A Neural Network for Predicting Protein Disorder using Amino Acid Hydropathy Values

Deborah A. Stoffer and L. Gwenn Volkert
Department of Computer Science
Kent State University
Kent, OH 44242, USA
Email: dstoffer@cs.kent.edu and volkert@cs.kent.edu

***Abstract*— Proteins have been discovered to contain ordered regions and disordered regions, where ordered regions have a defined three-dimensional (3D) structure and disordered regions do not. While in the past it was believed that proteins only function in a defined 3D structure, proteins with disordered regions have been discovered to have at least 28 distinct functions. It is now important to be able to determine the ordered and disordered regions in proteins. Several experimental techniques such as X-ray crystallography, NMR spectroscopy, circular dichroism, protease digestion, and Stokes radius determination, along with several computational techniques such as artificial neural networks (ANNs), support vector machines (SVMs), logistic regression, and discriminant analysis have so far been used to detect disordered proteins. Past research has shown that ANNs and amino acid properties are an effective tool at predicting protein disorder. This research uses a feed-forward neural network implemented using JavaNNS and the hydropathy values of amino acids to predict protein disorder. The results show that hydropathy is an important amino acid property for disorder.**

I. Introduction

In the dilemma of predicting protein function, it has generally been assumed that the three-dimensional (3D) structure of a protein determines its function. While this is accurate for some proteins [1], it does not hold for all proteins, as some proteins have been found to function without a defined 3D structure [2]–[5]. These proteins are referred to as natively unfolded [6], intrinsically unstructured [2], [3], or intrinsically disordered [5], [7]–[9]. Proteins that fold into a defined 3D shape to function are correspondingly referred to as folded, structured, or ordered. Disordered proteins can either be locally disordered, composed of regions that are disordered and other regions that are ordered, or globally disordered (e.g. where the entire protein does not have a defined 3D structure).

As ordered proteins are composed of different types of secondary structures such as alpha-helixes and beta-sheets, disordered proteins are also composed of distinct characteristics [10]. The Protein Trinity [5] proposes three functional states for proteins described as ordered, molten globule, and random coil, where molten globule and random coil are types of disorder. The molten globule state is defined as a form with regular secondary structure without a well defined tertiary structure [10], [11]. Random coil is defined as an extended flexible form without secondary structure [12]. The Protein Quartet model [6] takes this a step further adding a pre-molten globule state as another type of disorder. The disordered proteins and regions differ in attributes from ordered proteins including amino acid frequency, net charge, flexibility, sequence complexity, hydropathy, side chain polarity, surface area, and coordination number [9], [13]–[16].

Currently, the disordered regions have been associated with 28 different functions including cell cycle regulation, transcriptional and translational regulation, modulation of protein activity, assembly of other proteins, cell signaling, DNA recognition, protein-RNA recognition, membrane fusion and transport, and regulation of nerve cell function [2]–[4], [6], [7], [17]. Proteins associated with cancer may contain regions of disorder and are important in the development of anti-cancer drugs [17]. It is also suggested that disordered proteins can have different structural behaviors when functioning. Some completely disordered proteins can permanently bind to a partner, and for the rest of their lifetimes they exist in an ordered state [5], [7]. Others can change between disordered and ordered states when binding with partners [5]–[7]. Yet, in some cases disordered proteins carry out function without ever developing an ordered state [5]. For a more complete view of specific disordered proteins and their functions see [2]–[6].

It is obvious that disordered proteins are just as important biologically as ordered proteins. In fact, recent research, based on computational inference, has proposed that approximately 25-41% of eukaryotic proteins have at least one long disordered region consisting of more than 50 amino acids, 35-51% with more than 40 amino acids, and 48-63% with more than 30 amino acids [7]. Bacteria and Archaea also contain long disordered regions, but at lower percentages than eukaryotes [7]. With this discovery, it is increasingly important to develop techniques to accurately distinguish between regions of disorder and regions of order in proteins.

II. Background

Disordered proteins or regions of disorder in proteins can be identified experimentally by several methods, but these methods are labor intensive and costly. The most common lab methods are X-ray crystallography, NMR spectroscopy, circular dichroism (CD) spectroscopy, protease digestion, and Stokes radius determination [5], [8], [11], [12], [18], [19]. Each of these methods can detect different characteristics of disorder. In X-ray crystallography, some proteins have regions

of protein structure that lack electron density. This lack of electron density can be caused by ordered wobbly regions, disordered regions, or technical difficulties [20]. Additional experiments are usually needed to determine the structure of these regions. For example, the amino acid compositions of long regions that lack electron density are similar to the amino acid compositions of disordered regions found using NMR [20]. In NMR, disorder is indicated by sharp peaks, or using a ^{15}N-^{1}H heteronuclear nuclear Overhauser effect (NOE) measurement, ordered amino acids receive positive values and disordered amino acids receive negative values [5], [8]. However, while NMR can detect random coil disorder, it has difficulties in detecting molten globule regions [5]. Circular dichroism uses UV spectra to examine protein structure, and disorder is indicated by low intensity wavelength ranging between 210 to 240nm [8]. Protease digestion is used to break proteins into smaller components. Disordered regions experimentally tend to digest faster than ordered regions [5]. This method is useful in combination with the other methods. For example, it can be used to determine if a region lacking electron density from X-ray crystallography is a wobbly region or a disordered region [5] In Stoke's radius determination, disorder is indicated when a given molecular weight has abnormally large radii [5].

