Computational prediction of functions of intrinsically disordered regions

Akila Katuwawala[1], Sina Ghadermarzi[1] and Lukasz Kurgan[1*]

[1]Department of Computer Science, Virginia Commonwealth University

*Corresponding author
Address: Department of Computer Science, Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, Virginia 23284, USA
Email: lkurgan@vcu.edu; Phone: (804) 827-3986

Abstract

Intrinsically disorder regions (IDRs) and abundant in nature, particularly among Eukaryotes. While they facilitate a wide spectrum of cellular functions including signaling, molecular assembly and recognition, translation, transcription and regulation, only several hundred IDRs are annotated functionally. This annotation gap motivates the development of fast and accurate computational methods that predict IDR functions directly from protein sequences. We introduce and describe a comprehensive collection of 25 methods that provide accurate predictions of IDRs that interact with proteins and nucleic acids, that function as flexible linkers and that moonlight multiple functions. Virtually all of these predictors can be accessed online and many were developed in the last few years. They utilize a wide range of predictive architectures and take advantage of modern machine learning algorithms. Our empirical analysis shows that predictors that are available as webservers enjoy high rates of citations, attesting to their practical value and popularity. The most cited methods include DISOPRED3, ANCHOR, alpha-MoRFpred, MoRFpred, fMoRFpred and MoRFCHiBi. We present two case studies to demonstrate that predictions produced by these computational tools are relatively easy to interpret and that they deliver valuable functional clues. However, the current computational tools cover a relatively narrow range of disorder functions. Further development efforts that would cover a broader range of functions should be pursued. We demonstrate that a sufficient amount of functionally annotated IDRs that are associated with several other disorder functions is already available and can be used to design and validate novel predictors.

Keywords

Intrinsic disorder; intrinsically disordered regions; intrinsic disorder functions; DNA binding; RNA binding; flexible linker; moonlighting; protein binding; prediction.

1 Introduction

The notion that all proteins have unique and stable 3D structure has been challenged by the ubiquitous presence of intrinsically disordered regions (IDRs). IDRs lack well-defined 3D structure under physiological conditions and form dynamic ensembles of conformers without specific equilibria for their coordinates (1-4). Several large-scale computational studies suggest that IDRs are highly abundant in nature, particularly among the eukaryotic organisms and viral proteomes (5-15). The significance of IDRs stems from the diversity of the biological and molecular functions that they perform. These functions include translation, transcription, molecular assembly, signaling, regulation, programmed cell death, chromatin remodeling and compacting, and molecular recognition, to name just a few (16-32). Intrinsic disorder is also shown to be enriched in the alternative splicing and post-translational modification sites (33-37). These three phenomena are thought to drive the regulatory complexity that underpins eukaryotic organisms (34, 37). However, so far only several hundred IDRs were annotated functionally (38).

Computational prediction can be used to assist with closing the functional annotation gap for the millions of the unannotated protein sequences (39, 40). The underlying principle is to use the limited collection of functionally annotated IDRs to design and optimize predictive models that can be used to make accurate predictions for the currently unannotated protein sequences. Many computational tools that target prediction of various functional aspects of IDRs were developed and published during the last decade (41). They perform prediction in a high-throughput manner, i.e., a single protein sequence can be predicted in few seconds to a handful of minutes on a single CPU, depending on the method used. These methods were designed using a variety of machine learning algorithms, biophysical models, and empirically-derived scoring functions (41). This chapter sheds light on the myriad of these tools. It primarily focuses on several practical aspects that are commonly encountered by the end users, such as availability, popularity, methodology, and interpretation of results.

Section 2 introduces a commonly used categorization of the functions of IDRs. It also provides a detailed accounting of the currently available functional annotations for each of these categories. Section 3 motivates and explains computational prediction of the functions of IDRs. It categorizes the existing predictors based on their target functions, comments on their popularity and availability, and details the underlying predictive architectures. Section 4 demonstrates the working of these methods with two case studies that cover several types of disorder functions. These case studies also aim to familiarize the reader with the format of outputs generated by these predictors, and to illustrate their agreement with the native functional annotations. This chapter concludes by summarizing key observations concerning the current predictors and suggesting avenues for the future research and development.

2 Functional Annotations of Intrinsically Disordered Regions

The various functions of IDRs can be broadly categorized by the underlying molecular-level functions and by the molecular partners (42-45). This convention is implemented in the DisProt database, which is the global repository of the functionally annotated IDRs (46-48).

The molecular functions of the intrinsic disorder are categorized into six broad classes: entropic chains, display sites, chaperons, effectors, assemblers and scavengers (42). Entropic chains are the sequences that remains persistently unstructured to fulfill functions that require substantial levels of flexibility. A representative example are the IDRs present in the Titin protein (49). Display sites facilitate post-translational modifications (PTMs). PTMs are often placed inside IDRs (50, 51), and this placement facilitates interactions with catalytic site modifying enzymes and access to effector proteins that mediate downstream outcomes upon binding (52). Some proteins with IDRs act as chaperons to support folding of RNAs and proteins into their functional conformations (53). As many as half of known RNA chaperons and one third of protein chaperons are believed to include IDRs (54). The fourth molecular function category are the effectors which alter functions of other proteins after binding. These IDRs typically transform from the disordered to structured state upon binding, a process referred to as coupled folding and binding (55, 56). A couple of examples are the p21 and p27 proteins that associate with many cyclin dependent kinases for cell cycle regulation (52), and p53 that is known to interact with dozens of diverse protein partners (57). The next functional category, assemblers, are proteins that bring several proteins together to make a larger complex. The assembler IDRs work as either scaffolds or structural mortars that stabilize protein complexes. An example of the stabilizer function is the ribosomal complex (20, 58). An example for the scaffold function is Axin which co-localizes β-catenin, casein kinase Iα, and glycogen synthetase kinase 3β (59). The sixth and final molecular function category is the scavenger that ingests and neutralizes small ligands. Chromogranin A is a well-known example of a scavenger that targets ATP and adrenaline (60).

The other way to categorize functions of IDRs is by the type of their binding partners. There are seven generic types of partners: proteins, DNAs, RNAs, lipids, metals, inorganic salt and small molecules. This classification supplements information associated with some of the molecular functions categories, such as effectors, chaperons, assemblers and scavengers.

The current version 7.05 of the DisProt database provides access of 1996 experimentally annotated IDRs. This count excludes lower quality annotations that are supported by ambiguous experimental evidence. Figure 1 breaks down these IDRs into four groups: 1111 IDRs with no functional annotations, 202 with only the molecular function annotations, 216 with only the molecular partner annotations, and 467 that have both molecular function and partner annotations. Some of the IDRs in the latter three sets are associated with multiple functional annotations. This is concomitant with the observation that IDRs are known to be moonlighting

(61, 62). We also emphasize the fact that about 56% of IDR in DisProt entirely lack the functional annotations.

