Journal of Bioinformatics and Computational Biology

OPTIMIZING LONG INTRINSIC DISORDER PREDICTORS WITH PROTEIN EVOLUTIONARY INFORMATION

KANG PENG[1], SLOBODAN VUCETIC[1], PREDRAG RADIVOJAC[2]
CELESTE J. BROWN[3], A. KEITH DUNKER[4], ZORAN OBRADOVIC[1*]

[1]*Center for Information Science and Technology*
Temple University, Philadelphia, PA 19122, U.S.A.

[2]*School of Informatics*
Indiana University, Bloomington, IN 47408, U.S.A.

[3]*Initiative for Bioinformatics and Evolutionary Studies*
University of Idaho, Moscow, ID 83844, U.S.A.

[4]*Center for Computational Biology and Bioinformatics*
Indiana University School of Medicine, Indianapolis, IN 46202, U.S.A.

Received (11 August, 2003)
Revised (4 February, 2004)
Accepted (4 May 2004)

Protein existing as an ensemble of structures, called intrinsically disordered, has been shown to be responsible for a wide variety of biological functions and to be common in nature. Here we focus on improving sequence-based predictions of long (> 30 amino acid residues) regions lacking specific 3-D structure by means of four new neural-network-based Predictors Of Natural Disordered Regions (PONDRs): VL3, VL3H, VL3P, and VL3E. PONDR VL3 used several features from a previously introduced PONDR VL2, but benefitted from optimized predictor models and a slightly larger (152 versus 145) set of disordered proteins that were cleaned of mislabeling errors found in the smaller set. PONDR VL3H utilized homologues of the disordered proteins in the training stage, while PONDR VL3P used attributes derived from sequence profiles obtained by PSI-BLAST searches. The measure of accuracy was the average between accuracies on disordered and ordered protein regions. By this measure, the 30-fold cross-validation accuracies of VL3, VL3H, and VL3P were, respectively, 83.6±1.4%, 85.3±1.4%, and 85.2±1.5%. By combining VL3H and VL3P, the resulting PONDR VL3E achieved an accuracy of 86.7±1.4%. This is a significant improvement over our previous PONDRs VLXT (71.6±1.3%) and VL2 (80.9±1.4%). The new disorder predictors with the corresponding datasets are freely accessible through the web server at *www.ist.temple.edu/disprot*.

Keywords: Intrinsic Protein Disorder; PONDR; Neural Networks; Prediction; Evolutionary Information; PSI-BLAST

* Correspondence to: Zoran Obradovic, Center for Information Science and Technology, Temple University, 303 Wachman Hall (038-24), 1805 N Broad St, Philadelphia, PA 19122, U.S.A., Fax: +(215) 204-5082. E-mail: zoran@ist.temple.edu

1. Introduction

Intrinsic protein disorder is gaining increasing attention from the molecular biology community[1-6]. Disorder refers to the structural property that the amino acid sequence of a protein does not self-fold or lacks stable 3-D structure in its native state. A completely disordered protein does not fold from its N- to C-termini, while a partially disordered protein contains at least one unfolded or disordered region. Disordered regions longer than 30 consecutive amino acids are considered long disordered regions[3]. Disordered regions are characterized by various methods such as X-ray diffraction, nuclear magnetic resonance (NMR) spectroscopy, circular dichroism (CD) spectroscopy and limited proteolysis[7,8].

Contrary to the long accepted view that an amino acid sequence must first fold into a fixed 3-D structure in order to carry out protein function[9], disordered proteins have been associated with a wide range of important biological functions such as molecular recognition, molecular assembly/disassembly, protein modification and entropic chain activities. The lack of stable 3-D structure is crucial for these functions[3]. Following the prior work of Ptitsyn and Uversky[10], the protein trinity hypothesis was proposed[2] suggesting that proteins may exist in one of the three forms: ordered (fully-folded), collapsed-disordered (molten globule-like), or extended-disordered (random coil-like) and that protein function may arise from any of the three forms or the transitions between them. Proteins in collapsed-disordered forms still have regular secondary structure but lack fixed tertiary structure[11] and proteins in extended-disordered forms may also exhibit secondary structure. Thus, intrinsic disorder does not necessarily mean lack of secondary structure. More recently it has been proposed that extended disordered forms may have a less extended subset that was called premolten globules[4]. Both (or all three) disordered forms exist as flexible ensembles, and it would be expected that any given disordered protein would sample both (or all three) structural forms over time. In summary, disordered proteins or disordered regions are characterized by dynamically changing atomic coordinates or Ramachandran angles, whereas ordered proteins contain a single canonical set of atomic coordinates or Ramachandran angles.

While intrinsic disorder is clearly observed in some proteins *in vitro*, the existence of such disorder inside the cell is less certain. Disordered regions often evolve faster[12] and exhibit significantly different amino acid substitution frequencies[13] as compared to ordered proteins. These observations support the existence of protein disorder inside the cell. The existence of intracellular protein disorder recently received additional support when FlgM, previously shown to be an intrinsically disordered protein[14,15], was found by *in vivo* NMR experiments to gain some structure for only part of the molecule when in the cell, with about half of the protein remaining in the intrinsically disordered form [16].

In accordance with the hypothesis that amino acid sequence determines 3-D structure, we proposed that it also encodes the intrinsic disorder or lack of 3-D structure. To test this hypothesis, disordered proteins characterized by various methods as well as completely ordered proteins were collected by literature and database searches, and our first Predictor Of Natural Disordered Regions (PONDR) called XL1 was built from these

data to predict disordered regions based solely on the amino acid sequences. Using only 7 X-ray characterized partially disordered proteins, XL1 achieved 58% prediction accuracy[17]. Since this initial work, a series of PONDRs have been built and the overall prediction accuracy has been boosted to more than 80%. PONDR VLXT[18], whose accuracy reached 70%, integrated three feed-forward neural network predictors: VL1[18] trained on 25 variously characterized disordered proteins with 64% prediction accuracy, XN and XC[19] trained on N- and C-terminal disordered residues respectively. PONDR VL2[20] was a linear predictor built using ordinary-least-squares (OLS) regression on a much larger set of 145 disordered proteins and 130 completely ordered proteins. Its prediction accuracies were 73% on disordered residues and 88% on ordered residues. Naming conventions for various PONDRs were introduced in ref 8.

