Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization

Ethan Garner [1] egarner@wsunix.wsu.edu
Paul Cannon [2] pcannon@eecs.wsu.edu
Pedro Romero [2] promero@eecs.wsu.edu
Zoran Obradovic [2] zoran@eecs.wsu.edu
A. Keith Dunker [3] dunker@mail.wsu.edu

[1] Department of Biochemistry and Biophysics, Washington State University, Pullman, WA 99164-4660
[2] School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-4660
[3] Author to which all correspondence should be addressed.
Department of Biochemistry and Biophysics, Washington State University, Pullman, WA 99164-4660
Telephone: 509 335-5322, Fax: 509 335-9688

Abstract

Using ordered and disordered regions identified either by X-ray crystallography or by NMR spectroscopy, we trained neural networks to predict order and disorder from amino acid sequence. Although the NMR-based predictor initially appeared to be much better than the one based on the X-ray data, both predictors yielded similar overall accuracies when tested on each other's training sets, and indicated similar regions of disorder upon each sequence. The predictors trained with X-ray data showed similar results for a 5-cross validation experiment and for the out-of-sample predictions on the NMR characterized data. In contrast, the predictor trained with NMR data gave substantially worse accuracies on the out-of-sample X-ray data as compared to the accuracies displayed by the 5-cross validation during the network training. Overall, the results from the two predictors suggest that disordered regions comprise a sequence-dependant category distinct from that of ordered protein structure.

1 Introduction

Many regions of proteins and some whole proteins form ensembles of structures under native conditions, in essence lacking a fixed tertiary structure within a given functional domain. Such "disordered" (or "unfolded") proteins have been identified by several methods: 1. sensitivity to proteases; 2. missing electron density in structure determinations by X-ray diffraction; 3. NMR spectroscopy; and 4. CD spectroscopy coupled with other methods such as rapid protease digestion, gel exclusion chromatography, or survival of function following incubation at high temperatures.

Disordered regions characterized by the methods described above are often essential for function. Such regions have therefore been called 'natively unfolded' [51], or 'natively disordered' [20]. 'Unfolded' implies that the region of protein exists in an extended, flexible (random-coil-like) form, whereas 'disordered' includes not only these extended forms but in addition can also imply a collapsed, partially folded with secondary structure, but non-rigid (molten-globule-like) form. The disordered ensemble of structures can involve equilibria between random-coil-like and molten-globule-like forms.

Since amino acid sequence determines protein structure [2], we proposed that amino acid sequence should determine lack of tertiary structure or disorder as well [43]. To test this hypothesis, we identified disordered sequences from missing coordinates in Protein Data Bank (PDB) files and then developed and tested a collection of neural network predictors of order and disorder [44, 43, 45]. Following our initial studies on disordered regions of various lengths, our focus turned to long disordered regions (LDRs), where long is somewhat arbitrarily defined as > 40 contiguous amino [45]. Such long disordered regions have lower false positive error rates and are unlikely to be completely missed, even for a modestly accurate predictor. The first generation neural network predictor (NNP) of such long regions, called the LDR NNP but herein renamed the X-RAY NNP, was shown to predict order and disorder on a residue-by-residue basis with 73% accuracy as estimated by 5-cross validation. The false positive error rate for contiguous predictions of 40 or longer was found to be less than 7% on a sequence basis [43], which corresponds to less than 1 false positive prediction of disorder equal to or longer than 40 for every 1,000 amino acids [45].

Missing electron density in an X-ray-solved protein structure can result from experimental limitations or can be the result of structural disorder. Structural disorder can be static or dynamic [7, 25], which considerably weakens confidence in predictions based on this data. Thus, it would be useful to obtain data from other methods or at least to use X-ray data verified by other methods.

Several proteins with disordered regions have been characterized by NMR spectroscopy. Since this method does not suffer from an uncertainty regarding the type of disorder, application of our X-RAY NNP to several LDRs characterized by NMR provides a means to stringently test the validity of our predictor. Also, further validation of the X-RAY NNP training proteins could be accomplished by testing them with a predictor trained on NMR-characterized proteins, so a second predictor, herein called the NMR NNP, was developed. The two neural networks were then tested upon each other's training sets. Although the accuracies of the two predictors upon each other's set of disordered regions showed large variations, further study suggests that the variable accuracies are the result of different types of disordered regions. Overall, the results presented herein suggest that disordered or unfolded regions of sequence form a distinct category compared to ordered or folded regions of sequence.

2 Materials and Methods

2.1 The Proteins

In our previous work [43] disordered regions were identified from Protein Data Bank (PDB) [1] files as amino acids that were missing from the set of atomic coordinates. That is, disorder leads to incoherent X-ray scattering and subsequent absence of electron density in the solved structure [7, 30, 47]. These proteins containing X-ray-characterized regions of disorder, their PDB filenames, and their Swiss Protein [6] identifications (SW IDs), respectively were; 1. Tomato Busy Stunt Virus, 2tbv, COAT_TBSVB, 2. Tyrosyl tRNA synthetase, 2ts1, SYY_BACST, 3. Calcineurin, 1aui, P2BA_HUMAN, 4. Topoisomerase II, 1bgw, TOP2_YEAST, 5. Elongation factor G, 1elo, EFG_THETH, 6. Apoptosis regulator BCL-Xl, 1bcl, BCPA_PROAE, 7. Intact lactose operon repressor, 1lbh, LACI_ECOLI. The total number of disordered amino acids in this database is 449 aa.

NMR proteins with identified regions of disorder were identified by tedious literature searches. Several different NMR parameters identify regions of disorder [41, 54, 19, 55]. The proteins and their SW (or PIR) IDs were 1. 4e binding protein 1, s50866 (PIR), 2. Murine Prion, PRIO_MOUSE, 3. Histone H5, H5_CHICK, 4. Flagellum specific sigma factor (FlgM), FLGM_SALTY 5. Antitermination protein of bacteriophage λ (AT), REGN_LAMBD, 6. N term activator domain of Heat Shock Transcription Factor (HSTF), HSF_KLULA, 7.High Mobility Group-I (HMG-I), HMGI_HUMAN. The total number of disordered regions in this database is 677 aa.

