Comparing Predictors of Disordered Protein

Xiaohong Li[1] xiahong@livegrip.com
Zoran Obradovic[1] zoran@joda.cis.temple.edu
Celeste J. Brown[2] celesteb@disorder.chem.wsu.edu
Ethan C. Garner[2] egarner@itsa.ucsf.edu
A. Keith Dunker[2] dunker@disorder.chem.wsu.edu

[1]School of Electrical Engineering and Computer Sciences, Washington State University, Pullman, WA 99164-2752, USA
[2]School of Molecular Biosciences, Washington State University, Pullman, WA 99164-4660, USA

Abstract

More than 6,000 amino acid sequence attributes were ranked by their conditional probabilities for indicating ordered or disordered protein structure. The top 10 each from several different groups of attributes were merged with still other attributes and then subjected to selection by logistic regression. Evidently, the determination of order or disorder results from the interplay among several attributes, such as average Coordination Number, aromatic content and the numbers of non-polar amino acids, all of which favor the ordered state, and others like Net Charge, Flexibility Index, and the presence of certain polar amino acids, all of which favor disorder. The top 12 selected attributes were used as inputs for artificial neural network (ANN) predictors. Five predictors were developed, compared with each other, and with previous work. The best of these shows substantially improved generalization compared to our previously published predictor.

Keywords: disordered proteins, sequence attributes, prediction, artificial neural networks

Introduction

Current dogma holds that amino acid sequence determines 3-D protein structure [2] with the resulting 3-D structure being a prerequisite for function. However, many proteins remain unfolded under physiological conditions, yet carry out function. These "natively unfolded" [26] or "intrinsically disordered" proteins have led to a call for a re-assessment of the protein structure / function paradigm [27].

Amino acid sequence determines protein folding, so sequence should also determine non-folding [20]. If so, then the existence of intrinsically unstructured proteins implies a "protein non-folding problem," i.e. the prediction whether a given string of amino acids folds into a 3-D structure or remains partially or completely unfolded. Recently we are focusing on this non-folding problem [6, 18, 20, 7, 9, 21, 28, 10, 15, 19, 22].

Our published studies on the non-folding problem suffer from at least three limitations. First, small numbers of ordered and disordered amino acids were used, due mainly to the lack of appropriate data. Second, the predictor inputs were selected from a fairly small pool of sequence attributes, so many potentially important sequence features were not evaluated. Finally, attributes were calculated as their average values about a central position; these simple windows contain no directional information.

The present study focuses on the three aspects mentioned above. First, more data were used for predictor training. Second, a substantially larger attribute pool was explored. Finally, simple windows were compared with a triple-windowing process that included directional information. Overall, the results provide additional insight into the sequence basis of intrinsic disorder and yield improvements in prediction accuracy.

Materials and Methods

2.1. Data

Data on regions of disorder were from a non-redundant set of 57 putatively disordered segments of at least 21 residues in length from which only 898 residues were used, an increase in size of about 80% over the data in our initial study[20]. The first and last 5 residues of each protein chain were dropped to reduce the end-

effects noted previously [15], so disorder at the ends of chains had to be at least 21 + 5 = 26 residues in length to be used. Sliding windows of 21 residues in length were used to calculate sequence attributes. The disordered training set was balanced by ordered segments of length 21 randomly chosen from 130 non-homologous, putatively structured proteins having no disorder.

Several out-of-sample databases of order and disorder were used or constructed for evaluation of the various predictors. PDB_Select_25 contains proteins families based on 25% sequence identity from the Protein Data Bank [3], with the highest quality structure representing each family [11]. Starting with 931 X-ray-determined protein structures in the PDB_Select_25 of August, 1999, the 230,777 observed residues from 1,135 segments of at least 21 in length were gathered to form a database of ordered structure called O_PDB_Select_25. The 4,781 unobserved residues in 86 segments of at least 21 in length were likewise compiled to form a disordered set, called D_PDB_Select_25. NRL_3D [17] contains only the ordered (e.g. observed) segments of the proteins in PDB, but differs from O_PDB_Select_25 in having many redundant proteins. In total, the version of NRL_3D used here had 17,791 protein chains with a total of 2,636,570 residues. Intrinsic disorder has also been characterized by NMR and by far UV circular dichroism (CD). Such proteins were identified by literature searches, giving an NMR disorder dataset of 33 proteins having less than 25% sequence identity with 3,331 residues and a far UV CD disorder dataset of 45 proteins having less than 25% sequence identity with 6,438 residues.

All of the ordered and disordered data used here are described in detail at http://disorder.chem.wsu.edu, accessed by the Database button on the home page.

2.2. Attribute construction and ranking

Our previous study [15] used 51 attributes, 46 of which were based on composition and 5 of which were based on residue properties. These attributes were developed largely from domain knowledge.

