

## RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins

Zheng Rong Yang<sup>1</sup>, Rebecca Thomson<sup>2</sup>, Philip McNeil<sup>3</sup> and Robert M. Esnouf<sup>2,\*</sup>

<sup>1</sup>School of Engineering and Computer Science, Exeter University, Exeter EX4 4QF, UK, <sup>2</sup>Division of Structural Biology and Oxford Protein Production Facility, University of Oxford, Henry Wellcome Building for Genomic Medicine, Roosevelt Drive, Oxford OX3 7BN, UK and <sup>3</sup>European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on July 9, 2004; revised on June 7, 2005; accepted on June 8, 2005

Advance Access publication June 9, 2005

### ABSTRACT

**Motivation:** Recent studies have found many proteins containing regions that do not form well-defined three-dimensional structures in their native states. The study and detection of such disordered regions is important both for understanding protein function and for facilitating structural analysis since disordered regions may affect solubility and/or crystallizability.

**Results:** We have developed the regional order neural network (RONN) software as an application of our recently developed 'bio-basis function neural network' pattern recognition algorithm to the detection of natively disordered regions in proteins. The results of blind-testing a panel of nine disorder prediction tools (including RONN) against 80 protein sequences derived from the Protein Data Bank shows that, based on the probability excess measure, RONN performed the best.

**Availability:** RONN is available at <http://www.strubi.ox.ac.uk/RONN>. Requests for the RONN software and the database of disorder (XML format) can be directed to the corresponding author.

**Contact:** robert@strubi.ox.ac.uk

**Supplementary information:** Details of all predictions made during blind testing, also available at [http://www.strubi.ox.ac.uk/RONN3\\_Supplementary.pdf](http://www.strubi.ox.ac.uk/RONN3_Supplementary.pdf)

### 1 INTRODUCTION

It has been widely accepted that a protein's primary sequence determines its three-dimensional structure, which in turn determines its function (Anfinsen, 1973). However, the subtle interplay of atomic forces, solvent, environment and protein-folding machinery renders the *ab initio* determination of structure from primary sequence alone a remote possibility. Nevertheless, comparative studies of amino acid sequences have been widely and successfully used for analyzing biological data, for instance in the detection of homologous proteins and the recognition of functional sites (Bartlett *et al.*, 2002).

Recent studies have added complication by focusing attention on amino acid sequences that show no propensity to form specific three-dimensional structures, yet may still be functionally significant. Many proteins contain local regions of such disorder, and some appear to be totally unfolded in their native states (Dunker *et al.*,

2000). These studies have further found that many natively unfolded proteins or protein regions are involved in molecular recognition, depending on disorder-to-order transitions, to enable the natively unfolded proteins to form complexes with their cognate partners (Alber *et al.*, 1983; Weinreb *et al.*, 1996). Such molecular interactions can be enzyme–substrate, receptor–ligand, protein–protein, protein–RNA or protein–DNA (Huber, 1979). The transition upon binding can give rise to an interaction having the biologically desirable property of combining high specificity with modest binding affinity (Schulz, 1979), thereby avoiding irreversible binding, which is neither suitable nor acceptable for most biological processes. For example, it has been found that disordered regions play an important role in cell signaling pathways (Frankel and Kim, 1991; Weinreb *et al.*, 1996).

Accurate recognition of disordered regions in proteins is important in molecular biology for enzyme specificity studies, function recognition and drug design. The detection of such regions is also crucial in structural biology since structure determination by X-ray crystallography and NMR relies on ensembles of (near) identical structures to amplify the experimental signal. At best, disordered regions are invisible to these techniques, at worst they can disrupt the whole experiment by affecting solubility and/or crystallizability (Oldfield *et al.*, 2005).

One of the most direct ways of identifying disordered regions experimentally is by looking for the amino acids for which there is no electron density in the maps obtained by X-ray crystallography. However, a relatively small fraction of known proteins have had their structures determined and this method has some bias since only partly structured proteins that have been crystallized can be included. Furthermore, consider a terminal domain of a protein that is believed to be in an expression construct but is not observed in electron-density maps: (1) this whole domain may be disordered, (2) a short disordered linker may lead into an ordered domain having no fixed orientation relative to the rest of the protein rendering it invisible in the electron density or (3) the protein may have been unexpectedly proteolyzed and the domain may be absent from the crystallized entity. In this respect, structures determined by NMR have the advantage since independent domain structures may be still determined for domains connected by an 'invisible' flexible linker region. Classification of disordered regions in entries in the Protein Data Bank (PDB) (Berman *et al.*, 2000) is incomplete, although

\*To whom correspondence should be addressed.

since 1998 it has been possible to annotate unobserved residues using REMARK 465 records. However, even for some recent PDB entries such remarks have been found to be incomplete and/or inconsistent.

Garner *et al.* (1998) have shown that disordered regions comprise a sequence-dependent category distinct from that of ordered protein structure, i.e. certain amino acids have a higher frequency of occurrence in disordered regions than in ordered regions and vice versa. The aromatic amino acids (Trp, Tyr and Phe) are less likely to appear in long disordered regions (Kissinger *et al.*, 1995), the biological interpretation being that aromatic amino acids have a strong interaction capability that can develop structure and hence inhibit disorder (Burley and Petsko, 1985). Additionally, Cys and His are less common in disordered regions. On the other hand, Glu, Asp and Lys are more likely to appear in disordered regions since they are charged and charge imbalance tends to favor disorder. Another residue associated with disorder, Ser, increases solubility and provides a flexible locus, two properties inherent to disordered regions.