Several observations suggest that intrinsically disordered proteins indeed represent a sequence-encoded set of proteins distinct from the set of ordered proteins. The prediction accuracy exceeding 80% is significantly higher than the 50% expected by chance. This predictability arises from amino acid compositional distinctions between the two sets[8,18], and this predictability has been strongly supported in the recently completed fifth Critical Assessment of Structure Prediction experiment (*http://predictioncenter.llnl.gov/*).

In addition to our work on disorder prediction, several other groups have also investigated this topic. Williams[21] showed that two proteins gave NMR signals consistent with being unfolded under physiological conditions and that both of these proteins exhibited a low ratio of hydrophobic/charged residues. Uversky and co-workers independently developed their charge-hydropathy plot, which identifies natively unfolded proteins by their relatively higher net charge and lower hydropathy as compared to well ordered proteins[22]. Low complexity sequences were shown to be absent from globular regions of structured proteins in the Protein Data Bank (PDB)[18,23]. More recently, several other groups have developed predictors of disordered or similar regions[24-27] each of which has particular advantages.

A difficulty in comparing these related predictors is that different definitions of disorder are being used. For example, Liu and Rost[24] focused on long regions having no regular secondary structure (NORS); NORS have fixed 3D structure, and so by our definition, these NORS would not be considered to be disordered. Yet the amino acid compositions of the NORS segments share much in common with chains that remain unfolded under physiological conditions. As mentioned above, we combine both collapsed-disorder (molten globule-like) and extended-disorder (random coil-like) into a single disorder category, while others regard molten globules as partially folded and not in the natively unfolded category. Predictors of coil regions with high B-factors, called hot-loops, have also been developed[26,27]. Our studies showed that the amino acid compositions of high B-factor regions are distinct from those of disordered regions, and these distinctions were used to develop a predictor of high B-factor regions[28].

In this paper, we report on the development of four new predictors of long disordered regions, called VL3, VL3H, VL3P and VL3E, which are significantly more accurate than previous PONDRs. All four predictors are publicly accessible at our web site *www.ist.temple.edu/disprot*.

2. Materials and Methods

2.1. *Datasets*

The datasets used in this study consisted of 152 intrinsically disordered proteins and 290 completely ordered proteins (*www.ist.temple.edu/disprot*). The set of disordered proteins (*DIS152*) was developed based on the 145 disordered proteins used in our previous work[20]. This set was cleaned up by removing several incorrectly labeled chains as well as by adjusting order/disorder boundaries in a few others. In addition, several disordered proteins were added. The resulting set of disordered proteins consisted of 45 completely and 107 partially disordered chains with pairwise sequence identity limited to 25%. The total number of disordered residues was 22,434, while the partially disordered proteins also contained 26,188 ordered residues. In total, there were 162 disordered regions longer than 30 consecutive residues, 25 of which were longer than 200. A significant number of disordered residues came from two titin proteins (gi:1017427 with 2,174 disordered residues; gi:2136280 with 572 disordered residues). The average length of disordered regions shorter than 200 consecutive residues was 81.

The set of ordered proteins (*ORD290*) consisted of 290 non-redundant completely folded chains selected from the Protein Data Bank (PDB)[29] and first used in our previous study[30]. All proteins were required to be at least 80 residues long, to have resolution of ≤ 2Å, and an R-factor ≤ 20%. Sequence identity was limited to 25%. The total number of residues in the ordered set was 67,552.

Although the new predictors were all designed to predict disordered regions longer than 30 consecutive residues, their performance was also tested on *short* disordered regions having 30 or fewer consecutive residues. Thus, we also extracted 739 *putative* short disordered regions from 453 PDB chains, each identified as a stretch of 4-30 residues with missing backbone atom coordinates. All the chains were required to be at least 80 residues in length while pairwise sequence identity was limited to 25%. The resulting set (*SHORT453*) consisted of 7,543 disordered residues.

2.2. *PONDR VL3*

Sequence representation. The data for each position within a sequence consisted of an attribute vector and a class label. The attribute vector was derived from the subsequence covered by a moving window centered at the current position. The class label was 1 for examples from disordered regions and 0 for those from completely ordered proteins or ordered regions of partially disordered proteins from PDB. For position i of a sequence of length N, the attribute x_{ia} was calculated as the frequency 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 the input window of length W_{in},

$$x_{ia} = \frac{1}{w^r - w^l + 1} \sum_{j=w^l}^{w^r} d_{ja} \quad (1)$$

where $w^l = \max(1, i - (W_{in} - 1) / 2)$, $w^r = \min(N, i + (W_{in} - 1) / 2)$, $d_{ja} = 1$ if amino acid a

is at position j, and $d_{ja} = 0$ otherwise. Note that only 19 of the 20 amino acid attributes were necessary since the remaining one can be uniquely determined from the others.

Since low complexity regions are more likely to be disordered than ordered[18], an attribute called *K2*-entropy[23], which measures local sequence complexity, was calculated from the 20 amino acid frequencies within the same input window,

$$x_{iK2} = -\sum_{a=1}^{20} x_{ia} \log_2 x_{ia} \tag{2}$$

The flexibility[31] and hydropathy indices[32] and coordination number[33] are numeric attributes specific for each amino acid and are thought to be useful indicators of disorder[34]. The attribute for the flexibility index, x_{iFlex}, was calculated by averaging over positions within the input window:

$$x_{iFlex} = \frac{1}{w^r - w^l + 1} \sum_{j=w^l}^{w^r} Flex(j) \tag{3}$$

where $w^l = \max(1, i - (W_{in} - 1) / 2)$, $w^r = \min(N, i + (W_{in} - 1) / 2)$, $Flex(j)$ is the flexibility index for the amino acid at position j. The corresponding attributes for hydropathy and for coordination number were not included since their correlation coefficients to x_{iFlex} were larger than 0.7; thus, they would add little to the overall prediction accuracy and could cause collinearity problems[35]. Since the x_{iFlex} attribute is in fact a linear combination of the 20 amino acid frequencies, an additional amino acid frequency attribute was removed. Therefore, a total of 20 attributes were constructed for each position, consisting of 18 amino acid frequencies, *K2*-entropy x_{iK2} and flexibility x_{iFlex}. The excluded two amino acid frequencies were, quite arbitrarily, x_{iD}, and x_{iF}.

Predictor models. In a previous study[20], disorder predictors were built using ordinary-least-squares (OLS) regression, logistic regression (LR), and neural networks (NN). A surprising result, but consistent with previous experience in protein secondary structure prediction, was the relatively small difference between accuracies of linear predictors and neural networks.

