Predicting Protein Disorder for N-, C- and Internal Regions

Xiaohong Li[1] xli@eecs.wsu.edu
Pedro Romero [1] promero@eecs.wsu.edu
Meeta Rani [2] meeta_wsu@hotmail.com
A. Keith Dunker [2] dunker@disorder.chem.wsu.edu
Zoran Obradovic [1] zoran@eecs.wsu.edu

[1] School of Electrical Engineering and Computer Science
[2] School of Molecular Biosciences
Washington State University, Pullman, WA 99164, U.S.A.

Abstract

Logistic regression (LR), discriminant analysis (DA), and neural networks (NN) were used to predict ordered and disordered regions in proteins. Training data were from a set of non-redundant X-ray crystal structures, with the data being partitioned into N-terminal, C-terminal and internal (I) regions. The DA and LR methods gave almost identical 5-cross validation accuracies that averaged to the following values: 75.9 ± 3.1% (N-regions), 70.7 ± 1.5% (I-regions), and 74.6 ± 4.4% (C-regions). NN predictions gave slightly higher scores: 78.8 ± 1.2% (N-regions), 72.5 ± 1.2% (I-regions), and 75.3 ± 3.3% (C-regions). Predictions improved with length of the disordered regions. Averaged over the three methods, values ranged from 52% to 78% for length = 9-14 to ≥ 21, respectively, for I-regions, from 72% to 81% for length = 5 to 12-15, respectively, for N-regions, and from 70% to 80% for length = 5 to 12-15, respectively, for C-regions. These data support the hypothesis that disorder is encoded by the amino acid sequence.

1 Introduction

The current paradigm is that protein function depends on 3D structure [10, 16, 18], yet some proteins are partially or completely unfolded in their native states [2, 3, 7, 24, 26]. For such "natively unfolded" [30], "natively disordered" [9] or "intrinsically unstructured" [31] proteins, the lack of a fixed 3D structure can be an integral part of the function. Are such disordered proteins common or rare?

To estimate the commonness of disordered proteins, we applied predictors of disorder to appropriate databases [20]. The results suggested that intrinsic disorder is common [21], but lack of structural information limits confidence in these findings. Since the needed structural information will be slow in coming, we are revisiting the question of commonness by improving our disorder predictions.

A limitation of our previous studies was that only neural networks (NNs) were tried. By comparing NNs with discriminant analysis (DA) and logistic regression (LR), we can gain additional confidence in the suitability of prediction for identifying ordered and disordered protein.

Technical limitations of our previous algorithms resulted in absence of predictions on 15 residues at each end [20], resulting in non-prediction of a significant fraction of the residues. Here we modified the algorithms to extend the predictions to the N- and C-termini.

2 Materials and Methods

2.1 Data

Using missing electron density in X-ray structures as indicating disorder [19], we identified 115 N-terminal, 84 C-terminal and 69 internal (I) disordered regions (DRs) that were contained in 197 unrelated proteins listed in PDB-select-25 [11]. The minimum lengths used were 5 and 9 for termini

and I-regions, respectively. The various DRs contained the following numbers of residues 1,644 (N-regions), 1,347 (I-regions) and 1,250 (C-regions). A set of 130 unrelated, disorder-free proteins that were also from PDB-select-25 [11] was used to generate the ordered residues used for predictor training.

2.2 Attribute Generation

Composition-based and property-based attributes were calculated over sliding windows [20, 32] . A total 51 attributes were examined, where the sets of amino acids represented some property such as aromaticity, charge, sheet formers, etc (Table 1).

Var.	Attributes	Var.	Attributes	Var.	Attributes	Var.	Attributes
X1	FWY	X14	WCFIYVLHM	X27	WY	X40	P
X2	FWY(H/2)	X15	ATRGQSNPDEK	X28	A	X41	Q
X3	KR-D-E	X16	WYFAS	X29	C	X42	R
X4	KR-D-E(H/2)	X17	WYFKR	X30	D	X43	S
X5	KRDE	X18	WYFKRH	X31	E	X44	T
X6	KRDE(H/2)	X19	WYFDE	X32	F	X45	V
X7	WFYC	X20	WYFEDH	X33	G	X46	W
X8	WFYC(H/2)	X21	FWYKRDE	X34	H	X47	Y
X9	STQHNDERK	X22	FWYKRDEH	X35	I	X48	PEVK
X10	WEYCVILMP	X23	EMAL	X36	K	X49	Flexibility
X11	VILM	X24	YNPG	X37	L	X50	Hydropathy
X12	STQHN	X25	VIYFW	X38	M	X51	Coordination number
X13	GSA	X26	SGKPDE	X39	N		

Table 1: Attributes list.