The best-known, and most intensively developed, method for disorder detection is PONDR<sup>®</sup>, which employs pattern recognition algorithms using a set of features based on biological knowledge (Romero *et al.*, 1997, 2000; Garner *et al.*, 1998, 1999; Xie *et al.*, 1998; Li *et al.*, 1999, 2000; Vucetic *et al.*, 2001; Radivojac *et al.*, 2003, 2004). The first predictions analyzed the frequency measurements of eight amino acids (His, Glu, Lys, Ser, Asp, Cys, Trp and Tyr) and two average attributes (hydropathy and flexibility) from which a feed-forward neural network model with six hidden units was constructed (Romero *et al.*, 1997). Neural networks have also been trained using the disorder data derived from X-ray crystallographic and NMR methods separately (Garner *et al.*, 1998). Several pattern recognition models have been compared using different combinations selected from 51 possible features, and disorder has been predicted separately for N-terminal, C-terminal and internal regions (Li *et al.*, 1999). All the above methods are generally heuristic with rules derived from prior biological knowledge and therefore rely on the proper determination of an optimized subset of features for the characterization of disorder.

The prediction of disorder is now the focus of intensive research and disorder prediction assessment recently formed part of the CASP6 trial (<http://predictioncenter.llnl.gov/casp6/Casp6.html>), where 20 methods were compared in blind tests. Despite this effort, the accurate prediction of protein disorder remains problematic and we developed a novel approach to the problem based on an extension of our bio-basis function neural network method, a sequence alignment technique originally developed for the detection of protease cleavage sites (Thomson *et al.*, 2003; Yang and Thomson, 2005). An initial implementation demonstrated potential for detecting regions of disorder, but prediction accuracies per residue were low (Thomson and Esnouf, 2004). We have now substantially improved on this original method, which we term the regional order neural network (RONN), and here describe its underlying principles, and its training and validation based on a large set of disorder data derived from the PDB. Finally, we report the results of blind tests on nine disorder prediction methods (including RONN) against additional data derived from the PDB and show that based on our favored measure of accuracy, the probability excess, the current version of RONN (version 3) gives the best performance overall.

## 2 SYSTEMS AND METHODS

### 2.1 Bio-basis function neural networks

If two proteins have similar biological functions then the primary sequences usually demonstrate a significant similarity, which can be detected by sequence alignment techniques, usually using a mutation matrix to score the similarity. Conversely, the biological function of a query sequence is expected to be the same as that of a protein sequence of known function if the alignment score is high enough, with tools like BLAST and its variations commonly used for detecting this similarity (Altschul *et al.*, 1990, 1994, 1997). We apply this expectation to the detection of disordered regions: sequences are compared with a series of sequences of known folding state (ordered, disordered or a mixture of both) and the alignment scores against these sequences are used to classify each sequence as ordered or disordered using a suitably trained neural network.

To be more general, the decision about the likelihood of disorder is based on alignments to an ensemble of sequences of known folding state. Consider a database consisting of sequences corresponding to  $M$  ordered and  $N$  disordered regions of proteins. How can a decision on the folding state of an unknown protein be based on these  $M + N$  alignment scores? Finding the best alignment among the ensemble is one option, but this might not be the best way if some sequences are wrongly labeled or if none of the scores is above a threshold value. A natural generalization is to regard each known sequence as a prototype as in the prototype theory (Dasarathy, 1991, 1994). The recognition of a property is then based on the statistical relationships between a query and the prototypes, and an obvious method of implementation uses neural networks. We originally applied this idea to a sequence-based biological problem, the analysis of protease cleavage rates of specific sub-sequences, naming it the bio-basis function neural network (BBFNN) method (Thomson *et al.*, 2003; Yang and Thomson, 2005). A distinguishing feature of this approach is that individual amino acids and sequences are not represented in some arbitrary feature space (such as hydrophobicity and charge) according to known properties. Rather, 'distances' (determined by sequence alignment) from a subset of well-characterized prototype sequences are calculated and training of the neural network is performed in this 'distance' space.

The original BBFNN was developed for analyzing sub-sequences of fixed length, transforming a series of alignment scores to a similarity value in a monotonically decreasing way. However, since the length of disordered/ordered regions varies, for this implementation, the BBFNN had to be revised by using the concept of non-gapped homology alignment to maximize the alignment score between pairs of sequences (Thomson and Esnouf, 2004). In the version of RONN described here (version 3), the prototype sequences can have different lengths, although they must be at least as long as a pre-defined window size, and sub-sequences for a query sequence (of this window size and centered on each residue in turn) are then aligned to all the prototypes. The resulting homology scores are used for statistical pattern recognition to give a probability of disorder for each query sequence window, and these scores are averaged to give a probability of disorder for each residue in the query sequence.

### 2.2 The dataset of disordered proteins and regions

To develop disorder prediction techniques, it is essential to use a large and reliable dataset of ordered and disordered sequences for training, validation and testing. Annotation of disorder is fraught with problems for the reasons described above, and we have devoted significant effort to the compilation of a comprehensive dataset based on a complete analysis of entries in the Molecular Structure Database (MSD) (Boutselakis *et al.*, 2003; April 29, 2004 release) hosted at the European Bioinformatics Institute.