Composition-based attribute values are determined simply by summing the numbers of the specified amino acids in a given window. Here, we exhaustively screened every amino acid combination having 1-4 amino acids, giving 6195 composition-based attributes in all. Additional composition-based attributes with more than 4 amino acids were included for comparison with previous studies [28, 15]

In addition to the composition-based attributes, we investigated 14 property-based attributes. With the convention that a property will be capitalized if it is used as an attribute, these included Hydropathy [13], two scales of Flexibility [25](Smith, personal communication), Coordination Number [8] two different measures of formal Net Charge [28], two different scales of Residue Volumes [23, 4], two scales of Side Chain Polarity [12, 4], Surface Area [5], Bulkiness [12], Refractivity [12], Electron-Ion Interaction Potential (EIIP) [24, 14]. The pairs of attributes are qualified by the last names of the developers, yielding Flexibility-V and Flexibility-S, Volume-S and Volume-C, Polarity-J and Polarity-C.

From the collected data, the ln(odds-ratio) of a given site being ordered or disordered is fitted as:

$$\ln\left(\frac{p}{1-p}\right) = bx, \tag{1}$$

where p is the probability of being ordered, 1 - p the probability of being disordered, b is the constant to be estimated, and x is the given attribute value. Positive values of b are found for attributes that correlate positively with order and negative values for those that correlate with disorder.

Once the parameter b is estimated by fitting to the data, the probability of being ordered is calculated as

$$p = \frac{e^{bx}}{1+e^{bx}}, \tag{2}$$

with the probability of disorder being simply $1-p$.

The attributes were ranked by the absolute values of b. The top 20 attributes (10 for order and 10 for disorder) were picked from each of the 2, 3, and 4 residue combinations, giving 60 new composition-based attributes for further study. These 60 plus 51 attributes from our previous study and the 9 new property-based attributes were pooled, giving a total of 120 attributes for further analysis.

2.3. Windowing

Three windowing procedures were used, called left, right and whole. For Left Windows, attributes were calculated over 11 residues, the one being predicted and the 10 to its right. The 11 residue Right Windows

included the 10 to the left plus the position being predicted. Whole Windows included 10 on each side of the prediction site, for a total of 21 residues.

2.4. Attribute selection

The logistic regression method [1] was used for attribute selection as described in our previous study [15]. Briefly, let p be the probability of order (1-p is disorder) as defined above. Then the following model is established using the data:

$$\ln(\frac{p}{1-p}) = b_0 + b_1 x_1 + ...b_j x_j , \quad (3)$$

where the b_i values are the parameter estimates, and x_i are the attributes used in the selection process. Attributes were introduced or removed one-by-one at each step. To determine whether an attribute should be introduced or removed from the model, the Chi-square test was carried out at each step to test whether such action had a significant effect on the model or not. A 0.15 significance level was used. Any of the previously selected attributes, not just the last one, could be removed, thus enabling the discovery of synergistic effects for attributes that might be weak individually. If no significant effect is observed, then the attribute was removed from the model, or was not added to the model. The process is repeated until no attribute can be added to, or removed from, the model. Although more than 20 ranked attributes were identified by this step-wise logistic regression to the significance level mentioned, just the top 12 were used for training neural network predictors since the lower ranked attributes contribute less and less to the discriminating power [15].

2.5. Training the feedforward neural networks

A detailed description of the application of neural networks to the prediction of ordered and disordered proteins was previously reported [20, 15]. The feedforward artificial neural networks (ANNs) used in this study contained one input layer, one hidden layer and one output layer. During training, the data were partitioned into 5 disjoint subsets, with training on 4/5 and testing on 1/5; the accuracies obtained during this 5 cross-validation were averaged and are reported below as one way to characterize the different predictors.

Five predictors were developed and compared in this study. The first predictor used whole-window data; the next two used left-window or right-window data. The numbers of neurons in the input, hidden, and output layers of these ANN predictors were 12, 12, and 1, respectively. The output of the fourth predictor, Vote, was determined by the order / disorder assignment made by 2 of the 3 predictors. The fifth predictor, ANN-ANN, was a neural network developed from limited experimentation. This predictor used the output results of the first 3 predictors mentioned above as inputs, had 6 hidden neurons and one output neuron.

2.6 Evaluation of the Predictors

To evaluate the discrimination power of the predictors, receiver operating characteristic (ROC) curves [16] were generated. The curves plot the true positive prediction rates versus the false positive prediction rates for various classification-threshold values over the range from near 0 to near 1.

