

DISORDER: ASSESSMENT

Assessment of disorder predictions in CASP8

Orly Noivirt-Brik,¹ Jaime Prilusky,^{2,3} and Joel L. Sussman^{1,3*}

¹Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

²Bioinformatics Unit, Weizmann Institute of Science, Rehovot 76100, Israel

³The Israel Structural Proteomics Center, Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT

The interest in intrinsically disordered proteins has greatly increased, as it has become clear that they are very widespread, especially in eukaryotic organisms. Functionally, they appear to play a significant role in the control of many cellular processes and signaling pathways and have been, also, associated with a number of diseases ranging from cancer to Alzheimer's. Thus, there is enormous interest in attempts to predict disordered regions in proteins solely from knowledge of their amino acid sequences. In this study, we assess the quality of predictions for 25 groups on predicting disordered regions in 122 target proteins. In addition, we suggest the need of a "knowledge-independent" measure that would enable one to normalize the results of the different CASP experiments and to determine whether the disorder prediction field had improved across the years.

Proteins 2009; 77(Suppl 9):210–216.
© 2009 Wiley-Liss, Inc.

Key words: CASP8; protein disorder; protein structure prediction; intrinsically disordered proteins; intrinsically unfolded proteins.

INTRODUCTION

The biennial "critical assessment of structure prediction" (CASP) experiment is a crucial way to evaluate, in an unbiased way, the progress in predicting novel 3D protein structures. Although initially CASP focused solely on prediction of structured proteins, in CASP5 analysis of the possibility of predicting "unstructured" or "disordered" regions of proteins was for the first time formally addressed. This is due to the observation that a growing number of proteins have been found to be "natively unfolded" or "intrinsically disordered" under physiological conditions.^{1–9} In fact, one of the targets submitted was shown experimentally to be entirely unstructured in solution, that is, target 145, the intracellular domain of the neuro-adhesion protein, gliotactin.^{7,10}

As the amino acid sequence contains the information for protein folding, it was reasoned that for proteins that do not fold into 3D structures the amino acid sequence should also specify protein nonfolding. To test this hypothesis, methods were developed to predict regions of protein sequences that fail to fold.^{4,11} The fact that predictor accuracy was significantly better than expected by chance suggested that the information for failure to fold into a 3D structure is, indeed, likely to be inherent within the amino acid sequence.

The interest in intrinsically disordered proteins (IDPs) has greatly increased, as it has become clear that they are very widespread, especially in eukaryotic organisms.^{3,9} In addition, a number of studies have shown that IDPs appear to have key functional roles, with a rough rule of thumb being that for mammals about 75% of their signalling proteins are predicted to contain long disordered regions (>30 residues), about half of their total proteins are predicted to contain such long disordered regions, and about 25% of their proteins are predicted to be fully disordered.¹² Thus, there is enormous interest in the effort to predict disordered

The authors state no conflict of interest.

Grant sponsors: Erwin Pearl, the Divadol Foundation, the Nalvyco Foundation, the Bruce Rosen Foundation, the Jean and Julia Goldwurm Memorial Foundation, the Neuman Foundation, the Kalman and Ida Wolens Foundation.

*Correspondence to: Joel L. Sussman, Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: joel.sussman@weizmann.ac.il

Received 3 May 2009; Revised 5 August 2009; Accepted 7 August 2009

Published online 21 August 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22586

Table I

Summary of the Disorder Prediction Data in CASP8

Grp	Grp name	N_t	N_{res}	Probability type	'D' Probability range	'O' probability range
68	OnD-CRF	122	27489	Continuous	≥ 0.05	≤ 0.049
69	MULTICOM-CMFR	122	27489	Continuous	≥ 0.50	≤ 0.49
73	Distill	122	27489	Continuous	≥ 0.590	≤ 0.59
97	DISOclust	122	27489	Continuous	≥ 0.576	≤ 0.576
99	Distill-Punch1	121	27380	Continuous	≥ 0.238	≤ 0.23
113	Softberry	122	27489	Continuous	≥ 0.51	≤ 0.49
116	fais-server	122	27489	Continuous	≥ 0.51	≤ 0.49
118	Oka	121	27209	9 values	≥ 0.60	$= 0.00$
129	Casplta	119	26912	Continuous	≥ 0.50	≤ 0.49
133 ^a	GSmetaDisorder	118	26701	Continuous	$0 \leq p \leq 1$	$0 \leq p \leq 1$
153 ^a	GS-MetaServer2	122	27489	Continuous	≥ 0.21	≤ 0.30
157	3Dpro	117	25468	Continuous	≥ 0.50	≤ 0.49
161	CBRC_POODLE	122	27489	Continuous	≥ 0.50	≤ 0.49
167 ^a	Spritz2	122	27489	Continuous	≥ 0.04	≤ 0.50
229 ^a	CBRC-DP_DR	121	27381	Continuous	≥ 0.435	≤ 0.50
238	Distill-Punch2	121	27380	Continuous	≥ 0.201	≤ 0.201
293 ^a	LEE-SERVER	108	23306	Continuous	≥ 0.10	≤ 0.99
297 ^a	GeneSilicoMetaServer	122	27489	Continuous	≥ 0.18	≤ 0.3
359	Biomine	119	26224	Continuous	≥ 0.501	≤ 0.5
379	McGuffin	121	27209	Continuous	≥ 0.576	≤ 0.576
388	DISOPRED	122	27489	Continuous	≥ 0.50	≤ 0.49
407 ^a	LEE	118	26845	Continuous	≥ 0.50	$0 \leq p \leq 0.9$
433	metaprdos	116	25834	Continuous	≥ 0.51	≤ 0.49
450	mariner1	120	26619	Continuous	≥ 0.600	≤ 0.5
453	MULTICOM	121	27209	Continuous	≥ 0.50	≤ 0.49

