directions for these tools, stressing the importance to develop novel tools that would target interactions of disordered regions with other types of partners.

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contents

1. Introduction

Intrinsically disordered regions (IDRs) are absent a well-defined structure under physiological conditions and instead they take shape of heterogeneous conformational ensembles [1–3]. Recent computational analyses estimate that about 30–50% of eukaryotic proteins (depending on the specific organism) have one or more long (having at

* Corresponding author at: Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, Virginia 23284, USA.
E-mail address: lkurgan@vcu.edu (L. Kurgan).

https://doi.org/10.1016/j.csbj.2019.03.013
2001-0370/© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

least 30 consecutive residues) IDRs [4,5]. Intrinsic disorder is also one of the major factors that define dark proteomes [6,7]. The structural plasticity of IDRs facilitates efficient and promiscuous interactions with structurally distinct targets [8,9]. Correspondingly, functional repertoire of proteins with IDRs is largely driven by interactions with proteins and nucleic acids, and includes molecular assembly and recognition, signalling, regulation, transcription and translation [10–19]. These functions complement the cellular functions of structured proteins that are often involved in small molecule binding, transport and catalysis [20].

Proteins with IDRs are particularly important in the context of protein-protein interactions (PPIs). Hub proteins, which are defined as proteins that interact with a large number of proteins in the PPI networks, are enriched in IDRs when compared to the other proteins [21–26]. This stems from the conformation plasticity and the ability of IDRs to undergo disorder-to-order transitions (induced folding) concomitant with their functional activity [16,27–33]. Moreover, a single IDR is capable of interacting with several partners while potentially folding into different conformations [29,33–35]. Here, we focus on molecular recognition features (MoRFs), which are short binding regions (between 5 and 25 residues in length) that are located within longer IDRs and that undergo disorder-to-order transitions upon binding their protein partners [36]. While MoRFs are unstructured in their unbound state, upon binding they morph into well-defined structures that may include helical and strand conformations, often with partner-dependent conformational differences [33]. Correspondingly, MoRF regions are categorized into four types: α-MoRFs that fold into helical conformation, β-MoRFs that fold into β strands, γ-MoRFs transition into coils, and complex-MoRFs that fold into regions with multiple secondary structures [36]. Fig. 1 shows two MoRF regions located in the sequence of the T-cell surface glycoprotein CD3 (UniProt id: P20963), which is one of the key players in the adaptive immune response. These MoRFs were annotated using the structures of protein-protein complexes from the Protein Data Bank (PBD) [37] that are shown in Fig. 1. The first MoRF region (Ala-63 to Asp-87) participates in three diverse PPIs with the Nef protein (Ala-63 to Gly-78 segment that folds upon interaction into the α-MoRF), Tyrosine kinase (Leu-71 to Asp-87 segment that folds into the γ-MoRF) and Tyrosine phosphatase (Arg-80 to Val-85 segment that folds into the γ-MoRF). The second MoRF region (Gly-137 to Lys-150) interacts with the SH2 domain of SHC protein and folds into the γ-MoRF. This example clearly demonstrates that a single MoRF region is capable of binding to a structurally diverse set of protein partners by folding into multiple, different conformations.

A recent computational study has analyzed abundance of MoRFs in 868 complete proteomes, showing that estimated 21% of IDRs in Eukaryota and 29% in Bacteria and Archaea have MoRFs [9]. These abundant short disordered protein-binding regions have originally been studied using computational approaches that relied on the analysis of disorder predictions [38–41] and short sequence motifs associated with protein-binding [42–45]. The latter approach depends on finding over-represented short sequence patterns among a collection of different sequences that bind to a common protein partner [34,42,43,46]. The former approach, which pre-dates the motif-based methodology, is based on an observation that certain putative IDRs include regions with increased structural propensity. While initially these were treated as prediction errors, further analysis has revealed that they often correspond to protein binding sites [47]. MoRFs have unique sequence signatures that differ from the other disordered regions and structured regions, therefore allowing for accurate sequence-based computational prediction [9]. For instance, MoRF regions are enriched in amino acids with large hydrophobic side chains, especially aromatics, when compared with the flanking IDRs. These types of patterns motivated the development of computational predictors of MoRFs [48]. Experimentalists use these methods to support discovery of PPIs [49,50] and in fact MoRF predictions have been often used for this purpose on numerous occasions [51–62]. Knowledge of putative MoRFs also contributes to the elucidation of protein functions [63] and has been used to facilitate analysis of multiple viral proteomes [64–69], cell death pathways [70,71], interactomes of channel proteins [72], kinases [73], nucleosome [14] and ribosome [13].

While some of the MoRF predictors were mentioned in the context of a couple of recent articles that discuss prediction of functions of IDRs [48,74], they were never systematically surveyed. This is the first comprehensive review of computational MoRF predictors. We compare results generated by several MoRF predictors for the same T-cell surface glycoprotein CD3 and contrast them against a set of results produced by a few representative predictors of IDRs. We summarize availability and impact of 14 MoRF predictors, discuss their predictive models, and compare their predictive performance on two benchmark datasets. We also discuss several other computational tools that make predictions of similar types of disordered protein-binding regions.

2. Prediction of MoRFs

MoRF predictors identify putative MoRF regions in an input protein sequence. Fig. 2 visualizes such predictions for the sequence of the T-cell surface glycoprotein CD3 that was introduced in Fig. 1. This T-cell receptor is largely disordered based on experimental annotations [75,76] that we collected from the DisProt resource [77,78] (DisProt id: DP00200). The putative annotations of disorder that were produced with three state-of-the-art disorder predictors [79,80]: MFDp [81–83], VSL2B [84], and PrDOS [85], are in good agreement with each other and with the native annotations of IDRs. While these methods can

