PROTEINS: Structure, Function, and Genetics 53:573–578 (2003)

Prediction of Disordered Regions in Proteins From Position Specific Score Matrices

David T. Jones* **and Jonathan J. Ward**
Department of Computer Science, Bioinformatics Unit, University College London, London, United Kingdom

ABSTRACT **We describe here the results of using a neural network based method (DISOPRED) for predicting disordered regions in 55 proteins in the 5th CASP experiment. A set of 715 highly resolved proteins with regions of disorder was used to train the network. The inputs to the network were derived from sequence profiles generated by PSI-BLAST. A post-filter was applied to the output of the network to prevent regions being predicted as disordered in regions of confidently predicted alpha helix or beta sheet structure. The overall two-state prediction accuracy for the method is very high (90%) but this is highly skewed by the fact that most residues are observed to be ordered. The overall Matthews' correlation coefficient for the submitted predictions is 0.34, which gives a more realistic impression of the overall accuracy of the method, though still indicates significant predictive power. Proteins 2003;53:573–578.** © 2003 Wiley-Liss, Inc.

Key words: protein structure prediction; folding; disorder; neural networks; sequence analysis

INTRODUCTION

One of the central tenets of structural biology is that the function of a protein is determined by its three-dimensional structure. As a result, predicting protein structure has often been at the forefront of efforts to infer function. However, a large proportion of protein sequences appear not to code for globular structures at all. These regions may adopt a non-globular structure or even be unfolded (disordered) in solution. The abundance of these disordered regions suggests that they are, to some extent, evolutionarily conserved and therefore likely to possess biological function. Experiments have also shown that non-ordered regions are present in DNA-binding sites[1] and are involved in several other types of molecular recognition. It has been suggested that the absence of a well-defined structure allows disordered binding sites to interact with several different targets. Transitions between the native unfolded state and phosphorylation-induced nascent globular structure may also provide thermodynamic regulation of binding.[2] The automated prediction of disordered regions would provide a first step in high-throughput methods for identifying such disordered binding sites. The identification and removal of unstructured regions is also vital for the successful crystallisation of proteins prior to X-ray structure determination.

The premise that structure is determined by primary sequence might also be reasonably applied to lack of structure or disorder, and several methods for predicting disordered regions have been described in the literature.[3–4] There are also clear patterns that characterise disordered regions such as low sequence complexity, amino acid compositional bias (e.g. towards aromatic residues) and high flexibility.

The first attempt at predicting disordered regions used amino acid composition within a sliding window of residues along with mean hydropathy and flexibility.[3] A simple rule-based classifier was used to predict long (>44 residue) segments. Shorter segments were classified using feed-forward neural networks with features selected by a sequential forward search algorithm. This was followed by a method for predicting disorder at the N- and C- termini[4] along with improved feature selection.

The prediction of single-residue properties from amino acid sequence has a long history in bioinformatics with numerous algorithms for predicting secondary structure and solvent accessibility. Developments in secondary structure prediction have shown that accuracy can be improved by incorporating evolutionary information from sequences that have homology with the target.[5] Evolutionary information is likely to lead to improved prediction of secondary structure by indicating the level of conservation and properties of the substituted residues. The most accurate modern methods, such as PSIPRED,[6] use position-specific scoring matrices from iterated BLAST searches over large, filtered sequence databases.[7] The search can be carried out fairly efficiently and many distant homologues of the target sequence are recovered. Here we describe the application of a similar strategy to the prediction of protein disorder.

MATERIALS AND METHODS

The basic method used to predict disordered regions is similar in many respects to the PSIPRED method for secondary structure prediction,[6] which uses neural networks to analyse sequence profiles generated by PSI-BLAST.[7] In this instance, rather than a 3-state prediction of helix, strand or coil, a simple 2-state prediction is

*Correspondence to: David Jones, Department of Computer Science, Bioinformatics Unit, University College London, Gower St., London, WC1E 6BT. E-mail: dtj@cs.ucl.ac.uk

Received 26 February 2003; Accepted 17 June 2003

required: ordered or disordered. The method we employed for the CASP5 predictions, named DISOPRED, was very much a prototype and was based on just a single feed-forward neural network with 315 input units (window size of 15 residue positions and 21 units per residue position), 55 hidden units and two output units.