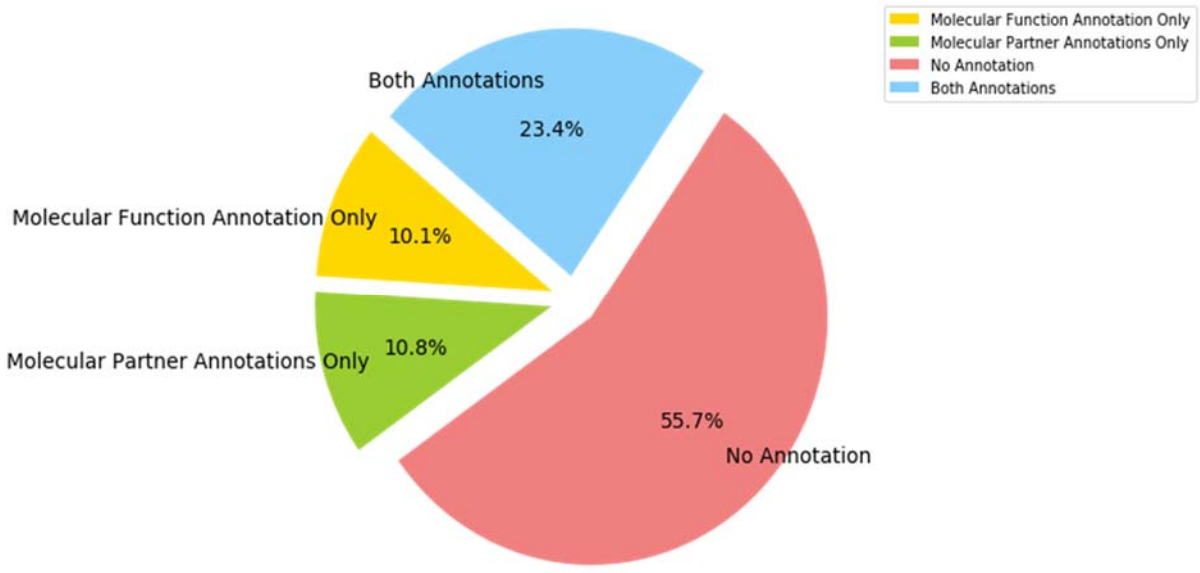

Figure 1. Breakdown of the different types of functional annotations for the IDRs in the DisProt database.

Table 1 provides further details for the molecular functions and their sub-categories that are annotated in DisProt. It lists all sub-categories together with the number of the corresponding IDRs and the number of proteins that have these IDRs. The most often annotated molecular function is the assembler, with 232 annotated IDRs in 146 proteins. The other two functions that are relatively common in DisProt are entropic chains (186 IDRs with 136 annotated as disordered linkers) and effectors (181 IDRs). To compare, there are relatively few IDRs annotated as chaperones (30 IDRs), display sites (23 IDRs) and scavengers (17 IDRs). Interestingly, some of the sub-categories of molecular functions were not yet annotated. These include entropic clock, solvate layer, entropy transfer, and methylation site. Table 2 provides the same analysis for the annotations of molecular partners. The most commonly annotated binding partners of IDRs are proteins, followed by DNAs, metals, RNAs, small molecules, lipids and inorganic salts. The latter ligand is associated with only one IDRs. The number of IDRs that are annotated to interact with proteins (417 IDRs) more than doubles the number of IDRs that are annotated for all other ligands combined (211 IDRs). Moreover, there are 55 IDRs located in 19 proteins that are annotated with multiple partners.

Table 1. Number of molecular function annotations for the functionally annotated IDRs in the DisProt database. Annotations tagged as ambiguous have been excluded. Functions are sorted the number of annotated IDRs.

Functional annotations		Number of annotated IDRs	Number of annotated proteins
Molecular functions	**Molecular function subcategories**		
Molecular Recognition: Assembler	Assembler	76	46
	Localization (targeting)	20	17
	Localization (tethering)	13	8
	Prion (self-assembly, polymerization)	10	7
	Liquid-liquid phase separation/demixing	3	3
	Total	**232**	**146**
Entropic Chain	Flexible linker/spacer	136	95
	Entropic bristle	10	5
	Entropic spring	2	2
	Self-transport through channel	2	2
	Structural mortar	1	1
	Entropic clock	0	0
	Total	**186**	**109**
Molecular Recognition: Effector	Inhibitor	44	29
	Activator	29	16
	cis-regulatory elements	6	5
	DNA bending	4	3
	Disassembler	3	1
	DNA unwinding	1	1
	Total	**181**	**109**
Molecular Recognition: Chaperone	Space filling	3	2
	Entropic exclusion	3	3
	Protein solvate layer	0	0
	Entropy transfer	0	0
	Total	**30**	**27**
Molecular Recognition: Display Site	Phosphorylation	14	12
	Glycosylation	4	4
	Fatty acylation	3	3
	Acetylation	2	2
	Ubiquitination	1	1
	Limited proteolysis	1	1
	Methylation	0	0
	Total	**23**	**21**
Molecular Recognition: Scavenger	Metal binding/metal sponge	10	8
	Neutralization of toxic molecules	4	3
	Water storage	1	1
	Total	**17**	**14**

Table 2. Number of molecular partner annotations for the functionally annotated IDRs in the DisProt database. IDRs annotated with multiple partners are counted for each listed partner. Annotations tagged as ambiguous have been excluded. Partners are sorted the number of annotated IDRs.

Molecular Partner	Number of annotated IDRs	Number of annotated proteins
Protein	417	267
DNA	76	48
Metal	39	26
RNA	33	24
Small molecule	32	27
Lipid	30	16
Inorganic Salt	1	1

Our analysis reveals only about half of the IDRs in the DisProt database have assigned functions or partners. The IDRs found in other resources, such as MobiDB (63) and PDB (64, 65), are entirely devoid of the functional annotations. Functional annotation of these regions as well as need to process annotations for the millions of uncharacterized protein sequences call for innovative and scalable solutions, one of which is the development of accurate predictive tools. The IDRs that are already functionally annotated can be utilized to design, optimize and test predictive models which than could be used to predict functional IDRs in the other protein sequences. While at this point many of the IDR functions lack sufficient numbers of annotated IDRs to merit these development efforts, several molecular functions and partners may have an adequate amount of data to perform optimization and testing. Our analysis suggests that this could be the case for the assembler, entropic chain and effector functions, as well as for IDRs that interact with protein and DNA partners. Correspondingly, majority of current development effort have concentrated on the prediction of protein-binding IDRs, with only a few methods that target other functions and partners (41). The next section overviews a comprehensive collection of computational predictors of IDR functions.

3 Prediction of Functions of Intrinsically Disordered Regions

We identified 25 computational predictors of functions of IDRs based on a comprehensive literature search. To the best of our knowledge this is a complete set of published method in this area at this point in time. Nearly all of these computational predictors were developed via a data-driven machine learning approach. These predictive models are generated by a machine learning algorithm using a functionally annotated training dataset. The models are optimized by the algorithm to minimize predictive error on the training dataset. After this optimization is completed, the models are assessed on a set aside set of test proteins where the predictions are compared against known native annotations. The test proteins are typically required to share low sequence identity (<30%) with the proteins in the training dataset. This is to ensure that these predictors are capable of producing accurate results in the absence of sequence similarity to the functionally annotated proteins. Each predictor discussed in this chapter has

undergone this assessment and was shown to offer relatively accurate predictions, even for the low sequence similarity proteins. This section lists the 25 computational tools and discusses their availability, impact and predictive architectures.

3.1 Overview and Impact

The 25 predictors can be divided into two main categories: methods that target prediction of molecular partners and methods that predict molecular functions. The current predictors in the former category address three types of partners: proteins, DNA and RNA. The available predictors of molecular functions are limited to the prediction of flexible linkers (a sub-category of the entropic chains) and moonlighting/multifunctional regions. A detailed list and classification of the 25 methods is presented Table 3. This table shows a substantial increase in the development of these predictors in recent years. Specifically, 13 methods were published in the last 3 year (2016-present), compared to 12 that were published in the preceding 10 years (between 2007 and 2016).

A significant majority of the current methods (21 out of 25) predict disordered protein-binding regions. These regions are by far the most annotated category of functional IDRs in DisProt (see Tables 1 and 2). The 21 methods can be further subdivided based on the particular type of the protein-binding IDRs they predict. The largest group of 16 methods focuses on predicting molecular recognition features (MoRFs). MoRFs are short protein-binding segments (typically between 5 and 25 consecutive residues) that undergo disorder-to-order transitions upon binding to their protein partners and which are localized inside longer IDRs (56, 66). Several notable examples of the MoRF predictors include the first method, alpha-MoRFpred (67, 68), which predicts MoRFs that fold into alpha-helical conformation, the first predictor that targets all MoRF types irrespective of their folded conformation, MoRFpred (69, 70), and several other popular tools including MoRFChiBi (71), fMoRFpred (56), and DISOPRED3 (72). The second type of the disordered protein-binding IDRs are short linear sequence motifs (SLiMs). SLiMs are short conserved motifs (3 to 12 consecutive amino acids) that are involved in protein-protein interactions (73). A list of currently known SLiMs can be collected from the Eukaryotic Linear Motif (ELM) resource (74). While majority of SLiMs are located in IDRs, about 20% of them are found within the structured protein domains (75). SLiMs can be predicted with the help of two sequence-based methods: SLiMpred (76) and PSSMpred (77). Another method, PepBindPred (78), predicts SLiMs in protein structures. Finally, three methods, ANCHOR (79, 80), disoRDPbind (81, 82) and ANCHOR2A (83), are designed to predict a generic set of disordered protein binding regions, which covers the short MoRFs and SLiMs and long protein-binding IDRs.

