Methods for Improving Protein Disorder Prediction

Slobodan Vucetic[1], Predrag Radivojac[3], Zoran Obradovic[3], Celeste J. Brown[2] and A. Keith Dunker[2]

[1] School of Electrical Engineering and Computer Science,
[2] Department of Biochemistry and Biophysics
Washington State University, Pullman, WA 99164
[3] Center for Information Science and Technology
Temple University, Philadelphia, PA 19122

Abstract

In this paper we propose several methods for improving prediction of protein disorder. These include attribute construction from protein sequence, choice of classifier and postprocessing. While ensembles of neural networks achieved the higher accuracy, the difference as compared to logistic regression classifiers was smaller then 1%. Bagging of neural networks, where moving averages over windows of length 61 were used for attribute construction, combined with postprocessing by averaging predictions over windows of length 81 resulted in 82.6% accuracy for a larger set of ordered and disordered proteins than used previously. This result was a significant improvement over previous methodology, which gave an accuracy of 70.2%. Moreover, unlike the previous methodology, the modified attribute construction allowed prediction at protein ends.

1 Introduction

Each protein is defined by its sequence of amino acids, corresponding to an alphabet of twenty symbols. Proteins fold into specific three-dimensional structures, with the folding information residing in their amino acid sequences [1]. Since biological function follows from protein 3-D structure, determining this structure is important for activities ranging from basic understanding to drug discovery. This sequence-to-structure-to-function paradigm is serving as the basis for the recently launched structural genomics initiative.

In contrast to the view that prior 3-D structure is necessary for function, several proteins have been shown to carry out function by means of regions that are incompletely folded or that are even completely unfolded, suggesting the need for a reassessment of the sequence-to-structure-to-function paradigm [17]. Thus, we recently proposed The Protein Trinity to explicitly link protein function with intrinsic disorder. In this view, a whole protein or a segment can exist in one of a trinity of structures, called the ordered state, the molten globule and the random coil, with function arising from any member of the trinity or from structural transitions between the members [6]. According to The Protein Trinity, intrinsic protein disorder rivals the importance of 3-D structure.

Starting from the assumption that amino acid sequence determines structure, we proposed that sequence also determines intrinsic disorder as well. To test this, we developed network predictors of order and disorder. The predictions of disorder, with ~ 70% accuracy compared to the 50% expected by chance for the balanced datasets, supported our suggestion that sequence determines intrinsic disorder [15]. Here the goal is to improve predictions of intrinsic disorder.

The first challenge is collecting an appropriate database of disordered proteins. Existing protein structural databases are strongly biased against disorder. As a result, in the previous work [14] just 32 proteins with disorder longer then 40 amino acids were available. Here, about 110 more disordered proteins were added. The next challenge is to represent the protein sequence appropriately. In previous work [15], attributes at a given position in the sequence were derived from a subsequence including 20 neighboring amino acids (with a window of size $W = 21$). Ten positions from the protein termini were not represented in the data set. In this study, we examine a range of possible window sizes without discarding information from the protein termini. We also enhanced the attribute set with a measure of entropy since there is evidence that low complexity amino acid sequences tend to be disordered [16].

The third challenge is finding an appropriate method of classification. Predicting disorder is a binary classification problem with order/disorder decisions. Neural network predictors represent highly nonlinear functions, and they are a natural choice if the discriminant function is expected to be nonlinear. However, there is some evidence [10] that linear classifiers achieve comparable accuracy to neural networks when the same set of attributes is used. Bagging [3] and boosting [7] are ensemble methods for the aggregation of unstable predictors known to be able to significantly improve accuracy as compared to single predictors. In this study, we compare different predictors in order to find the best choice for disorder prediction.

Finally, we are interested in prediction of disorder for regions longer then 40 consecutive amino acids, where post-processing can further improve prediction accuracy. Since neighboring amino acids are likely to be part of the same ordered/disordered region, we exploit this notion by filtering raw predictions using an output window of a proper length. In the Methods section, we explain in detail all proposed methods for improvement of protein disorder prediction. In the Experimental Results section, we explain the setup for building a disorder predictor and validating the results, and we present our results.