Composition-based attributes were the sums of the numbers of the indicated amino acids in a given window. For example, aromaticity, X1 = FWY, the number of phenylalanines (F) + tryptophans (W) + tyrosines (Y) within a given window. The number of histidines was sometimes divided by 2 (e.g. H/2) due to its small ring size or partial charge. For the net charge attributes, X3 and X4, the number of each negative residue was subtracted (e.g. -D, -E) from the number of positive residues.

Property-based attributes were the sums the residue property-values. For X49 = flexibility, the value for each residue was based on its backbone-atom B-factors averaged over 92 unrelated protein structures [28]. The values for X50 =l hydropathy were from the Kyte-Doolittle scale [15]. X51 = coordination number is the average number of side chain neighbors that are in contact with the given side chain when it is fully buried as determined from a set of 33 non-homologous proteins [8].

As in previous studies [20], a window of 21 was used for I-regions. A window of 11 was used for positions 6 onwards and for -6 backwards for N- and C-regions, respectively. Predictions at positions 1 to 5 and -1 to -5 used windows of size 6 to 10, respectively. For N-regions, these windows included residues from the end to 5 positions beyond the position being predicted, and for C-regions, from the end to 5 positions before.

2.3 Logistic Regression Model and Attribute Selection

The logistic regression (LR) model was developed for dealing with the situation in which the dependent variable is binary [5]. Here we used order = 0, and disorder = 1. SAS (Release 6.12, SAS Institute, Cary, NC) was used for the calculations.

For a given threshold probability, an observation is classified into the category with the probability higher than the threshold. In the logistic model, the probability is estimated from the following equation:

$$ln[\frac{p}{1-p}] = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_j X_{ij}$$

where $p = P(Y_i = 1$ for ordered) and $1 - p = P(Y_i = 0$ for disordered); i = 1,2...n, where n is the sample size; j = 1, 2, ..., j, where j is the attribute number; and $X_{i1}, \cdots, X_{ij}$ are attributes used for prediction.

The parameters b_i are estimated by maximizing the following function:

$$\sum_{i=1}^{n} P(B, Y_i) = \sum_{i=1}^{n} ln(\frac{1}{1 - e^{-BX_i}})$$

where B is the vector of parameters need to be estimated. After all b_i values are estimated, p can be calculated as:

$$p = \frac{1}{1 + e^{-BX_i}}$$

For order = 0 and disorder = 1, the threshold is set to be 0.5; if $p \geq 0.5$, then the amino acid is predicted to be disordered; otherwise, ordered. The LR is applied each time an attribute is introduced or removed, and the Chi-square test is executed [1]. The process is repeated until introduction or removal of an attribute leads to no change at a significance level of 0.05. Eight selected attributes were used in LR predictor even though a few more number passed the significance test.

2.4 Discriminant Analysis Model

For discriminant analysis (DA), it is assumed that prior probabilities are equal, that the variables (attributes) are independent, and that all attribute values satisfy the normal distribution. Since we used sliding windows to obtain data and since many of the attributes share dependencies on the same amino acids, the assumption that the data are independent is not true. However, this lack of independence didn't seem to cause serious problems since this approach gave results comparable to the other methods in this study. Again, SAS (Release 6.12, SAS Institute, Cary, NC) was used to carry out the calculations for this model.

