BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 201–208
doi:10.1093/bioinformatics/btu625

Structural bioinformatics

Advance Access publication September 21, 2014

Comprehensive large-scale assessment of intrinsic protein disorder

Ian Walsh[1,†], Manuel Giollo[1,2,†], Tomás Di Domenico[1], Carlo Ferrari[2], Olav Zimmermann[3] and Silvio C. E. Tosatto[1,*]

[1]Department of Biomedical Sciences, [2]Department of Information Engineering, University of Padua, Via Gradenigo 6, 35121 Padova, Italy and [3]Institute for Advanced Simulation, Forschungszentrum Juelich, Wilhelm-Johnen-Str., 52425 Juelich, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Intrinsically disordered regions are key for the function of numerous proteins. Due to the difficulties in experimental disorder characterization, many computational predictors have been developed with various disorder flavors. Their performance is generally measured on small sets mainly from experimentally solved structures, e.g. Protein Data Bank (PDB) chains. MobiDB has only recently started to collect disorder annotations from multiple experimental structures.

Results: MobiDB annotates disorder for UniProt sequences, allowing us to conduct the first large-scale assessment of fast disorder predictors on 25 833 different sequences with X-ray crystallographic structures. In addition to a comprehensive ranking of predictors, this analysis produced the following interesting observations. (i) The predictors cluster according to their disorder definition, with a consensus giving more confidence. (ii) Previous assessments appear over-reliant on data annotated at the PDB chain level and performance is lower on entire UniProt sequences. (iii) Long disordered regions are harder to predict. (iv) Depending on the structural and functional types of the proteins, differences in prediction performance of up to 10% are observed.

Availability: The datasets are available from Web site at URL: http://mobidb.bio.unipd.it/lsd.

Contact: silvio.tosatto@unipd.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 16, 2014; revised on August 11, 2014; accepted on September 15, 2014

1 INTRODUCTION

The rigid structure of proteins has been considered the determinant of function for many years. Recently, an alternative view is emerging with respect to non-folding regions, suggesting a reassessment of the structure-to-function paradigm (Dunker and Obradovic, 2001; Schlessinger *et al.*, 2011; Wright and Dyson, 1999). Flexible segments lacking a unique native structure, known as intrinsic disordered regions (Tompa, 2002), are widespread in nature, especially in eukaryotic organisms (Dunker *et al.*, 2000). These regions have been shown to play important roles in various biological processes such as cell signaling or regulation (Dunker *et al.*, 2002), DNA binding and molecular recognition (Tompa *et al.*, 2009). Their malleable properties allow multiple binding partners (Dosztányi *et al.*, 2006) with the flexible region often becoming folded on binding (Wright and Dyson, 2009).

Despite an emerging consensus regarding their existence, there is no single definition of disorder. As a result, various flavors of disorder have been proposed (Vucetic *et al.*, 2003). These disorder flavors have become diverse with some based on amino acid composition (Vucetic *et al.*, 2003), flexibility (Martin *et al.*, 2010) and functional roles coupled with conservation (Bellay *et al.*, 2011). Perhaps the simplest flavor distinction is the length of a disordered region, separated into short and long. Long regions seem to behave differently (Mohan *et al.*, 2009) and are difficult in structural determination, causing them to be underrepresented in the Protein Data Bank (PDB) (Rose *et al.*, 2013). The PDB contains structural information from X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, which can be used indirectly to study disorder. A plethora of computational predictors have also appeared, with special efforts to capture different flavors. Available methods can be broadly divided into three classes: biophysical, machine learning and consensus based. Biophysical methods (Dosztányi *et al.*, 2005; Galzitskaya *et al.*, 2006; Linding *et al.*, 2003a; Prilusky *et al.*, 2005) derive pseudo-energy functions from residue pairings in rigid structures (i.e. non-disorder) to recognize sequence regions with high energy as disordered. Machine learning, especially neural networks, has been widely used to predict protein disorder (Eickholt and Cheng, 2013; Hirose *et al.*, 2007; Ishida and Kinoshita, 2007; Linding *et al.*, 2003b; Walsh *et al.*, 2011, 2012; Ward *et al.*, 2004; Yang *et al.*, 2005). Many are tuned for the disorder style used in the Critical Assessment of techniques for protein Structure Prediction (CASP), where the goal is to detect missing residues in the X-ray crystal (Monastyrskyy *et al.*, 2014). Others attempting to move away from this disorder style measure some form of protein backbone flexibility. For example, ESpritz (Walsh *et al.*, 2012) can predict mobile NMR regions and DisEMBL (Linding *et al.*, 2003b) loops regions with high B-factor (high flexibility). The most recent disorder predictor category uses a consensus of various

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

biophysical and machine learning methods (Ishida and Kinoshita, 2008; Mizianty *et al.*, 2010; Schlessinger *et al.*, 2009; Walsh *et al.*, 2011; Xue *et al.*, 2010). Consensus approaches are frequently more accurate, but at the computational cost of running several predictors in parallel and averaging their output. Because there is no consensus on how to define disorder, predictors often vary in their parameter setting and disorder output. In nearly all cases, disorder is defined at the residue level, and the goal of the predictor is to maximize recovery of correct residues.

