PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessment of protein disorder region predictions in CASP10

Bohdan Monastyrskyy,[1] Andriy Kryshtafovych,[1] John Moult,[2] Anna Tramontano,[3] and Krzysztof Fidelis[1]*

[1] Genome Center, University of California, Davis, Davis, California 95616

[2] Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland 20850

[3] Department of Physics, Sapienza University of Rome, 00185 Rome, Italy

ABSTRACT

The article presents the assessment of disorder region predictions submitted to CASP10. The evaluation is based on the three measures tested in previous CASPs: (i) balanced accuracy, (ii) the Matthews correlation coefficient for the binary predictions, and (iii) the area under the curve in the receiver operating characteristic (ROC) analysis of predictions using probability annotation. We also performed new analyses such as comparison of the submitted predictions with those obtained with a *Naïve* disorder prediction method and with predictions from the disorder prediction databases D2P2 and MobiDB. On average, the methods participating in CASP10 demonstrated slightly better performance than those in CASP9.

Proteins 2014; 82(Suppl 2):127–137.

Key words: CASP; intrinsically disordered proteins; unstructured proteins; prediction of disordered regions; assessment of disorder prediction.

INTRODUCTION

A systematic analysis of intrinsic disorder in proteins started at the turn of the century[1–4] and still remains a hot research topic.[5] Only this year several papers covering general aspects of protein disorder have been published[5–9] and the discussion on the fundamental principles of disorder continues to unfold.[10,11] PubMed search with the keywords "intrinsically disordered protein 2012" and "intrinsically disordered protein 2013" returned 525 and 305 entries, respectively (as of April 2013). The number of experimentally verified intrinsically disordered proteins and regions is steadily increasing. The DisProt database[12] currently contains annotations for 684 intrinsically disordered proteins, 1513 disordered regions, and describes 38 different biological functions associated with disordered regions. The more recently established IDEAL database also has a number of useful annotations on disordered proteins.[13]

Such a high interest in this area of research triggered rapid development of computational methods for prediction of the location of disordered regions in proteins. The recently published reviews and assessment papers[14–18] altogether provide a comprehensive analysis of more than fifty disorder prediction methods. An independent assessment of the protein disorder methods within the scope of CASP started in 2002 and is now already in its sixth round.[18–22] This study analyzes the results obtained by the 28 disorder prediction groups participating in CASP10.

MATERIALS AND METHODS

Definition of disorder

As in all previous CASPs, a residue was considered as being in either ordered or disordered state based on the information provided in the protein coordinate file. If available at the time of the evaluation, the files from the

Additional Supporting Information may be found in the online version of this article.

Abbreviations: AUC_PR, area under the precision-recall curve; AUC, or AUC_ROC, area under the ROC curve; DR, disordered region; MCC, the Matthews correlation coefficient; ROC, the receiver operating characteristic; 3D, three-dimensional; SVM, support vector machine.

Grant sponsor: NIGMS/NIH; Grant number: R01GM100482 (to KF); Grant sponsor: King Abdullah University of Science and Technology (KAUST); Grant number: Award No. KUK-I1–012-43 (to AT); Grant sponsor: EMBO.

Bohdan Monastyrskyy and Andriy Kryshtafovych contributed equally to this work.

*Correspondence to: Krzysztof Fidelis, Genome Center, University of California, Davis, 451 Health Sciences Dr., Davis, CA 95616. E-mail: kfidelis@ucdavis.edu

Received 23 May 2013; Revised 14 June 2013; Accepted 18 June 2013
published online 14 August 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24391

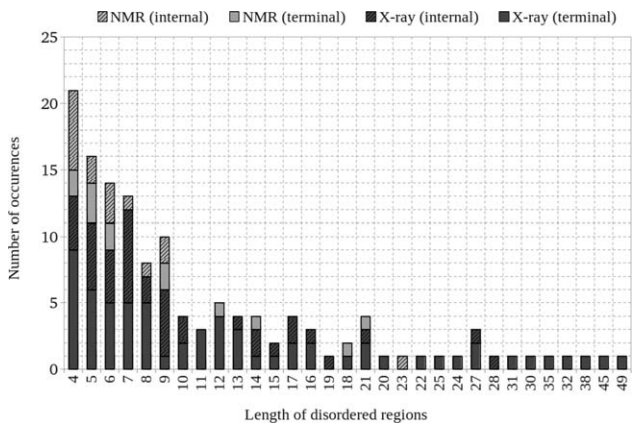

Figure 1

Distribution of disordered regions of different lengths in CASP10 targets. Internal regions are shown with darker shades.

PDB database[23] were used to define disordered regions; in all other cases, coordinate files provided directly by the experimentalists to the CASP organizers were used.

A residue was defined as disordered if it was present in the target's amino acid sequence but either (1) it lacked the spatial coordinates or (2) its coordinates were not well-defined, that is, the distances between positions of the same residue in any pair of models in the NMR ensemble or in any pair of X-ray chains in the asymmetric unit exceeded 3.5 Å.

Note that this definition of disorder is not ideal as lack of spatial coordinates in the PDB can arise from causes other than intrinsic disorder. For example, flexibly hinged structured domains with mobility within the crystal lattice have long been known to result in lack of spatial coordinates[24]; in some cases, lack of spatial coordinates in the PDB has resulted from simple annotation errors.[9] It is also worth mentioning that possible transitions of residues between disordered and ordered states with the change of physiological conditions were not taken into account as they are impossible to define solely from the coordinates, usually the only information available to the assessors.

Targets and test sets

The groups participating in the CASP10 DR prediction category were asked to predict disordered regions in all 114 released targets, including the all-group and server-only ones.[25] Eighteen targets were canceled by the organizers[26] and two more (T0677 and T0686) were excluded from the assessment as inappropriate for the disorder evaluation (see the quality assessment paper in this issue[27] for a detailed explanation). The remaining 94 targets were assessed.

According to the adopted disorder definition, 1664 residues in the assessed targets were identified as disordered. They constitute 6.8% of all residues, and this percentage is lower than that of the disordered residues present in the CASP9 targets (10.2%). Structures identified by NMR show a higher level of disorder, in particular due to difference in experimental conditions. A discussion of the implications of this observation is outside of scope of this paper, but some interesting thoughts on this subject can be found in recently published papers.[9–11] The CASP10 NMR-derived structures contained 277 disordered residues constituting 28.0% of all residues in the NMR targets versus 1387 disordered residues (or 5.9% of all residues) in the X-ray targets.