For some time, secondary structure has been predicted from amino acid sequence using computational techniques. It has been determined that disorder can also be predicted from amino acid sequence [12], [14], [16], [18], [21]–[23]. Disordered proteins and regions have been predicted with increasing accuracy using many computational techniques including artificial neural networks (ANNs) [7], [8], [10], [12], [15], [16], [18]–[22], [24]–[28], support vector machines (SVMs) [29], [30], logistic regression [20], [22], [25], and discriminant analysis [22]. Specific disorder predictors that have been developed include DisEMBL [19], RONN [27], GlobPlot [31], DISOPRED and DISOPRED2 [28], [30], [32], and an entire group named PONDRs [5], [7], [8], [10], [26].

This research was designed with the hypothesis that the hydropathy of the amino acids in a protein sequence significantly contributes to the ordered/disordered state of proteins. Previous research has demonstrated that straight amino acid composition or frequency is often enough to predict protein disorder [5], [10], [20], [29]. Other attributes have also been used, usually in combination, to predict protein disorder such as amino acid frequency, flexibility, sequence complexity, hydropathy, coordination number, net charge, amino acid volumes, side chain polarity, surface area, bulkiness, refraction, secondary structure attributes, and electron-ion interaction potential [8], [10], [12], [15], [18], [20]–[22], [24]–[26]. While it is certain that the many different properties of the amino acids contribute to the state of a protein, this project aims to examine the degree to which hydropathy affects the state of order and disorder and whether the attribute alone can be used to predict disorder. This hypothesis is further motivated by experimental findings that indicate that lower hydropathy is associated with disorder [33]. The hydropathy of amino acids is indicated by a numerical value where a positive value indicates the amino acid is hydrophobic (i.e. water hating) and a negative value indicates the amino acid is hydrophilic (i.e. water loving). For the purpose of this project, all hydropathy values have been normalized to positive values by increasing the lowest negative number to zero, and increasing all other values by the same amount.

TABLE I

FOR EACH SUBSET, THE NUMBER OF AMINO ACIDS IN EACH TYPE OF REGION IS GIVEN ALONG WITH THE TOTAL NUMBER OF AMINO ACIDS IN THE SUBSET.

Subset	Ordered	Disordered	Unknown	Total
1	5524	5523	12,138	23,185
2	5092	5092	10,586	20,770
3	5193	5195	10,002	20,390
4	5367	5368	12,871	23,606
5	5795	5795	12,269	23,859
Totals	26,971	26,973	57,866	111,810

III. METHODS

A. Dataset

The disordered dataset used for training and testing was obtained from DisProt [34], version 2.2 in FASTA format, which contains proteins with disordered regions and completely disordered proteins. After removing proteins from that version that were indicated on the web page to contain errors, the disordered dataset contained 183 protein sequences.

The sequences in the ordered dataset (i.e., containing completely ordered sequences) were obtained by manually searching PDB entries with structures determined using X-ray diffraction. Similar sequences were removed from the search to prevent redundancy and only proteins with a resolution of 2Å were examined to keep some similarity in the structure determination of the proteins in the dataset. Proteins were excluded from the dataset if the X-ray remarks indicated either that residues were missing or that residues were missing atoms, since this can either indicate disorder or procedural errors [12]. Proteins were included if no visible indication in the X-ray structure showed that residues were missing. For proteins with multiple chains, only one chain was selected since some chains in proteins can be identical or similar. The longest chain was chosen for the dataset to give more amino acids in the dataset; however, in the case that the chains were the same length, the first was chosen. Chains less than 81 amino acids in length were not considered since they would be smaller than the larger attribute window sizes.