Results

3.1. Ranking the composition-based attributes

By the ln(odds-ratio) ranking method, the top 10 attributes for ordered and disordered proteins for the composition-based attributes were determined (Table 1). These rankings are grouped by window type: left (L), right (R) or whole (W). Within each group, attributes favoring disorder (rows labeled 1) and those favoring order (rows labeled 0) are listed in order from highest-ranking (columns labeled 1) to lowest-ranking (columns labeled 10). The last row in each group, labeled overall, contains rankings based on $|b|$ without regard to number of residues/attribute, and so indicates the "best overall individual attributes" for each window type.

The higher-ranking attributes in Table 1 show good ability to discriminate order and disorder. As an example RES, the highest-ranking three amino acid attribute from the whole windows, is compared with RDS, the tenth ranked, RHL, the 168th ranked, and KCQ, the 594th ranked (Fig. 1). The y-axis for this graph indicates the probability of disorder given the number of residues belonging to the 3-mer set (x-axis) where

the probability of order, p, is calculated from (2), and the probability of disorder is 1 - p. Obviously, the higher-ranking attributes have better discrimination power than the lower-ranking ones.

Table 1. Ranks (high to low) of order- or disorder- promoting amino acid combinations for each windowing procedure.

Window	Order 0 Disorder 1	1	2	3	4	5	6	7	8	9	10
	1	R	P	K	S	E	Q	M	H	A	--
	0	W	C	Y	F	T	I	V	G	D	L
	1	RP	RE	RQ	RM	RS	ES	EP	KR	RA	RH
L	0	CW	CF	CY	FW	YW	FY	IC	IW	TC	TW
	1	RES	REP	REQ	RSQ	ESQ	RPM	RQP	KRP	KRS	REM
	0	CFW	CYW	CFY	FYW	ICW	TCW	TCY	ICF	ICY	IYW
	1	RES	RESP	RES	KRES	REAS	RDES	REQP	REHS	KRS	REP
	0	CFY	ICFW	TCY	VCF	ICY	CFW	TCFY	TCF	ICFY	IFYW
	overall	W	C	CW	CFW	CYW	CF	Y	CFY	CY	FW
	1	R	P	M	S	E	K	H	Q	N	D
	0	W	Y	C	F	T	G	V	I	L	A
R	1	RM	RP	RS	RH	RE	EP	PM	SP	ES	RW
	0	YW	CY	CW	FY	FW	CF	TY	GY	HY	TW
	1	RSM	REP	RES	RSP	RPM	ESP	RHP	RHM	RHS	RDP
	0	CY	FYW	CFW	CFY	TYW	TCY	HYW	GYW	TFY	GCY
	1	RES	RSP	RES	REP	RDSP	REHP	RHP	RHSP	RDE	KREP
	0	CFY	TCY	TFY	VCY	GCY	HFY	VFY	HCY	TCF	GFY
	overall	YW	W	Y	CYW	CY	CW	FYW	CFY	FY	C
	1	R	P	E	S	M	Q	K	H	A	D
	0	W	Y	C	F	T	I	G	V	N	L
W	1	RM	RS	RP	RE	ES	RH	RQ	EP	RD	SP
	0	CW	YW	CY	FW	CF	FY	TY	TC	TW	TF
	1	RES	RSM	REP	RSP	RPM	RSQ	ESP	RDP	RHP	RDS
	0	CY	FYW	CFW	CFY	TYW	TCY	TCW	TFY	TCY	ICW
	1	RES	RESQ	RSP	RES	RDES	KRES	RDSP	RDEP	RDP	REHS
	0	CFY	TCY	VCF	TCFY	VCY	TFY	CYW	TCF	ICY	VFY
	overall	W	CW	WY	CYW	Y	C	CY	FYW	CFW	CFY

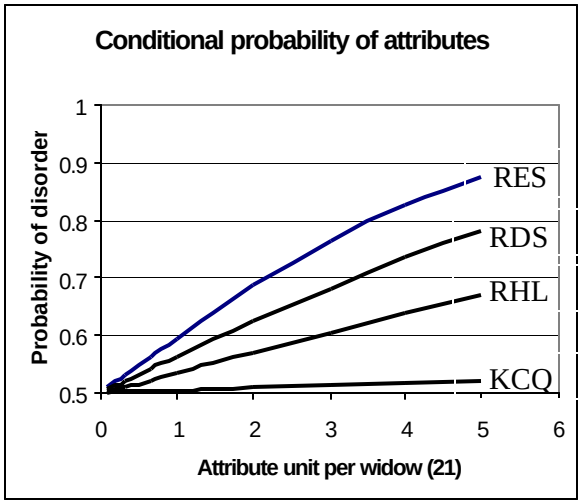

Figure 1: Conditional probability of disorder for attributes of different ranking.

The individual rankings of the property-based attributes were determined by their $|b|$ values using the ln(odds ratio) method of equation 1, and these were merged with the composition-based rankings to give the overall ranking of individual attributes (Table 2). Only one property-based attribute, Coordination Number, is among the 10 best individual attributes for distinguishing order and disorder (Table 2).