In addition, we collected a similar number of structured control proteins. These proteins were selected to be of similar overall size as the X-ray and NMR proteins, to be monomeric, and to be without cofactors. These proteins, their PDB filenames and their SW Ids were, respectively; 1. Hen egg-white lysozyme; 1hel, LYC_CHICK, 2. Ribonuclease A (Rnase A), 3rn3, RNP_BOVIN, 3. β-cryptogein (B-cryp), 2ctb, CBPA_BOVIN, 4. Elastase, 1lvy, EL1_PIG, 5. Profilin A (Pfln A), 1acf, PRO1_ACACA, 6. Haloalkane Dehalogenase (HDHase), 2edc, HALO_XANAU, 7. Azurin II (Az II), 1arn, AZU2_ALCXX, and 8. Carboxypepitidase A (CbPA), 2ctb, ELIB_PHYCR.

2.2 Feature Selection

We use the term 'attribute' to mean a value calculated over a specifies window and the term 'feature' for those attributes that are subsequently used to train the neural networks. Sequence attributes are numerical values calculated from an amino acid sequence over a specified window [4]. For these studies, 24 attributes provided the initial pool, the first 20 of which are the compositions of the 20 amino acids within the specified sequence windows. The last 4 are hydropathy [33], flexibility index [50], helix amphipathic moment [21] and sheet amphipathic moment [22].

The NMR and X-ray disordered datasets were each matched with an equal number of ordered amino acids taken from the NRL_3 [40], which is a subset of PDB containing only ordered structures. A feed-forward search with minimal error probability selection criterion was used on the balanced ordered and disordered NMR and X-ray datasets [43]. A quadratic Gaussian classifier using different covariance matrices for each class was used to calculate the minimal error probability during each of the searches. Experimentation with other dimensionality reduction methods, such as sequential backward search and branch-and-bound, yielded results quite similar to those presented here. Ten features were selected from the original pool of 24 attributes.

2.3 Neural Network Training

Several possible neural network architectures were investigated in the initial phase of these studies. A simple network with 10 inputs, 7 fully connected nodes in a single hidden layer, and one output was selected as being commensurate with the dataset size and as giving good results [43].

The X-ray and NMR disordered datasets, with their number-balanced datasets of ordered sequences, were scrambled in order to separate values from adjacent sequence positions, and then divided into 5 disjoint subsets by random selection. Experimentation indicated that similar prediction accuracies were achieved during training whether or not scrambling was used, but scrambling may serve to improve predictions for completely unrelated proteins.

For each training cycle, 4/5 of the data comprised the training set and 1/5 the test set. The training set was further separated into a proper training set (80%) and a validation set (20%). Three initializations were used and the number of epochs for each training was chosen as that which produced the highest accuracy on the validation set. Once training was investigated by 5-cross validation, the data were recombined and training was repeated using 5/5 of the data.

3 Results

3.1 Selected Features

The ten features selected on the basis of distinguishing order and disorder for the NMR and X-ray datasets are shown in Table 1. Six of the ten features were the same for both datasets: flexibility index, hydropathy, and mole fractions of Y, W, C, and S. These data indicate that the NMR-and-X-ray-characterized regions of disorder share important characteristics.

With regard to the selected features that were different for the two datasets, the compositions of H, D, K, and E distinguished ordered and disorder for the X-ray dataset, whereas the compositions of F, G, R, and P were useful for the NMR dataset. Thus, the features selected for the NMR-characterized regions of disorder show important differences from those selected for the X-ray-characterized regions of disorder.

X-RAY	H	D	K	E	S	C	W	Y	Hydropathy	Flexibility
NMR	F	G	R	P	S	C	W	Y	Hydropathy	Flexibility

Table 1: **Selected features.**

X-RAY NNP	NMR NNP
70%	86%
77%	84%
77%	89%
72%	86%
70%	89%
Average 73% ± 2%	Average 87% ± 4%

Table 2: **Five Cross Validation Results.**

3.2 Five Cross Validation

The evaluation of the training of the X-RAY NNP was described previously [43]. Here those data are compared with the results of a similar training exercise for the NMR NNP (Table 2.). Overall, the NMR NNP gives a significantly higher accuracy compared to the X-RAY NNP during the training exercises, e.g. 87%± 4% compared to 73%±2%.

3.3 Example Predictions

Example predictions are shown in Fig. 1. The X axis is the residue number while the Y axis is the prediction output. Anything above an output of 0.5 is considered a prediction of disorder. The solid horizontal line at the center of the graph indicates what regions are actually disordered. Fig. 1A and Fi. 1B are predictions on disordered proteins, while 1C and 1D are predictions upon the ordered control proteins. One of the best overall predictions (1A) and one of the worst (1B) on regions of disorder as well as the two worst overall predictions (1C, 1D) on the control proteins are provided. In (1A), the prion protein from the NMR dataset was subjected to analysis using both the NMR NNP and the X-RAY NNP. Notice how the X-ray prediction accuracy is relatively similar to that of the NMR predictor, which was trained on this protein's data (X-RAY NNP = 88.4% correct overall; NMR NNP = 97% correct overall). In (1B), the anti-termination (AT) protein from bacteriophage lambda from the NMR dataset was predicted upon by both predictors. Here, the accuracy of the X-RAY NNP (53.2%) seems very poor, but notice how its prediction is again somewhat similar to that of the NMR NNP, which, despite having this protein in its training set, still manages an accuracy of only 73%, much lower than that obtained on the prion protein in the previous example. Finally, in (1C) and (1D) the predictions on the ordered control proteins, profilin A and haloalkane dehalogenase, are presented, respectively. The false positive predictions of disorder are seen to be very short, especially for the NMR NNP.

For long disordered regions (LDRs) such as that for the AT protein, even modestly successful prediction rates (e.g. just 53.2% for the X-RAY NNP) still give an indication of protein disorder. For this reason, we are initially focussing our attention on proteins with such LDRs.