In this study, we used an ensemble of neural networks as the predictor model. Similar to bagging[36], a set of neural networks was first trained, each on a balanced set randomly sampled (with replacement) from the examples available for training. Then, a majority-voting scheme was employed to combine them into a single predictor: given an out-of-sample example, it is predicted to be *disordered* if more than half of the neural networks predict disorder. This simple mechanism has been shown to be very effective in increasing the accuracy of unstable and powerful predictors such as neural networks[37].

Post-filtering of predictions. Post-filtering techniques are frequently used in protein structure predictions since neighboring positions often share the same structural property. The final prediction for a given sequence position is calculated as a function of raw predictions from neighboring positions within a given window. In this way, information from neighboring positions is used to smooth out predictions and thus improve accuracy by removing occasional misclassifications. The moving-average approach simply averages the raw predictions across neighboring positions[20,38]. Another approach trains a

second-stage neural network to map the raw predictions across neighboring residues to the final prediction for the current position[25,39]. A third approach uses a set of explicit rules to decide the final prediction based on the raw predictions[40], which requires the proper integration of domain knowledge.

In this study, we used the moving-average approach since it proved effective and efficient in our previous studies[17,18,20]. With this approach, the final prediction z_i for position i is calculated as,

$$z_i = \frac{1}{w^r - w^l + 1} \sum_{j=w^l}^{w^r} y_j \tag{4}$$

where y_i is the raw prediction at position j, z_i is the final prediction at position i, where $w^l = \max(1,\ i - (W_{out} - 1) / 2)$, $w^r = \min(N,\ i + (W_{out} - 1) / 2)$, and N is the sequence length. Note that $W_{out} = 1$ corresponds to raw predictions with $z_i = y_i$.

2.3. *PONDR VL3 Homology (VL3H)*

VL3H used the same 20 attributes, predictor models, and post-filtering as VL3. Since the available dataset with 152 disordered proteins in *DIS152* was fairly small and likely to constrain the achievable accuracy of disorder prediction, the data set was enhanced by including homologues of the disordered sequences. Based on previous results showing that using evolutionary information as part of the prediction process improved the accuracy of secondary structure prediction[40-43], we hypothesized that including homologues in the disorder training set would also improve prediction of protein disorder.

In order to effectively enlarge the disorder data, we used PSI-BLAST[44] to search for homologues of each long disordered region from our dataset. The E-value[44] associated with each disorder homologue estimates the statistical significance of its alignment with the corresponding disordered region. In this study, we used two thresholds specifying both the maximum and minimum E-values for inclusion of disorder homologues in the training data. Sequences with low E-values are usually very similar to the query sequence and thus contain little new information, while sequences with high E-values might be unrelated and thus introduce noise. After selecting disorder homologues by their E-values, we further filtered the homologues of a disordered protein so that no two sequences have sequence identity greater than 90% thus reducing bias in the training data. The experimentally determined disordered region was always included in the training set. Since the number of ordered proteins in our data set was sufficiently large, we did not attempt to repeat this procedure on ordered proteins.

2.4. *PONDR VL3 Profile (VL3P)*

In a second attempt to use evolutionary information to improve prediction accuracy, family profiles or position-specific scoring matrices (PSSM) from PSI-BLAST were used to construct attributes for the VL3P predictor. For a sequence of length N, an $N \times 20$ family profile is constructed based on the multiple alignment of homologues found during a PSI-BLAST search. In the PSIPred[41] secondary structure predictor, an input

window of length 15 was applied and all elements of the PSSM within that window were used as attributes or inputs to a large neural network. Since the appropriate window length for long disorder predictors is longer and the training data set is significantly smaller, using this attribute construction scheme would cause overfitting of the disorder predictor. Thus we used the moving-average approach similar to the VL3 predictor to derive profile-based attributes p_{ia}, $a \in$ {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}, for sequence position i as

$$p_{ia} = \frac{1}{w^r - w^l + 1} \sum_{j=w^l}^{w^r} s_{ja} \tag{5}$$

where $w^l = \max(1, i - (W_{in} - 1) / 2)$, $w^r = \min(N, i + (W_{in} - 1) / 2)$, N is sequence length, s_{ja} is the PSSM element at residue (row) j. Note that the sum of the 20 profile-based attributes is not necessarily equal to 1.

From Eq. (5) we can see that p_{ia} is the averaged log-odds of amino acid a in the neighborhood of sequence position i. These profile-based attributes are measures of evolutionary bias towards different amino acids within an input window. Correlation analysis showed that the correlation of p_{ia} to amino acid frequencies was relatively high; e.g., for $W_{in} = 41$ the correlation coefficient between x_{iA} and p_{iA} of Ala (A) was 0.56, and it was 0.58 between x_{iK} and p_{iK} of Lys (K). In the special case where no homologues are available, the profile-based attributes are linear combinations of the 20 amino acid frequencies, and therefore analogous to the attributes used in VL3.

Besides the 20 profile attributes, the x_{iK2} for $K2$-entropy and x_{iFlex} for flexibility, described in Section 2.2, were also used since they proved to be good indicators of intrinsic disorder.

VL3P used the same predictor models and post-filtering as the VL3 and VL3H predictors. To predict disorder on a given protein, a PSI-BLAST search should be performed to build a PSSM and construct attributes for VL3P. Thus, predicting with VL3P is more computationally expensive than predicting with VL3 and VL3H.

2.5. *PONDR VL3 Evolutionary (VL3E)*

Both VL3H and VL3P utilized evolutionary information, but in very different ways. VL3H used the same attributes as VL3 but was trained on the homologues of disordered proteins. On the other hand, VL3P relied on attributes constructed from the PSI-BLAST profile based on the aligned homologues of the query sequence. Since these representations were quite different our hypothesis was that an appropriate combination of VL3H and VL3P could lead to higher prediction accuracy. Assuming both VL3H and VL3P consisted of B neural networks, in VL3E the final prediction for a given example was therefore obtained through the majority vote from the $2B$ predictions.

2.6. *Accuracy estimation*

In each of the following experiments, a 30-fold cross-validation procedure was performed to estimate the prediction accuracy. First 152 disordered proteins in *DIS152* were randomly divided into 15 disjoint subsets and for each i = 1, 2, …15, a predictor was

trained using disorder examples from the disordered subsets 1, 2, …, *i*–1, *i*+1, …, 15, and order examples from all 290 ordered proteins in *ORD290*. The predictor accuracy was then tested on all disordered proteins from the *i*-th disordered subset. To test the accuracy on completely ordered proteins, a similar procedure was repeated for the 290 ordered proteins after randomly dividing them into 15 subsets.

