Accuracy of protein-level disorder predictions

Akila Katuwawala[1], Christopher Oldfield[1], and Lukasz Kurgan[1]*
[1]Department of Computer Science, Virginia Commonwealth University

*corresponding author:
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

Experimental annotations of intrinsic disorder are available for 0.1% of 147,000,000 of currently sequenced proteins. Over 60 sequence-based disorder predictors were developed to help bridge this gap. Current benchmarks of these methods assess predictive performance on datasets of proteins, however predictions are often interpreted for individual proteins. We demonstrate that the protein-level predictive performance varies substantially from the dataset-level benchmarks. Thus, we perform first-of-its-kind protein-level assessment for 13 popular disorder predictors using 6,200 disorder-annotated proteins. We show that the protein-level distributions are substantially skewed toward high predictive quality while having long tails of poor predictions. Consequently, between 57% and 75% proteins secure higher predictive performance than the currently used dataset-level assessment suggests, but as many as 30% of proteins that are located in the long tails suffer low predictive performance. These proteins typically have relatively high amounts of disorder, in contrast to the mostly structured proteins that are predicted accurately by all 13 methods. Interestingly, each predictor provides the most accurate results for some number of proteins while the best-performing at the dataset-level method is in fact the best for only about 30% of proteins. Moreover, the majority of proteins are predicted more accurately than the dataset-level performance of the most accurate tool by at least four disorder predictors. While these results suggests that disorder predictors outperform their current benchmark performance for the majority of proteins and that they complement each other, novel tools that accurately identify the hard-to-predict proteins and that make accurate predictions for these proteins are needed.

Keywords: Intrinsic disorder; Intrinsically disordered proteins; Intrinsically disordered regions; Prediction; Protein sequence; Disorder content; Accuracy; Predictive performance.

Akila Katuwalala is a Ph.D. student in the Department of Computer Science at the Virginia Commonwealth University. His research interests are in the computational prediction and characterization of structural and functional properties of proteins.

Christopher Oldfield is a Postdoctoral Fellow in the Department of Computer Science at the Virginia Commonwealth University. His research focuses on sequence, structure, and functional analysis of intrinsically disordered proteins.

Lukasz Kurgan is a Robert J. Mattauch Endowed Professor of Computer Science at the Virginia Commonwealth University. He is a Fellow of AIMBE and has published close to 150 peer-reviewed journal articles that focus on structural and functional characterization of proteins and small RNAs. More details about his research group are available at http://biomine.cs.vcu.edu/.

1 Introduction

Intrinsically disordered regions form ensembles of conformations and they lack a stable tertiary structure in isolation [1-4]. Several large-scale computational studies estimate that about 19% of residues in eukaryotic proteins are disordered [5], and that depending on the specific eukaryotic organism between 30% and 50% of proteins (44% in human [6]) have at least one long disordered region with 30 or more consecutive amino acids [5, 7-9]. Intrinsic disorder is also a driving contributor to the dark proteome [10-12]. Disordered regions perform their functions while remaining disordered or via disorder-to-order transitions that occur upon binding to their physiological partner(s) [13-16]. Interaction with partner molecules is a common function of intrinsically disordered regions; they have been found to interact with DNA, RNA, other proteins, and small molecules [16]. Proteins with disordered regions are crucial for a wide range of cellular roles, including molecular assembly and recognition, signal transduction, cell cycle and cell death regulation, transcription, translation, and viral cycle [13, 15-37]. Due to their central role in cellular regulation, these proteins are associated with several human diseases [35, 38-40] and attract interest as potent drug targets [41-45]. Mechanistic descriptions of intrinsically disordered protein function are challenging since function arises from structural ensembles, rather than stable three-dimensional structure, but computational studies of the dynamics and structural ensembles of intrinsically disordered proteins shown promise in explaining functional mechanisms of these proteins [46-48].

Experimentally annotated disordered regions can be obtained from several databases: DisProt [49], MobiDB [50], IDEAL [51], and Protein Data Bank (PDB) [52-54], where they correspond to the regions with missing coordinates in crystal structures and the highly structurally varied regions in the NMR structures. However, these annotations cover only a very small fraction of proteins sequences in nature. DisProt and IDEAL covers only 803 [49] and 913 proteins [51], respectively, while a recent study estimates that about 26 thousand proteins with the disordered regions can be obtained from PDB [55]. The combined collection of these datasets constitutes only about 0.02% of the 147.4 million of unique proteins from the UniProt resource [56] (as of April 2019). This huge and growing annotation gap motivates the development of computational methods that accurately predict disorder in protein sequences. These methods rely on the fact that the amino acid composition and conservation of disordered regions are different from the structured regions [57-61].

Over 60 disorder predictors were developed over the last four decades [62-68]. They are divided into three broad categories [63-65, 67]: 1) *ab-initio* models; 2) methods with predictive models that were produced using machine learning algorithms; and 3) meta-predictors. The predictive models in the first category are derived from biophysical principles that are known to differentiate between the intrinsically disordered and ordered regions. Representative methods in this group are NORSP [69], GlobPlot [70] and IUPred [71-73]. The computational tools in the second category utilize predictive models that are produced by machine learning algorithms to maximize predictive performance on disorder-annotated training datasets. This category has a large collection of methods with several illustrative examples that include RONN [74], DisEMBL [75], DISpro [76, 77], DISOPRED [78, 79], VSL2B [80, 81], SPINE-D [82], DeepCNF-D [83], SPOT-Disorder [84], PrDOS [85], and SPOT-Disorder-Single [86]. The meta-predictors combine outputs produced by several disorder predictors with the underlying goal to improve predictive performance when compared to the results produced by their input single predictors [87, 88]. Example meta-predictors include MD [89], MetaDisorder [90], disCoP [88], DisMeta [91], PONDR-FIT [92], CSpritz [93], MFDp [94-96], ESpritz [97], metaPrDOS [98], MFDp2 [99], DISOPRED3 [100], and MobiDB-lite [101]. We note that pre-computed disorder predictions can be conveniently obtained from two databases: MobiDB [50] and D^2P^2 [102].

The predictive quality of disorder predictors was comparatively assessed in several studies [55, 103-112]. These studies compare predictive performance for selected sets of disorder predictors on various benchmark datasets that range in size from about a hundred to over 20 thousand proteins. Six assessments on smaller-sized datasets with about 100 proteins were included as part of the Critical Assessment of protein Structure Prediction (CASP) experiments between 2002 (CASP5) and 2012 (CASP10) [104-109]. Three more recent assessments include one that relies on the small dataset from the CASP10 experiment [112], another that uses a slightly larger set of 250 proteins collected from the DisProt resource [110], and the third that utilizes close to 500 proteins collected from PDB and DisProt [103]. One of the largest evaluations covers 13 predictors that were assessed on a large benchmark dataset of over 25 thousand proteins [55]. Moreover, a collection of 13 disorder predictors was assessed on a dataset of about 350 membrane proteins [111]. A recently published survey comprehensively summarizes results of multiple past assessments, but without collecting new results [67].

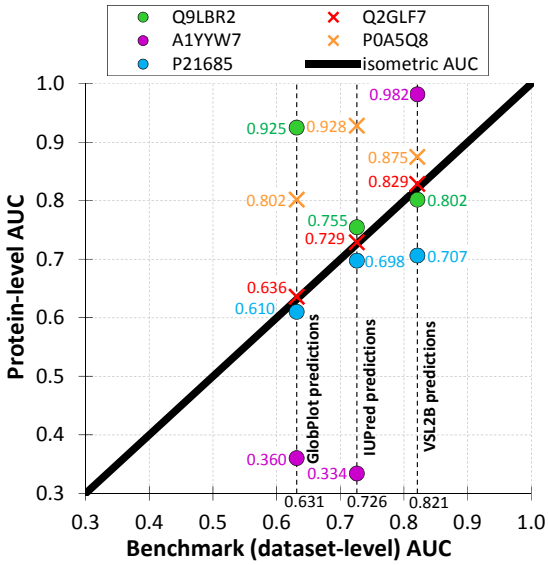

Figure 1. Comparison of the benchmark (dataset-level) and protein-level predictive performance measured with AUC for three disorder predictors (GlobPlot, IUPred, and VSL2B) and five color-coded proteins: non-toxic nonhemagglutinin type D (green marker; UniProt Q9LBR2), guanylate kinase (red marker; UniProt Q2GLF7), alkaline phosphatase (violet marker; UniProt A1YYW7), hydroxymethyltransferase (orange marker; UniProt P0A5Q8), and phytoene desaturase (blue marker; UniProt P21685). The benchmark AUC values are shown on the *x*-axis while the protein-level AUC values are color-coded, shown on the *y*-axis, and their values are given next to the corresponding markers. The black isometric AUC line shows equivalent dataset-level and protein-level values.

The results of the abovementioned comparative studies, which are comprehensively surveyed in [67], can be used to rank and compare dataset-level predictive performance of various disorder predictors. However, the past assessments fall short of addressing evaluation at a single protein level. This is an important shortcoming since the disorder predictors are used arguably more often to characterize intrinsic disorder for individual proteins than for datasets of hundreds of proteins. For instance, our MFDp predictor [94-96, 99] was recently used to predict disordered regions for the flagellar capping protein [113], Cia2 [114], SpSM30B [115], AP24 [116], and BRCA1 [117], to name just a few examples. Figure 1 illustrates some of the differences underlying the benchmark dataset-level vs. protein-level assessments. It compares the protein-level vs. the dataset-level performance for five proteins and three methods that were assessed as part of the recent large-scale evaluation [55]. The three predictors include one of the least accurate methods, GlobPlot, modestly accurate IUPred, and the most accurate VSL2B (based on their benchmark dataset-level performance). The

predictive performance is quantified with the popular area under the ROC curve (AUC) measure [67], which ranges between 0.5 (equivalent to random predictions) and 1 (always correct predictions). The dataset-level AUCs for GlobPlot, IUPred and VSL2B that were estimated in [55] equal 0.631, 0.726 and 0.821, respectively. These results suggest that the three tools perform at substantially different levels of predictive quality. The results for a couple of proteins, such as guanylate kinase (red marker in Figure 1; UniProt Q2GLF7) and phytoene desaturase (blue marker; UniProt P21685) are in relatively close agreement with the datasets-level AUCs. However, results for some other proteins could be very different from the dataset-level estimates. For instance, predictions generated by GloPlot for nonhemagglutinin type D (green marker in Figure 1; UniProt Q9LBR2) are substantially better than the corresponding dataset-level performance and much better than the results from the two other predictors, which also contradicts the ranking based on the datasets-level AUC. Similarly, by far the most accurate predictions for hydroxymethyltransferase (orange marker in Figure 1; UniProt P0A5Q8) are produced by IUPred, while the dataset-level assessment in [55] shows that VSL2B outperforms IUPred. Moreover, GlobPlot's results for this proteins are again much better than the same benchmark suggests. On the other hand, GlobPlot's and IUPred's predictions for alkaline phosphatase (violet marker; UniProt A1YYW7) have poor quality, much lower than the dataset-level benchmark suggests, while the predictions from VSL2B substantially outperforms its dataset-level AUC.