For the ordered and disordered data $\chi = \{\mathbf{x}_i, y_i\}, i = 1 \cdots n; y_i = \{0, 1\}$, where $y = 0$ for an ordered amino acid, $y = 1$ for a disordered one. The x_i values are the attributes data. We used Bayesian discriminant analysis method to predict the probability that a given amino acid belongs to an ordered or disordered regions. The posterior (conditional) probability that a residue belongs to an ordered or disordered region is given by the following equation:

$$P(C_j \mid x) = \frac{P(x \mid D)P(D)}{P(x \mid D)P(D) + P(x \mid O)P(O)}$$

where j = 0 (ordered) or 1 (disordered); P(O) and P(D) are the a priori probabilities of a residue being ordered and disordered residues, respectively. $P(x \mid D)$ and $P(x \mid O)$ are the conditional densities of disordered and ordered residues, respectively. $P(Cj \mid x)$ is given by the following relationship:

$$P(C_j \mid x) = \frac{e^{C_{j0} + \mathbf{b_j'x}}}{\sum_{k=1}^{m} e^{(b_{k0} + \mathbf{b_k'x})}} = \frac{1}{1 + e^{(b_{d0} - b_{o0}) + (\mathbf{b_d} + \mathbf{b_o})'\mathbf{x}}}$$

Using observed data, the parameters b_{d0} and b_{o0} and the vectors $\mathbf{b_d}$ and $\mathbf{b_o}$ can be estimated. The classification for a given pattern x is determined as: $Class = argmax\{P(C_j \mid x\}$, where class is 0 or 1 for ordered or disordered, respectively.

The attributes were repeatedly introduced or removed, and the F-test was applied after each operation, until no attributes could be introduced or removed at a significance level of 0.05 [6]. The top eight selected attributes were used for establishing the DA predictor even though a greater number were accepted at the significance level indicated.

2.5 Neural Network Model

The application of NNs to order/disorder prediction has been described elsewhere in more detail [20]. The feed forward NN used in this study is fully connected with an 8x8x1 architecture, which has eight inputs (selected by LR), one hidden layer with 8 nodes and one output layer with one node. The back propagation method was used for data training [23].

3 Results

3.1 Attribute selection

A list of 51 attributes was used in this study (Table 1). Many of the attribute values are correlated. In addition, some attributes make little contribution in distinguishing the ordered and disordered regions. Finally, 51 attributes is simply too many for the amount of disordered data. These characteristics necessitated the selection of a subset of the attributes for the predictors.

Stepwise DA and stepwise LR were used for attribute selection on ordered and disordered data from the N-, C- and I- regions. Although more than 8 attributes were selected for the data at a significance level of 0.05, the ninth and later selected attributes make relatively little contribution, as shown by the prediction accuracy upon addition of attributes in their order of importance (Fig. 1).

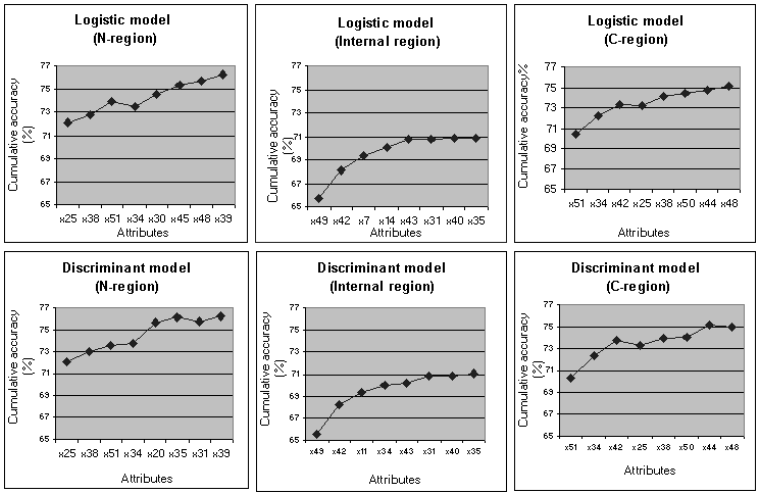

Figure 1: Contribution of selected attributes on prediciton

The selected attributes in Table 2 start with the most important. For the top 8 sequence attributes in a given protein region, the DA and LR models selected almost the same ones. That is, 5/8, 6/8, and 8/8 attributes were selected in common by the two methods for the N-, I-, and C-region data, respectively. In contrast, the selected attributes were very different for the different regions. Only 1 sequence attribute was selected in common for all three regions. For the 3 pairs of regions, only 4/8 were selected in common for N- and C-regions, just 3/8 for the N- and I-regions, and a mere 2/8 for the C- and I-regions. These results suggest that sequence characteristics leading to disorder depend on the location of the region in the sequence.