Fig. 1 also indicates several types of errors. Relative to prediction of disorder, false positive predictions are ordered regions incorrectly predicted to be disordered (peak labeled b in 1A) whereas false negative predictions are disordered regions predicted to be ordered (region a in 1A, regions c, d, and e in 1B and so on for 1C and 1D). Another useful classification is whether an errant prediction is false (for example peak a in 1A) throughout (e.g. a non-boundary error, for example peak b in 1A) or is correct over some region but then becomes false upon crossing an order/disorder junction in the

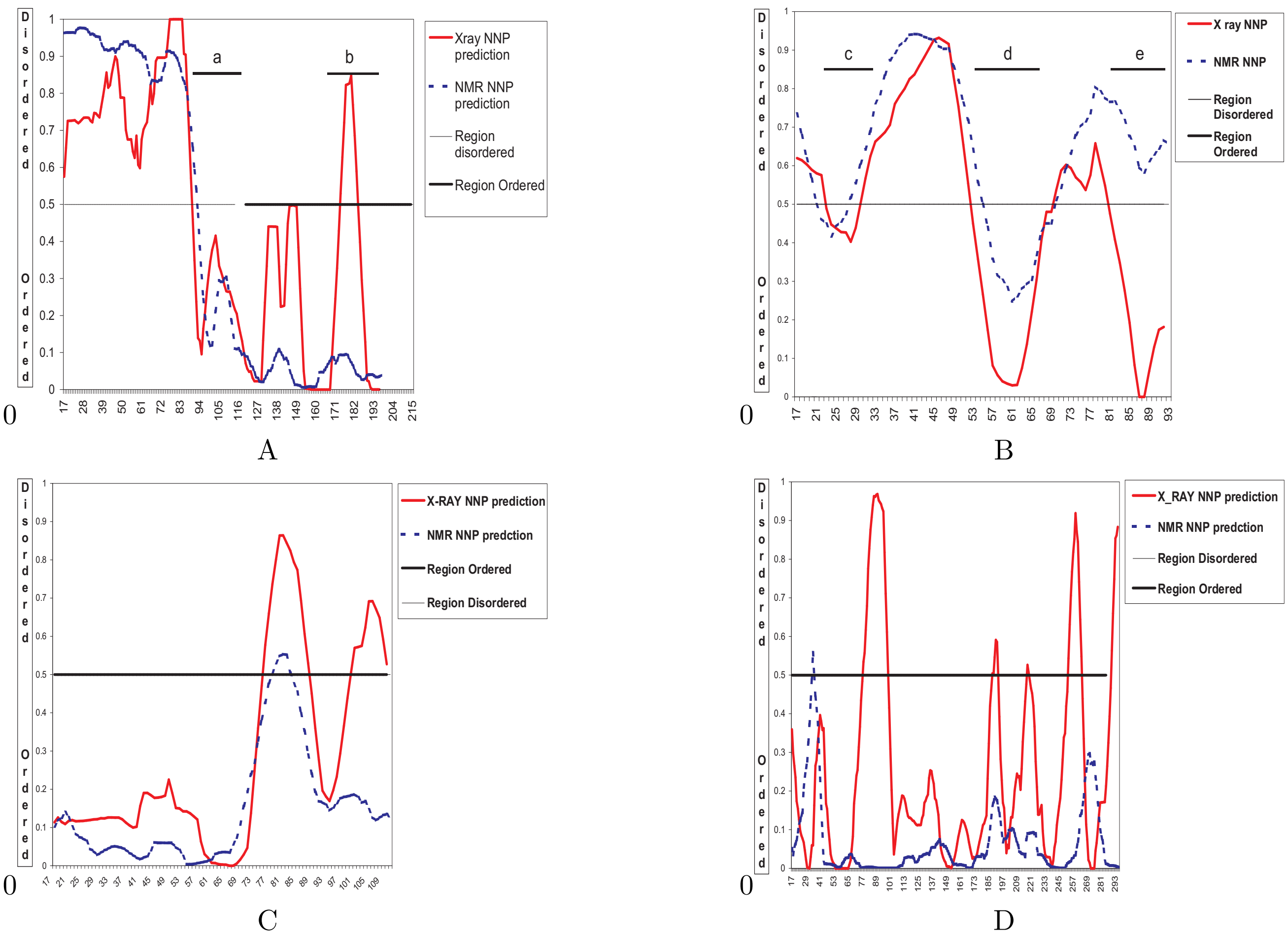

Figure 1: **Example Predictions using the X-RAY NNP and NMR NNP**.– Example predictions using the X-RAY and NMR NNPs. Both predictors were applied to the following proteins: murine prion (A); bacteriophage λ antitermination protein (B), profilin A (C) and haloalkane dehalogenase (D). The first two contain regions of disorder and the last two are ordered, control proteins. The X-axis is the residue number, while the Y axis is the prediction output. Values above 0.5 indicate disorder, below 0.5 indicate order. The solid line at 0.5 indicates an identified region of order, a dashed line a region of disorder. Various types of errors are marked and indicated by letters a, b, c, etc. (see text).

structure (e.g. a boundary error).

3.4 Prediction Accuracies

The X-RAY NNP was applied to the NMR-characterized proteins and the NMR NNP was applied to the X-ray-characterized proteins. The results of these out-of-sample predictions are presented in Table 3.

For the X-RAY NNP, the overall prediction accuracies range from 53.2% (AT) to 93.5% (HMGI(Y). The large range of error rates undoubtedly relates to a variation in the degree of similarity of the disordered regions in the different proteins to the disordered regions used to train the X-RAY NNP. For example, unlike most of the NMR proteins, the HMGI(Y) has local charge imbalance, thus having charge attributes commensurate with those of the X-ray training set and giving a very high overall prediction accuracy by the X-RAY NNP.