2 Methods

In this section, we describe the data sets used for building predictors of protein disorder, the representation of sequences suited for learning, and the various classification procedures for protein disorder prediction.

2.1 Data sets

The data set of disordered protein sequences is derived from a previously described data set [16]. Briefly, disordered regions identified by NMR, circular dichroism or protease digestion were found by keyword searches of PubMed (http://www.ncbi.nlm.nih.gov). Starting from a subset of the Protein Data Bank (PDB) called PDB_Select_25 [9], we identified disordered regions in X-ray crystal structures by searching for residues having backbone atoms that are absent from the ATOMS lists in their PDB files [15]. PDB_Select_25 contains a single representative structure from PDB for protein families whose members have greater than 25% sequence identity.

The disordered data set used in this study contains 145 of the 150 protein segments previously reported in [16]. The 5 eliminated sequences contained segments of disorder that were too short by the criterion used here. The remaining 145 segments contain a total of 16,705 disordered residues. Some of these disordered regions come from proteins that are completely disordered, and some come from proteins that also have ordered domains. The data set that includes the entire protein sequence for each protein is called the complete D_145 data set. To provide a representative data set of ordered proteins for training the disorder predictors, O_130 was used. This data set consists of 130 non-redundant proteins that are completely ordered from their N- to C-termini with a total of 32,506 residues. This data set was also extracted from PDB_Select_25 and therefore contains sequences with less than 25% sequence identity.

2.2 Data representation and attribute selection

In this study, order/disorder properties at a given position in a sequence are predicted using a subsequence of neighboring amino acids within a window of size W_{in}. In accord with previous work [15] and to prevent the "curse of dimensionality" [2], only first-order statistics of the 20 amino acids within a given window were used as the first within a given window were used as the first 20 attributes. For example, the value of attribute x_{ia} at position i within a sequence of length N is calculated as the fraction of amino acid a, $a \in$ {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W. Y}, within a window of length W_{in},

$$x_{ia} = \frac{1}{w''-w'+1} \sum_{j=i-w'}^{i+w''} \delta_{ja} \ ,$$

where $w' = \max(1,\ i-W_{in})$, $w'' = \min(N,\ i+W_{in})$, $\delta_{ja} = 1$ if amino acid a is at position j, and $\delta_{ja} = 0$ otherwise. This approach differs slightly from our previous approach where $W_{in}/2$ positions at the ends of each protein were not represented due to enforcing a fixed length W_{in} [14].

This simplistic approach is motivated by results known about the incompressibility of protein sequence [11], showing that very little information can be extracted from protein sequences beyond first order statistics. It is likely that useful information exists outside the window (some amino acids can be close in space but not close in sequence), but it is still an open question how to infer this from the protein sequence. Such simple knowledge representation is also favored because the small size of the training set may lead to problems resulting from the curse of dimensionality.

Flexibility, Hydropathy and Coordination Number are numerical attributes specific to each amino acid and are thought to be useful indicators of disorder [14]. We used only the Flexibility attribute x_{iFlex}, calculated as

$$x_{iFlex} = \frac{1}{w''-w'+1} \sum_{j=i-w'}^{i+w''} Flex(j) \ ,$$

where $Flex(j)$ is the flexibility value for amino acid at position j. Corresponding attributes x_{iHydro} for Hydropathy and x_{iCoord} for Coordination Number were not included since their the correlation to Flexibility is larger then 0.7. Therefore, they would add little to the overall prediction accuracy, and could cause collinearity problems [4].

Finally, it has been shown that sequence complexity is relatively highly correlated with disorder, i.e. low complexity protein sequences are likely to be disordered [16]. Therefore, we used the measure of sequence complexity called K2 entropy as the final attribute. When measured over a window, K2 entropy attribute at position i is given by

$$x_{iK2} = -\sum_{a=1}^{20} x_{ia} \log_2 x_{ia} \ ,$$

where a is the symbol for one of the 20 amino acids. Thus, this formula calculates complexity as the number of bits needed to encode a given protein segment. So, for an alphabet of 20 letters, $0 \le x_{iK2} \le \log_2(20) \approx 4.3$ bits.