The above analysis demonstrates that the protein-level performance may vary widely from the dataset-level results that were produced by the prior assessments. While past studies have focused on comparing the overall predictive performance of various disorder predictors based on the dataset-level assessments [55, 103-112], we investigate the protein-level predictive performance to provide practical insights. Our study tests a set of 13 diverse disorder predictors on a large dataset of 6271 proteins. Our aims are to quantify the spread of the protein-level performance for different methods, compare their protein-level predictive quality, measure divergence of these predictions from the dataset-level assessment, investigate characteristics of hard vs. easy to predict proteins, and study differences and similarities between protein-level results of different disorder predictors.

2 Materials and Methods

2.1 Benchmark dataset

The evaluation relies on a large benchmark dataset with native annotation of disorder that was originally published in [55]. The original dataset includes 25,717 proteins that were extracted from the MobiDB resource [50]. We improved this dataset by removing sequences with unknown/undetermined amino acid (AA) types, which is needed to secure disorder predictions, and by reducing within-dataset redundancy. We use BLASTCLUST [118] to reduce pairwise sequence similarity to 25% in order to minimize the redundancy. The resulting dataset has 6,271 protein sequences that share < 25% similarity and that include 105,709 disordered and 1,672,907 structured residues. This dataset was recently used in [119, 120] and is available at http://biomine.cs.vcu.edu/servers/QUARTER/. Analysis in [120] shows that the predictive performance of the disorder predictors on the original dataset of 25,717 proteins that was assessed in [55] is similar and consistent when compared with the evaluation based on the improved benchmark dataset with 6,271 proteins.

2.2 Selection of the disorder predictors

We cover a set of 13 popular, publically available and diverse disorder predictors. Ten predictors were selected from among the 13 methods that were included in the recent large-scale assessment [55]. They include three versions of ESpritz that predict intrinsic disorder annotated from X-ray

structures (ESpritz-Xray), NMR structures (ESpritz-NMR) and using DisProt database (ESpritz-DisProt) [97]; two versions of IUPred that are optimized to predict short (IUPred-short) and long (IUPred-long) disordered regions [71-73]; two versions of DisEMBL that were designed based on X-ray structures (DisEMBL-465) and propensity for loop conformations (DisEMBL-HL) [75]; GlobPlot [70]; RONN [74]; and VSL2B [80]. We excluded three methods that were originally covered in [55]: SEG [121], Pfilt [122] and FoldIndex [123]. These are older predictors that have secured the lowest dataset-level predictive quality in that assessment. We expanded the list of methods covered in [55] by adding three high-quality predictors: the top performing method from CASP10, DISOPRED3 [100], and two recently published deep learning-based methods: SPOT-Disorder [84] and DeepCNF-D [83]. We used the computationally tractable version of DeepCNF-D, DeepCNF-D(ami_only), which relies on the sequence only-derived inputs. Table 1 provides websites where these 13 selected methods can be found and summarizes key characteristics of these predictors including the year of publication, type of the predictive model that they apply, and their citation data. These tools were published between 2002 and 2016 and most use neural network-based predictive models. They are well-cited, with the annual number of citations ranging between 5 and 50. The selected predictors uniformly cover the three categories of methods including *ab-initio* tools (IUPred-short, IUPred-long and GlobPlot), machine learning-based predictors (RONN, DisEMBL-HL, DisEMBL-465, VSL2B, DeepCNF-D and SPOT-Disorder) and meta-predictors (DISOPRED3, ESpritz-Xray, ESpritz-NMR and ESpritz-DisProt). They were designed to address prediction of all major types of disorder annotations including annotations that rely on X-ray crystal structures, NMR structures and a variety of other experimental methods that are covered in the DisProt resource.

2.3 Assessment of predictive performance

Computational disorder predictors output putative propensity for intrinsic disorder and binary disorder prediction for every amino acid in the input protein sequence. This propensity is usually 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. The binary prediction categorizes each amino acid as either structured or disordered. This prediction is typically generated from the propensity scores, i.e., residues with scores > predictor-specific threshold are categorized as disordered and the remaining amino acids are categorized as structured.

We assess predictive quality for both types of outputs. The one measure that was used across all recent dataset-level assessments to quantify the predictive quality for the putative propensities is the AUC [55, 67, 103, 104, 110-112]. Calculation of AUC requires presence of both disordered and structured residues; they are needed to compute the true positive and false positive rates that make up the ROC curve. Consequently, we compute the protein-level AUCs for 1677 proteins from our dataset that have sufficient number of both structured and disordered residues (min. 20 residues of each type); this ensures that the protein-level AUCs can be estimated with high precision. The binary predictions are assessed with several measures that include accuracy, F1, sensitivity, specificity = 1 – false positive rate (FPR), and Matthews's correlation coefficient (MCC) [55, 67, 103, 104, 110-112]. The protein-level calculations of F1, sensitivity, specificity and MCC require presence of both disordered and structured residues in a given protein chain. The only two measures that be applied to all proteins, irrespective of the amount of their native disorder content, are accuracy (i.e., rate of correct predictions) and FPR (i.e., rate of incorrectly predicted disordered residues). Thus, we use accuracy and FPR to assess the protein-level binary predictions of intrinsic disorder.

Table 1. Summary of the 13 predictors that used on this comparative review. The number of citations was collected from the Google Scholar as of June 19th, 2019. The methods are sorted by the ascending order of year when they were published.

Predictor	Year Published	Ref.	Model Type	Number of Citations	Annual number of citations	Website
ESpritz-DisProt	2002	[97]	Bi-directional recursive neural network	388	22.8	http://protein.bio.unipd.it/espritz/
ESpritz-NMR						
ESpritz-Xray						
disEMBL-465	2003	[75]	Ensemble of feed-forward neural networks	802	50.1	http://dis.embl.de/
disEMBL-HL						
GlobPlot	2003	[70]	Derivative based curve optimization	657	41.1	http://globplot.embl.de/
JRONN	2005	[74]	Radial basis functional neural network	454	32.4	http://www.strubi.ox.ac.uk/RONN
VSL2B	2006	[80]	Support vector machine	395	30.4	http://www.dabi.temple.edu/disprot/predictor.php
IUPred long	2009	[71-73]	Scoring function derived using optimized energy minimization algorithm	311	31.1	https://iupred2a.elte.hu/
IUPred short						
DISOPRED3	2015	[100]	Ensemble of neural network, support vector machine and nearest neighbor models	165	41.3	http://bioinf.cs.ucl.ac.uk/psipred/
DeepCNF	2015	[83]	Deep convolutional neural network	20	5.0	https://ttic.uchicago.edu/~wangsheng/software.html
SPOT-DISORDER	2016	[84]	Deep bidirectional neural network	39	13.0	http://sparks-lab.org/server/SPOT-disorder/