In order to train the network a non-redundant (pairwise sequence identity less than 25%) training set was compiled from X-ray structures extracted from PDB.[8] Disordered regions were identified by aligning the sequence of the protein chain as specified by the SEQRES records in the PDB file with the sequence as specified by the ATOM records (alpha-carbon coordinates). Residues which were found in the SEQRES records but not in the ATOM records were assumed to be disordered. Only highly resolved structures (resolution < 2.0 Å) were used in order to ensure that missing coordinates were not caused by poor resolution. The final non-redundant set comprised 715 protein chains, in which a total of 176550 residues were classed as ordered and 4590 residues classed as disordered.

For each protein in the training set, a sequence profile was calculated using 3 iterations of PSI-BLAST searching against a filtered non-redundant sequence data bank as described earlier.

10% of the data was kept aside as a validation set to prevent over-fitting i.e. training was stopped as soon as the network prediction accuracy on the validation set began to get worse. The predicted state for each input window of 15 residues was determined by which of the two outputs (ordered or disordered) was highest, and the confidence of the prediction taken as the absolute difference between the two output values. According to the CASP submission requirements, confidence values should be expressed on a scale of 0 to 1 where 0 indicates high confidence that a residue is ordered and 1 indicates high confidence that it is disordered. The network outputs were scaled according to this scheme. This method was then used to predict the disordered regions in 67 CASP5 targets, of which 55 could be evaluated.

We noted very quickly that the network was making obvious over-predictions of disorder in regions that were confidently predicted to be in regular secondary structure by PSIPRED. An obvious idea would have been to include predicted secondary structure as an input to the neural network in addition to the sequence profiles. Unfortunately, there was insufficient time to implement this scheme before the first CASP5 closing dates, and so an *ad hoc* scheme was used to take predicted secondary structure into account. When deriving the confidence of the disorder prediction a note was made of the predicted secondary structure generated by PSIPRED for the residue being predicted. Where the secondary structure prediction indicated that the residue was in a helix or beta strand region with a confidence $> 50\%$, and the residue had been predicted as disordered, the prediction was changed to ordered, but with only a marginal confidence score of 0.49. This crude post-filtering of the disorder predictions was fairly effective, although it is likely that a more rigorous way of taking predicted secondary structure information into account will produce better results.

RESULTS

Table 1 summarizes the predictions made using DISOPRED for the 55 CASP5 targets solved by X-ray crystallography in time to be assessed. It is clear from the results that there is a tendency for DISOPRED to over-predict disordered regions, but the prediction rate is much more reasonable if only the disordered regions with high confidence (network output $>= 0.8$) are considered (see Figure 1).

Figure 2 shows the Receiver Operating Characteristic (RoC) curves (true positive rate plotted against false positive rate) for our method when applied to the targets with and without detectable sequence similarity to already-known structures. It is clear from these curves that DISOPRED is better at predicting the targets with sequence similarity to known structures than pure fold recognition and new fold targets. Interestingly, this bias towards sequence-similar targets is even apparent for targets with no homologues in the disordered protein training set. Presumably in these cases, the increased performance is due to the increased accuracy in the secondary structure predictions used in the post-filtering step.

Table 2 shows a number of overall performance measures for the DISOPRED CASP predictions. The Wilcoxon statistics (area under the RoC curve) are calculated by trapezium rule numerical integration for 1000 points. The correlation coefficients and percentage accuracy values were calculated at a decision threshold of 0.5 i.e. the threshold score between assigning a residue as ordered or disordered.