The total number of ordered, disordered, and unknown amino acids in each sequence was collected, and this information was used to manually split the disordered dataset into five subsets for use in a five-fold cross-validation experimental design. Each subset contained completely disordered sequences and partially disordered sequences (i.e., containing ordered and disordered regions or containing unknown and disordered

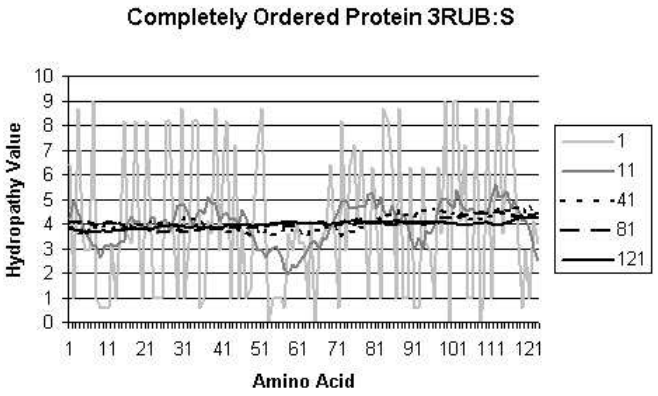

Fig. 1. The attributes of a completely ordered protein calculated using overlapping windows of various sizes

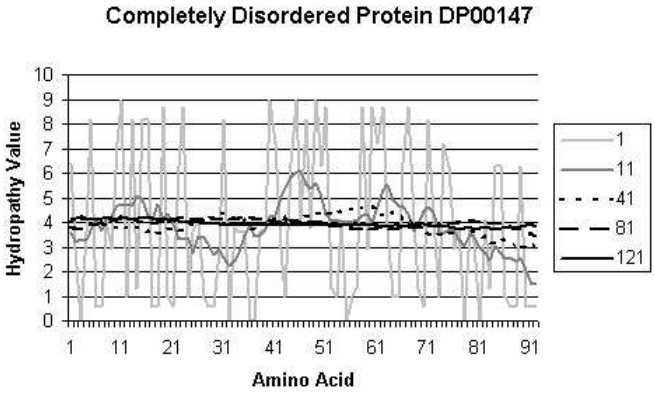

Fig. 3. The attributes of a completely disordered protein calculated using overlapping windows of various sizes

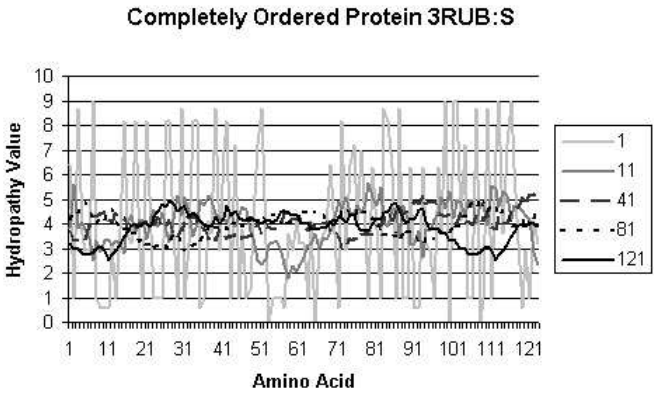

Fig. 2. The attributes of a completely ordered protein calculated using gapped windows of various sizes

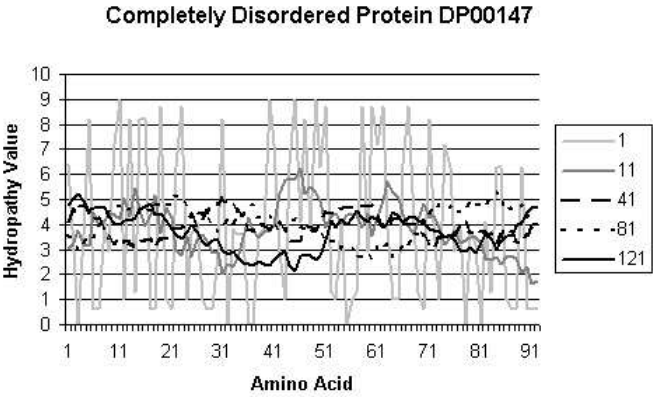

Fig. 4. The attributes of a completely disordered protein calculated using gapped windows of various sizes

regions). Ordered sequences were added to each subset so that each subset contained approximately the same number of ordered and disordered amino acids. Each subset contained sequences of various ranges of lengths. One of the problems with using all of the error free sequence data from the DisProt database is that many regions in these sequences contain amino acids labeled as unknown. At this initial stage of our project we have included all sequences containing these regions of unknown order/disorder and assigned the residues a third class label for training purposes. This is discussed in further detail in the next subsection. The goal of this dataset preparation was to reduce the likelihood that the neural network training would result in a particular state being predicted more often only because it was more common in the training set.

