Comprehensive comparative assessment of in-silico predictors of disordered regions

Zhen-Ling Peng[1] and Lukasz Kurgan[1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, CANADA
[*]corresponding author; email: lkurgan@ece.ualberta.ca; phone (780) 492-5488

Abstract

The intrinsic disorder is relatively common in proteins, plays important roles in numerous cellular activities, and its prevalence was implicated in various human diseases. However, annotations of the disorder lag behind the rapidly increasing number of known protein chains. Last decade observed development of a relatively large number of *in-silico* methods that predict the disorder using the protein sequence as their input. We perform a first-of-its kind comprehensive empirical evaluation of the disorder predictors which is characterized by three novel aspects, (1) we evaluate the quality of the disorder predictions at the residue, segment, and chain levels; (2) we consider a large number of published and accessible to the end user predictors that are evaluated on a relatively big dataset with close to 500 proteins; and (3) we assess statistical significance of differences between the considered methods. Our study reveals that there is no universally superior predictor and that the top-performing methods are complementary. We show that while recent consensus-based predictors outperform other considered methods for the residue-level predictions, some older methods perform better for the prediction of the disordered segments. Our analysis indicates that certain predictors are biased to under-predict the disorder, while some other solutions tend to over-predict the number of the disordered residues. We also evaluate the utility of the predicted residue-level disorder for prediction of proteins with long disordered segments and prediction of the chain-level disorder content. Lastly, we provide recommendations concerning development of a new generation of consensus-based methods and specialized methods for improved prediction of the disorder content.

Introduction

The intrinsically disordered proteins (IDPs), also called rheomorphic, natively denatured, natively unfolded, intrinsically unstructured, mostly unstructured, and natively disordered [[1]], lack stable tertiary structure under physiological conditions *in vitro*. IDPs take form of dynamic structural ensembles that undergo non-cooperative conformational changes, which means that the positions of their atoms and backbone angles have no specific equilibrium states and they vary largely over time. They also encompass random coil-like regions and collapsed, i.e., partially folded or molten/pre-molten globule-like, domains with poorly packed side chains [[2],[3]]. The intrinsic disorder is relatively common. For instance, eukaryotic genomes, such as *C. elegans*, *A. thaliana*, *Saccharomyces cerevisiae*, and D. *melanogaster*, were estimated to have between 52% and 67% of their proteins with long, over 40 consecutive residues, disordered regions [[4]]. IDPs play important roles in transcriptional regulation, translation, and cellular signal transduction [[5]]. Their prevalence was implicated in various human diseases [[6],[7]] and they were suggested as important targets for drug discovery [[8]]. However, their functional role is not as well understood when compared with the structured proteins. Furthermore, the disorder annotations lag behind the rapidly accumulating number of known protein chains. The disorder is frequently observed in regions with low sequence complexity and with low content of hydrophobic amino acids and high net charge, which would often form a core of a folded globular protein, and is often associated with lack of secondary structure and unique evolutionary profiles [[9]-[13]]. These sequence characteristics imply that disorder is predictable from the protein sequence. To this end, the past decade has seen development of a number of computational models for the prediction of the disordered regions from protein chains. The disorder prediction was included in the biannual CASP experiments since 2002 [[14]-[17]]. Inclusion of this prediction category in the CASP resulted in a substantial increase in the number of disorder predictors, with a few dozens of methods that were developed and published by now. A comprehensive listing and summary of these methods are included in three recent reviews [1,[13],[18]], as well as at http://www.disprot.org/predictors.php, which in a part of the DisProt database [[19]]. The disorder predictors allow for high-throughput annotations of protein chains and therefore they provide a viable solution to close the annotation gap. They could be categorized into 4 types:

1. methods based on relative propensity of amino acids to form disorder/ordered regions which include GlobPlot [[20]], IUPred [[21]], FoldIndex [[22]], and Ucon [[23]];
2. predictors built utilizing machine learning classifiers, such as DISOPRED [[24]], DisEMBL [[25]], DISOPRED2 [[26]], DISpro [[27]], RONN [[28]], Spritz [[29]], ProfBval [[30],[31]], PONDR family of predictors [[10],[32]-[36]], DisPSSMP [[37]], DisPSSMP2 [[38]], POODLE family of predictors [[39],[40]], NORSnet [[41]], IUP [[42]], and OnD-CRFs [[43]];
3. methods based on a meta-approach which combines predictions from multiple base predictors including PreDisorder [[27],[44]], metaPrDOS [[45]], MD [[46]], and most recently PONDR-FIT [[47]] and MFDp [[48]] predictors;
4. approaches based on analysis of predicted 3D structural models such as PrDOS [[49]] and DISOclust [[50]].

Prior works show that the assignment of the disordered regions performed using different experimental methods could be inconsistent [[32]]. Disorder predictors that were developed using regions identified by one experimental method could be less accurate for prediction of disorder characterized by other methods [[41]]. To date, there is no golden standard for the assignment of the disordered regions. In the past CASP experiments the disordered regions were defined as residues that lack coordinates in structures solved by X-ray crystallography and as residues that exhibit high variability within the structural ensembles or are annotated as disordered in REMARK 465 by experimentalists for the structures solved by NMR [[16],[17]]. Another commonly used source of the disorder annotations is the manually curated DisProt database [[19]], which includes annotations of experimentally verified and biologically relevant unstructured regions. We use both types of annotations to build our benchmark dataset, which contrasts our work with the evaluations performed at the CASP experiment. Although CASP experiments provide a useful forum to evaluate the state-of-the-art in disorder prediction and to compare various methods, they have a few drawbacks. They allow for submission of predictions from methods that are unpublished and inaccessible to the practitioners and researchers, they perform evaluations on relatively small datasets with up to 150 chains, and they concentrate on evaluation of the disorder predictions only at the residue level.

To this end, we perform a first-of-its kind comprehensive empirical evaluation of disorder predictors that (1) considers a large number of published and accessible to the end user, either as a standalone program or a web server, predictors; (2) comprehensively evaluates statistical significance of differences between all pairs of the considered methods; (3) utilizes a relatively large dataset with close to 500 chains; and (4) evaluates disorder predictions at the residues, segment, and chain levels. More specifically, we evaluate the disorder predictions for individual residues, including both the real-valued (probability of disorder) and the binary (assignment as either ordered or disordered) predictions; for segments of disordered residues, which is analogous to the segment-based evaluations performed for the secondary structure predictors [[51]]; and for the predicted amount and inclusion of long disordered segments in the entire protein chain. The segment-level evaluation is motivated by the fact that disordered residues are clustered together in the sequence forming segments. We quantify the amount of overlap between the predicted and the native disordered segments. We also evaluate whether the exiting methods accurately predict the overall amount of disorder, i.e., disorder content, in the protein chain. This is stimulated by the observation that the disorder content, which is computed from the per-residue disorder predictions, was extensively used to estimate the abundance of intrinsic disorder in protein databases [[52],[53]], protein families and classes [[54]-[62]], and in complete proteomes [4,[26],[63],[64]]. The content was also utilized to analyze intrinsic disorder-related protein functions [[65]-[67]]. Finally, we also investigate the quality the prediction applied to find proteins with long, ≥ 30 consecutive residues, disordered regions. This binary per-chain prediction is encouraged by the fact that this information is useful for target selection [[68],[69]] and protein-protein recognition [[70]]. The latter type of the evaluation was also performed on a smaller scale for the MD [[46]] and MFDp [[48]] predictors.