Table 2: Ranks of the property-based attributes and their merged ranking with the composition-based attributes for each of the windowing methods

Property Based Attributes Only			**Property and Composition Based Attributes**		
Left window	**Whole window**	**Right window**	**Left window**	**Whole window**	**Right window**
Coordination number	Coordination number	Coordination number	W	W	Coordination number
Flexibility-V	Flexibility-S	Flexibility-S	Coordination number	Coordination number	Y
Flexibility-S	Flexibility-V	Flexibility-V	C	CW	W
Hydropathy	Hydropathy	Hydropathy	CW	YW	Y
Refractivity	Refractivity	Net charge II	CFW	CYW	CYW
Net charge I	Net charge II	Net charge I	CYW	Y	CY
Bulkiness	Net charge I	Refractivity	CF	C	CW
Net-charge II	Bulkiness	Polarity-C	Y	CY	FYW
Volume-C	Volume-C	Polarity-S	CFYW	FYW	CFYW
Volume-J	Volume-J	Bulkiness	CY	CFY	C
Polarity-S	Polarity-C	Volume-C	FW	CFYW	CF
Polarity-C	Polarity-S	Volume-J	YW	FW	CFY
Surface area	Surface area	Surface area	CFY	CFY	FW
EIIP	EIIP	EIIP	FYW	CF	TYW

3.2. Attribute selection

The top 10 attributes in each length class from 2 to 4 for both order and disorder were pooled with the 51 previously used attributes [15] and with the 9 additional property-based attributes, giving a total of 120 attributes as described above. Step-wise logistic regression was carried out for each of the window types. The results are given in Table 3, with the first row for each window type ranking the top 6 (X1 – X6) and the second row ranking the next 6 (X7 – X12), for a total of 12 attributes for each.

Coordination Number was selected as the best overall for all three types of windows (Table 3) even though this attribute did not consistently rank first when the attributes were considered individually (Table 2). This attribute correlates positively with the order. Some selected attributes correlate positively with order and some with disorder.

Table 3: Selected Attributes for ANN predictor training.

Window	X1/X7	X2/X8	X3/X9	X4/X10	X5/X11	X6/X12
Left	Coordination No.	TCFW	RH	Net-Charge	KRSQ	VIFWY
	Flexibility-V	CW	RPM	Flexibility-S	D	ICFY
Right	Coordination No.	GFYW	TCYW	RSP	V	Net Charge
	ASFWY	RDSP	RS	Flexibility-V	Y	RESM
Whole	Coordination No.	RDEP	FYW	Net Charge	V	RSP
	ATRGQSNPDE	TFY	RESQ	VLICFYWPM	KDESPG	TW

3.3. Prediction accuracy

The accuracy of the 5 ANNs developed for predicting ordered and disordered regions are given in Table 4; also included in this table for comparison is our initial predictor based on bng disordered regions [20]; this predictor is herein called XL1. With regard to 5 cross-validation training accuracies, the predictor based on whole windows performed better than those based on left or right, the predictor based on voting gave a similar accuracy, while ANN-ANN appeared to be best.

When applied to out-of-sample ordered data, the 4 new predictors other than ANN-ANN generalized similarly and very well. That is, the decreases in prediction accuracies from training to testing were small, usually less than 4%. Only the ANN-ANN predictor failed to generalize well for prediction of order, showing a drop of more than 13% in accuracy from training to testing. The predictor based on whole windows showed much better generalization for the prediction of order than did the original XL1 predictor, with improvements in the 5% range.

In contrast, generalization for the prediction of disorder is much poorer for all of the predictors, with large drops in accuracy from training to testing being the rule rather than the exception. These large drops in prediction accuracy are discussed below.

Table 4. Accuracies of 6 predictors of protein disorder.

	Predictors					
Data	Left window	Whole window	Right window	Vote	ANN-ANN	XL1
5-cross validation	64.0%	74.3%	70.3%	75.0%	79.1%	73.0%
O_PDB_ Select_25	65.2%	73.0%	68.1%	72.6%	65.2%	67.4%
NRL_3D	66.5%	74.3%	66.0%	73.1%	66.5%	68.3%
D_PDB_ Select_25	60.6%	64.7%	59.6%	63.2%	60.6%	47.8%
NMR Disordered	62.2%	56.7%	61.8%	63.1%	65.3%	58.1%
Far UV CD Disordered	54.7%	48.6%	53.7%	52.6%	56.3%	48.5%

3.4. Evaluation of the predictors using ROC

Applying a predictor to data while varying the threshold yields, a receiver operating characteristic (ROC) curve can be generated [16]. The ROC curves of the five predictors are presented in Fig. 2. Five independent