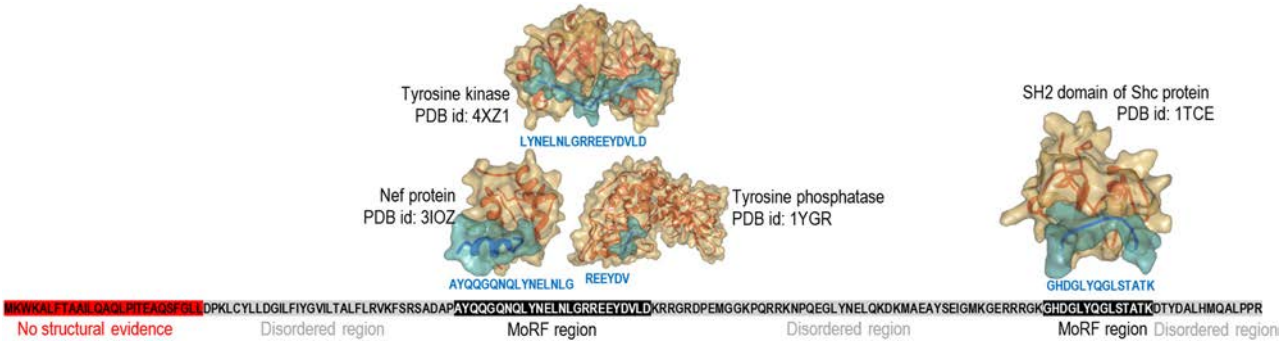

Fig. 1. Interactions between the MoRF regions in the T-cell surface glycoprotein CD3 zeta chain (UniProt id: P20963) and its interactors: Nef protein (UniProt id: Q5QGG3; PDB id: 3IOZ), Tyrosine kinase (UniProt id: P43403; PDB id: 4XZ1), Tyrosine phosphatase (UniProt id: P08575; PDB id: 1YGR), and SH2 domain of SHC protein (UniProt id: P29353; PDB id: 1TCE). The first MoRF region (Ala-63 to Asp-87) interact with the Nef protein (Ala-63 to Gly-78 segment), Tyrosine kinase (Leu-71 to Asp-87 segment) and Tyrosine phosphatase (Arg-80 to Val-85 segment). The second MoRF region (Gly-137 to Lys-150) interacts with the SHC protein. Top of the figure shows structures of the MoRF regions (in blue) in complex with the interactors (in orange). The bottom of the figure provides annotated sequence of the glycoprotein where red region has no evidence of presence/lack of structure, grey regions are annotated as disordered (DisProt id: DP00200), and black regions are MoRFs (annotated based on PDB complexes).

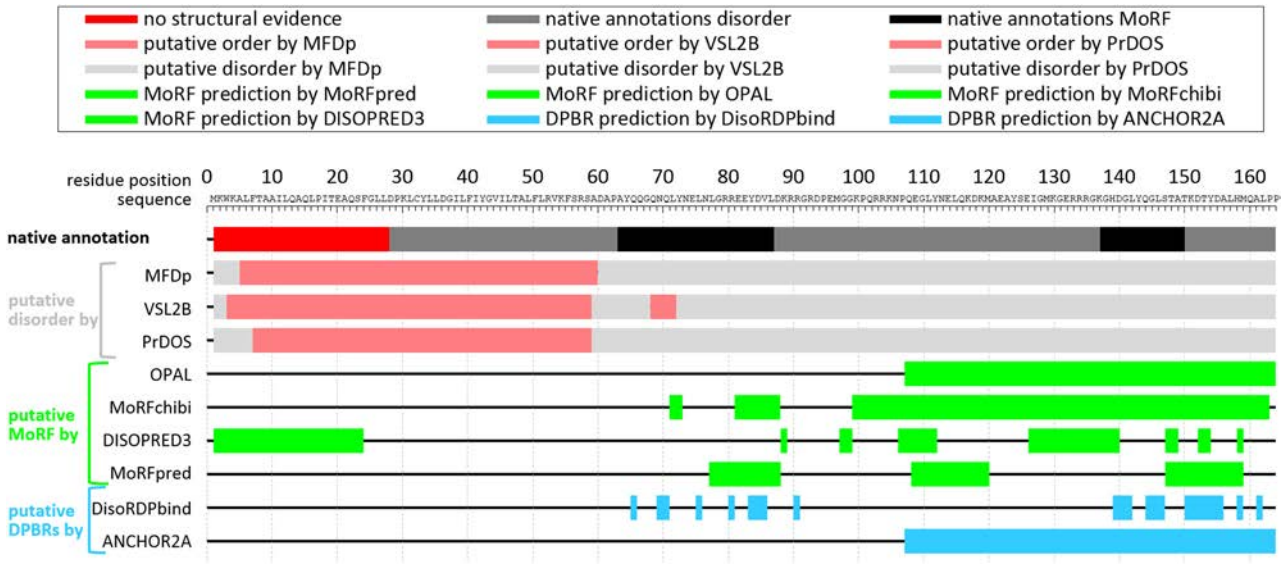

Fig. 2. Prediction of IDRs, MoRFs and disordered protein-binding regions (DPBRs) for the T-cell surface glycoprotein CD3 zeta chain (UniProt id: P20963). The native annotations are shown using horizontal line immediately below the protein sequence at the top of the figure where regions without structural evidence are in red, IDRs are in grey and MoRFs are in black. The following three lines show putative annotations of disorder produced with three leading predictors: MFDp, VSL2B and PrDOS where putative IDRs are in grey and putative structured regions in rose. The next four lines give the putative MoRFs generated with MoRFpred, OPAL, MoRFchibi and DISOPRED3 (in green). The two lines at the bottom correspond to putative DBPRs predicted with DisoRDPbind and ANCHOR2A (in blue).

accurately identify IDRs, they are clearly incapable of finding the two MoRFs that are identified in black in Fig. 2. We use four representative MoRF predictors to generate putative MoRF regions: MoRFpred [86,87], MoRHchibi [88,89], DISOPRED3 [90] and OPAL [91]. The corresponding green lines in Fig. 2 reveal that each of these methods identifies putative MoRF regions in this protein and that these predictions are inside the experimentally annotated IDR, except for DISOPRED3 that finds MoRF in the N-terminus that lacks structural/disorder annotations. While they correctly identify presence of the MoRF regions, only some of them localize these regions in good agreement with the native annotations. In particular, the predictions from MoRFchibi and MoRFpred overlap with the two native MoRFs, although neither of them finds the entire first MoRF region (Ala-63 to Asp-87). While not perfect, these predictions correctly suggest presence of MoRFs and even provide their approximate location in the sequence.

We also contrast MoRF predictions with results of two methods that target the prediction of a more generic set of disordered protein-binding regions (DPBRs), which cover MoRFs and other disordered protein binding domains that are longer than 25 residues [92]. The results generated by the two predictors: ANCHOR2A [39,41,93] and DisoRDPbind [94,95], are shown in blue in Fig. 2. ANCHOR2A finds a protein-binding region (positions Gln-107 to Arg-164) that overlaps with the second MoRF (positions Gly-137 to Lys-150). DisoRBPbind finds two clusters of disordered protein-binding residues that neatly overlap with the location of both MoRF regions. While these two tools are successful in identifying MoRFs for this protein, they are not meant to specifically predict MoRFs, and so these predictions could be misclassified as longer protein-binding domains.

2.1. MoRF Predictors

Over a dozen MoRF predictors was developed during the last 14 years. The first method, α-MoRFpred [40,96], which was published in 2005, is focused exclusively on the prediction of the α-MoRFs. This is motivated by an empirical observation that α-MoRFs can be relatively easily extracted from the disorder predictions generated by the VL3 [97] and VSL2 methods [84]. VL3 is a neural network-based model that was designed to address prediction of variously characterized long IDRs, and which improves over the VL2 version that applies a simpler regression-based model. The letters V and L in the name stand for variously and long, respectively, where the long regions are defined to have at least 30 consecutive residues. The VSL2 method combines two disorder predictors that were optimized to predict short (letter S in the name; these regions have <30 residues in length) and long IDRs. Authors of α-MoRFpred have found that α-MoRFs correspond to regions with high propensity for structural conformation localized inside longer IDRs produced by VL3 and VSL2, i.e., regions of lower putative propensity for disorder flanked by regions with high putative propensity for disorder.