Only four methods target predictions of the other functions of IDRs. Two predictors that are part of the DisoRDPbind method predict IDRs that have DNA and RNA partners (81, 82). Two methods predict molecular functions of IDRs, including DFLpred that predicts flexible linkers (84), and DMRpred that predicts multifunctional/moonlighting IDRs (85). The latter type of IDRs

have multiple different functions (e.g., they bind two different partners) and is commonly found in DisProt, i.e., about 37% of IDRs in DisProt are moonlighting (41).

We also investigate impact of these methods that is quantified with their citations in the Google Scholar. Table 3 includes total and annual numbers of citations, where the latter measure is more suitable when comparing between methods. In total, the 25 predictors were cited 1651 times with the median citation count of 22. Based on the annual citation numbers, the most popular predictors are DISOPRED3 (52 citations per year), ANCHOR (39 citations per year), alpha-MoRFpred (37 citations per year), MoRFpred (28 citations per year), and fMoRFpred and MoRFCHiBi (each with 12 citations per year). We note that DISOPRED3's citations could be overestimated in the context of predicting functional IDRs since this method also predicts generic IDRs; i.e., regions without functional annotations.

Table 3 reveals that 19 of the 25 predictors are available to the research community via a website. Among these methods, 16 are available as webservers and 11 as a source code that must be installed and run on the end user's hardware. Furthermore, eight methods are available as both webserver and source code. We found that the mode of the availability is connected with the citation levels. The median annual number of citations for the methods that do not offer webservers is only 2, while it goes up to 11 for the methods that have webservers. The methods that are available as the source code are cited at the annual median rate of 7, and those that have both code and webserver at 9 citations per year. Overall, this analysis suggests that availability of the webserver substantially boosts the usage of the corresponding predictors.

Table 3. Classification, citations and availability of the current predictors of IDR functions. The methods are classified based on their predictive target (molecular partner vs. molecular function) and sub-type of the target (protein, DNA and RNA for molecular partners vs. flexible linker and moonlighting region for molecular functions). Predictors are sorted within each sub-type by the year of publication. The citations and availability are based on information as of Feb 25, 2019. The citations were collected using Google Scholar, where the annual citations are computed as an average number of citations per year since publication. The type of availability, shown in column "type", is either through a webserver (WS), downloadable source code (SC), or both (WS+SC). Methods without any availability are listed as "not available" and those for which the websites cannot be found are denoted as "no longer available".

Predictive target			Year	Method	Ref.	Citations		Availability	
						Total	Annual	Type	URL
Partners	Proteins	MoRFs	2007	alpha-MoRFpred	(67, 68)	445	37	NA	Not Available
			2010	retro-MoRFs	(86)	27	3	NA	Not Available
			2012	MoRFpred	(69, 70)	194	28	WS	http://biomine.cs.vcu.edu/servers/MoRFpred/
			2013	MFSPSSMpred	(87)	32	5	NLA	No Longer Available
			2015	fMoRFpred	(56)	36	12	WS	http://biomine.cs.vcu.edu/servers/fMoRFpred/
			2015	DISOPRED3	(72)	206	52	WS+SC	http://bioinf.cs.ucl.ac.uk/disopred
			2015	MoRFCHiBi	(71)	35	12	WS+SC	https://gsponerlab.msl.ubc.ca/software/morf_chibi/downloads/
			2016	MoRFCHiBiLight	(75)	22	7	WS+SC	https://gsponerlab.msl.ubc.ca/software/morf_chibi/downloads/
			2016	MoRFCHiBiWeb	(75)	22	7	WS+SC	http://morf.chibi.ubc.ca:8080/mcw/index.xhtml
			2016	Predict-MoRFs	(88)	6	2	SC	https://github.com/roneshsharma/Predict-MoRFs
			2017	Wang et al. 2017	(89)	2	2	NA	Not Available
			2018	MoRFpred-plus	(90)	8	7	SC	https://github.com/roneshsharma/MoRFpred-plus/wiki/MoRFpred-plus
			2018	OPAL	(91)	8	6	WS+SC	http://www.alok-ai-lab.com/tools/opal/
			2018	OPAL+	(92)	0	0	WS+SC	http://www.alok-ai-lab.com/tools/opal_plus/
			2018	Fang et al 2018	(93)	0	0	NA	Not Available
			2019	Sharma et al. 2019	(94)	0	0	SC	https://github.com/roneshsharma/BMC_Models2018/wiki.
		SLiMs	2012	SLiMPred	(76)	54	8	WS	http://bioware.ucd.ie/~compass/biowareweb//Server_pages/slimpred.php
			2016	PSSMpred	(77)	0	0	NLA	No Longer Available
		ALL	2009	ANCHOR	(79, 80)	388	39	WS+SC	http://anchor.enzim.hu
			2015	disoRDPbind	(81, 82)	44	11	WS	http://biomine.cs.vcu.edu/servers/DisoRDPbind/
			2018	ANCHOR2	(83)	17	10	WS+SC	https://iupred2a.elte.hu/
	DNAs		2015	disoRDPbind	(81, 82)	44	11	WS	http://biomine.cs.vcu.edu/servers/DisoRDPbind/
	RNAs		2015	disoRDPbind	(81, 82)	44	11	WS	http://biomine.cs.vcu.edu/servers/DisoRDPbind/
Functions	Flexible linkers		2016	DFLpred	(84)	17	8	WS	http://biomine.cs.vcu.edu/servers/DFLpred/
	Moonlighting regions		2018	DMRpred	(85)	0	0	WS	http://biomine.cs.vcu.edu/servers/DMRpred/

Table 4. Architectures of the current predictors of IDR functions. The methods are classified based on their predictive target (molecular partner vs. molecular function) and sub-type of the target (protein, DNA and RNA for molecular partners vs. flexible linker and moonlighting region for molecular functions). Predictors are sorted within each sub-type by the year of publication. The "Class" column shows the overall class of the predictor: machine learning (ML), *ab-initio* (AI) and meta-predictors (Meta). The "Predictive Model" column specifies types of predictive models: neural network (NN), scoring function (SF), support vector machine (SVM), Bayesian model (Bayes), logistic regression (LR), and random forest (RF). The specific elements of the input profiles are encoded as "AA" (features computed directly from the amino acid sequence), "EVO" (evolutionary features including a position-specific scoring matrix and a hidden markov model profile), "PSS" (putative secondary structure), "PSA" (putative solvent accessibility), "PDIS" (putative disordered regions), "PMoRF" (putative MoRF regions), and "SQA" (sequence alignment).