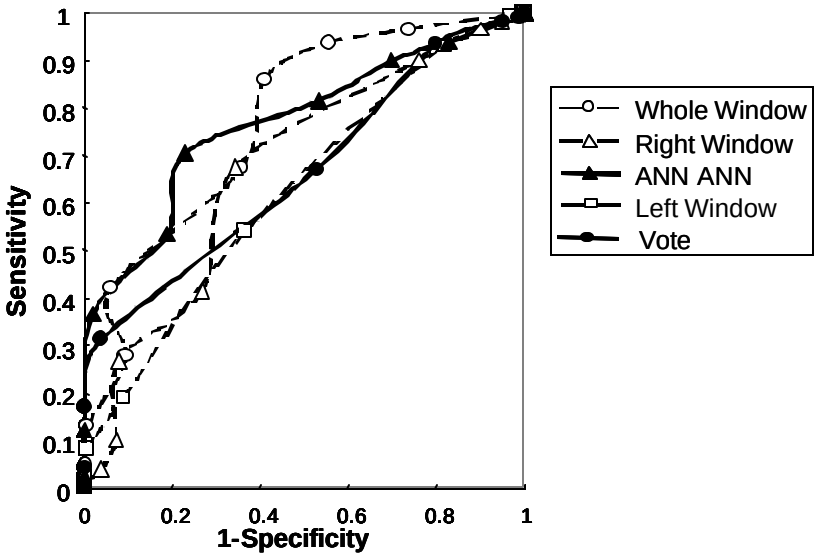

Figure 2. ROC curves for the 5 predictors

sets of calculations on five different sets of data were averaged to yield these curves. Sensitivity is the fraction of true positives predicted by the various methods. Specificity is the fraction of false positives. When specificity is low, the whole window approach has the best sensitivity. When specificity is high, the ANN-ANN procedure has as high sensitivity or better than the whole window method.

Discussion and Conclusions

4.1 Attribute Ranking

Except for Net Charge, every amino acid contributes to each property-based attribute, with the contribution being simply the value of the associated property. In contrast, for composition-based attributes, each amino acid is assigned a value of 1 (present) or 0 (absent), with only a few amino acids contributing to a given window. We expected property-based attributes to be superior to composition-based ones due to the finer grain of the former, but instead the composition-based attributes appear to be generally superior.

In our initial study [20], Flexibility-V, and Hydropathy were both found to be important for discrimination between order and disorder. Of 11 property-based attributes developed since the first study, only 1, Coordination Number, shows a better capacity for discriminating order and disorder than do Flexibility and Hydropathy (Table 2). Evidently Coordination Number depends on side chain size, shape, and surface area in a complex way that relates to a side-chain's capacity to pack with the other side chains. From this view, the very strong correlation of Coordination Number with the ordered state makes good structural sense. We recently carried out an evaluation of 27x attributes; Coordination Number ranked very high with the top ranking attribute being simply an alternative coordination number scale (Williams et al.)

In a previous study to evaluate the importance of 31 individual attributes for distinguishing order and disorder for windows of 21, which is directly comparable to the whole window data reported here, CFYW ranked 1, C ranked 2, FYW ranked 6 and Y ranked 17 [28]. Here the corresponding attributes ranked 11 rather than 1 (CFYW), 7 rather than 2 (C), 9 rather than 6 (FYW) and 6 rather than 17 (Y). There are several contributing factors for these ranking differences. Here we explored more than 6,000 amino acid combinations and residue properties whereas the previous study tested just 31 attributes. Several previously unexplored attributes are among the best, thus shifting the ranks of the attributes that appear in both studies. Second, very different data sets for order and disorder were used in the two studies. Third, different data modeling methods were used to determine the rankings in the two studies.

In the overall rankings of Tables 1 and 2, all of the highest-ranking attributes have strong positive correlations with order; none of the attributes that positively correlate with disorder were among the top. A likely explanation, to be more fully explored below, is that the ordered part of the training data set is much less noisy than the disordered part. In such a case, much stronger correlations would be observed between attribute values and order, and so only attributes that positively correlate with order would be top ranked.

4.2 Attribute Selection

We used a 2-step process for attribute selection. First, different groups of attributes were ranked. Next, a step-wise logistic regression protocol was used on a pooled set of top ranking attributes from each group. A weakness of this approach is obviously that synergy between lower ranking attributes could be missed; however, the current study still involved a much larger attribute pool, 120, than the previously used pools of 24 [20] or 51 [15].

Essentially all of the attributes used in this study are correlated with other attributes to some degree. The step-wise logistic regression model used here to select attributes might not yield the globally optimum set, but in general this method excludes highly correlated pairs or sets of attributes. Thus, the selection protocol yields sets of attributes with information being contributed from each member. We speculate that this approach might be more than a machine-learning protocol and might actually indicate the interplay among the attributes that determines order or disorder.