The second predictor, Retro-MoRF [98], was published in 2010. It combines disorder predictions with sequence alignment. In essense, Retro-MoRF extracts putative MoRF regions utilizing the idea behing α-MoRFpred, and next these regions are aligned against structured sequences from PDB and functionally annotated sequences from SwissProt [99]. These alignments are used to determine whether the original α-MoRF prediction should be accepted or refuted as a false positive. However, Retro-MoRF was tested on a small set of several proteins, the underlying algorithm and implementation were not released publically, and it remains unclear whether this method could predict the other types of MoRFs, besides the α-MoRFs.

The first publically available computational tool that covers prediction of all MoRF types is MoRFpred [86,87]. This method was released in 2012. MoRFpred applies predictive model that was produced with a machine learning algorithm, Support Vector Machine (SVM), from a large dataset of MoRFs extracted from protein-protein complexes in PDB. Interestingly, this training dataset relies on putative annotations of disordered regions; i.e., MoRFs in the training dataset are short protein-binding sequence segments which are located within longer predicted IDRs. The use of (arguably accurately) predicted IDRs to annotate MoRFs was necessary given that relatively few experimentally validated IDRs were known at that time. Subsequently published MoRF predictors were also trained using datasets that rely on putative IDRs. In fact, 11 out of the 14 MoRF predictors (except for α-MoRFpred, Retro-MoRF and DISOPRED3) use the same training dataset, which was introduced in [87]. Moreover, MoRFpred and the more recent predictors were evaluated using test datasets which cover all MoRF types. Some of these test datasets rely on the putative IDRs while some other utilize experimentally confirmed MoRFs; i.e., MoRFs located in the experimentally validated IDRs.

While only three predictors were developed between 2005 and 2012, 11 methods were published over the next seven years. Table 1 summarizes the corresponding collection of the 14 MoRF predictors. It reveals that the development efforts have accelerated in recent years, with four predictors that were released in 2016 and an additional four in 2018. The table provides year of publication and details concerning the predictive models, availability and impact. The predictive models can be broadly categorized into two classes: those that rely on machine learning algorithms and those that utilize scoring functions. The scoring function-based methods utilize an ab-initio derived empirical formula or a sequence alignment to make predictions. Only one MoRF predictor, namely Retro-MoRF, depends on this type of model. The machine learning-based methods compute predictive models from a training dataset annotated with MoRF regions. They use machine learning algorithms to optimize the architecture and parameters of the predictive models such that the differences between the outputs of these models and the native MoRF annotations in the training dataset are minimized. After completing the training, the resulting models can be used to predict MoRFs in sequences from outside of the training dataset. All but one MoRF predictor rely on the machine learning-generated models. However, they differ in the type of the machine learning algorithms that they use. The most frequently used algorithm is SVM, which is used by nine MoRF predictors. The remaining methods use the Naïve Bayes algorithm (two predictors) and the neural network algorithm (one predictor). Three of the machine learning-based methods are meta-predictors (Table 1). The meta-predictors use predictions of MoRFs generated by third-party methods as inputs to (re)predict MoRF regions. The underlying goal is to generate results that are more accurate than any of the input MoRF predictions. For instance, the newest OPAL+ method uses MoRF predictions generated by MoRFpred-plus and MoRFchibi as inputs to its SVM model. Correspondingly, the OPAL+ model is shown to be more accurate than these two input predictors on all test datasets [100].

The MoRF predictors are made accessible to the community in two ways: as webservers and/or standalone code. The webservers are arguably easier to use and they are more suitable for less computer savvy users who want to perform ad hoc predictions for a limited number of proteins. The only requirements for the webserver users are to have access to the Internet and to have a modern web browser to connect to the website of the webserver. The predictions are calculated on the server side and the results are returned via email or/and the web browser window. Nine out of 14 MoRF predictors offer this option. We note that the webserver for one of these methods, MFSPSSMpred, is no longer available. The source code option requires the end users to run the predictions on their own hardware. This could be attractive in situations when large datasets of proteins must to be predicted and when the end users would like to embed a given MoRF predictor into a larger bioinformatics pipeline. The source code is available for nine of the 14 MoRF predictors. We note that six methods, including MoRFCHiBi [88], DISOPRED3 [90], MoRFCHiBiLight [89], MoRFCHiBiWeb [89], OPAL [91] and OPAL+ [100], are currently offered as both webservers and source code. Table 1 gives the web links to the webservers and

Table 1
Methods for the prediction of MoRFs and related binding regions including SLiMs (short linear motifs that bind proteins) and disordered protein-binding regions (DPBRs). The methods 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. The 'Citations Total' column gives the number of citations collected from Google Scholar on March 20, 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 'Citations Annual' column gives an average number of citations per year since a given method was published. 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 machine learning models include neural network (NN), support vector machine (SVM), naïve Bayes (NB), and logistic regression (LR).

Target of predictions	Method name	Ref.	Year published	Predictive model	Meta predictor	Availability		Citations	
						Type	URL	Total	Annual
MoRF regions	α-MoRFpred	[40,96]	2005	ML (NN)	No	NA	NA	454	32
	retro-MoRFs	[98]	2010	SF (alignment)	No	NA	NA	27	3
	MoRFpred	[86,87]	2012	ML (SVM)	No	WS	http://biomine.cs.vcu.edu/servers/MoRFpred/	194	28
	MFSPSSMpred	[102]	2013	ML (SVM)	No	WS + SC	The website does not work as of January 2019	32	5
	MoRFCHiBi	[88]	2015	ML (SVM)	No	WS + SC	https://gsponerlab.msl.ubc.ca/software/morf_chibi/	37	9
	DISOPRED3	[90]	2015	ML (SVM)	No	WS + SC	http://bioinf.cs.ucl.ac.uk/disopred	218	54
	fMoRFpred	[9]	2016	ML (SVM)	No	WS	http://biomine.cs.vcu.edu/servers/fMoRFpred/	36	12
	MoRFCHiBiLight	[89]	2016	ML (NB)	No	WS + SC	https://gsponerlab.msl.ubc.ca/software/morf_chibi/	23	8
	MoRFCHiBiWeb	[89]	2016	ML (NB)	Yes	WS + SC	https://gsponerlab.msl.ubc.ca/software/morf_chibi/	23	8
	Predict-MoRFs	[103]	2016	ML (SVM)	No	SC	https://github.com/roneshsharma/Predict-MoRFs	6	2
	Fang et al.	[101]	2018	ML (SVM)	No	NA	NA	0	0
	MoRFPred-plus	[104]	2018	ML (SVM)	No	SC	https://github.com/roneshsharma/MoRFpred-plus/wiki/MoRFpred-plus	8	8
	OPAL	[91]	2018	ML (SVM)	Yes	WS + SC	http://www.alok-ai-lab.com/tools/opal/	9	9
	OPAL+	[100]	2018	ML (SVM)	Yes	WS + SC	http://www.alok-ai-lab.com/tools/opal_plus/	0	0
DPBRs	DisoRDPbind	[94,95]	2015	ML (LR)	No	WS	http://biomine.cs.vcu.edu/servers/DisoRDPbind/	47	12
	ANCHOR	[39,41,93]	2009	SF	No	WS + SC	http://anchor.enzim.hu	395	39
SLiMs	PepBindPred	[105]	2013	ML (NN)	No	WS	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/pepbindpred.php	17	3
	SLiMPred	[106]	2012	ML (NN)	No	WS	http://bioware.ucd.ie/~compass/biowareweb/Server_pages/slimpred.php	55	8
Semi-disorder	SPINE-D	[107]	2013	ML (NN)	No	WS + SC	http://sparks-lab.org/SPINE-D/	32	5
	SPOT-Disorder	[108]	2017	ML (NN)	No	WS + SC	http://sparks-lab.org/server/SPOT-disorder/	47	23