After the 30-fold cross-validation procedure, test accuracies were obtained for each protein sequence in the dataset. The true positive rate (percentage of disordered examples predicted to be disordered) was measured on 45 completely disordered proteins, averaged, and reported as TP_C as well as on 107 partially disordered proteins, averaged, and reported as TP_P. TP was calculated as the averaged true positive rate over all disordered proteins. The true negative rate (percentage of ordered examples predicted to be ordered) was measured on each ordered protein, averaged, and reported as TN_C. We also measured the true negative rate on ordered residues of the 54 partially disordered proteins (TN_P) that were obtained from PDB. For partially disordered proteins from other sources, little experimental information was available to confirm the nature of regions not characterized as disordered. Even for those from PDB, the ordered parts were less reliable since some of them had missing coordinate information. Thus, we excluded such ordered segments from the training and ranked overall prediction accuracy as

$$OVERALL = \frac{TP + TN_C}{2} \tag{6}$$

The standard error was calculated by bootstrapping[45] for each accuracy measure. More specifically, 1,000 bootstrap replicated samples were drawn from of the set of 152 disordered and 290 ordered proteins with replacement and the 6 accuracies TP_C, TP_P, TP, TN_P, TN_C and $OVERALL$ were calculated on each bootstrap sample. The standard errors were then reported as one standard deviation of the results obtained over the 1,000 runs.

Besides the *per-protein* accuracies described above, we also measured *per-residue* accuracies: the *per-residue* true positive/true negative rate was calculated as the percentage of correctly predicted disordered/ordered residues. Although commonly used in measuring the accuracy of secondary structure predictors, it is worth noting that this measure is biased towards very long proteins. Therefore, the *per-protein* accuracy was selected as the main accuracy measure for protein disorder prediction.

3. Experiments and Results

3.1. *Experimental setting*

All neural networks consisted of a single hidden layer with H neurons and an output neuron, all with sigmoid activation functions. Each neural network was trained on a balanced set with the same number of examples of order and disorder using the resilient backpropagation algorithm[46] with a maximum of 200 epochs and validation-based stopping. Since our dataset was imbalanced (biased toward ordered examples) and redundant (due to the sliding window pattern representation), a balanced set consisting of 2000 disorder examples and 2000 order examples was randomly sampled to construct a

training set. An equal number of examples were sampled with replacement from each available sequence to ensure that the training set was not biased toward long sequences such as titin.

For VL3 and VL3P, disorder examples were taken only from the experimentally confirmed disordered proteins, while for VL3H they were taken from the disordered protein families. Order examples were drawn only from completely ordered proteins, while examples from ordered parts of partially disordered proteins were not included in training as discussed in Section 2.6. To avoid overfitting, 70% of the 4000-example training set was used in actual training and the remaining 30% for validation.

To find disorder homologues and build family profiles by PSI-BLAST search, the initial scoring matrix was BLOSUM62[47], gap penalties were 11/1, and the E-value thresholds were 0.0005 for including a sequence in a family profile and 0.01 for including a sequence in the output. The maximum number of iterations was limited to 3 in order to constrain the influence of potential false positives. The protein database was the non-redundant protein database (*ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz*) at NCBI, which included more than one million proteins from the GenBank CDS translations, PDB, SwissProt, PIR and PRF. Although low complexity regions might produce statistically significant but biologically incorrect alignments[44,48], they were not removed since 60% of the disordered residues in our dataset were assigned to low complexity regions by the SEG program[23] using the standard parameters K(1) = 3.4 and K(2) = 3.75, and a window of length 45.

3.2. *Results*

PONDR VL3: choice of window lengths. In order to determine the most appropriate window lengths, experiments were performed for all combinations of W_{in} = {11, 21, 41, 61, 81} and W_{out} = {1, 21, 41, 61, 81, 101, 121}. Note that W_{in} longer than the protein length corresponds to representing a whole protein with a single data point. Similarly, W_{out} longer than the protein length also corresponds to a loss of resolution giving a constant prediction for the whole protein.

Table 1 shows the prediction accuracies of the VL3 predictor, which was built as an ensemble of 10 neural networks of 10 hidden neurons, for various window lengths. It is evident that prediction accuracy on completely ordered proteins was considerably higher than on disordered proteins (Table 1.a), despite the fact that all predictors were trained on balanced data sets. Also, post-filtering ($W_{out} > 1$) significantly improved the prediction accuracy on completely ordered proteins while the gain on disordered proteins was relatively small. For example, when W_{out} was changed from 1 (i.e., no filtering) to 61, the improvement on TN_c was 7.5% while the improvements on TP_c, TP_p, TN_p and TP were less than 3%. As W_{out} was further increased to 121, TP_c, TN_p and TN_c kept increasing while TP_p, TP and *OVERALL* started decreasing. Similar behavior was observed for other values of W_{in} (data not shown). Thus, we finally selected 61 as the optimal output window length. In Table 1.b we list the results for different W_{in} with W_{out} set to 61. Based

on the similar reasoning, we selected 41 as the optimal input window length. Therefore, values W_{in} = 41 and W_{out} = 61 were fixed in all the remaining experiments.

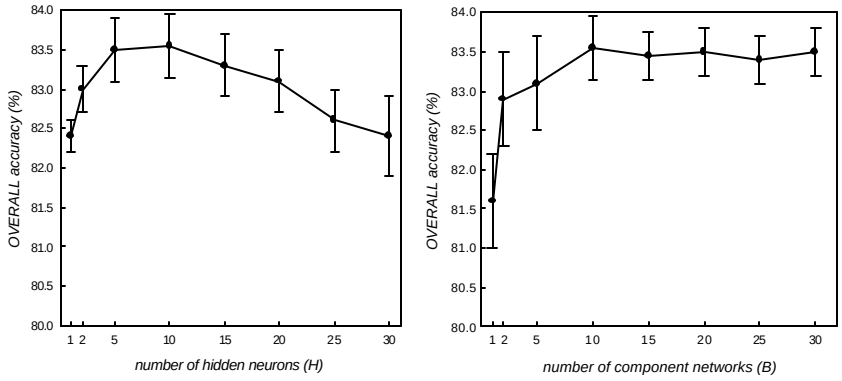
Table 1. Prediction accuracy of the VL3 predictor. (a) W_{in} fixed to 41 and W_{out} ranging from 1 to 121, (b) W_{out} fixed to 61 and W_{in} ranging from 11 to 81. The standard errors were estimated by bootstrapping described in Section 2.6.