The total fraction of disordered residues in individual targets varies from 0% in five targets (T0651, T0693, T0735, T0747, and T0757) to 57.3% in T0675 – a 75-residue-long NMR target. The distribution of disordered segments with respect to their length is shown in Figure 1. The longest continuous disordered region in CASP10 targets is 49 residues long and is located at the N-terminus of an X-ray target (T0652). Shorter disordered segments occur more often. In the evaluation, to reduce statistical noise due to experimental uncertainty, we have removed from consideration unstructured regions shorter than four residues (162 residues total).

The disordered regions are more likely to be found at the proteins' termini. Out of 134 disordered segments longer than three residues, 78 segments containing 998 residues (or 66.4% of all disordered residues) fall either at the beginning or at the end of the sequence. Figure 2 shows the fraction of ordered/disordered residues for the first and last 15 residues of the assessed targets. In more than half of the targets, the first five residues and the last residue are in the disordered state.

Format of predictions, participating groups

Disorder predictors were allowed to submit one model per target, and the format of the predictions did not change since the previous CASP. Every residue in the

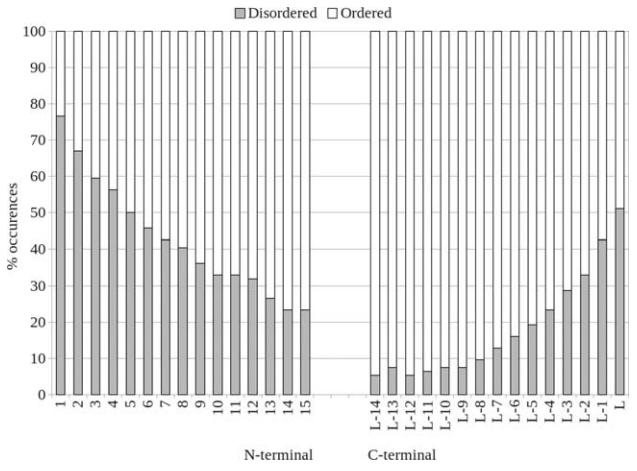

Figure 2

Fraction of disordered residues at the termini of CASP10 targets. *L* is the target length (in number of residues).

Table I
The Publicly Available Disorder Prediction Servers Participating in CASP10

CASP10 group name and URL	Description
Prdos-CNF, metaprdos2 http://prdos.hgc.jp/cgi-bin/top.cgi	Prdos-CNF: conditional neural fields. Metaprdos2: uses predictions from five servers.
DISOPRED3 http://bioinf.cs.ucl.ac.uk/disopred	SVM trained on high-resolution X-ray structures. Uses profiles from 15 positions around each residue as an input vector.
biomine_dr_mixed, biomine_dr_pdb-c http://biomine.ece.ualberta.ca/MFDp.html	Meta-servers which include additional information such as evolutionary profiles, secondary structure, solvent accessibility, and dihedral angles in SVM learning.
POODLE http://mbs.cbrc.jp/poodle	An SVM integrating three in-house SVM predictors: Poodle-S and Poodle-L specialized in short and long disorder regions, respectively, and Poodle-W targeting unfolded protein prediction.
MULTICOM-construct, MULTICOM-novel, MULTICOM-refine http://irirs.rnet.missouri.edu/dndisorder/ http://casp.rnet.missouri.edu/predisorder.html	MULTICOM-novel: deep neural networks. MULTICOM-refine: 1D recursive neural network. Input data include sequence profile, secondary structure, and solvent accessibility. MULTICOM-construct combines the predictions of both.
Espritz, Espritzv2, Cspritz http://protein.bio.unipd.it/espritz/ http://protein.bio.unipd.it/cspritz/	Espritz: recursive neural networks. Cspritz: additionally uses two SVM modules trained on different datasets.
IntFOLD2 http://www.reading.ac.uk/bioinf/IntFOLD/IntFOLD2_form.html	Uses 3D models of the ModFOLDclust2 server to identify the regions of high variability.
OnD-CRF69 http://babel.ucmp.umu.se/ond-crf/	Conditional random fields. Uses sequence and predicted secondary structure as inputs.
GSmetaDisorderMD, Gsmetaserver, GSMetadisorder, GSMetaDisorder3D http://iimcb.genesilico.pl/metadisorder/	The meta-servers use genetic algorithm and different weighting schemes to average models from 13 disorder predicting servers.

released target sequence had to be labeled as ordered or disordered and assigned a score in the range [0;1], estimating the probability of disorder. According to the requirements introduced in CASP9, probabilities above 0.5 were reserved for disordered residues.

Twenty-eight groups participated in the DR prediction category in CASP10, including 26 automatic servers and 2 human expert groups. These numbers are very similar to those of CASP9, where 22 servers and 8 expert groups participated. Classification and a brief description of publicly available CASP10 disorder prediction servers are provided in Table I.

Table II (Column "Targ") contains information on the total number of targets on which the CASP10 groups were evaluated. The majority of the groups submitted predictions for all 94 assessed targets; a few groups submitted predictions for a slightly smaller number of targets, but the coverage of the target dataset was always large enough to allow including all groups in the evaluation.

Datasets for comparison with the D2P2[28] and MobiDB[29] databases

Exact sequence matches for 53 and 83 out of 94 evaluated CASP10 targets were found in the D2P2 and MobiDB databases, respectively. Fifty-one CASP targets had 100% sequence identity with entries in both the D2P2 and MobiDB databases.

Naïve predictor

As mentioned in the "Targets and test sets" section above, the disordered residues are more likely to appear at the proteins' termini. This information can be easily incorporated in methods for statistical disorder prediction. To estimate how much better are CASP predictors compared to a simple predictor exploiting this tendency, we introduced a *Naïve* disorder predictor assigning the first nine and last four residues in the protein as disordered. These numbers were selected based on the average length of the disordered terminal regions in the CASP9 targets.

Evaluation criteria

A comprehensive analysis of strong and weak points of different measures historically used in CASP disorder evaluations is provided in the CASP9 assessment paper.[18] The measures and procedures that were identified there as the most suitable for the evaluation of disorder prediction were used here as basis for the CASP10 disorder assessment. Below we briefly discuss these measures.