B. Attributes

An attribute vector and a class vector were created for each amino acid a in a sequence. The attribute vector contained the attribute information calculated for a as described below. The class vector represented the expected prediction by the neural network for each type of region, disordered, ordered, or unknown. Specifically, class vector $v = 0, 1$ indicates that residue a is part of a disordered region, $v = 1, 0$ indicates that residue a is part of an ordered region, and $v = 0.5, 0.5$ indicates that residue a is part of an unknown region. The attribute vectors and class vectors make up the patterns used to train the neural network. The actual prediction is made by assigning to the residue the class of the output neuron receiving the highest activation when processing the test data. Keep in mind that this process allows us to make an initial guess as to the ordered/disordered propensity of those residues labeled as unknown in the test data. The representation of $v = 0.5, 0.5$ for unknown regions allowed us to transmit to the neural network during training that a residue is not $apriori$ known to be part of either a disordered or ordered region. Recall that the output layer of the neural network contains only two nodes representing the class label. The absolute value of the difference in the activation values of the two output neurons is recorded as a measure of confidence in the prediction and can thus also be used as an initial guess of the actual label of unknown residues.

The attributes in an attribute vector were calculated based upon the hydropathy values of amino acids. For a specific position i in a protein sequence, the attributes were determined using a subsequence inside a window of size W_{in} where i was the center of the window. Since short and long range

interactions between amino acids can contribute to the overall state of a protein, the hydropathy attributes were calculated using different sizes of W_{in}. The values of W_{in} were 1, 11, 41, 81, 121, giving a total of five attributes in an attribute vector. Note that the first attribute, constructed using $W_{in} = 1$, consists of the actual hydropathy value of amino acid a at position i in the sequence. The other attributes were averaged hydropathy values for all amino acids, (i.e. when the window size $W_{in} > 1$).

Two windowing techniques were used in this research. Overlapping windows included the previous windows in the next larger window, and gapped windows excluded the previous windows from the next larger window. Figures 1 and 2 show the attributes calculated for a completely ordered protein using overlapping and gapped windows respectively. Figures 3 and 4 show the attributes calculated for a completely disordered protein using overlapping and gapped windows respectively. As can be seen from the graphs, the attributes calculated using overlapping windows tended to average out as the windows sizes increased, while the attributes calculated using gapped windows retained more hydropathy information as the windows sizes increased. Experiments were conducted using each of the two windowing techniques to determine whether using gapped windows retained more information than the overlapping windows and would therefore produce poorer predictions.

For overlapping windows, the hydropathy of position i was calculated by summing the hydropathy values of all amino acids present within the window W_{in} and dividing by the window size. Formally, for a position i of an amino acid sequence of length N, the hydropathy of position i for a window of size W_{in} was calculated by

$$H_{ia} = \frac{1}{w^r - w^l + 1} \sum_{j=w^l}^{w^r} h_j, \qquad (1)$$

where w^l is the position of the left end of the window determined by max (0, i − (W_{in} − 1)/2), w^r is the position of the right end of the window determined by min ($N - 1$, i + (W_{in} − 1)/2), and h_j is the hydropathy of the amino acid at position j. Note that the window size used in the attribute calculations may not be the same as W_{in} if no amino acids are present in the beginning or the end of the window.

For gapped windows, the hydropathy of position i was calculated by summing the hydropathy values of all amino acids present within the window W_{in}, subtracting the sum of the hydropathy values from the previous window W_{in-1}, and dividing by the current window size minus the previous window size. Formally, for a position i of an amino acid sequence of length N, the hydropathy of position i for a window of size W_{in} was calculated by

$$H_{ia} = \frac{1}{(w^r - w^l + 1) - W_{in-1}} \sum_{j=w^l}^{w^r} h_j \; - \sum_{j=w^{l-1}}^{w^{r-1}} h_j, \quad (2)$$

```
for each sequence
    for i = 0 to sequence length − 1
        for each window size
            half = (window size − 1) / 2
            left = i − half
            if (left < 0)
                left = 0
            right = i + half
            if (right > sequence length − 1)
                right = sequence length − 1
            window = right − left + 1
            for j = left to right
                sum = sum + hydropathy value at position j
            attribute = sum / window
```

Fig. 5. Pseudocode showing how the hydropathy attributes were calculated for the overlapping windows.

where w^l is the position of the left end of the current window determined by max (0, i − (W_{in} − 1)/2), w^r is the position of the right end of the current window determined by min ($N-1$, i + (W_{in} − 1)/2), W_{in-1} equals $w^{r-1} - w^{l-1} + 1$, w^{l-1} is the position of the left end of the previous window determined by max (0, i − (W_{in-1} − 1)/2), w^{r-1} is the position of the right end of the previous window determined by min ($N - 1$, i + (W_{in-1} − 1)/2), and h_j is the hydropathy of the amino acid at position j.

Table II shows the hydropathy values from the Kyte-Doolittle scale [35] for 22 amino acids and the normalized values used to calculate the hydropathy attributes. A PERL script calculated the hydropathy attributes from the amino acid sequences and created the pattern files used to train and test a neural network. See fig. 5 for pseudocode to calculate the hydropathy attributes for the overlapping windows.