W_{out}	TP_C	TP_P	TP	TN_P	TN_C	*OVERALL*
1	78.4±3.2	73.5±2.9	74.9±2.2	76.2±3.2	83.6±1.0	79.2±1.2
21	78.8±3.5	74.2±3.0	75.6±2.4	78.5±3.4	86.6±1.0	81.1±1.3
41	79.5±3.8	74.7±3.3	76.0±2.5	79.8±3.6	89.2±1.0	82.6±1.3
61	**80.5±4.2**	**74.2±3.5**	**76.1±2.7**	**79.1±3.9**	**91.1±1.0**	**83.6±1.4**
81	82.0±4.5	72.3±3.7	75.2±3.0	79.5±4.2	92.8±1.0	84.0±1.6
101	83.8±4.7	70.5±3.8	74.4±3.1	79.7±4.3	93.9±1.0	84.2±1.6
121	84.8±4.8	68.5±4.0	73.2±3.2	79.8±4.4	94.6±1.0	83.9±1.7

(a) W_{in} = 41

W_{in}	TP_C	TP_P	TP	TN_P	TN_C	*OVERALL*
11	69.3±4.6	66.7±3.6	67.5±2.8	83.5±3.4	91.9±0.9	79.7±1.5
21	72.7±4.5	70.7±3.3	71.1±2.8	80.8±3.8	92.0±1.0	81.6±1.4
41	**80.5±4.2**	**74.2±3.5**	**76.1±2.7**	**79.1±3.9**	**91.1±1.0**	**83.6±1.4**
61	85.9±3.6	72.8±3.6	76.7±2.7	76.9±4.2	90.4±1.1	83.6±1.5
81	86.2±3.9	73.4±3.8	77.2±2.9	75.1±4.3	91.3±1.1	84.2±1.6

(b) W_{out} = 61



(a) B = 10 (b) H = 10

Fig. 1. Influence of number of component networks (B) and number of hidden neurons (H) on accuracy: (a) B was fixed to 10 (b) H was fixed to 10. The accuracies and error bars were calculated as the mean and standard deviation of the results over 40 repeat experiments.

PONDR VL3: choice of neural network and ensemble size. Given the window length combination W_{in} = 41 and W_{out} = 61, we also examined other choices for the number of component networks (B) and the number of hidden neurons (H). For each combination of $B \in \{1, 2, 5, 10, 15, 20, 25, 30\}$ and $H \in \{1, 2, 5, 10, 15, 20, 25, 30\}$, we repeated the 30-fold cross-validation procedure (Section 2.6) 40 times and obtained mean and standard deviation of the *OVERALL* accuracy. As the number of hidden neurons increased, the

accuracy increased first until H = 10 and then decreased with increasing variances, indicating possible overfitting (Fig. 1a). As the number of component networks increased, the accuracy first increased and then saturated at H = 10 with decreasing variances (Fig. 1.b). This result confirmed the known property of bagging in improving accuracy of unstable classification algorithms by reducing the variance[37]. Based on the results, in the remaining experiments we used ensembles of 10 neural networks with 10 hidden nodes.

PONDR VL3: characterization of accuracy. To better characterize the accuracy of VL3, we show histograms of accuracies on the 152 disordered proteins and 290 completely ordered proteins (Fig. 2). To be specific, accuracy was higher than 80% on 96 disordered proteins and 249 ordered proteins. It is interesting to note that disorder was completely missed on 19 disordered proteins. Subsequent analysis did not reveal any specific property characterizing these disordered proteins.

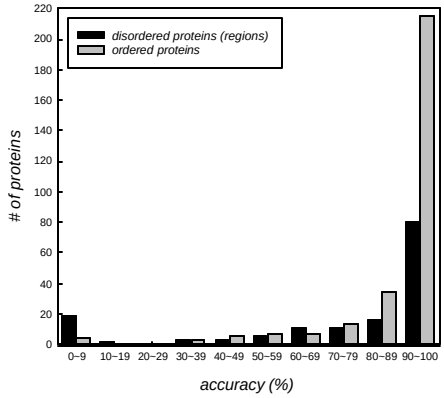


Fig. 2. Distribution of VL3 predictor accuracies on disordered proteins (regions) and completely ordered proteins.

PONDR VL3: comparison with linear predictors and single neural networks. Using the same training data and attributes as in VL3, we examined different predictor models, such as ordinary-least-squares (OLS) regression, logistic regression (LR) and single neural networks (NN). Furthermore, we also constructed an ensemble of 10 OLS predictors (OLS10) and an ensemble of 10 LR predictors (LR10). All of these predictors were less accurate than VL3 (Table 2). OLS was clearly the worst, but an unexpected result was that LR outperformed single NN. While the ensemble method helped improve performance of the neural network predictors (from NN to VL3), as expected, it did not improve OLS and LR significantly.

PONDR VL3: rejecting predictions with low confidence. Training by minimization of mean squared error leads to neural networks whose outputs approximate the posterior class probability[49]. Therefore, a high/low VL3 output indicates strong likelihood of

disorder/order. It is a standard practice, as in the experiments leading to Table 1, that a threshold of 0.5 is used to convert the output into the final prediction: if the output is higher than 0.5 disorder is predicted, and vice versa. However, the absolute difference between the actual output and a threshold of 0.5 also provides a useful measure of prediction confidence. Thus, we can use two separate thresholds, $q_d \geq 0.5$ for disorder and $q_o < 0.5$ for order, to achieve higher prediction confidence at the expense of lower prediction coverage, where no prediction is made when the predictor output is between the two thresholds.

Given $q_d = 1 - q$ and $q_o = q$, we examined several choices of q ranging from 0.5 to 0.05 in steps of 0.05 and plotted the prediction accuracy and coverage of VL3 (Fig. 3). As expected, the prediction coverage decreased while the prediction accuracies (*TP*, TN_C, *OVERALL*) increased as q decreased from 0.5 to 0.05 (right to left). For example, by accepting the coverage of 75%, accuracy was increased to *TP* = 80%, TN_C = 97%, and *OVERALL* = 89%, a significant improvement over accuracies reported in Table 1.

Table 2. Comparison of different predictor models. The standard errors were estimated by bootstrapping described in Section 2.6.