3.2 Prediction Accuracies

The prediction accuracies of the 3 models over the 3 regions are given in Table 3. The DA and LR models gave almost identical accuracies for each region, with the largest difference being 0.3% (for I-regions). Also, using the N-regions as an example, the 0.1% difference between the two methods is

Attributes	1	2	3	4	5	6	7	8
DA: N-terminal region	X25	X38	X51	X34	X20	X35	X31	X39
LR: N- terminal region	X25	X38	X51	X34	X30	X45	X48	X39
DA: Internal region	X49	X42	X11	X34	X43	X31	X40	X35
LR: Internal region	X49	X42	X7	X14	X43	X31	X40	X35
DA: C-terminal region	X51	X34	X42	X25	X38	X50	X44	X48
LR: C-terminal region	X51	X34	X42	X25	X38	X50	X44	X48

Table 2: Attributes selected according to the significance in DA and LR.

much less than the ± 3.5% and ± 2.7% variation among the 5-cross validation trials. Thus, the DA and LR models give essentially indistinguishable prediction accuracies overall.

Model	Region	1	2	3	4	5	Average
Neural Network	N region	79.0%	78.8%	78.7%	78.9%	78.7%	78.8% (±1.2%)
	I region	72.2%	72.6%	73.1%	72.2%	72.4%	72.5% (±1.2%)
	C region	75.1%	75.5%	74.9%	74.4%	76.5%	75.3% (±3.3%)
Discriminant Analysis	N region	74.2%	78.4%	75.9%	73.7%	77.2%	75.9% (±3.5%)
	I region	70.1%	71.3%	70.0%	71.8%	71.1%	70.9% (±1.4%)
	C region	72.7%	71.6%	77.0%	76.3%	75.9%	74.7% (±4.1%)
Logistic Regression	N region	74.0%	77.3%	76.3%	74.2%	77.2%	75.8% (±2.7%)
	I region	69.6%	70.62%	69.8%	71.7%	71.4%	70.6% (±1.6%)
	C region	72.0%	71.3%	77.3%	76.6%	75.5%	74.5% (±4.7%)

Table 3: Five-cross validations of the predictors developed by three methods.

The NN approach gives slightly higher predictions for all three regions. In the following, the first number in each pair is the NN accuracy and the second number is the average of the DA and LR accuracies: 78.8 ± 1.2% versus 75.9 ± 3.1% (N-regions), 72.5 ± 1.2% versus 70.7 ± 1.5% (I-regions), and 75.3 ± 3.3% versus 74.6 ± 4.4% (C-regions).

3.3 Cross Prediction

Each predictor was applied to the data from the regions not used for its training, here called cross prediction. In Table 4 accuracies observed during 5-cross validation (indicated by *) are compared with the accuracies for cross predictions (no *). For the most part, as expected, the accuracy of a given predictor drops when applied to the data from a region different from its training set. However, for both the LR and DA models trained on I-regions, the accuracies remain essentially the same when the predictors are applied to C-region data. That is, the LR model only changes from 70.6% on its I-regions training data to 70.9% when applied to C-region data, and the DA model, from 70.9% to 71.2%. This failure to drop in accuracy is especially surprising since I- and C-regions predictors share just 2/8 attributes.

3.4 Length dependence of prediction accuracy.

To estimate accuracy versus length, the prediction outputs were partitioned according to length with the number of residues in each class indicated in parenthesis (Table 5). For the DA and LR predictions in Table 5, the models from 5-cross validation were retrained on 5/5 of the data, whereas for the NN

Predictors	Region	N-terminal Data	Internal DR Data	C-terminal Data
Discriminant Model	N-region	75.9%*	52.9%	61.5%
	Internal region	64.9%	70.9%*	71.2%
	C-region	71.3%	68.8%	74.7%*
Logistic Model	N-region	75.8%*	44.6%	57.6%
	Internal region	66.3%	70.6%*	70.9%
	C-region	71.6%	68.9%	74.5%*

Table 4: Cross-prediction specificity for disordered regions.

predictions, retraining on the whole set of data was not performed. Instead, one of the NN models, which was trained on 4/5 of the data, was used. For DRs of 9 to 14, the roughly 52% accuracy (averaged over the 3 methods) corresponds to essentially random classification. For DRs of 15 to 20, the average accuracy increased to 74%, and for DRs ≥ 21 the average increased still further to about 78%. Since the windows are 21 in length, the shorter DRs fill only a fraction of their windows, and therefore the poor accuracies are expected.