C. Neural Network Architecture

The neural network was implemented using JavaNNS, a Java based neural network simulator developed at the Wilhelm-Schickard-Institute for Computer Science (WSI) in Tübingen, Germany based on the Stuttgart Neural Network Simulator (SNNS) 4.2 kernel. The neural network was a fully connected feed forward neural network with three layers, an input layer with the number of neurons equal to the number of input attributes (i.e., five, one for each window size), a single hidden layer with 10 neurons (10 because of the large number of amino acids in a training set), and an output layer with two neurons. Parameters for the neurons were a logistic activation function and identity for the output function. The neural network was randomly initialized using min equal to -1.0 and max equal to 1.0, and updating was by topological order. Training of the neural network was done using five-fold cross-validation, where four of the five subsets were used for training and the remaining subset was withheld for the testing phase.

D. Learning Algorithm

The learning algorithm used for training was the resilient propagation algorithm [36] using default parameters. Training

TABLE II

HYDROPATHY VALUES FOR 22 OF THE COMMON AMINO ACIDS AS DEFINED BY THE KYTE-DOOLITTLE SCALE.

Amino Acid	Hydropathy	Normalized
Alanine (A)	1.8	6.3
Cysteine (C)	2.5	7.0
Aspartate (D)	-3.5	1.0
Glutamate (E)	-3.5	1.0
Phenylalanine (F)	2.8	7.2
Glycine (G)	-0.4	4.1
Histidine (H)	-3.2	1.3
Isoleucine (I)	4.5	9.0
Lysine (K)	-3.9	0.6
Leucine (L)	3.8	8.2
Methionine (M)	1.9	6.4
Asparagine (N)	-3.5	1.0
Proline (P)	-1.6	2.9
Glutamine (Q)	-3.5	1.0
Arginine (R)	-4.5	0.0
Serine (S)	-0.8	3.6
Threonine (T)	-0.7	3.8
Valine (V)	4.2	8.7
Tryptophan (W)	-0.9	3.6
Tyrosine (Y)	-1.3	3.2
Aspartic Acid (B)	-3.5	1.0
Glutamic Acid (Z)	-3.5	1.0

was for 200 cycles using 1 step and shuffling of both patterns and subpatterns.

E. Performance Measure

A five-fold cross-validation scheme was used to assess the prediction accuracy of the neural network. True positives (TP) were the number of amino acids in disordered regions predicted to be disordered. True negatives (TN) were the number of amino acids in ordered regions predicted to be ordered. False positives (FP) were the number of amino acids in ordered regions predicted to be disordered. False negatives (FN) were the number of amino acids in disordered regions predicted to be ordered. These counts were used in the standard way to calculate sensitivity and specificity values. An additional measure, *Matthew's Correlation Coefficient* (MCC), which is a method of combining sensitivity and specificity into a single measure that is often used in the secondary structure prediction community, was also calculated as

$$MCC_i = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}}, \qquad (3)$$

where $p =$ patterns correctly assigned to class i, $n =$ patterns correctly assigned to not class i, $o =$ patterns incorrectly assigned to class i, $u =$ patterns incorrectly assigned to not class i. Since for our experiments we are only predicting two possible classes the MCC value for the disorder and order classes will be the same. Overall accuracy is measured by

TABLE III

AVERAGE PREDICTION ACCURACY FOR OVERLAPPING WINDOW AND GAPPED WINDOW NNS

Measure	Overlapping Windows	Gapped Windows
Sensitivity	0.8294	0.8282
Specifi city	0.6841	0.6960
Overall Accuracy	0.7606	0.7531
Matthew's CC	0.5153	0.5391

TABLE IV

UNKNOWN RESIDUE PREDICTIONS

Test Fold	Nbr of Unknowns	UD Rate	UO Rate
1	12,138	41.35%	58.65%
2	10,586	54.82%	45.18%
3	10,002	59.83%	40.17%
4	12,871	58.75%	41.25%
5	12,269	65.51%	34.49%
Mean	11,595	56.00%	44.00%

combining the TP and TN measures as follows:

$$OVERALL = \frac{TP + TN}{2} \qquad (4)$$

We point out that all of these measures are calculated on a *per-residue* basis. Accuracy measures for *per-protein* accuracy have not yet been analyzed.

Recall that some of the amino acids in our data were labeled as unknown, which was communicated to the neural network with a class label vector of $v = 0.5, 0.5$. The actual output of the predictor when processing unknown residues was calculated in the same fashion as was done for disordered residues and ordered residues (i.e. the output neuron with the highest activation level was taken as the prediction value). This procedure resulted in two additional measures: UD, the number of amino acids in unknown regions predicted to be disordered, and UO, the number of amino acids in unknown regions predicted to be ordered.

