Assessment of Disorder Predictions in CASP6

Yumi Jin and Roland L. Dunbrack, Jr.*
Institute for Cancer Research, Fox Chase Cancer Center, Philadelphia Pennsylvania

ABSTRACT **Natively disordered proteins or protein segments are those without stable secondary or tertiary structure in the absence of binding partners. Such disordered regions often are important functional sites in many biological processes, especially those involved in transcription, translation, and cell signaling. The prediction of such regions is therefore of great importance in focusing experimental efforts on regions of proteins that may be critical for function. In CASP6, held in 2004, twenty research groups participated in the prediction of disordered regions. Both binary predictions (ordered or disordered) and assigned scores for disorder were assessed. Several groups performed quite well in predicting regions of disorder in the X-ray and NMR structures available to the assessors. The best of these groups performed better than the best groups in CASP5, held in 2002. Proteins 2005;Suppl 7:167–175.** © 2005 Wiley-Liss, Inc.

Key words: CASP6; protein disorder; protein structure prediction

INTRODUCTION

The existence of natively disordered proteins or regions of proteins, those that have no fixed secondary or tertiary structure under physiological conditions, is now well documented. A number of recent reviews have surveyed both experimental and computational approaches to the study of these proteins.[1,2] They are found to participate in many biological processes and commonly occur in cell signaling pathways, DNA transcription and replication, protein translation, and endocytosis. Disordered proteins, or those with long, disordered segments (30 amino acids or longer) are more common in higher organisms, and as many as 30% of human proteins are at least in part natively disordered.[3] There are several hypotheses that attempt to explain the existence and evolution of such disordered regions. First, the flexibility of disordered regions allows such proteins to bind to multiple ligands (other proteins or DNA) with ease. Second, they may be easily regulated by post-translational modification such as phosphorylation,[4] and their concentrations may be easily modified by protease digestion. Third, they may in some instances form large contact areas with other proteins, while being much shorter in sequence than a globular protein forming a similarly sized interface.[5]

A number of groups have developed methods for the prediction of disordered regions of proteins, based on experimental data such as disorder in protein crystal structures and NMR experiments. Such methods are usually based on any of a number of classifiers from machine learning, such as neural networks and support vector machines. To compare these methods in a blind test, researchers in structure prediction were asked to submit predictions of disorder for CASP6, as they were at CASP5 in 2002. Predictions could be made on any or all of the targets in the CASP6 experiment. Twenty groups participated in disorder prediction in CASP6, compared to only six groups in CASP5.[6]

In this article, we present an assessment of the accuracy of disorder prediction in CASP6 using a number of parameters. Because the majority of targets at CASP6 were the subject of X-ray crystallographic structure determination, we were limited to assessing the prediction of disorder of segments within otherwise ordered structures. Some of these segments are rather short. The results are thus not necessarily indicative of how well these methods would perform on other kinds of disorder not present in these targets or identified by other experimental methods such as NMR. Nevertheless, the assessment of these methods on a common set of proteins with a common set of accuracy parameters provides a snapshot of how well these methods perform and an indication of the kinds of methods that seem to work better on targets similar to the CASP6 proteins.

METHODS

Disordered Residues in CASP Targets

Twenty groups participated in the disorder prediction experiment in CASP6, while only six groups submitted disorder predictions to CASP5 (two of which submitted predictions on only one target). As in CASP5, each group in CASP6 was allowed to submit up to five predictions for each target, of which the first one is considered by the predictor to be the best prediction. Only four out of the 20 groups submitted more than one prediction for some targets. Only assessment of the first predictions is reported here.

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/pages/0887-3585/suppmat/

Grant sponsor: National Institutes of Health; Grant number: R01-HG02302. Grant sponsor: Pennsylvania Tobacco Settlement.

*Correspondence to: Roland L. Dunbrack, Jr., Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111. E-mail: roland.dunbrack@fccc.edu

Received 18 April 2005; Accepted 21 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20734

The article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

The prediction groups were required to submit a binary prediction for each residue and a numerical value called P in this article. The format of submitted predictions was the same as those in CASP5. Each record consists of a single-letter amino acid type, a letter "O" or "D," indicating whether the residue is predicted to be ordered or disordered, respectively, and a real number between 0 and 1 for the value of P, which is supposed to be positively related to, if not exactly the same as, the probability that the residue is disordered.

Disordered regions of proteins are those lacking a fixed tertiary structure under native conditions, generally in the absence of specific binding partners. For disorder evaluation of CASP6 targets, we first considered those residues that appear in the sequence provided by the CASP6 organizers but have no atomic coordinates in the corresponding structure file as disordered. All the other residues are labeled as ordered. However, we discovered as the CASP6 target structures were deposited in the Protein Data Bank (PDB) that the actual constructs used in several structures were different from what was given to CASP6 and to predictors. In some cases, this seems to be because some regions were predicted to be disordered, and the crystallographers therefore created a construct without amino acid segments predicted to be disordered. Before the December meeting, we used the CASP6 sequences. Subsequently, for four CASP6 targets, we used the constructs given in the PDB entries, rather than by CASP6. These are (supposedly) what is actually the object in the experiment, and is therefore the entity for which we have actual experimental data.

Missing density is not a perfect definition of disorder. Crystallization may impose order on some regions that would otherwise be disordered, and whole domains may sometimes be missing in electron density (i.e., disordered with respect to the rest of the structure in the crystal, even though they may be internally ordered). As discussed above, regions that are likely to be disordered are often removed from the genetic construct used to express the protein for crystallographic or NMR experiments,[7] so that we lack experimental information on longer disordered regions that may disrupt experimental structure determination.