Although W is the top attribute for two of the three window types and second for the third (Table 2), this attribute is not selected among the top 12 for any of the window types using the logistic regression model. Instead, Coordination Number, which ranks 2, 2 and 1 for the three types of windows (Table 2), is selected as the top attribute for all three (Table 3). W and Coordination Number attributes are correlated in their specification of order/disorder, with W having a larger coordination number than any other residue [8]. Evidently Coordination Number is selected over W because of its superior synergy with the other important attributes.

Property-based attributes other than Coordination Number (e.g. Flexibility-V, Flexibility-S, and Net Charge II) and composition-based attributes that correlate positively with disorder (e.g. RSP, RDSP, RS, RESM, RH, RPM, REDP, RESQ) are all selected. These attributes are among the top 12 in combination with other attributes (Table 3), even though none of these are ranked among the top 10 on an individual basis (Table 2). This result shows that lower ranking attributes are selected by this method because of synergy with other attributes and because of removal of higher ranking, but correlated, attributes.

Flexibility (-V or -S) is the highest-ranking, property-based attribute after coordination number (Table 2). At least one of these is selected for two of the three window types but not the third. That is, Flexibility-V was selected 10^{th} for right windows, 7^{th} for left windows, but not selected for whole windows. However, the composition-based attribute that contains the 10 most flexible residues according to the Flexibility-V scale [25], namely ATRGQSNPDE, was selected 7^{th} for whole windows, so flexibility is represented for all three window types. Unexpectedly, both Flexibility-V and Flexibility-S were selected for one window type, left, at 7^{th} and 10^{th}, respectively. Evidently, these two flexibility scales are not so highly correlated that selection of one excludes the selection of the other in every case.

The next-ranking, property-based attribute, Hydropathy, was not selected for any of the windows. However, the composition-based attribute WFYCVILMP, which was selected for the whole window data, is the set of the amino acids with the highest hydropathy values. Thus, this attribute provides a simplified 1,0 representation of hydropathy.

Several composition-based attributes for individual and pairs of amino acids were selected (Table 2). These include V, RS, and Y at the 5^{th}, 9^{th}, and 11^{th} positions, respectively, for right-window data, RH, CW, and D at the 3^{rd}, 8^{th}, and 11^{th} positions, respectively, for left-windows data, and V and TW at the 5^{th} and 12^{th} positions, respectively, for whole-window data. It is interesting that such singles and pairs are selected as important in competition with the triplet-, quartet-, and property-based attributes.

4.2 Multi-ANN predictor development

In our previous studies, all the attributes for internal regions of sequence were assigned to the centers of the windows, thus including information from both the left and right of the position being predicted. However, a polypeptide chain has a direction, for example from the N- to the C-terminus. Thus, the influence of a given attribute on a given locus might differ on the C- compared to the N-side. In this case, the influence of the attributes would be unbalanced on the two sides of the locus being predicted. For this reason we developed attributes based on Left and Right Windows.

Because of the smaller size (11 residues) of the Left and Right Windows compared to the Whole Windows (21 residues), the Left and Right Window predictors would be expected to have lower accuracies than the Whole Window predictor, just as observed (Table 4). Two additional predictors were developed using combinations of the Left, Right, and Whole Window predictors. One of these, Vote, gave the output specified by 2 of the 3 predictors. The second one, ANN-ANN, used the outputs from the 3 predictors as inputs to a second neural network.

4.3 Comparisons of the Predictors

The training set of the original predictor, XL1, contained just 505 disordered residues balanced with an equal number of ordered residues. The training set used here is larger, but still small, with just 898 disordered residues balanced with ordered data. O_PDB_Select_25 has over 230,000 residues and NRL_3D has over 2,500,000 residues, and yet nearly all of the predictors generalize very well on these much larger datasets, with typically smaller than 4% decreases in accuracy compared to the 5-cross validation accuracies estimated during training. It is remarkable that predictors based on so few amino acids generalize so well. One possible explanation of such a result is that all ordered proteins are quite similar to each other.

The new predictor based on Whole Windows gives the best performance overall, with Vote coming in a very close second. From Table 4, the drop in accuracy from training to prediction on out-of-sample ordered data for the Whole Window predictor is just 1% for O_PDB_Select_25 and nil for NRL_3D, with accuracies of 73 amd 74% for these two datasets, respectively. The ROC curve (Figure 2) for this predictor is a second indicator that it is the best among the 5 new predictors. Compared to the original predictor, XL1, prediction of order on the large out-of-sample datasets shows > 5% improvement (Table 4).