For the NMR NNP, the overall prediction accuracies range from 55.1% (tyrosyl-tRNA synthetase) to 94.1% (Bcl-xL). Again, the low prediction accuracy on the synthetase signals a difference in the

X-RAY NNP on NMR-Characterized Data

Protein	Length of sequence	Prediction of Disordered Regions (DR) from amino acid sequence: Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization	Ref.
Antitermination protein of bacteriophage λ (ATa)	1-107	1-107	15-93	8, 23, 11	53.2%	46.8%	N/A	A, B, D	[36]
Histone H5 (H5)	1-189	1-21, 101-185	15-175	17, 16, 99	68.9%	0%	63.3%	A, C, D	[5]
Flagellum specific sigma factor (FlgM)	1-97	1-97	15-83	24, 37	88.4%	11.6%	N/A	A, D	[16]
N term activator domain of Heat Shock Transcription Factor (NAD-HSTF)	1-195	1-195	15-195	3, 58, 14, 31, 9	63.5%	36.5%	N/A	A, D	[14]
High Mobility Group-I (HMG-I (Y))	1-106	1-106	15-92	73	93.5%	6.4%	N/A	A, C, D	[31]
4e Binding Protein 1 (4e BP-1)	1-118	1-118	15-104	2, 15, 37	57.8%	42.2%	N/A	A, C	[26]
Murine Prion	1-254	23-120	15-240	73, 10	88.4%	22.5%	7.8%	A, D	[41]

NMR NNP on X-RAY-Characterized Data

Protein	Length of sequence	Prediction of Disordered Regions (DR) from amino acid sequence: Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization	Ref.
Tomato Busy Stunt Virus	1-387	1-66	15-373	65, 17	73.9%	50%	17.3%	B, D	[29]
Ligase (tyrosyl tRNA synthetase)	1-206	109-206	15-192	18, 21	55.1%	74.6%	18.9%	B	[11]
Calcineurin	1-179	74-168	15-165	24	55.6%	73.6%	0%	B	[32]
Topoisomerase II	1-400	225-273	15-385	9, 4, 8	85.8%	83%	4%	B, D	[8]
Elongation factor G	1-341	50-125	14-327	39, 20, 6, 14, 14	75.1%	36.8%	21.1%	B	[23]
Apoptosis regulator BCL-X	1-196	1, 31-80	15-182	50	94.1%	11.5%	3.4%	B, D, A	[46]
Intact lactose operon repressor	1-360	1-61	15-346	11, 10, 17, 39	69.3%	76.5	23.1	B, D, A	[34]

Structural Characterization
A= NMR, B=X-ray diffraction, C= CD, D= Protease hypersensitivity.

Table 3: **Cross Prediction Results.**

characteristics of its disordered regions compared to those in the NMR dataset. The details of this difference await further study. On the other hand, the disordered region in Bcl-xL must be more similar to those in the NMR training set. It is likely to be coincidental that the structure of Bcl-xL has also been determined by NMR [37].

Application of the NMR and X-RAY NNPs to the control proteins was carried out. The results on a protein-by-protein basis are shown in Table 4. The error rate ranges from a low of 72.7% (X-RAY NNP Haloalkane Dehalogenase) on to a high of 100% (many proteins).

Finally, the overall prediction accuracies are summarized in Table 5. The NMR and X-RAY NNPs give similar overall prediction rates near 74% on each other's training sets. The predictions on the fully ordered control proteins are considerably better, about 84% for X-RAY NNP and 98% for the NMR NNP. The high accuracy on the fully ordered proteins implies that the ordered part of our training sets is providing our predictors with information that allows for better generalization than that achievable from our disordered data.

4 Discussion

4.1 X-ray- and NMR-Characterized Regions of Protein Disorder

Our pilot studies indicated a definite relationship between amino sequence and the presence of ordered or disordered structure [43, 44, 45, 20] However, these initial studies had two, interrelated, acknowledged limitations related to their exploratory nature. First, the number of disordered examples was very small, just 449 amino acids in the original LDR set. Second, all the disordered examples were from PDB, which has considerable ambiguity with regard to the characterization of disordered proteins. Fortunately, the signals for the tendency for disorder seem to be so strong that these two limitations apparently haven't led to large errors. Work in progress with a 6-fold larger database, now about 2,500 disordered amino acids, yield results very similar to those of the pilot studies (manuscript in preparation).

The small number of examples of disordered proteins in our original studies resulted from the

X-RAY NNP on Control Proteins

		Prediction of Disordered Regions (DR) from amino acid sequence							
Protein	Length of sequence	Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization	Ref.
Hen egg-white lysozyme	1-129	None	15-115	4	96%	N/A	4%	B	[53]
Ribonuclease A (Rnase A)	1-124	None	15-110	3, 3	93.8%	N/A	6.2%	B	[10]
β-cryptogein (B-cryp)	1-98	None	15-84	None	100%	N/A	0%	B	[9]
Elastase	1-240	None	15-226	14, 7,4	87.4%	N/A	12.6%	B	[35]
Profilin A (Pfln A)	1-125	None	15-111	12, 11	75.3%	N/A	24.7%	B	[24]
Haloalkane Dehalogenase (HDHase)	1-310	None	15-297	22, 6, 2, 13, 7	72.7%	N/A	17.3%	B	[27]
Azurin II (Az II)	1-129	None	15-115	7, 10	83.2%	N/A	16.8%	B	[18]
Carboxypepitidase A (CbPA)	1-307	None	15-293	3, 9, 6, 9, 3, 2, 6, 21, 16	73.2%	N/A	26.8%	B	(not pub.)

NMR NNP on Control Proteins

Protein	Length of sequence	Known disorder	Region predicted	Predicted DR lengths	Percent correct	False negative	False positive	Structural characterization
Hen egg-white lysozyme	1-129	None	15-115	None	100%	N/A	0%	B
Ribonuclease A (Rnase A)	1-124	None	15-110	6	93.7%	N/A	6.3%	B
β-cryptogein (B-cryp)	1-98	None	15-84	None	100%	N/A	0%	B
Elastase	1-240	None	15-226	6, 1, 3	96.4%	N/A	3.6%	B
Profilin A (Pfln A)	1-125	None	15-111	6	93.8%	N/A	6.2%	B
Haloalkane Dehalogenase (HDHase)	1-310	None	15-297	1	99.6%	N/A	0.4%	B
Azurin II (Az II)	1-129	None	15-115	None	100%	N/A	0%	B
Carboxypepitidase A (CbPA)	1-307	None	15-293	0	100%	N/A	0%	B

Structural Characterization
A= NMR, B=X-ray diffraction, C= CD, D= Protease hypersensitivity.