Protein structure flexibility is also reflected by the atomic displacement factors, also known as the temperature or B-factors. The regions with high B-factors are also identified as disordered by some investigators.[8] However, there are many disadvantages in using B-factors as indicators of protein disorder. B-factors can vary greatly within a single structure as a result of the effects of local packing.[9] There is no significant correlation between the mean B-factor value of a CASP6 target and the percentage of residues without atomic coordinates. A protein can be very disordered based on the B-factor but rather ordered based on the number of residues missing atomic coordinates. For the disorder evaluation of CASP6 targets, we primarily used missing coordinates to identify disordered residues, rather than the B-factors. However, we calculated the correlation between the B-factor and values of P given by the predictors, which reflects whether residues with higher B-factors are more likely to be predicted as disordered by predictors.

The disorder predictions of 66 targets in CASP6 were evaluated. Other targets were canceled by the CASP6 organizers for a variety of reasons, including three structures due to unrefined coordinates or a large fraction of missing coordinates. Although failing to resolve the structures strongly suggests that these proteins are significantly disordered, it might also result from incomplete refinement or other experimental artifacts. We decided not to use these three targets (T0217, T0218, and T0245) in our assessment.

The fractions of disordered residues of these 66 targets range from 0% to about 40%. Target 206 is the most disordered target. It has 220 residues, 82 of which (37%) are disordered. It has a fully disordered N-terminal domain of 78 amino acids, composed mostly of a long sequence of repeating triplets with sequence Thr-Gly-Pro. Other targets have disorder rates of less than 25%. Altogether 743 out of 14,165 residues are disordered, yielding a mean disorder rate of 5.25%. The residues near both ends are more likely to be disordered than those in the middle of the sequence, as shown in Figure 1.

Evaluation Criteria

The receiver operating characteristic (ROC) curves have been used by some investigators to evaluate the discrimination power of disorder prediction methods.[3,8,10] We adopted this method for CASP6 disorder evaluation. The ROC curves were generated as follows. For each value of P in increments of 0.01, we assumed all the residues with values greater than or equal to P were predicted as disordered. Then we calculated the true positive rate (R_{TP}) and the false positive rate (R_{FP}) using the following equations:

$$R_{TP}(P) = \frac{N_{TP}(P)}{N_{disorder}}$$

$$R_{FP}(P) = \frac{N_{FP}(P)}{N_{order}}$$

where $N_{TP}(P)$ and $N_{FP}(P)$ are the numbers of true and false positives in the predictions with $P_{pred} \geq P$, respectively, $N_{disorder}$ is the total number of disordered residues as determined from the experimental structures, and N_{order} is the number of ordered ones. ROC curves were generated by a parametric plotting of $R_{TP}(P)$ against $R_{FP}(P)$. The area under the ROC curve, the ROC score, denoted S_{ROC}, was calculated using the trapezoid rule, and used as a measure of prediction accuracy.

The ROC score is a good indicator of the discrimination power if a predictor assigns values of P continuously (i.e., with a large enough number of distinct values) to make it possible to generate a smooth ROC curve. Unfortunately, as shown in Table I, not all prediction groups used sufficiently numerous distinct values of P. Five groups used only two values, one for predicted ordered residues and one for predicted disordered residues. Group 679

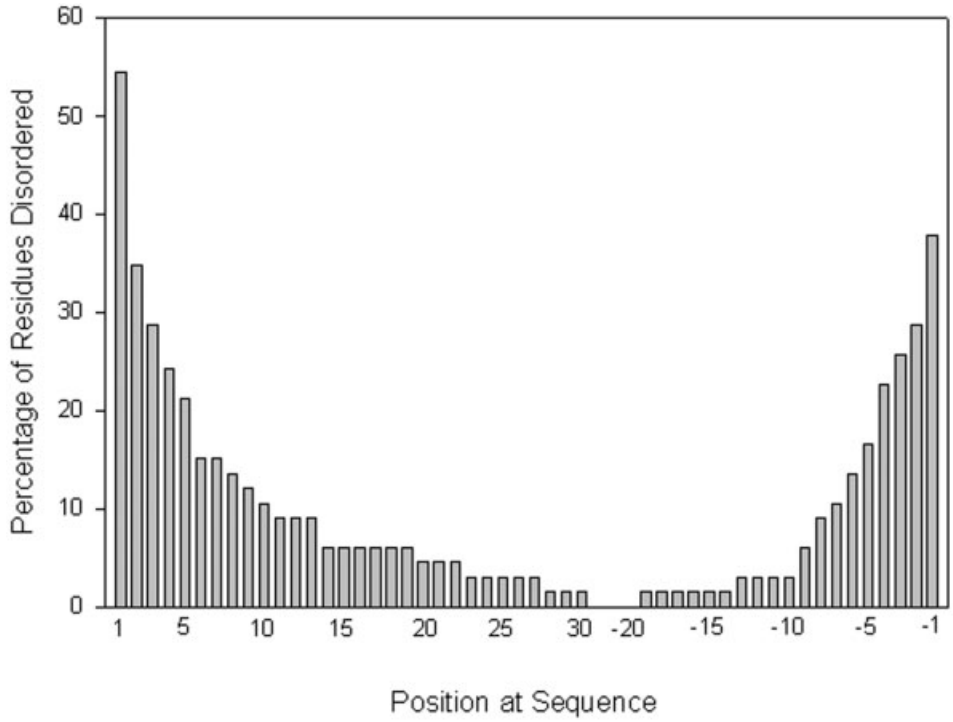

Fig. 1. Frequencies of disordered residues in terms of positions in the protein sequence. The positive numbers at the horizontal axis indicate positions counted from the N-terminus, while the negative numbers represent those from the C-terminus.