Predictor	TP_C	TP_P	*TP*	TN_P	TN_C	*OVERALL*
OLS	69.9±5.0	68.8±3.7	69.2±2.9	81.6±3.6	90.9±1.1	80.0±1.6
OLS10	70.7±5.2	70.6±3.8	70.6±3.0	81.0±3.7	90.4±1.1	80.5±1.6
LR	77.4±4.5	74.6±3.6	75.7±2.9	77.5±4.0	89.3±1.2	82.5±1.6
LR10	78.4±4.5	75.2±3.5	76.3±2.8	78.0±4.2	89.1±1.2	82.7±1.5
NN	78.5±4.4	71.0±3.5	73.2±2.8	77.5±4.0	90.2±1.4	81.7±1.5
NN10 (VL3)	80.5±4.2	74.2±3.5	76.1±2.7	79.1±3.9	91.1±1.0	83.6±1.4

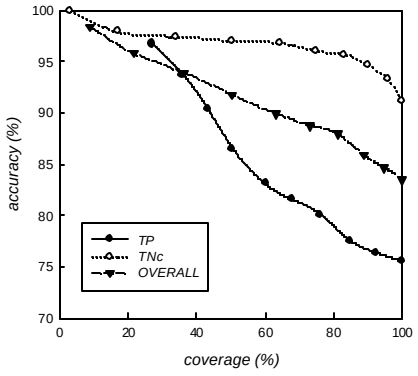


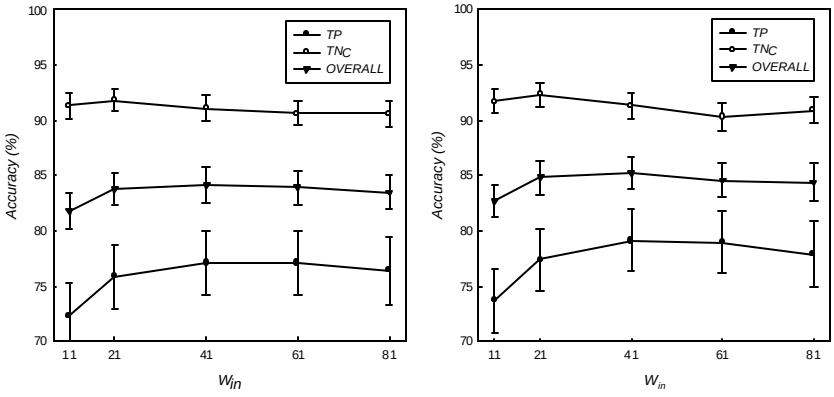
Fig. 3. Accuracy-coverage curves for the VL3 predictor

PONDR VL3H: choice of E-value thresholds. A total of 9,182 homologues were found for the 162 disordered regions by PSI-BLAST searches. Five disordered regions had no homologues, 24 had less than 5, 60 had less than 20, and 33 had more than 100 homologues. We examined all 21 possible pairs from a set {1e-02, 1e-05, 1e-10, 1e-20, 1e-30, 1e-40, 0} as the E-value range for selection of disorder homologues. The highest

OVERALL accuracy of 85.3±1.4% was achieved with the E-value range [1e-20, 1e-05], while the use of all available disorder homologues (i.e. E-value range [0, 1e-02]) produced a slightly worse accuracy of 84.7±1.4% (Table 3).

Table 3 Prediction accuracy of the VL3H predictor trained with different E-value ranges. The standard errors were estimated by bootstrapping described in Section 2.6.

E-value range	TP_C	TP_P	TP	TN_P	TN_C	*OVERALL*
[0, 1e-02]	**82.7±4.1**	**78.6±3.3**	**79.9±2.6**	**75.6±4.2**	**89.5±1.1**	**84.7±1.4**
[0, 1e-05]	83.8±3.9	79.2±3.2	80.6±2.5	75.8±4.2	89.4±1.1	85.0±1.4
[0, 1e-10]	84.4±3.9	79.5±3.2	80.9±2.6	75.2±4.3	89.3±1.1	85.1±1.4
[0, 1e-20]	80.2±4.3	77.2±3.4	78.0±2.7	76.4±4.2	90.1±1.1	84.0±1.4
[0, 1e-30]	82.3±3.8	74.6±3.5	76.9±2.7	78.1±4.1	90.9±1.1	83.9±1.5
[0, 1e-40]	78.9±4.4	74.5±3.5	75.7±2.8	77.9±4.0	91.2±1.0	83.5±1.5
[1e-40, 1e-02]	82.2±4.2	79.5±3.2	80.3±2.5	75.9±4.1	89.6±1.1	85.0±1.4
[1e-40, 1e-05]	81.9±4.0	79.3±3.2	80.1±2.5	75.5±4.3	89.3±1.1	84.7±1.4
[1e-40, 1e-10]	83.7±3.9	79.3±3.3	80.6±2.6	75.7±4.2	89.2±1.1	84.9±1.4
[1e-40, 1e-20]	81.3±4.3	77.4±3.3	78.5±2.7	77.0±4.3	90.2±1.1	84.4±1.5
[1e-40, 1e-30]	81.8±4.1	74.5±3.5	76.6±2.7	78.9±4.0	90.6±1.1	83.6±1.4
[1e-30, 1e-02]	83.9±3.8	79.6±3.2	80.8±2.6	76.0±4.2	89.6±1.1	85.2±1.4
[1e-30, 1e-05]	83.3±4.0	79.6±3.2	80.7±2.6	75.5±4.2	89.5±1.2	85.1±1.4
[1e-30, 1e-10]	83.7±3.8	79.9±3.2	81.0±2.6	75.4±4.2	89.3±1.1	85.2±1.4
[1e-30, 1e-20]	80.6±4.3	77.3±3.3	78.3±2.7	77.5±4.1	90.0±1.1	84.2±1.4
[1e-20, 1e-02]	83.1±3.8	78.2±3.3	79.7±2.5	76.5±4.1	89.7±1.1	84.7±1.3
[1e-20, 1e-05]	**84.7±3.6**	**79.0±3.2**	**80.7±2.5**	**75.7±4.2**	**89.9±1.1**	**85.3±1.4**
[1e-20, 1e-10]	83.6±3.7	77.6±3.3	79.4±2.5	74.8±4.2	89.5±1.1	84.5±1.4
[1e-10, 1e-02]	81.2±4.0	75.5±3.4	77.2±2.7	78.0±4.0	91.3±1.0	84.3±1.4
[1e-10, 1e-05]	81.9±3.8	75.1±3.4	77.1±2.6	78.0±4.1	91.1±1.0	84.1±1.4
[1e-05, 1e-02]	79.9±4.3	74.2±3.5	75.8±2.7	78.9±3.9	91.4±1.0	83.6±1.4



(a) 20 profile attributes only (b) 20 profile attributes + flexibility + *K2*-entropy

Fig. 4. Prediction accuracy of the VL3P predictor. The standard errors were estimated by bootstrapping described in Section 2.6.