Binary metrics

For evaluation of disorder predictors as binary classifiers we used the *precision*

$$precision = \frac{TP}{TP+FP},$$

the *balanced accuracy* (Acc)

$$Acc = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right),$$

and the *Matthews correlation coefficient* (MCC)

Table II
Performance of the Groups Participating in the DR Prediction Category and the *Naïve* Method

ID	Group name	Targ	TP	FP	TN	FN	prec	Acc	MCC	AUC (ROC)	AUC (PR)	Ranks				
												prec	Acc	MCC	AUC (ROC)	AUC (PR)
369	Prdos-CNF	94	657	287	22401	845	0.696	0.712	**0.529**	**0.907**	0.581	2	18	2	1	2
170	DISOPRED3	94	607	201	22487	895	0.751	0.698	**0.531**	**0.897**	0.603	1	22	1	2	1
478	biomine_dr_mixed	94	628	368	22320	874	0.631	0.701	**0.488**	**0.890**	0.526	4	20	3	3	3
288	biomine_dr_pdb_c	94	579	290	22398	923	0.666	0.686	**0.483**	**0.886**	0.526	3	25	4	4	3
340	metaprdos2	88	918	2228	18603	467	0.292	0.778	**0.385**	**0.879**	0.496	15	2	10	5	7
216	POODLE	94	980	2064	20624	522	0.322	0.781	**0.409**	**0.875**	0.416	12	1	6	6	15
222	MULTICOM-construct	94	940	1972	20716	562	0.323	0.769	**0.400**	**0.873**	0.502	11	4	7	7	5
180	Yang test	94	828	1702	20986	674	0.327	0.738	**0.376**	**0.872**	0.483	10	10	11	8	8
413	ZHOU-SPARKS-X	94	994	3065	19623	508	0.245	0.763	**0.340**	**0.870**	0.475	21	5	18	9	10
129	CASPITAv2	93	863	1610	20766	639	0.349	0.751	**0.400**	**0.859**	0.482	9	8	7	10	9
424	MULTICOM-novel	94	944	2630	20058	558	0.264	0.756	**0.349**	**0.856**	0.500	18	7	16	11	6
380	Espritz	94	916	2938	19750	586	0.238	0.740	**0.317**	**0.855**	0.465	23	9	20	12	11
327	Espritzv2	94	780	1639	21049	722	0.322	0.724	**0.360**	**0.852**	0.460	13	16	13	13	12
125	MULTICOM-refine	94	1029	3059	19629	473	0.252	0.775	**0.354**	**0.846**	0.459	20	3	14	14	13
003	GSmetadisorderMD	87	774	1895	18923	647	0.290	0.727	**0.341**	**0.844**	0.394	16	13	17	15	19
084	biomine_dr_pdb	94	814	2076	20612	688	0.282	0.725	**0.335**	**0.840**	0.409	17	15	19	16	16
484	CSpritz	94	1019	3655	19033	483	0.218	0.759	**0.316**	**0.829**	0.427	25	6	21	17	14
193	AIdisorder	83	731	1571	18172	674	0.318	0.720	**0.352**	**0.826**	0.405	14	17	15	18	18
273	IntFOLD2	94	1108	6359	16329	394	0.148	0.729	**0.239**	**0.821**	0.406	27	11	25	19	17
214	OWL2	90	566	714	20834	834	0.442	0.686	**0.387**	**0.821**	0.375	6	25	9	19	20
140	OnD-CRF2	92	819	2536	20027	628	0.244	0.727	**0.311**	**0.814**	0.248	22	13	22	21	26
496	GSmetadisorder	94	883	2989	19699	619	0.228	0.728	**0.300**	**0.808**	0.338	24	12	23	22	21
494	GSmetaserver	94	1059	6976	15712	443	0.132	0.699	**0.204**	**0.778**	0.332	28	21	27	23	22
183	sDisPred	94	980	5380	17308	522	0.154	0.708	**0.228**	**0.778**	0.316	26	19	26	23	23
384	GSmetadisorder3d	90	258	732	20717	1187	0.261	0.572	**0.173**	**0.753**	0.193	19	29	28	25	27
115	SIbio	87	575	899	20054	805	0.390	0.687	**0.362**	**0.699**	0.250	8	24	12	26	25
168	DisMeta	93	598	399	22201	890	0.600	0.692	**0.464**	**0.692**	0.310	5	23	5	27	24
167	Algorithmic_code	94	561	3999	18689	941	0.123	0.599	**0.122**	**0.599**	0.094	29	28	29	28	28
—	*Naïve*	94	366	526	22162	1136	0.410	0.610	**0.282**	—	—	7	27	24	—	—

The groups are ranked according to the AUC (ROC) score. The *Naïve* predictor is assessed on the binary predictions only. The two main evaluation scores (MCC for two-class estimations and AUC for probability-based predictions) are bold-faced.

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

where *TP* (true positives) and *TN* (true negatives) are the numbers of correctly predicted disordered and ordered residues, respectively, and *FP* (false positives) and *FN* (false negatives) are the numbers of misclassified disordered and ordered residues, respectively.

A good feature of these measures is that all of them place more weight on the prediction of the minority class (disordered residues). The *precision* is completely insensitive to the prediction of the dominant class (ordered state) and reports the ratio between the correctly predicted disordered residues and all predicted disordered residues. The balanced accuracy and the Matthews correlation coefficient take into account all parameters of the prediction quality (*TP*, *TN*, *FP*, and *FN*). The main conceptual difference is that *MCC* does not reward overprediction of disorder as much as the *Acc* does and is better suited to identify classifiers with higher *precision*. All three measures were shown to be appropriate for evaluation of the disorder data,[18] and in our assessment we provide scores for all of them. We consider the *MCC* as the main estimator of quality, as it is more balanced than the other two.

Probability-based metrics

The accuracy of identifying disorder by assigning per-residue disorder confidence scores can be evaluated by the receiver operating characteristic (ROC) or the precision-recall (PR) curve analysis.

The ROC analysis has been previously used in the assessment of protein disorder predictions in CASP[18–21] and elsewhere.[30] A classical ROC curve represents a monotonic function describing the balance between the true positive and false positive rates of a predictor. For a set of probability thresholds (from 0 to 1), a residue is considered as a positive example (disordered) if its predicted probability is equal to or greater than the threshold value. The area under the curve (*AUC*, or *AUC_ROC*) is used as an aggregate measure of the overall quality of a prediction method. A value of 1 corresponds to a perfect classifier, while 0.5 indicates a random prediction. Note that the ROC curve analysis works best for the probability estimates that are evenly distributed throughout the range of the allowed values. The "granularity" of the probability scores can affect