TABLE I. Prediction Groups, Numbers of Targets, and Residues Covered by Their Prediction, and Numerical Values Used for Assigning *P* Values

Group number	Targets predicted	Residues predicted	*P* values
003	66	14,165	Continuous[a]
060	66	14,165	Continuous
193	66	14,165	Continuous
245	60	12,914	Continuous
347	66	14,165	Continuous
472	61	12,994	Continuous
545	64	13,754	Continuous
633	64	13,725	Continuous
676	58	12,267	Continuous
686	57	12,141	Continuous
018	23	4791	0.3, 0.7
019	44	9383	0.3, 0.7
096	65	14,067	0.2, 0.8
461	65	13,986	0.0, 0.6, 0.8, 1.0 (occasionally other numbers)
536	66	14,165	0.0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0
667	59	12,411	0.0, 0.5, 1.0
673	59	12,411	0.0, 1.0
674	59	12,411	0.0, 1.0
675	59	12,411	0.0, 1.0
679	55	11,937	1.0

[a]Sufficiently numerous distinct values were used for assigning *P*.

assigned all residues a value of 1.0 and labeled residues as ordered or disordered completely independent of *P*.

Groups 667, 536, and 461 used fewer than 10 different values of *P*, and the resulting ROC curves are far from smooth. In total, 10 groups used fewer than 10 values of *P*, and these groups are not assessed well by the ROC score. In addition, it seems that not all of the predictors used a minimum value of $P = 0.5$ as a cutoff for prediction or disorder. Group 60 labeled disordered residues in a way that, for some targets, disordered residues correspond to those with $P > 0.5$, but for other targets, disordered residues are those with $P < 0.5$. Group 545 seems to have used $P = 0.42$ as the cutoff. Groups 245, 461, 536, and 633 used $P = 0.5$ as the cutoff, but with a few exceptions. Because of these problems, we also assessed the binary D/O assignments in addition to the ROC scores based on *P*.

For binary predictions, sensitivity (S_{sens}) and specificity (S_{spec}) are commonly used to evaluate predictive accuracy. S_{sens} and S_{spec} are defined in the following equations:

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

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

where N_{TP}, N_{FP}, N_{TN}, and N_{FN} are the numbers of true positives (i.e., predicted and experimentally disordered), false positives, true negatives, and false negatives, respectively, in the collective predictions for a particular group. In the disorder prediction, sensitivity represents the frac-

tion of disordered residues correctly identified in a prediction, while specificity indicates the fraction of ordered residues correctly identified. Obviously a prediction method is better than another if it has higher scores in both sensitivity and specificity. However, there is a tradeoff between sensitivity and specificity. A predictor can increase the sensitivity by predicting more "greedily" (i.e., predicting more residues as disordered). However, this usually will decrease the specificity. Similarly, predicting more "conservatively" will increase the specificity while sacrificing some sensitivity. Which prediction is better, a greedy prediction with higher sensitivity or a conservative one with higher specificity? The answer usually depends on the purpose of the prediction and on the costs of a false prediction.

For CASP6 targets, we used different measures to evaluate the disorder predictions in addition to the ROC score defined above. The simplest measure used for the evaluation of disorder predictions is the overall percentage of accuracy, or $Q2$ (similar to $Q3$ in secondary structure prediction)[6,8]:

$$Q2 = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{TN} + N_{FN}}.$$

However, $Q2$ is obviously not a good measure for disorder prediction because of the very unbalanced rates of ordered and disordered residues. For CASP6 targets, simply predicting all residues ordered would yield a $Q2$ of 95%. $Q2$ therefore strongly favors conservative predictions.

$Q2$ weighs the prediction of disordered and ordered residues equally. However, since disordered residues are rare and hard to predict, we should reward the predictors more generously for correctly predicting a disordered residue than for predicting an ordered one. We used a weighted score, S_w, in the evaluation of CASP6 disorder prediction. First we defined a score S:

$$S = W_{disorder}N_{TP} - W_{order}N_{FP} + W_{order}N_{TN} - W_{disorder}N_{FN},$$

where $W_{disorder}$ and W_{order} are weights assigned to experimentally defined disordered and ordered residues, respectively. S_w reaches a maximum when all the disordered residues are predicted disordered and all the ordered residues are predicted ordered:

$$S^{max} = W_{disorder}N_{disorder} + W_{order}N_{order}$$

The worst case is that none of the residues are correctly predicted. In this case we get a minimum score:

$$S^{min} = -S^{max}$$

To make the score comparable among different sets of data, we normalize the score as follows:

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

The value of S_w therefore varies between -1 and $+1$. $W_{disorder}$ and W_{order} can be set arbitrarily depending on the requirements of evaluation. For CASP6 evaluation, we set $W_{disorder}$ and W_{order} values based on the overall rates of disordered and ordered residues (i.e., we set $W_{disorder} = 94.75$ and $W_{order} = 5.25$). With these $W_{disorder}$ and W_{order} values, S_w is 0 if a predictor predicts that residues are all ordered or all disordered. The expected value of S_w is also 0 if the predictor predicts ordered or disordered residues randomly. So a positive S_w value indicates that a prediction is better than a random one. Basically, S_w rewards the correct disorder predictions more heavily than the correct order predictions, and penalizes the incorrect order predictions more heavily than the incorrect disorder predictions. In comparison with $Q2$, S_w favors more greedy predictions.