Predictors	9-14 AA (379)	15-20 AA (262)	21AA or longer (707)
NN	52.8%	73.7%	78.6%
DA	50.9%	74.4%	77.9%
LR	52.2%	74.4%	78.2%

Table 5: Prediction accuracies for different I-DR lengths.

The lowered prediction rates due to the short disordered windows probably helps to explain the surprising cross prediction results that occur when the predictors trained on I-regions are applied to C-region data as described above.

The N- and C-region data also show length-dependent accuracies (Table 6). For N-region data, the accuracies, averaged over the three methods, change from 72% (length = 5), to 83% (length = 6-8), to 77% (length = 9-11) to 81% (length = 12-15). For C-region data, the respective averaged accuracies are 69%, 78%, 72% and 80%.

DR Regions	Predictors for N and C regions	DR=5 AA (N:60; C:65)	DR=6-8AA (N:269; C:117)	DR=9-11 AA (N219; C135)	DR=12-15AA (N:137; C:163)
N	NN	75.0%	83.6%	77.1%	86.0%
	DA	71.7%	83.3%	78.1%	81.0%
	LR	70.0%	82.2%	76.3%	77.4%
C	NN	70.5%	73.1%	74.2%	85.2%
	DA	67.7%	74.4%	63.0%	75.5%
	LR	67.7%	74.4%	63.0%	76.1%

Table 6: Prediction accuracies for different N- or C-DR lengths.

3.5 Position-by-position accuracy for N- and C-regions

The position-by-position error rates were determined; all three predictors give similar outputs that result in fairly complex curves (figure 2). The data in figure 2 are incommensurate with the data in Table 6, so these should not be compared directly. This is discussed below in more detail.

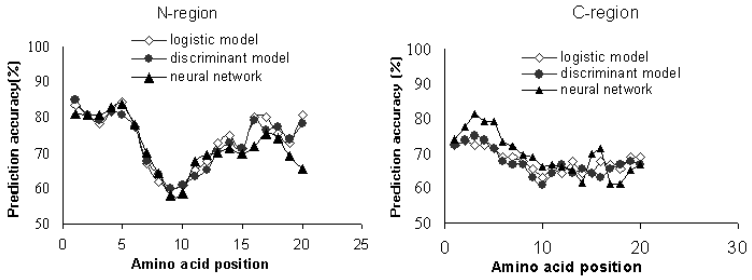

Figure 2: Prediction accuracy over AA positions in N- and C- regions.

4 Discussion

4.1 Data

Disorder characterized by X-ray diffraction can be static or dynamic [13]. In our previous studies we attempted to remove this ambiguity by finding independent information such as protease sensitivity or NMR spectra [20], but most often the information was lacking. As an alternative, we compared X-ray-characterized disorder with disorder characterized by other methods especially NMR [9]. The results indicate that ambiguity of X-ray characterized disorder is not fatal, but probably leads to the introduction of noise into the training data.

4.2 Comparison of Prediction models

There is no single best algorithm for pattern recognition problems. Performance for a given algorithm depends on the data set being investigated [14]. DA, LR and NN approaches are among the most commonly used, and all have been applied to sequence analysis problems. DA has been successfully used for predicting internal exons of DNA sequences [25] and protein secondary structure [27, 33]. LR has been used for identifying regulatory regions of genes [29]. NNs have also been used for predicting secondary structure [22]. Considering the characteristics of the three methods, we decided to try all of them in this study.

The LR and DA models exhibited nearly identical performance for the disorder predictions whereas the NN gave a slightly higher accuracy (Table 3). Application of Cochran's test [4] indicates a real significance for the superiority of the neural network. However, prediction accuracy is a simplistic indicator, so it seems inappropriate to rank the methods on this basis alone.