source codes. The implementations for three methods, the two earliest tools (α-MoRFpred and Retro-MoRF) and the method developed by Fang et al. in 2018 [101], are inaccessible to the public. They can be obtained only by directly contacting the authors.

Table 1 quantifies citations, which is one of the key measures of impact for the MoRF predictors. To avoid duplicate counting we use the reference with the highest citation count for the methods that were published in multiple articles. The table lists the total and the annual number of citations which we collected from Google Scholar. The 14 tools have accumulated a total of 1067 citations, with a respectable median of 25 citations (annual median = 8). Three methods were cited over 100 times: α-MoRFpred (total: 454, annually: 32), DISOPRED3 (total: 218, annually: 54) and MoRFpred (total: 194, annually: 28). Moreover, we found that predictors that are available as webservers are cited substantially more often compared to the methods that do not offer this option. Using the annual citation counts, which are more suitable for the comparisons between methods, the median of annual citations for the methods that have webservers is 9 vs. 3 for the other predictors. The difference in the corresponding medians of the total citations is even larger: 32 vs. 8.

2.2. Related Predictors of Disordered Protein-binding Regions

We also briefly discuss several related computational methods that target prediction of disordered protein-binding regions (DPBRs), short linear motifs (SLiMs), and semi-disordered regions.

DPBRs cover MoRF regions and longer disordered protein-binding domains. There are currently two predictors of DPBRs: ANCHOR [39,41,93] and DisoRDPbind [94,95]. Table 1 reveals that they are well-cited and available as webservers. ANCHOR is a scoring function-based method that implements an empirical calculation of propensity for protein binding in putative disordered regions, drawing from the methodology underlying a popular disorder predictor, IUpred [109]. In contrast, DisoRDPbind is a machine learning-based method that uses the logistic regression model. Besides predicting DPBRs, this is the first method that provides predictions of disordered RNA-binding and disordered DNA-binding regions. The RNA-binding regions generated with DisoRDPbind were recently used to derive the arguably most complete to date collection of putative RNA-binding proteins in the human proteome [110].

SLiMs are short sequence motifs in eukaryotic proteins that are associated with protein binding events. While most SLiMs are localized in IDRs, approximately 20% of them are associated with protein-protein interactions in structured regions [89]. A collection of over 3000 SLiMs curated from literature is available in the ELM resource [43,45]. The two SLiM predictors, PepBindPred [105] and SLiMPred [106], utilize machine learning-derived neural network models. PepBindPred's model was derived using training datasets of SLiMs that were filtered to be embedded within putative IDRs, representing a motif-associated subpopulation of MoRFs. The main difference is that MoRFs do not have to be associated with sequence motifs that, by definition, must occur across multiple proteins. PepBindPred relies on protein-protein docking and requires structure of the protein that binds to the SLiM region as the input. While this may improve quality of the predictions, it also increases computational cost of making predictions when compared to the MoRF and DPBR predictors that do not use docking. It also limits applications of PepBindPred to scenarios where the structure is available. SLiMPred predicts SLiMs in protein sequences (i.e., it does not need the structure as its input). However, its predictions do not distinguish between motifs located in IDRs and in structured regions. Consequently, SLiMPred's outputs partially cover MoRFs (those associated with motifs) and they also include short structured protein-binding regions.

Semi-disordered regions are the regions that are predicted midway between being disordered and structured; i.e., they are predicted with 50% probability to be disordered [107]. Recent study shows that the semi-disordered regions are partially collapsed and have intermediate levels of predicted solvent accessibility [107]. This work also suggests that these regions are linked to the induced folding and that the corresponding predictions can be used to identify MoRF regions. Two disorder predictors can be used to predict the semi-disordered regions: SPINE-D [107,111] and SPOT-Disorder [108]. Both methods rely on machine learning-derived neural network models, though SPOT-Disorder uses a more sophisticated deep recurrent network. Preliminary, small scale tests suggest that SPOT-Disorder can be used to accurately predict MoRFs [108].

3. Predictive Quality of the MoRF Predictors

Various MoRF predictors use different predictive models, different types of information extracted from the input sequence, and different training datasets. This results in different predictions for the same input sequence, where some methods are expected to be on average more accurate than others.

As we mentioned before, the MoRF predictors are trained using datasets of proteins with MoRF regions located within putative IDRs. A representative set of 11 MoRF predictors uses the same putative IDR-based training dataset from [87]. These predictors were also tested on a consistent set of test datasets that share low sequence similarity (<30%) to this training dataset, ensuring that the corresponding results can be compared across these methods. This also means that a simple sequence alignment could not be used to make accurate predictions on these test datasets. The set of 11 predictors excludes only the two earliest methods that target prediction of α-MoRFs (α-MoRFpred and Retro-MoRF) and DISOPRED3 that was designed to primarily target prediction of disordered regions. Testing of 10 out the 11 tools was done using two types of test datasets, with MoRFs located within the putative IDRs and with MoRFs inside the experimentally verified IDRs. Only the predictor by Fang et al. [101] was never tested using the experimentally verified IDRs, and thus we exclude this methods from our analysis. The most commonly used test dataset, TEST419 [87], includes 419 proteins, where MoRF regions are located within a larger sequence segment that is predicted to be disordered using a protocol from [112]. The two commonly used datasets that rely on the experimentally annotated IDRs are TEST45 [87] and TEST53 [89], which have 45 and 53 proteins, respectively. These three test datasets share the low similarity to the proteins from the training dataset used to develop the 10 MoRF predictors. We use the source references to collect measurements of the predictive performance for these datasets for the 10 MoRF predictors. Our aim is to investigate whether predictive performance have improved over the years and whether the results on the test datasets that rely on putative vs. native disorder annotations are different.