^aGroups that used overlapping values of probability for their disorder ('D') and order ('O') state predictions.

regions of proteins, and a large number of disorder predictions tools have been developed (see a listing at the Database of Protein Disorder: <http://www.ist.temple.edu/disprot/predictors.php>).

In this study, we analyze how well groups were able to predict disordered regions in a series of proteins given only their amino acid sequences. We also suggest the need of a “knowledge-independent” measure that would enable one to normalize the results of the different CASP experiments, and to determine whether the disorder prediction field had improved across the years.

METHODS

CASP8 dataset and disorder definition

Disorder regions of proteins are regions that have no unique 3D structure under physiological conditions. It is difficult to determine whether a certain residue is disordered *in vivo* using only X-ray crystal structures as order might be induced simply by the crystallization conditions. However, for the assessment of disorder of the X-ray targets in CASP8, and as in previous CASP experiments, the classification of residues to be ordered or disordered was defined only by that information. Thereby, all the residues of the X-ray targets of CASP8 that appeared at the sequence but lack atomic coordinates in the crystal structure were labelled as “disorder” whereas all other residues were labelled as “order”. On the other hand, the classification of disorder

for the NMR targets was done manually, which we will refer to as visual assessment (VA) by Jeremy Block and Jane Richardson. Specifically, mainchain atoms for the deposited ensemble of models were displayed in KiNG, and each end of each of the less-tight regions was examined in turn. The “cocentering” function in KiNG that translationally superimposes all models on a given atom¹³ was used at successive C_α atoms to see whether the local backbone conformations were closely similar or nearly uncorrelated. There was usually a fairly sharp transition in local agreement that was taken to define the order/disorder boundary. It is important to note that disorder segments shorter than four residues long were not considered in the evaluation process to reduce noise that derives from experimental uncertainties.

A total of 25 groups submitted predictions for the disorder category. Each group could submit up to five predictions for each target, but only four groups submitted more than one prediction. Therefore, only the first model was used for the disorder assessment. For each residue, the predictors were asked to submit a binary label of “O” or “D” (order or disorder state) and a probability that the specific residue is in a disordered region (a value in the range of 0–1). Regions that could not be predicted were to be assigned with a probability of 0.5. It can be seen from Table I that not all prediction groups followed the requested instructions and, thus, reduced their chances to be ranked at the top.

Disorder assessment in CASP8 consisted of the evaluation of 122 targets of which 103 were X-ray structures