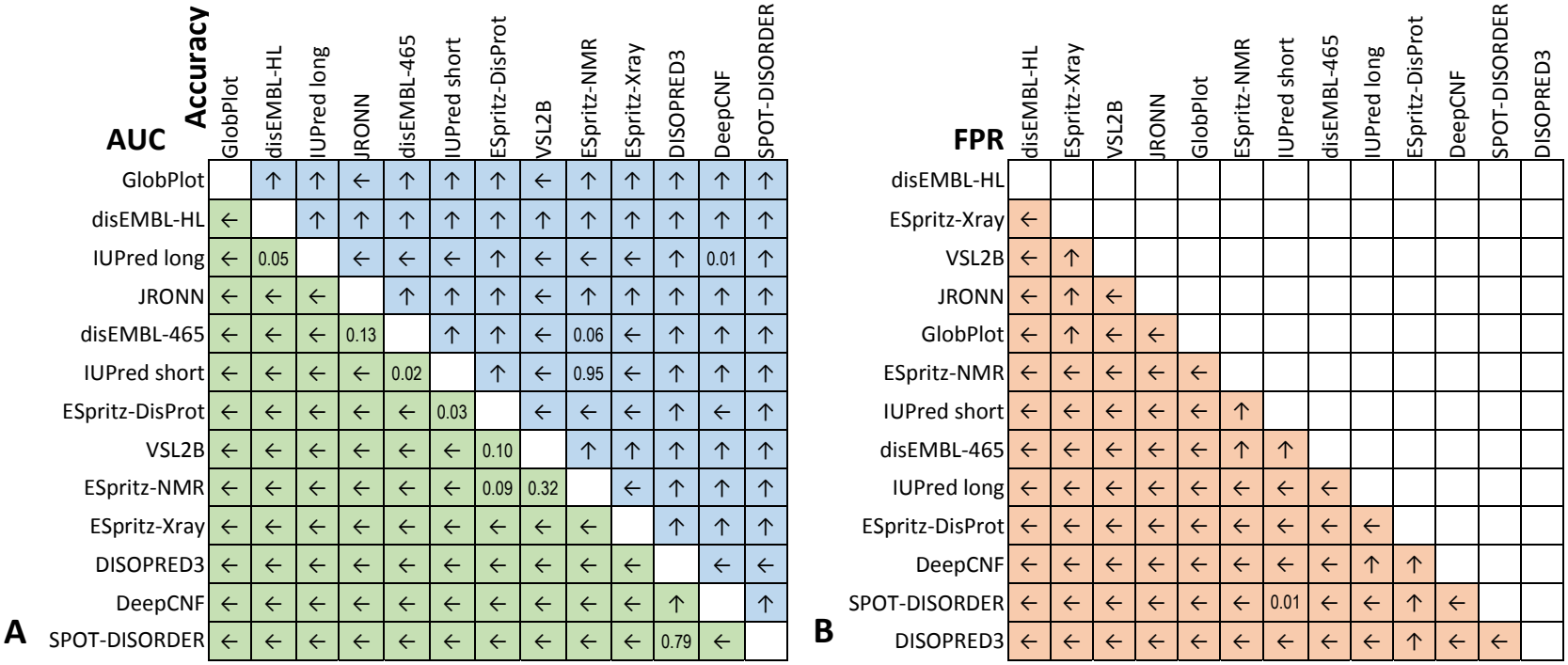

Figure 2. Comparison of the protein-level predictive performance between the 13 disorder predictors. Panel A summarizes comparison of the AUC values (on green background) and accuracies (on blue background). Panel B considers the false positive rates (on red background). Statistical significance of the differences between all pairs of methods was assessed with the *t*-test for normal measures and otherwise with the Wilcoxon rank-sum test. Normality was tested with the Anderson-Darling test at 0.05 significance. We assume that the difference in predictive performance for a given pair of predictors is significant if the corresponding *p*-value < 0.01. Arrows point to the methods that secure significantly better predictive performance (*p*-value < 0.01). The *p*-values are shown for the pairs of methods that are not significantly different.

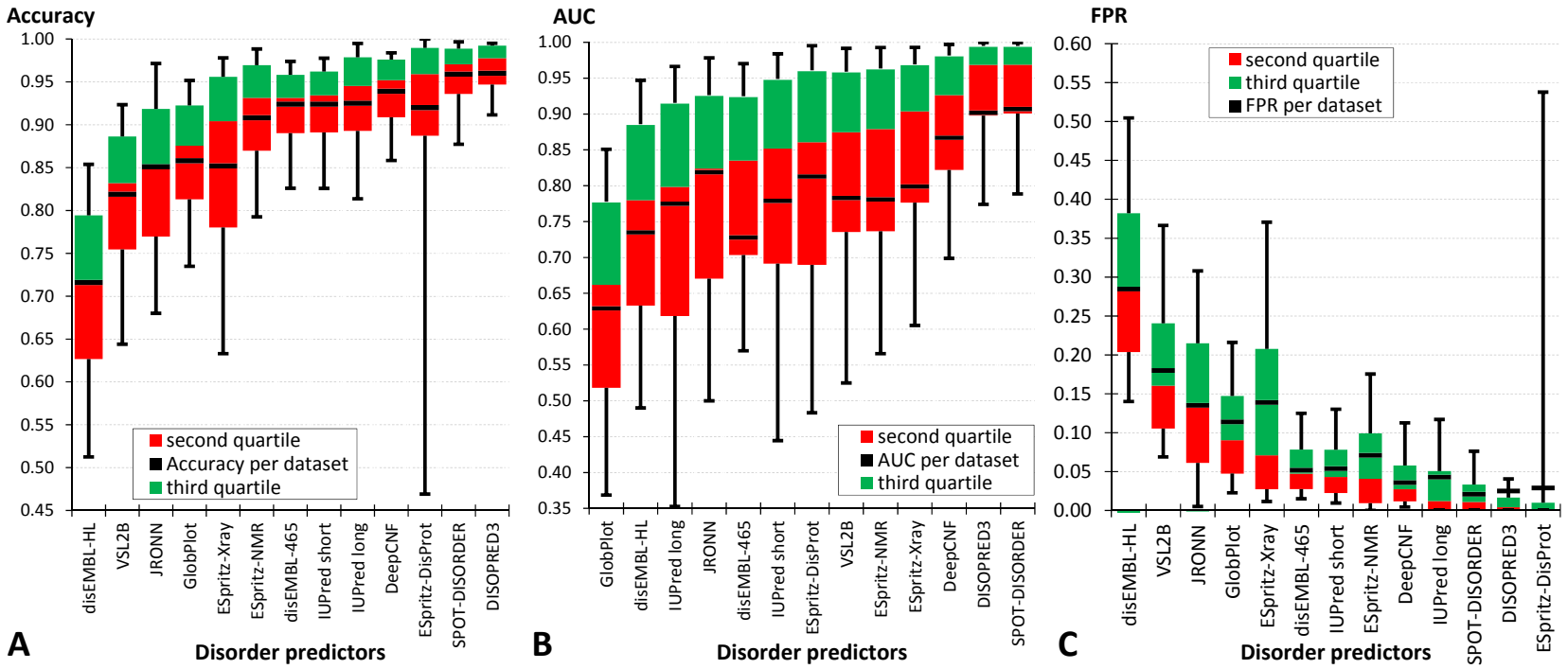

Figure 3. Distributions of the protein-level predictive quality measured with accuracy (panel A), AUC (panel B) and false positive rate (panel C) for the 13 disorder predictors. Box plots show the second quartile (in red), median (between red and green boxes), and third quartile (in green) for the distribution of the protein-level values. The whiskers denote the corresponding 10th and 90th percentiles. The black horizontal lines show the benchmark dataset-level performance. The predictors are sorted by their median values of the predictive performance.

Table 2. Dataset- and protein-level predictive quality for the 13 considered disorder predictors. The dataset-level accuracy and AUC are compared against the previously published results. We note that false positive rates are typically not reported in the past studies. Protein-level results are summarized with the median value. The methods are sorted by their dataset-level AUC on our benchmark dataset.

Disorder predictor	Accuracy (binary predictions)				AUC (putative propensities)				FPR (binary predictions)	
	Protein-level median	Dataset-level	Previously reported dataset-level	Difference dataset-level	Protein-level median	Dataset-level	Previously reported dataset-level	Difference dataset-level	Protein-level median	Dataset-level
GlobPlot	0.876	0.855	0.847	0.8%	0.662	0.626	0.631	0.5%	0.090	0.111
disEMBL-HL	0.715	0.713	0.721	0.8%	0.780	0.725	0.727	0.2%	0.282	0.277
IUPred-long	0.945	0.922	0.921	0.1%	0.798	0.732	0.726	0.6%	0.012	0.040
JRONN	0.848	0.847	0.839	0.8%	0.824	0.772	0.759	1.3%	0.132	0.131
ESpritz-NMR	0.931	0.905	0.903	0.2%	0.879	0.776	0.770	0.6%	0.041	0.068
IUPred-short	0.934	0.921	0.924	0.3%	0.852	0.778	0.778	0.0%	0.043	0.051
disEMBL-465	0.931	0.921	0.925	0.4%	0.835	0.780	0.787	0.7%	0.047	0.049
ESpritz-Xray	0.904	0.849	0.840	0.9%	0.904	0.796	0.778	1.8%	0.071	0.136
VSL2B	0.832	0.816	0.805	1.1%	0.874	0.810	0.821	1.1%	0.161	0.177
ESpritz-DisProt	0.959	0.917	0.934	1.7%	0.861	0.816	0.791	2.5%	0.000	0.034
DeepCNF	0.952	0.936	0.944	0.8%	0.926	0.871	0.898	2.7%	0.027	0.033
DISOPRED3	0.977	0.957	0.955	0.2%	0.969	0.899	0.897	0.2%	0.004	0.016
SPOT-Disorder	0.971	0.956	0.950	0.6%	0.969	0.904	0.891	1.3%	0.011	0.018

3 Predictive performance of disorder predictions

3.1 Dataset-level performance

Table 2 compares the dataset-level results on our benchmark dataset against the previously published results. The prior results for DISOPRED3 are taken from CASP10 [104], for SPOT-Disorder from [84], for DeepCNF from [83], and for the other 10 methods from [55]. The two "Difference" columns show the magnitude of the differences between the previously reported results and the assessment on our dataset for the accuracy and AUC measures. The average differences (across the 13 predictors) in accuracy and AUC are 0.67% and 1.04%, respectively. This demonstrates that our dataset closely reflects the current state of the per-dataset assessments. Our results are also in agreement with [67, 83, 84], showing that DeepCNF, SPOT-Disorder and DISOPRED3 outperform the other disorder predictors. We conclude that these results validate reliability of our subsequent per-protein analysis which is performed in the context of the dataset-level results.

3.2 Protein-level performance

Figure 2 compares the protein-level predictive performance between the 13 disorder predictors. We assess statistical significance of the differences in predictive performance between all pairs of methods. We assume that the differences are significant if the corresponding p-value < 0.01. This comparison reveals that most of the differences are significant, which suggests that improvements between methods are largely consistent across different proteins. When using AUC for the assessments, the top performing SPOT-Disorder (see Table 2) is similar to DISOPRED3 and significantly outperforms the other 11 methods. DISOPRED3 has the highest accuracy (Table 2) and it significantly improves over the other 12 predictors based on this measure. Finally, ESpritz-DisProt that has the lowest median protein-level FPR (Table 2) provides predictions with significantly lower FPR values when contrasted with each of the 12 other predictors. The main reason why this method secures such low FPR is that it under-predicts the amount of disorder. This was shown in [119] where ESpritz-DisProt predicts 2.6% disorder content in a large dataset with 5% native disorder content, while the other 9 methods considered in that article predict between 6% (IUpred-long) and 29% (DisEMBL-HL) disorder content.