So far we have constructed a set of 22 attributes thought to be useful for disorder prediction. The next question is

whether attribute reduction or transformation can result in a smaller, but still effective set of attributes. The protein incompressibility result indicates that, statistically, as a first approximation each protein sequence can be considered as a random sample from the underlying distribution. Since the 20 amino acid attributes represent their fractions within a window, $\sum x_{ia} = 1$, and 20-dimensional vector $\{x_{ia}\}$ can be considered as a random sample from a 20-dimensional multinomial distribution. As a consequence, knowledge of 19 amino acid attributes uniquely determines the remaining amino acid attribute, and an arbitrary attribute should be excluded from the data set to obtain a nonredundant set of amino acid attributes.

Additionally, Flexibility x_{Flex} is a linear combination of amino acid attributes. Therefore, another amino acid attribute should be eliminated from the data set of attributes. Note that Entropy x_{K2} is a nonlinear combination of amino acid attributes, and it would not cause collinearity problems. In this study, we (rather arbitrarily) excluded amino acid attributes x_D and x_F. Therefore, this resulted in the set DD_145 with 16,705 disordered 20-dimensional data points, and the set OD_130 with 32,506 ordered 20-dimensional data points. These data sets were used in building the protein disorder predictors.

Regarding the choice of window length W_{in}, it is not clear which length is optimal for disorder prediction. A similar problem occurs in predicting protein secondary structure. Qian and Sejnowski [13], using slightly different sequence representation, showed that a window size of 13 resulted in the best accuracy for this problem. However, it should be noted that the length of secondary structures (α-helices, β-strands, coils) could be just several amino acids long. Since we are examining long regions of protein disorder (>40 amino acids), it is likely that longer windows than in our previous studies (with $W_{in} = 21$) would be better. One goal of our study is to determine the optimal length of W_{in}.

An interesting question regarding the DD_145 disordered data set with 16,705 examples, and OD_130 ordered data set with 32,506 examples is what is the information contained in them. The sampling theorem (e.g., [12]) claims that a time signal with spectral bandwidth F can be perfectly reconstructed by its sampling at each $1/2F$ time units. It should be noted that amino acid attributes x_a can be approximated by low-pass filtering of a binary signal $\{\delta_{ia}\}$ to a signal with spectral bandwidth $\sim 1/W_{in}$, where $\delta_{ia} = 1$ if amino acid a is at position i, and $\delta_{ia} = 0$ otherwise. Thus, it can be stated that a data set containing 20-dimensional data points from each $W_{in}/2$-th position within a sequence would contain approximately the same information as the full data set with data points taken from each sequence position. As a result, the effective size of DD_145 is $2*16{,}705/W_{in}$, while the effective size of OD_130 is $2*32{,}506/W_{in}$.

2.3 Accuracy measures and building predictors of protein disorder

Before building a disorder predictor, a reasonable performance measure must be defined. Previous pessimistic estimates [5] on the commonness of disorder in proteins indicated that at least 11% of all the residues in the Swiss Protein data bank belong to disordered regions of length 40 or longer, while this estimate is much higher in some eucaryotic genomes (up to 30%). Since ordered protein sequences outnumber examples of protein disorder, we report separately on the percent of true negatives (TN, ordered positions predicted to be ordered) and true positives (TP, disordered positions predicted to be disordered). However, for convenience, we also report on the average accuracy between order and disorder where an accuracy above 50% indicates that the predictor is better then a random guess.

Since proteins and their disordered regions can be of very different lengths we measure accuracy on each protein sequence, and report an average TN and TP rate over a set of ordered and disordered proteins. Occasionally, we will also use a standard accuracy measured as an average TN and TP rate over available examples of order and disorder. Protein disorder predictors can be useful for a range of applications starting from aiding protein crystallization to measuring the commonness of protein disorder, where costs of false positive and false negative prediction can differ. A popular accuracy measure suited for applications where costs of misclassification are not well defined is the ROC curve.

The ROC curve assumes that classifiers give real-valued predictions, where the user can chose the decision threshold. If the problem is binary classification and the two classes are encoded as 0 and 1, the threshold $\theta = 0.5$ is assumed to be the default value. However, depending on application some other threshold can be suitable. The ROC curve is a plot obtained by changing the value of the threshold and measuring true positive (TP) against false negative (FN) prediction rates. A predictor can be regarded as successful if its ROC curve is high.