Materials and Methods

Considered prediction methods

The primary selection criteria were that each of the included methods has to be accessible to the end user as either standalone software or a web server and that it has to be published in a reputable peer-reviewed scientific venue. Our assessment considers sixteen disorder predictors that cover the fours types of the methods, including four relative propensity-based methods including GlobPlot [[20]], IUPred [[21]], FoldIndex [[22]], and Ucon [[23]]; eight machine learning-based solutions such as DisEMBL [[25]], DISOPRED2 [[26]], DISpro [[27]], RONN [[28]], Spritz [[29]], ProfBval [[30],[31]], VSL2B [[36]], NORSnet [[41]]; three most recent consensus-based methods including MD [[46]], PONDR-FIT [[47]] and MFDp [[48]]; and the most recent 3D-prediction based DISOclust [[50]]. The IUPred was used in both of its modes, one for prediction of short disordered segments, IUPredS, and the other for long segments, IUPredL. The DisEMBL method consists of three predictors, DisEMBL-H designed to detect hot-loop, DisEMBL-C that finds coils and loops, and DisEMBL-R that predicts residues annotated with REMARK 465. As a result, including two versions of IUPred and three versions of DisEMBL, we test total of 19 methods. These methods are summarized in Table 1. We include information concerning their inputs, the predictive models that they utilize, and their availability. We observe a few trends of how the methods have progressed over the time by analyzing the table in the bottom-up fashion. The newer methods include more inputs, they rely more on machine-learning algorithms to implement the predictive model, and the most recent method are based on an ensemble of multiple disorder predictors. The most commonly used inputs include various propensities of the amino acids, evolutionary profiles in the form of the position specific scoring matrix (PSSM) and multiple alignments, predicted secondary structure, solvent accessibility, and flexibility, and most recently predicted globular domains and torsion angles. The two dominant machine-learning algorithms that are utilized are neural networks and support vector machines.

Table 1. Summary of the considered disorder predictors. The prediction methods are sorted by the year of publication in the descending order.

Prediction method		Input information					Prediction method		standalone program (SP); web server (WS); upon request (UR)	URL of the web server or standalone implementation
name	Year of publication	Amino acid type, propensity or position	PSSM profile	Secondary structure prediction	Solvent accessibility prediction	Other inputs	Algorithm used	Meta prediction		
MFDp	2010	X	X	X	X	Predicted flexibility, globular domains, torsion angles, and disorder	Support vector machine	X	WS	http://biomine-ws.ece.ualberta.ca/MFDp.html
PONDR-FIT	2010					Predicted disorder	Neural network	X	UR	http://www.disprot.org/predictors.php
MD	2009	X	X	X	X	Chain length, predicted disorder	Neural network	X	SP+WS	http://www.rostlab.org/services/md/
DISOCLUST	2008					Alignment of predicted folds	Scoring function		SP+WS	http://www.reading.ac.uk/bioinf/DISOclust/
Norsnet	2007	X	X	X	X	Predicted flexibility	Neural network		SP+WS	http://cubic.bioc.columbia.edu/newwebsite/services/NORSp/
Ucon	2007					Predicted residue-residue contacts	Scoring function		WS	http://cubic.bioc.columbia.edu/services/ucon/
ProfBval	2006	X	X	X	X	Chain length	Neural network		SP+WS	http://cubic.bioc.columbia.edu/services/profbval/
Spritz	2006		X	X			Support vector machine		WS	http://distill.ucd.ie/spritz/
VSL2B	2006	X					Support vector machine		SP+WS	http://www.ist.temple.edu/disprot/Predictors.html
DISpro	2005		X	X	X		Neural network		SP+WS	http://scratch.proteomics.ics.uci.edu/
FoldIndex	2005	X					Scoring function		WS	http://bioportal.weizmann.ac.il/fldbin/findex
IUPred	2005	X				Interaction energy	Scoring function		SP+WS	http://iupred.enzim.hu/
RONN	2005					Sequence alignment	Neural network		SP+WS	http://www.strubi.ox.ac.uk/RONN
DISOPRED2	2004	X	X				Support vector machine		SP+WS	http://bioinf.cs.ucl.ac.uk/disopred/
DisEMBL	2003	X					Neural network		SP+WS	http://dis.embl.de/
GlobPlot	2003	X					Scoring function		SP+WS	http://globplot.embl.de/

Benchmark dataset and prediction protocol

We test the considered methods on a large dataset that combines the CASP-like and the DisProt disorder annotations. This dataset was originally developed to validate the MFDp meta-predictor [[48]]. The sequences were obtained from the PDB and the DisProt databases. The PDB sequences were filtered using the culled PDB list [[71]] to extract a high-quality and low sequence identity subset. We selected sequences that have structures with R-factor < 0.2 and resolution < 2.0Å, and that are characterized by sequence identity < 25%. Since most protein chains in PDB are completely ordered we kept randomly selected 20% of the fully structured proteins. We extracted the entire set of 523 proteins from the release 4.9 of the DisProt and we merged them with the PDB chains. The combined set was filtered at 25% sequence identity as follows. For a given pair of sequences that share >25% identity, we remove the chain that has fewer disordered residues (a less complete annotation). Among the remaining 514 chains we removed 20 for which some of the considered methods failed to produce predictions. Specifically, MD predictor could not produce results for chains 15, 177, 531, and 277 from the DisProt, and Dispro does not predict chains with over 1500 residues, which resulted in excluding proteins 81, 102, 122, 181, 228, 238, 269, 348, 440, 467, 517, 519, 557, 560, 573, and 591 from the DisProt. The resulting dataset includes 289 chains from DisProt and 205 from PDB, among which 248 have long segments with 30 or more consecutive disordered residues and 246 are without such segments. The original dataset with the 514 chains is freely available at http://biomine.ece.ualberta.ca/MFDp.html.

All considered methods, except MFDp, were tested on the entire benchmark dataset, i.e., we supplied the protein chains, one by one, into their web servers or we predicted them using the standalone program. The MFDp's predictions were performed using five-fold cross validation on this benchmark set, i.e., they were taken directly from [[48]]. This means that in the case of MFDp the training chains (chains that were used to compute the predictive model) share below 25% identity with the test sequences, while the other methods likely use more similar chains in their training datasets.

Evaluation criteria

The assessment of the per-residue predictions uses the same criteria as in the CASP experiments [[16],[17]]. The same as in the CASP8, we discard the native disordered regions with 3 or fewer residues (private correspondence with authors) [[17]], i.e., these residues are ignored when computing the quality measures. These predictions take two forms: 1) the binary value that defines whether a given residue is disordered or not; and 2) the real value that quantifies probability of disorder. The binary predictions were assessed using four measures:

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

Sensitivity = SENS = $TP / (TP + FN) = TP / N_{disorder}$

Specificity = SPEC = $TN / (TN + FP) = TN / N_{order}$

S_W = $(W_{disorder}*TP - W_{order}*FP + W_{order}*TN - W_{disorder}*FN)$ /
$(W_{disorder}* N_{disorder} + W_{order}* N_{order})$

where TP is the number of true positives (number of correctly predicted disordered residues), FP denotes false positives (number of ordered residues that were predicted as disordered), TN denotes true negatives (number of correctly predicted ordered residues), FN stands for false negatives (number of disordered residues that were predicted ordered), $W_{disorder}$ is the percentage of ordered residues, W_{order} is the percentage of disordered residues, and N_{order} and $N_{disorder}$ are the total number of ordered and disordered residues, respectively. The S_w and MCC values range between -1 and 1 and they are equal zero when all residues are predicted to be ordered or disordered. Higher values of the above four measures indicate better predictions. The receiver operating characteristic (ROC) curve was used to examine the predicted probabilities. For each value of probability p achieved by a given method (between 0 and 1), all the residues with probability equal or greater than p are set as disordered, and all other residues are set as ordered. Next, the TP-rate = TP / (TP + FN) and the FP-rate = FP / (FP + TN) are calculated and we use the area under the curve (AUC) to quantify the predictive quality. We note that four of the considered predictors, namely FoldIndex, Spritz, DisEMBL, and GlobPlot do not provide the probabilities (their outputs is only binary, or in case of the GlobPlot there are too few probability values, i.e., it only outputs probability equal 0, 0.5, and 1), and thus we were unable to compute their AUC values.

The segment-level evaluation uses the segment overlap (SOV) measure that was originally developed to quantify the quality of the prediction of secondary structure segments [[51]]. We compute the SOV values to compare the amount of overlap between the segments formed by the binary per-residue predictions and the native disorder segments; we note that the native segments are at least 4 residues long.