Among applications of disorder prediction, we can distinguish at least two different scenarios. The first is the CASP experiment (Monastyrskyy *et al.*, 2014), i.e. methods are used to predict a relatively small number of proteins with maximum accuracy and consensus predictors aiming for maximum accuracy should clearly excel. A more practical scenario is represented by high-throughput analysis of protein disorder on entire genomes (Schlessinger *et al.*, 2011). Over the years, most prediction methods have addressed the first problem, with comparatively little attention to the practicalities of large-scale predictions (Walsh *et al.*, 2012). MobiDB (Di Domenico *et al.*, 2012) is a large-scale disorder database containing experimental information on the entire PDB and predictions for all UniProt (The UniProt Consortium, 2012) sequences. Here, we use the vast quantity of disorder data for a first large-scale assessment. While most assessments are performed with hundreds or a few thousand examples, we have analyzed >25 000 UniProt sequences combining all available X-ray crystallographic structures. All disorder assessments so far are carried out on single PDB chains, whereas here the UniProt sequence is the final target. The UniProt annotation is unique and we compare it with standard PDB chain analysis for further insights.

2 METHODS

2.1 Datasets and classifications

All UniProt (The UniProt Consortium, 2012) sequences with at least one X-ray annotation in MobiDB (Di Domenico *et al.*, 2012) were downloaded on the May 13, 2013 (25 833 entries). Where more than one MobiDB annotation was available, a majority vote was used (Fig. 1) to produce a more stable disorder definition, filtering rare conflicts due to experimental conditions. Where MobiDB cannot find annotation for part of the UniProt sequence, residues are annotated as unknown and ignored. Each PDB (Rose *et al.*, 2013) chain that covered UniProt entries was also extracted for comparison and identical chains majority voted (Fig. 1). Similar chains were removed at 90% pairwise sequence identity using CD-HIT (101 338 chains reduced to 24 669). See Table 1 for statistics. Each UniProt entry was assigned to CATH using SIFTS (Velankar *et al.*, 2013). Gene Ontology (GO) terms (Ashburner *et al.*, 2000) were downloaded from UniProt and expanded to the ontology root. For a deeper analysis, the UniProt dataset was further split according to the following rules (Supplementary Table S1): removing short (<30 residue) PDB fragments, excluding conflicting residues, up to 10 non-consecutive disordered residues, >10 disordered residues.

2.2 Predictors

Predictors were selected with the condition that they must be available as an executable and fast, ideally returning predictions in <1 min. The following 11 programs were used (disorder definition used in parenthesis): ESpritz (X-ray, NMR and DisProt; Walsh *et al.*, 2012), IUPred (short and long; Dosztányi *et al.*, 2005), DisEMBL (hot loops and remark 465; Linding *et al.*, 2003b), RONN (X-ray; Yang *et al.*, 2005) and VSL2b (combination of X-ray and Disprot; Peng *et al.*, 2006), GlobPlot (globularity; Linding *et al.*, 2003a) and FoldIndex (folding; Prilusky *et al.*, 2005). This resulted in a total of 11 predictors with different disorder flavors. A short description of the predictors is given in the Supplementary Material. Predictor similarity was calculated on their residue scores (e.g. probability of disorder) was shown as a dendogram based on SOV performance. Low-complexity regions are parts of the sequence with strongly biased compositions (e.g. polyQ), which are thought to correlate with intrinsic disorder (Romero *et al.*, 2001). The low-complexity predictors SEG (Wootton, 1994) and Pfilt (Jones and Swindells, 2002) were used both as disorder predictors and to analyze disorder predictor performance in low-complexity regions.

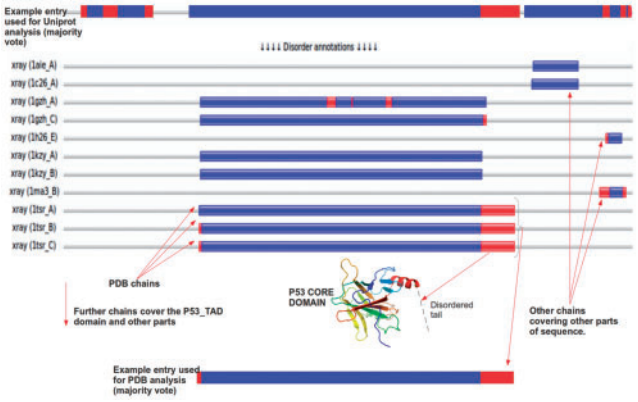

Fig. 1. Human P53 (UniProt ID: P04637) disorder annotation. The top bar shows the majority voting scheme, with blue for order and red for disorder. For simplicity, only a subset of PDB hits was shown. Missing regions are not considered. The bottom bar shows an example majority voted chain used in the PDB chain analysis

Table 1. Number of proteins, residues and region size

Dataset	Proteins	Residues			Disorder		Order
		Disorder	Structured	Unknown	Short	Long	
UniProt	25 833	350 858	6 731 814	3 655 566	23 566	3439	6271
PDB90	24 669	339 603	6 168 717	0	22 324	3576	5732
CASP10	95	1597	22 673	1186	139	20	19

Notes: UniProt contains full sequences annotated from MobiDB. PDB90 contains PDB chains at ≤90% sequence identity. CASP10 is shown for comparison purposes. Short disorders are proteins with at least three and long at least 20 consecutive residues. Order lists completely ordered proteins.