For each entry in the MSD, the sequence of observed residues (derived from ATOM coordinate records in the PDB entry) is aligned both against the complete protein sequence as shown by the SEQRES records and against the sequence in the corresponding UniProt entry. These two alignments are merged to give the complete residue-level mapping between the sequence of the polypeptide used in the experiment and its UniProt counterpart, thereby

**Table 1.** Datasets for training and blind testing of RONN and comparison between different disorder prediction methods

	Training set	Main blind set	Secondary set
Proteins (PDB entries)	—	80	190
Ordered regions	891	—	—
Disordered regions	530	—	—
Ordered residues	170 923	29 909	64 838
Disordered residues	19 427	3 649	5 680
Total residues	190 350	33 558	70 518

The training set provides prototypes and is used for training and validation. Proteins in the main blind set contain at least one region of disorder of at least 21 consecutive residues. Proteins in the secondary blind set contain at least one region of disorder of at least five consecutive residues.

detecting errors, avoiding any dependence on the REMARK 465 records and allowing entries pre-dating 1998 to be included.

The analysis was performed on all the 25 931 entries in the MSD and it found 7327 entries for which there were unobserved regions of at least five consecutive residues. These entries were divided into two sets: those with at least one disordered region of >20 consecutive residues (long set; 1573 entries) and the rest (short set; 5754 entries). Further filters were applied to each set:

- Only structures obtained by X-ray crystallography or NMR were included (i.e. excluding theoretical models and structures determined by electron diffraction, electron microscopy or unspecified methods).
- Multi-component complexes were excluded as disordered regions may have undergone disorder-to-order transitions in these structures.
- Only the entry with the highest number was included for entries with the same 3-letter code (e.g. 1MOB and 2MOB) since, in general, this entry is the most recent. (Although not always true, this acts as a useful filter for very similar sequences.)
- For entries containing multiple chains (of identical sequence), only residues that were not observed in all the chains were used.

After filtering, the long set contained 872 entries from which 105 entries were randomly selected to form the main blind test set. A secondary test set was created by adding randomly selected members of the filtered short set to the main blind set. From the remainder of the long set, sequences for 1691 ordered regions and 983 disordered regions were extracted for training and validation of RONN.

At this point, each of the four datasets (ordered training set, disordered training set, main blind set and secondary blind set) was filtered using Cd-Hit to remove the similar sequences (Li *et al.*, 2001, 2002). For a sequence identity cut-off of 90% many sequences were removed. But as this cut-off was reduced to ~60%, relatively few additional sequences were excluded. Below ~60%, the rate of removal increased again as matches were found between quite distantly related sequences. Thus, a 70% sequence identity cut-off was selected as the best compromise between maximizing the number of sequences included and minimizing the redundancy in the data. The sizes and compositions of these different datasets are summarized in Table 1.

### 2.3 Training and validation of RONN

*Step 1.* The 1421 training sequences are partitioned into 10 ‘folds’. For each run of 10-fold cross-validation, nine folds are used as prototypes (in some cases only disordered sequences are used as prototypes) and the remaining fold is randomly divided into two parts: one for training to estimate model parameters and one for validation to estimate the probability density functions

for both ordered and disordered classes, to allow Bayes’ rule to be used for decision making.

*Step 2.* The BBFNN is trained using the training sequences. A window of pre-determined size is used to scan through the training sequences to generate training sub-sequences, which are then aligned with each prototype in turn to maximize the non-gapped homology alignment score. The bio-basis function is then used to transform this score into a normalized similarity measurement. Each prototype is considered to be an independent variable and the relationship between these independent variables and the class labels (ordered or disordered) of sub-sequences is assumed to be simple, allowing a linear classifier to be constructed. Parameters weighting the similarity measurements are determined using a pseudo-inverse method (Thomson *et al.*, 2003; Yang and Thomson, 2005).

*Step 3.* The discrimination threshold between ordered and disordered sequences is optimized using the validation sequences. These sequences are divided into validation sub-sequences in the same way as for training sub-sequences. After the input into the model constructed in Step 2, the probability density functions of the outputs are assumed to follow two Gaussian distributions, one for ordered and the other for disordered sequences. A parametric method is used to approximate these two functions. Based on the approximated probability density functions, an optimum threshold for discrimination can be determined using Bayes’ rule (Duda *et al.*, 2002). Cost functions are used for the final decision making. Let the probabilities of order and disorder generated by the trained model be  $P_o$  and  $P_d$ , respectively, with associated cost functions  $C_o$  and  $C_d$  defined such that  $C_o + C_d = 1$ . The final probability of disorder is then  $C_d P_d / (C_o P_o + C_d P_d)$  and is assigned to all the residues within the current window.

*Step 4.* For each position of the sliding window, all residues within the window are assigned the same probability of disorder. When the window is moved on one residue, another probability of disorder is generated. Thus, for a window size of  $W$  residues, every residue will have up to  $W$  probabilities of disorder assigned. To arrive at a single estimate for each residue these estimates are simply averaged and the residue is predicted to be disordered if the averaged probability of disorder is >0.5.

*Step 5.* The training and validation process (Steps 2–4) is repeated 10 times using the non-overlapping folds created in Step 1.