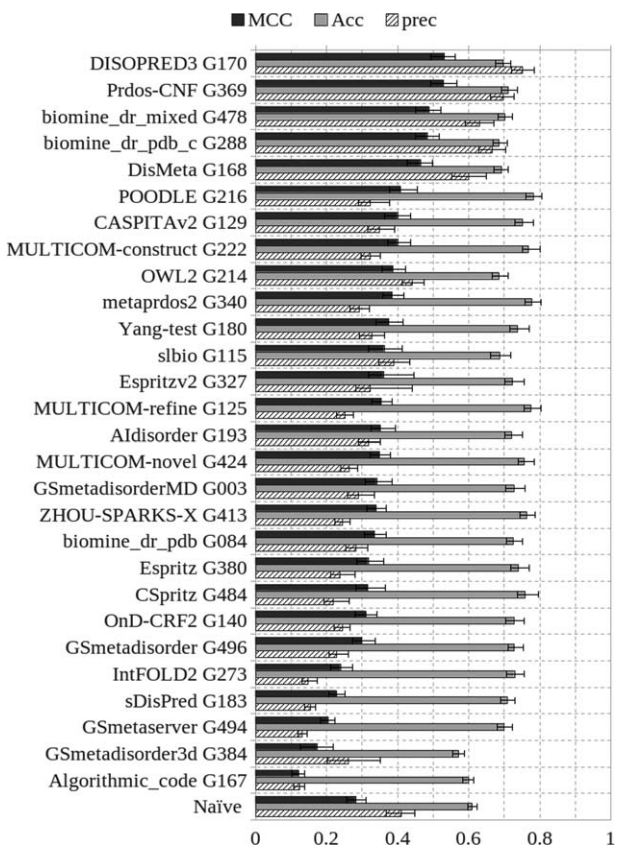

Figure 3

Performance of groups as binary disorder/order classifiers according to the *MCC*, *Acc*, and *precision* scores. The error bars indicate boundaries of the 95% confidence intervals for each measure. Groups are ordered according to decreasing MCC score.

smoothness of the ROC curves and, subsequently, the accuracy of the *AUC* scores.

The ROC curves are known to overestimate performance of predictors on the imbalanced data.[31] To address this potential issue, we complemented the ROC analysis with the PR curve analysis, which is particularly suitable for statistical evaluations on disproportional datasets.[32–35] The PR curves are conceptually similar to the ROC curves,[36] but differ in that they are plotted in the $(recall = TP/(TP + FN),\ precision = TP/(TP + FP))$ coordinates and are not necessarily monotonous. As in ROC curve analysis, the area under the PR curve, *AUC_PR*, is indicative of the classifier's accuracy, with a value of 1 corresponding to a perfect predictor.

Statistical significance of differences in group results

The statistical significance of the differences in group performance was estimated in different ways in the binary and probability-based analyses.

For binary predictions, we used a resampling technique, where we randomly drew 80% of the targets and calculated the scores on the selected subset. The procedure was repeated 1000 times and the obtained distributions of scores were used to compute the confidence intervals at the 95% level. The conclusion regarding the statistical difference in performance of the groups was inferred by the comparison of the corresponding confidence intervals.

The statistical differences in the probability-based evaluation of results were assessed by the nonparametric DeLong tests calculated with the statistical package *R*.[37,38]

RESULTS

The assessment results of the CASP10 disorder prediction methods and the *Naïve* predictor (see "Methods") are summarized in Table II and Figures 3 and 4. For each group, Table II reports the values of *TP*, *TN*, *FP*, and *FN*; the assessment scores (*precision*, *Acc*, *MCC*, *AUC_ROC*, and *AUC_PR*); and the rank of the methods according to all five measures. Figure 3 provides a visual summary of the group performance according to the binary evaluation scores together with their 95% confidence intervals; and Figure 4 illustrates the ROC and PR curves used for calculation of the *AUC_ROC* and *AUC_PR* scores in the probability-based analysis, respectively. Note that the *AUC_PR* scores are highly correlated with the classical *AUC* scores from the ROC analysis (Pearson's correlation coefficient of 0.9) and we use only one of them (*AUC_ROC*) to describe the probability-based evaluation results in what follows.

Group *prdos-CNF* (G369) is the best performing group according to the *AUC* score. It is statistically indistinguishable from the second group in the ranking, *DISOPRED3*, and better than all other groups based on the results of the nonparametric DeLong test (Table III). The results of *DISOPRED3* (G170) are, in turn, statistically indistinguishable from those of the next two groups, *biomine_dr_mixed* (G478) and *biomine_dr_pdb_c* (G288), and better than the rest of the results.

The best four groups according to the *AUC* score maintain their leading positions also in the *MCC*-based rankings, and in general, the correlation between the *AUC* and *MCC* scores is quite high (Spearman's $\rho = 0.7$), even though the two measures are conceptually different. The *MCC*-based confidence intervals for the top five groups (the four groups mentioned above plus *DisMeta*, G168) overlap, and, as a result, the statistical significance of the differences in their performance cannot be established. At the same time, comparing their confidence intervals with those of other methods allows us to conclude that these five groups are statistically better than the remaining ones. We note that even though the *DisMeta* group is ranked high according to the *MCC* (fifth), it is way down the list according to the *AUC*. The reason is that this group used only two values of disorder