The TP, TN, FP, and FN rates indicated the percentage of each that were predicted. The UD and UO rates were calculated to indicate what percentage of amino acids in unknown regions were predicted as disordered and ordered respectively. The average prediction confidence was determined for each test, where the confidence of each single residue prediction was determined by subtracting the lowest output value from the highest. Ideally, an output of 1 0 or 0 1 would give a 100% prediction confidence and an output of 0.5 0.5 would give a 0% prediction confidence.

IV. RESULTS

A feed-forward neural network was trained and tested using hydropathy attributes calculated with overlapping windows and then with gapped windows. Table III shows the results of the experiments in terms of the previously defined prediction

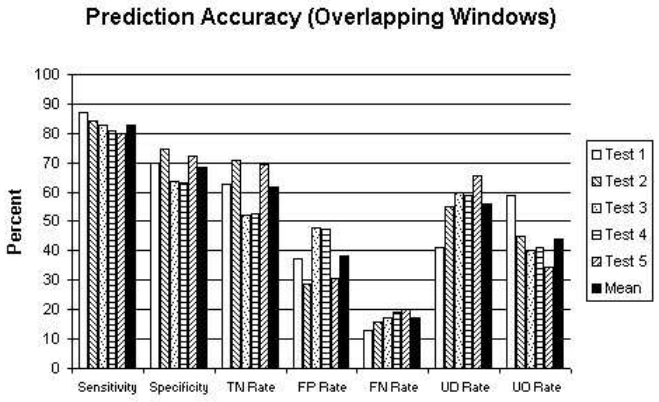

Fig. 6. The prediction accuracy of each test including the mean is given for the attributes calculated using overlapping windows

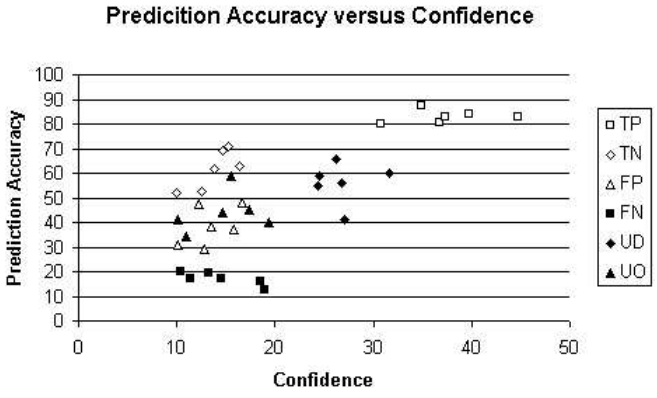

Fig. 8. The prediction accuracy plotted against the average confi dence is given for the attributes calculated using overlapping windows

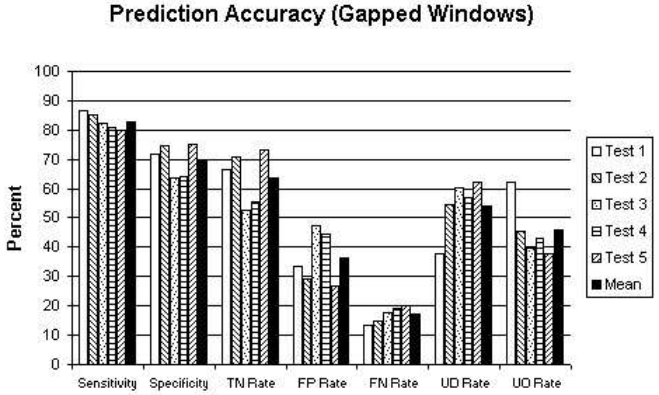

Fig. 7. The prediction accuracy of each test including the mean is given for the attributes calculated using gapped windows

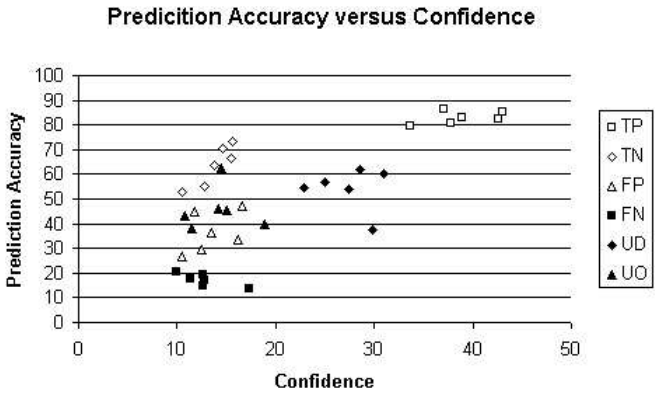

Fig. 9. The prediction accuracy plotted against the average confi dence is given for the attributes calculated using gapped windows

measures, *Sensitivity*, *Specificity*, *Overall Accuracy*, and MCC. The results shown are the average values as determined from the five test folds. For both methods, the measure of sensitivity was between 0.8 and 0.85 and the measure of specificity was above 0.60 for all folds, with a mean of 0.7.