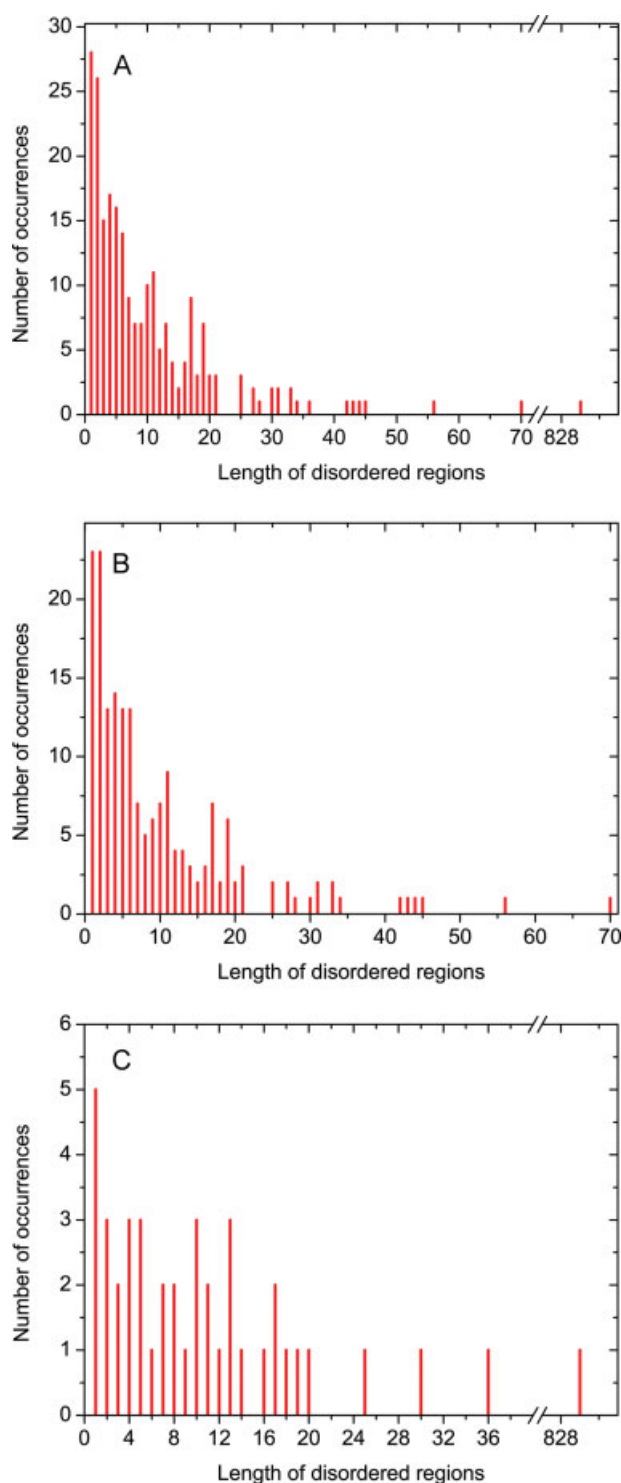


Figure 1

Length distribution of the disordered regions in the CASP8 data set. Length distribution of the disorder regions of all 122 targets (A), of the 103 X-ray targets (B), and of the 19 NMR targets (C). Note that disorder segments shorter than four residues long were not considered in the evaluation process.

and 19 were NMR structures. Six targets were eliminated from the disorder evaluation. Targets T0387, T0403, and T0439 were cancelled by the organizers. Targets T0498 and T0499 were cancelled, as they share very similar sequences. Although not identical, however, their structures are very different, and they contain different disordered regions. These two targets, in hindsight, would be very interesting to examine, as they are heavily engineered and the natural selection of disorder supporting sequences would not have occurred. Thus, it would have been interesting to see how the disorder predicting methods behaved on these targets. Target T0390, which is a ligand that was crystallized with its receptor, was cancelled as the disorder profile might have changed because of that interaction. The full set of targets was partitioned to “only X-ray” and “only NMR” structure sets, and the disorder analysis was applied separately for each of these sets. Overall 27,489 residues were considered out of which 10.7% were classified as disordered. The fraction of disorder in specific targets ranged from 0 to 100%. Target T0500 is fully disordered and very long (i.e., it contains 829 residues) and thus played a central role on the assessment results.

Evaluation criteria

The disorder assessment in CASP8 was implemented according to the same criteria as in previous CASP experiments,^{14,15} and therefore, we will describe it briefly. The assessment was based on per-residue level predictions of the entire set of targets. The statistical significance of the evaluation scores was determined by a bootstrapping procedure: 80% of the targets were ran-

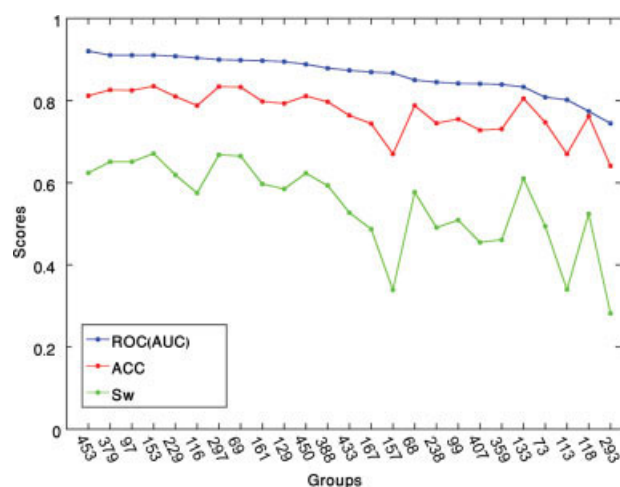


Figure 2

Assessment scores of all prediction groups participating in CASP8. ACC (in red), S_w (in green) and AUC (in blue) scores are sorted in descending order of the AUC score.

