Computational prediction of intrinsic disorder in proteins

Fanchi Meng[1], Vladimir Uversky[2,3], and Lukasz Kurgan[4]*

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada

[2]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA

[3]Laboratory of Structural Dynamics, Stability and Folding of Proteins, Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia

[4]Department of Computer Science, Virginia Commonwealth University, Richmond, USA.

* Corresponding author
Email: lkurgan@vcu.edu; Phone: 804-827-3986

How to cite this article:
Meng, F., Uversky, V., & Kurgan, L. (2017). Computational prediction of intrinsic disorder in proteins. Current Protocols in Protein Science, 88, 2.16.1–2.16.14. doi: 10.1002/cpps.28

ABSTRACT

Computational prediction of intrinsically disordered proteins (IDPs) is a mature research field. These methods predict disordered residues and regions in an input protein chain. Over 60 predictors of IDPs were developed so far. This unit defines computational prediction of intrinsic disorder, summarizes major types of predictors of disorder, and provides details about three accurate and recently released methods. We illustrate their predictions using a few sample proteins, provide insights how these predictions should be interpreted, and discuss and quantify their predictive performance. We comment on how easy it is to collect these predictions using freely and conveniently accessible webservers. Lastly, we point to the availability of databases that provide access to annotations of native and pre-computed putative intrinsic disorder and we summarize a few experimental methods that can be used to validate computational predictions.

Key words: intrinsic disorder; intrinsically disordered protein; prediction;

INTRODUCTION

While many proteins maintain a well-defined tertiary structure, many lack structure under physiological conditions and take the form of dynamic conformational ensembles. These intrinsically disordered proteins (IDPs) (Dunker et al., 2013; van der Lee et al., 2014) lack the structure along their entire amino acid chain or in specific regions. IDPs are highly abundant in nature. According to a few recent estimates, 19, 6, and 4% of amino acids are disordered in eukaryotes, bacteria, and archaea (Peng et al., 2015b), respectively, between 30 and 50% of eukaryotic proteins (depending on an organism) have at least one long (≥ 30 consecutive amino acids) intrinsically disordered region (IDR) (Dunker et al., 2000; Ward et al., 2004b; Xue et al., 2012), and between 6 and 17% of proteins encoded by various genomes are fully disordered (Tompa, 2002). Furthermore, 44% of protein-coding genes in human include long disordered regions (Oates et al., 2013). The IDPs participate in a diverse range of cellular functions (Peng et al., 2015b; van der Lee et al., 2014). They play important functional roles in transcription (Fuxreiter et al., 2008; Liu et al., 2006), translation (Peng et al., 2014), protein-protein interactions (Dunker et al., 2005; Fuxreiter et al., 2014), protein-RNA interactions (Varadi et al., 2015; Wang et al., 2016) and cell signaling (Dyson and Wright, 2005; Galea et al., 2008; Uversky et al., 2005; Xie et al., 2007), to name but a few. IDPs are also associated with various human diseases (Uversky et al., 2008) and they were recently suggested to be attractive targets for drug discovery (Hu et al., 2015). Several databases of IDPs are available, such as DisProt (Sickmeier et al., 2007), the largest database of manually curated and functionally annotated IDRs, and IDEAL (Fukuchi et al., 2014), which includes information about binding partners of IDPs. Moreover, IDRs can be found in the Protein Data Bank (PDB) (Berman et al., 2000) as residues with missing coordinates in crystal structures and highly flexible residues in NMR structures (Martin et al., 2010). However, these repositories of experimentally annotated intrinsic disorder represent only a small fraction of sequences in nature. The total number of IDPs in IDEAL and DisProt is only 713 and 803, respectively, while the number of currently known proteins that are included in the UniProt resource has already reached 68 million.

Interestingly, sequences of IDRs are different when compared to structured regions and proteins. For instance, the disordered regions have specific composition of amino acids, lower sequence complexity, and lower propensity to form alpha and beta secondary structure (Li et al., 1999; Pentony et al., 2010; Romero et al., 1997; Romero et al., 2001). Given these differences, the experimentally annotated IDRs and IDPs are used to empirically derive predictive models, which in turn are used to predict intrinsic disorder for the millions of the unannotated proteins. These methods use protein sequence as their input and generate propensity for intrinsic disorder for each residue in this sequence as their output. A study in 2012 indicated that there were approximately 60 computational predictors of disorder (Kozlowski and Bujnicki, 2012). In 2016, we found approximately 70 such predictors. We classify these methods into four categories:

1) Scoring function-based methods. The propensity for disorder is calculated using a function that takes physiochemical properties of individual amino acid in the input protein chain as its inputs. Examples methods include NORSP (Liu and Rost, 2003), GlobPlot (Linding et al., 2003b) and IUPred (Dosztányi et al., 2005a; Dosztányi et al., 2005b).
2) Machine learning-based methods. The propensity for disorder is computed using a predictive model generated by a machine learning algorithm (such as neural network and support vector machine (SVM)) using inputs derived based on physiochemical properties of amino acid, evolutionary conservation, and putative secondary structure and solvent accessibility. Examples are DisEMBL (Linding et al., 2003a), DISOPRED (Jones and Cozzetto, 2015; Jones and Ward, 2003), and a family of VLS predictors (Obradovic et al., 2003; Obradovic et al., 2005).
3) Meta methods. These methods combine prediction of multiple predictors of disorder with information extracted from protein sequence and putative structural properties of the sequence (secondary structure and solvent accessibility) to predict propensity for disorder. They include MFDp (Mizianty et al., 2010), MetaDisorder (Kozlowski and Bujnicki, 2012) and PONDR-FIT (Xue et al., 2010).
4) Hybrid methods. These predictors combine the abovementioned machine learning approach with structural modelling, typically using template-based structure predictions. Examples are PrDOS (Ishida and Kinoshita, 2007) and Disoclust3 (McGuffin et al., 2015).

This unit defines computational prediction of disorder, summarizes three arguably most accurate predictors, and present a case study that explains and compares their predictions.

PREDICTION OF INTRINSIC DISORDER FROM SEQUENCE

Computational predictors of intrinsic disorder use protein sequence as their only input. They generate putative propensity for intrinsic disorder for every residue in the input protein sequence. Typically, this propensity is expressed as a numeric score where a low value denotes high propensity for a structured conformation and a high value denotes propensity for the disordered state. Besides this numeric propensity, most of the predictors also offer a binary prediction where each residue is categorized as either structured or disordered.