For out-of-sample disorder prediction, comparisons on the D_PDB_Select_25 are the most useful because these disordered regions, like the training data, were characterized by X-ray diffraction. When the prediction accuracies on this database are compared with the accuracy estimated by the 5-cross validation during

training, all of the predictors show large drops, ranging from ~ 10 to ~ 20% (Table 4). Again, the Whole Windows predictor performs the best, with Vote a close second. This new predictor performs almost 17% better than the our first predictor of long disordered regions, here called XL1.

For the NMR- and CD-characterized out-of-sample disorder, prediction accuracies drop even further for the Whole Window predictor. For the other 5 predictors, accuracies on these out-of-sample disordered data are sometimes better but more often worse than the accuracies on D_PDB_Select_25 (Table 4).

Overall, the Whole Window predictor performs slightly better than the 4 other new ones and much better than the originally published one, XL1. Including directional information did not improve the predictions as we had hoped it would.

4.4 Prediction errors

Apparent prediction errors result from two obvious sources: 1. actual errors due to failures of the predictors; and 2. false errors due to miss-classification within the order or disorder data. A preliminary error analysis of the present and earlier predictors provides some insight into these two possibilities.

Several proteins in NRL_3D are predicted to be entirely disordered. Since entirely disordered proteins would not be expected to form crystals, such a result seems anomalous. Examination of several of these anomalous proteins reveals that they exist as co-crystals with DNA, with other proteins, or with other ligands. We have not confirmed that these proteins are actually disordered in the absence of their ligands. However, disorder-to-order transitions upon ligand binding is a common occurrence, and so the existence of disorder in the absence of the ligand seems to be a reasonable supposition.

In other cases, short regions of strong disorder predictions correspond to metal ion or other small co-factor binding sites. Again, it is reasonable to suppose that these segments are disordered in the absence of their obligatory ligands.

Finally, in still other cases regions of actual disorder are found to become ordered by the crystallization process and correspond to contacts between the proteins in the crystal. Such examples are just special cases of disorder-to-order transitions upon binding.

The finding of likely disorder in the ordered datasets in the examples given above suggests that the prediction error rate on ordered proteins is actually lower than that reported in Table 3. However, so far we have not developed a general strategy for finding such miss-classifications, and the study of examples one-by-one is simply too time consuming and expensive. At this time it is unclear how much the error rate is affected by this miss-classification of disorder as order, but the effects cannot be very large because the error rate for prediction on ordered data is already fairly low. Clearly, more work on this problem is needed.

Miss-classification of order as disorder is a much larger problem than the reverse as discussed above. Such miss-classification might be a significant cause of the lower success in predicting disorder (Table 4). For example, disorder identified by missing coordinates in X-ray structures might actually be the result of domain wobble [9] rather the result of the existence of a dynamic structural ensemble. This would lead to significant miss-classification of order as disorder and would thus yield significant numbers of false errors.

The indication of disorder by CD spectroscopy relies on a global estimate of folding with no positional information. Therefore, disorder based on this measure seems especially prone to miss-classification of order as disorder.

Compared to X-ray- or CD-characterized disorder, NMR-characterized disorder would seem to be less subject to uncertainties and so would seem to provide very unambiguous disorder data. However, examination of predictions of order in regions of NMR-characterized segments of disorder reveal a correlation with ligand binding sites [10]. Evidently, some NMR-characterized disordered regions can be involved in disorder-to-order transitions upon binding to specific ligands. In this circumstance, prediction of order in a segment of NMR-characterized disorder might not be a true error after all.

The indication of ligand binding sites by prediction of disorder in an ordered protein and by prediction of order in a disordered protein seems to be a contradiction. Further examination shows that the ligand binding by locally disordered regions typically involves the folding of the protein to fit complex surfaces or to even surround the ligand. Thus, disorder in the unbound state helps to solve steric problems associated with complex formation. In contrast, binding by a local region of a disordered protein often involves the formation of structure by fitting into a groove of the partner. In this case, a local region with a high tendency for order would be appropriate for the formation of such complexes. Thus, the indication of ligand binding sites by disorder prediction in ordered proteins and by order prediction in disordered proteins has a reasonable explanation.

Of course the second source of error is simply prediction mistakes. This might be a greater problem for disordered regions compared to ordered ones. That is, our previous work suggests that, while ordered sequences seem to occupy a more local, more specific region of attribute space, disorder may occupy a much more extensive region of this space, with various local regions corresponding to different "flavors" of disorder [22]. Another way of saying this is that compositional bias can indicate order or disorder, with the type of ordered structure largely determined by the specific sequence of the residues, but with the flavors of the disorder largely determined by the type of compositional bias. If this is indeed true, then the small sample size used to train our predictors is a much greater problem for the prediction of disorder as compared to the prediction of order. This is consistent with the results displayed in Table 3.