Table II

Summary of Evaluation Scores for All Disorder Prediction Groups in CASP8

Grp	S_{sens}	S_{spec}	ACC	S_w	AUC
153	0.758 ± 0.048	0.904 ± 0.004	0.831 ± 0.024	0.662 ± 0.048	0.9078 ± 0.0186
297	0.741 ± 0.050	0.920 ± 0.003	0.830 ± 0.025	0.661 ± 0.050	0.8967 ± 0.0208
69	0.796 ± 0.039	0.864 ± 0.004	0.830 ± 0.020	0.660 ± 0.039	0.8958 ± 0.0185
97	0.727 ± 0.047	0.917 ± 0.004	0.822 ± 0.024	0.644 ± 0.047	0.9083 ± 0.0176
379	0.706 ± 0.053	0.938 ± 0.004	0.822 ± 0.026	0.644 ± 0.053	0.9084 ± 0.0158
450	0.694 ± 0.040	0.927 ± 0.003	0.811 ± 0.020	0.621 ± 0.040	0.8857 ± 0.0159
453	0.641 ± 0.061	0.978 ± 0.001	0.809 ± 0.030	0.619 ± 0.061	0.9184 ± 0.0152
229	0.657 ± 0.049	0.955 ± 0.003	0.806 ± 0.025	0.612 ± 0.049	0.9052 ± 0.0180
133	0.711 ± 0.044	0.894 ± 0.004	0.802 ± 0.022	0.605 ± 0.044	0.8295 ± 0.0245
161	0.646 ± 0.066	0.942 ± 0.004	0.794 ± 0.033	0.588 ± 0.066	0.8953 ± 0.0206
388	0.626 ± 0.067	0.957 ± 0.002	0.792 ± 0.033	0.583 ± 0.067	0.8764 ± 0.0218
129	0.629 ± 0.059	0.951 ± 0.004	0.790 ± 0.030	0.579 ± 0.059	0.8912 ± 0.0207
68	0.853 ± 0.030	0.719 ± 0.015	0.786 ± 0.016	0.572 ± 0.032	0.8484 ± 0.0158
116	0.600 ± 0.055	0.966 ± 0.001	0.783 ± 0.027	0.566 ± 0.055	0.9011 ± 0.0174
433	0.556 ± 0.063	0.964 ± 0.003	0.760 ± 0.031	0.520 ± 0.063	0.8711 ± 0.0224
99	0.634 ± 0.026	0.879 ± 0.005	0.756 ± 0.013	0.513 ± 0.026	0.8427 ± 0.0088
118	0.574 ± 0.083	0.935 ± 0.004	0.755 ± 0.041	0.509 ± 0.083	0.7678 ± 0.0413
238	0.602 ± 0.028	0.893 ± 0.005	0.748 ± 0.014	0.495 ± 0.028	0.8456 ± 0.0087
73	0.689 ± 0.047	0.801 ± 0.007	0.745 ± 0.024	0.490 ± 0.047	0.8044 ± 0.0289
167	0.513 ± 0.054	0.969 ± 0.002	0.741 ± 0.027	0.482 ± 0.054	0.8667 ± 0.0220
359	0.507 ± 0.019	0.954 ± 0.002	0.731 ± 0.010	0.461 ± 0.019	0.8396 ± 0.0104
407	0.465 ± 0.077	0.982 ± 0.001	0.724 ± 0.039	0.447 ± 0.077	0.8371 ± 0.0274
157	0.349 ± 0.018	0.990 ± 0.001	0.670 ± 0.009	0.339 ± 0.018	0.8669 ± 0.0104
113	0.366 ± 0.120	0.953 ± 0.005	0.659 ± 0.060	0.319 ± 0.120	0.7954 ± 0.0381
293	0.306 ± 0.014	0.976 ± 0.002	0.641 ± 0.007	0.282 ± 0.013	0.7441 ± 0.0127

Highest value of each score is highlighted in bold.

domly selected for each group 1000 times, and the scores standard error was calculated.

There are two levels for disorder predictions: (i) the ability to determine absolute disorder by a binary score and (ii) the ability to evaluate the confidence level of a disorder prediction by the disorder probability. The binary classification of each group was assessed by the following scores:

$$\text{Sensitivity} = S_{\text{sens}} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{N_{\text{disorder}}}$$

$$\text{Specificity} = S_{\text{spec}} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{N_{\text{order}}}$$

where TP is the number of true positives (disordered residues that were classified correctly), FP false positives (ordered residues that were classified as disordered), TN true negatives (ordered residues that were classified correctly), and FN false negative (disordered residues that were classified as ordered), respectively. The higher these two scores, the better the predictions; therefore, they were combined into a single score, which is the average of the two:

$$\text{ACC} = \frac{S_{\text{sens}} + S_{\text{spec}}}{2}.$$

As was previously pointed out, the calculation of a simple percentage of accuracy score (Q2) will not be very in-

dicative of the disorder prediction power, as disordered residues are very rare. For that reason, a weighted score rewarding correctly disorder prediction more than order prediction was developed¹⁴:

$$S_w = \frac{S}{S_{\text{max}}} = \frac{W_{\text{disorder}}\text{TP} - W_{\text{order}}\text{FP} + W_{\text{order}}\text{TN} - W_{\text{disorder}}\text{FN}}{W_{\text{disorder}}N_{\text{disorder}} + W_{\text{order}}N_{\text{order}}}$$

where the W_{disorder} was set to the total percent of order and W_{order} was set to the total percent of disorder (for groups that predicted all 122 targets, $W_{\text{disorder}} = 0.893$ and $W_{\text{order}} = 0.107$). S_w is in the range of -1 to 1 and is equal to zero, when all residues are predicted to be ordered.

The receiver operating characteristic (ROC) curve was used to examine the ability of the predictors to estimate the confidence level of their predictions. This is reflected in the ability of the predictors to rank their predictions correctly by the disorder probability parameter. As was previously described for each value of P in increments of 0.01 (from 0 to 1), all the residues with probability equal or greater than P are set as disordered, and all other residues are set as ordered. Then, the TP-rate and the FP-rate are calculated, and a full curve is obtained. The area under the curve (AUC) is the measure that is used to evaluate the different groups. It is important to note that this score is a more indicative of quality of predictions, as the number of probabilities set by the predictors is larger (i.e. more continuous). The ROC curve analysis is

based on the disorder probability parameter given by the predictors, and thus, it is most important that the range of probabilities that are designated as “disordered residues” must not overlap those designated as “ordered residues”. Unfortunately, several groups had overlapping ranges of probabilities (Table I) and hence made it impossible for us to evaluate their predictions using this measure.

RESULTS

As in previous CASP experiments, the dataset is dominated by short disorder segments (see Fig. 1); however, it is important to note that due to the existence of one very long disordered region (above 800 residues long) the results are also dramatically influenced by the ability to predict long disorder segments. In addition, the probability to be disordered is much higher at both tails of the polypeptide chain than in the middle (data not shown).

Figure 2 shows the ACC, S_w , and AUC scores for the 25 groups that participated in the disorder category of CASP8 sorted by the AUC score. It can be seen that the correlation between the AUC score and the S_w or the ACC scores is low. This may stem from the fact that the AUC score fails in the assessment of groups that used small number of distinct probability values (as group 118) or used an overlapping probability ranges both for order and disorder residues (these groups are marked with asterisk in Table I) whereas the binary scores are not affected by that. Another explanation for the low correlation can be a nonoptimal selection of the probability threshold that distinguishes disorder state from the order state in the binary prediction.

Table II summarizes all the prediction evaluation scores sorted by the ACC score. According to both the ACC and the S_w measure, group 153 (Kaminski, GS-MetaServer2) shows the best performance tightly followed by groups 297 (Kaminski, GeneSilicoMetaServer), 69 (Cheng & Wang, MULTICOM-CMFR), and 97 (McGuffin, DISOclust.). Although, according to the AUC score, group 453 (Cheng & Wang, MULTICOM) was ranked at the top. Both group 68, that had the highest sensitivity, and group 157, that had the highest specificity, have quite low rankings due to the fact that there is a trade-off between these two parameters. Classifiers that set disorder or order in a biased manner, that is, biased toward one of the states, tend to outperform when considering only one of these parameters but fail in the other. Figure 3(A) displays the confusion matrix of each prediction group calculated over all the 122 targets as bar plots sorted by the sum of true predictions (TP+TN). Here, we can see that ordered residues are much easier to predict than disordered residues, and therefore, group 69, which has relatively low level of true predictions