The sequence-level evaluations concern two predictive objectives, prediction of proteins with the long disordered segments and prediction of the disorder content. We compute the MCC, sensitivity, specificity and AUC to evaluate predictions of proteins with long disordered segments. Similarly as in [[46], [48]], each protein is categorized into one of two classes, protein with or without at least one long disordered segment that has at least 30 consecutive disordered residues. These sequence-level annotations, which are generated using the binary residue-level predictions, are compared against the sequence-level annotations computed from the native disorder. In the case of the AUC measure, we threshold the residue-level predictions to re-compute the sequence-level annotations.

The prediction of the disorder content compares the normalized amount of the disordered residues in the protein chain, i.e., number of disordered residues divided by the number of all residues in the chain, between the predicted and the native disorder annotations. Following studies on the prediction of the content of secondary structures [[72]-[74]], the disorder content predictions are evaluated using three measures:

$$\text{Pearson Correlation Coefficient (PCC)} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

$$\text{Mean Squared Error (MSE)} = \sum_{i=1}^{n} (y_i - x_i)^2 \Big/ n$$

$$\text{Mean Absolute Error (MAE)} = \sum_{i=1}^{n} \left|y_i - x_i\right| \Big/ n$$

where *n* is the total number of chains, and y_i and x_i are the native and the predicted disorder content for the i^{th} protein chain, respectively. The higher values of PCC and the lower values of MSE and MAE correspond to better predictions.

Statistical significance

We also evaluate statistical significance of the differences between each pair of the considered predictors. We performed these tests for the S_W and MCC measures for the binary predictions, the AUC measure for the predicted probabilities, the SOV for the segment-level evaluations, the MCC and AUC for the prediction of chains with long disordered segments, and MAE and PCC for the prediction of disorder content. We contrast a given pair of methods by comparing their results over 100 datasets with 100 chains each that were selected at random from the benchmark dataset; the same randomized datasets were used to compare all pairs of methods. We first evaluate whether these 100 values computed for a given predictor follow normal distribution, as tested using Shapiro-Wilk test [[75]] at the 0.05 significance. If the values are normal for both of the being compared predictors then we utilize the paired t-test (we compare results of the two methods on the same datasets); otherwise we use the non-parametric Wilcoxon rank sum test [[76]]. We annotate the significance of the differences at the 0.05 and the 0.001 levels. A pair of methods for which the *p*-value is greater that 0.05 is assumed to be equivalent, i.e., the difference between these two methods is not significant.

Results and Discussion

The results for the 19 considered disorder predictors, including two versions of IUPred and three versions of DisEMBL, on the benchmark datasets with 494 proteins are summarized in Table 2.

Evaluation of residue and segment level predictions

The values of the S_w measure, which was used as the main residue-level evaluation criterion during the most recent completed CASP8 experiment [[17]], vary between 0.17 and 0.52. Several of the methods with the lower scores are designed to target specific types of disorder, like DisEMBL that has three versions that specifically address prediction of hot-loops, coils and loops, and REMARK 465 based annotations of disorder, and NORSnet that targets unstructured loops, which explains their lower overall performance when applied to predict all types of disorder. The GlobPlot and FoldIndex use simple scoring functions that are based on propensities of residues to be in random coil conformation and based on hydrophobicity and charge of residues, respectively, which is the likely reason for their relatively low predictive quality. The ProfBval [[30]] is designed to predict flexibility, expressed as B-factors, rather than disorder, which may explain why this method obtains relatively low S_w. Both, ProfBval and DisEMBL-C over-predict the disorder, which results in the low specificity and high sensitivity. We hypothesize that the moderate quality of predictions of Ucon stems from the fact that it uses only the information about the density of the predicted residue-residue contacts. The top-performing methods, which obtain the $S_w > 0.45$, MCC > 0.41, and AUC > 0.79 include MFDp, MD, and PONDR-FIT. These are the most recent predictors and the likely explanation for their superior predictive quality is the fact that they are based on a consensus of multiple disorder predictors. This is also in agreement with the results of the recent CASP experiments that show that consensus-based solutions generate favorable quality of the predictions [[17]]. We also note the high quality of the predictions generated by an older VSL2B method, which obtains high S_w and AUC, and moderately high MCC. The downside of this approach is the fact that it has relatively low specificity, which means that it moderately over-predicts the disorder. The ROC curves for the VSL2B and the three methods with the highest AUC (MFDp, MD, and PONDR-FIT) are shown in Figure 1A. They reveal that MD provides the highest TP-rates for the low FP-rates up to about 0.13, while MFDp outperforms the other methods for the larger values of the FP-rates and for the TP-rates of above 0.6. The above suggests that these two top-performing solutions complement each other. Although they are both based on a consensus, they utilize a different set of the base disorder predictors and different architectures and prediction algorithms (see Table 1), which could result in complementarity.

Table 2. Summary of the results for predictions performed with the 19 considered disorder predictors, including two versions of IUPred and three versions of DisEMBL, on the benchmark datasets with 494 proteins. The results include the per-residue evaluations (columns 2 to 6), per-segment evaluation (column 7), and per-sequence evaluation of prediction of proteins with long, ≥ 30 consecutive residues, disordered segments (columns 8 to 11), and prediction of disorder content (columns 12 to 14). The methods are sorted in the descending order by their per residue S_w values and the best result, for each quality index, is shown in bold font. The N/A in the two AUC columns means that the corresponding method does not output the real-value disorder probability and thus AUC could not be computed.

Predictor	Disorder prediction per-residue					Disorder prediction per-segment	Prediction of proteins with long disordered segments				Prediction of disorder content		
	Sw	AUC	SENS	SPEC	MCC	SOV	MCC	SENS	SPEC	AUC	MSE	MAE	PCC
MFDp	**0.515**	**0.821**	0.746	0.768	**0.451**	60.974	0.525	0.815	0.707	0.853	0.069	0.167	**0.622**
MD	0.486	**0.821**	0.673	0.813	0.444	45.055	0.484	0.661	0.817	0.796	0.086	0.193	0.606
VSL2B	0.473	0.793	0.774	0.698	0.401	**62.982**	**0.589**	0.883	0.695	**0.869**	0.076	0.207	0.611
PONDR-FIT	0.452	0.790	0.631	0.821	0.419	54.520	0.565	0.706	0.854	0.851	0.058	0.155	0.619
DISOPRED2	0.447	0.781	0.647	0.800	0.406	49.892	0.506	0.690	0.813	0.811	0.071	0.164	0.543
IUPredL	0.422	0.784	0.581	0.841	0.405	33.322	0.500	0.581	0.894	0.821	0.073	0.166	0.551
RONN	0.418	0.764	0.664	0.754	0.368	50.544	0.525	0.810	0.711	0.828	0.068	0.184	0.559
DISOCLUST	0.411	0.775	0.779	0.632	0.344	59.517	0.463	0.831	0.622	0.811	0.105	0.257	0.529
IUPredS	0.387	0.781	0.522	0.866	0.389	46.159	0.508	0.617	0.874	0.819	0.063	**0.151**	0.585
NORSnet	0.361	0.738	0.532	0.829	0.347	20.882	0.477	0.500	0.931	0.791	0.097	0.192	0.461
Ucon	0.340	0.741	0.554	0.787	0.313	26.522	0.521	0.512	0.955	0.842	**0.057**	0.163	0.617
FoldIndex	0.319	N/A	0.602	0.717	0.278	36.745	0.346	0.867	0.447	N/A	0.089	0.225	0.504
Spritz	0.307	N/A	0.494	0.812	0.293	36.968	0.394	0.411	0.927	N/A	0.096	0.187	0.279
DisEMBL-R	0.252	N/A	0.316	0.936	0.323	35.144	0.377	0.351	0.951	N/A	0.088	0.172	0.516
DISpro	0.243	0.775	0.303	**0.940**	0.318	31.660	0.402	0.371	**0.955**	0.833	0.092	0.179	0.499
DisEMBL-H	0.227	N/A	0.435	0.792	0.216	44.483	0.366	0.601	0.760	N/A	0.074	0.198	0.453
ProfBval	0.222	0.697	**0.835**	0.387	0.196	48.895	0.377	0.718	0.659	0.753	0.227	0.437	0.346
GlobPlot	0.179	N/A	0.353	0.826	0.182	33.242	0.338	0.452	0.858	N/A	0.090	0.198	0.291
DisEMBL-C	0.174	N/A	0.760	0.414	0.150	54.219	0.135	**0.927**	0.159	N/A	0.223	0.428	0.222