We illustrate predictions of intrinsic disorder and contrast these predictions with the native annotations of disorder using the ICln protein (Figure 1). This protein is a chloride channel that is involved in regulation of several cellular processes including membrane ion transport and RNA splicing. Structure of ICln, which was solved using NMR, is composed of several

superimposed conformations. The regions colored in blue converge to the same confirmation and constitute structured regions. The disordered regions that are colored in red form an ensemble of diverse conformations. The annotation of disordered regions was collected from DisProt version 6.02 and is shown below the image in Figure 1 ("Native Dis" line). The figure includes eight IDRs which are numbered in the structure as they appear along the amino acid sequence. The structure excludes parts of both termini of this protein, including residues 1 to 18 and residues 134 to 235, which are disordered. The bottom part of Figure 1 includes predictions of three methods: MFDp (meta method), DISOPRED (machine-learning method) and PrDOS (hybrid method). Both binary and numeric scores are included where the numeric scores that range between 0 and 1 are represented by the first digit after the decimal point. A given residue is predicted as disordered if its predicted numeric propensity is high: 0.5 and above for PrDOS and DISOPRED, and >0.37 for MFDp; otherwise it is predicted as structured.

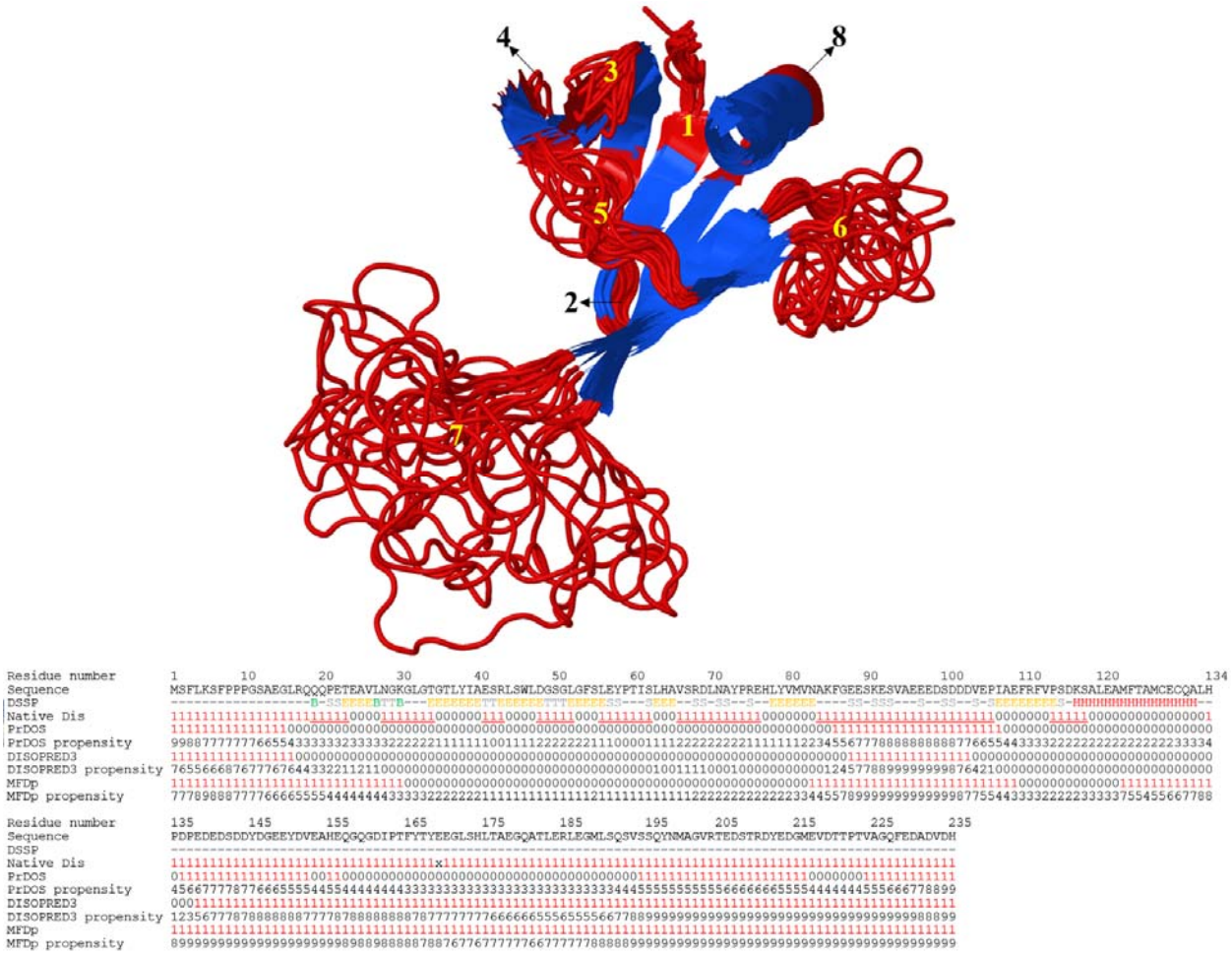

Figure 1. Native intrinsic disorder and putative disorder for the ICln protein (PDB ID: 1ZYI; DisProt ID: DP000717). The top portion of the figure is a cartoon view of multiple superimposed NMR structures of this protein taken from PDB. Ordered regions where all structures converge to the same conformation and disordered regions that form conformational ensembles are colored in blue and red, respectively. The eight disordered regions are numbered 1 to 8, and they correspond to eight underlined disorder regions in the "Native Dis" line. The bottom part shows native disorder annotations collected from DisProt along with putative disorder annotations generated with PrDOS, DISOPRED3 and MFDp methods. The first line shows residue number which is followed by the amino acid sequence. The third line shows 8-state secondary structure assigned with DSSP, where B is an isolated β-strand residue, E is an extended strand that forms β sheets, G is 3_{10} helix, H is α-helix, I is π helix, S is a bend, T is a

hydrogen bonded turn and – denotes residue without specific secondary and tertiary structure. The forth line shows native annotation of disorder where 0 denotes structured residues, 1 denotes disordered residues, and x denotes a residue that lacks annotation. The following six lines show binary predictions and propensity scores. The propensity that ranges between 0 and 1 is represented by the first digit after the decimal point.

To interpret results produced by the computational predictors, users should first analyze the binary predictions in order to extract the corresponding putative IDRs and structured regions. Next, each predicted IDRs should be assessed using the numeric propensities. Residues that have high scores are more likely to be disordered and the corresponding predictions are more likely to be accurate. Users can also analyze the scores of all residues in a given putative IDR, which is annotated based on binary predictions, to quantify the likelihood of this entire region to be correctly identified. On the other hand, low scores can be used to identify structured residues and regions. The predictions for residues with scores close to 5 for PrDOS and DISOPRED, and close to 4 for MFDp (these values are used to convert the propensity into the binary prediction) are arguably less accurate than the predictions with either high or low scores. We also recommend that, if possible, multiple methods should be used and the users should rely on a consensus-based prediction. In other words, IDRs and disordered residues predicted by multiple methods are more likely to be correct compared with predictions that disagree between different methods. The favourable predictive performance of a consensus-based approach was shown empirically in a few recent studies (Fan and Kurgan, 2014; Peng and Kurgan, 2012a).