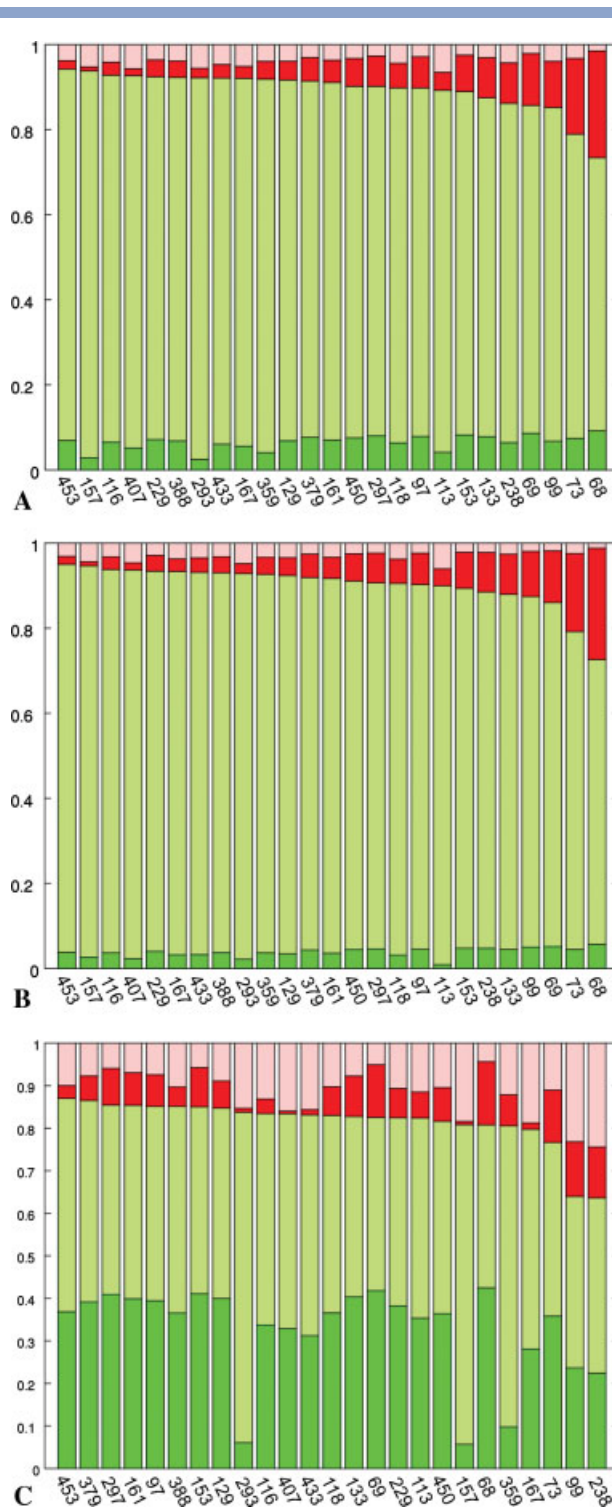


Figure 3

Bar plot of the confusion matrices for all prediction groups in CASP8. The confusion matrix displays the percentage of TP (dark green), TN (light green), FP (dark red) and FN (light red) for all prediction groups and are based on all the 122 targets (A), on the 103 X-ray targets (B) and on the 19 NMR targets classified by the VA procedure (C). The groups are sorted by the fraction of true predictions (TP+TN).

Table III

Summary of Evaluation Scores for All Disorder Prediction Groups in CASP8 by the 19 NMR Targets Classified Visually (i.e. T0437, T0460, T0462, T0464, T0466, T0467, T0468, T0469, T0471, T0472, T0473, T0474, T0475, T0476, T0480, T0482, T0484, T0492, T0500)