Fig. 3A reports values of the area under the ROC curve (AUC), which ranges between 0.5 (equivalent to random predictions) and 1 (always correct predictions). Fig. 3B compares values of the other two popular measures: sensitivity, which quantifies rate of correct predictions of MoRF residues among all native annotations of MoRF residues; and specificity that quantifies the rate of correct predictions among the native non-MoRF residues. These three measures were used to assess majority of the MoRF predictors [87–91,100–104]. Inspired by the comparative analyses in [87–89,91,100,103,104], we report sensitivity values that are calibrated between different methods to the same value of the false positive rate, which we set to 0.1. We similarly calibrate the specificity values to the same true positive rate = 0.5. This way, these measurements can be directly compared between different predictors. Both figures compare the predictive performance on the TEST419 (using putative IDRs) against the results on the test datasets that rely on experimental IDRs (either TEST45 or TEST53, whichever results is available) across the 10 MoRF predictors.

Fig. 3A shows a relatively wide range of AUC values, from about 0.65 to 0.84. Fig. 3B reveals that the 10 MoRF predictors secure high sensitivity values between 0.31 and 0.55, relative to the corresponding low false positive rate = 0.1. It also demonstrates that they obtain high specificity

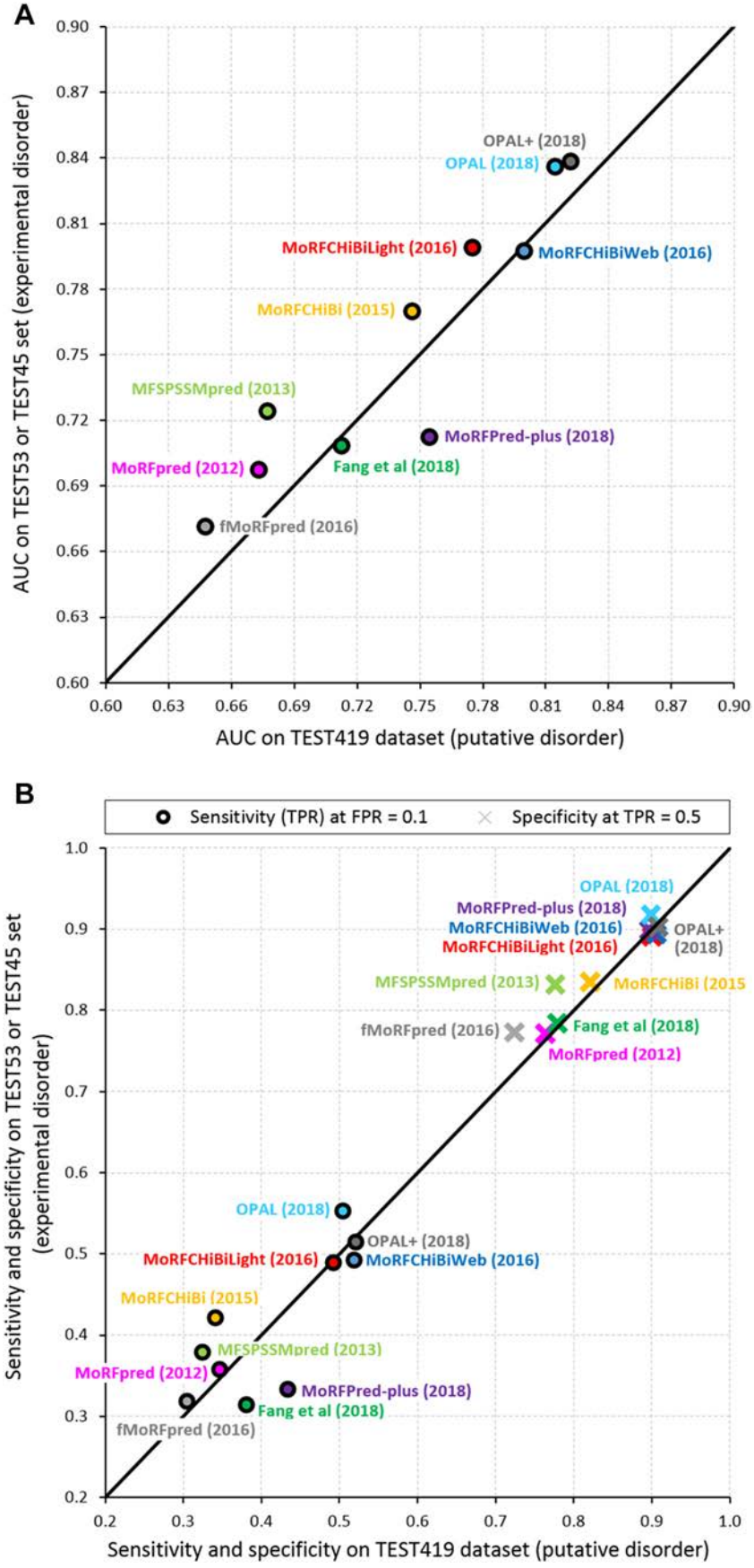

Fig. 3. Predictive quality for the predictors of MoRF regions measured on the TEST419 dataset (for which MoRF annotations are based on putative disorder) and TEST53/TEST45 (for which MoRF annotations are based on experimentally verified disorder). Panel A shows the AUC values. Panel B gives the values of sensitivity (measured for FRP = 0.1 and shown using circles) and specificity (measured for TRP = 0.5 and shown with crosses). The results were taken from the original publications. All predictors were developed using the same training dataset (TRAIN419) that shares low sequence similarity (<30%) with these test datasets.

that ranges between 0.72 and 0.92, relative to the corresponding true positive rate = 0.5. We argue that all ten methods provide reasonably accurate predictions; i.e., AUC ≥ 0.65; sensitivity that is much higher than the corresponding false positive rate, and specificity that is much higher than the corresponding true positive rate. The most accurate predictors on these benchmark test datasets are OPAL+, OPAL, MoRFchibiLight and MoRFchibiWeb. These methods were developed recently and they offer high values for all three measures of predictive performance. Three of these methods (OPAL+, OPAL and MoRFchibiWeb) are meta-predictors (Table 1), suggesting that this type of predictive architecture provides promising results for the prediction of MoRF regions. Moreover, as expected, our analysis reveals that the predictive performance continues to improve. The three predictors that were published between 2012 and 2015 secure an average AUC = 0.70 on the TEST419 dataset, compared to 0.74 for the three methods published in 2016 and 0.78 for the four methods from 2018. The corresponding average AUCs that were measured on the experimentally annotated test datasets are 0.73, 0.76 and 0.77.