SELECTED COMPUTATIONAL PREDICTORS OF INTRINSIC DISORDER

We introduce three accurate predictors of intrinsic disorder, DISOPRED, MFDp and PrDOS. These methods were ranked as the top three in predictive performance among 28 methods that were assessed during the most recent Critical Assessment of protein Structure Prediction (CASP) experiment, CASP10 (Monastyrskyy et al., 2014), that featured evaluation of predictions of disorder. CASP is a biannual worldwide event in which predictions submitted by research labs across the world are assessed on a blind dataset of proteins (these proteins that were not available to the participants ahead of time) by a group of independent assessors who not participate in the event. The three predictors were also ranked among the top three in other recent comparative reviews (Deng et al., 2012; Peng and Kurgan, 2012b). We list them in chronological order and discuss their origin, key architectural characteristics, and several practical aspects. The latter include their inputs, outputs, and availability.

PrDOS (2007)

PrDOS was created by Ishida and Kinoshita at the University of Tokyo (Ishida and Kinoshita, 2007). This is a hybrid method that combines a machine learning approach that relies on an SVM model with a template-based model. The machine learning model uses an evolutionary profile of the input sequence as its input. The template-based model searches for homologues in PDB. The prediction is based on a weighted average of the results produced by the machine learning and template-based models.
Input: One FASTA-formatted or raw amino acid sequence.
Output: Putative binary disorder annotation and propensity scores for each residue.
Availability: PrDOS is available as a webserver at http://prdos.hgc.jp/cgi-bin/top.cgi

MFDp (2010)

MFDp was developed by Kurgan's group at the University of Alberta (currently at the Virginia Commonwealth University) (Mizianty et al., 2010). This meta predictor combines three SVM models that are specialized to predict long, short, and all-size IDRs. Each SVM utilizes a diverse set of inputs that include information extracted directly from the amino acid sequence and from putative disorder predicted by three predictors, evolutionary profile, putative B-factors, putative secondary structure and backbone dihedral torsion angles, putative solvent accessibility, and putative annotation of globular domains. This method was upgraded to a new version, MFDp2, in 2013 (Mizianty et al., 2013; Mizianty et al., 2014). MFDp2 combines predictions generated by MFDp with predictions computed based on alignment against a database of disordered proteins extracted from DisProt. These predictions are corrected such that the number of predicted disordered residues matches the number of putative disordered residues output by DisCon method (Mizianty et al., 2011).
Input: Up to 5 FASTA-formatted amino acid sequences for MFDp. Up to 100 FASTA-formatted amino acid sequences for MFDp2.
Output: Putative binary disorder annotation and propensity scores for each residue.
Availability: MFDp is available as a webserver at http://biomine-ws.ece.ualberta.ca/MFDp. MFDp2 is available as a webserver at http://biomine-ws.ece.ualberta.ca/MFDp2

DISOPRED3 (2015)

DISOPRED was released by Jones's group at the University College London (Jones and Cozzetto, 2015). The first version of this method was published in 2003 (Jones and Ward, 2003), the second version, DISOPRED2, in 2004 (Ward et al., 2004a) and the newest third version, DISOPRED3, in 2015 (Jones and Cozzetto, 2015). DISOPRED3 is a machine learning method implemented as a two-stage neural network which uses predictions from three predictors: DISOPRED2, a specialized predictor of long IDRs, and a nearest neighbor-based model that uses similarity to a set of proteins annotated with IDRs. This design is to some extend similar to MFDp2 that also includes a module that predicts long IDRs and an alignment-based module. The main differences are the input information that consists of an evolutionary profile and the second stage that combines these three predictions using a neural network. Moreover, DISOPRED3 also predicts protein binding sites, defined as protein binding regions located inside IDRs. DISOPRED3 is a part of a comprehensive protein sequence analysis workbench PSIPRED that includes predictors of tertiary and secondary protein structure, membrane helices and topology of transmembrane helices, protein domains, and protein functions.
Input: One FASTA-formatted or raw amino acid sequence, or multiple sequence alignment.
Output: Putative binary disorder annotation and propensity scores for each residue. Putative binary annotations and propensity scores for disordered protein binding sites.
Availability: DISOPRED3 is available as a webserver and standalone package running on Linux platform at http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1

Each of these methods offers a convenient and user-friendly webserver. A user only needs a web browser and internet connection to use these webservers. After arriving at the specific URL that is listed above, a user only needs to provide the sequences of the protein and request the

prediction by clicking on the "run" button. The computations are performed on the server side and delivered back via the web site and/or to a user-provided email.

Besides these webservers, users have an option to employ databases that provide access to pre-computed predictions of intrinsic disorder. An advantage of these databases is that the predictions are available instantly, while the predictors require up to a few minutes to predict one protein. However, the databases are limited to a specific list of proteins while the predictors can generate putative disorder for any sequence provided by the users. Two largest databases of putative intrinsic disorder are MobiDB at http://mobidb.bio.unipd.it/ (Di Domenico et al., 2012; Potenza et al., 2015) and D^2P^2 at http://d2p2.pro/ (Oates et al., 2013). MobiDB offers access the putative disorder generated by ten predictors and a consensus of these predictions. It also provides access to experimental annotations of disorder collected from DisProt and PDB. The current MobiDB's version 2.3.2014.07 includes over 80 million proteins which are cross-referenced to UniProt (Consortium, 2010). D^2P^2 stores results of nine predictors, is linked to the experimental data from DisProt and IDEAL (Fukuchi et al., 2012), and includes annotations of putative disordered protein binding regions. It covers over 10 million proteins from complete proteomes of 1,765 distinct species. The main difference between these two resources is that MobiDB includes a larger set of proteins while D^2P^2 focuses on complete proteomes.

ASSESSMENT OF PREDICTIVE PERFORMANCE OF COMPUTATIONAL PREDICTORS OF INTRINSIC DISORDER

One of important aspects in the context of the prediction of intrinsic disorder is predictive performance. Since the predictions are in two formats: binary and propensity, we define two corresponding and widely accepted metrics of predictive performance: Matthews's correlation coefficient (MCC) and area under Receiver operating characteristic (AUC). These measures were used in CASP and other comparative evaluations.