### 2.4 Blind testing of disorder prediction methods

Nine methods for predicting protein disorder were tested against the 80 proteins forming the smaller blind test set, each of which contains at least one significant region of disorder. (Since there was no way of identifying the sequences that were used for training each method, the random selection procedure described above was the most appropriate.) The methods tested were RONN, PONDR<sup>®</sup> (VL-XT method), FoldIndex (<http://biportal.weizmann.ac.il/fldbin/findex>), all three variants of DisEMBL 1.4 (coils/loops, hot loops and REMARK 465) (Linding *et al.*, 2003a), GlobPlot 2 (Linding *et al.*, 2003b), DISOPRED2 (Ward *et al.*, 2004) and PreLink (Coeytaux and Poupon, 2005). Except for PONDR<sup>®</sup>, which was licensed and installed locally, all predictions were made by submitting sequences to the public web servers. For each prediction, only the binary classification of each residue as either ordered or disordered was recorded and analyzed (rather than the probability of disorder) since this most closely reflects the use of these tools by structural biologists.

A final blind test set was compiled from two external sources to test whether the common selection process used for both training and testing sequences affected the results. A set of disordered sequences where the disorder had been established experimentally by methods other than X-ray crystallography was obtained from Uversky *et al.* (2000). Of the 91 disordered regions described it was possible to identify precise sequences for 79, which formed the disordered part of the test set. To balance these, a set of 80 protein sequences that was annotated as being ordered on the PONDR<sup>®</sup> website (retrieved in February 2003) was added. This final blind test set comprised 16 568 ordered residues and 14 462 disordered residues (see Supplementary information).

### 3 RESULTS AND DISCUSSION

#### 3.1 Measuring the accuracy of prediction

Quantifying the accuracy of binary classification procedures (such as the prediction of order or disorder for a residue) is a common problem in computer science and many different measures have been devised. In each case the most appropriate measure will depend, among other things, on the relative frequencies of the two classes in the dataset and the relative importance of incorrect classification either way. For comparisons against the reference data the results can be described by four integers:

- (1) TN: true negatives, e.g. the number of correctly classified ordered residues.
- (2) FN: false negatives, e.g. the number of disordered residues incorrectly classified as ordered.
- (3) FP: false positives, e.g. the number of ordered residues incorrectly classified as disordered.
- (4) TP: true positives, e.g. the number of correctly classified disordered residues.

Further measures are derived from these values, for example the sensitivity, also referred to as the recall, is  $TP/(TP + FN)$ ; the specificity is  $TN/(TN + FP)$  and the precision is  $TP/(TP + FP)$ . However, none of these measures is adequate in isolation (e.g. predicting all residues to be disordered would give sensitivity = 1, but specificity = 0 and precision = fraction of disordered residues in the test set) and more complex measures are required.

An ROC curve is obtained by plotting sensitivity against (1 – specificity) for a classifier as a function of decision cut-off (e.g. probability of disorder required for a residue to be classified as disordered). The area under this curve is particularly useful as an indicator of algorithm robustness rather than for the comparison of different algorithms. Three further measures are defined: the accuracy is  $(TN + TP)/(TN + FN + FP + TP)$ ; the Matthews' correlation coefficient is  $(TN \cdot TP - FN \cdot FP)/\sqrt{[(FN + TP)(TN + FP)(FP + TP)(FN + TN)]}$  (Matthews, 1975) and the probability excess is  $(TN \cdot TP - FN \cdot FP)/[(FN + TP)(TN + FP)]$ . The accuracy is heavily affected by the relative frequencies of the two classes (e.g. if 90% of residues are ordered then randomly classifying 90% of residues as ordered will give an accuracy of 0.82 for an algorithm that only identifies 10% of disordered residues). The Matthews' correlation coefficient is a better measure since all random algorithms have a coefficient of 0 while a perfect algorithm has a coefficient of 1, but it is still strongly influenced by the relative class frequency in the test set. The probability excess also varies between 0 (for random picking) and 1 (perfect prediction) and in addition is formally independent of the relative class frequency in the test set. This value is also conveniently measured from a plot of sensitivity against specificity, it being the distance of a point from the diagonal line corresponding to random algorithms. Indeed, the probability excess can be shown to be simply sensitivity + specificity – 1. These properties make the probability excess our favored measure of performance and the one we use for ranking different algorithms.

CASP6 used a weighted score,  $S$ , defined as  $100(w_{TP}TP + w_{FP}FP + w_{TN}TN + w_{FN}FN)/(TP + FP + TN + FN)$ , where  $w_{TP}$  is the number of ordered residues divided by the total number of residues;  $w_{TN}$  is the number of disordered residues divided by the total number of residues;  $w_{FN} = -w_{TP}$  and  $w_{FP} = -w_{TN}$ .

For comparisons based on the same test set,  $S$  is directly proportional to the probability excess and is reported here for convenience.

#### 3.2 Optimizing, training and testing RONN

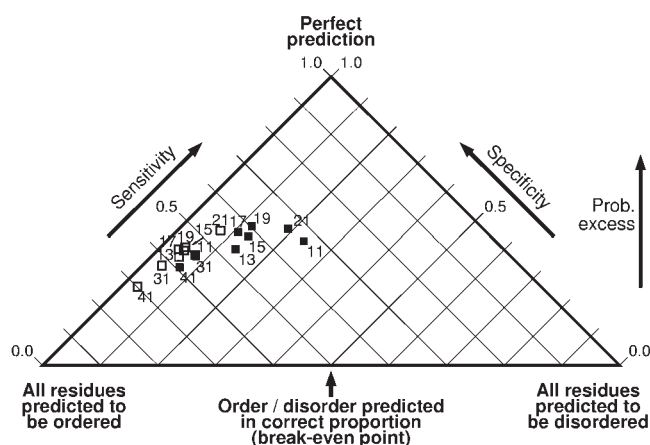
Within the framework of the RONN algorithm outlined above, several choices remain such as (1) what mutation matrix to use for alignments, (2) what prototypes to use, (3) what window size to use and (4) what choice of cost function to use. All these choices were explored by training and validating RONN and then measuring its performance against the two blind test sets.