The correlation between the values of two measures for the binary prediction, S_w and MCC, obtained across the results of the tested predictors, see Table 2, is relatively high and equals 0.94. This suggests that these two measures are closely related. On the other hand, the correlations between the AUC that evaluates the predicted probabilities and S_w and MCC are 0.81 and 0.91, respectively. This shows that MCC better captures the relation between the binary and real-valued predictions when compared with S_w. This is perhaps due to the fact that S_w is designed to concentrate on the prediction of native disordered residues, while putting substantially smaller emphasis on the correct prediction of the ordered residues. This is evidenced through the use of the $W_{disorder}$ and W_{order} weights and the fact that they differ by close to an order of magnitude.

The statistical significance of the differences between all pairs of the considered predictors for the S_w and MCC measures is summarized in Table 3. We observe that MFDp generates the best and significantly better S_w values when compared with the other 18 predictors. The runner up and equivalent with each other MD and VSL2B have S_w values that significantly outperform the other 16 methods. The results based on the MCC measure indicate that MFDp and MD, which have comparable MCC, significantly improve over the remaining 17 methods. The runner up group of equivalent methods includes PONDR-FIT, DISOPRED2 and IUPredL, which are followed by VSL2B and IUPredS. The above seven predictors significantly outperform the other 12 methods when evaluated using the MCC. Table 4 shows the statistical significance of the differences for the predictions of probabilities that are scored using AUC. Once again, the MFDp and MD, which obtain very similar AUC of about 0.82 on the entire benchmark dataset, are shown not to be significantly different between each other and significantly better than the other 11 methods. Six methods including FoldIndex, Spritz, GlobPlot and DisEMBL-R, DisEMBL-H, and DisEMBL-C could not be evaluated since they produce only the binary predictions; GlobPlot also provides probabilities but they are limited to only three values, 0, 0.5, and 1. The second-best group of equivalent methods includes PONDR-FIT and VSL2B, and they significantly outperform the remaining 9 predictors, except for the IUPredL which is similar to PONDR-FIT.

The segment level evaluation using SOV measure reveals that the average overlap between the predicted and the native disordered segments ranges between 21% and 63%. The methods that that have the SOV > 50 include RONN, DisEMBL-C, PONDR-FIT, DISOCLUST, MFDp, and VSL2B. To compare, the SOV of recent secondary structure predictors (the average segment overlap for coils, strand, and helix segments) is around 0.8 [[77]]. This shows that the prediction of the disordered segments requires further work and should be considered when evaluating the current and future predictors. The statistical significance of the differences between all pairs of the 19 predictors that are evaluated utilizing SOV is provided in Table 4. The VSL2B method that achieves SOV = 63 is significantly better than the second best MFDp that has SOV = 61, which in turn significantly outperforms the remaining 17 methods. The third best DISOCLUST is also shown to provide significantly higher SOV, which is close to 60, when compared with the lower ranked methods that obtain SOV < 55.

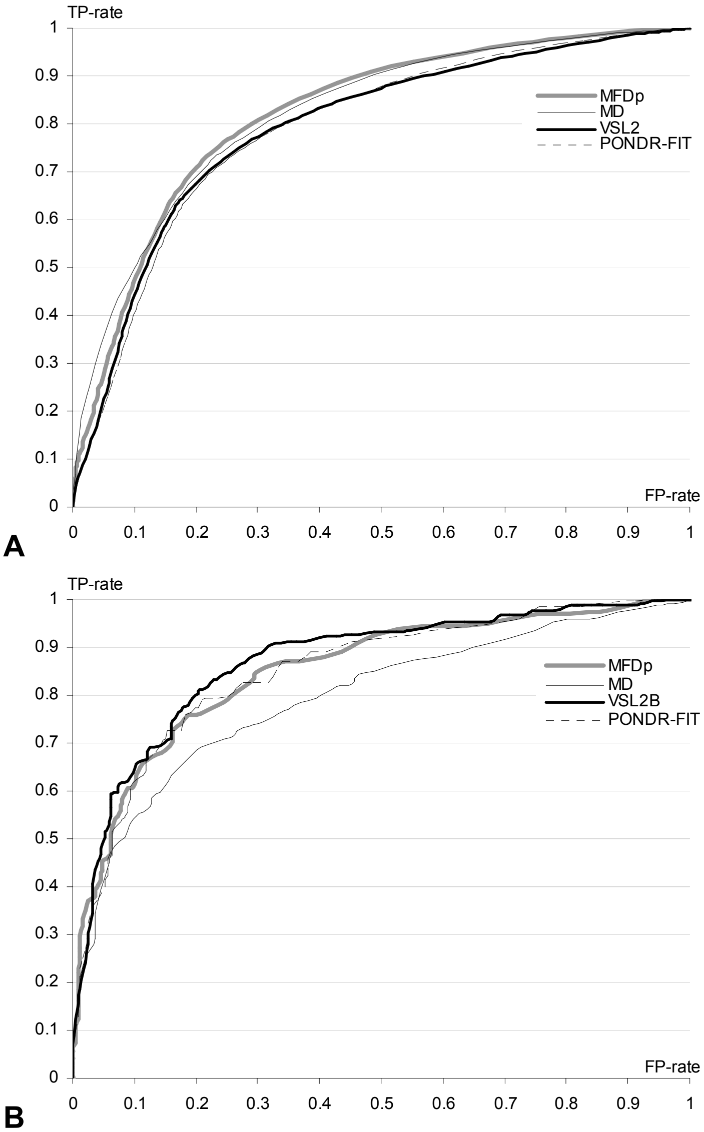

Figure 1. The ROC curves of the four predictors, MFDp (in blue), PONDR-FIT (in green), MD (in red), and VSL2B (in black) for the prediction of (A) disordered residues; and (B) proteins with long disorder segments. The *x*-axis and *y*-axis show the FP- and TP-rates, respectively.

Table 3. The results of the statistical significance tests for the two measures for the binary, residue-level predictions, S_w (results are shown in the upper triangle) and MCC (in the lower triangle). The tests compare all pairs of predictors, among the considered 19 methods, based on results obtained on 100 datasets with 100 chains each that were selected at random from the benchmark dataset; the same randomized datasets were used to compare all pairs of methods. For measures that are normal (tested using Shapiro-Wilk test at the 0.05 significance), we use paired t-test; otherwise we use the non-parametric Wilcoxon rank sum test. The ++ / -- indicate that a method in a given row is significantly better / worse than the method in a given column with $p < 0.001$; the + / - denote the same but when $p \leq 0.05$ and ≥ 0.001. Otherwise, = is used to denote that the difference between a given pairs of methods is not significant, i.e., $p > 0.05$.