Table 4: **Prediction on Controls.**

Overall Accuracy

Set	Total aa predicted on	Total ordered aa predicted on	Total disordered aa predicted on	Total false negative aa	Total false positive aa	Percent false negative	Percent false positive	Percent overall correct
X-RAY on NMR	884	207	677	179	60	26.4%	28.9%	72.9%
NMR on X-RAY	1873	1424	449	204	265	59.0%	14.3%	75.0%
X-RAY on Control	1238	1238	0	N/A	201	N/A	16.2%	83.8%
NMR on Control	1238	1238	0	N/A	29	N/A	2.4%	97.6%

Table 5: **Summary Tables.**

lack of organized data on non-folding amino acid sequences. PDB is the largest organized source of information about proteins with regions of disorder, but as our initial studies clearly demonstrate, PDB is strongly biased against the presence of disorder [44], so even this source does not have very many examples.

In addition to being few in number, X-ray-characterized regions of structural disorder have alternate possible causes for the observed missing electron density, including the following possibilities; 1. A locally *structured* domain could be moving; 2. a locally *structured* domain could be occupying several alternative positions; 3. a local region of sequence could comprise an ensemble of interconverting shapes; and 4. A local region of sequence could comprise an ensemble of static shapes. From our perspective, the important distinction is whether a region of sequence folds into a single structure (e.g. either 1 or 2) or comprises an ensemble of structures (e.g. 3 or 4). The distinctions between 1 and 2 and the distinctions between 3 and 4 are less important; indeed, 1 versus 2 and the 3 versus 4 distinctions are a continuum that depends only on the timescale. Here, we refer to locally structured domains that are disordered by movement or by occupancy of different positions (e.g. either 1 or 2) as 'domain wobble.'

Others have used 'dynamic,' 'static,' 'hinged' and 'flexible' to describe the various possible causes of structural disorder that leads to missing electron density in X-ray crystal structures [7, 25] but these previous terms do not correspond in any precise way to the 4 possibilities listed above. It is for this reason that we are proposing the terms 'domain wobble' and 'ensembles of structures' to contrast the distinctions that we believe are important to this work. In our initial studies we attempted to eliminate wobbly-domains by literature studies on each protein. However, not only does PDB lack clear information regarding disorder, but attempts to find information from the original literature often fail because the needed experiments have simply not been done or have not been reported. We hope that, as the importance of disordered protein becomes more generally realized, the key information about such disordered regions will become more readily available.

Given the above, extending the studies of disorder to include regions characterized by NMR is important for overcoming both limitations of our initial studies. NMR-characterized disordered regions include information about the extent of folding of the disordered region and at the same time increase the number of examples.

Unfortunately, there are relatively few examples of NMR-characterized regions of disorder, and these are scattered in the literature and not collected at one location. So, just as for development of a disordered regions database using PDB, an intensive effort is required for each new entry characterized by NMR. The amount of effort required will continue to slow the rate of enlargement of our database of disordered protein. Nevertheless, the amount of disordered data in this paper has more than doubled the data compared to that of the pilot studies.

4.2 Feature Selection

Our initial work [45, 44, 43] emphasized the use of sequence attributes based on amino acid composition. We reasoned that LDRs could be considered to be a new "structural class," and amino acid compositions had been shown to be successful for protein class prediction [38]. We are aware that considerations of coupling effects among different amino acids has led to much improved prediction of protein class [15], and we would like to apply such approaches to disorder prediction. However, consideration of amino acid pair frequencies requires much more data than we currently have, so such approaches are simply out of reach at the present time.

The feature selection experiments in the development of the X-RAY and NMR NNPs (Table 1.) suggest substantial similarities for the disordered regions characterized by these two methods. While 6 out of 10 of the features are identical to each other, the remaining 4 contribute enough information to cause the differences noticed between the predictors. The fact that the NMR NNP has both a higher false negative rate and a lower false positive rate than the X-RAY NNP suggests that the NMR NNP has a higher threshold to which it ascribes its disordered features. This may be due to the fact that the NMR NNP's training set contains more extreme values for the attributes specific to disorder within its training set (see Fig. 1), values not found as frequently within the X-ray data set correlating with disorder/order predictions .

We have developed a substantially larger database, having approximately 2,500 disordered amino acids in windows of 21 matched with an equal number of ordered amino acids in windows of the same size. Studies in progress on this larger database indicate that charge imbalance, when it exists, is a very strong determinant of local disorder (manuscript in preparation). The selection of H, D, K and E for the X-ray dataset is the result of substantial charge imbalance in several of the disordered regions in the X-ray-characterized proteins. In contrast, charge imbalance is not so important for the current NMR dataset.

Flexibility index, hydropathy and the mole fraction of S were found to be relatively higher in disordered regions as compared to ordered regions for both the NMR and X-ray data in the enlarged dataset, which is in complete agreement with the pilot studies [44, 42, 43, 45]. More flexible, more polar regions are more likely to be disordered. Not only does S promote disorder by its polarity, but it is special owing to its generally high abundance coupled with its ability to stabilize multiple local backbone conformations by side-chain-backbone hydrogen bonding [48].

On the other hand, Y, W, and C were lower in the disordered regions as compared to the ordered regions, both in the new enlarged database and in the data used for pilot studies. In several different datasets of ordered and disordered regions, W, Y, and C have always been found to be lower in disordered regions: indeed, these three appear to be the most order-forming of the natural amino acids (manuscript in preparation). The order-forming tendencies of W and Y may be related to the extra stability arising from aromatic/aromatic interactions [12], while the ability to form disulfide bonds is an obvious reason for the order-forming potential of C. Interestingly, W, Y, and C also evidently have the highest tendency to be conserved as judged, for example, by the various values in the PAM 250 matrix [17].

With regard to the features selected for the NMR dataset, the studies in progress on the larger dataset (manuscript in preparation) suggest that disorder is found to be associated with high mole fractions of R and P and with low fractions for F. R is typically an uncommon amino acid, but when there are high local concentrations R, it likely induces disorder by charge imbalance. High levels of proline prevent compact folding; indeed, proline-rich regions are common in proteins and seem to have function associated with their ill-folded conformations [52, 3, 39]. Higher local concentrations of F probably encourage order for the same reasons as Y and W - due to extra stability from aromatic/aromatic interactions [12].