Another commonly used measure of prediction accuracy is the Matthews correlation coefficient,[11] which is given by

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

The value of S_{MCC} also ranges between -1 and $+1$, with random prediction scoring 0. For a prediction with unequal class frequencies like disorder/order predictions, S_{MCC} also favors correct predictions of disordered residues, but not as strongly as S_w.

We can also combine both the sensitivity and specificity to make a single measurement criterion. In CASP6 disorder prediction, we used the product of sensitivity and specificity

$$S_{product} = S_{sens}S_{spec} = \frac{N_{TP}N_{TN}}{N_{disorder}N_{order}}$$

as a measure of the prediction accuracy. Since $N_{disorder} \ll N_{order}$ and presumably $N_{TP} \ll N_{TN}$, predicting an additional disordered residue correctly (i.e., increasing N_{TP} by 1) can raise $S_{product}$ much more quickly than predicting an additional ordered residue correctly (increasing N_{TN} by 1). Like S_w, $S_{product}$ also favors greedy predictions strongly.

In addition to evaluating the accuracy of prediction of disordered residues, we also evaluated the distinguishing powers of each prediction in terms of the B-factors. The B-factor values of the CASP6 targets vary from 0 to 140, and their distribution is shown in Figure 2. To correlate P with the B-factors, we calculated the mean P and mean B-factor for each bin in the histogram shown in Figure 2. Then we plotted the mean P against the mean B-factor for all the groups and calculated the correlation coefficient for each curve.

RESULTS

Figure 3 shows the ROC curves of predictions submitted by some prediction groups, and the areas under these curves (S_{ROC}) are shown in Table II. Most of these groups used a sufficient number of distinct values of P to make smooth ROC curves. Group 679 assigned all the P to 1.0. So their ROC curve is a straight line from (0,0) to (1,1) and the ROC score is 0.5, which is the score expected with

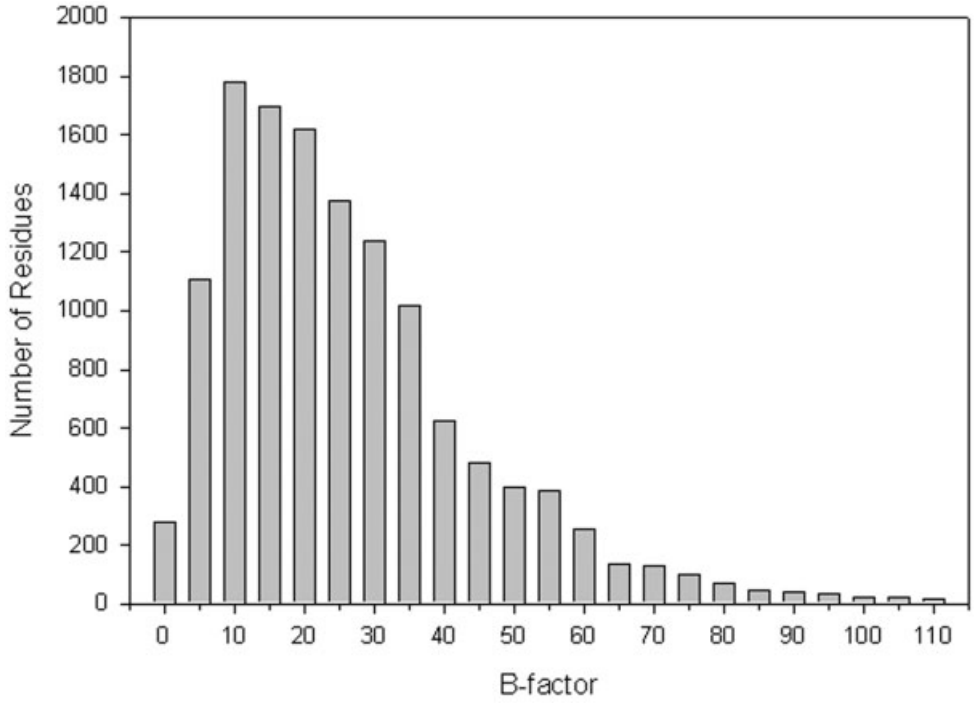

Fig. 2. Distribution of B-factors of CASP6 targets. Bars in the graph are the number of residues whose B-factors are within each B-factor range. The last bar shows the number of residues whose B-factors are 100 and above.

random predictions. The ROC curves of almost all the other groups are above this line, indicating these predictions are all better than random. It is interesting that the first half of Group 60′s curve is below the diagonal. As mentioned above, Group 60 set P lower than 0.5 for all the residues they identified as disordered for some targets. Apparently the strange shape of the ROC curve is due to this error in their prediction procedure.

Figure 3 clearly shows that the prediction results of Group 193 are the best, as assessed by the ROC score, with true positive rate better than all other groups at all false positive levels. Group 96 has one point close to the ROC curve of Group 193. This group used only two values of P, so their ROC curve is composed of two straight line segments, connecting (0,0) and (1,1). For this group and other groups that used limited distinct values of P, the ROC score is not a good measure of prediction success.