In this study we examine logistic regression (LR) (e.g, [4]) and neural networks (NN) to build predictors of protein disorder. LR is a linear predictor suited for binary classification requiring an iterative parameter fitting procedure. NN are known as universal approximators, capable of representing highly nonlinear relationships. They require abundant data, have relatively slow training, and give unstable solutions sensitive to the initial conditions and training procedures.

In this study we also examined bagging [3] and boosting [7] which are popular methods for combining predictors known to reduce prediction variance while retaining a small bias of unstable predictors. In bagging, an ensemble of predictors is learned on bootstrap samples of the training data, and a

classification is a majority vote from the ensemble. In boosting, an ensemble of predictors is constructed so that subsequent predictors are trained on samples biased towards examples where previous predictors failed. Both methods are known to significantly improve the accuracy of neural networks on a large range of problems, while boosting appears to achieve better accuracy on the majority of benchmark machine learning data sets.

Regardless of the applied prediction algorithm, we propose to perform a post-processing step on the raw predictions. Since we are interested in predictions of long protein disorder/order regions, it is to be expected that neighboring amino acids belong to the same disordered/ordered region. However, raw predictions allow cases where, for example, the region predicted to be overly disordered has scattered positions predicted to be ordered. These scattered positions are likely to be misclassified. Therefore, we propose to use the output window of length W_{out} to smooth-out the raw predictions and to correct occasional predictions that differ from neighboring predictions as

$$z_i = \frac{1}{w''-w'+1} \sum_{j=i-w'}^{i+w''} y_j$$

where $\{y_i\}$ are raw predictions, $\{z_i\}$ are smoothed-out predictions, $w' = \max(1, i-W_{out})$, $w'' = \min(N, i+W_{out})$, and N is the sequence length.

3 Experiments

3.1 Experimental parameters

In all the experiments we used balanced data sets with the same number of examples of order and disorder. Since moving averaging over window of length W_{in} for attribute construction effectively reduces the information contained in the data sets OD_130 and DD_145, all training sets used in this study had 1,000 examples of order and 1,000 examples of disorder randomly chosen from available examples of order and disorder.

We used cross-validation to test accuracy in the following way. The D_145 set of proteins was randomly partitioned into 15 subsets, while the O_130 set was randomly partitioned into 13 subsets. Then, a predictor was trained using 1,000 random disorder examples from 14 disorder subsets, and 1,000 random order examples from OD_130. The accuracy of this predictor was measured on the remaining disorder subset. This procedure was repeated 15 times each time using different disorder subset for testing. Similarly, 13 predictors were trained using 2,000 random disorder examples from DD_145, and 2,000 random order examples from 12 order subsets. The accuracy of these predictors was measured on the corresponding remaining order subset. In this way, by learning 15+13 predictors it was possible to test the accuracy on the complete OD_130 and DD_145 data sets.

We examined different combinations of windows W_{in}, W_{out} with lengths taken from the set {1, 9, 21, 41, 61, 81, 121}, except that $W_{in} = 1$ was not examined. For the majority of proteins in the database, the length of 121 covered all or nearly all of their disordered regions. Only 6 of 145 available disordered regions were longer then 200 amino acids, with one region being 1,800 amino acids long. Note that W_{in} longer than the protein length corresponds to representing a whole protein with only one data point, so resolution is lost. Similarly, W_{out} longer then the protein length also corresponds to loss of resolution, giving an average prediction for the whole protein.

For each length W_{in}, we performed one run of cross-validation with logistic regression predictors, 10 runs with neural network predictors, one run with a bagging ensemble of 10 neural networks, and one run with a boosting ensemble of 30 neural networks. Neural networks used in this study had one hidden layer with 5 hidden nodes and a sigmoid output node. A relatively small number of 5 hidden nodes was proper for the relatively small available data sets. One hundred iterations of resilient backpropagation were preformed to train the neural networks, since this number was determined to be sufficient for successful training without significant overfit.