2.3 Performance assessment

Disorder prediction is a binary classification problem. As such, the standard measures accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC) and area under the curve (AUC) are used (see Supplementary Material). All these measures are calculated both *per residue* and as average on a *per protein* basis. MCC and AUC were replaced by SOV and FPreg in the per protein analysis. SOV is the mean of the segment overlap for disorder and structure, in analogy

to secondary structure (Zemla *et al.*, 1999). FPreg counts the number of predicted false-positive disordered regions. Disorder content measures the ability to recover the fraction of disordered residues in a protein independent of residue position. We adopted two previously used measures (Mizianty *et al.*, 2011), root mean square error (RMSE) and Pearson Correlation Coefficient (PCC), with predicted and observed disorder content normalized by the number of annotated residues. As a large number of measures hinders a global view of performance, we established an overall ranking as the average over all 12 quality measures. The Welch *t*-test was used to compute statistical significance.

3 RESULTS

3.1 First large-scale disorder assessment

We report the first large-scale assessment of disorder predictions on UniProt sequences through a comprehensive assessment of 11 fast predictors with new performance measures and a statistical evaluation. With respect to diversity, the UniProt set has 15 942 unique clusters at 40% identity cutoff. The 24 699 PDB chains are non-redundant by design (see Methods). Therefore, in both sets there is no large cluster of similar sequences, guaranteeing no bias in the analysis.

Figure 1 shows human P53 sequence (UniProt ID: P04637) covered by different disorder and structure definitions from the PDB. Using the majority voting approach, all structural and disorder information is combined and a more reliable global picture of the full p53 complex is constructed. Table 1 shows the number of proteins and residues using this annotation strategy. A total of 25 833 UniProt entries are annotated and a dataset of unique PDB chains is constructed for comparison purposes. For all sets, there is a clear imbalance between disorder and structured residues. Similar to the CASP10 experiment (Monastyrskyy *et al.*, 2014) where 20% of the data was completely ordered, 6271 sequences 24.3% of the UniProt dataset are completely ordered. Long disordered regions (>20 residues) are also abundant with 3439 examples.

Table 2 shows the per-residue performance on the UniProt dataset. Supplementary Tables S3 and S4 show similar results when excluding short PDB fragments and conflicting positions. The same trends are also found when separating the UniProt dataset for disorder content (see Supplementary Tables S5 and S6 and Section 3.4 below). Table 3 shows the per protein and disorder content performance. Most predictors have disorder scores significantly above random (AUCs >70). Depending on the prediction style, e.g. high coverage (overprediction) or highly confident (underprediction), one could argue for and against different predictors. For example, VSL2b has a lower residue specificity (81.16) predicting many false positives (1 268 274 residues), yet its AUC is the highest. On the contrary, IUPred-short has the best MCC (31.43) due to its high specificity. Table 2 can reveal detailed future objectives such as the need to retune the VSL2b decision threshold for higher specificity. For the SOV measure, DisEmbl-465 has the best performance (50.23). FPreg measures overprediction on segments as opposed to single residues. Again VSL2b clearly over predicts compared with DisEmbl-465.

Table 2. UniProt per-residue performance

Method	Accuracy	Sensitivity	Specificity	MCC	AUC
DisEmbl-465	67.42	39.56	95.28	30.80	78.73
DisEmbl-HL	66.17	59.59	72.76	15.49	72.69
ESpritz-Disprot	54.08	10.39	97.76	11.03	73.12
ESpritz-NMR	68.37	44.00	92.75	27.76	77.00
ESpritz-X-ray	69.93	54.32	85.54	23.34	77.76
FoldIndex	59.73	37.12	82.34	10.85	60.79
Globplot	59.61	31.76	87.46	12.21	63.15
IUPred long	63.14	30.98	95.29	23.99	72.59
IUPred short	68.16	41.26	95.06	**31.43**	77.81
RONN	68.57	51.53	85.59	21.85	75.87
VSL2b	**74.15**	**67.14**	81.16	25.62	**81.21**
SEG	54.15	16.69	95.45	11.91	54.15
Pfilt	50.75	2.16	**99.34**	3.80	50.75

Notes: All values are shown as percentages. The top performing method in each category is shown in bold and the second best underlined.

Table 3. UniProt per-protein and content performance

Method	Accuracy	Sensitivity	Specificity	SOV	FPreg	RMSE	PCC
DisEmbl-465	**79.60**	65.49	93.71	**50.23**	22 681	7.60	0.376
DisEmbl-HL	73.72	77.51	69.94	29.54	131 073	26.89	0.223
ESpritz-Disprot	62.05	31.35	92.75	43.97	1889	21.11	0.171
ESpritz-NMR	76.81	62.03	91.60	49.03	30 388	10.10	0.337
ESpritz-X-ray	78.26	72.33	84.19	48.39	54 411	17.89	0.241
FoldIndex	62.78	45.85	79.72	34.24	48 419	21.03	0.220
Globplot	68.65	50.04	87.26	28.37	55 433	12.73	0.100
IUPred long	67.33	41.03	93.64	36.58	16 601	8.72	0.367
IUPred short	78.69	64.14	93.24	48.84	18 904	8.17	**0.387**
RONN	70.21	56.34	84.07	39.14	45 177	14.56	0.331
VSL2b	78.96	**80.02**	77.90	38.75	72 125	19.60	0.338
SEG	63.23	29.76	96.70	42.98	6908	**7.55**	0.197
Pfilt	62.12	25.22	**99.03**	45.31	**1017**	8.27	0.081

Notes: All values are shown as percentages, except RMSE and PCC. The top-performing method in each category is shown in bold and the second best underlined.