PONDR VL3P: accuracy. For VL3P, we examined two sets of attributes: (a) 20 profile

attributes only ; (b) 20 profile attributes plus attributes for flexibility and *K2*-entropy (Fig. 4). VL3P achieved similar accuracy (85.2±1.5% using W_{in} = 41 and W_{out} = 61) to VL3H with E-values in the [1e-20, 1e-05] range (Table 3), although the latter did better on disordered proteins.

PONDR VL3E: accuracy. As described in Section 2.5, we constructed the VL3E predictor as a combination of VL3H and VL3P predictors. The E-value range for VL3H was [1e-20, 1e-05] and attributes for VL3P were 20 profile attributes, flexibility and *K2*-entropy. The resulting accuracy was 86.7±1.4%, considerably higher than those of VL3H and VL3P.

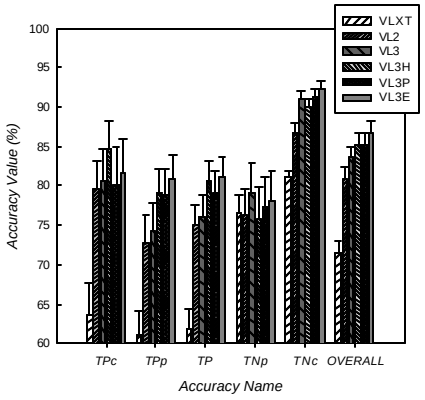


Fig. 5. Performance comparison of 6 disorder predictors on *DIS152* and *ORD290 datasets*. The standard errors were estimated by bootstrapping described in Section 2.6.

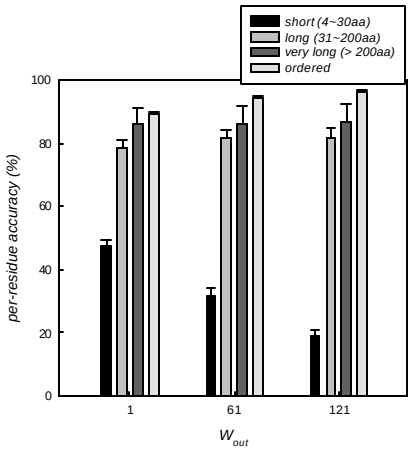


Fig. 6. Prediction accuracy on disordered regions of various lengths: *short* (4-30 residues), *long* (31-200 residues), *very long* (>200 residues), with W_{in} = 41 and several choices for W_{out}. Also shown is the accuracy on completely ordered proteins (*order*). Error bars are one standard error estimated by bootstrapping.

Comparison between VL3-type predictors and our previous disorder predictors. In Fig. 5, we compare the performance of the three new predictors with the VLXT and VL2 predictors, developed previously. The improvement from VL2 to VL3 was 2.7%, close to the sum of their standard errors. The actual difference is likely even larger because the accuracies of VLXT and VL2 might be overestimated since some of the test proteins were also used in their training. Therefore, a cleaner dataset, improved sequence representation, and more powerful prediction models used for VL3 resulted in a significantly higher accuracy over VL2 predictor. Furthermore, the improvement from VL3 to VL3E was 3.1%, larger than the sum of their standard errors. Therefore, using evolutionary information with two different approaches and combining the resulting

predictors resulted in additional significant improvement in accuracy.

Accuracy on short disordered regions. In Fig. 6 we compare VL3E prediction accuracy on disordered regions from three length groups: *very long* (>200 residues), *long* (31-200 residues), and *short* (4-30 residues), as well as on ordered proteins. A total of 24 very long disordered regions and 138 long disordered regions were selected from *DIS152*; 739 putative short disordered regions were obtained from 453 PDB chains in *SHORT453*. Note that, due to the fact that short regions are often either completely hit or completely missed, here we reported *per-residue* accuracies, while in previous tables we reported *per-protein* accuracies. For the whole range of W_{out}, prediction accuracies on the long and very long disordered regions were much higher than on putative short disordered regions. As W_{out} increased, prediction accuracy on short disorder decreased while accuracy on order, long and very long disorder kept increasing.

Comparison with other disorder predictors. We also compared VL3 and VL3E with disorder predictors developed by other research groups, including GlobPlot[27] V2.0 (http://globplot.embl.de/) and DisEMBL[26] V1.3 (http://dis.embl.de/) by Linding *et al.*, and FoldIndex© (http://bip.weizmann.ac.il/fldbin/findex) by Prilusky *et al.* based on an algorithm proposed by Uversky *et al.*[22] While NORSp [24] predictor is also related to the protein disorder it was not considered due to difficulties with obtaining the predictions from the server (http://cubic.bioc.columbia.edu/services/NORSp/). In Table 4 we show the results (*per-protein* and *per-residue* accuracies) for the 5 predictors on the two datasets used in this study: *DIS152+ORD290* and *SHORT453*. Although there are several versions of GlobPlot and DisEMBL predictors based on different disorder definitions, we show only the results corresponding to the Remark-465 disorder definition (i.e. a disordered region is identified as a stretch of residues with missing coordinates for backbone atoms), since this definition was also used in the CASP5 experiment[50]. When using DisEMBL and FoldIndex, we treated regions as ordered if they were not predicted to be disordered. For GlobPlot, however, only the predicted *potential globular domains* were treated as ordered. Furthermore, since GlobPlot might predict short overlaps between disordered regions and globular domains, we did not evaluate such predictions. Thus, the prediction coverage for DisEMBL and FoldIndex was always 100% while for GlobPlot it was usually less than 100%.

As in Table 4, VL3E achieved the highest *OVERALL* accuracy on *DIS152+ORD290* (long disorder) while GlobPlot performed best on *SHORT453* (short disorder). However, these results should not be viewed as a definitive measure of predictor quality. First, VL3/VL3E was designed for disordered regions longer than 30 residues, while disordered regions in *SHORT453* were shorter (see Discussion for details). Second, the disorder definitions corresponding to different predictors are somewhat different; non-VL3 predictors were trained on sequences with slightly different properties than *DIS152+ORD290*, and are therefore likely to be inferior on this data. Finally, accuracies of non-VL3 predictors on *DIS152+ORD290* might be overestimated since some proteins might be used in their training, while accuracies for VL3 and VL3E were obtained

through the 30-fold cross-validation procedure described in Section 2.6. Thus, these results should only be used as information about the differences between the properties of the examined predictors. This also highlights the need for unification of various definitions of protein disorder in order to allow future research to focus on development of disorder predictors that would have an even higher impact on proteomics research.