Table II also summarizes the results of the assessment of binary predictions (*ordered* or *disordered*) of all the groups evaluated using different criteria. Group 193 has the best weighted score (S_w = 0.582), the best product of sensitivity and specificity ($S_{product}$ = 0.624), as well as the best ROC score (S_{ROC} = 0.882). Group 18 tops at $Q2$ (0.960), although this group submitted predictions for only about one third of the targets. Group 60 is best at Matthews correlation coefficient (S_{MCC} = 0.415). Group 96 did nearly as well (S_{MCC} = 0.413) on 65 of the 66 targets. The data in the table also show that $Q2$ is not a good evaluation criterion, since even a group with sensitivity as low as 4% has a $Q2$ of 90%. Group 193 correctly identified 75% of disordered residues, at the cost of rather high overprediction: About 17% of ordered residues were incorrectly predicted as disordered. Groups 676, 96, 347, 60, and 3 all maintain specificity higher than 90%, while correctly identifying about 50% of disordered residues. Some groups that did not do well with S_{ROC} because of using only a few distinct values of P did better with the binary scores.

As noted earlier, those prediction groups that used continuous scales for P generally used a cutoff of P = 0.5 to determine their binary predictions. We examined whether other choices for this cutoff might produce higher scores. In Table III, we show the choices for cutoff that produce the maximum values for each assessment score, and we also list the "new" score using this cutoff, as well as the "old" score based on the binary predictions of the predictors. For all the groups, the cutoffs needed to maximize $Q2$ are much higher than 0.50. This occurs because $Q2$ weights TP and TN predictions equally, and for most methods increasing the TN score is easier than increasing the TP score. Similarly, the conservative S_{MCC} score often requires a higher value of the cutoff to maximize the score. The greedier $S_{product}$ and S_w scores require lower cutoffs, because they emphasize higher TP scores than $Q2$ or S_{MCC}. Notably, group 193 has optimal cutoffs not far from 0.5 (0.45 in both cases) for the greedier assessment scores.

To examine the correlation between P and actual disorder rates further, we binned the predictions for each group in ranges of 0.1 (i.e., 0.0–0.1, 0.1–0.2, 0.2–0.3, etc.) and calculated the proportion of experimentally disordered residues in each bin. The results for the top four groups from Table III are shown in Figure 4. For instance, for

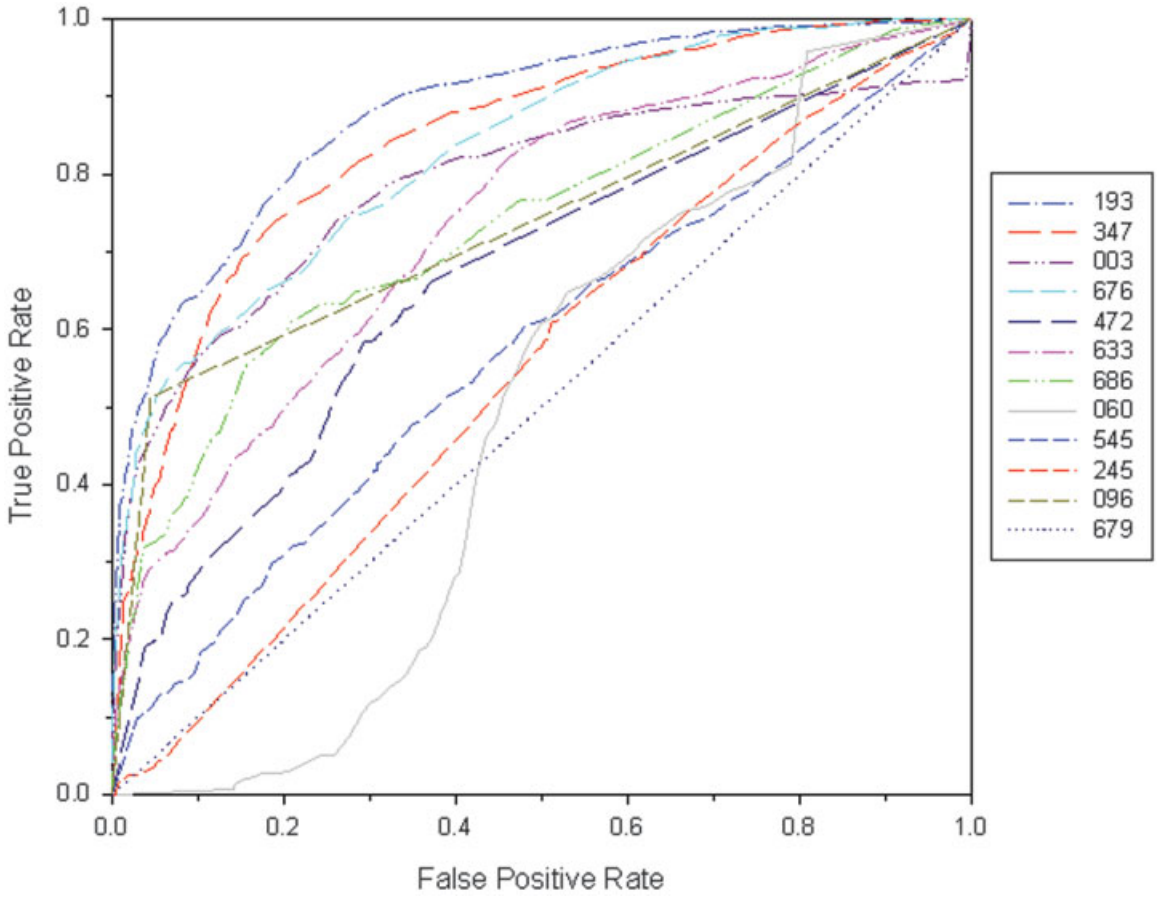

Fig. 3. ROC curves of the predictions submitted by some predictor groups. Note that Group 96 only used two numbers (0.2, 0.8) to assign *P* values, and Group 679 only used one number (1.0). All the other groups used continuous real numbers between 0 and 1 (to two significant digits) to assign values to *P*.

group 347, the residues with $0.4 < P \leq 0.5$ have an average P of 0.44 (x axis) and 13% (y axis) of these residues are actually disordered. The figure indicates that for these groups (and the others, not shown) the actual rates of disorder are significantly lower that the values of P provided by the predictors at most values of P. We conclude that the probabilities provided by the methods used by the predictors are not accurate assessments of the probability of disorder.