In the study on the larger dataset mentioned above, the mole fraction of G when considered alone is found to be essentially uncorrelated with either order or disorder, so it is unclear why this amino acid was selected for the NMR dataset. On the other hand, in the development of flexibility index, G was found to change markedly, showing high flexibility index values when next to flexible neighbors, but showing low flexibility values when next to less flexible neighbors [50]. Thus, the selection of G may reside in its behavior in conjection with neighboring amino acids in the ordered and disordered sequences. This is consistent with the way feature selection works, trying to select not only the best features but also the best combinations of features. The mole fraction of G may not be correlated with order/disorder, but its combination with other variables improves the discriminatory power of the predictor.

Since the NMR NNP has a lower rate of false positives than the X-RAY NNP we plan to be use it to complement other NNP's by helping weed out false positive predictions. Future predictors based upon a larger training set of diverse proteins may yield better results, as characteristics of disorder from differing families of proteins are incorporated into the predictors. A new predictor is currently under development that combines both of these training sets, and selects from a greater number of features (Unpublished). Preliminary results show greater accuracy as judged by 5-cross validation, suggesting that enlarging the data set can lead to greater prediction precision as more features indicative of disorder are included. Also, a larger data set can provide our predictors with enough disorder data to allow for better generalization.

4.3 Out-of-Sample Predictions

The X-RAY NNP and the NMR NNP give similar overall prediction accuracies on the proteins used in each other's training sets: 72.9% for the X-RAY NNP on the NMR data and 75% for the NMR NNP on the X-ray data. However, significant differences become evident when the types of errors are considered.

The X-RAY NNP on the NMR data exhibits similar false positive (28.9%) and false negative (26.4%) rates, whereas the NMR NNP on the X-ray data exhibits large false negative rates (59%) and small (14.3%) false positive rates. Additional insight follows from noting that the more or less uniform performance of the X-RAY NNP on the NMR data with a overall accuracy of 72.9% closely matches the 5-cross validation results (73%), whereas the much more variable performance of the NMR NNP is associated with a large drop-off in the out-of-sample predictions (about 75%) as compared to the 5-cross validation results (87%). Overall, these data suggest that the NMR NNP is much more specific for the disordered regions characterized by NMR, whereas the X-RAY NNP appears to be more general.

One possibility for the poor performance of the NMR NNP on the X-ray-characterized disordered regions that these disordered regions are misclassified, e.g. they are actually wobbly domains.

For example, a region of missing electron density in a tyrosyl t-RNA synthetase different from the one in the present studies was later shown to have a considerable part that is ordered [28]. This observation on another t-RNA synthetase coupled with the high false negative error rate (74.6%) for the

NMR NNP (Table 3.) could be an indication that we have misclassified the disordered region identified by X-ray diffraction in this protein. On the other hand, the NMR NNP's worse false negative prediction (e.g., 83% for topoisomerase II) would seem to indicate that this disordered region surely must be misclassified. However, the disordered region of this protein has been well-characterized to lack ordered structure: the putatively disordered region is extremely rich in charged residues and hypersensitive to protease digestion [13, 49]. Indeed, most of the disordered regions of the X-ray characterized data exhibit hypersensitivity to protease digestion at multiple sites, which argues strongly against ordered structure for these regions. The NMR NNP shows very poor performance on several of the X-ray-characterized regions of disorder with false negative values well over 50%. The overall accuracies of the NMR NNP on these proteins gives reasonable values, above 50% in every case, because the predictions on the ordered parts of these proteins have good accuracies.

An alternative possibility to explain the very poor predictions of the NMR NNP on the X-ray-characterized proteins is that their disorder depends on charged residues, which are utilized by the X-RAY NNP but not by the current NMR NNP. As mentioned above, the disordered region of topisomerase II is highly charged and therefore consistent with this suggestion. Examination of the other disordered regions with high false negative error rates shows that these regions are also highly charged.

4.4 Control Predictions

The false positive rates of 16.2% (X-RAY NNP) is much lower than the evaluation of this same predictor on NRL_3D which gave a false positive error rate of 31.5%. The reasons for this large discrepancy are unclear, but may relate to the selection criteria for the control proteins in this study (e.g. small overall size to match the sizes, monomeric, no ligands, except for the metal ion in carboxypeptidase). NRL_3D contains large proteins, oligomeric proteins, and proteins with bound co-factors. All of these factors could contribute to false positive predictions. For example, a local tendency for disorder would be more likely to be over-ridden by non-local interactions in a larger protein as compared to a smaller one simply by chance, so larger proteins are more likely to have false positive predictions of disorder. We are testing this possibility. Oligomeric proteins might associate via disordered regions, in which case a prediction of disorder would appear as a false positive in the crystal structure of the oligomer. Finally, binding of co-factors could involve disorder-to-order transitions, so again predictions of disorder would appear as false positives in the structure of the holoprotein.

The NMR NNP exhibited an especially low false positive error rate on the control proteins, just 2.4%. This is more than 6 times smaller than the false positive error rate, 14.3%, on the ordered parts of the X_RAY NNP training set. For the NMR NNP on the X-RAY NNP training set, most of the false positive errors related to improper placement of the order / disorder boundary. Because of our windowing procedure, disorder information is carried into the ordered regions with the resulting tendency that the NMR NNP predicts disorder farther along the sequence than it actually occurs when a region of disorder is present. A further possibility is that in solution the regions of disorder actually extend farther along their respective sequences than indicated in the X-ray structures, which are more ordered than the solution state due to the ordering effects of crystal formation.

4.5 Summary

An NMR NNP and an X-RAY NNP were developed and tested on each other's training set proteins. The X-RAY NNP seems gives similar results across the ordered and disordered regions for both its own training set (as measured by 5-cross validation) as for the out-of-sample NMR NNP training set. In contrast, the NMR predictor does much better on its own training set (as measured by 5-cross validation) as compared to the out-of-sample X-RAY NNP training set. These data support the validity of the X-RAY NNP as a general predictor of protein disorder, suggesting that the uncertain interpretation of disorder characterized by X-ray diffraction in principle did not lead to a significant problem. On the other hand, the NMR NNP appears to be much more specific, for reasons that are not fully understood. Overall, these data provide additional evidence that disordered sequence are distinct from ordered ones and thus merit recognition as a separate category of protein structure.