The relatively low predictive performance of fMoRFpred (Fig. 3) can be explained by fact that it was designed to provide fast predictions [9]. Runtime analyses reveal that the three fastest MoRF predictors: fMoRFpred, MoRFCHiBi and MoRFCHiBiLight, predict an average size protein chain (300 amino acids long) in about 1 s [9], 1.6 s [89] and 1.7 s [89], respectively. To compare, MoRFPred-plus, MoRFCHiBiWeb, OPAL and OPAL+ would take approximately 34 s [100], 36 s [89], 84 s [100] and 2 min [100], respectively. These longer runtimes are primarily caused by the high computational cost of running multiple sequence alignments, which are required to produce some of the inputs used by these predictors. Moreover, we observe that three of the most accurate predictors (MoRFCHiBiWeb, OPAL and OPAL+) require at least an order of magnitude more runtime compared to the fastest fMoRFpred.

Interestingly, a majority of the results are located at or above the diagonal line in Fig. 3A and B. This means that these AUCs, sensitivities and specificities are the same or better on the test dataset that relies on the experimentally validated disorder annotations when compared to the test dataset that uses putative IDRs. This trend reveals that the current MoRF predictors accurately identify experimentally annotated MoRFs in spite of the fact that they are trained using the dataset with the putative annotations. The lower AUCs on the TEST419 datasets are possibly because some of the MoRF annotations in this dataset might be incorrect resulting in partially incorrect measurement of predictive quality, which in turn effectively depresses AUC, sensitivity and specificity values.

4. Summary and Outlook

MoRF regions are highly abundant across all domains of life. They have unique sequence signatures that facilitate the development of accurate computational predictors of MoRFs. These predictions were used to assist experimental discovery of PPIs, generate putative protein functions, and facilitate computational analysis of a variety of proteomes, pathways, and protein families. We survey a comprehensive collection of 14 MoRF predictors. Our study reveals that the development of these methods has accelerated in recent years, resulting in the release of eight tools in the last three years. MoRF predictors rely primarily on machine learning-derived predictive models that are generated using training datasets where MoRFs are annotated using putative IDRs. We demonstrate that these computational tools are well-cited and that most of them are available as convenient to use webservers. Our analysis also shows that they produce accurate predictions on test datasets that use both putative and experimental annotations of disorder. We highlight the empirical observation that they accurately identify experimentally annotated MoRFs in spite of the fact that they were trained using datasets with putative annotations. The most accurate methods are meta-predictors but they also require the longest runtime. On the other hand, the fastest method, fMoRFpred, is shown to generate the least accurate results.

Our survey reveals that the underlying predictive models are rather homogeneous, as they almost always use the SVM model. This is true for all methods that were published in 2018. With the advent of deep learning models in bioinformatics [113], we believe that these neural network architectures should be tried to further improve the accuracy of the MoRF predictions. This claim is supported by the fact that a few accurate deep learning models that predict residue-level protein interactions were recently published, including the predictor of residue-residue contacts in protein structures [114], and the predictor of residue-residue interactions in protein complexes [115].

We stress the fact that IDRs carry out many cellular functions that require interactions with a wide range of partners. IDRs are involved in protein-protein, protein-DNA, protein-RNA, protein-lipid, and a variety of protein-small ligand interactions. Numerous examples of these interactions are available in the DisProt resource [77,78]. A substantial collection of disordered protein-protein and protein-nucleic acids interfaces was recently studied [116], suggesting that large training datasets can be assembled. While over a dozen predictors of MoRFs regions is available, we note that there are very few methods that address prediction of interactions of IDRs with the other partners. Notable examples include DisoRDPbind that predicts disordered protein-RNA and protein-DNA binding regions [94,95], DFLpred that predicts disordered linker regions [117], and DMRpred that predicts disordered moonlighting (multifunctional) regions [118]. More methods that would cover the other types of partners are needed.

Finally, recent research advocates the development of quality assessment scores for the disorder predictions [119]. These scores indicate which residue-level predictions are more likely to be accurate, therefore suggesting which parts of the predictions are more trustworthy. The scores are calculated by a separate predictive model that uses the predicted disorder as the input. We observe that the development of the quality assessment tools is already a well-researched and developed topic in the protein structure prediction area [120–122]. A recently developed method, QUARTER, generates the quality assessment scores for ten different predictors of IDRs [123]. We believe that the MoRF predictors would also benefit from the availability of the quality assessment scores.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgment

This research was supported in part by the National Science Foundation, USA (grant 1617369), the National Natural Science Foundation of China (grant 61832019, 11871290 and 61873185), China Scholarship Council, KLMDASR of China, and the Robert J. Mattauch Endowment Funds.

References

[1] Lieutaud P, Ferron F, Uversky AV, Kurgan L, 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:e1259708.
[2] Dunker AK, Babu MM, Barbar E, Blackledge M, et al. What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins 2013;1:e24157.
[3] Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev 2014;114:6561–88.
[4] Peng Z, Yan J, Fan X, Mizianty MJ, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 2015;72:137–51.
[5] Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J Biomol Struct Dyn 2012;30:137–49.