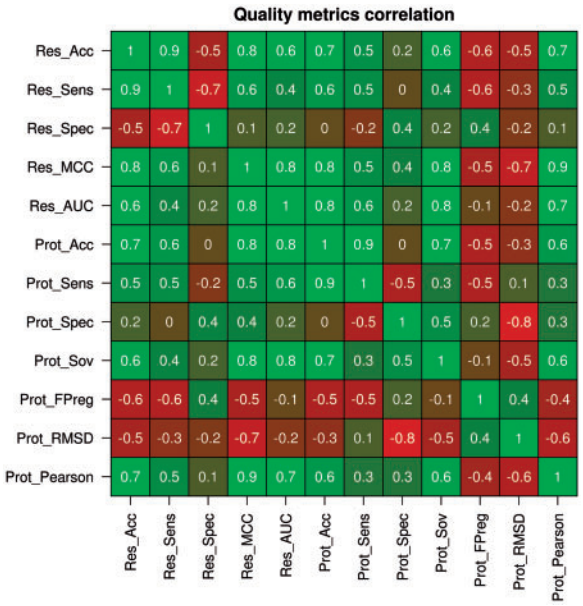

Fig. 2. PCC among performance measures. Each cell shows the PCC for the corresponding measures, with colors varying from green (+1) to red (–1). Res denotes per residue and Prot per protein measures

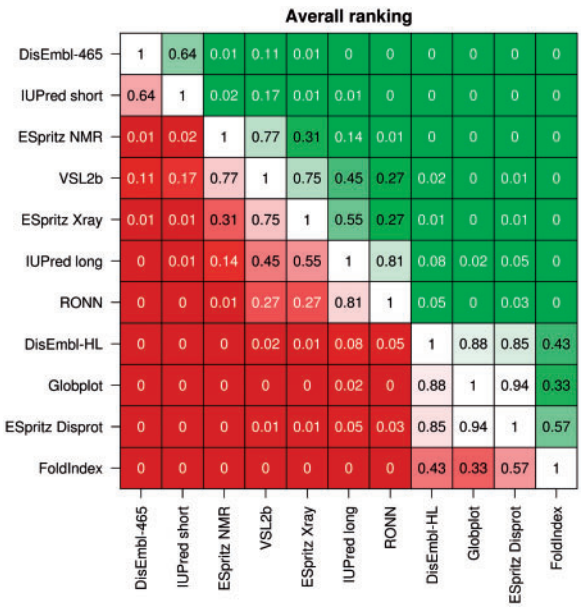

Fig. 3. Average ranking over all performance measures with statistical tests. The predictors to the left and bottom are ranked with *P*-value separating groups (Welch *t*-test) in each cell. Colors for the left predictor range from green (better) to red (worse), passing through white (tied). *P*-values >0.05 mean the performance distributions are similar and the difference between two predictors is not statistically significant

3.2 Similarity between measures and predictors

While the evaluation complexity arises due to predictor variability and the quantity of performance measures, it is useful to understand the deeper predictor behavior. Although many more observations could be made, for the sake of brevity a summarized ranking was chosen to give a clearer performance summary. Before ranking, it is important that the measures are evaluating different aspects of the predictions. Figure 2 shows that no measure correlates highly and most are diverse (pairwise correlation value >0.7 or <–0.7 in only 9 of 66 cases). This diversity ensures the ranking procedure is fair. Interestingly, per-residue MCC correlates highly with both disorder content measures (–0.7 with RMSD and 0.9 with PCC), suggesting they could be captured effectively by residue-level MCC. Both were kept because they are not completely redundant and the content measures do not depend on residue positions. From Figure 3, the top-ranked predictors are DisEmbl-465 and IUPred-short, with no statistically significant difference between performance (*P*-value 0.64). A second group consists of Espritz-NMR, VSL2b and Espritz-X-ray. In Supplementary Figure S2, these top five predictors are analyzed using receiver operating characteristic curves at low 0–5% false-positive rates (FPR). VSL2b starts outperforming the rest at around 2% FPR, again suggesting high-quality residue scores but a need for recalibration of its decision threshold.