Distributions of the protein-level predictive performance for the 13 disorder predictors are shown in the Supplementary Figure S1. The distributions for all predictors for accuracy and AUC are left-skewed with long tails. The corresponding distributions of the FPR values are right-skewed with similarly long tails; this is because larger values of FPR indicate lower predictive quality. These distributions demonstrate that while majority of the proteins are predicted with above average predictive performance, minority of proteins that are located in the long tails are predicted with relatively low performance. The distributions are summarized and compared to the dataset-level results in Figure 3A for accuracy, Figure 3B for AUC, and Figure 3C for FPR. The box plots show the first quartile (in red), second quartile (median; where red and green meet), and third quartile (in green), with whiskers that denote the 10^{th} and 90^{th} percentiles. The long tails are represented by the long bottom whiskers for accuracy and AUC (long top whiskers for FPR) when compared to the corresponding top whiskers (bottom whiskers for FPR). The dataset-level values are denoted by the black horizontal lines. Figure 3 reveals that majority of the proteins secure higher levels of predictive performance than their corresponding dataset-level assessment suggests. In other words, protein-level medians for accuracy and AUC are consistently higher or at worst similar to the dataset-level values, while the protein-level medians for FPR are consistently lower or at worst similar when compared to the dataset-level values. These values can be directly compared in Table 2. We emphasize that this trend is consistent across all disorder predictors and the three measures of predictive performance.

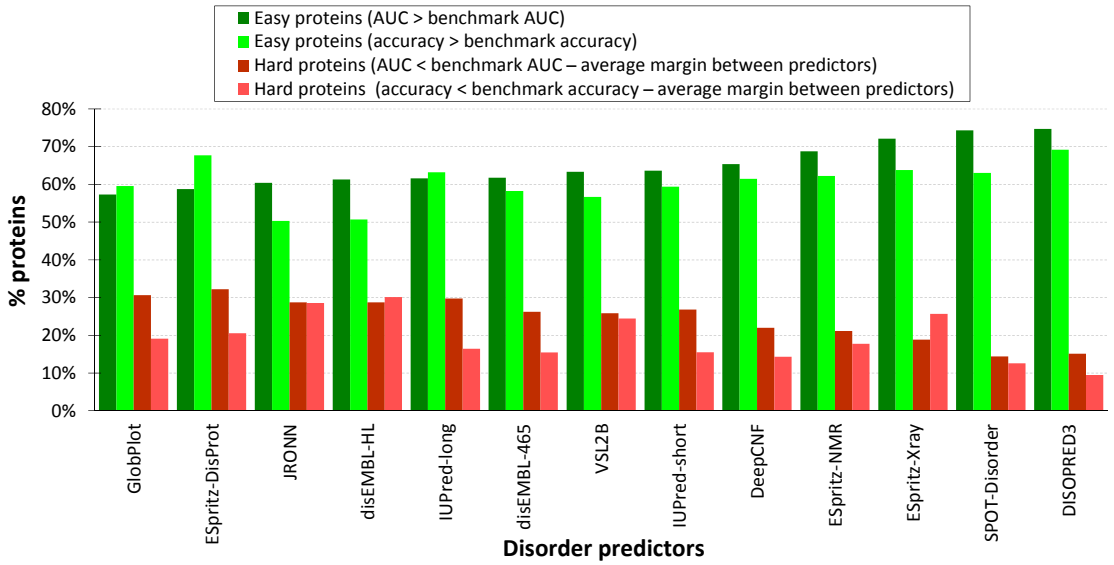

Figure 4. Analysis of the easy and hard to predict proteins for the 13 disorder predictors. The easy proteins are predicted with higher than expected accuracy or AUC, i.e., their protein-level accuracy (AUC) > dataset-level accuracy (AUC). The hard proteins are predicted with relatively low accuracy or AUC, i.e., their protein-levels accuracy (AUC) < (dataset-level accuracy (AUC) – average margin of difference between disorder predictors). Bars represent the fraction of the easy proteins (green bars) and the hard proteins (red bars) when predictive performance is quantified with AUC (dark shade) and accuracy (light shade). Predictors are sorted by fraction of the easy proteins quantified with AUC (dark green bars).

Green bars in Figure 4 quantify the number of "easy"-to-predict proteins, defined as proteins for which the protein-level predictive quality is higher than the corresponding expected value, i.e., the dataset-level performance. When considering AUC, between 57% of proteins (for GlobPlot) and 75% of proteins (for DISOPRED3) secure predictive performance that is higher than the expected value. Similarly, between 50% (for JRONN) and 69% (for DISOPRED3) of proteins have better than expected accuracy. This means that the users should expect that the predictive quality of binary predictions and propensities is better than the current dataset-level benchmarks estimate for majority of the proteins, irrespective of the predictor that they apply.

However, the above good news come at a cost. The distributions have long tails (Figure 3), which means that the predictive performance drops to low levels for some number of proteins. Red bars in Figure 4 quantify the abundance of these "hard"-to-predict proteins. We define the hard proteins as those for which the protein-level predictive performance is lower than the corresponding dataset-level performance minus an average margin of difference between the considered 13 predictors. In other words, these proteins are predicted with accuracy/AUC that is lower than the expected (dataset-level) value by as much as the average difference to the dataset-level performance of the next worse predictor. The value of the margin (average difference in the dataset-level performance between all pairs of the considered 13 predictors) equals 0.067 and 0.071 for accuracy and AUC, respectively. Figure 4 shows that between 14% of proteins (for SPOT-Disorder) and 32% of proteins (for ESpritz-DisProt) are hard-to-predict with respect to their AUCs. Similarly, our analysis reveals that between 9% (for DISOPRED3) and 30% (for disEMBL-HL) of proteins have low accuracy. This means that the users should expect low quality protein-level predictions for anywhere between 10% (for accurate predictors like DISOPRED3 and SPOT-Disorder) and 30% (for less accurate predictors like JRONN and disEMBL-HL) of proteins that they submit.

Supplementary Figure S2 provides a more detailed analysis of the differences between the protein-level and the dataset-level performance for each of the 13 disorder predictors. It shows that accuracy for half of the proteins differs from their expected (dataset-level) value by over 0.03 for the best performing SPOT-Disorder and DISOPRED3. This difference becomes as high as over 0.10 for ESpritz-Xray. Similarly, AUC for half of the proteins differs from the dataset-level value by more than 0.09 for the best performing SPOT-Disorder and DISOPRED3, and by as much as over 0.18 for IUPred-long. The average (across the 13 predictors) difference from the expected value for half of the proteins exceed 0.06 in accuracy and 0.13 in AUC, which is comparable or higher than the average margin of difference in the dataset-level performance between the disorder predictors. Overall, we conclude that the dataset-level assessment does not provide a reliable estimate of the expected protein-level performance.

3.3 Hard to predict proteins have high levels of intrinsic disorder

Several studies have observed that disorder predictors, and especially the more accurate methods, provide substantially weaker predictive performance for dataset of proteins that have long disordered regions [55, 84, 104, 105]. This does not imply a relation between the chain length and predictive performance but rather that proteins with long disordered regions (that have large amount of disordered residues) are harder to predict accurately. We tested this assertion by comparing the Pearson correlation coefficients (PCCs) between the sequence length and accuracy vs. PCCs between the native disorder content (% of disordered residues) and accuracy. The former PCCs computed over the 13 disorder predictors are low and range between 0 and 0.37, with the average of 0.12. In contrast, the PCCs between disorder content and accuracy are relatively high and vary between -0.14 and -0.75, with the average of -0.42. The negative sign implies that, as expected, proteins with more disorder are more difficult to predict with high accuracy.

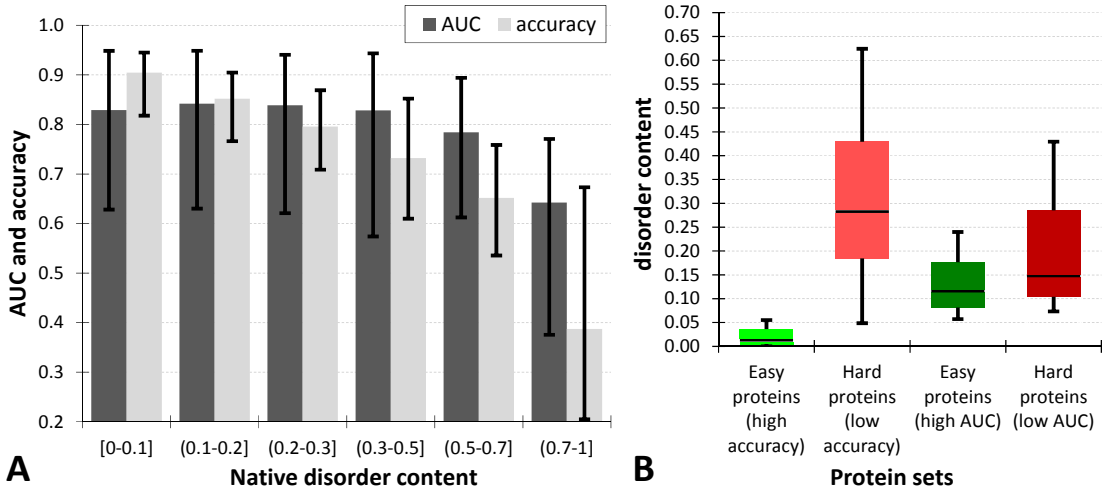

Figure 5. Relation between the protein-level predictive performance and the native disorder content. Panel A shows medians of the average (over the 13 predictors) accuracy and AUC for proteins grouped by their native disorder content, defined as the fraction of disordered residues in the sequence. The whiskers give the 10th and 90th percentiles of these averages. Panel B gives the distribution of the disorder content for the easy and hard proteins that are in common across the 13 predictors. The box plots show the second quartile, median (black horizontal line), and third quartile for the distribution of the protein-level disorder content values. The whiskers denote the corresponding 10th and 90th percentiles.

The above observation prompted a more detailed analysis of the relation between the native disorder content and the protein-level predictive performance, see Figure 5A. The figure show the relation between the average predictive performance (over the 13 methods) and the native disorder content that is binned into six ranges. While the predictive performance measured with both

accuracy and AUC remains relatively high for the structured and modestly disordered proteins (disorder content < 0.5), Figure 5A demonstrates that it substantially drops for the mostly disordered proteins and in particular for the proteins with the native disorder content > 0.7. The median values of the average protein-level accuracy and AUC equal only 0.387 and 0.642, respectively, for the latter set of proteins. The whiskers, which denote the 10^{th} and 90^{th} percentiles reveal that many proteins with over 0.7 disorder content secure near random levels of accuracy and AUC. Figure 5B flipsides the analysis. It shows the distribution of the disorder content for the hard vs. easy protein sets that were defined in Section 3.2. The easy proteins that secure above-expected predictive performance (green box plots) have low amounts of disorder while the hard proteins that obtain substantially below expected predictive performance (red box plots) have relatively much higher disorder content. The differences in the disorder levels between the results based on accuracy (left side of Figure 5A) and based on AUC (right side of Figure 5B) stem from the fact that AUC can be calculated only for the proteins that have large enough number of disordered residues (see Section 2.3), thus excluding proteins with little or no disorder.