The correlation of P with B-factors is given in Table II for all the groups. For about half (9 out of 20) groups, P is positively correlated with the B-factors at the 5% significance level (p value $< .05$), which means their algorithms can somehow distinguish high B-factor residues from the lower B-factor ones. Groups 193, 3, and 472 are the best groups in B-factor prediction, with correlation coefficients of .86 or higher. Interestingly, the P's submitted by Group 60 are very significantly, but negatively correlated with the B-factors, due to the reversal in assigning P in some of their predictions.

Finally, we investigated whether it is easier to predict disorder in proteins that are homologous to proteins of known structure. We separated the CASP6 targets into two categories: homology and nonhomology, as defined by the assessors for structure prediction. We combined comparative modeling and folding recognition (homology) targets into a "homology" group, and fold recognition (analogy) and new fold targets into a "nonhomology" group. For those targets with domains in different categories, we labeled them either homology or nonhomology according to the longer domain. The disorder prediction scores of the targets in these two categories are shown in Table IV. The prediction scores averaged over all groups in terms of S_{MCC}, $S_{product}$, and S_w are all better for the homology group than for the nonhomology group. A one-tailed, paired Student's t-test shows that the differences in the S_{MCC} ($p < .001$) and $S_{product}$ ($p = .038$) values are statistically significant, but the difference in S_w values are not significant ($p = .082$).

We also compared the mean prediction scores (averaged on all prediction groups) of the individual targets (see Table I in the Supplementary Material). Obviously the prediction difficulty is positively correlated with the rate of the disordered residues in the protein (p value $\ll .01$). S_{MCC} scores are not shown, since this score is not defined when either the targets or the predictions do not contain both ordered and disordered residues. All the targets with mean S_w score above 0.76 have less than 1% disordered

TABLE II. Disorder Predictions Evaluated With Different Criteria

Group	N_f[a]	$N_?$[b]	S_{sens}	S_{spec}	$Q2$	S_{MCC}	$S_{product}$	S_w	S_{ROC}	r (p value)[c]
193	66	14,165	**0.754**	0.828	0.824	0.323	**0.624**	**0.582**	**0.882**	**0.92** (<.001)
676	58	12,267	0.505	0.952	0.932	0.368	0.480	0.501	0.831	0.61 (.002)
096[d]	65	14,067	0.513	0.955	0.932	0.413	0.490	0.467	0.734	0.61 (.002)
347	66	14,165	0.522	0.915	0.895	0.315	0.478	0.437	0.845	0.78 (<.001)
060	66	14,165	0.471	0.965	0.939	**0.415**	0.454	0.436	0.488	−0.64 (.001)
003	66	14,165	0.471	0.949	0.924	0.360	0.447	0.420	0.787	0.89 (<.001)
018[d]	23	4791	0.301	0.990	**0.960**	0.403	0.299	0.358	0.646	0.30 (.163)
686	57	12,141	0.318	0.964	0.936	0.269	0.306	0.345	0.735	0.25 (.251)
019[d]	44	9383	0.255	0.987	0.955	0.321	0.252	0.321	0.621	0.34 (.108)
667[d]	59	12,411	0.356	0.903	0.879	0.170	0.322	0.313	0.656	0.06 (.791)
633	64	13,725	0.595	0.713	0.707	0.150	0.424	0.308	0.737	0.75 (<.001)
461[d]	65	13,986	0.393	0.885	0.859	0.184	0.348	0.280	0.652	0.63 (.002)
674[d]	59	12,411	0.208	0.980	0.946	0.229	0.204	0.264	0.594	−0.11 (.602)
679[d]	55	11,937	0.197	**0.995**	0.959	0.338	0.196	0.262	0.500	0.00 (.000)
536[d]	66	14,165	0.265	0.983	0.945	0.322	0.261	0.248	0.642	−0.15 (.494)
675[d]	59	12,411	0.501	0.715	0.705	0.096	0.358	0.237	0.608	0.25 (.247)
673[d]	59	12,411	0.432	0.743	0.729	0.081	0.321	0.206	0.588	−0.08 (.732)
472	61	12,994	0.302	0.891	0.858	0.136	0.269	0.177	0.668	0.86 (<.001)
545	64	13,754	0.424	0.691	0.678	0.054	0.293	0.120	0.575	−0.12 (.591)
245	60	12,914	0.044	0.942	0.904	−0.012	0.042	0.092	0.545	0.50 (.016)

The highest scores for each parameter are shown in bold type.
[a]Number of targets predicted.
[b]Number of residues predicted.
[c]Correlation coefficient between predicted P value and B-factor (p values in parentheses).
[d]These groups used fewer than 10 values of P.