Figures 6 and 7 show the prediction accuracy using overlapping windows and gapped windows, respectively. Overall, the prediction accuracies using the two different methods resulted in very similar prediction accuracies. The TP rate, or percentage of disordered amino acids that were predicted as disordered, ranged from 79% to 87%, with a mean of 83%. The TN rate, or percentage of ordered amino acids that were predicted as ordered, ranged from 52% to 73%, with a mean of 63%. There appears to be a very slight increase in the overall per-residue accuracy, computed as $(TP + TN)/2$, for the overlapping windows method, given that the overall accuracy of the most recent PONDR predictors [26] ranges from 79% to 83% with a W_{in} size of 41 and W_{out} sizes ranging from 1 to 121.

The FP rate, or percentage of ordered amino acids that were predicted as disordered, ranged from 28% to 47%. The FN rate, or percentage of disordered amino acids that were predicted as ordered, ranged from 12% to 20%. The UD rate, (i.e., the number of residues from unknown regions predicted to be in disordered regions) was typically higher than the UO rate, (i.e., the number of residues from unknown regions predicted to be in ordered regions) indicating that amino acids in unknown regions are more likely to be predicted as disordered. The actual counts for UD and UO are shown in table IV. The UD and UO measures showed the most variability across folds which may at least be partially due to there being up to a 22% difference in the number of unknown residues in different folds.

Figures 8 and 9 show the prediction rates plotted against the average confidence. The TP had both a high prediction accuracy and high average confidence compared to the others. The TN, FP, UO, and FN had average confidence between 10% and 20%, but varying prediction accuracies. The UD had a higher average confidence of being disordered, even though the amino acids were in unknown regions. We note that additional analysis of the confidence data is warranted based on an initial review of a random sampling of the distribution of confidence predictions for different proteins in the first fold.

Overall we note that although our system utilizes information based on a single type of attribute, hydropathy, our residue based prediction results are competitive with other disorder

predictors such as the PONDR series [26]. Perhaps the most surprising result is that our system achieved higher prediction accuracy for disorder region predictions, where as in general disorder predictors typically achieve slightly higher accuracies for ordered region predictions.

The prediction of disordered regions in proteins has become an important sub-problem in determining protein structure and function. Many computational techniques using different properties of amino acids have been employed to aid in this difficult task. This research using a feed-forward neural network and the hydropathy values of amino acids as attributes, confirms our hypothesis that hydropathy information can be used to predict regions of disorder. With regards to the two different windowing techniques tested, our results were not conclusive as the overall prediction accuracies were within the standard deviation of each other for the two windowing techniques. More analysis is clearly needed including testing for statistical significance between the results of these two windowing approaches and over all results in general. Such analysis is currently in process.

Future research includes plans to improve our training data and investigate different window sizes. The training data will be improved by replacing proteins containing unknown amino acids with proteins for which the certainty of a disorder or order class is already known. The current training set contains large quantities of amino acids in unknown regions as compared to amino acids in ordered/disordered regions (see table I). Training using only clearly identified (i.e. ordered/disordered) attribute vectors would make the current method for confidence measure more meaningful for residues with known class labels by eliminating the unknown class. It will be interesting to then compare the predictions obtained for unknown residues to see how they differ from the current system. Systematic investigation of different windowing sizes may help elucidate important areas of interaction between residues far apart from each other in the primary sequence. Most importantly we are currently working on the analysis of our results as computed on a per-protein basis in addition to the per-residue basis presented here.

ACKNOWLEDGMENT

The authors would like to thank the members of the DisProt team for their kind and efficient assistance to our queries with regards to the DisProt database. This work was supported by a grant from the Ohio Board of Regents Research Challenge Funds.

REFERENCES

[1] C. A. Orengo, A. E. Todd, and J. M. Thornton, "From protein structure to function," *Current Opinion in Structural Biology*, vol. 9, no. 3, pp. 374–382, 1999.

[2] P. Tompa, "Intrinsically unstructured proteins," *Trends in Biochemical Sciences*, vol. 27, no. 10, pp. 527–533, 2002.

[3] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: reassessing the protein structure-function paradigm," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 321–331, 1999.

[4] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[5] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26–59, 2001.

[6] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Science*, vol. 11, no. 4, pp. 739–756, 2002.

[7] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown, "Intrinsic protein disorder in complete genomes," *Genome Informatics Series: Workshop on Genome Informatics*, vol. 11, pp. 161–171, 2000.

[8] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker, "Sequence complexity of disordered protein," *Proteins*, vol. 42, no. 1, pp. 38–48, 2001.

[9] C. Bracken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker, "Combining prediction, computation and experiment for the characterization of protein disorder," *Current Opinion in Structural Biology*, vol. 14, no. 5, pp. 570–576, 2004.

[10] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder," *Proteins*, vol. 52, no. 4, pp. 573–584, 2003.