We further test robustness of the above analysis. Some of the long disordered regions that are defined based on the crystal structures may include stable substructures, which could affect results for proteins with the large disorder amount. We find such potentially unreliably annotated disordered regions and repeat our analysis after excluding the corresponding proteins from our dataset. We use a three-step process to identify these problematic disorder annotations among the long disorder regions ($n \geq 30$) collected from proteins with disorder content > 0.5 (Figure 5A shows that these proteins are relatively poorly predicted). First, we download protein sequences (≥ 30 residues long) for crystal structures in PDB and mask residues with missing electron densities (i.e., disordered regions) in these sequences, consequently generating the set of structured sequences. Second, we use BLAST to align the long disordered regions against the masked PDB chain sequences. Third, we remove the proteins that have the long disordered regions which are similar to the structured sequences, i.e., disordered regions which have E-value < 0.01 and identity > 30% based on the alignment against at least one structured sequence. Next, we repeat the analysis of the relation between the content of the native disorder and the protein-level performance using the dataset that excludes the corresponding 15 proteins that harbour 22 potentially problematic long disordered regions that share similarity to the structured sequences. Supplementary Figure S3A that uses this smaller dataset reveals that proteins with large disorder content, both > 0.7 and between 0.5 and 0.7, are predicted with substantially lower levels of accuracy and AUC. Similarly, Supplementary Figure S3B shows that the hard proteins have much more disorder than the easy proteins. These results are in agreement with the results on the complete datasets (Figures 5A and 5B), confirming robustness of our analysis. Altogether, our analysis reveals that current disorder predictors struggle to provide accurate predictions for the mostly disordered proteins.

3.4 Complementarity of disorder predictors at the protein level

We quantify the complementarity with the Pearson correlation coefficients computed between protein-level accuracies and AUCs for each pair of the 13 considered disorder predictors. Figure 6A summarizes results when the performance is measured with accuracy. Protein-level accuracy of every predictor has at least modest levels of correlation (PCC > 0.3) with at least six other predictors (green and yellow cells in Figure 6A). Accuracies of six methods (disEMBL-HL, ESpritz-DisProt, ESpritz-Xray, ESpritz-NMR, GlobPlot and DISOPRED3) lack high correlations (PCC > 0.66) with any other predictors. Overall, only 17% pairs of different methods show lack of correlation (PCC < 0.3) between their protein-level accuracies, while 71% shows modest correlation (0.3 > PCC > 0.66), and 12% has high correlations (PCC > 0.66). We found three clusters of predictors for which the protein-level accuracies are highly correlated (green cells in Figure 6A): 1) VSL2B and JRONN; 2) IUPred-long, IUPred-short, disEMBL-465 and DeepCNF; and 3) disEMBL-465, DeepCNF and SPOT-Disorder. The complementarity of the protein-level AUCs is summarized in Figure 6B. We observe similar trends

when compared to the accuracy. AUCs for every disorder predictor have modest or high correlation with at least four other methods. AUCs for only 23% pairs of different methods lack correlation (PCC < 0.3), compared to 56% and 21% that have modest (0.3 > PCC > 0.66) and high correlations (PCC > 0.66), respectively. There are also three clusters of highly correlated predictors: 1) ESpritz-Xray and DeepCNF; 2) disEMBL-465, JRONN, VSL2B, IUPred-long, IUPred-short and disEMBL-HL; and 3) SPOT-Disorder and DISOPRED3.

Accuracy

	disEMBL-HL	ESpritz-DisProt	ESpritz-Xray	ESpritz-NMR	GlobPlot	DISOPRED3	VSL2B	JRONN	IUPred-long	IUPred-short	disEMBL-465	DeepCNF	SPOT-Disorder
disEMBL-HL	1.00	0.44	0.16	0.20	0.10	0.14	0.49	0.29	0.27	0.39	0.38	0.36	0.34
ESpritz-DisProt	0.44	1.00	0.17	0.21	0.22	0.26	0.49	0.31	0.37	0.45	0.52	0.51	0.44
ESpritz-Xray	0.16	0.17	1.00	0.09	0.36	0.28	0.33	0.36	0.41	0.44	0.40	0.43	0.35
ESpritz-NMR	0.20	0.21	0.09	1.00	0.48	0.50	0.38	0.44	0.60	0.62	0.57	0.56	0.44
GlobPlot	0.10	0.22	0.36	0.48	1.00	0.43	0.31	0.38	0.52	0.52	0.48	0.44	0.43
DISOPRED3	0.14	0.26	0.28	0.50	0.43	1.00	0.24	0.33	0.60	0.59	0.65	0.60	0.61
VSL2B	0.49	0.49	0.33	0.38	0.31	0.24	1.00	0.68	0.52	0.58	0.59	0.56	0.49
JRONN	0.29	0.31	0.36	0.44	0.38	0.33	0.68	1.00	0.64	0.60	0.55	0.53	0.40
IUPred-long	0.27	0.37	0.41	0.60	0.52	0.60	0.52	0.64	1.00	0.87	0.69	0.71	0.58
IUPred-short	0.39	0.45	0.44	0.62	0.52	0.59	0.58	0.60	0.87	1.00	0.78	0.79	0.64
disEMBL-465	0.38	0.52	0.40	0.57	0.48	0.65	0.59	0.55	0.69	0.78	1.00	0.81	0.70
DeepCNF	0.36	0.51	0.43	0.56	0.44	0.60	0.56	0.53	0.71	0.79	0.81	1.00	0.72
SPOT-Disorder	0.34	0.44	0.35	0.44	0.43	0.61	0.49	0.40	0.58	0.64	0.70	0.72	1.00

A

AUC

	ESpritz-DisProt	ESpritz-NMR	GlobPlot	ESpritz-Xray	DeepCNF	disEMBL-465	JRONN	VSL2B	IUPred-long	IUPred-short	disEMBL-HL	SPOT-Disorder	DISOPRED3
ESpritz-DisProt	1.00	0.28	0.02	0.34	0.49	0.29	0.27	0.29	0.23	0.24	0.23	0.37	0.32
ESpritz-NMR	0.28	1.00	0.50	0.38	0.64	0.62	0.54	0.66	0.60	0.64	0.53	0.45	0.36
GlobPlot	0.02	0.50	1.00	0.41	0.27	0.42	0.43	0.54	0.63	0.64	0.53	0.12	0.04
ESpritz-Xray	0.34	0.38	0.41	1.00	0.69	0.60	0.53	0.59	0.52	0.58	0.49	0.52	0.43
DeepCNF	0.49	0.64	0.27	0.69	1.00	0.67	0.50	0.65	0.50	0.58	0.60	0.66	0.59
disEMBL-465	0.29	0.62	0.42	0.60	0.67	1.00	0.71	0.84	0.75	0.79	0.67	0.39	0.34
JRONN	0.27	0.54	0.43	0.53	0.50	0.71	1.00	0.73	0.71	0.69	0.47	0.38	0.31
VSL2B	0.29	0.66	0.54	0.59	0.65	0.84	0.73	1.00	0.84	0.88	0.66	0.35	0.28
IUPred-long	0.23	0.60	0.63	0.52	0.50	0.75	0.71	0.84	1.00	0.94	0.66	0.21	0.12
IUPred-short	0.24	0.64	0.64	0.58	0.58	0.79	0.69	0.88	0.94	1.00	0.72	0.27	0.18
disEMBL-HL	0.23	0.53	0.53	0.49	0.60	0.67	0.47	0.66	0.66	0.72	1.00	0.25	0.20
SPOT-Disorder	0.37	0.45	0.12	0.52	0.66	0.39	0.38	0.35	0.21	0.27	0.25	1.00	0.80
DISOPRED3	0.32	0.36	0.04	0.43	0.59	0.34	0.31	0.28	0.12	0.18	0.20	0.80	1.00

B

Figure 6. Correlations between the protein-level predictive performance for each pair of the considered 13 disorder predictors. Panels A and B panels quantify the performance with accuracy and AUC, respectively. Both correlation matrices are symmetric. The sorting of the predictors differs between the two panel and was optimized to highlight clusters of highly correlated methods. Values of the Pearson correlation coefficient (PCC) are color-coded where red denotes no correlation (PCC < 0.3), yellow denotes modest correlation (0.3 ≤ PCC ≤ 0.66) and green corresponds to high correlation (PCC > 0.66).