[6] Hu G, Wang K, Song J, Uversky VN, Kurgan L. Taxonomic landscape of the dark proteomes: Whole-proteome scale interplay between structural darkness, intrinsic disorder, and crystallization propensity. Proteomics 2018;18(21-22) e1800243.
[7] Kulkarni P, Uversky VN. Intrinsically disordered proteins: the dark horse of the dark proteome. Proteomics 2018;18:e1800061.
[8] Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–31.
[9] Yan J, Dunker AK, Uversky VN, Kurgan L. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst 2016;12:697–710.
[10] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
[11] Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 2005;18:343–84.
[12] Liu J, Perumal NB, Oldfield CJ, Su EW, et al. Intrinsic disorder in transcription factors. Biochemistry 2006;45:6873–88.
[13] Peng Z, Oldfield CJ, Xue B, Mizianty MJ, et al. A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell Mol Life Sci 2014;71:1477–504.
[14] Peng Z, Mizianty MJ, Xue B, Kurgan L, Uversky VN. More than just tails: intrinsic disorder in histone proteins. Mol Biosyst 2012;8:1886–901.
[15] Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323:573–84.
[16] Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets: the roles of intrinsic disorder in protein interaction networks. FEBS J 2005;272:5129–48.
[17] Meng F, Na I, Kurgan L, Uversky VN. Compartmentalization and functionality of nuclear disorder: intrinsic disorder and protein-protein interactions in intra-nuclear compartments. Int J Mol Sci 2016;17:24.
[18] Na I, Meng F, Kurgan L, Uversky VN. Autophagy-related intrinsically disordered proteins in intra-nuclear compartments. Mol Biosyst 2016;12:2798–817.
[19] Wang C, Uversky VN, Kurgan L. Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. Proteomics 2016;16:1486–98.
[20] Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, et al. Intrinsic disorder and functional proteomics. Biophys J 2007;92:1439–56.
[21] Hu G, Wu Z, Uversky VN, Kurgan L. Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. Int J Mol Sci 2017;18:2761.
[22] Haynes C, Oldfield CJ, Ji F, Klitgord N, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2006;2:890–901.
[23] Ekman D, Light S, Bjorklund AK, Elofsson A. What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biol 2006;7:R45.
[24] Kim PM, Sboner A, Xia Y, Gerstein M. The role of disorder in interaction networks: a structural analysis. Mol Syst Biol 2008;4:179.
[25] Higurashi M, Ishida T, Kinoshita K. Identification of transient hub proteins and the possible structural basis for their multiple interactions. Protein Sci 2008;17:72–8.
[26] Patil A, Kinoshita K, Nakamura H. Domain distribution and intrinsic disorder in hubs in the human protein-protein interaction network. Protein Sci 2010;19:1461–8.
[27] Pontius BW. Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association. Trends Biochem Sci 1993;18:181–6.
[28] Dunker AK, Obradovic Z. The protein trinity–linking function and disorder. Nat Biotechnol 2001;19:805–6.
[29] Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60.
[30] Plaxco KW, Gross M. Cell biology. The importance of being unfolded. Nature 1997;386:657–9.
[31] Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 2005;18:343–84.
[32] Fuxreiter M, Toth-Petroczy A, Kraut DA, Matouschek AT, et al. Disordered proteinaceous machines. Chem Rev 2014;114:6806–43.
[33] Oldfield CJ, Meng J, Yang JY, Yang MQ, et al. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genom 2008;9(Suppl. 1):S1.
[34] Hsu WL, Oldfield CJ, Xue B, Meng J, et al. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Sci 2013;22:258–73.
[35] Dunker AK, Garner E, Guilliot S, Romero P, et al. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pac Symp Biocomput 1998:473–84.
[36] Mohan A, Oldfield CJ, Radivojac P, Vacic V, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006;362:1043–59.
[37] Burley SK, Berman HM, Kleywegt GJ, Markley JL, et al. Protein Data Bank (PDB): the single global macromolecular structure archive. Methods Mol Biol 1607;2017:627–41.
[38] Oldfield CJ, Cheng Y, Cortese MS, Romero P, et al. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 2005;44:12454–70.
[39] Meszaros B, Simon I, Dosztanyi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 2009;5:e1000376.
[40] Cheng Y, Oldfield CJ, Meng J, Romero P, et al. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry 2007;46:13468–77.
[41] Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 2009;25:2745–6.
[42] Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31:3635–41.
[43] Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, et al. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–30.
[44] Davey NE, Edwards RJ, Shields DC. The SLiMDisc server: short, linear motif discovery in proteins. Nucleic Acids Res 2007;35:W455–9.
[45] Dinkel H, Van Roey K, Michael S, Kumar M, et al. ELM 2016–data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res 2016;44:D294–300.
[46] Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev 2014;114:6733–78.
[47] Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. Predicting binding regions within disordered proteins. Genome Inform Ser Workshop Genome Inform 1999;10:41–50.
[48] Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 2017;74:3069–90.
[49] Ehrenberger T, Cantley LC, Yaffe MB. Computational prediction of protein-protein interactions. Methods Mol Biol 2015;1278:57–75.
[50] Valencia A. Pazos, F. In: Panchenko A, Przytycka TM, editors. Protein-protein interactions and networks. London: Springer-Verlag; 2008. p. 67–81.
[51] Callaghan AJ, Aurikko JP, Ilag LL, Gunter Grossmann J, et al. Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E. J Mol Biol 2004;340:965–79.
[52] Bourhis JM, Johansson K, Receveur-Brechot V, Oldfield CJ, et al. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. Virus Res 2004;99:157–67.
[53] Dolan PT, Roth AP, Xue B, Sun R, et al. Intrinsic disorder mediates hepatitis C virus core-host cell protein interactions. Protein Sci 2015;24:221–35.
[54] Nyarko A, Song Y, Novacek J, Zidek L, Barbar E. Multiple recognition motifs in nucleoporin Nup159 provide a stable and rigid Nup159-Dyn2 assembly. J Biol Chem 2013;288:2614–22.
[55] Micaroni M, Giacchetti G, Plebani R, Xiao GG, Federici L. ATP2C1 gene mutations in Hailey-Hailey disease and possible roles of SPCA1 isoforms in membrane trafficking. Cell Death Dis 2016;7:e2259.
[56] O'Shea C, Staby L, Bendsen SK, Tidemand FG, et al. Structures and short linear motif of disordered transcription factor regions provide clues to the interactome of the cellular hub protein radical-induced cell Death1. J Biol Chem 2017;292:512–27.
[57] Ulrich AKC, Seeger M, Schutze T, Bartlick N, Wahl MC. Scaffolding in the spliceosome via single alpha helices. Structure 2016;24:1972–83.
[58] Canales A, Rosinger M, Sastre J, Felli IC, et al. Hidden alpha-helical propensity segments within disordered regions of the transcriptional activator CHOP. Plos One 2017;12.
[59] Jamsheer KM, Shukla BN, Jindal S, Gopan N, et al. The FCS-like zinc finger scaffold of the kinase SnRK1 is formed by the coordinated actions of the FLZ domain and intrinsically disordered regions. J Biol Chem 2018;293:13134–50.
[60] Pozo PN, Cook JG, et al. Regulation and Function of Cdt1. A Key Factor in Cell Proliferation and Genome Stability, Genes 2018;293:13134–50.
[61] Pujols J, Santos J, Pallares I, Ventura S. The disordered C-terminus of yeast Hsf1 contains a cryptic low-complexity amyloidogenic region. Int J Mol Sci 2018;19:1384.
[62] Shiina N, Nakayama K. RNA granule assembly and disassembly modulated by nuclear factor associated with double-stranded RNA 2 and nuclear factor 45. J Biol Chem 2014;289:21163–80.
[63] Cozzetto D, Jones DT. The contribution of intrinsic disorder prediction to the elucidation of protein function. Curr Opin Struct Biol 2013;23:467–72.
[64] Mishra PM, Uversky VN, Giri R. Molecular recognition features in Zika virus proteome. J Mol Biol 2018;430:2372–88.
[65] Meng F, Badierah RA, Almehdar HA, Redwan EM, et al. Unstructural biology of the dengue virus proteins. FEBS J 2015;282:3368–94.
[66] Fan X, Xue B, Dolan PT, LaCount DJ, et al. The intrinsic disorder status of the human hepatitis C virus proteome. Mol Biosyst 2014;10:1345–63.
[67] Singh A, Kumar A, Uversky VN, Giri R. Understanding the interactability of chikungunya virus proteins via molecular recognition feature analysis. RSC Adv 2018;8:27293–303.
[68] Charon J, Theil S, Nicaise V, Michon T. Protein intrinsic disorder within the Potyvirus genus: from proteome-wide analysis to functional annotation. Mol Biosyst 2016;12:634–52.
[69] Xue B, Blocquel D, Habchi J, Uversky AV, et al. Structural disorder in viral proteins. Chem Rev 2014;114:6880–911.
[70] Uversky AV, Xue B, Peng Z, Kurgan L, Uversky VN. On the intrinsic disorder status of the major players in programmed cell death pathways. F1000Res 2013;2:190.
[71] Peng Z, Xue B, Kurgan L, Uversky VN. Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. Cell Death Differ 2013;20:1257–67.
[72] Peng Z, Sakai Y, Kurgan L, Sokolowski B, Uversky V. Intrinsic disorder in the BK channel and its interactome. PLoS One 2014;9:e94331.
[73] Kathiriya JJ, Pathak RR, Clayman E, Xue B, et al. Presence and utility of intrinsically disordered regions in kinases. Mol Biosyst 2014;10:2876–88.
[74] Dosztányi Z, Tompa P, in: J. Rigden, D., editors. From protein structure to function with bioinformatics. Dordrecht: Springer Netherlands; 2017. p. 167–203.

