PROTEINS: Structure, Function, and Genetics 53:561–565 (2003)

Evaluation of Disorder Predictions in CASP5

Eugene Melamud[1,2] **and John Moult**[1*]
[1]*Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, Rockville, Maryland*
[2]*Molecular and Cell Biology Program, University of Maryland, College Park, Maryland*

ABSTRACT **This paper reports an analysis of the accuracy of predictions of structural disorder received as part of the CASP5 experiment. Six groups made predictions of disorder. The predictions of the four most active groups have been compared with the experimental results, in terms of the sensitivity and specificity of the methods. All four methods succeed in detecting over half the disordered residues in the targets, with a generally low rate of over-prediction. Two of the methods perform significantly better when the structure of a related protein is available. There is a trade-off between the fraction of disordered residues detected and the extent of over-prediction, and groups have adopted different compromises in this respect. Comparison of performance at the same over-prediction rates highlights the role of related structures in some methods rather than others, with different groups achieving the highest sensitivtiy for different target sets. Over-all, the methods are clearly of considerable use in identifying potential disorder. Proteins 2003;53:561–565.** © 2003 Wiley-Liss, Inc.

Key words: protein structure disorder; protein structure prediction; communitywide experiment; CASP

INTRODUCTION

The CASP experiments measure the ability of current computational methods to model protein structure.[1] Implicit in that goal is the assumption that there is a single structure to determine. While this is clearly true of the majority of naturally occurring proteins, there are a number of well documented exceptions.[2] There are several possible explanations for a lack of structure. For example, the protein may interact with multiple ligands, and adopt correspondingly different structures to complement each,[3] or the functional structure of the protein may require the presence of a partner before a single conformation is of sufficiently low free energy to dominate the population.[4] For some disordered regions, function depends on the ensemble of unfolded structures, rather than on the formation of a unique structure induced by some factor.[5] Disorder may affect the whole of a structure, or only part of it. Now that large scale efforts to determine the structure of all protein domains are underway,[6] it is important to determine how many proteins may exhibit partial or complete disorder. Therefore, CASP5 included a category to measure the effectiveness of current methods for predicting protein disorder.

PREDICTION SET

Predictors were invited to submit predictions on as many CASP5 targets as they wished. As with other classes of prediction, up to five predictions per target were allowed, and predictions were ranked by the submitors as models 1 through 5, with model 1 understood to be considered the most accurate. All analysis has been performed on model 1 submissions. The format for a prediction contained one record for each residue in the target. Each such record consists of a residue identifier, a one bit prediction of order or disorder for that residue (yes/no) and a number between zero and one, indicating the probability that the residue is disordered. All prediction data are available at the CASP5 web site: http://predictioncenter.llnl.gov/casp5.

IDENTIFYING DISORDER IN CASP TARGETS

'Disorder' is a rather soft concept: Under what conditions must disorder be present to qualify? Must there be no preferred conformation, or just no dominant conformation? For evaluation purposes, a residue was considered to be disordered if it was included in the sequence provided by the target submitor, but had no atomic co-ordinates in the corresponding structure file. All other residue was considered ordered. We also experimented with considering every residue with an average temperature factor above some cut-off as disordered. As shown in the results, this criterion did not correlate with absent co-ordinates, and was abandoned. There are some caveats with the definition adopted. It could be that the sequence provided for a target was not that actually present in the material from which the structure was obtained. Disorder might also occur at a chain terminus because of the presence of a tag of extra residues (used for purification perhaps) or removal of part of the natural sequence as part of the processing. Or a domain may appear disordered in a crystal because of a single hinge motion between it and the rest of the protein. Or only a portion of a single polypeptide chain may have been expressed, and the absence of the rest of protein may lead to an unstable fold. Most of these effects will tend to over-estimate the extent of *in vivo* disorder in the targets.