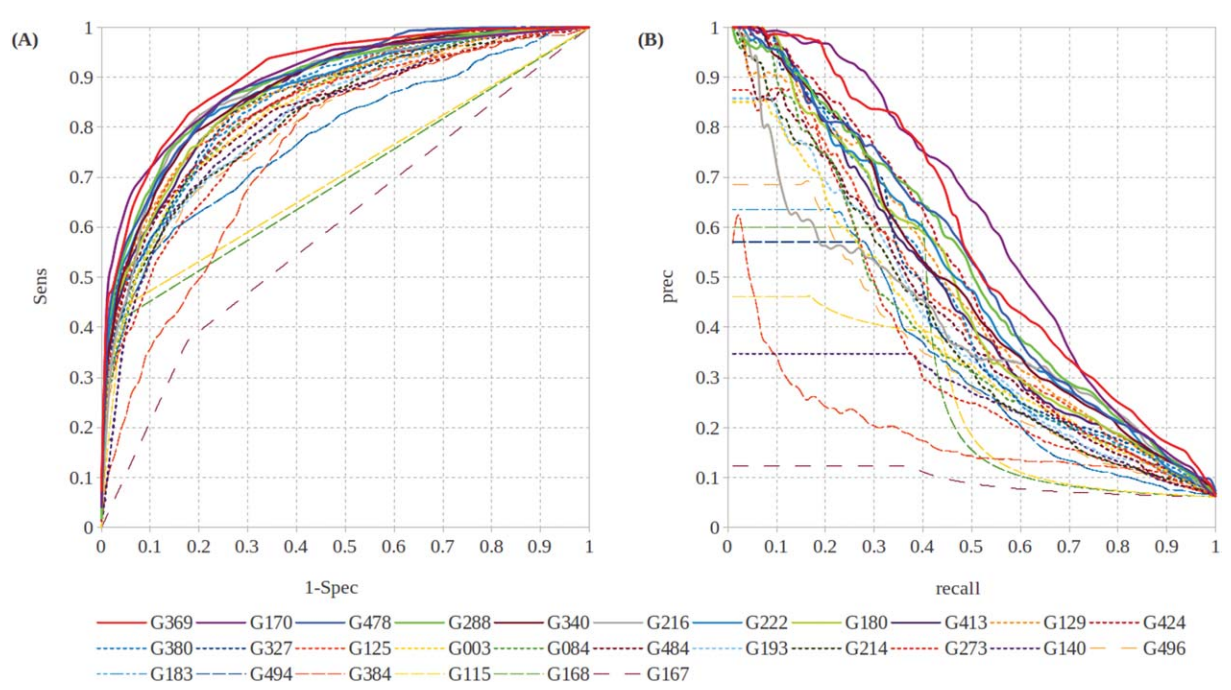

Figure 4

(A) ROC and (**B**) PR curves for the probability-based disorder region predictions for all CASP10 groups. The three groups with the atypical ROC curves in (A) used only a very limited number of disorder probability values: Groups G167 and G168 used only two different values (one for ordered and another for disordered residues); Group G115 used only five different values (two for ordered and three for disordered residues, and two out of the five values were only assigned to a very small number of residues). Groups in the legend are sorted according to decreasing *AUC_ROC* score.

probability (0.3 for the ordered residues and 0.7 for disordered) and therefore was penalized in the ROC analysis (see "Methods").

It should also be mentioned that the top four groups according to the *AUC* and *MCC* scores rank low in the *Acc*-ordered list. This difference can be explained by comparing the numbers of true and false positives for these groups with those of the subsequent groups. It can be noticed that the first four *MCC*-ranked groups have many fewer false positives than the next few groups, while maintaining comparable numbers of the true positives. This results in higher *precision* of these methods and higher *MCC* scores, which are known to adequately favor the higher-precision groups. At the same time, the *Acc* scores are higher for the second-tier groups as *Acc* strongly favors "greedy" classifications (i.e., predicting more residues as disordered, even at the cost of a larger fraction of wrong predictions). This type of behavior is a well-known feature of predictions on highly imbalanced data.

Table III

Results of Nonparametric DeLong Tests of Comparison of the Performance of the Best 12 Groups According to the AUC (ROC) Scores

	1	2	3	4	5	6	7	8	9	10	11	12
1. Prdos-CNF	x											
2. DISOPRED3	*0.093*	x										
3. biomine_dr_mixed	<0.01	*0.247*	x									
4. biomine_dr_pdb_c	<0.01	*0.090*	*0.541*	x								
5. metaprdos2	<0.01	<0.01	*0.095*	*0.311*	x							
6. POODLE	<0.01	<0.01	0.021	*0.104*	*0.534*	x						
7. MULTICOM-construct	<0.01	<0.01	0.012	*0.065*	*0.387*	*0.798*	x					
8. Yang test	<0.01	<0.01	<0.01	0.036	*0.281*	*0.661*	*0.867*	x				
9. ZHOU-SPARKS-X	<0.01	<0.01	<0.01	0.015	*0.157*	*0.438*	*0.619*	*0.730*	x			
10. CASPITAv2	<0.01	<0.01	<0.01	<0.01	<0.01	0.033	*0.064*	*0.079*	*0.151*	x		
11. MULTICOM-novel	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.019	0.023	*0.052*	*0.654*	x	
12. Espritz	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	0.01	0.012	0.03	*0.527*	*0.857*	x

The cells contain the *P* values. Values in italics represent the statistically indistinguishable cases at the 0.05 confidence level.

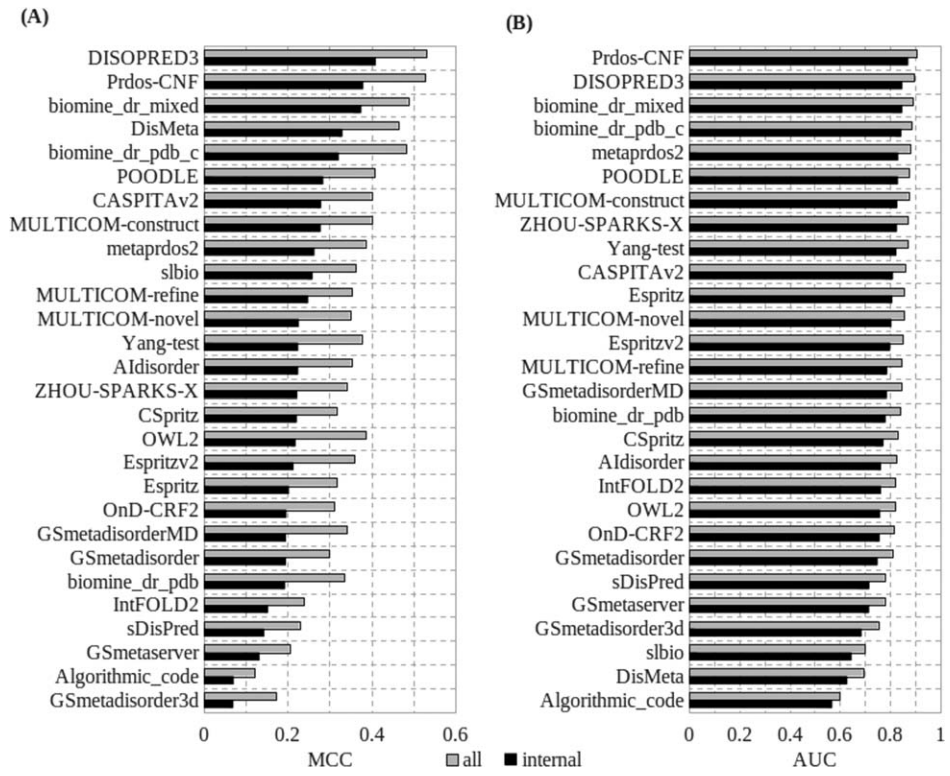

Figure 5

Comparison of group performance on the full-length and termini-trimmed targets according to the (**A**) *MCC* and (**B**) *AUC* scores. Scores on the trimmed targets are marked as "internal."

The comparison of CASP10 methods with the *Naïve* disorder predictor shows that practically all participating groups are better binary classifiers than the *Naïve*, which is near the bottom of Table II according to the *MCC* and *Acc* scores. The *Naïve* method is statistically worse than all 26 higher ranked methods when the *Acc* confidence intervals are compared and worse than the majority of the methods when the *MCC* confidence intervals are compared. It does have a quite high *precision* score, but this score is overinflated as, by definition, the method identifies only $9 + 4 = 13$ termini residues as disordered for each target, and the probability of finding disordered residues at termini is relatively high.