[75] Call ME, Schnell JR, Xu C, Lutz RA, et al. The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. Cell 2006;127:355–68.
[76] Sigalov A, Aivazian D, Stern L. Homooligomerization of the cytoplasmic domain of the T cell receptor zeta chain and of other proteins containing the immunoreceptor tyrosine-based activation motif. Biochemistry 2004;2049-2061:43.
[77] Vucetic S, Obradovic Z, Vacic V, Radivojac P, et al. DisProt: a database of protein disorder. Bioinformatics 2005;21:137–40.
[78] Piovesan D, Tabaro F, Micetic I, Necci M, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res 2016;D1:D219–27.
[79] Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. Proteins 2014;82(Suppl. 2):127–37.
[80] Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 2012;13:6–18.
[81] Mizianty MJ, Peng ZL, Kurgan L. MFDp2: accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. Intrinsic Disorder Proteins 2013;1:e24428.
[82] Mizianty MJ, Uversky V, Kurgan L. Prediction of intrinsic disorder in proteins using MFDp2. Methods Mol Biol 2014;1137:147–62.
[83] Mizianty MJ, Stach W, Chen K, Kedarisetti KD, et al. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 2010;26:i489–96.
[84] Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics 2006;7:208.
[85] Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids Res 2007;35:W460–4.
[86] Oldfield CJ, Uversky VN, Kurgan L. Predicting functions of disordered proteins with MoRFpred. Methods Mol Biol 1851;2018.
[87] Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, 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:i75–83.
[88] Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. Bioinformatics 1738-1744;2015:31.
[89] Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res 2016;44:W488–93.
[90] Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 2015;31:857–63.
[91] Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A. OPAL: prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics 1850-1858;2018:34.
[92] Tompa P, Fuxreiter M, Oldfield CJ, Simon I, et al. Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays 2009;31:328–35.
[93] Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res 2018;46:W329–37.
[94] Peng Z, Wang C, Uversky VN, Kurgan L. Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. Methods Mol Biol 2017;1484:187–203.
[95] Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 2015;43:e121.
[96] Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, et al. Comparing and combining predictors of mostly disordered proteins. Biochemistry 2005;44:1989–2000.
[97] Obradovic Z, Peng K, Vucetic S, Radivojac P, et al. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53(Suppl. 6):566–72.
[98] Xue B, Dunker AK, Uversky VN. Retro-MoRFs: identifying protein binding sites by Normal and reverse alignment and intrinsic disorder Prediction. Int J Mol Sci 2010;11:3725–47.
[99] Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol 2007;406:89–112.
[100] Sharma R, Sharma A, Raicar G, Tsunoda T, Patil A. OPAL+: length-specific MoRF prediction in intrinsically disordered protein sequences. Proteomics 2018;19:e1800058.
[101] Fang C, Moriwaki Y, Zhu D, Shimizu K. Proceedings of the 2018 6th international conference on bioinformatics and computational biology. Chengdu, China: ACM; 2018; 50–4.
[102] Fang C, Noguchi T, Tominaga D, Yamana H. MFSPSSMpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinform 2013;14:300.
[103] Sharma R, Kumar S, Tsunoda T, Patil A, Sharma A. Predicting MoRFs in protein sequences using HMM profiles. Bmc Bioinform 2016;17(Suppl 19):504.
[104] Sharma R, Bayarjargal M, Tsunoda T, Patil A, Sharma A. MoRFPred-plus: computational identification of MoRFs in protein sequences using physicochemical properties and HMM profiles. J Theor Biol 2018;437:9–16.
[105] Khan W, Duffy F, Pollastri G, Shields DC, Mooney C. Predicting binding within disordered protein regions to structurally characterised peptide-binding domains. Plos One 2013;8:e72838.
[106] Mooney C, Pollastri G, Shields DC, Haslam NJ. Prediction of short linear protein binding regions. J Mol Biol 2012;415:193–204.
[107] Zhang T, Faraggi E, Li Z, Zhou Y. Intrinsically semi-disordered state and its role in induced folding and protein aggregation. Cell Biochem Biophys 2013;67:1193–205.
[108] Hanson J, Yang YD, Paliwal K, Zhou YQ. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics 2017;33:685–92.
[109] Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–4.
[110] Chowdhury S, Zhang J, Kurgan L. In silico prediction and validation of novel RNA binding proteins and residues in the human proteome. Proteomics 2018;18:e1800064.
[111] Zhang T, Faraggi E, Xue B, Dunker AK, et al. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J Biomol Struct Dyn 2012;29:799–813.
[112] Gunasekaran K, Tsai CJ, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J Mol Biol 2004;341:1327–41.
[113] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851–69.
[114] Stahl K, Schneider M, Brock O. EPSILON-CP: using deep learning to combine information from multiple sources for protein contact prediction. BMC Bioinformatics 2017;18:303.
[115] Zhao Z, Gong X. Protein-protein interaction interface residue pair prediction based on deep learning architecture. IEEE/ACM Trans Comput Biol Bioinform 2017. https://doi.org/10.1109/TCBB.2017.2706682.
[116] Wu Z, Hu G, Yang J, Peng 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;589:2561–9.
[117] Meng F, Kurgan L. DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics 2016;32:i341–50.
[118] Meng F, Kurgan L. High-throughput prediction of disordered moonlighting regions in protein sequences. Proteins 2018;86:1097–110.
[119] Wu Z, Hu G, Wang K, Kurgan L. 6th international conference on artificial intelligence and soft computing. Poland: Zakopane; 2017; 722–32.
[120] Kihara D, Chen H, Yang YD. Quality assessment of protein structure models. Curr Protein Pept Sci 2009;10:216–28.
[121] Cao R, Bhattacharya D, Adhikari B, Li J, Cheng J. Large-scale model quality assessment for improving protein tertiary structure prediction. Bioinformatics 2015;31:i116–23.
[122] Skwark MJ, Elofsson A. PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. Bioinformatics 2013;29:1817–8.
[123] Hu G, Wu Z, Oldfield C, Wang C, Kurgan L. Quality assessment for the putative intrinsic disorder in proteins. Bioinformatics 2018. https://doi.org/10.1093/bioinformatics/bty881.