The mutation matrix is central to RONN in that it determines how similar a window of a query sequence is to each of the prototype sequences. In sequence alignment tools like BLAST, the mutation matrix measures the effective likelihood of any given mutation occurring by chance during evolution, such matrices include the Dayhoff and Blosom62 matrices (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992). For disorder prediction the criterion of similarity might be considered to be the physical similarity between residues, which in turn determines the relative propensities for being in ordered/disordered regions. RONN was tested using the Dayhoff matrix, the Blosom62 matrix and a matrix devised to encode size, charge and hydrophobicity differences. In all cases, the Blosom62 matrix was found to perform the best (data not shown) and so was used for all subsequent testings.

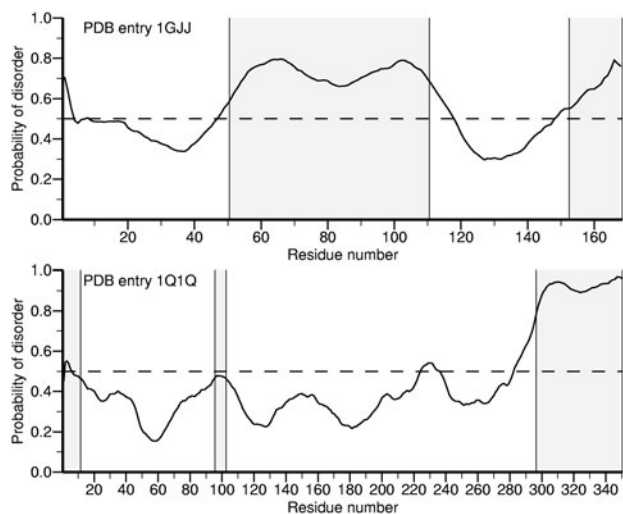
The number of prototype sequences used is obviously critical to the success of the algorithm and also affects its execution speed significantly. In general, the more sequences the better the chance of finding a significant match to any given query sequence, but redundancy in prototypes should also be avoided since it could lead to inappropriate weighting. The second question is whether to use just one class of prototypes or a mixture (ordered and disordered). It was found that including ordered prototype sequences did not improve the accuracy of the method (data not shown) and all subsequent testing solely used disordered prototype sequences, which also resulted in significantly faster predictions.

Blind tests for cost functions of 0.5 and 0.6 and for window sizes from 11 to 41 residues were explored systematically (Fig. 1). They showed that better predictions were consistently obtained with a cost function of 0.6, the optimum window size being ~15–21 residues with a window size of 19 residues giving the best prediction overall. Based on these blind tests, this version of RONN has a sensitivity of 0.603, a specificity of 0.878, an accuracy of 0.849, a Matthews' correlation coefficient of 0.395, a CASP  $S$ -score of 9.33 and a probability excess of 0.481.

In common with other disorder prediction methods, the result of submitting a protein sequence to the RONN web server is a graph of the predicted disorder probability per residue (Fig. 2). RONN has been trained and validated to optimize performance based on a probability cut-off of 0.5. However, in some situations it is useful to consider the raw probability predictions when analyzing a sequence, particularly when looking for short regions of disorder such as loops and linkers. Using a 19-residue sliding window the final prediction for each residue is an average of (up to) 19 predictions. Thus, the sensitivity to short disordered regions (<10 residues) forming loops and linkers is somewhat compromised. A different manifestation of the same effect is that RONN has some difficulty in precisely defining the first and last residues of disordered regions, as shown in the plot for the PDB entry 1GJJ (Fig. 2). Tests with the secondary blind set (Table 1) containing proportionately more short regions showed that



**Fig. 1.** Effect of varying the cost function,  $C_d$ , and the window size on RONN predictions for the main blind test set. The plot shows specificity versus sensitivity, but is rotated anticlockwise by 45° to get a plot of probability excess, our favored performance measure, against the ratio of ordered to disordered residues in our prediction. Thus, the best prediction is highest up the plot. Predictions are marked either by open squares ( $C_d = 0.5$ ) or by closed squares ( $C_d = 0.6$ ) with labels showing the size of the window in residues.



**Fig. 2.** RONN plot of disorder probability per residue for two proteins from the main blind test set (PDB accession codes 1GJJ and 1Q1Q). The horizontal dashed lines mark the threshold for disorder prediction and the shaded regions show the disorder identified from the PDB entries.

the problem is not too severe. Nevertheless, if RONN is being used specifically to search for short regions of disorder then inspection of the probability plot may be necessary. The plot for the PDB entry 1Q1Q (Fig. 2) gives an example of this for a disordered loop from residues 96 to 102, where, although the predicted probabilities are high for the loop region, they never quite reach the threshold required for the prediction of a disorder.