At first sight, the overall percentages of accuracy (Q_2) seem very impressive. A total of 90% of residues in the targets are correctly assigned as ordered or disordered. However, the high accuracy scores are not particularly informative because of the very unbalanced frequencies of ordered and disordered residues (i.e. similar accuracies could be achieved by classifying all residues as ordered). The Matthews' correlation coefficient (MCC)[9] is generally considered to be a more useful performance measure for problems such as this with unequal prediction class frequencies. For the targets in the non-homologous category, the MCC is 0.30, increasing to 0.34 when all targets are considered. For comparison, a good modern secondary structure prediction method will give a MCC of approximately 0.5 for coil prediction, and methods for beta-turn prediction will give a MCC of around 0.3, indicating a similar level of accuracy to DISOPRED.

For comparing a number of methods with relatively small differences with respect to their RoC curves, the Wilcoxon statistic may well prove to be the best metric. Generally speaking, the best prediction method is the one with the greatest area under the RoC curve, with a perfect predictor giving an area of 1.0. The advantage of the Wilcoxon statistic is that a straightforward statistical test

TABLE I. Summary of the Prediction Results for the Evaluated CASP Targets

Target	Length	No. of disordered residues	No. of high B-value residues	No. predicted	No. predicted (conf. > 0.8)
T0129	182	12	34	36	21
T0130	114	14	31	10	8
T0132	154	7	28	11	5
T0133	312	19	20	61	20
T0134	251	18	35	43	13
T0135	108	2	9	11	4
T0136	523	3	7	49	9
T0137	133	0	0	10	8
T0138	135	0	0	19	10
T0139	83	21	21	12	8
T0140	103	17	17	26	15
T0141	187	0	0	21	10
T0142	282	2	63	31	9
T0143	216	0	0	27	7
T0146	325	26	201	32	13
T0147	245	11	13	7	4
T0148	163	1	35	23	5
T0149	318	1	111	29	6
T0150	102	5	6	8	8
T0151	164	58	127	72	60
T0152	210	12	17	18	13
T0153	154	20	20	31	19
T0154	309	21	22	23	10
T0155	133	16	16	9	8
T0156	157	1	1	11	6
T0157	138	18	30	21	9
T0159	309	0	0	26	11
T0160	128	2	2	18	12
T0161	156	2	39	12	5
T0162	286	11	177	26	9
T0165	318	0	0	35	9
T0167	185	5	5	7	5
T0168	327	16	41	29	7
T0169	156	0	0	8	6
T0170	69	0	0	19	9
T0172	299	6	19	33	10
T0173	303	16	16	35	5
T0174	417	65	67	89	40
T0177	240	20	42	26	13
T0178	219	0	1	8	5
T0179	276	2	24	11	5
T0181	111	0	0	26	21
T0182	250	1	45	13	7
T0183	248	1	5	14	10
T0184	240	3	14	28	16
T0185	457	29	182	16	10
T0186	364	1	92	7	3
T0187	417	3	3	35	8
T0188	124	17	17	34	17
T0189	319	0	22	7	2
T0190	114	3	3	15	8
T0191	282	0	158	14	6
T0192	171	1	13	6	5
T0193	211	6	59	15	7
T0195	299	9	9	36	10

The number of residues assigned as disordered by identifying missing regions of the structure is compared with the number of residues with high B-values, along with the number of residues predicted to be disordered by DISOPRED. The number of residues confidently predicted as disordered (network output > 0.8) is also shown.