Predictor	MFDp	MD	VSL2B	PONDR-FIT	DISO PRED2	IUPredL	RONN	DISO CLUST	IUPredS	NORSnet	Ucon	FoldIndex	Spritz	DisEMBL-R	DISpro	DisEMBL-H	ProfBval	GlobPlot	DisEMBL-C
MFDp		++	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++
MD	=		=	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++
VSL2B	--	--		+	++	++	++	++	++	++	++	++	++	++	++	++	++	++	++
PONDR-FIT	--	-	+		=	++	++	++	++	++	++	++	++	++	++	++	++	++	++
DISOPRED2	--	--	=	-		+	+	++	++	++	++	++	++	++	++	++	++	++	++
IUPredL	--	--	=	=	=		=	+	++	++	++	++	++	++	++	++	++	++	++
RONN	--	--	--	--	--	--		=	++	++	++	++	++	++	++	++	++	++	++
DISOCLUST	--	--	--	--	--	--	--		+	++	++	++	++	++	++	++	++	++	++
IUPredS	--	--	=	--	=	=	++	++		++	++	++	++	++	++	++	++	++	++
NORSnet	--	--	--	--	--	--	-	=	--		+	++	++	++	++	++	++	++	++
Ucon	--	--	--	--	--	--	--	--	--	--		+	++	++	++	++	++	++	++
FoldIndex	--	--	--	--	--	--	--	--	--	--	--		+	++	++	++	++	++	++
Spritz	--	--	--	--	--	--	--	--	--	--	-	=		++	++	++	++	++	++
DisEMBL-R	--	--	--	--	--	--	--	-	--	-	+	++	++		=	++	++	++	++
DISpro	--	--	--	--	--	--	--	-	--	-	=	++	+	=		+	++	++	++
DisEMBL-H	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--		=	++	++
ProfBval	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--		++	++
GlobPlot	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	-		=
DisEMBL-C	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	

Table 4. The results of the statistical significance tests for the real-valued, residue-level disorder probability, AUC (results are shown in the upper triangle) and the segment-level SOV (in the lower triangle). The tests compare all pairs of predictors, among the considered 19 methods, based on results obtained on 100 datasets with 100 chains each that were selected at random from the benchmark dataset; the same randomized datasets were used to compare all pairs of methods. For measures that are normal (tested using Shapiro-Wilk test at the 0.05 significance), we use paired t-test; otherwise we use the non-parametric Wilcoxon rank sum test. The ++ / -- indicate that a method in a given row is significantly better / worse than the method in a given column with $p < 0.001$; the + / - denote the same but when $p \leq 0.05$ and ≥ 0.001. Otherwise, = is used to denote that the difference between a given pairs of methods is not significant, i.e., $p > 0.05$. The N/A denotes that the AUC results are not available since the corresponding predictors, including FoldIndex, Spritz, GlobPlot and DisEMBL, do not generate the probabilities.

Predictor	MFDp	MD	VSL2B	PONDR-FIT	DISO PRED2	IUPredL	RONN	DISO CLUST	IUPredS	NORSnet	Ucon	FoldIndex	Spritz	DisEMBL-R	DISpro	DisEMBL-H	ProfBval	GlobPlot	DisEMBL-C
MFDp		=	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	N/A	N/A	N/A	+ +	N/A	+ +	N/A	N/A
MD	- -		+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	N/A	N/A	N/A	+ +	N/A	+ +	N/A	N/A
VSL2B	+ +	+ +		=	+	+	+ +	+ +	+	+ +	+ +	N/A	N/A	N/A	+ +	N/A	+ +	N/A	N/A
PONDR-FIT	- -	+ +	- -		+	=	+ +	+ +	+	+ +	+ +	N/A	N/A	N/A	+ +	N/A	+ +	N/A	N/A
DISOPRED2	- -	+ +	- -	- -		=	+ +	+	=	+ +	+ +	N/A	N/A	N/A	=	N/A	+ +	N/A	N/A
IUPredL	- -	- -	- -	- -	- -		+ +	+	=	+ +	+ +	N/A	N/A	N/A	=	N/A	+ +	N/A	N/A
RONN	- -	+ +	- -	- -	+ +	+ +		=	- -	+ +	+ +	N/A	N/A	N/A	-	N/A	+ +	N/A	N/A
DISOCLUST	- -	+ +	- -	+ +	+ +	+ +	+ +		-	+ +	+ +	N/A	N/A	N/A	=	N/A	+ +	N/A	N/A
IUPredS	- -	=	- -	- -	- -	+ +	- -	- -		+ +	+ +	N/A	N/A	N/A	=	N/A	+ +	N/A	N/A
NORSnet	- -	- -	- -	- -	- -	- -	- -	- -	- -		=	N/A	N/A	N/A	- -	N/A	+ +	N/A	N/A
Ucon	- -	- -	- -	- -	- -	- -	- -	- -	- -	+ +		N/A	N/A	N/A	- -	N/A	+ +	N/A	N/A
FoldIndex	- -	- -	- -	- -	- -	+ +	- -	- -	- -	+ +	+ +		N/A	N/A	N/A	N/A	N/A	N/A	N/A
Spritz	- -	- -	- -	- -	- -	+ +	- -	- -	- -	+ +	+ +	=		N/A	N/A	N/A	N/A	N/A	N/A
DisEMBL-R	- -	- -	- -	- -	- -	+ +	- -	- -	- -	+ +	+ +	- -	- -		N/A	N/A	N/A	N/A	N/A
DISpro	- -	- -	- -	- -	- -	- -	- -	- -	- -	+ +	+ +	- -	- -	- -		N/A	+ +	N/A	N/A
DisEMBL-H	- -	=	- -	- -	- -	+ +	- -	- -	-	+ +	+ +	+ +	+ +	+ +	+ +		N/A	N/A	N/A
ProfBval	- -	+ +	- -	- -	=	+ +	- -	- -	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +		N/A	N/A
GlobPlot	- -	- -	- -	- -	- -	=	- -	- -	- -	+ +	+ +	- -	- -	- -	+ +	- -	- -		N/A
DisEMBL-C	- -	+ +	- -	=	+ +	+ +	+ +	- -	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	+ +	

Table 5. The results of the statistical significance tests for the prediction of proteins with long, ≥ 30 consecutive residues, disordered segments, measured using AUC (results are shown in the upper triangle) and MCC (in the lower triangle). The tests compare all pairs of predictors, among the considered 19 methods, based on results obtained on 100 datasets with 100 chains each that were selected at random from the benchmark dataset; the same randomized datasets were used to compare all pairs of methods. For measures that are normal (tested using Shapiro-Wilk test at the 0.05 significance), we use paired t-test; otherwise we use the non-parametric Wilcoxon rank sum test. The ++ / -- indicate that a method in a given row is significantly better / worse than the method in a given column with p < 0.001; the + / - denote the same but when p ≤ 0.05 and ≥ 0.001. Otherwise, = is used to denote that the difference between a given pairs of methods is not significant, i.e., p > 0.05. The N/A denotes that the AUC results are not available since the corresponding predictors do not generate the probabilities.

Predictor	MFDp	MD	VSL2B	PONDR-FIT	DISO PRED2	IUPredL	RONN	DISO CLUST	IUPredS	NORSnet	Ucon	FoldIndex	Spritz	DisEMBL-R	DISpro	DisEMBL-H	ProfBval	GlobPlot	DisEMBL-C
MFDp		++	--	=	++	++	++	++	++	++	+	N/A	N/A	N/A	++	N/A	++	N/A	N/A
MD	-		--	--	--	--	--	-	--	=	--	N/A	N/A	N/A	--	N/A	++	N/A	N/A
VSL2B	++	++		++	++	++	++	++	++	++	++	N/A	N/A	N/A	++	N/A	++	N/A	N/A
PONDR-FIT	++	++	-		++	++	++	++	++	++	+	N/A	N/A	N/A	++	N/A	++	N/A	N/A
DISOPRED2	=	=	--	--		=	-	=	=	++	--	N/A	N/A	N/A	--	N/A	++	N/A	N/A
IUPredL	-	=	--	--	=		=	=	=	++	--	N/A	N/A	N/A	--	N/A	++	N/A	N/A
RONN	=	+	--	--	=	=		+	+	++	-	N/A	N/A	N/A	-	N/A	++	N/A	N/A
DISOCLUST	--	=	--	--	--	-	--		=	++	--	N/A	N/A	N/A	--	N/A	++	N/A	N/A
IUPredS	-	=	--	--	=	=	-	+		++	--	N/A	N/A	N/A	--	N/A	++	N/A	N/A
NORSnet	--	=	--	--	-	=	-	=	=		--	N/A	N/A	N/A	--	N/A	++	N/A	N/A
Ucon	=	+	--	--	=	=	=	++	+	++		N/A	N/A	N/A	=	N/A	++	N/A	N/A
FoldIndex	--	--	--	--	--	--	--	--	--	--	--		N/A	N/A	N/A	N/A	N/A	N/A	N/A
Spritz	--	--	--	--	--	--	--	--	--	--	--	++		N/A	N/A	N/A	N/A	N/A	N/A
DisEMBL-R	--	--	--	--	--	--	--	--	--	--	--	++	=		N/A	N/A	N/A	N/A	N/A
DISpro	--	--	--	--	--	--	--	--	--	--	--	++	=	+		N/A	++	N/A	N/A
DisEMBL-H	--	--	--	--	--	--	--	--	--	--	--	+	-	=	--		N/A	N/A	N/A
ProfBval	--	--	--	--	--	--	--	--	--	--	--	++	=	=	-	=		N/A	N/A
GlobPlot	--	--	--	--	--	--	--	--	--	--	--	=	--	--	--	-	--		N/A
DisEMBL-C	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	