The window size giving the best performance might have been affected by the average size of disordered regions in the main blind test set. This was investigated by further trials using the secondary blind test set (Table 1) that was both substantially larger and contained

**Table 2.** Raw results from the blind testing of nine disorder prediction methods against the main blind test set of 80 proteins

Method	Predicted order (TN)	Missed disorder (FN)	Missed order (FP)	Predicted disorder (TP)
RONN	26 275	1449	3634	2200
DISOPRED2 <sup>a</sup>	25 890	2104	734	1432
PONDR <sup>®</sup>	24 391	1616	5518	2033
DisEMBL (hot)	25 121	1854	4788	1795
DisEMBL (465)	29 345	2432	564	1217
FoldIndex	24 245	1867	5664	1782
PreLink	28 312	2786	1597	863
GlobPlot	24 255	2292	5654	1357
DisEMBL (coils)	12 687	947	17 222	2702

<sup>a</sup>Prediction based on 77 proteins (comprising 30 160 residues).

proportionately many more short regions of disorder. With this dataset the best results were again obtained with a cost function of 0.6 and a 19 residue window, giving a similar result overall, although the sensitivity, in particular, was somewhat lower (0.562 compared with 0.603). Thus, this combination of options performed the best overall and is the one used in our comparisons with other disorder prediction methods given below and is also used for the RONN server (<http://www.strubi.ox.ac.uk/RONN>).

### 3.3 Blind testing of disorder prediction methods

Publicly available web servers for a panel of disorder prediction methods were used to make predictions. The methods tested were RONN, PONDR<sup>®</sup>, FoldIndex, DisEMBL, GlobPlot, DISOPRED2 and PreLink. The server for DisEMBL provides three choices for prediction based on the source of the disorder data. All three were tested independently and these methods are referred to below as DisEMBL (coils) for the definitions based on coils and loops; DisEMBL (hot) for the definitions based on loops with high  $B$  factors and DisEMBL (465) for the definitions based on the residues annotated in the REMARK 465 records. Thus, nine disorder prediction methods were separately analyzed using all the 80 proteins in the main blind test dataset. (The DISOPRED2 server has a limit of 1000 residues per protein, so in this case the results were based on 77 proteins.) A list of the PDB entries that form the main blind test set and a complete, compact graphical representation of the predictions (indicating both correct and erroneous predictions) are given in the Supplementary information. The results can be conveniently summarized by the number of residues in each prediction class (Table 2).

From the raw prediction data in Table 2 it is possible to derive all the different performance measures discussed above (Table 3). The order of the algorithms in these tables is determined by the probability excess, our favored performance measure. The sensitivity and specificity are further used to create a plot of probability excess against the balance of prediction outcomes (Fig. 3).

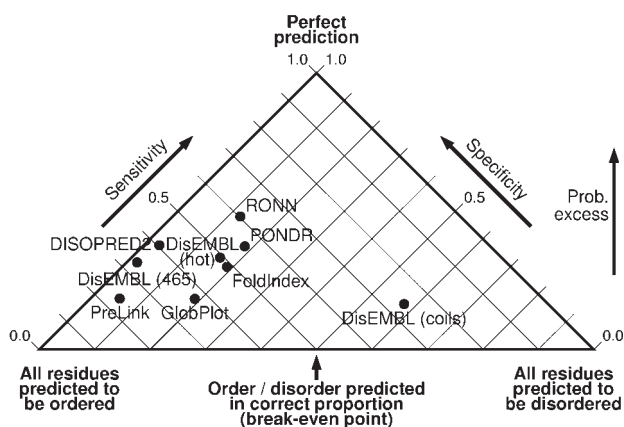
The blind tests show that the algorithms perform very differently, roughly dividing into three groups based on specificity. DisEMBL (coils) has an extremely low specificity revealing a significant overprediction of disordered residues (59% of all residues in the test set are predicted disordered, whereas only 11% are actually annotated as such), shown by its position toward the right-hand

**Table 3.** Performance measures calculated from the blind testing of nine disorder prediction methods against the main blind test set of 80 proteins

Method	Sens.	Spec.	Acc.	MCC	CASP S-score	Prob. excess
RONN	0.603	0.878	0.849	0.395	9.33	0.481
DISOPRED2 <sup>a</sup>	0.405	0.972	0.906	0.470	7.81	0.377
PONDR <sup>®</sup>	0.557	0.816	0.787	0.278	7.22	0.373
DisEMBL (hot)	0.492	0.840	0.802	0.260	6.43	0.332
DisEMBL (465)	0.334	0.981	0.911	0.437	6.10	0.315
FoldIndex	0.488	0.811	0.776	0.224	5.79	0.299
PreLink	0.237	0.947	0.869	0.219	3.55	0.183
GlobPlot	0.372	0.811	0.763	0.140	3.54	0.183
DisEMBL (coils)	0.740	0.424	0.459	0.104	3.19	0.165

The performance measures, defined and discussed in Section 3.1, are sensitivity (Sens.), specificity (Spec.), accuracy (Acc.), Matthews' correlation coefficient (MCC), CASP S-score and our favored measure, probability excess (Prob. excess).

<sup>a</sup>Prediction based on 77 proteins (comprising 30 160 residues).

**Fig. 3.** Plot comparing the outcome of the blind testing of nine different disorder prediction algorithms. The detailed form of the plot is described in the legend to Figure 1, but prediction quality increases vertically up the plot. Predictions are marked by closed circles with a label giving the method by which that prediction was produced.

side of Figure 3. This translates directly into a very low accuracy for this algorithm, although the more sophisticated measures suggest an overall performance nearly comparable to some of the other algorithms.