Olson [17] reported that, with proper selection of attributes, both statistical and neural network classifiers yield essentially identical accuracies for a given test case. From this, there are two implications that arise from the possible superiority of the NN predictors. First, other factors not included in Table 1 might affect the determination of order or disorder. To test this, other attributes need to be investigated. Alternatively, the predictors might not be optimized.

DA is fast and performs well except for very skewed data [14]. LR was developed for binary data and so might be the most robust for predicting two states, order and disorder. DA and LR methods need much less computation time than NN, and produce results that are easier to interpret.

Back propagation NN, in most cases, performs well especially for noisy data. Noisy data is of particular concern due to the ambiguity of X-ray-characterized disorder. With appropriate architecture,

a back propagation neural network can be a universal approximator for arbitrary finite inputs [12]. No assumptions are required for the input and output parameters.

There are some general disadvantages, however, in using NNs. For example, the selection of the architecture (number of layers, number of neurons) is empirical. If too few hidden neurons are used, training convergence is often poor, whereas if too many are used, the network might converge well, but generalization is typically poor. A further shortcoming of NNs is the failure to provide insight. That is, there is no deterministic way to carry out attribute selection. For these reasons, we carried out an entirely separate study to gain understanding of our problem [32]. A significant advantage of the LR and DA methods is the ability to carry out step-wise addition of the various attributes.

4.3 Attribute Selection

Both our previous studies and the studies on I-region data presented here used windows of 21 residues. Despite the very different databases in the two studies, the previously selected attributes closely resemble those reported here. That is, 6 of 8 attributes were selected in common by the LR and DA methods; these were X49 (flexibility), X42 (R), X43 (S), X31 (E), X40 (P), and X35 (I) as shown in Table 2. Of the 6 attributes in common, 5 were selected in our previous studies on completely different databases of order and disorder; only the last, and least important attribute found here, X35, was not selected previously. Of the 4 attributes not selected in common, e.g. X11 (VILM) and X34 (H) by DA and X7 (WFYC) and X14 (WCFIYVLHM) by LR, all are identical to, or share amino acids with, attributes selected previously on completely different data.

The prediction of order or disorder for I-regions depends on a balance of different types of attributes. X49 (flexibility), X42 (R), X43 (S), X31 (E), and X40 (P) are attributes that, at high value, favor disorder, whereas X35 (I), X11 (VILM), X7 (WFYC), and X14 (WCFIYVLHM) all favor order.

This is the first study of the relationship between amino acid sequence and disorder at the ends of proteins. Comparing attributes for N- and C-regions with each other and with attributes for I-regions provides insight regarding disorder at the ends of proteins.

Although just 4/8 attributes are in common between the two ends, these include the top two attributes for each (Table 2). That is, the top two attributes, X25 (VIYFW) and X 38 (M), for N-regions data rank fourth and fifth, respectively, for C-regions data. Also, the first, X51 (Coordination Number), and second, X34 (H) for C-regions rank third and fourth, respectively, for N-regions. From figure 1, these top attributes are the most important. Of the attributes specific for each end, some of these contain residues with charges opposite to the charge at the termini (Table 2). For example, the positive charge at the N-terminus is opposite to the negative charges (E and D) in X20 (WYFEDH) and to that of X31 (E). Likewise, the negative charge of the C-terminus is opposite to the positive charge of X42 (R).

The attributes selected for the N- and C-regions can for the most part be described as being associated with the formation of ordered structure, whereas the attributes selected for I-regions appear to be more balanced between attributes favoring order and those favoring disorder. Even the charged attributes, X31 (E), and X42 (R), which are associated with disorder in I-regions, are selected at the ends in a manner that brings about charge balance and so could be promoting order in these regions. Perhaps I-regions are neutral with respect to order or disorder, whereas perhaps N- and C-regions tend to be naturally disordered. If so, order or disorder in I-regions is determined by the overall balance of various types of attributes, whereas overcoming the natural disorder tendency at the ends may require the presence of order-inducing amino acids in these regions.

4.4 Prediction accuracies

If only the longer I-regions data are considered, the estimated accuracy here (Table 5) is slightly better than we found earlier. That is, here we find about 78% (average of DA and LR) versus about 73% - 74% (NN) reported previously [20]. The slight improvement probably relates to the increased number

of attributes surveyed, 51 here versus 24 previously. More specifically, only single amino acids were used in the original study, whereas the expanded set used here contains combinations of amino acids. Several of the selected combinations include groups of the single amino acids selected in the original study, thus creating space for additional inputs that bring more information to bear on the problem.