Predictive target			Year	Method	Class	Predictive model	Contents of the input profile							
							AA	EVO	PSS	PSA	PDIS	PMoRF	SQA	Other information
Partners	Proteins	MoRF	2007	alpha-MoRFpred	ML	NN	✓				✓			
			2010	retro-MoRFs	AI	SF	✓				✓		✓	
			2012	MoRFpred	ML	SVM	✓	✓						
			2013	MFSPSSMpred	ML	SVM			✓		✓			
			2015	fMoRFpred	ML	SVM		✓						
			2015	DISOPRED3	ML	SVM	✓	✓		✓	✓			
			2015	MoRFCHiBi	ML	SVM	✓		✓	✓	✓			
			2016	MoRFCHiBiLight	Meta	Bayes	✓				✓	✓		
			2016	MoRFCHiBiWeb	Meta	Bayes	✓				✓	✓		
			2016	Predict-MoRFs	ML	SVM		✓					✓	
			2017	Wang et al. 2017	ML	SVM	✓							
			2018	MoRFpred-plus	Meta	SVM	✓	✓				✓	✓	Predicted B-factors
			2018	OPAL	Meta	SVM	✓	✓		✓		✓		
			2018	OPAL+	Meta	SVM	✓	✓				✓		
			2018	Fang et al 2018	ML	SVM		✓						
			2019	Sharma et al. 2019	Meta	SVM	✓	✓	✓	✓		✓		
		SLiMs	2012	SLiMPred	ML	NN	✓							
			2016	PSSMpred	ML	SVM		✓						
		All	2009	ANCHOR	AI	SF	✓	✓			✓			
			2015	disoRDPbind	ML	LR	✓						✓	Sequence complexity
			2018	ANCHOR2	AI	SF	✓	✓			✓			
	DNAs		2015	disoRDPbind	ML	LR	✓						✓	Sequence complexity
	RNAs		2015	disoRDPbind	ML	LR	✓						✓	Sequence complexity
Functions	Flexible linkers		2016	DFLpred	ML	LR	✓		✓		✓			
	Moonlighting regions		2018	DMRpred	ML	RF	✓	✓		✓	✓			

3.2 Predictive Architectures

The predictive architecture of the 25 methods can be divided into three classes: methods that rely on machine learning (ML) models, *ab-initio* (AI) models, and meta-predictors (Meta). The ML models are derived with the help of machine learning algorithms. These algorithms parametrize the models to maximize predictive quality on a functionally annotated training datasets. Several different ML algorithms have been applied, such as neural networks, support vector machines, Bayesian algorithms, logistic regressions, and random forests. The *ab-initio* models are developed utilizing biophysical principles that are known to differentiate between the functional regions and other parts of the protein sequences. They are typically implemented as scoring functions. Several recently published predictors rely on meta-architectures that combine outputs generated by several predictors of IDR functions, typically using a predictive model derived with ML algorithms. The underlying motivation for this class of methods is that the meta-models are expected to improved predictive performance when compared to the use of individual predictors (95-98).

The prediction generally consists of two steps, irrespective of the architectural class. First, the input protein sequence is converted into a profile that includes the sequence itself and a set of selected sequence-derived characteristics. These characteristics may include evolutionary information (such as a measures of conservation), sequence alignment, as well as putative IDRs, putative secondary structure, and/or putative solvent accessibility. In case of the meta-architectures, the profile consists of outputs generated by several predictors of IDR functions. Second, the profile is input into a predictive model (ether a ML model or a scoring function) which produces numeric scores that quantify propensity for the specific function for each residue from the input protein sequence.

Table 4 shows that majority of the predictors fall into the ML class (16 out of 25). There are also three ab-initio predictors and six meta-predictors. The table also reveals that different methods utilize different information in the profile and different types of predictive models. All but three predictors rely on the ML models, with the most popular being the support vector machine (13 out of 22 ML predictors). However, support vector machine models are used to predict only MoRFs and SLiMs. The meta architecture-based methods are exclusively used to predict MoRFs. This is motivated by the fact that MoRF prediction is the most mature and most populated sub-area, which results in availability of several strong predictors that can be used as inputs for these meta-predictors.

The scope of the profiles varies widely between different predictors, as shown in Table 4. We break down the profiles into seven major components that are used across the 25 predictors: information computed directly from the amino acid sequence, evolutionary information, putative secondary structure, putative solvent accessibility, putative MoRF regions, putative disordered regions, and sequence alignment. Virtually all predictors use information obtained directly from the input protein sequence, which typically includes amino acid composition and

physicochemical properties of the amino acids, such as hydrophobicity and polarity. The second most popular element of the profile is the evolutionary information, which is encoded in the form a position-specific scoring matrix or a hidden Markov model profile. The least popular components of the profile include putative secondary structure and solvent accessibility. Furthermore, we observe that the contents of the profiles are unique to each predictor, ranging from simple architectures that use a single element to complex profiles that include as many as five elements.

4 Case Studies

We explain and illustrate the disorder function predictions using two proteins that feature different types of functional IDRs. Both case studies show results generated by the same set of five predictors: ANCHOR2, DISOPRED3 and the three predictive models included in DisoRDPbind, the only method capable of predicting DNA- and RNA-binding IDRs.

In general, these and other predictors output a numerical propensity score, which quantifies likelihood for a given function for each residue in the input protein sequence. These propensity scores are converted into a binary prediction (functional *vs*. non-functional residue) with a help of predictor-specific thresholds. More specifically, residues with propensities exceeding the threshold are assumed to be functional, while the remaining residues are annotated as non-functional.

The first case study is the RNA polymerase subunit 13 from *Saccharolobus shibatae* (UniProt id: B8YB65), which is a regulator of transcription. This protein has two IDRs (positions Met-1 to Glu-32 and Lys-83 to Gly-104) that were annotated using circular dichroism and NMR spectroscopy (DisProt id: DP01001) (99). Figure 2 shows predictions of disorder for this protein using several popular disorder predictors (41, 100, 101) that include VSL2B (102), IUPred2A (103), disCoP (95, 96) and DISOPRED3 (104). Both IDRs are found by each of the four predictors, however, these methods are unable to annotate these regions functionally. The two IDRs were shown to interact with DNA and proteins, including the transcription initiation factors TFIIF and TFIIE (105). Figure 3 shows that the three predictors of protein partners (ANCHOR2, DISOPRED3 and DisoRDPbind) correctly identify the protein-binding IDR at the N-terminus. The ANCHOR2's predictions (in pink) is fragmented into two regions, one of which extends beyond the native annotation. However, the propensities generated by ANCHOR2 in this part of the sequence are in general high (above or marginally below the threshold) suggesting a high likelihood for protein binding. The protein-binding IDR at the C-terminus is detected only by DISOPRED3 (in orange). DisoRDBbind successfully find the DNA-binding region at the C-terminus (in teal), while it fails to identify that the IDRs at the N-terminus also interacts with DNA. Overall, this example reveals that both IDRs can be found and functionally annotated by the current methods, with the exception of the DNA interaction at the N-terminus.

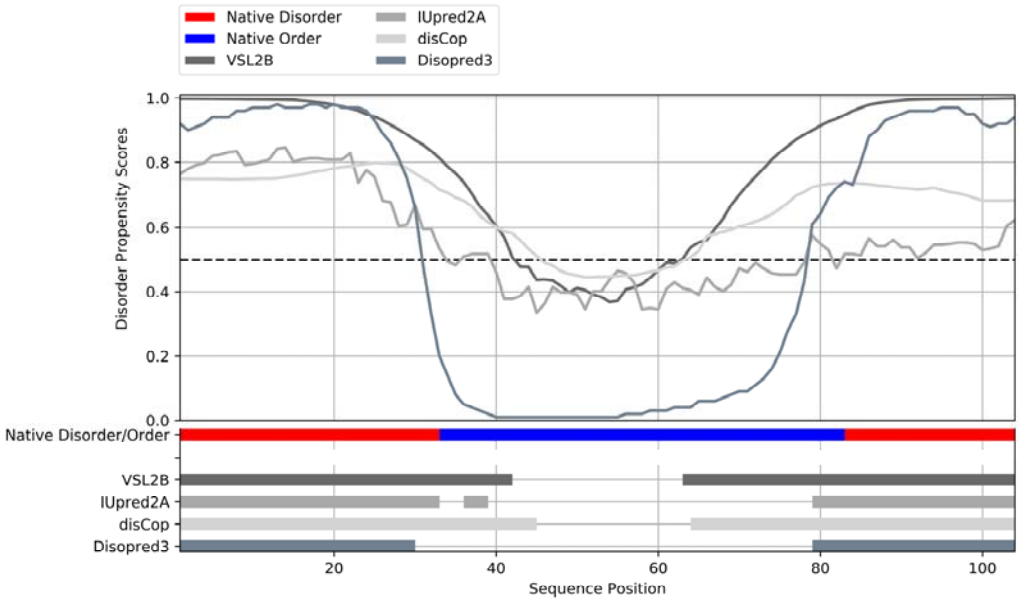

Figure 2. Predictions and native annotations of the disorder for the RNA polymerase subunit 13 (UniProt id: B8YB65). Native annotations, which are shown in red (disordered regions) and blue (structured regions), were collected from the DisProt database (DisProt id: DP01001). Predictions of IDRs (shown using shades of gray) were generated with VSL2B, IUPred2A, disCoP and DISOPRED3. The solid line curves denote the numerical propensities, dashed horizontal line gives the threshold used to convert propensities into the binary predictions for all four predictors, and the horizontal bands at the bottom correspond to the binary predictions and native annotations.