TABLE III. Maximal Scores and Corresponding *P* Cutoffs

Group	pQ2	nQ2	oQ2	pSm	nSm	oSm	pSp	nSp	oSp	pSw	nSw	oSw
193	0.88	0.958	0.824	0.85	**0.484**	0.323	0.45	**0.637**	**0.624**	0.45	**0.596**	**0.582**
676	0.81	**0.962**	0.932	0.56	0.401	**0.368**	0.27	0.539	0.480	0.45	0.511	0.501
347	0.99	0.948	0.895	0.99	0.336	0.315	0.33	0.599	0.478	0.33	0.550	0.437
003	0.58	0.953	0.924	0.51	0.409	0.360	0.16	0.548	0.447	0.16	0.480	0.420
686	0.94	0.957	**0.936**	0.50	0.269	0.269	0.20	0.482	0.306	0.27	0.425	0.345
633	0.99	0.948	0.707	0.94	0.249	0.150	0.37	0.450	0.424	0.30	0.361	0.308
472	1.00	0.945	0.858	0.99	0.165	0.136	0.01	0.416	0.269	0.01	0.292	0.177
245	1.00	0.958	0.904	0.38	0.039	−0.012	0.38	0.297	0.042	0.71	0.124	0.092
545	1.00	0.949	0.678	0.99	0.087	0.054	0.17	0.314	0.293	0.27	0.129	0.120

Highest scores for each parameter are shown in bold type.
pQ2, cutoff P for maximizing $Q2$; nQ2, (new) maximal $Q2$ value; oQ2, (old) $Q2$ calculated on the binary predictions; pSm, cutoff P for maximizing S_{MCC}; nSm, maximal S_{MCC} value; oSm, S_{MCC} calculated on the binary prediction; pSp, cutoff P for maximizing $S_{product}$; nSp, maximal $S_{product}$ value; oSp, $S_{product}$ calculated on the binary prediction; pSw, cutoff P for maximizing S_w; nSw, maximal S_w value; oSw, S_w calculated on the binary prediction.

residues. In contrast, all the targets with S_w score less than 0.2 have more than 10% disordered residues. The fraction of correctly identified disordered residues seems not to be related to the disorder rate in each protein (p value $> .5$). However, there is a very significant negative correlation between the specificity and the disorder rate (p value $< .01$), indicating that it is easier to identify ordered residues for less disordered proteins. It should be noted that the same $W_{disorder}$ (94.75) and W_{order} (5.25) values were used when calculating the S_w scores for individual targets, even though fractions of disordered residues vary greatly from one protein to another.

In order to compare the current disorder prediction techniques with those of 2 years ago, we evaluated CASP5 disorder predictions with the same criteria used in CASP6 disorder prediction assessment. Table V shows the performances of CASP5 disorder predictions evaluated using the various scores we have used for CASP6. Since many more groups participated in the CASP6 disorder prediction, it is hard to compare the overall performances between CASP5 and CASP6 disorder predictions. However, it is clear that the best groups in CASP6 performed better than the best ones in CASP5, as measured by any of the scores we have used.

SUMMARY

In summary, Group 193 (K. Peng, S. Vucetic, and Z. Obradovic, Temple University, Philadelphia, PA) achieves the best performance in terms of almost all the criteria, including ROC score (S_{ROC}), product of sensitivity and

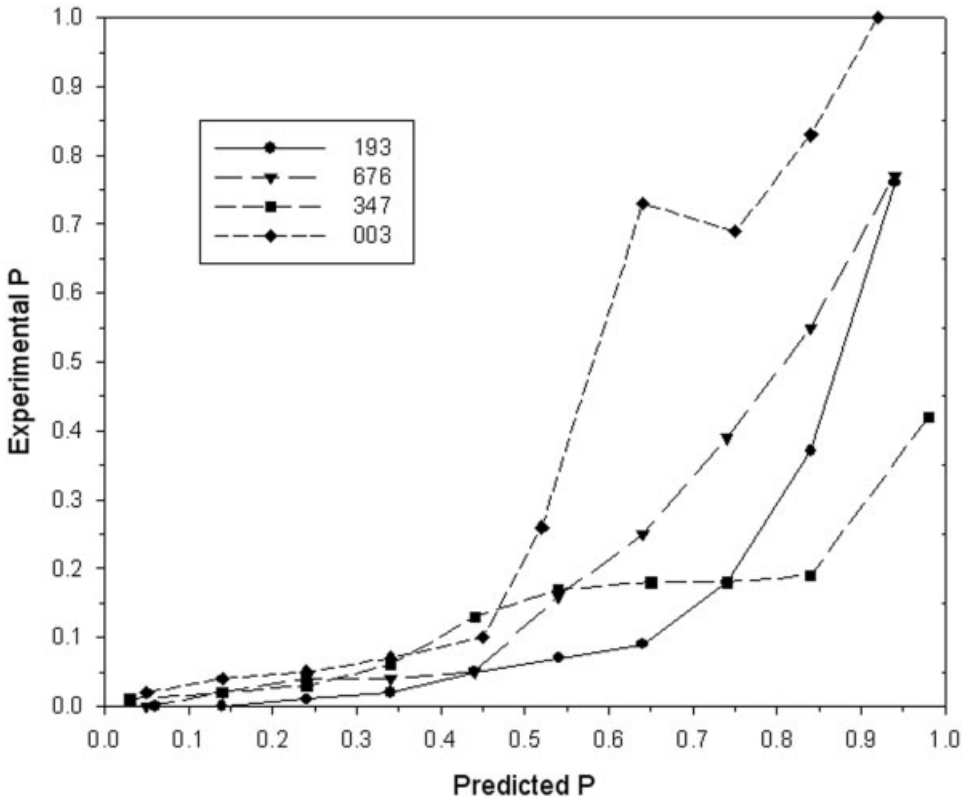

Fig. 4. Probability assigned by predictors versus experimentally determined rates of disorder. For each group, residues with probabilities (*P*) in each range (0–0.1, 0.1–0.2, etc.) were binned and the average *P* was calculated and plotted on the *x* axis. For the same group of residues, the actual rate of disorder (the fraction of residues in the bin that are disordered) is plotted on the *y* axis.