The length-dependence of I-region predictions shows a very large gradient, from almost random predictions (near 52% averaged over the three methods) for length = 9-14 to fairly strong predictions (about 78% averaged over the three methods) for length $\geq$ 21. Because windows of 21 were used, the shorter lengths only partially filled the windows and so the essentially random predictions are a reasonable outcome when the disorder training examples contain large amounts of order.

Here we report our first attempt to predict to the ends of the protein. We included down to very short DRs (5 amino acids) with the expectation that we would find some minimum length below which the predictions would fail completely. Such failure would give random predictions like those observed for the shortest I-regions data, although for different reasons. To our surprise, even DRs as short as 5 amino acids at the ends yielded good prediction accuracies, about 72% (N-regions) and 70% (C-regions) when averaged over the three methods (Table 6). Although not monotonic, increases in accuracy reached 82% (N-regions) and 80% (C-regions) for DRs of length 12-15. These high values suggest the possibility of special effects at the ends of proteins.

The NN, LR, and DA methods give similar curves for the position-dependent accuracies at each end (figure 2), with high value followed by minima that are very noticeable for the N-region curves and barely noticeable for the C-regions curves. The causes of these minima near positions 9-10 are uncertain. One possibility is that windows at the 9-10 positions for the disorder data contain substantial fractions of ordered residues, resulting from a combination of the distribution of disorder lengths in the training data and the way in which the windows were specified. Based on this idea, we are exploring alternative window specifications with the goal of reducing these minima.

The data in figure 2 were grouped differently from the data in Table 2. This leads to false discrepancies such as the $>$ 80% accuracies for positions 1-5 (N-regions, figure 2) which appear to be better than the 72% accuracy for N-region DRs = 5 AA (averaged over the 3 methods from Table 6). The false discrepancy arises because the data for Table 6 come from the specified lengths whereas the data for figure 2 are predictions at particular positions from DRs of all different lengths. So, the higher accuracy of $>$ 80% for the first 5 positions results from contributions from DRs longer than 5, which yield predictions over the first 5 positions better than the 72% observed for DRs of length = 5.

4.5 Implications for Future Research

The high accuracy of prediction of very short DRs at the termini might be special, due to end effects, or the high accuracy might be simply the result of the use of very short windows. If the latter is true, then use of shorter windows might be of benefit for I-region predictions as well.

A second task will be to merge our various predictors into one, making it possible to predict disorder from the amino to the carboxyl terminus of a protein. This will open the way for a variety of projects, such as improving predictions of disorder on a genomic basis and such as using disorder predictions to indicate which proteins are likely to crystallize and which ones are not.

Acknowledgments

Support from NSF research grant NSF-CSE-IIS-9711532 to Z. Obradovic and A. K. Dunker is gratefully acknowledged. Dr. R. Drossu's neural network simulator was used in part of the study.

References

[1] Anderson, E.B., "The statistical analysis of categorical data.," 2ed ed. Springer-Verlag.1991.

[2] Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., Klug, A., "Protein Disk of Tobacco Mosiac Virus at 2.8 Resolution Showing the Interactions Within and Between Subunits," *Nature,* 276:362-368, 1978.

[3] Bode, W., Schwager, P., Huber, R., "The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 A resolution," *J. Mol. Biol.,* 118(1):99-112, 1978.

[4] Cochran, W.G., "The comparison of percentages in matched samples.," *Biometrika,* 37:256-266, 1950.

[5] Cox, D.R., "The analysis of binary data.," London:Methuen and Co.1970.

[6] Eisenbeis, R.A., Avery, R.B., "Discriminant analysis and classification procedures: theory and applications.," Lexington, Mass.:Health. 1972.

[7] Fletcher, C.M., Wagner, G., "The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein " *Protein Sci.,* 7(7):1639-1642, 1998.

[8] Galaktionov, S.G., Marshall, G.R., "Prediction of Protein Structure inTerms of Intraglobular Contacts: 1D to 2D to 3D.," Technical Report, Center for Molecular Design, Washington University, St. Louis, 1996.