Acknowledgements

Partial support by the NSF research grant NSF-CSE-IIS-9711532 to Z. Obradovic and A.K. Dunker is gratefully acknowledged. Garner thanks Dr. John Paznokas for providing support via an institutional grant for undergraduate research from the Howard Hughes Foundation.

References

[1] Abola, E.E., Sussman, J.L., Prilusky, J., Manning, N.O., Protein Data Bank archives of three-dimensional macromolecular structures, *Methods Enzymol.*, 277:556–571, 1997.

[2] Anfinsen, C.B., Principles that govern the folding of protein chains, *Science*, 181(96):223–230, 1973.

[3] Ann, D.K., Carlson, D.M., The structure and organization of a proline-rich protein gene of a mouse multigene family, *J. Biol. Chem.*, 260(29):15863–15872, 1985.

[4] Arnold, G.E., Dunker, A.K., Johns, S.J., Douthart, R.J., Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure, *Proteins: Structure, Function and Genetics*, 12(4):382–399, 1992.

[5] Aviles, F.J., Chapman, G.E., Kneale, G.G., Crane-Robinson, C., Bradbury, E.M., The conformation of histone H5. Isolation and characterisation of the globular segment, *Eur. J. Biochem.*, 88(2):363–371, 1978.

[6] Bairoch, A., Apweiler, R., The SWISS-PROT protein sequence data bank and its new supplement TREMBL, *Nucleic Acids Res.*, 24(1):21–25, 1996.

[7] Bennet, W.S., Hubber R., Structural and Functional Aspects of Domain Motions in Proteins, *CRC Critical Reviews on Biochemistry*, 15(4):291–369, 1984.

[8] Berger, J.M., Gamblin, S.J., Harrison, S.C., Wang, J.C., Structure and mechanism of DNA topoisomerase II, *Nature*, 379(6562):225–232, 1996.

[9] Boissy, G., de La Fortelle, E., Kahn, R., Huet, J.C., Bricogne, G., Pernollet, J.C., Brunie, S., Crystal structure of a fungal elicitor secreted by Phytophthora cryptogea, a member of a novel class of plant necrotic proteins, *Structure*, 4(12):1429–1439, 1996.

[10] Borkakoti, N., Moss, D. S., Palmer, R. A, *Acta. Crys. B.*, 38:2210–2217, 1982.

[11] Brick, P., Bhat T.N., and Blow D.M., Structure of Tyrosyl-tRNA Synthetase Refined at 2.3 Resolution, 1989.

[12] Burley, S.K., Petsko, G.A., Aromatic-aromatic interaction: a mechanism of protein structure stabilization, *Science*, 229(4708):23–28, 1985.

[13] Caron, P.R., Watt, P., Wang, J.C., The C-terminal domain of Saccharomyces cerevisiae DNA topoisomerase II, *Mol. Cell Biol.*, 14(5):3197–3207, 1994.

[14] Cho, H.S., Liu, C.W., Damberger, F.F., Pelton, J.G., Nelson, H.C., Wemmer, D.E., Yeast heat shock transcription factor N-terminal activation domains are unstructured as probed by heteronuclear NMR spectroscopy, *Protein Sci.*, 5(2):262–269, 1996.

[15] Chou, K.C., Does the folding type of a protein depend on its amino acid composition?, *FEBS Lett.*, 363(1–2):127–131, 1995.

[16] Daughdrill, G.W., Hanely, L.J., Dahlquist, F.W., The C-terminal half of the anti-sigma factor FlgM contains a dynamic equilibrium solution structure favoring helical conformations., *Biochemistry*, 37:1076–1082, 1998.

[17] Dayhoff, M.O., Ech, R.V., *Atlas of Protein Sequence and Structure*, 5, Natl. Biomed. Res. Found., Silver Spring, MD, 1978.

[18] Dodd, F.E., Hasnain, S.S., Hunter, W.N., Abraham, Z.H., Debenham, M., Kanzler, H., Eldridge, M., Eady, R.R., et al., Evidence for two distinct azurins in Alcaligenes xylosoxidans (NCIMB 11015): potential electron donors to nitrite reductase, *Biochemistry*, 34(32):10180–10186, 1995.

[19] Donne, D.G., Viles, J.H., Groth, D., Mehlhorn, I., James, T.L., Cohen, F.E., Prusiner, S.B., Wright, P.E., et al., Structure of the recombinant full-length hamster prion protein PrP(29-231): the N terminus is highly flexible, *Proc. Natl. Acad. Sci. U.S.A.*, 94(25):13452–13457, 1997.

[20] Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., et al., Protein Disorder and the Evolution of Molecular Recognition: Theory, Predictions and Observations, *Pacific Symposium on Biocomputing*, 3:471–782, 1998.

[21] Eisenberg, D., Weiss, R.M., Terwilliger, T.C., The helical hydrophobic moment: a measure of the amphiphilicity of a helix, *Nature*, 299(5881):371–374, 1982.

[22] Eisenberg, D., Weiss, R.M., Terwilliger, T.C., The hydrophobic moment detects periodicity in protein hydrophobicity, *Proc. Natl. Acad. Sci.*, 81:140–144, 1984.

[23] AEvarsson, A., Brazhnikov, E., Garber, M., Zheltonosova, J., Chirgadze, Y., al-Karadaghi, S., Svensson, L.A., Liljas, A., Three-dimensional structure of the ribosomal translocase: elongation factor G from Thermus thermophilus, *EMBO J.*, 13(16):3669–3677, 1994.

[24] Fedorov, A.A., Magnus, K.A., Graupe, M.H., Lattman, E.E., Pollard, T.D., Almo, S.C., X-ray structures of isoforms of the actin-binding protein profilin that differ in their affinity for phosphatidylinositol phosphates, *Proc. Natl. Acad. Sci. U S A*, 91(18):8636–8640, 1994.