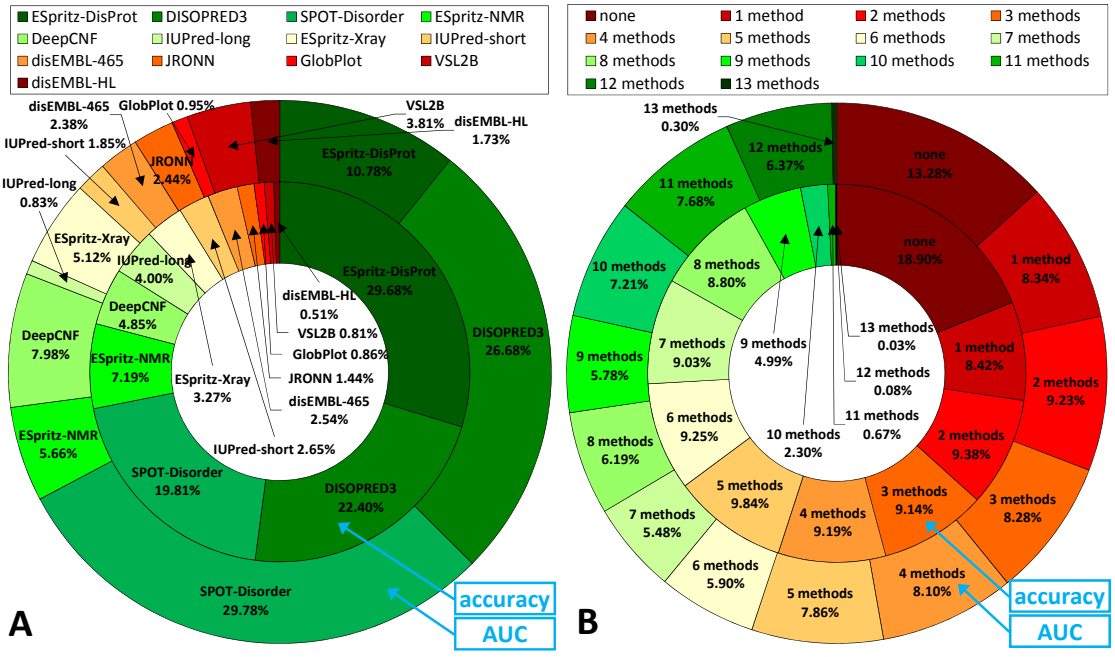

Figure 7. Contributions of the 13 disorder predictors to the highly accurate predictions. Panel A quantifies the fraction of proteins for which a given method generates the highest predictive performance compared to all other disorder predictors. Panel B show the fraction of proteins for which a given number of predictors offers

highly accurate predictions, i.e., predictive performance that is higher than the expected performance of the best method (the dataset-level performance of the best method). The inner and outer rings show results when using accuracy and AUC, respectively.

Our analysis points to several interesting observations. First, protein-level predictive performance is at least modestly correlated for vast majority of pairs of methods. This suggests that there are no clear winners or losers and that using multiple different methods should provide relatively good results. Second, results for AUC and accuracy are in close agreement. The average absolute difference between the corresponding PCCs is only 0.17, suggesting that similarity in the propensity scores is associated with the similarity in binary predictions. In other words, different methods assign the binary predictions from the propensity scores in similar way. Third, methods that are placed in the abovementioned clusters will produce similar protein-level results and should not be used together. Using methods that have weaker mutual correlations would provide a more diverse set of results, therefore improving chances to secure a more accurate prediction.

Figure 7 further expands the analysis of the first finding. It summarizes contributions of different disorder predictors to the production of highly accurate protein-level predictions. Figure 7A shows the fraction of proteins for which a given disorder predictor generates the most accurate result, as quantified with the accuracy (inner ring) and AUC (outer ring). The best performing at the dataset-level DISOPRED3 and SPOT-Disorder provide the best protein-level results for only about half of the proteins. Moreover, each of the 13 tools generates the most accurate protein-level predictions for some proteins. This includes the worst dataset-level performers, GlobPlot and disEMBL-HL, which provide the best accuracies for 0.9% and 0.5% of proteins, respectively, and the highest AUCs for 0.9% and 1.7%, respectively. Figure 7B breaks down the proteins for which a given number of predictors offers highly accurate predictions, i.e., predictive quality that is higher than the expected quality of the best method (the dataset-level performance of the best method). The figure reveals that only less than 19% of proteins lack highly accurate predictions (dark red regions in Figure 7B). Furthermore, over half of the proteins secure highly accurate predictions (measured with either accuracy or AUC) by at least four disorder predictors, while 25% of proteins (when using accuracy) and 39% of proteins (when using AUC) have such accurate predictions produced by the majority of the 13 disorder predictors. The bottom line is that high-quality protein-level predictions can be often obtained from several disorder predictors. This suggests that the end users should not limit themselves to using only the most accurate (at the dataset-level) methods.

3.5 Case study

Our care study visualizes and compares outputs produced by several disorder predictors for the same protein, hydroxymethyltransferase from *Mycobacterium tuberculosis* (Uniprot id: P9WIL7). The native annotations of disorder for this protein were obtained from its crystal structure (PDB ID: 1OY0). Figure 8 shows that this protein has short disordered regions at both termini and another short disordered region in the middle of the chain (positions 161 to 174). Figure 8 visualizes results generated by five disorder predictors that cover the entire spectrum of the predictive performance including the two most accurate methods: DISOPRED3 and SPOT-Disorder, two modestly accurate VSL2B and IUPred-long, and the least accurate methods, GlobPlot. The five predictors substantially outperform their dataset-level AUCs for this protein: DISOPRED3 (AUC = 0.96 for this protein vs. 0.90 at the dataset-level), SPOT-Disorder (0.96 vs. 0.90), VSL2B (0.88 vs. 0.81), IUPred-long (0.93 vs. 0.73) and GlobPlot (0.80 vs. 0.63). The five methods correctly predict the longest disordered region at the N-terminus and they also correctly show that this protein is primarily structured. This is why their predictive performance is so high. Only VSL2B and DISPORED3 were able to find the disordered region at the other terminus, and only GlobPlot and IUPred-long found the disordered region in the middle of the sequence. The GlobPlot's success for the latter region comes at the expense of over-predicting disorder in several other regions, which is why GlobPlot secures the lowest AUC among

the five computational tools. The very high AUC of SPOT-Disorder and DISOPRED3 are the result of the fact that these methods predict high propensities for the disordered region at the N-terminus while also predicting the lowest propensities for the structured regions. Moreover, outputs of these two methods show spikes in their predicted propensities in the vicinity of the two other disordered regions. While these spikes are not high enough to trigger generation of the binary disorder prediction, they suggest that disorder is more likely in these regions than in the other parts of this sequence. Overall, this case study shows how the disorder predictors can beat their dataset-level predictive performance, which is a typical scenario that is revealed by our analysis.

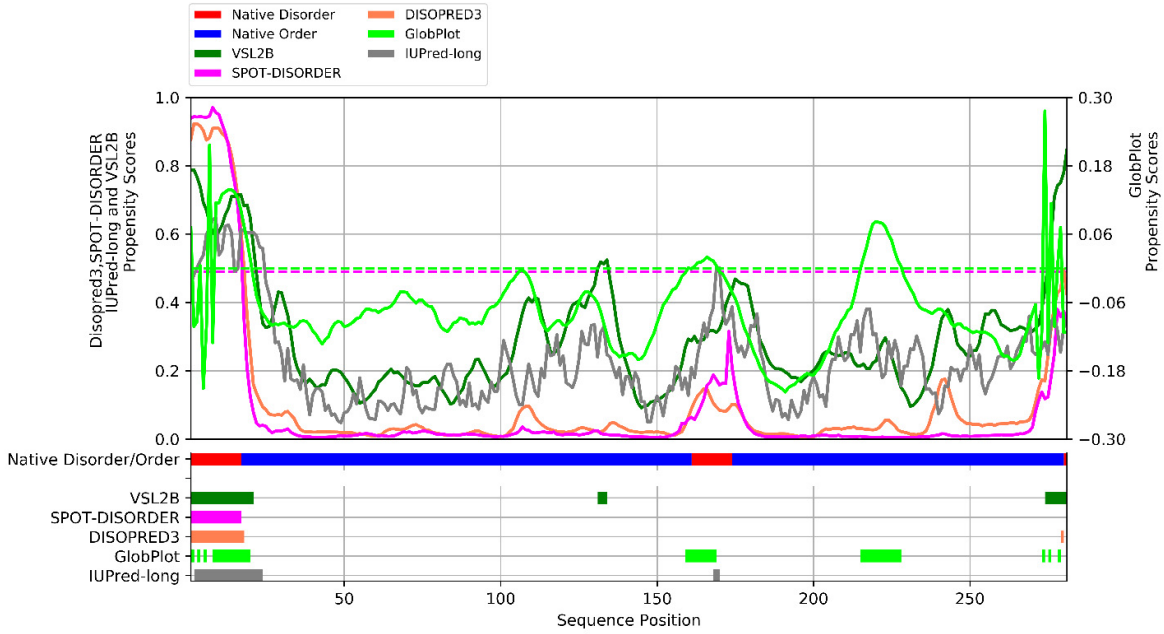

Figure 8. A case study that compares disorder predictions for the hydroxymethyltransferase protein from *Mycobacterium tuberculosis* (Uniprot id: P9WIL7) that were generated by five methods: VSL2B (dark green), SPOT-DISORDER (magenta), DISOPRED3 (orange), GlobPlot (lime) and IUPred-long (grey). The putative propensities are shown using the solid, color-coded lines. The corresponding binary predictions are given using the color-coded horizontal bars at the bottom of the figure; thresholds that are used to convert the propensities into the binary predictions are visualized with the dashed horizontal lines in the top part of the figure. The red and blue horizontal bar denotes the native annotation of disordered and structured regions, respectively, which were annotated using crystal structure (PDB ID: 1OY0).

4 Summary

Accurate prediction of intrinsic disorder for the millions of the currently unannotated protein sequences is facilitated by the many predictors of disordered regions [63-65, 67]. Users who navigate the selection of these predictive methods benefit from the many comparative assessments that were released over the last decade [55, 103-106, 110-112]. However, these studies focus on the dataset-level analysis of predictive quality while users often apply these tools to make predictions for individual proteins. To this end, this article provides a detailed evaluation and analysis of the protein-level results and extends the knowledge that we can glean from the current dataset-level benchmarks.

Our first-of-its-kind large scale analysis of 13 representative disorder predictors shows that the quality of the protein-level predictions is often very different from the dataset-level results. We demonstrate that the protein-level predictive performance is in fact higher than the corresponding

dataset-level assessments suggest for a substantial majority of the proteins, as many as over 70% of proteins for the ESpritz-Xray, SPOT-Disorder and DISOPRED3 methods. This observation is consistent for all 13 predictors and for both types of their outputs: putative propensities for disorder and binary disorder predictions. However, this advantage comes at the cost of under-performing predictive quality for between 10% and 30% of proteins, depending on the particular disorder predictor. These proteins are predicted with accuracy/AUC that is substantially below the expected (dataset-level) estimates.