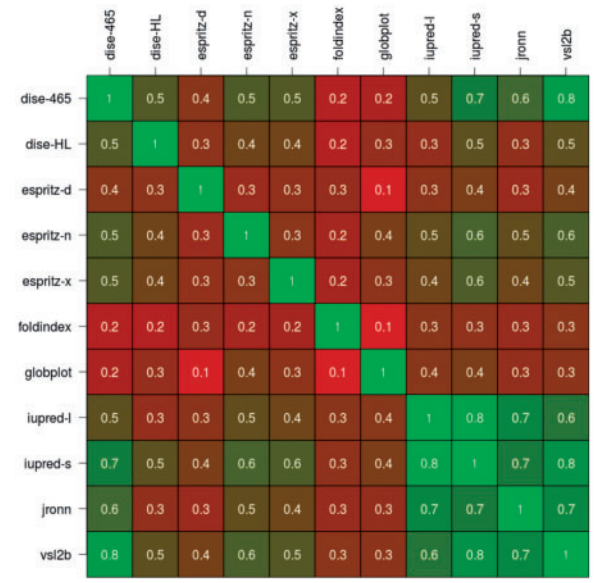

Fig. 4. PCC among predictors. Each cell shows the PCC for the corresponding measures with colors varying from green (+1) to red (–1)

Combining several good but complementary predictors is the heart of most consensus methods. Figure 4 shows the Pearson correlation between predictors. Examining similarity and performance is the first step in designing a consensus. Both are not necessarily related, e.g. IUPred-short, IUPred-long and RONN form a group of highly correlated predictors (PCC range 0.7–0.8) with different performances. Figure 5 shows a dendrogram of predictors grouped by SOV. This SOV difference of correlated predictors is mainly due to their selected decision threshold with residue scores remaining similar. Three methods, FoldIndex, GlobPlot and DisEmbl-HL correlate poorly with all others (PCC < 0.5) and also perform poorly on SOV.

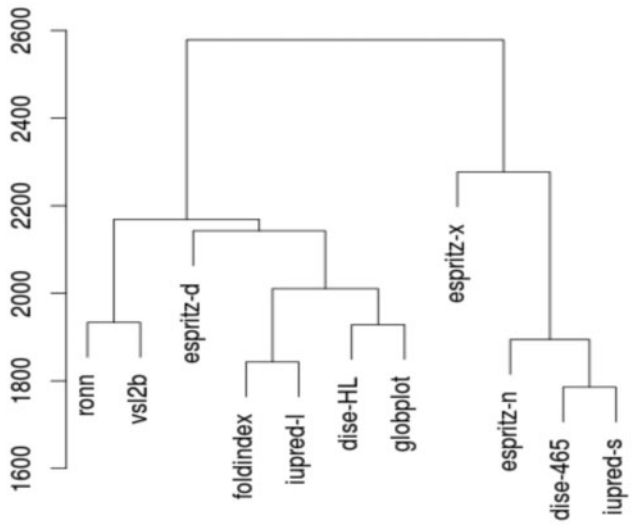

Fig. 5. Dendogram for disorder residue score and performance using the SOV measure. On the Y axis, the cumulative SOV score difference is plotted

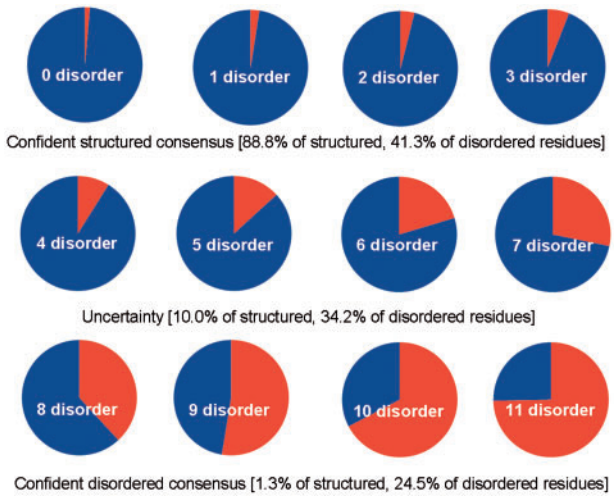

Fig. 6. Proportion of data that can be assigned confidently using a consensus. The pie chart shows each of the 11 possible scenarios (i.e. from 0 to 11 disorder predictions) and the corresponding fraction of truly disordered residues (in red). Each row corresponds to a situation (structure, uncertain, disorder) for which the percentages of occurrence are summarized

They nevertheless define different disorder flavors, which may be useful in certain situations.

IUPred-short and DisEmbl-465 have high correlation (>0.7 PCC) in conjunction with similarly high SOV, suggesting they detect the same disordered regions well. For a consensus approach, however, it is more interesting to combine, for example, Espritz-NMR/X-ray with DisEmbl-465, as they have low correlation (PCC 0.5) and quality SOV. To investigate consensus further, we measure agreement among predictors. Figure 6 shows the residues split into three equal groups: consensus structure, consensus disorder and uncertain. Uncertain is defined when there is disorder agreement for