[9] Garner, E., Cannon, P., Romero, P., Obradovic, Z., Dunker, A., "Predicting disordered regions from amino acid sequence: common theme despite differing structural characterization.," *Genome Informatics,* 9:201-214, 1998.

[10] Hagerman, P.J.I., "From sequence to structure to function," *Curr Opin Struct Biol,* 6(3):277-280, 1996.

[11] Hobohm, U., Sander, C., "Enlarged representative set of protein structures.," *Protein Sci,* 3(3):522-524, 1994.

[12] Hornik, K., "Approximation capabilities of multilayer feedforward networks.," *Neural network*, 4:251-257, 1991.

[13] Huber, R., "Conformational flexibility in protein molecules.," *Nature (London)*, 280:538-539, 1979.

[14] King, R.D., Feng, C., Sutherland, A., "Statlog: comparison of classification algorithms on large real-world problems.," *Applied artificial intelligence,* 9:289-333, 1995.

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

[16] Mirsky, A.E., Pauling, L., "On the Structure of Native, Denatured and Coagulated Proteins," *Proc. Natl. Acad. Sci. U.S.A ,* 22:439-447, 1936.

[17] Olson, K.M., Ybarra, G.A., "A performance comparison of neural network and statistical pattern recognition approaches to automatic target recognition of ground vehicles using SAR imagery.," Miceli, W. J. *Proceeding of SPIE: Radar processing, technology, and applications II.* San Diego, CA, USA. August 1997, 3161:159-170, 1997.

[18] Orengo, C.A., Todd, A.E., "From protein structure to function.," *Curr. Opin. Struct. Biol.,* 9:374-382, 1999.

[19] Rani, M.P.R., Z. Obradovic, and A. K. Dunker, "Annotation of PDB with respect to "disordered regions" in proteins.," *Genome Informatics,* 9:240-241, 1998.

[20] 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.

[21] 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," *Pac. Symp. Biocomp.,* 3:435-446, 1998.

[22] Rost, B., Sander, C., "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proc. Natl. Acad. Sci. U.S.A.,* 90(16):7558-7562, 1993.

[23] Rumelhart, D.E., Durbin, R., Dolden, R., Chauvin, Y., "Backpropagation: the basic theory.," Y. Chauvin, D. E. Rumelhart (ed): Back propagation: theory, and applications., Hillside, NJ: Lawrence Erlbaum:1-34, 1985.

[24] Schulz, G.E., "Nucleotide Binding Proteins," Molecular Mechanism of Biological Recognition, Elsevier/North-Holland Biomedical Press: pp79-94, 1979.

[25] Solovyev, V.V., Salamov, A.A., Lawrence, C.B., "Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames," *Nucleic Acids Res,* 22(24):5156-5163, 1994.

[26] Spolar, R.S., Record II, M.T., "Coupling of Local Folding to Site-Specific Binding of Proteins to DNA," *Science,* 263:777-784, 1994.

[27] Stolorz, P., Lapedes, A., Xia, Y., "Predicting protein secondary structure using neural net and statistical methods," *J Mol Biol,* 225(2):363-377, 1992.

[28] Vihinen, M., Torkkila, E., Riikonen, P., "Accuracy of protein flexibility predictions," *PROTEINS: Struct. Funct. Genet,* 19(2):141-149, 1994.

[29] Wasserman, R., Felix, C.A., McKenzie, S.E., Shane, S., Lange, B., Finger, L.R., "Identification of an altered immunoglobulin heavy-chain gene rearrangement in the central nervous system in B-precursor acute lymphoblastic leukemia," *Leukemia,* 7(8):1294-1299, 1993.

[30] 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.

[31] Wright, P.E., Dyson, H.J., "Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm," *J. Mol. Biol.,* 293(2):321-331, 1999.

[32] Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E., Dunker, A.K., "The Sequence Attribute Method For Determining Relationships between Sequence and Protein Disorder," *Genome Informatics,* 9:193-200, 1998.

[33] Zhang, X., Mesirov, J.P., Waltz, D.L., "Hybrid system for protein secondary structure prediction [published erratum appears in J Mol Biol 1993 Aug 20;232(4):1227]," *J. Mol. Biol.,* 225(4):1049-1063, 1992.