DISOPRED2, DisEMBL (465) and PreLink form a group of methods with very high specificities, in other words they predict relatively few ordered residues to be disordered. Given the preponderance of ordered residues in the main blind test set, these methods achieve very high accuracy and indeed they all score higher than any of the other algorithms based on this measure (Table 3). However, they achieve this at the expense of missing a significant number of disordered residues (these methods predict 7, 5 and 7%, respectively, of residues to be disordered rather than 11%; shown by their position toward the left-hand edge of Fig. 3), and in terms of sensitivity these methods perform moderately (in the case of DISOPRED2) to poorly. As discussed above, the Matthews' correlation coefficient is also strongly dependent on relative class frequencies, and based on

**Table 4.** Performance measures calculated from the blind testing of eight disorder prediction methods against the final blind test set of 159 sequences

Method	Sens.	Spec.	Acc.	MCC	CASP S-score	Prob. excess
RONN	0.675	0.888	0.789	0.580	28.02	0.563
DISOPRED2 <sup>a</sup>	—	—	—	—	—	—
PONDR <sup>®</sup>	0.632	0.782	0.712	0.420	20.60	0.414
DisEMBL (hot)	0.502	0.749	0.634	0.260	12.49	0.251
DisEMBL (465)	0.348	0.978	0.685	0.430	16.27	0.327
FoldIndex	0.722	0.815	0.771	0.540	26.68	0.536
PreLink	0.319	0.991	0.678	0.430	15.43	0.310
GlobPlot	0.308	0.821	0.582	0.151	6.42	0.129
DisEMBL (coils)	0.719	0.446	0.573	0.170	8.21	0.165

The performance measures are the same as for Table 3.

<sup>a</sup>Predictions for DISOPRED2 not done.

this measure DISOPRED2 and DisEMBL (465) are the best performing algorithms. Despite its underprediction of disorder DISOPRED2 does give some very good predictions and is ranked second overall based on probability excess (and the CASP S-score), while DisEMBL (465) is ranked fifth due to it missing such a high proportion of disordered residues.

The final group of algorithms is characterized by a better balance between order and disorder prediction (to the left of center in Fig. 3) and the ranking of performance within the group is RONN (first), PONDR<sup>®</sup>, DisEMBL (hot), FoldIndex and GlobPlot. Although the balance of these prediction algorithms appears to be better than those of the other methods studied, the error rate is still quite high: a disordered residue has a 37% chance of being predicted disordered with GlobPlot, while even with RONN the chance is only 60%. Based on either probability excess or CASP S-score, RONN is ranked first among all the nine methods examined while PONDR<sup>®</sup> is ranked third.

Our main blind test set is more than twice the size of the one used in the CASP6 trial (which comprised a total of 14 490 residues, of which 13 422 were ordered and 1068 were disordered). Furthermore, it contains a higher proportion of disordered residues, thereby reducing the distortion of some performance estimates that arises from the relative class frequencies. Thus, to our knowledge, the results presented here form the largest and most balanced blind test of disorder prediction algorithms in the literature.

Nevertheless, to demonstrate the independence of our training and testing tests, the disorder prediction methods were also evaluated against a final test set of sequences compiled from other workers in the field. Many sequences in this set correspond to fully disordered proteins that, by definition, could not be found by analysis of the PDB. Against this panel of 159 sequences, RONN again gave the best predictions (Table 4 and Supplementary information). Several methods showed significantly improved performance compared to tests with the main blind set, with rules-based methods showing the largest gains. For example, FoldIndex gave predictions second only to those of RONN. The improved performance against these test data can be ascribed to the presence of a significant number of low-complexity disordered sequences that are relatively easy to detect. We believe that our main blind test set offers a more demanding and realistic test of disorder prediction ability for the important

'real world' application of disorder prediction as an aid to rational construct design for structural analysis.

As well as providing a stern test for the different algorithms, these blind tests also highlight the inadequacies of the performance measures. Sensitivity and specificity can only properly be considered in combination. Accuracy can be very misleading where the relative class frequencies are very different, and this is also a problem for the Matthews' correlation coefficient as shown by the results for DISOPRED2 and DisEMBL (465). For a given test set, the CASP *S*-score is directly proportional to the probability excess and the proportionality constant is  $200 f(1 - f)$ , where *f* is the fraction of disordered residues. This constant is the limiting value of this score and so the CASP *S*-score for a perfect prediction is determined by the proportion of disordered residues in the test set.

#### 4 CONCLUDING REMARKS

The demand for accurate prediction of natively disordered regions in proteins is being driven by structural genomics initiatives worldwide in an effort to increase the success rate, particularly for studies on eukaryotic proteins. Incorporating accurate disorder prediction into the construct design process can reduce the amount of time and resources devoted to non-crystallizable or otherwise badly behaved proteins (Oldfield *et al.*, 2005), allowing researchers in the wet lab to tackle the real technological challenges posed by high-throughput studies. In the current state of the art, no single disorder prediction technique is so accurate that it can be entirely trusted, and a strategy of using several well-performing methods and looking for common prediction features may be the best way to attempt the reliable identification of regions of disorder in unknown protein sequences.

The RONN algorithm presented here is an application of our revised BBFNN technique to the prediction of natively disordered regions. We have also described the collation of large datasets of protein disorder from the PDB data and their use for training RONN and for blind testing a panel of freely available disorder prediction techniques. These blind tests show that, based on either probability excess or the CASP *S*-score, RONN is ranked as the best algorithm for disorder prediction with DISOPRED2 and PONDR<sup>®</sup> also performing well (Table 3 and Fig. 3).