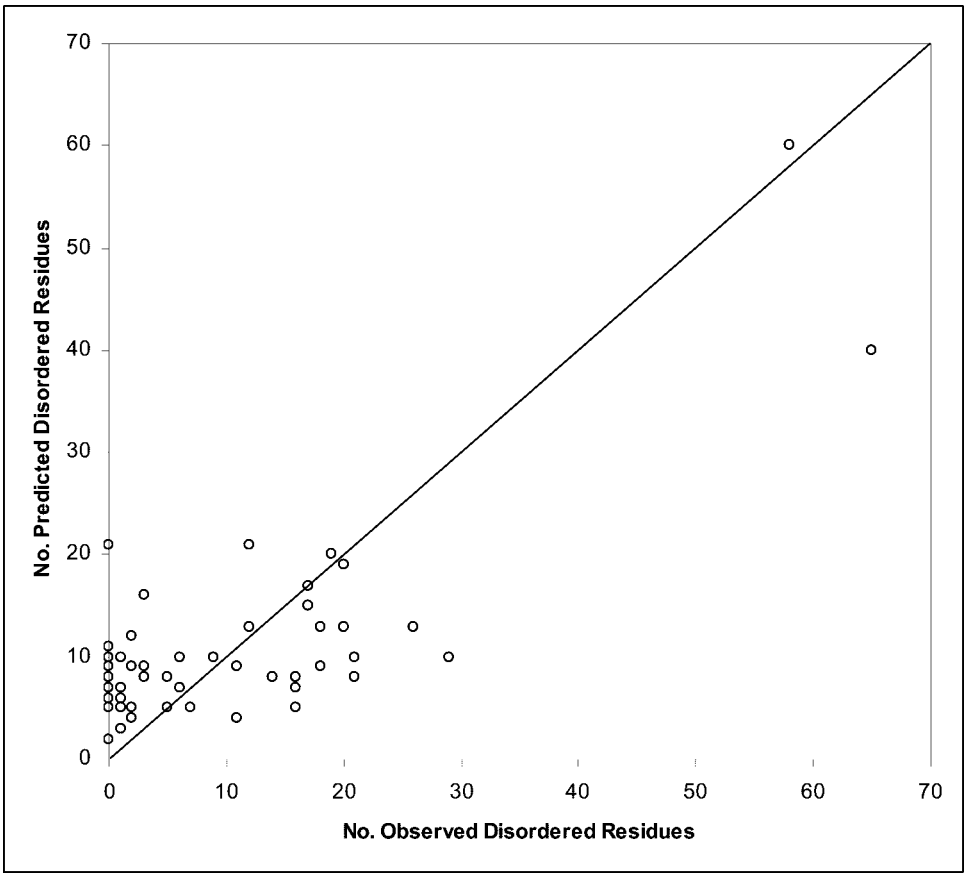

Fig. 1. The number of confidently predicted disordered residues (neural network output > 0.8) is plotted against the observed number of disordered residues for the targets listed in Table 1. The diagonal line represents the ideal relationship.

can be applied to decide whether or not two predictors are significantly different in their performance.

Although all of the prediction results discussed above are based on X-ray crystallographic data, it is interesting to see how DISOPRED fares when compared to NMR structure data. Figure 3 shows the results of applying DISOPRED to Target T0170 which is an NMR structure comprising an ensemble of 25 structures. Regions where there is a large variation between NMR models may result from a lack of sufficient experimental restraints, or actual motion or disorder of the molecule in solution. Although it is generally possible to distinguish these two cases in modern NMR experiments, no information was available as to whether the variable regions in T0170 are really caused by disorder. Nevertheless, the variable regions in the structure do seem to match up with the regions that are predicted to be disordered by DISOPRED, but we cannot generalise too much from just one example.

DISCUSSION

As is customary for CASP reports, we will consider what went right, what went wrong and any reasons for the successes and failures. Clearly, in terms of what went right, we have shown that a very simple method for predicting disordered regions in proteins, based on a well-tried technique for secondary structure prediction, and trained on a data set compiled in a very simplistic fashion, was in fact able to predict the majority of disordered regions in the 55 CASP targets we tried.