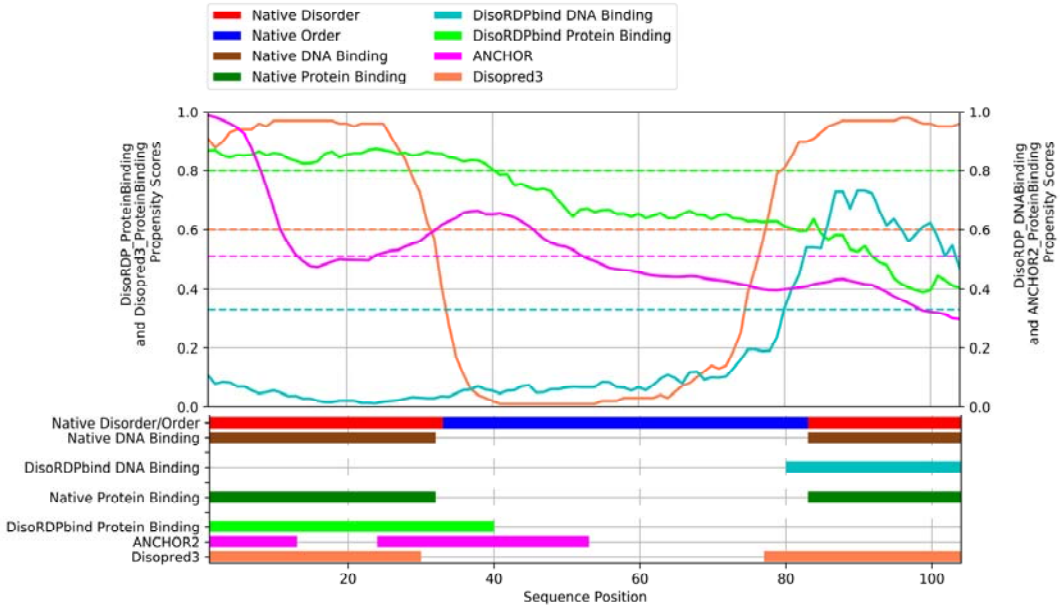

Figure 3. Predictions and native annotations of the disorder functions for the RNA polymerase subunit 13 (UniProt id: B8YB65). Native annotations, which are shown in red (disordered regions), blue (structured regions), brown (IDRs interacting with DNA), and dark green (IDRs interacting with proteins), were collected from the DisProt database (DisProt id: DP01001). Predictions of protein-binding IDRs were generated with DisoRDPbind (light green), ANCHOR2 (pink) and DISOPRED3 (orange). Predictions of DNA-binding IDRs were generated with DisoRDPbind (teal). The solid line curves denote the numerical propensities, dashed horizontal lines correspond to the thresholds and the horizontal bands at the bottom give the binary predictions and native annotations.

Prediction of functions of intrinsic disorder

The second case study is the 50S ribosomal protein L11 from *Geobacillus stearothermophilus* (UniProt id: P56210). This protein forms ribosomal stalk that facilitates interaction with the GTP-bound translation factor. In contrast to the first case study, where IDRs are located at the sequence termini, this protein has two short IDRs in the middle of the chain (positions Thr-59 to Lys-63, and positions Glu-76 to Thr-91). Both disordered regions were characterized by Nuclear Magnetic Resonance and the second region is known to interact with RNA (DisProt id: DP00512) (106, 107). Results of disorder predictions with VSL2B, IUPred2A, disCoP and DISOPRED3, which are shown in Figure 4, reveal that the second IDRs is predicted by all four methods. VSL2B over-predicts disorder in this protein and three predictors (VSL2B, IUPred2A and disCoP) incorrectly identify disorder at the N-terminus. Sequence of this protein was used to make predictions of IDR functions with DisoRDPbind, ANCHOR2, and DISOPRED3 and the results are visualized in Figure 5. DisoRDPbind correctly identifies the RNA-binding IDRs (in teal). The DisoRDPbind's propensity scores in this region are very high suggesting high confidence for this prediction. Results generated by the three predictors of protein-binding IDRs reveals that they do not predict these regions; i.e., propensities generated by these methods are relatively low and well below the corresponding thresholds that are shows with dashed horizontal lines. The only exception are the propensity values generated by DISOPRED3 in the vicinity of positions 77I to 79S which are relatively high, although they still remain below the cut-off value. In the nutshell, this example demonstrates that the second IDR that interacts with RNA is correctly identified by DisoRDPbind. This method also makes a correct determination that the first IDRs does not interact with RNA. Moreover, the three predictors of IDRs that partner with proteins correctly identify that neither of the two IDRs is binding proteins.

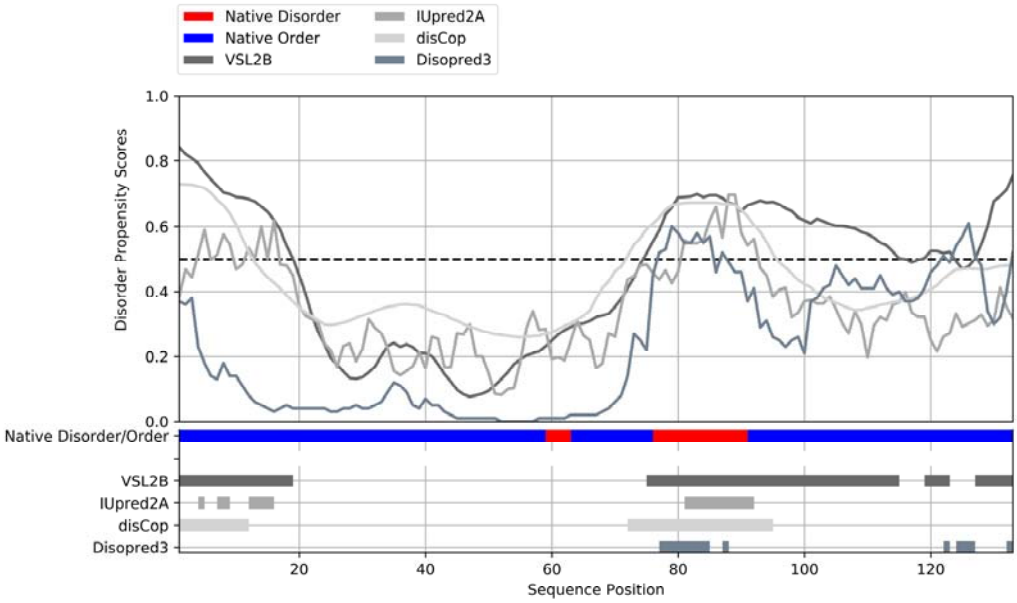

Figure 4. Predictions and native annotations of the disorder for the 50S ribosomal protein L11 (UniProt id: P56210). Native annotations, which are shown in red (disordered regions) and blue (structured regions), were collected from the DisProt database (DisProt id: DP00512). Predictions of IDRs (shown using shades of gray) were generated with VSL2B, IUPred2A, disCoP and DISOPRED3. The solid line curves denote the numerical propensities, dashed horizontal line gives the threshold used to convert propensities into the binary predictions for all four predictors, and the horizontal bands at the bottom correspond to the binary predictions and native annotations.