Nevertheless, RONN still has weaknesses, particularly in the detection of short regions of disorder and in defining the first and last residues of disordered regions (Fig. 2). The use of a shorter window improves the sensitivity, but also increases the noisiness of the prediction and the overall performance is degraded (Fig. 1). A more complex scheme might involve the use of two (or more) window sizes, effectively making multiple predictions and then combining them to return a single overall prediction for the sequence. While such algorithms are necessarily more complex, they could also potentially cope with any dependence of the amino acid composition of disordered regions on their lengths. The potential of such algorithms is currently being investigated.

#### ACKNOWLEDGEMENTS

PONDR<sup>®</sup> is copyright by the WSU Research Foundation. Access to PONDR<sup>®</sup> was provided by Molecular Kinetics (IUETC, 351 West 10th Street, Suite 318, Indianapolis, IN 46202, USA; email: main@molecularkinetics.com). The OPPF is funded by the UK Medical Research Council with additional support from the European Commission Integrated Programme SPINE, contract

number QL2-CT-2002-00988. The work of P.M. was funded by the MRC eFamily project.

*Conflict of Interest:* none declared.

#### REFERENCES

- Alber, T. *et al.* (1983) The role of mobility in the substrate binding and catalytic machinery of enzymes. *Ciba Found. Symp.*, **93**, 4–24.
- Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Altschul, S.F. *et al.* (1994) Issues in searching molecular sequence databases. *Nat. Genet.*, **6**, 119–129.
- Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
- Bartlett, G.J. *et al.* (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.
- Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Boutselakis, H. *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, **31**, 458–462.
- Burley, S.K. and Petsko, G.A. (1985) Aromatic–aromatic interaction: a mechanism of protein structure stabilization. *Science*, **229**, 23–28.
- Coeytaux, K. and Poupon, A. (2005) Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics*, **21**, 1891–1900.
- Dasarathy, B.V. (1991) *Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.
- Dasarathy, B.V. (1994) Minimal consistent set (MCS) identification for optimal nearest neighbour decision systems design. *IEEE Trans. Systems Man, Cybernetics*, **24**, 511–517.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In Dayhoff, M.O. (ed.), *Atlas of Protein Sequence and Structure Volume 5*. National Biomedical Research Foundation, Washington DC, pp. 345–358.
- Duda, R.O., Hart, P.E. and Stork, D.G. (2002) *Pattern Classification*, 2nd edn. Wiley-Interscience, Hoboken, NJ.
- Dunker, A.K. *et al.* (2000) Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 161–171.
- Frankel, A.D. and Kim, P.S. (1991) Modular structure of transcription factors: implications for gene regulation. *Cell*, **65**, 717–719.
- Garner, E. *et al.* (1998) Predicting disordered regions for amino acid sequence: common themes despite differing structural characterization. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 201–213.
- Garner, E. *et al.* (1999) Predicting binding regions within disordered proteins. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 41–50.
- Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
- Huber, R. (1979) Conformational flexibility in protein molecules. *Nature*, **280**, 538–539.
- Kissinger, C.R. *et al.* (1995) Crystal structures of human calcineurin and the human FKBP12–FK506–calcineurin complex. *Nature*, **378**, 641–644.
- Li, W. *et al.* (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
- Li, W. *et al.* (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
- Li, X. *et al.* (1999) Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 30–40.
- Li, X. *et al.* (2000) Comparing predictors of disordered protein. *Genome Inform. Ser. Workshop Genome Inform.*, **11**, 172–184.
- Linding, R. *et al.* (2003a) Protein disorder prediction: implications for structural proteomics. *Structure (Camb.)*, **11**, 1453–1459.
- Linding, R. *et al.* (2003b) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
- Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
- Oldfield, C.J. *et al.* (2005) Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins*, **59**, 444–453.
- Radivojac, P. *et al.* (2003) Prediction of boundaries between intrinsically ordered and disordered protein regions. *Pac. Symp. Biocomput.*, 216–227.
- Radivojac, P. *et al.* (2004) Protein flexibility and intrinsic disorder. *Protein Sci.*, **13**, 71–80.
- Romero, P. *et al.* (1997) Identifying disordered regions in proteins from amino acid sequence. *Proc. IEEE Int. Conf. Neural Networks*, **1**, 90–95.

- Romero,P. et al. (2000) Intelligent data analysis for protein disorder prediction. *Artif. Intell. Rev.*, **14**, 447–484.
- Schulz,G.E. (1979) Nucleotide binding proteins. In Balaban,M. (ed.), *Molecular Mechanism of Biological Recognition*. Elsevier/North-Holland Biomedical Press, Amsterdam, pp. 79–94.
- Thomson,R. and Esnouf,R. (2004) Prediction of natively disordered regions in proteins using a bio-basis function neural network. *Lecture Notes in Computer Science*, **3177**, 108–116.
- Thomson,R. et al. (2003) Characterizing proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics*, **19**, 1741–1747.
- Uversky,V.N. et al. (2000) Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins*, **41**, 415–427.
- Vucetic,S. et al. (2001) Methods for improving protein disorder prediction. *Int. Joint INNS-IEEE Conf. Neural Networks*, **4**, 2718–2723.
- Ward,J.J. et al. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.
- Weinreb,P.H. et al. (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. *Biochemistry*, **35**, 13709–13715.
- Xie,Q. et al. (1998) The sequence attribute method for determining relationships between sequence and protein disorder. *Genome Inform. Ser. Workshop Genome Inform.*, **9**, 193–200.
- Yang,Z.R. and Thomson,R. (2005) Bio-basis function neural network for prediction of protease cleavage sites in proteins. *IEEE Trans. Neural Networks*, **16**, 263–274.