Grp	S_{sens}	S_{spec}	ACC	S_w	AUC
379	0.790 ± 0.136	0.892 ± 0.013	0.841 ± 0.068	0.682 ± 0.136	0.915 ± 0.055
453	0.730 ± 0.167	0.944 ± 0.009	0.837 ± 0.084	0.675 ± 0.167	0.914 ± 0.060
297	0.836 ± 0.118	0.837 ± 0.02	0.837 ± 0.059	0.674 ± 0.119	0.915 ± 0.055
161	0.804 ± 0.141	0.857 ± 0.03	0.831 ± 0.071	0.662 ± 0.142	0.901 ± 0.070
153	0.834 ± 0.119	0.827 ± 0.022	0.83 ± 0.06	0.661 ± 0.12	0.921 ± 0.056
97	0.794 ± 0.133	0.86 ± 0.018	0.827 ± 0.067	0.654 ± 0.134	0.904 ± 0.061
129	0.774 ± 0.127	0.876 ± 0.015	0.825 ± 0.063	0.65 ± 0.127	0.903 ± 0.059
388	0.726 ± 0.182	0.914 ± 0.013	0.82 ± 0.091	0.639 ± 0.182	0.863 ± 0.088
133	0.800 ± 0.106	0.817 ± 0.024	0.809 ± 0.054	0.617 ± 0.107	0.858 ± 0.064
69	0.850 ± 0.107	0.767 ± 0.022	0.808 ± 0.054	0.617 ± 0.108	0.890 ± 0.065
229	0.744 ± 0.128	0.867 ± 0.03	0.805 ± 0.065	0.611 ± 0.13	0.891 ± 0.065
116	0.662 ± 0.163	0.936 ± 0.007	0.799 ± 0.082	0.599 ± 0.163	0.895 ± 0.058
68	0.878 ± 0.087	0.718 ± 0.028	0.798 ± 0.046	0.597 ± 0.091	0.877 ± 0.050
118	0.712 ± 0.207	0.874 ± 0.02	0.793 ± 0.104	0.586 ± 0.208	0.826 ± 0.118
407	0.597 ± 0.211	0.988 ± 0.003	0.793 ± 0.105	0.585 ± 0.211	0.876 ± 0.073
450	0.734 ± 0.127	0.851 ± 0.014	0.793 ± 0.064	0.585 ± 0.127	0.855 ± 0.056
433	0.603 ± 0.191	0.977 ± 0.006	0.79 ± 0.096	0.579 ± 0.191	0.910 ± 0.056
113	0.685 ± 0.227	0.884 ± 0.033	0.785 ± 0.114	0.57 ± 0.228	0.848 ± 0.100
167	0.531 ± 0.195	0.973 ± 0.009	0.752 ± 0.098	0.504 ± 0.196	0.865 ± 0.069
73	0.709 ± 0.165	0.769 ± 0.019	0.739 ± 0.083	0.478 ± 0.166	0.806 ± 0.100
359	0.445 ± 0.036	0.907 ± 0.009	0.676 ± 0.019	0.352 ± 0.037	0.772 ± 0.020
293	0.283 ± 0.021	0.989 ± 0.003	0.636 ± 0.011	0.271 ± 0.022	0.657 ± 0.016
99	0.502 ± 0.018	0.758 ± 0.026	0.63 ± 0.015	0.26 ± 0.031	0.713 ± 0.020
238	0.481 ± 0.017	0.776 ± 0.024	0.628 ± 0.014	0.256 ± 0.028	0.716 ± 0.021
157	0.236 ± 0.024	0.99 ± 0.002	0.613 ± 0.012	0.226 ± 0.023	0.785 ± 0.020

Highest value of each score is highlighted in bold.

(meaning, TP+TN) but relatively high TP predictions, was ranked at the top three by both the ACC and the S_w measures. For comparison, group 68, which is the only prediction group that had higher values of TPs, has a very strong bias toward the disorder state prediction, and therefore, failed by these measures. In addition, although all the top-ranked groups have relatively high FP values, which could be misinterpreted as overprediction, it is important to note that the FP values should be only considered relative to the value of all positive predictions (meaning TP + FP). For example, group 157, which has a very low FP value, could not be ranked at the top because its predictions are very biased toward the ordered state (reflected in its very low TPs value), whereas group 153, which has higher values of FP, ranked at the top by both the ACC and the S_w measures, as its ratio between FP and TP is much larger.

For comparison, we repeated the analysis also for the “only X-ray” [Fig. 3(B)] and the “only NMR” sets [Fig. 3(C) and Table III]. It can be seen that the X-ray set contains smaller fraction of disorder residues than the NMR set, where about 45% of its residues were classified as disorder. It is important to emphasize that the definition of disorder in the “only NMR” set, which was done manually, is very reliable, and therefore, the top-ranked groups in this section should be highlighted. Although the groups ranking results obtained by the NMR dataset assessment are not drastically different from the full dataset, they cannot be neglected. Group 379 (McGuffin,

McGuffin) was ranked first, whereas in the full dataset it was ranked only fifth. Group 161 (Noguchi, CBRC_POODLE) that was previously ranked tenth is ranked fourth by the NMR set, and group 69 (Cheng & Wang, MULTICOM-CMFR) dropped from the third place to the tenth. It is interesting to note that although the fraction of disorder is very different between the sets, the magnitude of all scores remained approximately the same.

In addition, we calculated the S_w scores for each target separately to check the dependence between the fraction of disordered residues in a protein and the prediction accuracy, and we saw that as the fraction of disorder increases classification abilities decrease. However, the fully or mostly disordered proteins (T0500 and T0484) are outliers of this tendency. These targets had, on average, relatively high values of S_w , indicating that they were easier to predict. This may suggest that disorder residues are easier to identify as part of long continuous segments, and indeed, when we repeated our calculation of the S_w score using disordered segments of at least 10 or 20 residues long, almost all prediction groups got better results.