Prediction of internal disorder

As unstructured residues are more abundant at the termini (see "Methods"), prediction of the internal disordered regions is expected to be a harder problem. To assess the ability of methods to predict disordered residues inside the protein sequence, we repeated our analysis after trimming 10 residues from each terminal of the targets. Indeed, this analysis showed that the group performance dropped as the average *MCC* score decreased from 0.35 to 0.23 and the average *AUC* decreased from 0.83 to 0.77 (see Fig. 5). The relatively small change in *AUC* scores is an indication that the groups' algorithms are almost equally accurate in assigning disorder probabilities to residues inside the proteins and at the termini. Figure 5 also shows that the ranks of the groups remained essentially unchanged.

Prediction of long disordered regions

The functional roles of proteins containing short or long unstructured regions are likely to be different. To test the predictive power of methods to identify longer disordered regions, we have re-evaluated predictions after setting a minimum length for the disordered regions.

Figure 6 shows the *MCC* and *AUC* scores for four minimum length cutoffs: 4, 20, 30, and 40 residues. There is no clear general tendency in the data. According to the *AUC*, the average score (corresponding to the thicker "AVRG" line) remains unchanged for the cutoffs of 4, 20, and 30 residues and drops considerably for the cutoff of 40. The MCC scores show a trend to decrease with the increase of the disorder segment length. The average *MCC* is at the 0.35 level when calculated on segments longer than three residues and decreases to 0.25 and 0.20 on those not shorter than 20 and 30 residues, respectively. Further, it falls to a level just above that for a random predictor ($MCC = 0$) for segments of 40

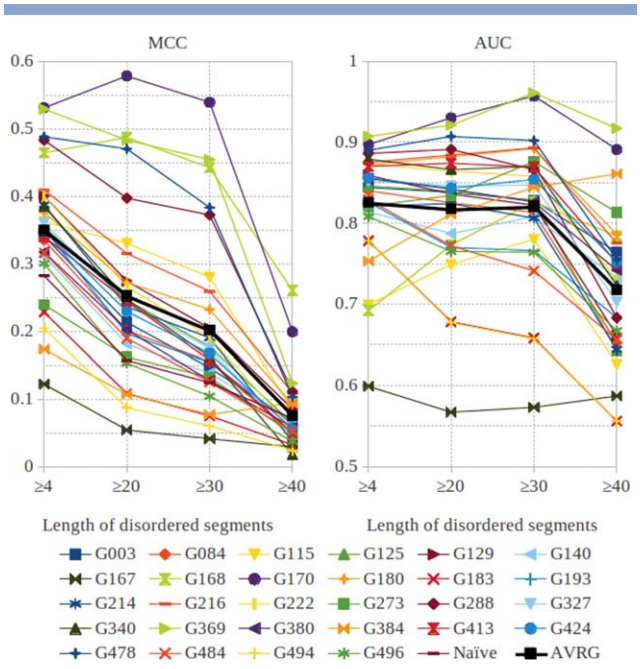

Figure 6

Comparison of group performance for four different thresholds of the minimum length of disordered regions. The two panels show data for two different evaluation measures (*MCC* and *AUC*). Each group is shown with a different color; groups in the legend are sorted according to the *AUC* score (across and then down); the artificial average group ("AVRG," black thicker line) is added to the graph for reference.

residues or longer. It should be noted, though, that the results for the latter case should be interpreted with abundance of caution as the dataset consists of only two segments.

The *Disopred3* (G170) and *prdos-CNF* (G369) are the two best performing groups across a wide range of disorder region lengths according to both the *MCC* and *AUC* scores. Their results are shown to get better with the increase of the disorder region length cutoff from 4 to 20 to 30 according to the *AUC* measure. The *Disopred3* (G170) and *DisMeta* (G168) groups show better [comparable] results on the ≥20- [≥30-] residue-long disorder regions than on the ≥4-residue-long segments according to the *MCC*.

Comparison of CASP10 predictions with those from the D2P2[28] and MobiDB[29] databases

In 2012, the year of the CASP10 experiment, two databases of protein disorder were made available in addition to the already existing repository of experimentally determined disordered regions DisProt[12] and IDEAL.[13] The Database of Disordered Protein Prediction (D2P2) contains disorder predictions for protein sequences from 1765 complete proteomes and their variants generated by six disorder prediction methods: VL-XT, VSL2b, PrDOS, PV2, ESpritz, and IUPred. The MobiDB database contains proteins with experimental disorder annotations (covering the entire PDB) and predictions for all SwissSprot[39] sequences from three disorder predictors: ESpritz, IUpred, and DisEMBL. It is interesting to compare disorder predictions in CASP10 with the corresponding entries in these two databases. The results on the common set of targets (see "Methods") are provided in Table IV (comparisons on the maximum sets of CASP10 targets overlapping with entries in each of the two databases are also provided in the Supporting Information Tables S1 and S2).

The scores of the best CASP10 methods according to all evaluation measures are higher than the scores of the methods used to produce the data stored in the databases. Interestingly, a few of the prediction methods participating in CASP have also contributed to the databases (ESpritz and prDOS series of methods). Table IV shows that the scores achieved by these techniques in CASP are different (usually substantially higher) than the scores calculated for the corresponding entries in the databases. This indicates that the methods might have been tuned differently for CASP10 and the databases. Our communications with the authors revealed that indeed this was the case. The methods were trained on different databases, with different disorder definitions, and different optimization functions. For example, for the databases, both Espritz and prDOS were trained to keep the false positive rate at a level of around 5%, while for CASP Espritz was tweaked to maximize the *Acc* score and prDOS was adjusted to gain balance between the high *MCC* and *Acc* scores. Besides being trained differently, the methods were also run using different modes of operation. For example, the Spritz series of methods was run using the PSI-BLAST sequence profiles for CASP and without them (to speed up the calculations) for the databases. All this information highlights the importance of using different flavors of methods depending on which aspect of the results is more important for the specific purpose.

Progress in the recent CASPs?

To address the question of whether there is progress in the field, we compared the performance of groups participating in the last four rounds of CASP. Since the definition of disorder was slightly modified after CASP8 (i.e., differences in different chains of X-ray structures were treated similarly to the differences in models from NMR ensembles), we reevaluated the CASP7 and CASP8 results according to the procedures and measures used in this article.