3.2 Experimental results

In Table 1 we compare predictions of logistic regression (LR), neural networks (NN), bagging, and boosting for $W_{in} \in \{9, 21, 41, 61, 81, 121\}$ with $W_{out} = 1$. True positive (TP), true negative (TN), and average (TN+TP)/2 accuracy are given for each predictor. As seen, a surprising result is that LR is superior to NN for all examined W_{in} lengths. While ensembles of neural networks were superior to individual NN, they appeared to achieve only up to 1% better accuracy then LR. The difference between bagging and boosting appeared to be small. These results indicate that the discriminant function between order and disorder can be considered as approximately linear.

Another result is the increased accuracy of all the models with the increase of W_{in} that appears to level out at $W_{in} = 81$. We note that NN learned with $W_{in} = 21$ and without using the entropy attribute corresponds to our previous predictors [14, 16]. An average accuracy of 10 such NN predictors was 70.2% (TP = 67.8%, TN = 73.6%). Interestingly, the true positive accuracy (accuracy on disordered protein regions) was considerably lower than the true negative accuracy (accuracy on ordered protein regions), although all predictors were trained on balanced data sets. This phenomenon supports our hypothesis that ordered sequences occupy a more compact region of the attribute space. As a consequence, the decision boundary is expected to be more biased towards ordered examples, which are less scattered.

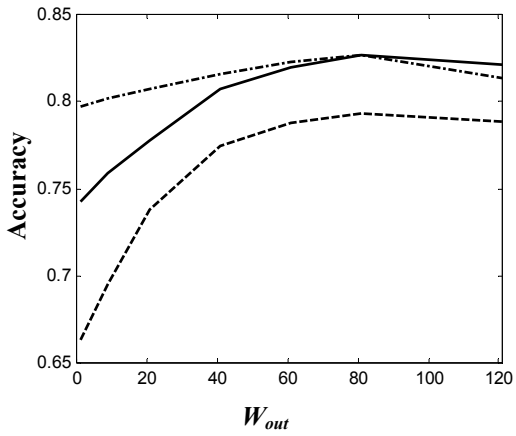

Figure 1. *Results for bagging predictors comparing (TP+TN)/2 accuracies averaged over proteins for W_{in} = {9: dashed, 21: solid, 61: dash-dotted} over a range of W_{out} = {1, 9, 21, 41, 61, 81, 121}.*

Table 1. *Accuracies of each predictor of protein disorder measured as the average accuracy over all data sets*

Model	W_{in}	Accuracy		
		TN	TP	Average
LR	9	68.9	63.3	66.1
	21	76.0	68.6	72.3
	41	**79.7**	**69.9**	**73.5**
	61	81.2	75.9	78.6
	81	83.0	76.8	79.9
	121	83.3	75.6	79.4
NN	21	74.1±1.0	67.6±0.7	70.8
	41	**79.2±1.3**	**72.5±1.4**	**75.8**
	81	82.0±1.2	76.0±1.4	79.0
Bagging	9	69.7	60.9	65.3
	21	78.0	68.5	73.3
	41	**81.4**	**72.8**	**77.1**
	61	82.8	74.5	78.7
	81	83.2	78.5	80.9
	121	82.7	77.8	80.3
Boosting	**41**	**81.5**	**73.1**	**77.3**

In Figure 1 we present average accuracies of the bagging predictors for different choices of (W_{in}, W_{out}) pairs. Results of a similar nature were obtained for LR and NN, but are not included due to lack of space. In Table 2, we report on the best accuracy results for different choices of bagging predictors and W_{in}. We report both on the accuracy averaged over proteins and on the accuracy averaged over amino acids. Note that the accuracy averaged over proteins is slightly lower. This is mainly caused by one protein with a 1,800 residue disordered region, that composes more than 10% of the *D*_145 amino acids. The accuracy on this protein increases as the lengths W_{in} and W_{out} increase for each of the predictors. This protein biases the accuracy results averaged over amino acids slightly upwards for large window lengths.