Figure 1 shows the fraction of disordered residues and the fraction of residues with temperature factors greater

*Correspondence to: John Moult, Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute, 9600 Gudelsky Dr., Rockville, MD 20850. E-mail: jmoult@tunc.org

Received 20 May 2003; Accepted 23 May 2003

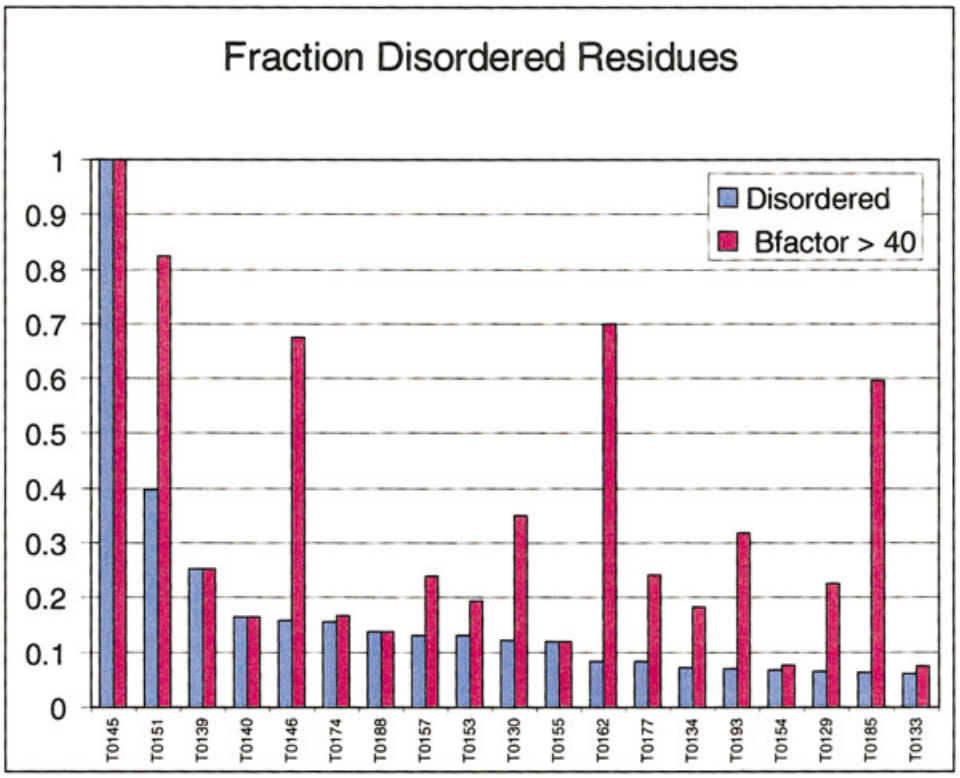
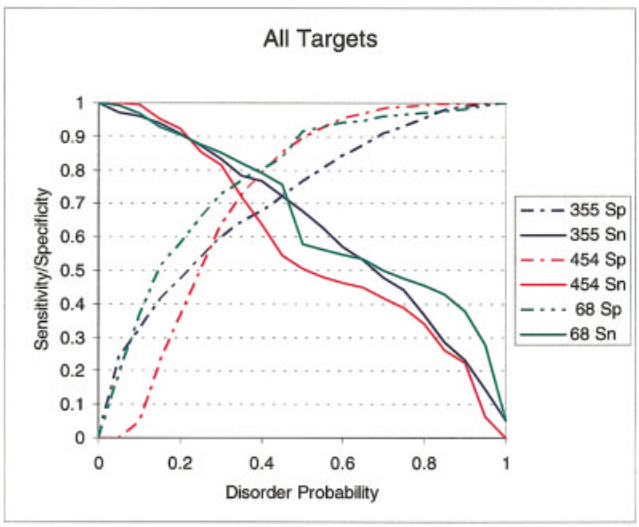
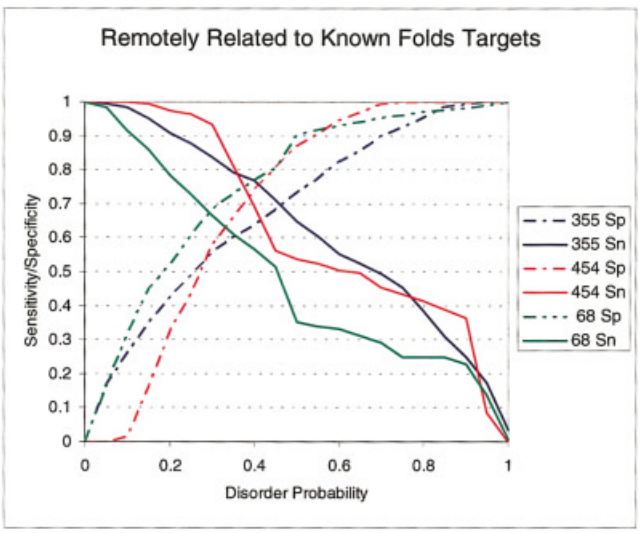