Table 6. The results of the statistical significance tests for the prediction of disorder content measured using PCC (results are shown in the upper triangle) and MAE (in the lower triangle). The tests compare all pairs of predictors, among the considered 19 methods, based on results obtained on 100 datasets with 100 chains each that were selected at random from the benchmark dataset; the same randomized datasets were used to compare all pairs of methods. For measures that are normal (tested using Shapiro-Wilk test at the 0.05 significance), we use paired t-test; otherwise we use the non-parametric Wilcoxon rank sum test. The ++ / -- indicate that a method in a given row is significantly better / worse than the method in a given column with $p < 0.001$; the + / - denote the same but when $p \leq 0.05$ and ≥ 0.001. Otherwise, = is used to denote that the difference between a given pairs of methods is not significant, i.e., $p > 0.05$.

Predictor	MFDp	MD	VSL2B	PONDR-FIT	DISO PRED2	IUPredL	RONN	DISO CLUST	IUPredS	NORSnet	Ucon	FoldIndex	Spritz	DisEMBL-R	DISpro	DisEMBL-H	ProfBval	GlobPlot	DisEMBL-C
MFDp		=	=	=	++	++	++	++	+	++	=	++	++	++	++	++	++	++	++
MD	--		=	-	++	++	++	++	=	++	=	++	++	++	++	++	++	++	++
VSL2B	--	--		=	++	++	++	++	=	++	=	++	++	++	++	++	++	++	++
PONDR-FIT	++	++	++		++	++	++	++	+	++	=	++	++	++	++	++	++	++	++
DISOPRED2	=	++	++	--		=	=	=	--	++	--	++	++	=	+	++	++	++	++
IUPredL	=	++	++	--	=		=	+	-	++	--	++	++	+	++	++	++	++	++
RONN	--	++	++	--	--	--		+	-	++	--	++	++	+	++	++	++	++	++
DISOCLUST	--	--	--	--	--	--	--		--	++	--	+	++	=	=	++	++	++	++
IUPredS	++	++	++	=	++	++	++	++		++	=	++	++	++	++	++	++	++	++
NORSnet	--	=	++	--	--	--	-	++	--		--	-	++	--	-	=	++	++	++
Ucon	=	++	++	--	=	=	++	++	--	++		++	++	++	++	++	++	++	++
FoldIndex	--	--	--	--	--	--	--	++	--	--	--		++	-	=	++	++	++	++
Spritz	--	=	++	--	--	--	=	++	--	=	--	++		--	--	--	--	=	++
DisEMBL-R	-	++	++	--	--	-	++	++	--	++	--	++	++		=	++	++	++	++
DISpro	--	++	++	--	--	--	=	++	--	++	--	++	+	-		++	++	++	++
DisEMBL-H	--	=	++	--	--	--	--	++	--	=	--	++	-	--	--		++	++	++
ProfBval	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--		++	++
GlobPlot	--	-	+	--	--	--	--	++	--	-	--	++	--	--	--	=	++		++
DisEMBL-C	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	+	--	

Evaluation of the predictions of proteins with long disordered segments

We evaluate whether the binary residue-level predictions could be used to accurately find proteins that include at least one long, with 30 or more consecutive residues, disordered segment. The results shown in Table 2 reveal that some of the methods, including VSL2B, PONDR-FIT, MFDp, RONN, Ucon, IUPredS, DISOPRED2, and IUPredL, provide high quality predictions with MCC > 0.5 and AUC > 0.81. Ucon has the best specificity of 0.96 with sensitivity equal 0.51. This means that is can correctly find 51% of the proteins with the long segments with a low false positive rate. On the other hand, VSL2B provides the second highest sensitivity of 0.88 combined with moderate specificity of 0.71. This means that is accurately predicts 88% of the chains with the long segments (219 out of 248 such chains) but as a trade-off for misclassifying 75 proteins without the long segments. The DisEMBL-C which has the highest sensitivity also has the lowest specificity, which means that is predicts long disordered segments in almost all proteins. Potentially the best trade-off is achieved by PONDR-FIT that obtains 0.71 sensitivity, i.e., it correctly predicts 175 out of 248 proteins with the long segments, combined with 0.85 specificity, i.e., it incorrectly classifies only 36 out of 246 proteins with no long segments. The ROC curves for the three predictors with the highest AUC values (VSL2B, MFDp, and PONDR-FIT) and the remaining consensus-based MD (for consistency with the Figure 1A) are shown in Figure 1B. They reveal that VSL2B provides favorable TP-rates for virtually the entire range of the FP-rates and that the runner-up MFDp and PONDR-FIT have overlapping curves. Analysis of the statistical significance of the differences in AUC and MCC between all pairs of the considered predictors is given in Table 5. The VSL2B, which obtains the highest AUC = 0.87 and the highest MCC = 0.59 is shown to be significantly better than all other included methods. Based on the AUC values, the runner up methods, MFDp and PONDR-FIT, are not significantly deferent with each other and they significantly outperform the remaining predictors. Using the MCC values, the second best PONDR-FIT provides significantly improved predictions when compared with all methods, except for VSL2B; the third cluster of equivalent methods includes MFDp, RONN, Ucon and DISOPRED2.

Evaluation of the predictions of disorder content

We also investigate whether the binary residue-level predictions could be used to estimate the overall, per-chain, amount of disorder. We measure the differences between the native and the predicted disorder content, using MAE and MSE, and the correlation between these two contents, see Table 2. Well-performing methods should be characterized by low differences and high correlations. We observe that the average, over the entire benchmark dataset, differences expressed using MAE range between 15.1% and 43.7%. The two methods that over-predict the residue-level disorder, namely ProfBval and DisEMBL-C, also obtain the largest errors for the prediction of the disorder content. The methods that obtain errors below 17% include IUPredS, PONDR-FIT, Ucon, DISOPRED2, IUPredL, and MFDp. The correlations also vary widely between 0.22 for DisEMBL-C and 0.62 for MFDp. The methods that obtain correlation > 0.6 include MFDp, PONDR-FIT, Ucon, VSL2B, and MD; 12 of out the 19 methods achieve correlation > 0.5. The statistical significance of the differences between different predictors is shown in Table 6. The IUPredS that obtains the lowest MAE is statistically equivalent to PODR-FIT, and the mean absolute errors of these two methods are significantly lower when compared with the other 17 predictors. The second-best group of methods with respect to their MAE values includes Ucon, DISOPRED2, IUPredL, and MFDp. The MFDp, PONDR-FIT, Ucon, VSL2B, and MD are shown to achieve the best and comparable correlations between their predictions and the native disorder content. The best-performing MFDp and PONDR-FIT have significantly higher PCC values when compared with the other 14 methods, while the Ucon, VSL2B, and MD are comparable to IUPredS and they significantly outperform the remaining 13 methods.