TABLE IV. Comparison of Disorder Predictions for Targets in Homology and nonhomology Categories

	Homology category					Nonhomology category				
Group	N_t	$N_?$	S_{MCC}	$S_{product}$	S_w	N_t	$N_?$	S_{MCC}	$S_{product}$	S_w
193	51	11,244	0.350	**0.640**	**0.602**	15	2921	0.232	**0.561**	0.498
676	44	9585	0.391	0.479	0.489	14	2682	0.285	0.496	**0.566**
096	50	11,146	**0.443**	0.520	0.498	15	2921	0.265	0.341	0.312
060	51	11,244	0.436	0.462	0.444	15	2921	0.331	0.420	0.395
347	51	11,244	0.333	0.509	0.468	15	2921	0.213	0.313	0.273
003	51	11,244	0.386	0.450	0.426	15	2921	0.276	0.436	0.396
018	16	3602	0.381	0.296	0.448	7	1189	**0.444**	0.302	0.121
686	44	9585	0.279	0.288	0.315	13	2556	0.264	0.432	0.517
019	36	7982	0.326	0.257	0.319	8	1401	0.262	0.193	0.386
667	45	9729	0.182	0.323	0.301	14	2682	0.125	0.319	0.379
679	43	9475	0.376	0.224	0.271	12	2462	−0.011	0.000	0.209
674	45	9729	0.246	0.213	0.251	14	2682	0.132	0.140	0.326
536	51	11,244	0.334	0.266	0.254	15	2921	0.266	0.231	0.217
461	50	11,065	0.237	0.391	0.336	15	2921	0.010	0.161	0.019
633	49	10,804	0.151	0.422	0.303	15	2921	0.143	0.434	0.324
472	49	10,740	0.163	0.297	0.213	12	2254	−0.007	0.098	−0.025
675	45	9729	0.124	0.391	0.276	14	2682	−0.028	0.158	0.032
673	45	9729	0.066	0.286	0.160	14	2682	0.169	0.554	0.471
245	45	9993	−0.023	0.029	0.106	15	2921	0.024	0.084	0.028
545	49	10,833	0.059	0.298	0.127	15	2921	0.037	0.271	0.085
Mean	46	9997	0.262	0.352	0.330	14	2624	0.172	0.297	0.276

Highest scores for each parameter are shown in bold type.

TABLE V. CASP5 Disorder Predictions Evaluated With Different Criteria

G[a]	N_t[b]	$N_?$[c]	S_{MCC}	$S_{product}$	S_{ROC}	S_w	r[d]
068	56	12,596	0.347	**0.566**	**0.534**	**0.846**	**0.830**
355	50	11,066	0.208	0.520	0.445	0.756	0.738
454	56	12,596	0.204	0.379	0.321	0.729	0.433
020	36	7487	**0.499**	0.549	0.531	0.392	−0.033

Highest scores for each parameter are shown in bold type.
[a]Group number.
[b]Number of targets predicted.
[c]Number of residues predicted.
[d]Correlation coefficient between P and B-factor.

specificity ($S_{product}$), weighted score (S_w), and the correlation between P and B-factor. Group 60 (S. Nakamura, K. Shimizu, and T. Ishida, University of Tokyo, Tokyo, Japan) is best at S_{MCC}. Group 18 (J. Cheng, M. Sweredoski, and P. Baldi, University of California Irvine, Irvine, CA) scores best in $Q2$, with more conservative predictions and on only one third of the targets. Other high-scoring groups include Group 676 (Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, Hungarian Academy of Science, Budapest, Hungary), 096 (S. Tosatto, O. Bortolami, A. Cestaro, G. Cozza, M. Lexa, S. Toppo, G. Valle, and S. Moro, University of Padova, Padova, Italy), 347 (R. MacCallum, Stockholm University, Sweden), and 003 (D. Jones, University College London, London, UK). The best prediction groups can successfully identify nearly half of the disordered residues with false positives rates [i.e., $100(1 - S_{spec})$] less than 20%.

REFERENCES

1. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
2. Fink AL. Natively unfolded proteins. Curr Opin Struct Biol 2005;15:35–41.
3. 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.
4. Iakoucheva LM, Dunker AK. Order, disorder, and flexibility: prediction from protein sequence. Structure (Camb) 2003;11:1316–1317.
5. Gunasekaran K, Tsai CJ, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J Mol Biol 2004; 341:1327–1341.
6. Melamud E, Moult J. Evaluation of disorder predictions in CASP5. Proteins 2003;53(Suppl 6):561–565.
7. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure (Camb) 2003;11:1453–1459.
8. Jones DT, Ward JJ. Prediction of disordered regions in proteins from position specific score matrices. Proteins 2003;53(Suppl 6):573–578.
9. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12: 1060–1072.
10. Li X, Obradovic Z, Brown CJ, Garner EC, Dunker AK. Comparing predictors of disordered protein. Genome Inform Ser Workshop Genome Inform 2000;11:172–184.
11. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.