Comparison with previous CASP experiments

It would be very useful to be able to simply compare the results of CASP8 with previous CASP experiments, as

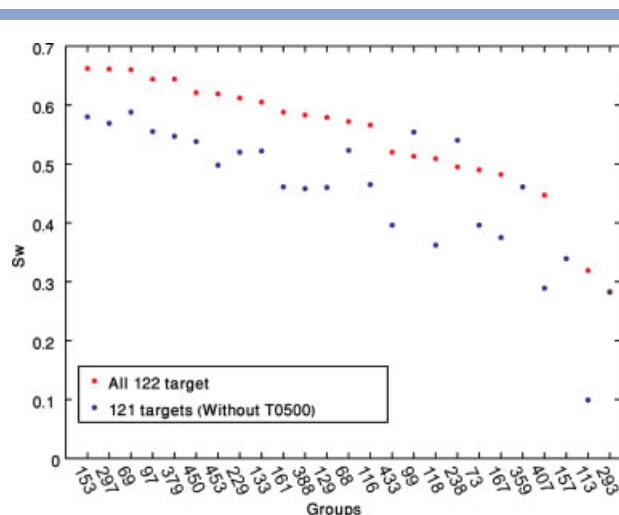


Figure 4

The effect of the dataset on the S_w score. The S_w scores that were calculated over all 122 targets (red stars) and without target T0500 (blue circles) for all prediction groups participating in CASP8 are compared. It shows that elimination of target T0500 had a dominant effect on the range of values of this score.

a way to see whether the prediction abilities were improved during the years. However, the question whether it is valid to compare different datasets without, in some way, normalizing the results for the different targets must be addressed. To approach this question, we repeated our analysis using only 121 targets eliminating target T0500, which is a long and fully disordered target. Figure 4 shows the S_w score for the two sets with and without target T0500. It shows that this single target has a dramatic effect on the score. The S_w values of most prediction groups change significantly when target T0500 was eliminated. Therefore, we suggest the need of a quantitative and objective way to compare the intrinsic properties of targets used between CASP experiments. This measure will make it possible to normalize scores between the different sets of targets and allow for an unbiased indication whether the ability to predict disorder regions has significantly improved across the years. However, this is not a simple task as virtually all the methods used for prediction of disorder are constantly being updated based on newly determined 3D structures, disordered regions in these structures, as well as proteins that are observed, in solution, to have little or no structure. In addition, the algorithms themselves are being constantly modified and improved. As currently there is no depository that stores these methods “frozen in time”, it is not obvious to compare a set of target, for example, from the current CASP8 versus CASP7 using the same program. A similar principle has been used in CASP structure prediction by Kevin Karplus (personal communication), who keeps an old version of his methods as well as the sequence and structure databases, so as to be

able to put each CASP on a difficulty scale. It would be very worthwhile for the Protein Prediction Center to consider establishing such a depository, which could be of use for all aspects of the CASP experiment. This would then provide an objective and “knowledge-independent” measure that would enable one to normalize the results of the different CASP experiments.

ACKNOWLEDGMENTS

The authors thank Jeremy Block and Jane Richardson for carefully examining the NMR targets ensembles and making a visual assessment as to which regions of the targets should be designated as disordered, and the members of the Protein Structure Prediction Center for all their help in preparation of this manuscript. J.L.S. is the Morton and Gladys Pickman Professor of Structural Biology.

REFERENCES

1. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999;293:321–331.
2. Zhang O, Forman-Kay JD. NMR studies of unfolded states of an SH3 domain in aqueous solution and denaturing conditions. *Biochemistry* 1997;36:3959–3970.
3. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Informatics* 2000;11:161–171.
4. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 2000;41:415–427.
5. Bell S, Klein C, Muller L, Hansen S, Buchner J. p53 contains large unstructured regions in its native state. *J Mol Biol* 2002;322:917–927.
6. Schweers O, Schonbrunn-Hanebeck E, Marx A, Mandelkow E. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *J Biol Chem* 1994;269:24290–24297.
7. Zeev-Ben-Mordehai T, Rydberg EH, Solomon A, Tokar L, Botti S, Auld VJ, Silman I, Sussman JL. The intracellular domain of the *Drosophila* cholinesterase-like neural adhesion protein, gliotactin, is natively unfolded. *Proteins* 2003;53:758–767.
8. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. DisProt: a database of protein disorder. *Bioinformatics* 2004;21:137–140.
9. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* 2008;18:756–764.
10. Melamud E, Moulton J. Evaluation of disorder predictions in CASP5. *Proteins* 2003;53(Suppl 6):561–565.
11. Romero P, Obradovic Z, Kissinger C, Villafrance JE, Dunker AK. Identifying disordered regions in proteins from amino acid sequence. 1997;1:91–95.
12. 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.
13. Block JN, Zielinski DJ, Chen VB, Davis IW, Vinson EC, Brady R, Richardson JS, Richardson DC. KinImmense: macromolecular VR for NMR ensembles. *Source Code Biol Med* 2009;4:3.
14. Jin Y, Dunbrack RL, Jr. Assessment of disorder predictions in CASP6. *Proteins* 2005;61(Suppl 7):167–175.
15. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. *Proteins* 2007;69:129–136.