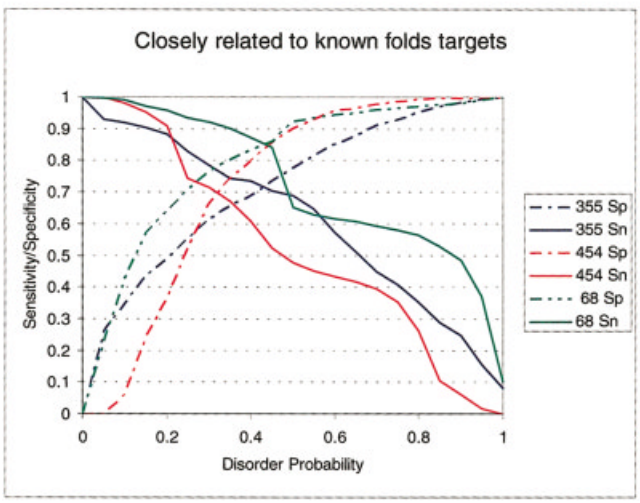
Fig. 1. Extent of experimentally observed disorder for the most disordered of the CASP5 targets. Red bars show the fraction of residues with Cα atom temperature factors greater than 40 or with no reported atomic co-ordinates. Blue bars show the fraction of residues with no reported atomic co-ordinates. The latter set was considered disordered for the present analysis. T0145 is seen as fully disordered experimentally. In general, a number of targets display a significant amount of disorder, providing a basis for the evaluation of the methods.



(A)



(C)



(B)

Fig. 2. Sensitivity and specificity for three of the disorder prediction methods as a function of the reported probability of residue level disorder. Sensitivity (Sn) curves are continuous lines, specificity (Sp) curves are dash-dot. All methods show the expected trade-off of improved sensitivity (ability to detect disordered residues) versus decreased specificity (equivalent to over prediction of disorder). (A) Predictions on all CASP targets. (B) Predictions on targets most related to a known structure. (C) Predictions on targets least related to a known structure. The relative quality of the methods varies depending on the target set considered. At reasonably low false positive rates, group 68 has the superior performance on all targets and on those most related to a known structure, while group 355 has the superior performance on the targets least related to a known structure.

TABLE I. Prediction Groups Submitting Disorder Data in CASP5 and Number of Targets Included by Each

Group	Targets
20	35
41	1
68	55
131	2
355	56
454	56

than 40 in the 18 most disordered CASP5 targets. Target 145 is fully disordered, based on extensive characterization in solution (Sussman, J, personal communication). Target 151, with 40% disorder, contains an apparently fully disordered domain. Other targets contain 30 – 10% of disordered residues. Thus, there is a considerable amount of disorder to predict in the CASP5 target set. In general, there is no correlation between the extent of disorder, and the fraction of residues with high temperature factors.

Six groups submitted predictions on one or more targets. Table 1 shows the group IDs, and the number targets each submitted predictions for. The evaluation has been carried on data from the four groups who made a substantial number of predictions.

EVALUATION CRITERIA

Most residues in the CASP targets are ordered, so that a simple Q2 measure (analogous to the Q3 measure popular in secondary structure prediction evaluation) – fraction of residues correctly predicted as either ordered or disordered, is not useful: Simply predicting everything as ordered would yield a Q2 of about 90%. More appropriate are standard definitions of specificity and sensitivity:

$$\text{Sensitivity } S_n = TP/(TP+FN) = TP/N_d$$

$$\text{Specificity } S_p = TN/(TN+FP) = TN/N_o$$

where TP is the number of true positives in a prediction – residues that are predicted disordered and are experimentally disordered, FN is the number of false negatives (predicted as ordered when disordered experimentally), TN is the number of true negatives (predicted ordered, and observed ordered) and FP the number of false positives (predicted disordered, but observed ordered). N_d is the total number of residues observed as disordered, and N_o is the total number of residues observed as ordered. Then, sensitivity (S_n) is the fraction of experimentally disordered residues identified as such, and the specificity, S_p, is the fraction of ordered residues identified as such.