The *MCC* and *AUC* scores for the best twelve performing groups are presented in Figure 7. The data indicate that the best CASP10 methods are moderately more

Table IV
Comparison of CASP10 Disorder Predictors with the Methods Contributing to the D2P2 and MobiDB Databases on the Common Set of 51 CASP10 Targets

											Ranks			
ID	Group name	Targ	TP	FP	TN	FN	prec	Acc	MCC	AUC	prec	Acc	MCC	AUC
170	DISOPRED3	51	359	119	12531	431	0.751	0.723	**0.565**	**0.92**	1	15	1	1
478	biomine_dr_mixed	51	347	187	12463	443	0.65	0.712	**0.511**	**0.898**	4	18	2	3
369	Prdos-CNF	51	313	158	12492	477	0.665	0.692	**0.491**	**0.918**	2	22	3	2
168	DisMeta	51	356	262	12388	434	0.576	0.715	**0.483**	**0.715**	5	16	4	29
216	POODLE	51	528	914	11736	262	0.366	0.798	**0.453**	**0.89**	9	1	5	4
288	biomine_dr_pdb_c	51	259	135	12515	531	0.657	0.659	**0.442**	**0.877**	3	28	6	5
222	MULTICOM-construct	51	489	1083	11567	301	0.311	0.767	**0.39**	**0.871**	15	4	7	8
129	CASPITAv2	50	405	758	11580	385	0.348	0.726	**0.378**	**0.842**	11	12	8	13
180	Yang test	51	442	952	11698	348	0.317	0.742	**0.374**	**0.871**	13	8	9	7
424	MULTICOM-novel	51	493	1250	11400	297	0.283	0.763	**0.368**	**0.867**	18	6	10	11
340	metaprdos2	47	448	1227	10317	246	0.267	0.77	**0.363**	**0.876**	22	3	11	6
115	SIbio	48	314	535	11497	426	0.37	0.69	**0.356**	**0.707**	8	23	12	30
003	GSmetadisorderMD	48	389	852	11003	362	0.313	0.723	**0.354**	**0.867**	14	13	13	10
214	OWL2	48	256	371	11475	438	0.408	0.669	**0.354**	**0.82**	6	27	14	18
327	Espritzv2	51	332	605	12045	458	0.354	0.686	**0.344**	**0.838**	10	24	15	15
125	MULTICOM-refine	51	536	1748	10902	254	0.235	0.77	**0.338**	**0.836**	27	2	16	16
413	ZHOU-SPARKS-X	51	528	1718	10932	262	0.235	0.766	**0.336**	**0.868**	28	5	17	9
—	D2P2_Espritz-X	51	265	389	12261	525	0.405	0.652	**0.333**	**0.652**	7	29	18	—
193	AIdisorder	47	359	834	10810	392	0.301	0.703	**0.329**	**0.83**	16	21	19	17
380	Espritz	51	463	1388	11262	327	0.25	0.738	**0.325**	**0.84**	24	10	20	14
—	D2P2_PrDOS	51	463	1439	11211	327	0.243	0.736	**0.319**	**0.736**	26	11	21	—
140	OnD-CRF2	51	432	1276	11374	358	0.253	0.723	**0.315**	**0.812**	23	14	22	20
484	CSpritz	51	485	1620	11030	305	0.23	0.743	**0.314**	**0.808**	29	7	23	21
084	biomine_dr_pdb	51	403	1247	11403	387	0.244	0.706	**0.295**	**0.817**	25	20	24	19
496	GSmetadisorder	51	420	1485	11165	370	0.22	0.707	**0.279**	**0.792**	33	19	25	23
273	IntFOLD2	51	562	2870	9780	228	0.164	0.742	**0.261**	**0.845**	37	9	26	12
—	MobiDB_Dise-465	51	238	583	12067	552	0.29	0.628	**0.251**	**0.7**	17	34	27	31
—	D2P2_PV2	51	497	2544	10106	293	0.163	0.714	**0.241**	**0.714**	38	17	28	—
—	D2P2_IUPred-S	51	215	556	12094	575	0.279	0.614	**0.231**	**0.614**	19	36	29	—
—	MobiDb_IUPed-S	51	215	556	12094	575	0.279	0.614	**0.231**	**0.614**	20	37	30	33
***	*Naïve*	51	161	314	12336	629	0.339	0.589	**0.228**	**0.788**	12	39	31	—
—	D2P2_Espritz-N	51	289	1021	11629	501	0.221	0.643	**0.226**	**0.643**	32	31	32	—
—	MobiDB_Espritz-N	51	272	964	11686	518	0.22	0.634	**0.218**	**0.727**	34	33	33	28
—	MobiDB_Espritz-X	51	290	1185	11465	500	0.197	0.637	**0.206**	**0.745**	35	32	34	25
384	GSmetadisorder3d	49	167	450	11587	601	0.271	0.59	**0.2**	**0.805**	21	38	35	22
183	sDisPred	51	474	2949	9701	316	0.138	0.683	**0.198**	**0.74**	41	25	36	26
—	D2P2_VSL2b	51	355	1979	10671	435	0.152	0.646	**0.182**	**0.646**	40	30	37	—
494	GSmetaserver	51	519	3814	8836	271	0.12	0.678	**0.179**	**0.74**	43	26	38	27
—	D2P2_IUPred-L	51	148	518	12132	642	0.222	0.573	**0.159**	**0.573**	30	41	39	—
—	MobiDB_IUPred-L	51	148	518	12132	642	0.222	0.573	**0.159**	**0.57**	31	42	40	35
—	MobiDB_Dise-HL	51	416	3415	9235	374	0.109	0.628	**0.134**	**0.653**	45	35	41	32
—	D2P2_IUPred-A	51	112	484	12166	678	0.188	0.552	**0.118**	**0.552**	36	44	42	—
167	Algorithmic_code	51	265	2134	10516	525	0.11	0.583	**0.102**	**0.583**	44	40	43	34
—	MobiDB_Espritz-D	51	78	413	12237	712	0.159	0.533	**0.083**	**0.757**	39	45	44	24
—	D2P2_VLXT	51	236	2155	10495	554	0.099	0.564	**0.079**	**0.564**	46	43	45	—
—	D2P2_Espritz-D	51	78	496	12154	712	0.136	0.53	**0.069**	**0.53**	42	46	46	—

The names of the database methods are precluded with the corresponding database name. The groups are ranked according to the MCC score. The *Naïve* predictor and methods from the D2P2 database are assessed in the binary mode only. The database methods are marked with the gray background. The two main evaluation scores (MCC for two-class estimations and AUC for probability-based predictions) are bold-faced.

accurate than the best CASP9 and CASP7 methods, and approximately as accurate as the best CASP8 methods.

CONCLUSIONS

The CASP10 experiment tested performance of 28 disorder prediction methods on 94 test sequences. Four prediction groups—Prdos-CNF, DISOPRED3, biomine_dr_mixed, and biomine_dr_pdb_c—perform better than the others according to the majority of the evaluation measures. The scores of the best CASP10 groups are slightly higher than those of the best CASP9 groups, potentially indicating a (modest) progress. It should be mentioned, though, that this conclusion should be taken with a grain of salt, as measuring progress is always a tricky business as targets, methods, and databases change in time.