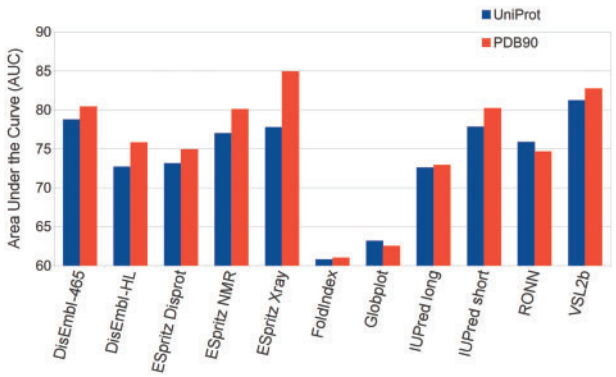

Fig. 7. Comparison between PDB chains and UniProt

4–7 predictors because the accuracy continually decreased below 61.2% (see Supplementary Fig. S3). When there is confident agreement, accuracy increases as expected, i.e. both tails of Supplementary Figure S3: 0–3 structure and 8–11 disorder agreement. In these regions, a consensus can recover 88.8 and 24.5% highly confident structured and disordered residues, respectively. Applying a simple majority vote in analogy to secondary structure (Albrecht *et al.*, 2003) produced 43.3% sensitivity, 95.6% specificity and an AUC of 78.8 per residue (see Table 2 for comparison).

3.3 Uniprot versus PDB chains

The most surprising result is the large decrease in performance compared with previously published performances. As most of the assessments in the literature are based on PDB chains, we examine whether assessments on PDB chains behave differently from UniProt sequences. Figure 7 shows the per-residue AUC differences between the UniProt and PDB chain datasets. Most predictors perform better on PDB chains and start approaching their published values (e.g. ESpritz-X-ray AUC 86.58 on CASP9). This is possibly due to the fact that predictor parameters are optimized on PDB chains. Another possible reason may be the positional dependence in PDB sequences, i.e. missing atoms or disordered residues in solved structures are often located at the N and C termini. This effect was recently noted for CASP-10 (Monastyrskyy *et al.*, 2014). Given that most methods encode the sequence context (e.g. using sliding windows in neural networks) they will implicitly learn the position of the termini. This information is lost when the PDB sequence is assigned to a part, often the middle, of the UniProt sequence. Moreover, the definition is different in UniProt because it is a majority combination of multiple experimental sources. Supplementary Table S7 shows the full set of performance measures on the PDB chain set.

3.4 Sequence and structure variability

Fluctuations in performance given different protein properties are often overlooked. To our knowledge, this has never been examined comprehensively. In all cases, performance is assessed using SOV and there are indeed some striking performance differences. Proteins are grouped into bins of low complexity content (1% intervals), and the top five predictors are analyzed for

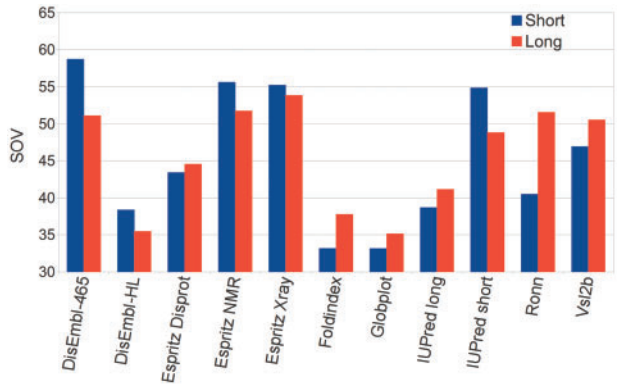

Fig. 8. Comparison between predictors on long and short disorder proteins

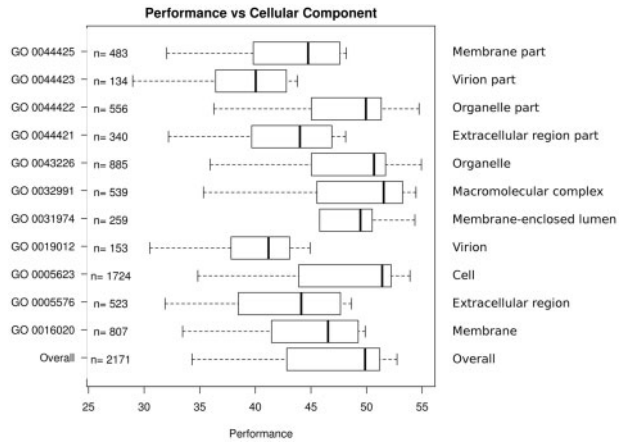

Fig. 9. Relationship between GO cellular component and disorder performance of top five predictors. All proteins had at least some disorder. GO terms only considered if the number of proteins is >50

performance changes in Supplementary Figure S4. Increasing low complexity content increases SOV logarithmically with the largest gains in the 5–20% low complexity range. A similar concept was already investigated (Romero *et al.*, 2001), suggesting that disorder predictors are generally using low complexity patterns to predict unstructured regions. At first glance, this seems to contradict Table 2, which shows SEG and Pfilt producing almost random predictions. It can be explained by the fact that no low complexity regions were detected for 40.1% of the disordered proteins. From Supplementary Figure S4 it is clear that low complexity has a significant relationship with disorder performance whenever present.