Clearly, DISOPRED is better at predicting proteins with detectable sequence similarity to proteins of known 3-D structure, but in our tests this seems to be mainly due to the increased accuracy of the secondary structure predictions used to post-filter the predictions made by the neural network. We are currently investigating how much value is added by the evolutionary information provided by the sequence profile. Although there are clear reasons why this information is valuable for secondary structure prediction, it is less clear how much sequence conservation can tell us about protein disorder.[10] Preliminary results from further tests on a more sophisticated disorder prediction method (J. Ward et al., paper in preparation) suggest that the sequence profile information improves the prediction of regions that have marginal confidence scores.

Despite the fact that DISOPRED is clearly a useful predictor of disordered regions in proteins, there is still a lot of scope for improvement. The relatively high accuracy but low MCC values suggest that over-prediction of disordered regions is clearly a problem. In particular, the N- and C-termini of proteins are generally cases where disor-

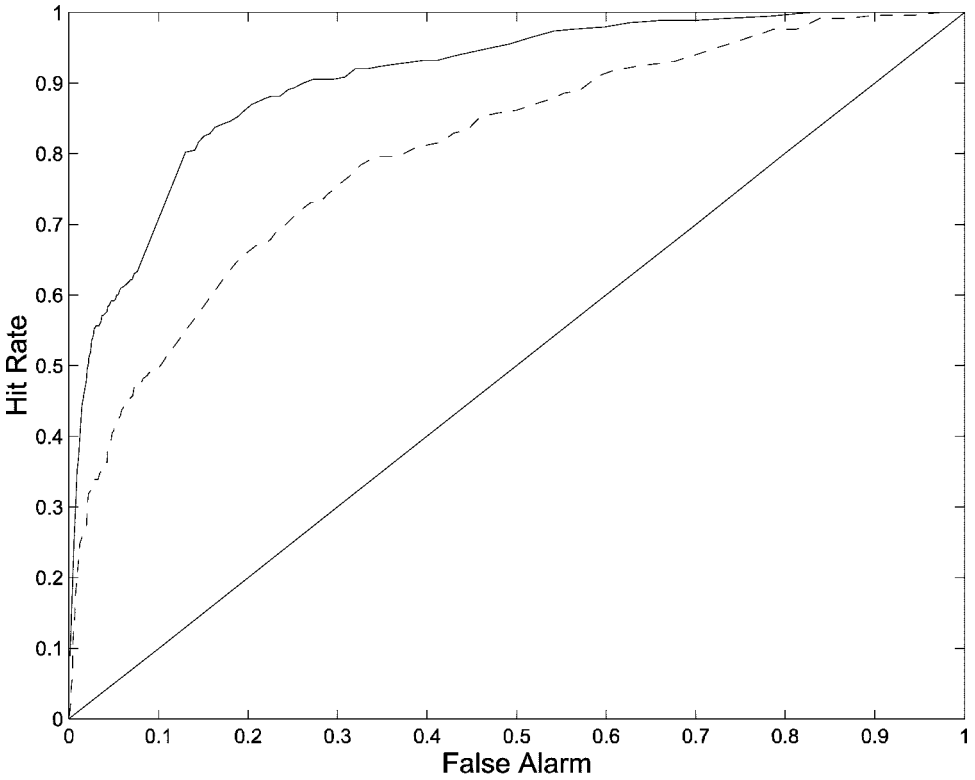

Fig. 2. The hit (true positive) rate is plotted against false alarm (false positive) rate to give RoC curves for the targets both with (solid curve) and without (dashed curve) close homologues of known 3-D structure. The result expected for a completely random predictor is shown as a solid diagonal line.

TABLE II. Statistics for Targets With Sequence Similarity to Known Structures

	CM targets	Non-CM targets	All targets
Wilcoxon	0.91	0.81	0.87
Matthews Correlation Coefficient	0.36	0.30	0.34
% Accuracy (Q_2)	91.30	86.80	90.02

CM, Comparative Modelling, non-CM targets, and the whole set.

der is over-predicted. This is to be expected as many proteins have disordered termini, and so a learning algorithm will tend to learn this as a rule, but unfortunately not all proteins have disordered termini. It may be possible to attenuate the disorder predictions at the ends of the sequence, or perhaps to train separate networks to specifically handle the prediction of disorder at the sequence ends in a similar fashion to the approach of Li *et al.*[4]

We expect that the rather *ad hoc* usage of predicted secondary structure as a post-filtering step is sub-optimal. Retraining the neural network with the predicted secondary structure of the target protein as an explicit additional input is very likely to produce better results than those described here.