As in previous CASPs, prediction targets were not optimal for the evaluation of disorder prediction as the

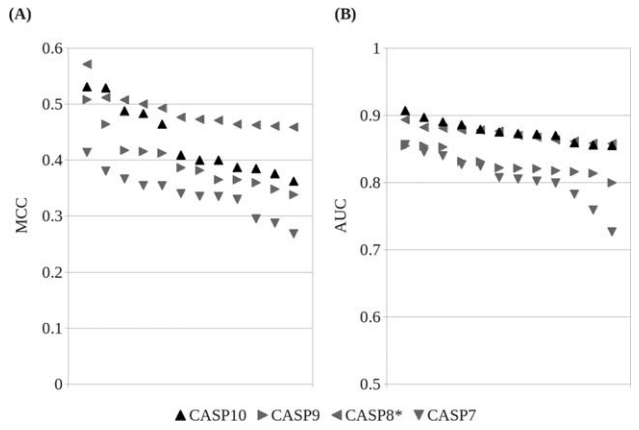

Figure 7

Comparison of performance for the best 12 groups in the last four CASPs according to the *MCC* (**A**) and *AUC* (**B**) scores. Groups in each panel and are sorted according to decreasing scores in each CASP. CASP8* scores are calculated without a very long and completely unstructured target T0500, the correct prediction of which would over-inflate the evaluation scores.

vast majority of CASP10 targets were solved by X-ray crystallography and typically contained relatively short disorder regions. Obtaining a test dataset better representing the type of disorder observed in functionally relevant proteins (longer disordered regions) still remains a challenging task for the CASP organizers.

In CASP10, the standard assessment of disorder prediction was complemented with the analysis of capacity to recognize disorder regions of different lengths and at different locations along the sequence (at termini or inside the protein). We also compared CASP predictions with the entries stored in the two recently established databases for disorder predictions—D2P2 and MobiDB. The CASP10 prediction methods show better performance than methods contributing to the databases, perhaps due to differences in training (e.g., using datasets reflecting shorter disorder regions), tuning (reflecting CASP assessment criteria more closely), and execution (e.g., allowing for more elaborate calculations). This analysis shows that using a problem-tuned approach can enhance performance by a substantial margin.

ACKNOWLEDGMENTS

The authors thank Matt Oates (D2P2) and Tomas Di Domenico (MobiDB) for extracting predictions for the CASP10 sequences from the D2P2 and MobiDB databases. We also thank Silvio Tosatto (Spritz series of methods) and Takashi Ishida (prDOS series of methods) for sharing details of the training/running procedures for their methods. We are grateful to Keith Dunker for his comments on the manuscript.

REFERENCES

1. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. Identifying disorder regions in proteins from amino acid sequence. Proc IEEE Int Conf Neural Netw 1997;1:90–95.
2. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999; 293:321–331.
3. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 2000;11:161–171.
4. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.
5. Uversky VN. A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci 2013;22:693–724.
6. Hsu WL, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, Uversky VN, Dunker AK. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Sci 2013;22:258–273.
7. Ota M, Koike R, Amemiya T, Tenno T, Romero PR, Hiroaki H, Dunker AK, Fukuchi S. An assignment of intrinsically disordered regions of proteins based on NMR structures. J Struct Biol 2013;181:29–36.
8. Cozzetto D, Jones DT. The contribution of intrinsic disorder prediction to the elucidation of protein function. Curr Opin Struct Biol 2013;23:467–472.
9. Oldfield CJ, Xue B, Van YY, Ulrich EL, Markley JL, Dunker AK, Uversky VN. Utilization of protein intrinsic disorder knowledge in structural proteomics. Biochim Biophys Acta 2013;1834:487–498.
10. Uversky VN, Dunker AK. The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. F1000 Biol Rep 2013;5:1.
11. Janin J, Sternberg MJ. Protein flexibility, not disorder, is intrinsic to molecular recognition. F1000 Biol Rep 2013;5:2.
12. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK. DisProt: the Database of Disordered Proteins. Nucleic Acids Res 2007;35(database issue):D786–793.
13. Fukuchi S, Sakamoto S, Nobe Y, Murakami SD, Amemiya T, Hosoda K, Koike R, Hiroaki H, Ota M. IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. Nucleic Acids Res 2012;40(database issue):D507–511.
14. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Res 2009;19:929–949.
15. Dosztanyi Z, Meszaros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform 2010;11:225–243.
16. Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst 2012;8:114–121.
17. Peng ZL, Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 2012;13:6–18.
18. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. Proteins 2011;79 Suppl 10:107–118.
19. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. Proteins 2009;77 Suppl 9:210–216.
20. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. Proteins 2007;69 Suppl 8:129–136.
21. Jin Y, Dunbrack RL, Jr. Assessment of disorder predictions in CASP6. Proteins 2005;61 Suppl 7:167–175.
22. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003;53 Suppl 6:561–565.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
24. Huber R. Conformational flexibility in protein molecules. Nature 1979;280:538–539.

25. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82(Suppl 2):7–13.
26. Taylor T, Tai C-H, Bai H, Kryshtafovych A, Montelione G, Lee B. Definition and classification of evaluation units for CASP10. Proteins 2014;82(Suppl 2):14–25.
27. Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A. Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins 2014;82(Suppl 2):112–126.
28. Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J. D(2)P(2): Database of Disordered Protein Predictions. Nucleic Acids Res 2013;41(database issue):D508–516.
29. Di Domenico T, Walsh I, Martin AJ, Tosatto SC. MobiDB: a comprehensive database of intrinsic protein disorder annotations. Bioinformatics 2012;28:2080–2081.
30. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004;337:635–645.
31. Davis J, Goadrich, M. The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning. New York: ACM; 2006. pp 233–240.
32. Bunescu R, Ge R, Kate R, Marcotte E, Mooney R, Ramani A, Wong Y. Comparative experiments on learning information extractors for proteins and their interactions. J Artif Intell Med 2004: 139–155.
33. Kok S, Domingos, P. Learning the structure of Markov logic networks. In 22nd International Conference on Machine Learning. New York: ACM Press; 2005. pp 441–448.
34. He HB, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–1284.
35. Goadrich M, Oliphant L, Shavlik J. Learning ensembles of first-order clauses for recall-precision curves: a case study in biomedical information extraction. In 14th International Conference on Inductive Logic Programming (ILP); 2004. pp 421–456.
36. Fawcett T, Flach PA. A response to Webb and Ting's On the application of ROC analysis to predict classification performance under varying class distributions. Mach Learn 2005;58:33–38.
37. The R development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2006.
38. http://ca.expasy.org/tools/pROC/. 2011
39. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011;2011:bar009.