Disorder region length is perhaps the most obvious sequence property, and the majority (75.8%) of sequences in our dataset contain at most 10 disordered residues (Supplementary Fig. S5). The performance separated on this threshold is shown in Supplementary Tables S5 and S6. This skewed distribution may not reflect the truth in nature, especially for long proteins where long disordered regions may be missed owing to lack of evidence, but is nevertheless interesting, as we are using a common disorder definition (Monastyrskyy *et al.*, 2014). Figure 8 shows the performance of each method separated into two sets. Proteins containing at least one long disorder stretch (i.e. >20 residues) or not. Detection of disorder with long regions had a decreased performance in 5 of 11 methods, but 4 of these were the top ranked ones. Conversely, methods trained to take into account long disorder (e.g. IUPred-long, Espritz-Disprot, RONN and VSL2b) and the folding predictors (GlobPlot and Foldindex) showed better performance on proteins with long disorder. This suggests that the top-ranked methods can be improved by taking into account long disorder regions in their training.

Using CATH, 9378 proteins with disorder were extracted form the UniProt set. Supplementary Figure S6 shows the performance of the top five methods on the four main CATH classes. The CATH *few secondary structure* class is predicted considerably below average. This could be due to the high quantity of disorder in this class (Supplementary Fig. S7). *Mainly alpha* structures are clearly easier to detect, perhaps due to alpha helices being dependent on local sequence. Conversely, *mainly beta* structures are harder to predict perhaps because beta sheet hydrogen bonds are dependent on distant residues. This difficulty in capturing distant sequential dependencies is a common problem in secondary structure prediction (Rost, 2001) and likely also true for disorder.

3.5 Functional variability

The top five predictor SOV performances for proteins with at least one disordered residue are separated into the three GO classes (Ashburner *et al.*, 2000). Cellular Component covers 3458 proteins, Biological Process 4696 and Molecular Function 5260. Figure 9 shows how SOV varies significantly for different cellular component terms. Virion-related proteins have the most interesting performance drop, probably due to an increased level of disorder in these proteins (Supplementary Fig. S8). Perhaps more interestingly, performance for *membrane* and *extracellular* proteins is generally lower than average, even though the amount of disorder was not enriched (Supplementary Fig. S8). This may be a consequence of disorder having different amino acid composition in these proteins (Xue *et al.*, 2009). For GO *Molecular Function* (Supplementary Fig. S9), *binding* and *transporter activity* are predicted well but the activity relationships *structural molecule* and *receptor* have the lowest performance. Disorder performance also varies with *Biological Process* (Supplementary Fig. S10). *Signaling* and *regulation* were easier to predict while *biological adhesion* has a glaringly poor performance. In each of the three GO classes, SOV performance varies by up to 10% for different GO terms.

4 DISCUSSION

Efficient disorder predictions are vital for understanding large collections of proteins and entire proteomes. In this work, 11 predictors are evaluated on 25 833 UniProt sequences with disorder annotations from X-ray crystallographic structures. The evaluation procedure consists in measuring performances using 12 different scores and ranking the predictors while highlighting statistically similar groups. Although in some cases the disorder definition used will not represent true functional

mobility, we feel that it should capture most aspects of intrinsic disorder. The assumption is that in most cases, missing backbone atoms in PDB structures correlate with intrinsic disorder defined in DisProt (Sickmeier *et al*., 2007). Our definition is also arguably more stable because it is based on a majority vote on all PDB structures covering a UniProt entry. Thus experimental errors in X-ray crystallography (e.g. missing residues due to low resolution) should be removed, as disorder is only considered if it occurs most frequently in the PDB.

The evaluation reveals a strong variability in predictors across the 12 measures, indicating different prediction styles (e.g. overprediction or confident underprediction). Ranking each predictor with the 12 measures shows both DisEmbl-465 and IUPred-short performing consistently well on each measure. The ranking was robust because the 12 measures show little correlation (Fig. 2). Predictors that ranked poorly still contain a good signal across our disorder definition. In most cases, they fall behind because they offer a different interpretation, which may be useful in alternative settings. At CASP, the best disorder predictors are widely known to be meta-predictors combining orthogonal information (i.e. unique predictors performing well). A correlation analysis on the predictors produced similar clusters as well as unique predictors. Some predictors showed both uniqueness and good performance, indicating a consensus predictor may be beneficial. Using predictor combinations, 88.8% of structured and 24.5% of disordered residues are found with highly confident agreement. The remaining 10.0% structured and 34.2% disordered residues classified as uncertain may be decided with more sophisticated heuristics (e.g. high residue scores).

Highly accurate predictors on UniProt sequences are vital, considering that users are invariably trying to understand disorder properties of unannotated proteins and not the PDB, which is already annotated with quality structural information. Despite this, the literature largely concentrates on PDB chain assessments. The performance on the UniProt disorder definition is substantially lower than the equivalent evaluation on PDB chains. A similar effect was recently noted in the CASP-10 assessment, where database predictions were worse than the direct submissions by the same methods (Monastyrskyy *et al*., 2014). In general, increases in PDB chains are observed across all measures (Supplementary Table S7), suggesting that the prediction of the more desirable UniProt disorder may be worth considering for training new predictors.