MCC that is used to evaluate the binary predictions is defined as:

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

where *TP* (true positives) is the number of correctly predicted disorder residues, *FN* (false negatives) is the number of disorder residues predicted as structured, *FP* (false positives) is the number of structured residues predicted as disordered, and *TN* (true negatives) is the number of correctly predicted structured residues. Like other correlation coefficients, MCC ranges between –1 and +1, where 0 denotes lack of correlation (predictions are no better than random) and a larger positive value denotes higher predictive performance. Negative correlation, which does not happen in practise, indicates that structured residues are predicted primarily as disordered and vice versa, meaning that the predictions are inverted.

AUC is used to assess propensity scores and it quantifies area under the ROC curve defined as a relation between true positive rate, $TPR = TP/(TP + FN)$, and false positive rate, $FPR = FP/(FP + TN)$. The curve is composed of multiple points that correspond to the *TPR* and *FPR* values computed at different thresholds imposed over the propensity scores, where the residues with scores above (below) the threshold are assumed to be predicted as disordered (structured). AUC

values range between 0.5, which is equivalent to a random predictor, and 1 that implies perfect prediction.

Next, we analyze results from the DISOPRED3, MFDp and PrDOS methods for the ICln protein (Figure 1) and quantify their predictive quality using MCC and AUC. The predictions from DISOPRED3 correctly identify the three largest disordered regions: the region at the N-terminus, the seventh IDRs and the region at the C-terminus. However, they fail to identify the remaining shorter IDRs. The predictions from MFDp similarly cover the three longest native IDRs, where the two putative regions at the termini are elongated. Finally, PrDOS also identifies the three largest IDRs but it makes mistakes for parts of the IDR at the C-terminus. Overall, these methods identify majority of disordered residues correctly and miss a few shorter disordered regions. The MCC values of DISOPRED3, MFDp and PrDOS are 0.57, 0.38, and 0.39, respectively. These values suggest that the predictions are correlated with the native disorder, with the strongest correlation attributed to DISOPRED3. These numbers agree with our analysis. The propensity scores from DISOPRED3, MFDp and PrDOS methods follow a similar pattern. The three long IDRs identified by the three predictors have high scores, which suggest that the user should be confident that the predictions are correct. One exception is the IDR predicted by PrDOS at the C-terminus that was scored in the 5 to 7 range compared to the 7 to 8 range for the other two IDRs predicted by this method. DISOPRED also provides low scores for the structured regions (annotated using 0 in the "Native Dis" line in Figure 1). Such low scores suggest that these predictions are likely correct. The AUC values that quantify predictive performance of these scores are 0.92, 0.83 and 0.84 for DISOPRED3, MFDp and PrDOS, respectively. They again agree with our observations, in particular pointing to the high quality of scores generated by DISOPRED3.

Next, we assess these three methods on three proteins to provide insights on how their predictive performance varies depending on the input protein. The three proteins were selected from DisProt version 6.02 and were deposited into this database after the version 5.9 was released. This means that they were not available where the three predictions were developed and thus can be used to perform a blind test (test on proteins that were not used to design these methods). Table 1 summarizes the AUC and MCC values of PrDOS, DISOPRED3 and MFDp for the three proteins. The average AUC values of these predictors are similar and equal 0.80 for PrDOS and 0.83 for MFDp and DISOPRED3. The average MCC values are also comparable and equal 0.52 for DISOPRED3 and 0.39 for MFDp and PrDOS. These values agree with the results of recent comparative reviews of predictors of intrinsic disorder. In a study by Cheng's group DISOPRED, MFDp and PrDOS were shown to achieve AUC = 0.85, 0.82 and 0.85, respectively (Deng et al., 2012). In another study by Kurgan's group DISOPRED and MFDp were shown to secure AUC = 0.78 and 0.82, and MCC = 0.41 and 0.45, respectively (Peng and Kurgan, 2012b). Interestingly, results in Table 1 reveal that there is no universally best method. For instance, while AUC of DISOPRED is the highest for the ICln and PPARG proteins, the results of this predictor are outperformed by both MFDp and PrDOS for the CRK protein. This observation supports our advice to use and combine results from multiple methods in order to secure predictions that are characterized by higher predictive performance.

Table 1. Comparison of predictive performance of three predictors of intrinsic disorder (PrDOS, DISOPRED3 and MFDp) and their consensus for three disordered proteins collected from DisProt. The highest AUC and MCC values for each protein are shown in bold font. The consensus binary prediction is based on a majority vote (residue is assumed disordered if most methods predict it as disordered, otherwise it is predicted as structured). The consensus propensities are calculated as average of the propensities for methods that predict a given residue as disordered (structured) if the binary prediction for this residue is that it is disordered (structured). The runtime, which is measured in minutes, is computed as an average over five predictions for the same protein on the same webserver. Proteins are sorted by their length from shortest to longest to demonstrate that runtime increases with the protein length.

Protein name	Protein	AUC				MCC				Runtime [minutes]		
(DisProt ID)	length	PrDOS	DISOPRED3	MFDp	Consensus	PrDOS	DISOPRED3	MFDp	Consensus	PrDOS	DISOPRED3	MFDp
ICln (DP000717)	235	0.84	**0.92**	0.83	0.88	0.39	**0.57**	0.38	0.59	12±8.7	55±15.1	6±0.0
CRK (DP00748)	304	0.75	0.69	**0.79**	0.77	0.49	**0.51**	0.35	0.54	15±13.5	62±24.5	7±0.5
PPARG (DP00718)	477	0.79	**0.88**	0.87	0.88	0.28	**0.49**	0.44	0.45	16±15.3	196±111.8	11±0.4
Average (± standard deviation)	338±124	0.80±0.05	0.83±0.13	0.83±0.04	**0.84**±0.07	0.39±0.10	0.52±0.04	0.39±0.05	**0.53**±0.76	14±2.1	104±79.5	8±2.6