Conclusions

Our empirical study that compares 19 disorder predictors on a dataset with close to 500 chains points out to several interesting observations:

- The top performing methods for the prediction of the residue-level disorder obtain S_w = 0.52, AUC = 0.82 and MCC = 0.45. While these results are substantially better than a random predictor, which would obtain S_w = 0, AUC = 0.5, and MCC = 0, there is a substantial margin for the future improvements.
- The recent consensus-based predictors like MFDp, MD and PONDR-FIT outperform other methods for the residue-level predictions. Interestingly, an older VSL2B method also produces high quality residue-level predictions.
- The VSL2B is shown to be superior to other methods for the prediction of the disordered segments, with the MFDp and DISOCLUST being the second- and third-best solutions.
- The top performing predictions of chains with the long disordered segments are generated by the VSL2B, followed by MFDp and PONDR-FIT.
- We also show that the disorder content computed by PONDR-FIT, MFDp and IUPredS significantly outperforms most of the other predictors.
- Some of the predictors, such as DisEMBL-R, DisEMBL-H, GlobPlot, and DISpro under-predict the disorder, i.e., their sensitivity is relatively low and lower than 0.45.
- On the other hand, a few of the studied method, including ProfBval and DisEMBL-C over-predict the number of the disordered residues (their specificity is below 0.45) and the overall disorder content.

The facts that there is no universally superior predictor and that the top-performing methods are complementary, as shown in Figure 1A and Table 2, support further work towards developing a new generation of consensus-based methods. We note that the existing and considered here consensus methods, including MFDp, PONDR-FIT and MD are shown to complement and outperform each other, depending on a given quality measure and the objective of the prediction. We believe that further progress requires a careful study that would investigate and quantify complementarity between existing predictors to perform an informed selection of the base methods to be included in the consensus. This is in contrast to the current developments that are based on a non-quantitative and sometimes ad-hock selection, e.g., based on availability, of the methods that are included in the ensemble.

We also observe that the errors in the disorder content prediction are relatively high, between 15% and 44%. These error levels and the fact that the content is relatively widely used [[4],[26],[47],[52]-[67],[78]] call for the development of specialized methods that would improve the content prediction. These methods would predict content directly from the sequence using both residue-level inputs as well as inputs that are aggregated over the entire chain, e.g., the total predicted coil content or number of predicted secondary structure segments. We believe that chain-level aggregation could reveal certain biases of some proteins to be mostly or fully disordered. This cannot be accomplished with the current methods that focus on the residue-level predictions using information from a relatively small, when compared to the chain length, window centered over the residue of interest.

References

[1] Uversky, V.N.; Dunker, A.K. Understanding protein non-folding. *Biochim Biophys Acta.*, **2010**, *1804*, 1231-64.

[2] Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; Ausio, J.; Nissen, M.S.; Reeves, R.; Kang, C.; Kissinger, C.R.; Bailey, R.W., Griswold, M.D.; Chiu, W.; Garner, E.C.; Obradovic, Z. Intrinsically disordered protein. *J Mol Graph Model.*, **2001**, *19*, 26-59.

[3] Uversky, V.N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.*, **2002**, *11*, 739-56.

[4] Dunker, A.K.; Obradovic, Z.; Romero, P.; Garner, E.C.; Brown, C.J. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform.*, **2000**, *11*, 161-71.

[5] Dunker, A.K.; Oldfield, C.J.; Meng, J.; Romero, P.; Yang, J.Y.; Chen, J.W.; Vacic, V.; Obradovic, Z.; Uversky, V.N. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, **2008**, *9*, S1-26.

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

[7] Uversky, V.N.; Oldfield, C.J.; Midic, U.; Xie, H.; Vucetic, S.; Xue, B.; Iakoucheva, L.M.; Obradovic, Z.; Dunker, A.K. Unfoldomics of human diseases: Linking protein intrinsic disorder with diseases. *BMC Genomics*, **2009**, *10*, S7-23.

[8] Cheng, Y.; LeGall, T.; Oldfield, C.J.; Mueller, J.P.; Van, Y.Y.; Romero, P.; Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Rational drug design via intrinsically disordered protein. *Trends Biotechnol*, **2006**, *24*, 435-42.

[9] Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **2000**, *1*, 415-27.

[10] Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. Sequence complexity of disordered protein. *Proteins*, **2001**, *42*, 38-48.

[11] Liu, J.; Tan, H.; Rost, B. Loopy proteins appear conserved in evolution. *J Mol Biol*, **2002**, *322*, 53-64.

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

[13] Dosztányi, Z.; Mészáros, B.; Simon, I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform*, **2010**, *11*, 225-43.

[14] Melamud, E.; Moult, J. Evaluation of disorder predictions in CASP5. *Proteins*, **2003**, *53*(S6), 561-65.

[15] Jin, Y.; Dunbrack, R.J. Assessment of disorder predictions in CASP6. *Proteins*, **2005**, *61*(S7), 167-75.

[16] Bordoli, L.; Kiefer, F.; Schwede, T. Assessment of disorder predictions in CASP7. *Proteins*, **2007**, *69*(S 8), 129-36.

[17] Noivirt-Brik, O.; Prilusky, J.; Sussman, J.L. Assessment of disorder predictions in CASP8. *Proteins*, **2009**, *77*(Suppl 9), 210-6.

[18] He, B.; Wang, K.; Liu, Y.; Xue, B.; Uversky, V.N.; Dunker, A.K. Predicting intrinsic disorder in proteins: an overview. *Cell Res*, **2009**, *19*(8), 929-49.

[19] Sickmeier, M.; Hamilton, J.A.; LeGall, T.; Vacic, V.; Cortese, M.S.; Tantos, A.; zabo, B.; Tompa, P.; Chen, J.; Uversky, V.N.; Obradovic, Z.; Dunker, A.K. DisProt: the database of disordered proteins. *Nucleic Acids Res*, **2007**, *35*, D786-93.

[20] Linding, R.; Russell, R.B.; Neduva, V.; Gibson, T.J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, **2003**, *31*, 3701-8.

[21] Dosztányi, Z.; Csizmok, V.; Tompa, P.; Simon, I. IUPred: web server for the pre-diction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **2005**, *21*, 3433-4.

[22] Prilusky, J.; Felder, C.E.; Zeev-Ben-Mordehai, T.; Rydberg, E.H.; Man, O.; Beckmann, J.S.; Silman, I.; Sussman, J.L. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **2005**, *21*, 3435-8.

[23] Schlessinger, A.; Punta, M.; Rost, B. Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **2007**, *23*, 2376-84.

[24] Jones, D.T.; Ward, J.J.; Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **2003**, *53*(S6), 573-8.

[25] Linding, R.; Jensen, L.J.; Diella, F.; Bork, P.; Gibson, T.J.; Russell, R.B. Protein disorder prediction: implications for structural proteomics. *Structure*, **2003**, *11*, 1453-9.

[26] Ward, J.J.; McGuffin, L.J.; Bryson, K.; Buxton, B.F; Jones, D.T. The DISOPRED server for the predic-tion of protein disorder. *Bioinformatics*, **2004**, *20*, 2138-9.

[27] Cheng, J.; Sweredoski, M.; Baldi, P. Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Know. Disc.*, **2005**, *11*, 213-222.

[28] Yang, Z.R.; Thomson, R.; McNeil, P.; Esnouf, R.M. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **2005**, *21*,3369-76.

[29] Vullo, A.; Bortolami, O.; Pollastri, G.; Tosatto, S.C. Spritz: a server for the predic-tion of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res*, **2006**, *34*, W164-8.

[30] Schlessinger, A.; Rost, B. Protein flexibility and rigidity predicted from sequence. *Proteins*, **2005**, *61*, 115-26.

[31] Schlessinger, A.; Yachdav, G.; Rost, B. PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **2006**, *22*, 891-3.

[32] Vucetic, S.; Brown, C.J.; Dunker, A.K.; Obradovic, Z. Flavors of protein disorder. *Proteins*, **2003**, *52*,573-84.