The fraction of truly disordered residues that are predicted (the Sensitivity) can always be increased by over-predicting, that is, predicting more residues as disordered than actually exist in the data set. But over-predicting inevitably reduces specificity. The impact of this can be seen most directly by considering the fraction of false positive predictions, related to the specificity by $f_p = (1 - S_p)$. In general, a successful method will achieve a high sensitivity, with a low fraction of false positives.

TABLE II. Disorder Prediction Results, All Targets

CASP group no.	Sensitivity	Fraction of false positives	Total no. of predictions
20	0.56	0.03	7387
68	0.57	0.08	12496
355	0.67	0.23	12712
454	0.50	0.10	12712

TABLE III. Disorder Prediction Results, Comparative Modeling Targets

CASP group no.	Sensitivity	Fraction of false positives	Total no. of predictions
20	0.62	0.02	5507
68	0.65	0.07	6714
355	0.69	0.22	6714
454	0.47	0.09	6714

We have applied these measures to all residues in all predicted targets. Because disorder evaluation is a new area in CASP, we considered it appropriate to consult with the groups who had made predictions. The procedure adopted is consistent with the responses received. We calculated specificity and sensitivity on the binary predictions (yes/no predictions of whether each residue is disordered), and as a function of the declared probability with which a residue was predicted to be disordered.

In other areas of structure prediction, knowledge of a related structure greatly influences the quality of any model. It is therefore interesting to determine whether that is the case for disorder predictions as well. To investigate this, we also examined sensitivity and specificity for two subsets of targets. One set contains those targets most similar to a known structure (comparative modeling ('CM')). The other is the set of targets least similar to any known structure ('new fold' ('NF'), folds remotely or partially related to a known structure ('FR/NF'), and apparently analogous fold relationships ('FR(A)'). Target categories are those used for the other CASP5 assessments.

RESULTS

Table 2 summarizes the results for yes/no predictions of disorder on all targets. The highest fraction of correctly predicted residues is for group 355, at about 67%. This value was obtained at the expense of substantial over-prediction – 23% of ordered residues were assigned as disordered. Groups 20 and 68 both have approximately the same fraction of disordered residues identified, 56 and 57% respectively. However, group 20 achieved this performance with only 3% of false positives, while group 68 has 8% false positives.

Table 3 shows the results for the targets with close structural homologs. Sensitivity increases significantly for group 20 and 68, by 6 and 8% respectively, and there is also a slight drop the fraction of false positives, indicating that these methods are able to take advantage of knowledge of related structures in some way. Results for the other two groups are similar to those with all targets included.

TABLE IV. Disorder Prediction Results, Targets Least Related to Known Structures

CASP group no.	Sensitivity	Fraction of false positives	Total no. of predictions
20	0.63	0.08	483
68	0.34	0.09	3037
355	0.64	0.26	3253
454	0.54	0.12	3253

Table 4 shows the results for the set of targets least related to known structures. The sensitivity for group 68 is substantially lower here, falling to 34%, confirming the effective use of knowledge of a related structure. Results for 355 and 454 are again similar to those obtained for all targets. Group 20 submitted few targets in this subset, so the results are not well determined.

As noted earlier, there is an inevitable trade off between increasing sensitivity, and decreasing specificity. To investigate this factor, the sensitivity and specificity were calculated for different thresholds of the probability of disorder of each residue. I.e. considering all residues with a declared probability of zero or higher probability as disordered, all with a probability of 0.05 and higher, 0.1 and higher, and so on. Figure 2 shows these data, for the predictions on the three different target sets (parts A, B, and C) by groups 68, 355 and 454. Group 20 submited probability values in a form that could not be accommodated in this analysis, and so could not be included.

When all predictions are considered to be of 'disorder' (probabilities of 0 and higher), the sensitivity is 1 – all disordered residues are identified, At the other extreme, very few instances of disorder are assigned a probability of near 1 by these methods, so the sensitivity for such predictions is very low. As expected, specificity shows the reverse trend – predictions with a high probability are usually correct, so that there are few false positives and a specificity approaching 1. If low probabilities are considered, there are many false positives, and the specificity approaches zero.