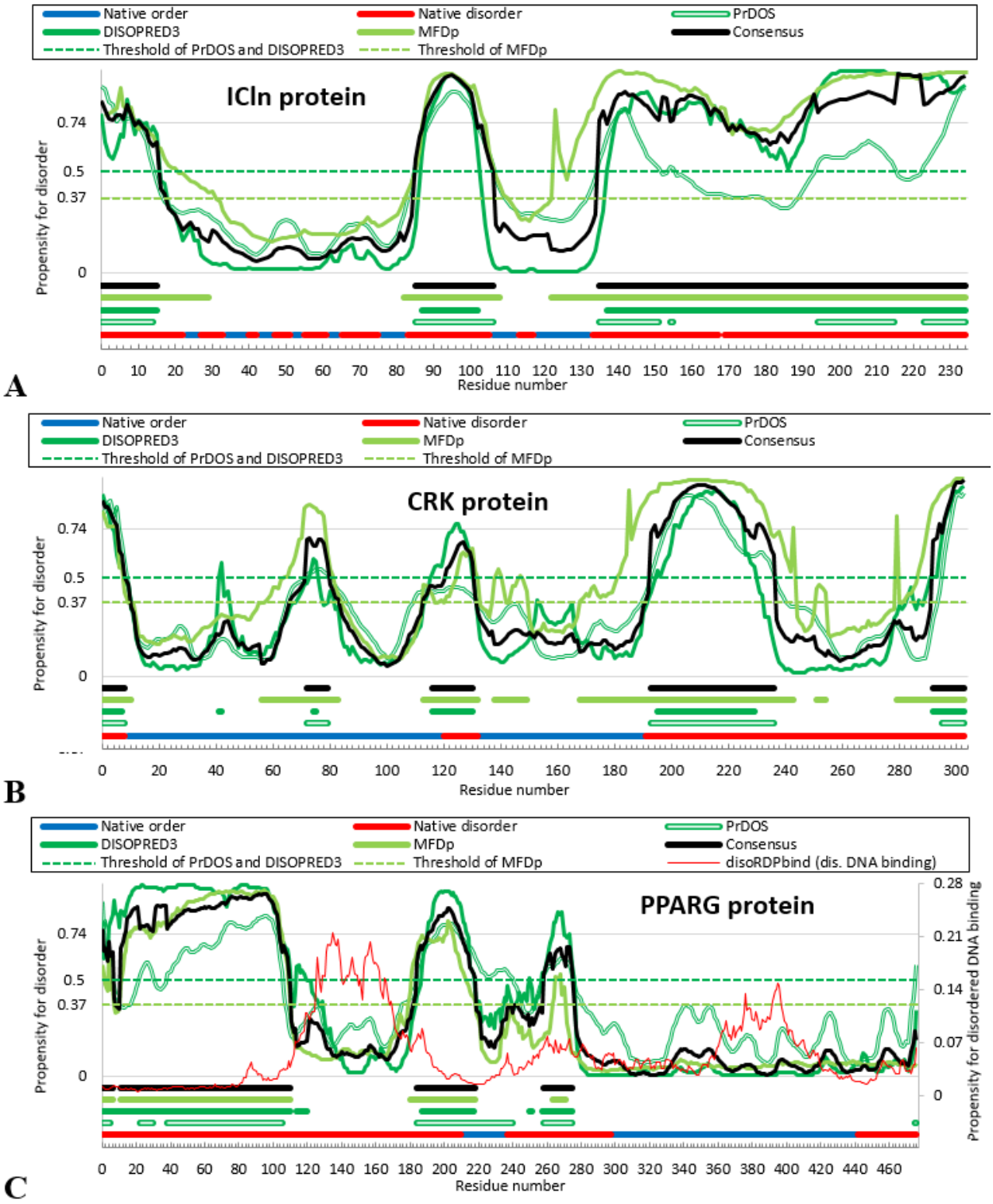

Figure 2. Visualization of predictions generated by PrDOS, DISOPRED3 and MFDp and a consensus-based prediction for ICln protein (Panel A; DisProt ID: DP00717), CRK protein (Panel B; DisProt ID: DP00748), and Peroxisome proliferator-activated receptor gamma (PPARG) protein (Panel C; DisProt ID: DP00718_A001). The *x*-axis denotes the protein sequence and plots at the top of each panel show the propensity scores; higher propensity values indicate higher likelihood for disorder. Propensities generated by PrDOS, DISOPRED3 and MFDp and consensus-based method are shown using hollow dark green, dark green, light

green and black lines, respectively. The binary annotations of both native and putative disorder (each residue is categorized as either disordered or structured) are shown with horizontal lines at the bottom of each panel. The native binary annotations of disordered and structured residues that were collected from DisProt are shown using red and blue line, respectively. The putative binary annotations for the three predictors and the consensus-based predictor follow the color scheme of the propensity plots. They were obtained from the putative propensities by using thresholds = 0.5 for PrDOS and DISOPRED3 and = 0.37 for MFDp that were suggested by the authors, i.e., residues with propensities above the corresponding threshold are predicted as disordered and below the threshold as structured. The threshold values are shown using dashed horizontal lines. Panel C also includes prediction of disordered DNA binding residues shown using thin red line that were generated with the DisoRDPbind method.

We also measure the runtime required to obtain predictions for these three proteins using the public webservers of the selected three methods. We predict each protein five times on the same webserver and we report the average time and the corresponding standard deviation in Table 1. MFDp requires the least amount of time, with an average of about 8 minutes per protein. The second fastest is PrDOS which needs 14 minutes per protein while DISOPRED3 needs over 100 minutes per protein. We caution the reader that this estimate includes the time to run a given method on the input protein and possibly also wait time in a queue of other jobs on a given webserver. This is evident by the relatively high values of the standard deviations, in particular the deviation of 80 minutes for DISOPRED3. The runtime also depends on the length of a given protein meaning that, as expected, longer proteins require more time (Table 1). The PPARG protein that is twice as long as the ICln protein requires about twice the runtime when using MFDp, 25% longer runtime using PrDOS and about three time longer runtime using DISOPRED3. Overall, the users should expect that a single prediction takes typically several minutes with more time needed for longer proteins.

Figure 2 visualizes the predictions from PrDOS, DISOPRED3 and MFDp on the three proteins and compares these predictions with the native annotations of disorder. The three methods generate relatively similar predictions for the ICln protein (Figure 2A). They correctly annotate the two IDRs at both termini and also the large IDR between positions 85 and 107, except for PrDOS that under-predicts the region at the C-terminus. At the same time, the three methods equally struggle to find several short IDRs located between positions 29 and 77.
In contrast to the results for the two above proteins, the predictions for CRK protein differ between the three methods (Figure 2B). MFDp predicts seven IDRs, DISOPRED3 predicts six IDRs, and PrDOS annotates four putative disordered regions. The propensity scores generated by these methods diverge particularly between positions 110 and 180. The three predictors identify the short native IDRs located at the N-terminus and parts of the long IDR at the C-terminus. The short IDR between positions 122 and 134 is correctly predicted by MFDp and DISOPRED3 and is missed by PrDOS. Moreover, MFDp and PrDOS identify a longer false IDR near position 75 while DISOPRED3 incorrectly predicts only two disordered residues there. MFDp also predicts another false IDR near position 140, and DISOPRED incorrectly annotates a couple of disordered residues near position 40. The PPARG protein includes three IDRs, one at each termini and one between positions 239 and 299 (Figure 2C). Again, we observe that the three predictions are in agreement with each other. They correctly find the IDR at the N-terminus and a fragment of the internal to the sequence IDR, while they fail to identify the IDR at the C-terminus. The propensities generated by DISOPRED3 and MFDp are better than the propensities generated by PrDOS since they have higher values for the disordered regions and lower values for the structured regions. This is why DISOPRED and MFDp secure higher AUCs than PrDOS for this protein (Table 1). We note that the three methods consistently predict a fragment of the IDR at the N-terminus as structured. This IDR was shown to interact with DNA and the nature of