There are large performance variations when splitting the data into groups of proteins. As expected, predictors prefer large amounts of low sequence complexity, but the performance seems to plateau after 20% low complexity. On the other hand, long disorder detection seems to be more difficult, especially for the predictors we find to be accurate. While both trends are somewhat expected, the dependence of performance on structure and function are less obvious. At the structural level, *beta-only* proteins seem to be more difficult to predict compared with *alpha-only* or mixed *alpha/beta*. The *few secondary structure* class is certainly the poorest, but this may be due to long disordered regions being poorly detected. Given that functional disorder analysis using predictors is gaining attention (Ward *et al*., 2004; Xue *et al*., 2014), prediction error is shown relative to GO. The analysis shows that the average error rate is not universal across all functions. It is possible that enriched functions found in genome analysis may have a slight bias. For example, the association with disorder and *binding*, *signaling* and *regulation* is known, but here we found that they are more easily detected, possibly inferring enrichment. Compared with virion sequences, which are more abundant in experimental disorder (Supplementary Fig. S8) and supported by the literature (Xue *et al*., 2014), the error rates on their predictions are higher than average. One of the main reasons for performance variation could be the distribution of protein types in predictor training sets. Binding, signalling and regulation proteins together constitute a large fraction of known disorder datasets, and it is reasonable to assume that the same distributions are used in each predictor. It is therefore possible that predictors are optimized for these common families. Optimistically, the use of this prior knowledge could enhance predictor training or motivate the development of specific tools.

To our knowledge, this is not only the first large scale analysis of disorder predictions from X-ray crystallographic structures but also the first attempt to provide error rates on sequence, structure and functional protein types. We are in the process of developing this evaluation into an automatic evaluation server and plan to integrate it in the new version of MobiDB (Di Domenico *et al*., 2012).

ACKNOWLEDGEMENT

The authors are grateful to members of the BioComputing UP lab for insightful discussions.

Funding: FIRB Futuro in Ricerca grant (RBFR08ZSXY to S.T.).

Conflict of Interest: none declared.

REFERENCES

Albrecht,M. *et al*. (2003) Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng*., **16**, 459–462.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology consortium. *Nat. Genet*., **25**, 25–29.

Bellay,J. *et al*. (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol*., **12**, R14.

Di Domenico,T. *et al*. (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.

Dosztányi,Z. *et al*. (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res*., **5**, 2985–2995.

Dosztányi,Z. *et al*. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol*., **347**, 827–839.

Dunker,A.K. *et al*. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.

Dunker,A.K. *et al*. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Workshop Genome Inform*., **11**, 161–171.

Dunker,A.K. and Obradovic,Z. (2001) The protein trinity—linking function and disorder. *Nat. Biotechnol*., **19**, 805–806.

Eickholt,J. and Cheng,J. (2013) DNdisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics*, **14**, 88.

Galzitskaya,O.V. *et al*. (2006) Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput. Biol*., **2**, e177.

Hirose,S. *et al*. (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23**, 2046–2053.

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

Ishida,T. and Kinoshita,K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**, 1344–1348.
Jones,D.T. and Swindells,M.B. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci.*, **27**, 161–164.
Linding,R. *et al.* (2003a) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
Linding,R. *et al.* (2003b) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
Martin,A.J.M. *et al.* (2010) MOBI: a web server to define and visualize structural mobility in NMR protein ensembles. *Bioinformatics*, **26**, 2916–2917.
Mizianty,M.J. *et al.* (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.
Mizianty,M.J. *et al.* (2011) In-silico prediction of disorder content using hybrid sequence representation. *BMC Bioinformatics*, **12**, 245.
Mohan,A. *et al.* (2009) Influence of sequence changes and environment on intrinsically disordered proteins. *PLoS Comput. Biol.*, **5**, e1000497.
Monastyrskyy,B. *et al.* (2014) Assessment of protein disorder region predictions in CASP10. *Proteins*, **82** (Suppl. 2), 127–137.
Peng,K. *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.
Prilusky,J. *et al.* (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21**, 3435–3438.
Romero,P. *et al.* (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
Rose,P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
Schlessinger,A. *et al.* (2009) Improved disorder prediction by combination of orthogonal approaches. *PloS One*, **4**, e4433.
Schlessinger,A. *et al.* (2011) Protein disorder—a breakthrough invention of evolution? *Curr. Opin. Struct. Biol.*, **21**, 412–418.
Sickmeier,M. *et al.* (2007) DisProt: the database of disordered proteins. *Nucleic Acids Res.*, **35**, D786–D793.
The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
Tompa,P. *et al.* (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *BioEssays*, **31**, 328–335.
Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
Velankar,S. *et al.* (2013) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.
Vucetic,S. *et al.* (2003) Flavors of protein disorder. *Proteins*, **52**, 573–584.
Walsh,I. *et al.* (2011) CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. *Nucleic Acids Res.*, **39**, W190–W196.
Walsh,I. *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.
Ward,J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.
Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.*, **293**, 321–331.
Wright,P.E. and Dyson,H.J. (2009) Linking folding and binding. *Curr. Opin. Struct. Biol.*, **19**, 31–38.
Xue,B. *et al.* (2009) Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol. Biosyst.*, **5**, 1688–1702.
Xue,B. *et al.* (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*, **1804**, 996–1010.
Xue,B. *et al.* (2014) Structural disorder in viral proteins. *Chem. Rev.*, **114**, 6880–6911.
Yang,Z.R. *et al.* (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.
Zemla,A. *et al.* (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.