[11] H. J. Dyson and P. E. Wright, "Unfolded proteins and protein folding studied by nmr," *Chemical Reviews*, vol. 104, no. 8, pp. 3607–3622, 2004.

[12] E. Garner, P. Cannon, P. Romero, Z. Obradovic, and A. K. Dunker, "Predicting disordered regions from amino acid sequence: Common themes despite differing structural characterization," *Genome Informatics Series: Workshop on Genome Informatics*, vol. 9, pp. 201–213, 1998.

[13] R. M. Williams, Z. Obradovic, V. Mathura, W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown, and A. K. Dunker, "The protein non-folding problem: amino acid determinants of intrinsic order and disorder," *Pacific Symposium on Biocomputing*, pp. 89–100, 2001.

[14] L. M. Iakoucheva and A. K. Dunker, "Order, disorder, and flexibility: prediction from protein sequence," *Structure (Camb)*, vol. 11, no. 11, pp. 1316–1317, 2003.

[15] X. Li, Z. Obradovic, C. J. Brown, E. C. Garner, and A. K. Dunker, "Comparing predictors of disordered protein," *Genome Informatics Series: Workshop on Genome Informatics*, vol. 11, pp. 172–184, 2000.

[16] P. Romero, Z. Obradovic, and A. K. Dunker, "Intelligent data analysis for protein disorder prediction," *Artificial Intelligence Review*, vol. 14, no. 6, S2, pp. 447–484, 2000.

[17] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins," *Journal of Molecular Biology*, vol. 323, no. 3, pp. 573–584, 2002.

[18] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker, "Predicting intrinsic disorder from amino acid sequence," *Proteins*, vol. 53 Suppl 6, pp. 566–572, 2003.

[19] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure (Camb)*, vol. 11, no. 11, pp. 1453–1459, 2003.

[20] P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker, "Protein flexibility and intrinsic disorder," *Protein Science*, vol. 13, no. 1, pp. 71–80, 2004.

[21] P. Romero, Z. Obradovic, C. Kissinger, J. E. Villafranca, and A. K. Dunker, "Identifying disordered regions in proteins from amino acid sequence," *Proceedings IEEE International Conference on Neural Networks*, vol. 1, pp. 90–95, 1997.

[22] X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic, "Predicting protein disorder for n-, c-, and internal regions," *Genome Informatics Series: Workshop on Genome Informatics*, vol. 10, pp. 30–40, 1999.

[23] S. Lise and D. T. Jones, "Sequence patterns associated with disordered regions in proteins," *Proteins*, vol. 58, no. 1, pp. 144–150, 2005.

[24] P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot, and A. K. Dunker, "Thousands of proteins likely to have long disordered regions," *Pacific Symposium on Biocomputing*, pp. 437–448, 1998.

[25] S. Vucetic, P. Radivojac, Z. Obradovic, C. J. Brown, and A. K. Dunker, "Methods for improving protein disorder prediction," *Proc. 2001 IEEE/INNS International Joint Conference on Neural Networks*, vol. 4, pp. 2718–2723, 2001.

[26] K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Optimizing long intrinsic disorder predictors with protein evolutionary information," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 1, pp. 35–60, 2005.

[27] R. Thomson, T. C. Hodgman, Z. R. Yang, and A. K. Doyle, "Characterizing proteolytic cleavage site activity using bio-basis function neural networks," *Bioinformatics*, vol. 19, no. 14, pp. 1741–1747, 2003.

[28] D. T. Jones and J. J. Ward, "Prediction of disordered regions in proteins from position specific score matrices," *Proteins*, vol. 53 Suppl 6, pp. 573–578, 2003.

[29] E. A. Weathers, M. E. Paulaitis, T. B. Woolf, and J. H. Hoh, "Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein," *FEBS Letters*, vol. 576, no. 3, pp. 348–352, 2004.

[30] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[31] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson, "Globplot: Exploring protein sequences for globularity and disorder," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3701–3708, 2003.

[32] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The disopred server for the prediction of protein disorder," *Bioinformatics*, vol. 20, no. 13, pp. 2138–2139, 2004.

[33] V. N. Uversky, J. Gillespie, and A. L. Fink, "Why are 'natively unfolded' proteins unstructured under the phsiological conditions?" *Proteins*, vol. 42, no. 3, pp. 415–427, 2000.

[34] S. Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L. M. Iakoucheva, M. S. Cortese, J. D. Lawson, C. J. Brown, J. G. Sikes, C. D. Newton, and A. K. Dunker, "Disprot: a database of protein disorder," *Bioinformatics*, vol. 21, no. 1, pp. 137–140, 2005.

[35] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, 1982.

[36] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm," in *Proceedings of the IEEE International Conference on Neural Networks*, San Fransisco, CA, 1993, pp. 586–591.