this interaction that stabilizes protein structure is perhaps the reason for this incorrect prediction (Chandra et al., 2008). To this end, we use DisoRDPbind method (Peng and Kurgan, 2015; Peng et al., 2015a) to predict disordered DNA-binding regions and found that its predictions (shown using thin red line in Figure 2B) complement the disorder predictions for that fragment of the IDR at the N-terminus. More specifically, DisoRDPbind identifies a disordered region that binds DNA and this region fills in the gap in the disorder predictions. This suggests that methods that predict specific functions of disordered regions, such as DisoRDPbind that predicts disordered protein-, DNA- and RNA-binding regions, may offer information that complements the results produced by the predictors of "generic" disordered regions.

Overall, these examples demonstrate that the predictions of disorder are relatively accurate and can be used to identify IDRs in the input protein chains. Our examples convey the richness of the information that can be obtained from the disorder predictions and illustrate how to analyze and understand these predictions.

CONSENSUS-BASED PREDICTIONS

One of aspects related to the interpretation of the results generated by different predictors is how to proceed when these methods are in disagreement. This is particularly relevant to the CRK protein where our three predictors diverge (Figure 2B). We suggest to use a consensus approach where the final prediction is determined by a majority of the results generated by the considered methods. In the case of the binary predictions, a given residue should be assumed disordered if most methods predict it as disordered, otherwise it should be predicted as structured. The propensities should be calculated as an average of the propensities generated by the methods that predict a given residue as disordered (structured) if the binary prediction is disordered (structured). This approach is inspired by favourable empirical results of the consensus-based approaches when compared to the predictions of individual methods (Fan and Kurgan, 2014; Peng and Kurgan, 2012a) and the fact that consensus predictions are provided in the D^2P^2 and MobiDB databases. We visualize the consensus predictions in Figure 2 using black lines and include the corresponding predictive quality in Table 1. The consensus prediction for the ICln protein (Figure 2A) avoids the pitfalls of the PrDOS's prediction of the IDR at the C-terminus, trims the incorrectly elongated putative IDR generated by MFDp at the N-terminus, and more accurately delineates the long disordered regions at position 82 compared to the output of DISOPRED3. Consequently, it secures the highest MCC value for this protein (Table 1). Similarly, for the other two proteins the consensus resolves the disagreements between the three predictors in a way that generally is in a better agreement with the native annotations of disorder. For instance, the consensus trims a number of disordered residues that were overpredicted by MFDp and adds the disordered region at position 120 that was missed by PrDOS for the CRK protein (Figure 2B). It also fixes the incorrect predictions from PrDOS at positions 220 to 238 where a structured region is incorrectly predicted as disordered (Figure 2C). Overall, the consensus secures a higher AUC and MCC values compared with each of the three predictors (Table 1). However, we note that this improvement comes at a cost of running the three predictors which takes more runtime than running a single method.

EXPERIMENTAL MEANS FOR VALIDATION OF PREDICTED DISORDER

There are multiple experimental approaches that may be used to validate and support predictions of the intrinsic disorder. Similar to the outputs of different predictors that either generate information on the overall disorder status of a whole protein molecule or provide the per-residue disorder score, experimental techniques also describe the whole protein or give a residue-level information. There are several reviews and books that describe a wide range of experimental techniques that can be used to characterize intrinsic disorder in proteins (Daughdrill et al., 2005; Eliezer, 2009; Receveur-Brechot et al., 2006; Uversky, 2015; Uversky and Dunker, 2012a; Uversky and Dunker, 2012b; Uversky and Dunker, 2012c; Uversky and Longhi, 2010). The number of such experimental techniques amounts to almost 70. Detailed description of these approaches is outside the scope of this unit and here we summarize four selected techniques: X-ray crystallography, NMR, limited proteolysis and hydrogen-deuterium exchange. These methodologies provide information on the intrinsic disorder at the residue level. From the viewpoint of natural propensity of an amino acid sequence for the intrinsic disorder, these four techniques are non-invasive since their application does not require introduction of the amino acid substitutions, which can affect predisposition of a protein for the intrinsic disorder.

Although X-ray crystallography is traditionally used to describe atomic-level structures of structured proteins, the increased flexibility of atoms in the structured regions is reflected in high values of their B-factor, whereas high flexibility of atoms in the disordered regions is responsible for the non-coherent X-ray scatter in the crystallographic experiments. As a consequence of the non-coherent X-ray scatter the corresponding atoms become “invisible”, giving rise to the missing electron density regions (Le Gall et al., 2007; Radivojac et al., 2004). Therefore, if a crystal structure of a protein that contains both structured and disordered regions is available, then the validity of the predicted disorder of some of its regions can be verified by looking for the presence of regions with missing electron density (remark 465) in the corresponding PDB entry. The NMR spectroscopy is the technique of choice for providing high-resolution, residue-level structural information on the intrinsically disordered proteins. In fact, heteronuclear multidimensional NMR can generate precise structural information on IDPs/IDRs via assignment of their resonances and can also provide direct measurement of the mobility of IDRs (Daughdrill et al., 2005; Eliezer, 2009; Jensen et al., 2010; Nodet et al., 2009; Salmon et al., 2010). Both, limited proteolysis and hydrogen-deuterium exchange are based on the solvent accessibility of corresponding target sites. A high solvent accessibility of the potential cleavage sites makes non-folded proteins highly susceptible to proteolytic degradation *in vitro* (Fontana et al., 2004). Therefore, limited proteolysis can be used to indirectly confirm the increased conformational flexibility of IDPs and IDRs (Fontana et al., 2012) and thereby confirm the results of a disorder prediction. Similarly, structural information and detailed description of the dynamics of a protein chain can be obtained from the analysis of the efficiency and rates of incorporation of deuterium into a protein’s backbone amide. This is achieved via monitoring hydrogen/deuterium exchange in proteins by mass spectrometry combined with the high performance liquid chromatography (Smith et al., 1997). The ability of this technique to distinguish between structured and disordered protein regions by their level of protection against hydrogen/deuterium exchange makes it suitable to detect intrinsic disorder and to validate predictions of disorder (Bobst and Kaltashov, 2012).