Our analysis reveals a relationship between the accuracy of the protein-level predictions and the native disorder content. We show that accurately predicted proteins, which secure above-expected (better than dataset-level) predictive performance, typically have relatively low amounts of disorder. On the other hand, the inaccurately predicted proteins for which predictive quality is substantially below the expected/dataset-level value have high amounts of disorder. This finding parallels the published observations that disorder predictors performs relatively poorly for proteins with long disordered regions [55, 84, 104, 105]. Consequently, we recommend that the development of novel predictors that target accurate prediction of the mostly disordered proteins should be pursued.

We investigate similarities of the protein-level predictive performance between different predictors. We found several clusters of methods that produce highly correlated protein-level results. Our recommendation is that users who want to benefit from application of multiple predictors should apply predictors that come from different clusters. This would ensure more diversity in predictions and would increase likelihood of securing a highly accurate prediction. This benefit is premised on the availability of a tool capable of predicting correctness of disorder predictions. A relevant computational tool that predicts quality assessment scores for the residue-level predictions of ten popular disorder predictors is QUARTER [120]. However, further research is needed to assess whether QUARTER's scores can be used to identify accurate protein-level predictions. Furthermore, we advocate for the development of new family of predictive tools capable of identifying hard-to-predict proteins. These methods could be used to suggest disorder predictors that produce accurate predictions for a given input protein chain.

Our statistical test show that the SPOT-Disorder and DISOPRED3 provide significantly higher protein-level AUCs when compared with the other considered here predictors. Moreover, ESpritz-DisProt provides predictions with the lowest protein-level FPRs but at the cost of substantially under-predicting the amount of disorder. However, we also empirically demonstrate that the protein-level predictive performance is at least modestly correlated for substantial majority of methods. This result indicates that multiple predictors can provide relatively good results and that no single method is universally superior. Detailed analysis shows that accurate protein-level predictions for a given protein of interest are in fact usually produced by several disorder predictors, Moreover, the predictor with the most accurate dataset-level results outperforms the other methods for only about 30% of proteins. These findings support the need and feasibility of the development of the new family of predictive tools and they suggest that the users should explore using multiple disorder predictors, instead of simply using one (the most accurate) method.

Finally, recent research in this area focuses on the development of predictors of specific types of functional disordered regions. Several methods that target prediction of disordered protein-binding regions [73, 124-129], nucleic acids-binding regions [124, 130], linker regions [131], and multifunctional regions [132] were released in recent years. While the currently low numbers of these tools that address a specific region type prohibit comparative analysis of the protein-level performance, we anticipate that such analysis will be needed in the near future.

Key Points

- Protein-level performance of disorder predictions varies widely and differs from the currently used dataset-level benchmark results
- A large majority of proteins secure predictive performance that is higher than the dataset-level result suggest
- Disorder predictors perform poorly for between 10% and 30% of proteins, and these proteins typically have a substantial amount of intrinsic disorder
- No disorder predictor is universally superior and highly accurate disorder predictions can be typically generated with multiple methods
- Novel computational tools that accurately identify the hard-to-predict proteins and that make accurate predictions for these proteins are needed

Funding

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

Conflict of Interest: none declared.

References

1. Lieutaud P, Ferron F, Uversky AV et al. How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe, Intrinsically Disord Proteins 2016;4:e1259708.
2. Dunker AK, Babu MM, Barbar E et al. What's in a name? Why these proteins are intrinsically disordered, Intrinsically Disordered Proteins 2013;1:e24157.
3. Habchi J, Tompa P, Longhi S et al. Introducing Protein Intrinsic Disorder, Chemical Reviews 2014;114:6561-6588.
4. Uversky VN. Introduction to intrinsically disordered proteins (IDPs), Chem Rev 2014;114:6557-6560.
5. Peng Z, Yan J, Fan X et al. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life, Cell Mol Life Sci 2015;72:137-151.
6. Oates ME, Romero P, Ishida T et al. D(2)P(2): database of disordered protein predictions, Nucleic Acids Res 2013;41:D508-516.
7. Xue B, Dunker AK, Uversky VN. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, J Biomol Struct Dyn 2012;30:137-149.
8. Ward JJ, Sodhi JS, McGuffin LJ et al. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J Mol Biol 2004;337:635-645.
9. Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions, Proteins 2014;82:145-158.
10. Bhowmick A, Brookes DH, Yost SR et al. Finding Our Way in the Dark Proteome, J Am Chem Soc 2016;138:9730-9742.
11. Hu G, Wang K, Song J et al. Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity, Proteomics 2018:e1800243.
12. Kulkarni P, Uversky VN. Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome, Proteomics 2018;18.
13. Yan J, Dunker AK, Uversky VN et al. Molecular recognition features (MoRFs) in three domains of life, Mol Biosyst 2016;12:697-710.

14. Mohan A, Oldfield CJ, Radivojac P et al. Analysis of molecular recognition features (MoRFs), J Mol Biol 2006;362:1043-1059.
15. Xie H, Vucetic S, Iakoucheva LM et al. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, J. Proteome Res. 2007;6:1882-1898.
16. Dunker AK, Brown CJ, Lawson JD et al. Intrinsic disorder and protein function, Biochemistry 2002;41:6573-6582.
17. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions, Nat Rev Mol Cell Biol 2005;6:197-208.
18. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, J Mol Recognit 2005;18:343-384.
19. Liu J, Perumal NB, Oldfield CJ et al. Intrinsic disorder in transcription factors, Biochemistry 2006;45:6873-6888.
20. Peng Z, Oldfield CJ, Xue B et al. A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome, Cell Mol Life Sci 2014;71:1477-1504.
21. Peng Z, Mizianty MJ, Xue B et al. More than just tails: intrinsic disorder in histone proteins, Mol Biosyst 2012;8:1886-1901.
22. Wang C, Uversky VN, Kurgan L. Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea, Proteomics 2016;16:1486-1498.
23. Meng F, Na I, Kurgan L et al. Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments, Int J Mol Sci 2015;17.
24. Fuxreiter M, Toth-Petroczy A, Kraut DA et al. Disordered proteinaceous machines, Chem Rev 2014;114:6806-6843.
25. Na I, Meng F, Kurgan L et al. Autophagy-related intrinsically disordered proteins in intra-nuclear compartments, Mol Biosyst 2016;12:2798-2817.
26. Uversky AV, Xue B, Peng Z et al. On the intrinsic disorder status of the major players in programmed cell death pathways, F1000Res 2013;2:190.
27. Peng Z, Xue B, Kurgan L et al. Resilience of death: intrinsic disorder in proteins involved in the programmed cell death, Cell Death Differ 2013;20:1257-1267.
28. Fan X, Xue B, Dolan PT et al. The intrinsic disorder status of the human hepatitis C virus proteome, Mol Biosyst 2014;10:1345-1363.
29. Charon J, Theil S, Nicaise V et al. Protein intrinsic disorder within the Potyvirus genus: from proteome-wide analysis to functional annotation, Molecular Biosystems 2016;12:634-652.
30. Dolan PT, Roth AP, Xue B et al. Intrinsic disorder mediates hepatitis C virus core-host cell protein interactions, Protein Sci 2015;24:221-235.
31. Xue B, Uversky VN. Intrinsic disorder in proteins involved in the innate antiviral immunity: another flexible side of a molecular arms race, J Mol Biol 2014;426:1322-1350.
32. Xue B, Mizianty MJ, Kurgan L et al. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1, Cell Mol Life Sci 2012;69:1211-1259.
33. Meng F, Badierah RA, Almehdar HA et al. Unstructural biology of the Dengue virus proteins, FEBS J 2015;282:3368-3394.
34. Kjaergaard M, Kragelund BB. Functions of intrinsic disorder in transmembrane proteins, Cellular and Molecular Life Sciences 2017;74:3205-3224.
35. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease, Biochem Soc Trans 2016;44:1185-1200.
36. Varadi M, Zsolyomi F, Guharoy M et al. Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins, PLoS One 2015;10:e0139731.
37. Dunker AK, Silman I, Uversky VN et al. Function and structure of inherently disordered proteins, Curr Opin Struct Biol 2008;18:756-764.

38. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept, Annu Rev Biophys 2008;37:215-246.
39. Uversky VN, Dave V, Iakoucheva LM et al. Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases, Chem Rev 2014;114:6844-6879.
40. Uversky VN. The triple power of D(3): protein intrinsic disorder in degenerative diseases, Front Biosci (Landmark Ed) 2014;19:181-258.
41. Cheng Y, LeGall T, Oldfield CJ et al. Rational drug design via intrinsically disordered protein, Trends Biotechnol 2006;24:435-442.
42. Uversky VN. Intrinsically disordered proteins and novel strategies for drug discovery, Expert Opin Drug Discov 2012;7:475-488.
43. 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;12:221-233.
44. Ambadipudi S, Zweckstetter M. Targeting intrinsically disordered proteins in rational drug discovery, Expert Opin Drug Discov 2015:1-13.
45. Dunker AK, Uversky VN. Drugs for 'protein clouds': targeting intrinsically disordered transcription factors, Curr Opin Pharmacol 2010;10:782-788.
46. Peterson LX, Roy A, Christoffer C et al. Modeling disordered protein interactions from biophysical principles, PLoS Comput Biol 2017;13:e1005485.
47. Huang J, Rauscher S, Nawrocki G et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins, Nat Methods 2017;14:71-73.
48. Choi JM, Pappu RV. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics, J Chem Theory Comput 2019;15:1367-1382.
49. Piovesan D, Tabaro F, Micetic I et al. DisProt 7.0: a major update of the database of disordered proteins, Nucleic Acids Res 2016;D1:D219-D227.
50. Piovesan D, Tabaro F, Paladin L et al. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins, Nucleic Acids Res 2018;46:D471-D476.
51. Fukuchi S, Amemiya T, Sakamoto S et al. IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners, Nucleic Acids Res 2014;42:D320-325.
52. Berman HM, Westbrook J, Feng Z et al. The Protein Data Bank, Nucleic Acids Research 2000;28:235-242.
53. DeForte S, Uversky VN. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree, Protein Sci 2016;25:676-688.
54. Le Gall T, Romero PR, Cortese MS et al. Intrinsic disorder in the Protein Data Bank, J Biomol Struct Dyn 2007;24:325-342.
55. Walsh I, Giollo M, Di Domenico T et al. Comprehensive large-scale assessment of intrinsic protein disorder, Bioinformatics 2015;31:201-208.
56. The UniProt C. UniProt: the universal protein knowledgebase, Nucleic Acids Res 2017;45:D158-D169.
57. Dunker AK, Lawson JD, Brown CJ et al. Intrinsically disordered protein, J Mol Graph Model 2001;19:26-59.
58. Uversky VN. Natively unfolded proteins: a point where biology waits for physics, Protein Sci 2002;11:739-756.
59. Uversky VN. What does it mean to be natively unfolded?, Eur J Biochem 2002;269:2-12.
60. Uversky VN, Dunker AK. Understanding protein non-folding, Biochim Biophys Acta 2010;1804:1231-1264.
61. Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions?, Proteins 2000;41:415-427.
62. He B, Wang K, Liu Y et al. Predicting intrinsic disorder in proteins: an overview, Cell Res 2009;19:929-949.

63. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, Cell Mol Life Sci 2017;74:3069-3090.
64. Meng F, Uversky V, Kurgan L. Computational Prediction of Intrinsic Disorder in Proteins, Curr Protoc Protein Sci 2017;88:2 16 11-12 16 14.
65. Dosztányi Z, Tompa P. Bioinformatics Approaches to the Structure and Function of Intrinsically Disordered Proteins. In: J. Rigden D. (ed) From Protein Structure to Function with Bioinformatics. Dordrecht: Springer Netherlands, 2017, 167-203.
66. Dosztanyi Z, Meszaros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins, Briefings in Bioinformatics 2010;11:225-243.
67. Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction, Brief Bioinform 2019;20:330-346.
68. Li J, Feng Y, Wang X et al. An Overview of Predictors for Intrinsically Disordered Proteins over 2010-2014, Int J Mol Sci 2015;16:23446-23462.
69. Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure, Nucleic Acids Research 2003;31:3833-3835.
70. Linding R, Russell RB, Neduva V et al. GlobPlot: exploring protein sequences for globularity and disorder, Nucleic Acids Res 2003;31:3701-3708.
71. Dosztányi Z, Csizmok V, Tompa P et al. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics 2005;21:3433-3434.
72. Dosztányi Z, Csizmók V, Tompa P et al. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins, J Mol Biol 2005;347:827-839.
73. Meszaros B, Erdos G, Dosztanyi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res 2018;46:W329-W337.
74. Yang ZR, Thomson R, McNeil P et al. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, Bioinformatics 2005;21:3369-3376.
75. Linding R, Jensen LJ, Diella F et al. Protein Disorder Prediction: Implications for Structural Proteomics, Structure 2003;11:1453-1459.
76. Cheng J, Sweredoski M, Baldi P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data, Data Mining and Knowledge Discovery 2005;11:213-222.
77. Hecker J, Yang JY, Cheng J. Protein disorder prediction at multiple levels of sensitivity and specificity, BMC Genomics 2008;9:1-7.
78. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices, Proteins: Structure, Function, and Bioinformatics 2003;53:573-578.
79. Ward JJ, McGuffin LJ, Bryson K et al. The DISOPRED server for the prediction of protein disorder, Bioinformatics 2004;20:2138-2139.
80. Peng K, Radivojac P, Vucetic S et al. Length-dependent prediction of protein intrinsic disorder, BMC Bioinformatics 2006;7:208.
81. Obradovic Z, Peng K, Vucetic S et al. Exploiting heterogeneous sequence properties improves prediction of protein disorder, Proteins 2005;61 Suppl 7:176-182.
82. Zhang T, Faraggi E, Xue B et al. SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method, Journal of biomolecular structure & dynamics 2012;29:799-813.
83. Wang S, Weng S, Ma J et al. DeepCNF-D: Predicting Protein Order/Disorder Regions by Weighted Deep Convolutional Neural Fields, International Journal of Molecular Sciences 2015;16:17315.
84. Hanson J, Yang Y, Paliwal K et al. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks, Bioinformatics 2017;33:685-692.
85. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence, Nucleic Acids Res 2007;35:W460-W464.

86. 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.
87. Peng Z, Kurgan L. On the complementarity of the consensus-based disorder prediction, Pac Symp Biocomput 2012:176-187.
88. Fan X, Kurgan L. Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus, J Biomol Struct Dyn 2014;32:448-464.
89. Schlessinger A, Punta M, Yachdav G et al. Improved Disorder Prediction by Combination of Orthogonal Approaches, PLoS ONE 2009;4:e4433.
90. Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins, BMC Bioinformatics 2012;13:1-11.
91. Huang YJ, Acton TB, Montelione GT. DisMeta: a meta server for construct design and optimization, Methods Mol Biol 2014;1091:3-16.
92. Xue B, Dunbrack RL, Williams RW et al. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids, Biochim Biophys Acta 2010;1804:996-1010.
93. Walsh I, Martin AJM, Di Domenico T et al. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs, Nucleic Acids Research 2011;39:W190-W196.
94. Mizianty MJ, Stach W, Chen K et al. Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, Bioinformatics 2010;26:i489-i496.
95. Mizianty MJ, Uversky V, Kurgan L. Prediction of intrinsic disorder in proteins using MFDp2, Methods Mol Biol 2014;1137:147-162.
96. Mizianty MJ, Peng ZL, Kurgan L. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles, Intrinsically Disordered Proteins 2013;1:e24428.
97. Walsh I, Martin AJM, Di Domenico T et al. ESpritz: accurate and fast prediction of protein disorder, Bioinformatics 2012;28:503-509.
98. Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach, Bioinformatics 2008;24:1344-1348.
99. Mizianty MJ, Peng Z, Kurgan L. MFDp2-Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles, Intrinsically Disordered Proteins 2013;1:e24428.
100. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 2015;31:857-863.
101. Necci M, Piovesan D, Dosztanyi Z et al. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins, Bioinformatics 2017;33:1402-1404.
102. Oates ME, Romero P, Ishida T et al. D2P2: database of disordered protein predictions, Nucleic Acids Res 2013;41:D508-D516.
103. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions, Curr Protein Pept Sci 2012;13:6-18.
104. Monastyrskyy B, Kryshtafovych A, Moult J et al. Assessment of protein disorder region predictions in CASP10, Proteins 2014;82 Suppl 2:127-137.
105. Monastyrskyy B, Fidelis K, Moult J et al. Evaluation of disorder predictions in CASP9, Proteins 2011;79 Suppl 10:107-118.
106. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8, Proteins 2009;77 Suppl 9:210-216.
107. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7, Proteins 2007;69 Suppl 8:129-136.
108. Jin Y, Dunbrack RL, Jr. Assessment of disorder predictions in CASP6, Proteins 2005;61 Suppl 7:167-175.
109. Melamud E, Moult J. Evaluation of disorder predictions in CASP5, Proteins 2003;53 Suppl 6:561-565.
110. Necci M, Piovesan D, Dosztanyi Z et al. A comprehensive assessment of long intrinsic protein disorder from the DisProt database, Bioinformatics 2018;34:445-452.

111. Pryor EE, Jr., Wiener MC. A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder, Biophys J 2014;106:1638-1649.
112. Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods, Mol Biosyst 2012;8:114-121.
113. Postel S, Deredge D, Bonsor DA et al. Bacterial flagellar capping proteins adopt diverse oligomeric states, Elife 2016;5.
114. Vo AT, Fleischman NM, Marquez MD et al. Defining the domains of Cia2 required for its essential function in vivo and in vitro, Metallomics 2017;9:1645-1654.
115. Jain G, Pendola M, Rao A et al. A Model Sea Urchin Spicule Matrix Protein Self-Associates To Form Mineral-Modifying Protein Hydrogels, Biochemistry 2016;55:4410-4421.
116. Chang EP, Perovic I, Rao A et al. Insect Cell Glycosylation and Its Impact on the Functionality of a Recombinant Intracrystalline Nacre Protein, AP24, Biochemistry 2016;55:1024-1035.
117. Yadav LR, Rai S, Hosur MV et al. Functional assessment of intrinsic disorder central domains of BRCA1, J Biomol Struct Dyn 2015;33:2469-2478.
118. Camacho C, Coulouris G, Avagyan V et al. BLAST+: architecture and applications, BMC Bioinformatics 2009;10:421.
119. Wu Z, Hu G, Wang K et al. Exploratory Analysis of Quality Assessment of Putative Intrinsic Disorder in Proteins. 6th International Conference on Artificial Intelligence and Soft Computing. Zakopane, Poland, 2017, 722-732.
120. Hu G, Wu Z, Oldfield C et al. Quality Assessment for the Putative Intrinsic Disorder in Proteins, Bioinformatics 2018.
121. Wootton JC. Nonglobular Domains in Protein Sequences - Automated Segmentation Using Complexity-Measures, Computers & Chemistry 1994;18:269-285.
122. Jones DT, Swindells MB. Getting the most from PSI-BLAST, Trends in Biochemical Sciences 2002;27:161-164.
123. Prilusky J, Felder CE, Zeev-Ben-Mordehai T et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded, Bioinformatics 2005;21:3435-3438.
124. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder, Nucleic Acids Res 2015;43:e121.
125. Disfani FM, Hsu WL, Mizianty MJ et al. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, Bioinformatics 2012;28:i75-83.
126. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences, Nucleic Acids Res 2016.
127. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 2015;31:857-863.
128. Yan J, Dunker AK, Uversky VN et al. Molecular recognition features (MoRFs) in three domains of life, Molecular BioSystems 2015.
129. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins, Bioinformatics 2009;25:2745-2746.
130. Peng Z, Wang C, Uversky VN et al. Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind, Methods Mol Biol 2017;1484:187-203.
131. Meng F, Kurgan L. DFLpred: High-throughput prediction of disordered flexible linker regions in protein sequences, Bioinformatics 2016;32:i341-i350.
132. Meng F, Kurgan L. High-throughput prediction of disordered moonlighting regions in protein sequences, Proteins 2018.