4.5 Future Directions

Two themes for future research are suggested from the discussion of prediction errors. First, strategies are needed, other than one-by-one study of examples, to reveal order / disorder miss-classification. Second, the flavor concept for disorder needs to be more fully tested and utilized. If the concept of disorder flavors is true, then classification by flavors could help to reduce miss-classification of order as disorder. Further, it is likely that flavor-specific predictors of disorder would be superior to one global predictor within a particular flavor domain. In this case, the long-term goal would be to identify the set of disorder flavors used in nature and then develop predictors for each one.

Acknowledgements

Support from NSF research grant CSE-IIS9711532 and NIH research grant RO1-LM06916 to ZO and AKD is gratefully acknowledged.

References

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

[2] Anfinsen, C.B., Principles that govern the folding of protein chains, Science, 181: 223-230, 1973.

[3] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., The protein data bank, Nucleic Acids Res., 28: 235-242, 2000.

[4] Chechetkin, V.R. and Lobzin, V.V., Characterization and comparison of protein structures. Part I-characterization, J. Theor. Biol., 198: 197-218, 1999.

[5] Chothia, C., The nature of accessible and buried surfaces in proteins, J. Mol. Biol., 105: 1-14, 1975.

[6] Dunker, A., Obradovic, Z., Romero, P., Kissinger, C., and Villafranca, E., On the importance of being disordered, PDB Newsletter, 81: 3-5, 1997.

[7] Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J.E., Protein disorder and the evolution of molecular recognition: theory, predictions and observations, Pacific Symp. Biocomputing, 3: 473-484, 1998.

[8] Galaktionov, S.G. and Marshall, G.R., Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D, presented at Fourth International Conference on Computational Biology, St. Louis, MO, 1996.

[9] Garner, E., Cannon, P., Romero, P., Obradovic, Z., and Dunker, A.K., Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization, Genome Informatics, 9: 201-213, 1998.

[10] Garner, E., Romero, P., Dunker, A.K., Brown, C., and Obradovic, Z., Predicting binding regions within disordered proteins, Genome Informatics, 10: 41-50, 1999.

[11] Hobohm, U., Scharf, M., Schneider, R., and Sander, C., Selection of representative protein data sets., Protein Sci., 1: 409-417, 1992.

[12] Jones, D.D., Amino acid properties and side chain orientation in proteins: A cross correlation approach., J. Theor. Biol., 50: 167-183, 1975.

[13] Kyte, J. and Doolittle, R.F., A simple method for displaying the hydropathic character of a protein, J. Mol. Biol., 157: 105-132, 1982.

[14] Lazovic, J., Selection of amino acid parameters for Fourier transform-based analysis of proteins, Comput. Appl. Biosci., 12: 553-562, 1996.

[15] Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z., Predicting protein disorder for N-, C-, and internal regions, Genome Informatics, 10: 30-40, 1999.

[16] Metz, C.E., Basic Principle of ROC analysis, Seminars in Nucleic Medicine, 8: 283-298, 1978.

[17] Pattabiraman, N., Namboodiri, K., Lowrey, A., and 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: 387-405, 1990.

[18] Romero, P., Obradovic, Z., and Dunker, A.K., Sequence data analysis for long disordered regions prediction in the calcineurin family, Genome Informatics, 8: 110-124, 1997.

[19] Romero, P., Obradovic, Z., and Dunker, A.K., Folding minimal sequences: the lower bound for sequence complexity of globular proteins., FEBS Lett., 462: 363-367, 1999.

[20] Romero, P., Obradovic, Z., Kissinger, C.R., Villafranca, J.E., and 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., and Dunker, A.K., Thousands of proteins likely to have long disordered regions, Pacific Symp. Biocomputing, 3: 437-448, 1998.

[22] Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K., The Complexity of Disorder, Proteins: Struc., Funct., Gen., 42:38-48, 2001.

[23] Schneider, G. and Wrede, P., Development of artificial neural filters for pattern recognition in protein sequences, J. of Molecular Evolution, 36: 586-595, 1993.

[24] Veljkovic, V., Cosic, I., Dimitrijevic, B., and Lalovic, D., Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?, IEEE Trans. Biomed. Eng., 32: 337-341, 1985.

[25] Vihinen, M., Torkkila, E., and Riikonen, P., Accuracy of protein flexibility predictions, Proteins, 19: 141-149, 1994.

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

[27] Wright, P.E. and Dyson, H.J., Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, J. Mol. Biol., 293: 321-331, 1999.

[28] Xie, Q., Arnold, G.E., Romero, P., Obradovic, Z., Garner, E., and Dunker, A.K., The sequence attribute method for determining relationships between sequence and protein disorder, Genome Info., 9: 193-200, 1998.