CONCLUSIONS

The first predictor of intrinsic disorder was developed over 35 year ago. With dozens of new predictors that were developed over the last three decades, their predictive performance and availability has substantially improved. Modern predictors are characterized by sophisticated designs that are based on meta and hybrid approaches, utilize state-of-the-art machine learning algorithms, and are available to the users as convenient webservers. Most importantly, predictions generated by these methods are accurate, with AUC values at about 0.8 and MCC values in the 0.4 to 0.5 range. We describe and illustrate inputs, outputs, architectures, predictive performance, and runtime of three popular and accurate predictors. We also discuss how to proceed when the predictions of different methods disagree and suggest several experimental methods that can be used to validate the predictions. Moreover, we describe several databases of native and putative annotations of disordered residues. While these methods and databases reaches the point of maturity, research in this area has recently shifted to the prediction of various functions of the disordered regions. These functions include protein-protein binding regions (Disfani et al., 2012; Dosztanyi et al., 2009; Jones and Cozzetto, 2015; Malhis et al., 2016; Peng and Kurgan, 2015; Peng et al., 2015a; Yan et al., 2015), protein-RNA and protein-DNA binding regions (Peng and Kurgan, 2015; Peng et al., 2015a), and disordered linkers (Meng and Kurgan, 2016), to name a few that were already addressed.

REFERENCES

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research* 28:235-242.

Bobst, C.E. and Kaltashov, I.A. 2012. Localizing flexible regions in proteins using hydrogen-deuterium exchange mass spectrometry. *Methods Mol Biol* 896:375-385.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P., and Rastinejad, F. 2008. Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. *Nature* 456:350-356.

Consortium, T.U. 2010. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* 38:D142-D148.

Daughdrill, G.W., Pielak, G.J., Uversky, V.N., Cortese, M.S., and Dunker, A.K. 2005. Natively Disordered Proteins. *In* Handbook of Protein Folding (J. Buchner and T. Kiefhaber, eds.) pp. 271-353. Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany.

Deng, X., Eickholt, J., and Cheng, J. 2012. A comprehensive overview of computational protein disorder prediction methods. *Molecular BioSystems* 8:114-121.

Di Domenico, T., Walsh, I., Martin, A.J.M., and Tosatto, S.C.E. 2012. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* 28:2080-2081.

Disfani, F.M., Hsu, W.L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N., and Kurgan, L. 2012. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 28:i75-83.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. 2005a. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21:3433-3434.

Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. 2005b. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *Journal of Molecular Biology* 347:827-839.

Dosztanyi, Z., Meszaros, B., and Simon, I. 2009. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25:2745-2746.

Dunker, A.K., Babu, M.M., Barbar, E., Blackledge, M., Bondos, S.E., Dosztányi, Z., Dyson, H.J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., Han, K.-H., Jones, D.T., Longhi, S., Metallo, S.J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R.V., Rost, B., Selenko, P., Subramaniam, V., Sussman, J.L., Tompa, P., and Uversky, V.N. 2013. What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins* 1:e24157.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J* 272:5129-5148.

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., and Brown, C.J. 2000. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 11:161-171.

Dyson, H.J. and Wright, P.E. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6:197-208.

Eliezer, D. 2009. Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol* 19:23-30.

Fan, X. and Kurgan, L. 2014. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. *J Biomol Struct Dyn* 32:448-464.

Fontana, A., de Laureto, P.P., Spolaore, B., and Frare, E. 2012. Identifying disordered regions in proteins by limited proteolysis. *Methods Mol Biol* 896:297-318.

Fontana, A., de Laureto, P.P., Spolaore, B., Frare, E., Picotti, P., and Zambonin, M. 2004. Probing protein structure by limited proteolysis. *Acta Biochim Pol* 51:299-321.

Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S.D., Koike, R., Hiroaki, H., and Ota, M. 2014. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res* 42:D320-325.

Fukuchi, S., Sakamoto, S., Nobe, Y., Murakami, S.D., Amemiya, T., Hosoda, K., Koike, R., Hiroaki, H., and Ota, M. 2012. IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Research* 40:D507-D511.

Fuxreiter, M., Tompa, P., Simon, I., Uversky, V.N., Hansen, J.C., and Asturias, F.J. 2008. Malleable machines take shape in eukaryotic transcriptional regulation. *Nat Chem Biol* 4:728-737.

Fuxreiter, M., Toth-Petroczy, A., Kraut, D.A., Matouschek, A., Lim, R.Y., Xue, B., Kurgan, L., and Uversky, V.N. 2014. Disordered proteinaceous machines. *Chem Rev* 114:6806-6843.

Galea, C.A., Wang, Y., Sivakolundu, S.G., and Kriwacki, R.W. 2008. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. *Biochemistry* 47:7598-7609.

Hu, G., Wu, Z., Wang, K., Uversky, V.N., and Kurgan, L. 2015. Untapped potential of disordered proteins in current druggable human proteome. *Current drug targets*.

Ishida, T. and Kinoshita, K. 2007. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Research* 35:W460-W464.

Jensen, M.R., Salmon, L., Nodet, G., and Blackledge, M. 2010. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc* 132:1270-1272.

Jones, D.T. and Cozzetto, D. 2015. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31:857-863.

Jones, D.T. and Ward, J.J. 2003. Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Bioinformatics* 53:573-578.

Kozlowski, L.P. and Bujnicki, J.M. 2012. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics* 13:1-11.

Le Gall, T., Romero, P.R., Cortese, M.S., Uversky, V.N., and Dunker, A.K. 2007. Intrinsic disorder in the Protein Data Bank. *J Biomol Struct Dyn* 24:325-342.

Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z. 1999. Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform Ser Workshop Genome Inform* 10:30-40.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. 2003a. Protein Disorder Prediction: Implications for Structural Proteomics. *Structure* 11:1453-1459.

Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. 2003b. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research* 31:3701-3708.

Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N., and Dunker, A.K. 2006. Intrinsic disorder in transcription factors. *Biochemistry* 45:6873-6888.

Liu, J. and Rost, B. 2003. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Research* 31:3833-3835.

Malhis, N., Jacobson, M., and Gsponer, J. 2016. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res*.

Martin, A.J.M., Walsh, I., and Tosatto, S.C.E. 2010. MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics* 26:2916-2917.

McGuffin, L.J., Atkins, J.D., Salehe, B.R., Shuid, A.N., and Roche, D.B. 2015. IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Research* 43:W169-W173.