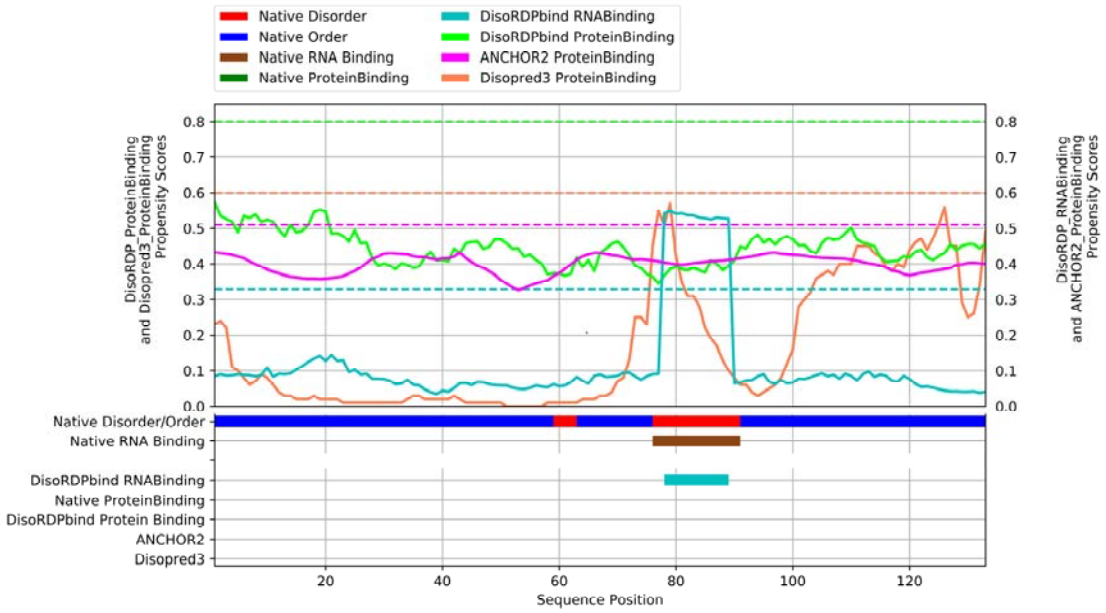

Figure 5. **Predictions and native annotations of the disorder functions for the 50S ribosomal protein L11 (UniProt id: P56210).** Native annotations, which are shown in red (disordered regions), blue (structured regions), and brown (IDRs interacting with RNA), were collected from the DisProt database (DisProt id: DP00512). Predictions of protein-binding IDRs were generated with DisoRDPbind (light green), ANCHOR2 (pink) and DISOPRED3 (orange). Predictions of RNA-binding IDRs were generated with DisoRDPbind (teal). The solid line curves denote the numerical propensities, dashed horizontal lines correspond to the thresholds and the horizontal bands at the bottom give the binary predictions and native annotations.

5 Summary and Prospective Advances

The growing body of experimental and computational studies demonstrates the diversity and abundance of functions that are carried out by IDRs (5, 17, 19, 21, 27, 43, 51, 73, 108-113). This chapter focuses on the computational predictions of these functions from protein sequences. We cover 25 methods that address prediction of IDRs that interact with proteins, DNA and RNA, which form flexible linkers and that moonlight molecular functions. We show that most of these methods are available online, are relatively well-cited and many were developed over the last few years. Our empirical analysis reveals that methods that are available as webservers attract substantially more citations, suggesting that they are utilized as higher rate than the methods that do not offer this option. We also argue that the predictions generated by these tools are relatively easy to interpret and that they provide useful functional clues, as we demonstrated in the two case studies.

While these methods utilize a wide range of predictive architectures, a significant majority of them rely on machine learning-derived models. The most often used machine learning algorithms are support vector machines and regression. Given the recent rapid advances in the deep learning technologies (114) and their adoption in the bioinformatics area (115, 116), we believe that prediction of functions of IDRs would also benefit from the introduction of these models. The deep neural networks are already used to predict generic IDRs showing that these models produce accurate predictions (117-120). We anticipate that a similar boost in predictive performance could be gained by adopting these models for the prediction of functional IDRs. One potential obstacle is the fact that the training datasets in this area are relatively small. The limited amount of functionally annotated IDRs may hamper learning of accurate deep network models that typically require large training datasets.

In spite of the fact that many new predictors were published in the past few years, they focus on a very narrow range of functions. We found that 84% of current methods (21 out of 25) focus on the predictions of IDRs that partner with proteins. While this can be explained by the availability of the larger number of the corresponding functionally annotated IDRs (Table 2), other types of molecular functions and partners of IDRs deserve an equal amount of attention. This is arguably a particularly acute problem in the case of IDRs that bind DNA and RNA, given the fact that IDRs are known to be heavily involved in the protein-nucleic acids interactions (16, 18, 20, 27, 31, 112, 121-124). While dozens of computational tools are available to predict RNA- and DNA-binding regions in structured proteins and protein regions (125-138), there is only one such method for the disordered regions, DisoRDPbind (81, 82). Similar problem is apparent for the prediction of molecular functions associated with IDRs. There are currently only two such tools that were developed very recently, DFLpred (84) and DMRpred (85). There seems to be sufficient amount of functionally annotated IDRs to develop new methods that would target prediction of assemblers and effectors (Table 1), besides the currently covered entropic chains. Therefore, we encourage the bioinformatics community to focus their efforts on the development of a broad ranges of computational predictors of disorder functions.

Finally, prediction of disorder functions could be a time-consuming and difficult, especially when working with large datasets of proteins (i.e., proteins families or whole proteomes) and when wanting to predict multiple functions. Most of the current predictors can make predictions for one protein at the time. Moreover, prediction of multiple functions would require using several different webservers. The first problem is already solved for the prediction of generic disorder (41, 100), where the end users can take advantage of two large-scale databases of disorder predictions: D^2P^2 (139) and MobiDB (63, 140). The putative disorder functions should be included in these databases in the near future. In fact, MobiDB already includes predictions of protein-binding IDRs generated with the ANCHOR method. The second issue could be accommodated by the development of a large predictive resource that would provide access to a comprehensive set of predictors. Several of these resources are already available for the prediction of structural aspects of proteins, including SCRATCH (141), PredictProtein (142) and MULTICOM (143). A similar solution should be released for the prediction of disorder and disorder functions.

Acknowledgement

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

References

1. Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev. 2014;114(13):6561-88.
2. Lieutaud P, Ferron F, Uversky AV, Kurgan L, Uversky VN, Longhi S. 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):e1259708.
3. A. Keith Dunker MMB, Elisar Barbar, Martin Blackledge, Sarah E. Bondos, Zsuzsanna Dosztányi, H. Jane Dyson, Julie Forman-Kay, Monika Fuxreiter, Jörg Gsponer, Kyou-Hoon Han, David T. Jones, Sonia Longhi, Steven J. Metallo, Ken Nishikawa, Ruth Nussinov, Zoran Obradovic, Rohit V. Pappu, Burkhard Rost, Philipp Selenko, Vinod Subramaniam, Joel L. Sussman, Peter Tompa & Vladimir N Uversky. What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins. 2013;1(1):e24157
4. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of Intrinsically Disordered Regions and Proteins. Chemical Reviews. 2014;114(13):6589-631.
5. Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci. 2015;72(1):137-51.
6. 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(2):137-49.
7. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol. 2004;337(3):635-45.

8. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161-71.
9. Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions. Proteins. 2014;82(1):145-58.
10. Fan X, Xue B, Dolan PT, LaCount DJ, Kurgan L, Uversky VN. The intrinsic disorder status of the human hepatitis C virus proteome. Mol Biosyst. 2014;10(6):1345-63.
11. Charon J, Theil S, Nicaise V, Michon T. Protein intrinsic disorder within the Potyvirus genus: from proteome-wide analysis to functional annotation. Molecular Biosystems. 2016;12(2):634-52.
12. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci. 2012;69(8):1211-59.
13. Meng F, Badierah RA, Almehdar HA, Redwan EM, Kurgan L, Uversky VN. Unstructural biology of the Dengue virus proteins. FEBS J. 2015;282(17):3368-94.
14. Yan J, Mizianty MJ, Filipow PL, Uversky VN, Kurgan L. RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. Biochim Biophys Acta. 2013;1834(8):1671-80.
15. 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:e1800243.
16. 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(10):1486-98.
17. 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(12).
18. Na I, Meng F, Kurgan L, Uversky VN. Autophagy-related intrinsically disordered proteins in intra-nuclear compartments. Mol Biosyst. 2016;12(9):2798-817.
19. Xue B, Blocquel D, Habchi J, Uversky AV, Kurgan L, Uversky VN, et al. Structural disorder in viral proteins. Chem Rev. 2014;114(13):6880-911.
20. Peng Z, Oldfield CJ, Xue B, Mizianty MJ, Dunker AK, Kurgan L, et al. A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell Mol Life Sci. 2014;71(8):1477-504.
21. Fuxreiter M, Toth-Petroczy A, Kraut DA, Matouschek A, Lim RY, Xue B, et al. Disordered proteinaceous machines. Chem Rev. 2014;114(13):6806-43.
22. 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(9):1257-67.
23. Peng Z, Mizianty MJ, Xue B, Kurgan L, Uversky VN. More than just tails: intrinsic disorder in histone proteins. Mol Biosyst. 2012;8(7):1886-901.
24. Dyson HJ. Roles of intrinsic disorder in protein-nucleic acid interactions. Mol Biosyst. 2012;8(1):97-104.
25. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 2008;18(6):756-64.
26. Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN. Close encounters of the third kind: disordered domains and the interactions of proteins. Bioessays. 2009;31(3):328-35.
27. Varadi M, Zsolyomi F, Guharoy M, Tompa P. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. PLoS One. 2015;10(10):e0139731.
28. Pancsa R, Tompa P. Coding Regions of Intrinsic Disorder Accommodate Parallel Functions. Trends Biochem Sci. 2016;41(11):898-906.
29. 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(5):343-84.

30. Tantos A, Kalmar L, Tompa P. The role of structural disorder in cell cycle regulation, related clinical proteomics, disease development and drug targeting. Expert Rev Proteomics. 2015;12(3):221-33.
31. Sandhu KS. Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins. J Mol Recognit. 2009;22(1):1-8.
32. 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.
33. Buljan M, Chalancon G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, et al. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. Current opinion in structural biology. 2013;23(3):443-50.
34. Zhou JH, Zhao SW, Dunker AK. Intrinsically Disordered Proteins Link Alternative Splicing and Post-translational Modifications to Complex Cell Signaling and Regulation. Journal of Molecular Biology. 2018;430(16):2342-59.
35. Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, et al. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Molecular cell. 2012;46(6):871-83.
36. Colak R, Kim T, Michaut M, Sun M, Irimia M, Bellay J, et al. Distinct types of disorder in the human proteome: functional implications for alternative splicing. PLoS Comput Biol. 2013;9(4):e1003030.
37. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, et al. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc Natl Acad Sci U S A. 2006;103(22):8390-5.
38. Piovesan D, Tabaro F, Micetic I, Necci M, Quaglia F, Oldfield CJ, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. 2016;D1:D219-D27.
39. Cozzetto D, Jones DT. The contribution of intrinsic disorder prediction to the elucidation of protein function. Curr Opin Struct Biol. 2013;23(3):467-72.
40. Uversky VN, Radivojac P, Iakoucheva LM, Obradovic Z, Dunker AK. Prediction of intrinsic disorder and its use in functional proteomics. Methods Mol Biol. 2007;408:69-92.
41. 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(17):3069-90.
42. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114(13):6589-631.
43. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry. 2002;41(21):6573-82.
44. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci. 2002;11(4):739-56.
45. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002;27(10):527-33.
46. Piovesan D, Quaglia F, Minervini G, Mičetić I, Necci M, Tabaro F, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Research. 2016;45(D1):D219-D27.
47. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, et al. DisProt: a database of protein disorder. Bioinformatics. 2005;21(1):137-40.
48. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the Database of Disordered Proteins. Nucleic Acids Res. 2007;35(Database issue):D786-93.
49. Tskhovrebova L, Trinick J. Titin: properties and family relationships. Nat Rev Mol Cell Biol. 2003;4(9):679-89.
50. Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. Molecular & cellular proteomics : MCP. 2008;7(7):1331-48.

51. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. J Proteome Res. 2007;6(5):1917-32.
52. Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of Cell Division by Intrinsically Unstructured Proteins: Intrinsic Flexibility, Modularity, and Signaling Conduits. Biochemistry. 2008;47(29):7598-609.
53. Schroeder R, Barta A, Semrad K. Strategies for RNA folding and assembly. Nat Rev Mol Cell Biol. 2004;5(11):908-19.
54. Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. FASEB J. 2004;18(11):1169-75.
55. Sugase K, Dyson HJ, Wright PE. Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature. 2007;447:1021.
56. Yan J, Dunker AK, Uversky VN, Kurgan L. Molecular recognition features (MoRFs) in three domains of life. Mol Biosyst. 2016;12(3):697-710.
57. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics. 2008;9 Suppl 1:S1.
58. Adilakshmi T, Ramaswamy P, Woodson SA. Protein-independent Folding Pathway of the 16S rRNA 5′ Domain. Journal of Molecular Biology. 2005;351(3):508-19.
59. Xue B, Romero PR, Noutsou M, Maurice MM, Rüdiger SGD, William AM, Jr., et al. Stochastic machines as a colocalization mechanism for scaffold protein function. FEBS letters. 2013;587(11):1587-91.
60. Daniels AJ, Williams RJP, Wright PE. The character of the stored molecules in chromaffin granules of the adrenal medulla: A nuclear magnetic resonance study. Neuroscience. 1978;3(6):573-85.
61. Tompa P. Moonlighting by disordered proteins. Biophysical Journal. 2007:1a-2a.
62. Tompa P, Szasz C, Buday L. Structural disorder throws new light on moonlighting. Trends in Biochemical Sciences. 2005;30(9):484-9.
63. Piovesan D, Tabaro F, Paladin L, Necci M, Micetic I, Camilloni C, et al. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res. 2018;46(D1):D471-D6.
64. Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. Methods Mol Biol. 2017;1607:627-41.
65. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.
66. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol. 2006;362(5):1043-59.
67. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and Combining Predictors of Mostly Disordered Proteins†. Biochemistry. 2005;44(6):1989-2000.
68. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. Mining α-Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments†. Biochemistry. 2007;46(47):13468-77.
69. Oldfield CJ, Uversky VN, Kurgan L. Predicting functions of disordered proteins with MoRFpred. Computational Methods in Protein Evolution: Springer; 2019. p. 337-52.
70. Disfani FM, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, 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):i75-i83.
71. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. Bioinformatics. 2015;31(11):1738-44.

72. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015;31(6):857-63.
73. Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev. 2014;114(13):6733-78.
74. Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, et al. ELM 2016--data update and new functionality of the eukaryotic linear motif resource. Nucleic Acids Res. 2016;44(D1):D294-300.
75. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res. 2016.
76. Mooney C, Pollastri G, Shields DC, Haslam NJ. Prediction of Short Linear Protein Binding Regions. Journal of Molecular Biology. 2012;415(1):193-204.
77. Fang C, Noguchi T, Yamana H, Sun F, editors. Identifying Protein Short Linear Motifs by Position-Specific Scoring Matrix. International Conference on Swarm Intelligence; 2016: Springer.
78. 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(9):e72838.
79. Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 2009;25(20):2745-6.
80. Mészáros B, Simon I, Dosztányi Z. Prediction of Protein Binding Regions in Disordered Proteins. PLoS Comput Biol. 2009;5(5):e1000376.
81. 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.
82. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res. 2015;43(18):e121.
83. Mészáros B, Erdös G, Dosztányi Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Research. 2018;46:W329-W37.
84. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. Bioinformatics. 2016;32(12):i341-i50.
85. Meng F, Kurgan L. High-throughput prediction of disordered moonlighting regions in protein sequences. Proteins: Structure, Function, and Bioinformatics. 2018;86(10):1097-110.
86. Xue B, Dunker AK, Uversky VN. Retro-MoRFs: Identifying Protein Binding Sites by Normal and Reverse Alignment and Intrinsic Disorder Prediction. International Journal of Molecular Sciences. 2010;11(10):3725-47.
87. 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 Bioinformatics. 2013;14:300.
88. Sharma R, Kumar S, Tsunoda T, Patil A, Sharma A. Predicting MoRFs in protein sequences using HMM profiles. BMC Bioinformatics. 2016;17.
89. Wang H, Feng L, Webb GI, Kurgan L, Song J, Lin D. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. Brief Bioinform. 2017:https://doi.org/10.1093/bib/bbx018.
90. 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. Journal of Theoretical Biology. 2018;437:9-16.
91. Sharma R, Raicar G, Tsunoda T, Patil A, Sharma A. OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics. 2018;34:1850-8.
92. Sharma R, Sharma A, Raicar G, Tsunoda T, Patil A. OPAL+: Length-Specific MoRF Prediction in Intrinsically Disordered Protein Sequences. Proteomics. 2018;1800058:1800058.

93. Fang C, Moriwaki Y, Zhu D, Shimizu K. Identifying MoRFs in Disordered Proteins Using Enlarged Conserved Features2018.
94. Sharma R, Sharma A, Patil A, Tsunoda T. Discovering MoRFs by trisecting intrinsically disordered protein sequence into terminals and middle regions. BMC bioinformatics. 2019;19(13):378.
95. Peng Z, Kurgan L. On the complementarity of the consensus-based disorder prediction. Pac Symp Biocomput. 2012:176-87.
96. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Struct Dyn. 2014;32(3):448-64.
97. Necci M, Piovesan D, Dosztanyi Z, Tosatto SCE. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics. 2017;33(9):1402-4.
98. Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics. 2012;13:111.
99. Wojtas MN, Mogni M, Millet O, Abrescia NGA, Bell SD. Structural and functional analyses of the interaction of archaeal RNA polymerase with DNA. Nucleic Acids Research. 2012;40(19):9941-52.
100. Meng F, Uversky V, Kurgan L. Computational Prediction of Intrinsic Disorder in Proteins. Curr Protoc Protein Sci. 2017;88:2 16 1-2 4.
101. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci. 2012;13(1):6-18.
102. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006;7(1):208.
103. 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(W1):W329-W37.
104. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015;31(6):857-63.
105. Geiger SR, Lorenzen K, Schreieck A, Hanecker P, Kostrewa D, Heck AJR, et al. RNA Polymerase I Contains a TFIIF-Related DNA-Binding Subcomplex. Molecular Cell. 2010;39(4):583-94.
106. Hinck AP, Markus MA, Huang S, Grzesiek S, Kustonovich I, Draper DE, et al. The RNA binding domain of ribosomal protein L11: three-dimensional structure of the RNA-bound form of the protein and its interaction with 23 S rRNA. J Mol Biol. 1997;274(1):101-13.
107. Markus MA, Hinck AP, Huang S, Draper DE, Torchia DA. High resolution solution structure of ribosomal protein L11-C76, a helical protein with a flexible loop that becomes structured upon binding to RNA. Nat Struct Biol. 1997;4(1):70-7.
108. Keith Dunker A, Obradovic Z. The protein trinity - Linking function and disorder2001. 805-6 p.
109. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, et al. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. J Proteome Res. 2007;6(5):1899-916.
110. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res. 2007;6(5):1882-98.
111. Kjaergaard M, Kragelund BB. Functions of intrinsic disorder in transmembrane proteins. Cellular and Molecular Life Sciences. 2017;74(17):3205-24.
112. 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(1).
113. Marin M, Ott T. Intrinsic disorder in plant proteins and phytopathogenic bacterial effectors. Chem Rev. 2014;114(13):6912-32.
114. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

115. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851-69.
116. Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics. 2018;16(1):17-32.
117. 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(5):685-92.
118. Wang S, Ma JZ, Xu JB. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics. 2016;32(17):672-9.
119. Gao JZ, Yang YD, Zhou YQ. Grid-based prediction of torsion angle probabilities of protein backbone and its application to discrimination of protein intrinsic disorder regions and selection of model structures. Bmc Bioinformatics. 2018;19.
120. Hanson J, Paliwal KK, Zhou Y. Accurate Single-Sequence Prediction of Protein Intrinsic Disorder by an Ensemble of Deep Recurrent and Convolutional Architectures. J Chem Inf Model. 2018.
121. Dyson HJ. Roles of intrinsic disorder in protein-nucleic acid interactions. Molecular bioSystems. 2012;8(1):97-104.
122. Wu Z, Hu G, Yang J, Peng Z, Uversky VN, Kurgan L. 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):2561-9.
123. Basu S, Bahadur RP. A structural perspective of RNA recognition by intrinsically disordered proteins. Cell Mol Life Sci. 2016;73(21):4075-84.
124. Dunker AK, Uversky VN. Drugs for 'protein clouds': targeting intrinsically disordered transcription factors. Curr Opin Pharmacol. 2010;10(6):782-8.
125. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. 2017.
126. Kauffman C, Karypis G. Computational tools for protein–DNA interactions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;2(1):14-28.
127. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein–RNA interactions. Journal of Structural Biology. 2012;179(3):261-8.
128. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. Brief Bioinform. 2016;17(1):88-105.
129. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. Brief Bioinform. 2017.
130. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res. 2017;45(10):e84.
131. Zhang T, Zhang H, Chen K, Ruan J, Shen S, Kurgan L. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. Curr Protein Pept Sci. 2010;11(7):609-28.
132. Chowdhury S, Zhang J, Kurgan L. In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome. Proteomics. 2018:e1800064.
133. Si J, Cui J, Cheng J, Wu R. Computational Prediction of RNA-Binding Proteins and Binding Sites. Int J Mol Sci. 2015;16(11):26303-17.
134. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein-RNA interactions. J Struct Biol. 2012;179(3):261-8.
135. Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. Mol Biosyst. 2013;9(10):2417-25.
136. Ding XM, Pan XY, Xu C, Shen HB. Computational prediction of DNA-protein interactions: a review. Curr Comput Aided Drug Des. 2010;6(3):197-206.

137. Si J, Zhao R, Wu R. An overview of the prediction of protein DNA-binding sites. Int J Mol Sci. 2015;16(3):5194-215.
138. Zhao H, Wang J, Zhou Y, Yang Y. Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome. PLoS One. 2014;9(5):e96694.
139. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, et al. D(2)P(2): database of disordered protein predictions. Nucleic Acids Res. 2013;41(Database issue):D508-16.
140. Di Domenico T, Walsh I, Martin AJM, Tosatto SCE. MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics. 2012;28(15):2080-1.
141. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res. 2005;33(Web Server issue):W72-6.
142. Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Research. 2014;42(W1):W337-W43.
143. Cheng J, Li J, Wang Z, Eickholt J, Deng X. The MULTICOM toolbox for protein structure prediction. BMC Bioinformatics. 2012;13:65.