Table 2. *The best accuracies for the bagging predictor and different W_{in}. We report accuracies averaged by protein and the accuracies averaged by residue. W_{out}^* is the optimal length for a given W_{in}.*

W_{in}	W_{out}^*	Accuracy by protein			Accuracy by residue
		TN	TP	Average	
9	81	93.5	65.2	79.3	81.1
21	81	93.5	71.5	82.5	84.3
41	81	90.3	73.7	82.0	84.5
61	81	88.8	76.5	82.6	85.3
81	61	86.1	77.9	82.0	85.3
121	61	85.3	76.8	81.0	85.4

The main observation from Figure 1 and Table 2 is that, for each choice of W_{in}, the accuracy can be significantly improved by increasing W_{out}. Another result is that the output window with length W_{out} = 81 or 61 seems to be the optimal value for all choices of W_{in}, and that the accuracies achieved by these lengths are similar over a range W_{in} lengths. This is a rather interesting result as compared to Table 1, where the choice of W_{in} seemed to be very important for accuracy.

In Figure 2, we compare ROC curves for bagging neural networks with two sets of window lengths: (W_{in} = 21, W_{out} = 1) corresponding to our previous predictors, and (W_{in} = 61, W_{out} = 81) corresponding to the best predictor from our experiments. The accuracies were measured by averaging over residuals. Since ROC curves depict the accuracy for different choices of threshold, they can be very useful for determining the optimal threshold given misclassification cost. If we are interested in accuracy (TP+TN)/2, the optimal predictor can be obtained by threshold 0.41 which achieves 86.5% accuracy as compared to 85.3% accuracy with threshold 0.5 (see Table 2).

In our previous work we noted that the accuracy of predicting disorder is stronger at protein ends then elsewhere within a sequence [10]. We wanted to examine to determine if the same effect occurs for our new predictors. We com-

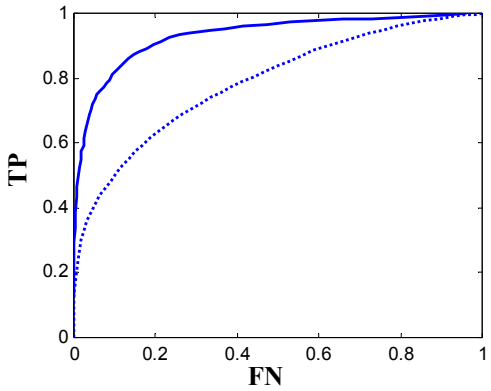

Figure 2. *ROC curves for bagging prediction with (W_{in} = 21, W_{out} = 1, dotted) and (W_{in} = 61, W_{out} = 81, solid).*

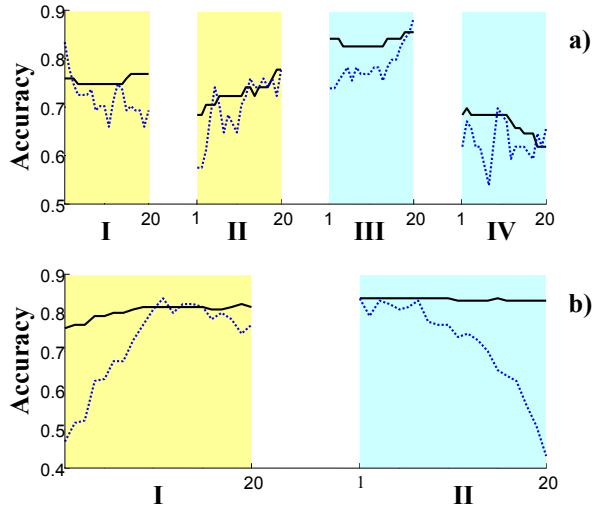

Figure 3. *Comparing prediction accuracies of bagging with (W_{in}=21, W_{out}=1, dashed lines) and (W_{in}=61, W_{out}=81, solid lines) near order/disorder boundaries and at protein ends. a) Comparison on D_145 proteins. We show averaged accuracies of the first 20 positions of 91 disorder regions that start at the beginning of protein sequence (Region I) and 54 disordered regions that do not start at the beginning of protein sequence (Region II). We also show averaged accuracies of the last 20 positions of 76 disordered regions that do not end at the end of protein sequence (Region III) and 69 disorder regions that end at the end of protein sequence (Region IV). b) Comparison of accuracies at the first 20 (Region I) and last 20 (Region II) positions of O_130 proteins.*