Meng, F. and Kurgan, L. 2016. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics* 32:i341-i350.

Mizianty, M.J., Peng, Z.L., and Kurgan, L. 2013. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disordered Proteins* 1:e24428.

Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., and Kurgan, L. 2010. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics* 26:i489-i496.

Mizianty, M.J., Uversky, V., and Kurgan, L. 2014. Prediction of intrinsic disorder in proteins using MFDp2. *Methods Mol Biol* 1137:147-162.

Mizianty, M.J., Zhang, T., Xue, B., Zhou, Y., Dunker, A.K., Uversky, V.N., and Kurgan, L. 2011. In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinformatics* 12:245.

Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. 2014. Assessment of protein disorder region predictions in CASP10. *Proteins* 82:127-137.

Nodet, G., Salmon, L., Ozenne, V., Meier, S., Jensen, M.R., and Blackledge, M. 2009. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J Am Chem Soc* 131:17908-17918.

Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztanyi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L., Dunker, A.K., and Gough, J. 2013. D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res* 41:D508-516.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., and Dunker, A.K. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins* 53 Suppl 6:566-572.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A.K. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 61 Suppl 7:176-182.

Peng, Z. and Kurgan, L. 2012a. On the complementarity of the consensus-based disorder prediction. *Pac Symp Biocomput* 176-187.

Peng, Z. and Kurgan, L. 2015. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res* 43:e121.

Peng, Z., Oldfield, C.J., Xue, B., Mizianty, M.J., Dunker, A.K., Kurgan, L., and Uversky, V.N. 2014. A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol Life Sci* 71:1477-1504.

Peng, Z., Wang, C., Uversky, A.V., and Kurgan, L. 2015a. Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol Biol* accepted.

Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., and Kurgan, L. 2015b. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci* 72:137-151.

Peng, Z.L. and Kurgan, L. 2012b. Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr Protein Pept Sci* 13:6-18.

Pentony, M., Ward, J., and Jones, D. 2010. Computational Resources for the Prediction and Analysis of Native Disorder in Proteins. *In* Proteome Bioinformatics, vol. 604 (S.J. Hubbard and A.R. Jones, eds.) pp. 369-393. Humana Press.

Potenza, E., Domenico, T.D., Walsh, I., and Tosatto, S.C.E. 2015. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Research* 43:D315-D320.

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., and Dunker, A.K. 2004. Protein flexibility and intrinsic disorder. *Protein Sci* 13:71-80.

Receveur-Brechot, V., Bourhis, J.M., Uversky, V.N., Canard, B., and Longhi, S. 2006. Assessing protein disorder and induced folding. *Proteins* 62:24-45.

Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., and Dunker, A.K. 1997. Identifying disordered regions in proteins from amino acid sequence. Neural Networks,1997., International Conference on, 9-12 Jun 1997.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. 2001. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics* 42:38-48.

Salmon, L., Nodet, G., Ozenne, V., Yin, G., Jensen, M.R., Zweckstetter, M., and Blackledge, M. 2010. NMR characterization of long-range order in intrinsically disordered proteins. *J Am Chem Soc* 132:8407-8418.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z., and Dunker, A.K. 2007. DisProt: the Database of Disordered Proteins. *Nucleic Acids Research* 35:D786-D793.

Smith, D.L., Deng, Y., and Zhang, Z. 1997. Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry. *J Mass Spectrom* 32:135-146.

Tompa, P. 2002. Intrinsically unstructured proteins. *Trends in Biochemical Sciences* 27:527-533.

Uversky, V.N. 2015. Biophysical Methods to Investigate Intrinsically Disordered Proteins: Avoiding an "Elephant and Blind Men" Situation. *Adv Exp Med Biol* 870:215-260.

Uversky, V.N. and Dunker, A.K. 2012a. Intrinsically Disordered Protein Analysis: Volume I. Methods and Experimental Tools, vol. 895. Humana Press, Totowa, NJ, USA.

Uversky, V.N. and Dunker, A.K. 2012b. Intrinsically Disordered Protein Analysis: Volume II. Methods and Experimental Tools. Humana Press, Totowa, NJ, USA.

Uversky, V.N. and Dunker, A.K. 2012c. Multiparametric analysis of intrinsically disordered proteins: looking at intrinsic disorder through compound eyes. *Anal Chem* 84:2096-2104.

Uversky, V.N. and Longhi, S. 2010. Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure and Conformation. John Wiley & Sons, New Jersey, USA.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. 2005. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 18:343-384.

Uversky, V.N., Oldfield, C.J., and Dunker, A.K. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 37:215-246.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., Kim, P.M., Kriwacki, R.W., Oldfield, C.J., Pappu, R.V., Tompa, P., Uversky, V.N., Wright, P.E., and Babu, M.M. 2014. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews* 114:6589-6631.

Varadi, M., Zsolyomi, F., Guharoy, M., and Tompa, P. 2015. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One* 10:e0139731.

Wang, C., Uversky, V.N., and Kurgan, L. 2016. Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* 16:1486-1498.

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., and Jones, D.T. 2004a. The DISOPRED server for the prediction of protein disorder. *Bioinformatics* 20:2138-2139.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004b. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337:635-645.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., and Obradovic, Z. 2007. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 6:1882-1898.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., and Uversky, V.N. 2010. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta* 1804:996-1010.

Xue, B., Dunker, A.K., and Uversky, V.N. 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* 30:137-149.

Yan, J., Dunker, A.K., Uversky, V.N., and Kurgan, L. 2015. Molecular Recognition Features (MoRFs) in three domains of life. *Mol Biosyst*.

Key References

van der Lee et al. 2014
Defines intrinsic disorder and discusses the relevant experimental and computational tools

Ishida and Kinoshita, 2007
Describes PrDOS, one of the most accurate hybrid method for the prediction of intrinsic disorder

Jones and Cozzetto, 2015
Describes DISOPRED3, one of the most accurate machine learning method for the prediction of intrinsic disorder and disordered protein binding regions

Mizianty et al. 2010
Describes MFDp, one of the most accurate meta method for the prediction of intrinsic disorder

Peng and Kurgan 2012b
Provides comprehensive empirical assessment of predictive performance of modern methods for the prediction of intrinsic disorder

Sickmeier et al. 2007
Introduces and describes the DisProt database of the intrinsically disordered proteins

Internet Resources

http://d2p2.pro/
D^2P^2 database

http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1
DISOPRED3's webserver

http://biomine-ws.ece.ualberta.ca/MFDp
MFDp's webserver

http://mobidb.bio.unipd.it/
MobiDB database

http://prdos.hgc.jp/cgi-bin/top.cgi
PrDOS's webserver