[33] Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Brown, C.J.; Dunker, A.K. Predicting intrinsic disorder from amino acid sequence. *Proteins*, **2003**, *53*(S6), 566-72.

[34] Obradovic, Z.; Peng, K.; Vucetic, S.; Radivojac, P.; Dunker, A.K. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **2005**, *61*(S7), 176-82.

[35] Peng, K.; Vucetic, S.; Radivojac, P.; Brown, C.J.; Dunker, A.K.; Obradovic, Z. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.*, **2005**, *3*, 35-60.

[36] Peng, K.; Radivojac, P.; Vucetic, S.; Dunker, A.K.; Obradovic, Z. Length-dependent prediction of protein intrinsic disorder, *BMC Bioinformatics*, **2006**, *7*, 208-224.

[37] Su, C.T.; Chen, C.Y.; Ou, Y.Y. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics*, **2006**, *7*, 319-334.

[38] Su, C.T.; Chen, C.Y.; Hsu, C.M. iPDA: integrated protein disorder analyzer. *Nucleic Acids Res*, **2007**, *35*, 465-72.

[39] Shimizu, K.; Muraoka, Y.; Hirose, S.; Tomii, K.; Noguchi, T. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics* 2007; 8:78-92.

[40] Hirose, S.; Shimizu, K.; Kanai, S.; Kuroda, Y.; Noguchi, T. POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **2007**, *23*, 2046-53.

[41] Schlessinger, A.; Liu, J.; Rost, B. Natively unstructured loops differ from other loops. *PLoS Comput Biol*, **2007**, *3*, e140-51.

[42] Yang, J.Y.; Yang, M.Q. Predicting protein disorder by analyzing amino acid sequence. *BMC Genomics*, **2008**, *9*, S8-15.

[43] Wang, L.; Sauer, U.H. OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. *Bioinformatics*, **2008**, *24*, 1401-2.

[44] Deng, X.; Eickholt, J.; Cheng, J. PreDisorder: Ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*, **2009**, *10*, 436-441.

[45] Ishida, T.; Kinoshita, K. Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **2008**, *24*, 1344-8.

[46] Schlessinger, A.; Punta, M.; Yachdav, G.; Kajan, L.; Rost, B. Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **2009**, *4*, e4433-42.

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

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

[49] Ishida, T.; Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence, *Nucleic Acids Res*, **2007**, *35*, W460-464.

[50] McGuffin, L.J. Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, **2008**, *24*,1798-804.

[51] Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **1999**, *34*, 220-3.

[52] Romero, P.; Obradovic, Z.; Kissinger, C.R.; Villafranca, J.E.; Garner, E.; Guilliot, S.; Dunker, A.K. Thousands of proteins likely to have long disordered regions, *Pac Symp Biocomput*, **1998**, 437-448.

[53] LeGall, T.; Romero, P.; Cortese, M.S.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder in the Protein Data Bank. *J. Biomol. Struct. Dyn.*, **2007**, *24*, 303-428.

[54] Dosztányi, Z.; Chen, J.; Dunker, A.K.; Simon, I.; Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J Proteome Res*, **2006**, *5*, 2985-95.

[55] Haynes, C.; Oldfield, C.J.; Ji, F.; Klitgord, N.; Cusick, M.E.; Radivojac, P.; Uversky, V.N.; Vidal, M.; Iakoucheva, L.M. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*., **2006**, *2*, e100-11.

[56] Liu, J.; Perumal, N.B.; Oldfield, C.J.; Su, E.W.; Uversky, V.N.; Dunker, A.K. Intrinsic disorder in transcription factors. *Biochemistry*, **2006**, *45*, 6773-6888.

[57] Goh, G.M.; Dunker, A.K.; Uversky, V.N. A comparative analysis of viral matrix proteins using disorder predictors. *Virology J*., **2008**, *5*, 126-135.

[58] Balázs, A.; Csizmok, V.; Buday, L.; Rakács, M.; Kiss, R.; Bokor, M.; Udupa, R.; Tompa, K.; Tompa, P. High levels of structural disorder in scaffold proteins as exemplified by a novel neuronal protein, CASK-interactive protein1. *FEBS J*., **2009**, *276*, 3744-56.

[59] Hébrard, E.; Bessin, Y.; Michon, T.; Longhi, S.; Uversky, V.N.; Delalande, F.; Dorsselaer, A.V.; Romero, P.; Walter, J.; Declerk, N.; Fargette, D. Intrinsic disorder in viral proteins genome-linked: Experimental and predictive analyses. *Virology J*., **2009**, *6*, 23-35.

[60] Hegyi, H.; Buday, L.; Tompa, P. Intrinsic structural disorder confers cellular viability on oncogenic fusion proteins. *PLoS Comput Biol*, **2009**, *5*, e1000552-61.

[61] Tompa, P.; Csermely, P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J*., **2004**, *18*, 1169-75.

[62] Xue, B.; Williams, R.W.; Oldfield, C.J.; Dunker, A.K.; Uversky, V.N. Archaic chaos: Intrinsically disordered proteins in Archaea. *BMC Systems Biol*., **2010**, *4*, S1-21.

[63] Tompa, P.; Dosztányi, Z.; Simon, I. Prevalent structural disorder in E. coli and S. cerevisiae proteomes. *J Proteome Res*, **2006**, *5*, 1996-2000.

[64] Xue, B.; Williams, R.W.; Oldfield, C.J.; Goh, G.M.; Dunker, A.K.; Uversky, V.N. Viral disorder or disordered viruses: Do viral proteins possess unique features? *Prot. Pept. Lett*., **2010**, *17*, 932-951.

[65] Vucetic, S.; Xie, H.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions, *J Proteome Res*, **2007**, *6*, 1899-1916.

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

[67] Xie, H.; Vucetic, S.; Iakoucheva, L.M.; Oldfield, C.J.; Dunker, A.K.; Obradovic, Z.; Uversky, V.N. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins, *J Proteome Res*, **2007**, *6*, 1917-1932.

[68] Slabinski, L.; Jaroszewski, L.; Rodrigues, A.P.; Rychlewski, L.; Wilson, I.A.; Lesley, S.A.; Godzik, A. The challenge of protein structure determination--lessons from structural genomics. *Protein Sci*., **2007**, *16*, 2472-82.

[69] Punta, M.; Love, J.; Handelman, S.; Hunt, J.F.; Shapiro, L.; Hendrickson, W.A.; Rost, B. Structural genomics target selection for the New York

consortium on membrane protein structure. *J Struct Funct Genomics*, **2009**, *10*, 255-68.

[70] Tompa, P.; Fuxreiter, M.; Oldfield, C.J.; Simon, I.; Dunker, A.K.; Uversky, V.N. Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **2009**, *31*, 328-35.

[71] Wang, G.; Dunbrack, R.L. PISCES: a protein sequence culling server. *Bioinformatics*, **2003**, *19*, 1589-1591.

[72] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. *Protein Pept. Lett.*, **2009**, *16*, 27-31.

[73] Chen, K.; Stach, W.; Homaeian, L.; Kurgan, L. iFC(2): an integrated web-server for improved prediction of protein structural class, fold type, and secondary structure content. *Amino Acids*, **2011**, *40*, 963-973.

[74] Homaeian, L.; Kurgan, L.; Ruan, J.; Cios, K.J.; Chen, K. Prediction of protein secondary structure content for the twilight zone sequences. *Proteins*, **2007**, *69*, 486-98.

[75] Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika*, **1965**, *52*, 591-611.

[76] Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, **1945**, *1*, 80-3.

[77] Montgomerie, S.; Cruz, J.A.; Shrivastava, S.; Arndt, D.; Berjanskii, M.; Wishart, D.S. PROTEUS2: a web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Res*, **2008**, *36*, W202-9.

[78] Mizianty, M.; Zhang, T.; Xue, B.; Zhou, Y.; Dunker, A.K.; Uversky, V.N.; Kurgan, L. In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinformatics*, **2011**, *12*, 245-297.