pared bagging predictors with two sets of window lengths: (W_{in} = 21, W_{out} = 1) and (W_{in} = 61, W_{out} = 81), and the results are presented in Figure 3. We measured the accuracy of predicting disorder at the first and last 20 positions in each disordered region. In Figure 3, two cases are examined: (i) disordered regions are located at one or both termini of a protein, and (ii) disordered regions start or end within a protein sequence. We also measured accuracies at the first and last 20 positions in ordered proteins. The new predictors significantly improve accuracy in all cases examined, however, the highest improvement is in the prediction of ordered proteins termini (Figure 3b).

4 Conclusions

By studying various-sized input and output windows, by altering data representations, and by postprocessing, we have improved the prediction accuracy of order and disorder from about 70% to about 83%, while at the same time adding the capacity to predict to the ends of proteins. At the same time, averaging over such large windows as used here inevitably leads to loss of resolution. For some applications, less accuracy but higher resolution might be useful. For example, we showed previously that short predictions of order within long regions of disorder are indicative of binding sites [8]. Thus, combined use of the current, more accurate predictors with less accurate ones developed previously may provide a useful strategy for the initial characterization of putative disorder.

5 Acknowledgements

Support from NSF-CSE-IIS-9711532 to Z.O. and A.K.D. and from N.I.H. 1R01 LM06916 to A.K.D. and Z.O is gratefully acknowledged.

6 References

[1] Anfinsen, C.B., Principles that govern the folding of protein chains, *Science*, **181**, 223-230, 1973.
[2] Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.
[3] Brieman, L., Bagging predictors, *Machine Learning*, **24**, 123-140, 1996.
[4] Davidson, R., and MacKinnon, J., *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993.
[5] Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., and Brown, C.J., Intrinsic protein disorder in complete genomes. *Genome Informatics,* **11**, 161-171, 2000.
[6] Dunker, A.K., Romero, P., Lawson, D., Oh, J., Oldfield, C., Campen, A., Ratliff, C., Hipps, K.W., Ausio, J., Nissen, M., Reeves, R., Kang, C., Kissinger, C., Bailey, R., Griswold, M., Brown, C., Chiu, W., Garner, E., and Obradovic, Z., Intrinsically disordered protein, *J. Mol. Graphics and Modeling*, **19**, 26-59, 2001.
[7] Freund, Y., and Schapire, R.E., Experiments with a new boosting algorithm, *Proc. Thirteenth International Conference on Machine Learning*, 148-156, 1996.
[8] Garner, E., Romero, P., Dunker, A. K., and Obradovic, Z. Predicting Binding Regions Within Disordered Proteins, *Genome Informatics*, **10**, 41-50, 1999.
[9] Hobohm, U., and Sander, C., Enlarged representative set of protein structures. *Protein Sci.,* **3**, 522-524, 1994.
[10] 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.
[11] Nevill-Manning, C.G., and Witten, I.H., Protein is incompressible, *Proc. Data Compression Conference*, Snowbird, Utah, 257-26, 1999.
[12] Proakis, J.G., and Manolakis, D.G., *Digital Signal Processing: Principles, Algorithms, and Applications*, 3rd ed., Prentice Hall, 1996.
[13] Qian, N. and Sejnowski, T. J., Predicting the secondary structure of globular proteins using neural network models, *J. Mol.Biol.*, **202**, 865-884, 1988.
[14] Romero, P., Obradovic, Z., and Dunker, A.K., Intelligent data analysis for protein disorder prediction. *Artificial Intelligence Reviews,* **14**, 447-484, 2001.
[15] Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., and Dunker, A. K., Identifying disordered regions in proteins from amino acid sequences. *Proc. IEEE International Conference on Neural Networks,* **1**, 90-95, 1997.
[16] Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K., Sequence complexity of disordered protein. *Proteins: Struc., Funct., Gen.,* **42**, 38-48, 2001.
[17] Wright, P.E. and Dyson, H.J., Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J. Mol. Biol.*, **293**, 321-331, 1999.