In the evaluation based on yes/no predictions of disorder, group 355 achieved a relatively high specificity at the price of a relatively high rate of false positives. Using the data in figure 2, it is possible to compare performance of the groups at a particular false positive rate. Inspection of figure 2(A) shows groups 68 and 355 have approximately the same dependence of sensitivity on probability threshold – the green and blue continuous lines follow approximately the same course as a function of probability threshold. That is, at a given a threshold, both are able to predict approximately the same fraction of truly disordered residues. However, the specificity curve (dashed blue) for group 355 runs significantly below that of group 68 (dashed green), so that at a given threshold, group 355 will have more false positives than group 68. For example, at a false positive rate of 10% (close to that for group 68 on yes/no predictions), group 68 has a sensitivity of about 58%, and group 355 about 47%. At a false positive rate of 20% (close to that for yes/no on group 355), group 68 has a sensitivity of about 80%, and group 355, about 62%. By these criteria, then, group 68 delivers a significantly better performance. However, the picture depends sharply on which set of targets are considered. Figure 2(C) shows the same data for the targets least related to known structures. At a 10% false positive rate, group 68 has a sensitivity of about 34%, and group 355, a rate of about 47%. At a 20% false positive rate, the equivalent numbers are 51% and 57%. For such targets, the method of group 355 is superior, particularly at low false positive rates.

DISCUSSION

This is the first time disorder has been evaluated in CASP, so there are no prior performances to compare with. Overall, the methods clearly have value. In the best case, over half of the disordered residues in all the CASP5 targets were identified, with little over-prediction. A second method correctly identifies a higher fraction of the disordered residues, but at the cost of substantial over prediction – about 22%. Which method might be preferred depends some-what on the application. However, over-prediction can be misleading. In choosing which portion of a polypeptide chain to include in a construct for structure determination, omitting residues that are part of the ordered structure may result in an unstable fold. When estimating the extent of disorder in complete genomes, over-prediction results in substantial over-estimates. For example, a method with a false positive rate of 20% applied to a proteome that is in fact fully ordered would return a 20% disorder prediction.

The methods vary in utilization of knowledge of homologous structures. Most pronounced is that of group 68, with an 8% increase in sensitivity for targets with a close structural relative, and a 23% decrease in sensitivity when targets with at best remote structural relatives are considered. On the other hand, group 355's method is rather insensitive to knowledge of a structural relative, and this may make it more suitable for use on sets of proteins where there are many new folds, such as complete genomes, provided the false positive rate is carefully considered.

Consideration of performance as a function of the declared probability of disorder for each residue allows comparison of the methods at the same false positive rates. From this view-point, which method is superior depends sharply on whether or not there is a known structural relative of the target structure. Group 68 has the best performance for targets with close structural relatives, and group 355 the best when there are no close structure relatives known.

As noted earlier, there are issues with defining disorder based on experimental structures. All the identified problems tend to increase the apparent disorder beyond that which may exist *in vivo*. The possible effect of this on the accuracy of the methods is hard to estimate. All the methods train on experimental structures, and are tuned to minimize false negatives, so will try to compensate for any training set problems in some average way. In the longer term, NMR may offer the best hope of extensive and reliable data on disorder, and when more such data are available, accuracy may improve.

After the evaluation for the CASP meeting was complete, group 355 suggested that a probable reason for the over-prediction of their method was because it had been trained with relatively long disordered regions. The implication is that the method would perform more reliably for predicting longer windows of disorder than a single residue. Examination of the data suggests that the over-prediction is primarily but not solely due to over-extending short regions of disorder to longer regions, rather than predicting disorder where none is present. There are insufficient long windows in the target set to evaluate performance on those alone.

ACKNOWLEDGMENTS

We thank the prediction teams who contributed disorder predictions, so making possible the first objective evaluation of these methods.

REFERENCES

1. Moult, J., Fidelis, K., Zemla, A. & Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Suppl 5, 2–7.
2. Dyson, H. J. & Wright, P. E. (2002). Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. Adv Protein Chem 62, 311–40.
3. Kim, A. S., Kakalis, L. T., Abdul-Manan, N., Liu, G. A. & Rosen, M. K. (2000). Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. Nature 404, 151–8.
4. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 12, 54–60.
5. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. Biochemistry 41, 6573–82.
6. Norvell, J. C. & Machalek, A. Z. (2000). Structural genomics programs at the US National Institute of General Medical Sciences. Nat Struct Biol 7 Suppl, 931.