[25] Fehlhammer, H., Bode, W., The refined crystal structure of bovine beta-trypsin at 1.8 A resolution. I. Crystallization, data collection and application of patterson search technique, *J. Mol. Biol.*, 98(4):683–692, 1975.

[26] Fletcher, C.M., McGuire, A.M., Gingras, A.C., Li, H., Matsuo, H., Sonenberg, N., Wagner, G., 4E binding proteins inhibit the translation factor eIF4E without folded structure, *Biochemistry*, 37(1):9–15, 1998.

[27] Franken, S.M., Rozeboom, H.J., Kalk, K.H., Dijkstra, B.W., Crystal structure of haloalkane dehalogenase: an enzyme to detoxify halogenated alkanes, *EMBO J.*, 10(6):1297–1302, 1991.

[28] Guez-Ivanier, V., Bedouelle, H., Disordered C-terminal domain of tyrosyl transfer-RNA synthetase: evidence for a folded state, *J. Mol. Biol.*, 255(1):110–120, 1996.

[29] Hopper, P., Harrison, S.C., Sauer, R.T., Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications, *J. Mol. Biol.*, 177(4):701–713, 1984.

[30] Huber, R., Conformational flexibility and its functional significance in some protein molecules., *TIBS*, 4:271–276, 1979.

[31] Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M., Clore, G.M., The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif, *Nat. Struct. Biol.*, 4(8):657–665, 1997.

[32] Kissinger, C.R., Parge, H.E., Knighton, D.R., Lewis, C.T., Pelletier, L.A., Tempczyk, A., Kalish, V.J., Tucker, K.D., et al., Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex, *Nature*, 378(6557):641–644, 1995.

[33] Kyte, J., Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.*, 157(1):105–132, 1982.

[34] Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., Lu, P., Crystal structure of the lactose operon repressor and its complexes with DNA and inducer [see comments], *Science*, 271(5253):1247–1254, 1996.

[35] Meyer, E., Cole, G., Radhakrishnan, R., Epp, O., Structure of native porcine pancreatic elastase at 1.65 A resolutions, *Acta. Crystallogr. B.*, 44(Pt 1):26–38, 1988.

[36] Mogridge, J., Legault, P., Li, J., Van Oene, M.D., Kay, L.E., Greenblatt, J., Independent ligand-induced folding of the RNA-binding domain and two functionally distinct antitermination regions in the phage lambda N protein, *Mol Cell*, 1(2):265–275, 1998.

[37] Muchmore, S.W., Sattler, M., Liang, H., Meadows, R.P., Harlan, J.E., Yoon, H.S., Nettesheim, D., Chang, B.S., et al., X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death, *Nature*, 381(6580):335–341, 1996.

[38] Nakashima, H., Nishikawa, K., Ooi, T., The folding type of a protein is relevant to the amino acid composition, *J. Biochem.*(Tokyo), 99(1):153–162, 1986.

[39] Ohba, T., Ishino, M., Aoto, H., Sasaki, T., Interaction of two proline-rich sequences of cell adhesion kinase beta with SH3 domains of p130Cas-related proteins and a GTPase-activating protein, Graf, *Biochem. J.*, 330(Pt 3):1249–1254, 1998.

[40] Pattabiraman, N., Namboodiri, K., Lowrey, A., Gaber, B.P., NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment, *Protein Seq. Data Anal.*, 3(5):387–405, 1990.

[41] Riek, R., Hornemann, S., Wider, G., Glockshuber, R., Wuthrich, K., NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231), *FEBS Lett.*, 413(2):282–288, 1997.

[42] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Dunker, A.K., A Calcinuerin Sequence with Few Planar Side Chains: Implicationsfor Structure and Function, *Protein Science*, August Meeting Abstract, 1996.

[43] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Dunker, A.K., Identifying Disordered Regions in Proteins from Amino Acid Sequences, *Proc. I.E.E.E. International Conference on Neural Networks*, 1:90–95, 1997.

[44] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., Guilliot, S., Garner, E., Dunker, A.K., Thousands of Proteins Likely to have Long Disordered Regions, *Pacific Symposium on Biocomputing*, 4-9 January 1998.

[45] Romero, P.Z., Obradovic, C., Dunker, A.K., Intelligent data analysis for protein disorder prediction, *Artificial Intelligence Review*, :in press, 1998.

[46] Sattler, M., Liang, H., Nettesheim, D., Meadows, R.P., Harlan, J.E., Eberstadt, M., Yoon, H.S., Shuker, S.B., et al., Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis, *Science*, 275(5302):983–986, 1997.

[47] Schulz, G.E., Nucleotide Binding Proteins, *Molecular Mechanism of Biological Recognition*, Elsevier/North-Holland Biomedical Press:79–94, 1979.

[48] Schulze-Kremer (ed) (1994) Development and Application of Three-Dimensional Description od Amino Acid Envioments in Proteins. *Advances in Molecular Bioinformatics*, IOS Press, Tokyo

[49] Shiozaki, K., Yanagida, M., A functional 125-kDa core polypeptide of fission yeast DNA topoisomerase II, *Molecular and cellular biology*, 11(12):6093–6102, 1991.

[50] Vihinen, M., Torkkila, E., Riikonen, P., Accuracy of Protein Flexibility Predictions, *Proteins: Structure, Function, and Genetics*, 19:141–149, 1994.

[51] Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., Lansbury, P.T., Jr., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded, *Biochemistry*, 35(43):13709–13715, 1996.

[52] Williamson, M.P., The structure and function of proline-rich regions in proteins, *Biochem. J.*, 297(Pt 2):249–260, 1994.

[53] Wilson, K.P., Malcolm, B.A., Matthews, B.W., Structural and thermodynamic analysis of compensating mutations within the core of chicken egg white lysozyme, *J. Biol. Chem.*, 267(15):10842–10849, 1992.

[54] Yao, J., Dyson, H.J., Wright, P.E., Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins, *FEBS Lett.*, 419(2–3):285–289, 1997.

[55] Zhang, O., Forman-Kay, J.D., Shortle, D., Kay, L.E., Triple-resonance NOESY-based experiments with improved spectral resolution: applications to structural characterization of unfolded, partially folded and folded proteins, *J. Biomol. NMR*, 9(2):181–200, 1997.