One practical failure in our efforts to predict disorder in CASP was that we submitted predictions for all the targets except target 145, which of course turned out to be a natively disordered protein. In fact we did make a prediction for this target (125 out of 210 residues were predicted as disordered) but thought this to be caused by an error in the program or failure in the profile alignment, and so did not submit it to the CASP server. This is a clear example where human intervention was detrimental to success.

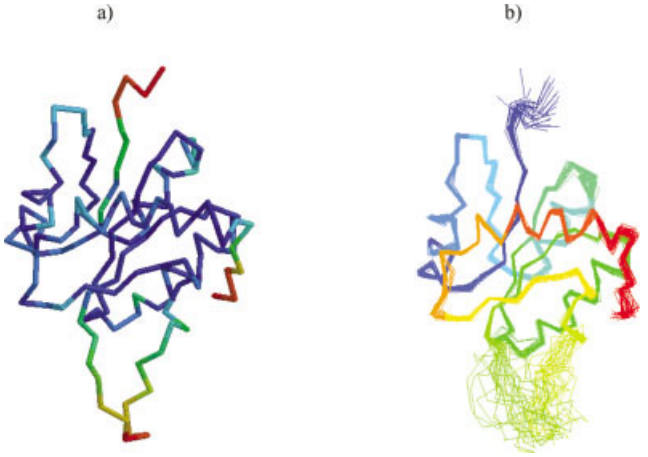

Fig. 3. The DISOPRED prediction for a structure solved by NMR is shown (Target 170). a) shows the DISOPRED prediction. Residues are coloured according to the estimated confidence that a residue is disordered, with red indicating the regions predicted with highest confidence and blue indicating the regions least likely to be disordered. b) shows the structural ensemble for the target. The local variability of the NMR ensemble clearly demarks regions which could be disordered, with good agreement with the disorder prediction. Colouring in this case is according to position in the sequence from residue 1 (blue) to residue 69 (red).

CONCLUSION

DISOPRED has been shown to be a useful prediction method for determining disordered regions in proteins, although close examination of the results shows some scope for improvement, particularly in the over-prediction of disorder at the ends of proteins. In terms of the benchmarking and evaluation of the method, it is clear that some further thought needs to be given to the best way of measuring the accuracy of disorder predictions, given that the two classes (ordered/disordered) are so unequally represented, though the Matthews correlation coefficient appears to be one usable measure.

The DISOPRED source-code can be downloaded from the URL ftp://bioinf.cs.ucl.ac.uk/pub/DISOPRED

ACKNOWLEDGMENTS

Jonathan Ward is supported by a Medical Research Council Studentship.

REFERENCES

1. Weiss MA, Ellenberger T, Wobbe CR, Lee JP, Harrison SC and Struhl K. Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature. 1990;347:575–578.
2. Wright PE and Dyson HJ. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.
3. Romero P, Obradovic Z, Kissinger CR, Villafranca JE and Dunker AK. Identifying disordered regions in proteins from amino acid sequences. Proc IEEE International Conference on Neural Networks 1997: 90–95.
4. Li X, Romero P, Rani M, Dunker AK, and Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. Genome Informatics 1999;10:30–40.
5. Rost B and Sander C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 1993;232:584–99.
6. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999;292:196–202.
7. Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res. 1997;25(7):3389–3402.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl. Acids Res. 2000; 28:235–242.
9. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta. 1985; 405: 442–451.
10. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 2002; 55: 104–110.