Table 4 Comparison for VL3, VL3E, GlobPlot, DisEMBL and FoldIndex predictors on two sets of proteins: *DIS152+ORD290 and SHORT453*. The standard errors were estimated by bootstrapping described in Section 2.6. Also shown is the prediction coverage.

Dataset	*Predictor*	*TP*	TN_c	*OVERALL*	*Coverage*
DIS152+ORD290	GlobPlot	53.7±2.5	71.9±1.2	62.8±1.4	88.5
	DisEMBL	32.2±2.5	97.5±0.3	64.8±1.3	100
	FoldIndex	59.8±3.2	86.4±1.4	73.1±1.7	100
	VL3	76.1±2.7	91.1±1.0	83.6±1.4	100
	VL3E	81.0±2.6	92.4±1.0	86.7±1.4	100
SHORT453	GlobPlot	56.3±2.2	76.8±0.8	66.5±1.0	89.5
	DisEMBL	31.4±1.9	97.4±0.2	64.4±0.9	100
	FoldIndex	19.3±1.6	85.0±1.0	52.1±0.6	100
	VL3	30.2±2.0	87.7±0.9	58.9±0.9	100
	VL3E	27.6±2.0	87.9±0.9	57.8±0.8	100

(a) *per-protein accuracy*

Dataset	*Predictor*	*TP*	TN_c	*OVERALL*	*Coverage*
DIS152+ORD290	GlobPlot	71.3±4.1	81.6±0.9	76.4±2.1	88.5
	DisEMBL	33.7±3.0	97.9±0.2	65.8±1.5	100
	FoldIndex	65.0±3.5	88.7±1.0	76.9±1.8	100
	VL3	79.3±3.8	93.2±0.7	86.3±2.0	100
	VL3E	84.0±3.4	94.6±0.7	89.3±1.7	100
SHORT453	GlobPlot	56.8±2.3	79.0±0.7	67.9±1.1	89.5
	DisEMBL	31.0±1.9	97.6±0.2	64.3±0.9	100
	FoldIndex	22.9±2.1	86.8±0.7	54.9±0.9	100
	VL3	33.7±2.3	89.8±0.6	61.8±1.1	100
	VL3E	31.7±2.3	90.6±0.7	61.2±1.1	100

(b) *per-residue* accuracy

4. Discussion

Over the last several years we and others have been developing predictors of natural disordered regions from amino acid sequence[17,19,20,22-27]. Indeed, prediction of disorder was added to the Critical Assessment of Protein Structure prediction in the 2002 cycle (CASP5) [30]. The general predictability of intrinsic disorder [50,51] and the specific prediction accuracies of VL3, VL3H, VL3P and VL3E reported here were all supported by the blind-prediction format of the CASP5 experiment. The small size of the CASP5 prediction set, however, limits the utility of a quantitative comparison with the results presented herein (the CASP5 target set contained only 2 proteins with long disordered regions).

Data quality and prediction accuracy. As in secondary structure prediction, a challenging problem affecting predictor quality came from a significant fraction of

incorrectly labeled residues. Long regions of CD-characterized disorder could easily contain structured subregions whose signals are missed because they are overwhelmed by the signals arising from the disordered regions. Long regions of missing coordinates in X-ray structures could be structured but wobbly domains[52]. NMR-characterized regions of disorder often contain local subregions with strong predictions of order that correlate with binding domains, suggesting that the lack of suitable non-local interactions causes intrinsically ordered regions to become disordered[53]. Similarly, the ordered dataset could also contain intrinsically disordered regions which undergo disorder to order transitions upon binding with partners or when in the crystal. Overall, the ordered and disordered datasets apparently contain a significant amount of noise, but the exact amount is very difficult to determine. Accordingly, it is possible that our current predictors are already approaching the limits set by the noise in the data, in which case only marginal improvements in accuracy could be expected in the future.

Consistent with our previous observations, prediction of order is always significantly more accurate as compared to disorder for all four new predictors. As discussed in more detail elsewhere[3], the lower accuracy for disorder prediction has multiple causes, with two probably being the most important. First, given the greater difficulty in characterizing non-folding compared to self-folding proteins, disordered data are likely to be much noisier than ordered data as discussed above. Second, attribute space spanned by disordered proteins is almost certainly larger than that of ordered proteins and so a larger collection of disorder examples is needed to cover this larger space.

The lower prediction accuracy for disordered protein might have an important biological root as indicated by the following observations. Structurally characterized regions of disorder frequently exhibit short, spike-shaped predictions of order. A short segment with order-forming tendencies typically would not fold into 3D structure - unless stabilized by a binding partner. A few of these spike-shaped predictions of order in structurally characterized disordered proteins have been shown to correspond to signaling regions that become ordered upon complex formation with their signaling partners[54]. A prototypical example is p53's acidic activation domain. In the absence of its binding partner, hdm2, this acidic activation domain undergoes rapid interconversion between helix and coil, with the equilibrium favoring the coil. The hdm2 molecule captures the helical form, thus leading to an overall coil to helix transitions upon binding[55]. Thus, the lower prediction accuracy of intrinsic disorder could result at least in part from such signaling regions. We are calling such regions “molecular recognition elements” or MoREs.

Cost-effective enlargement of the disordered protein database. Since the development of the initial XL1 predictor, a major problem has been and continues to be the paucity of disorder information. Information about disordered protein has not been organized in any fashion and remains scattered over various protein databases and literature citations. Also, disorder has been described with inconsistent terminology. As a result, our current disordered protein database is small compared with some other structural databases such as those used to develop secondary structure predictors. A more serious problem is that

the currently known disordered proteins might not be a good representative of those in nature, possibly even contributing to the discrepancy in accuracy between ordered and disordered regions. To effectively enlarge our current database of disordered proteins, we are currently developing methods for cost-effective techniques for extraction of knowledge about disorder from the literature. These methods also assist in developing a better understanding of available information about protein disorder[56]. With regard to new types of disordered proteins, we are also developing methods for identifying sequences that are distinct from both ordered proteins and also from currently known disordered proteins, and that yet are likely to be disordered[57,58]. Of course, such putative novel disordered proteins would have to be studied by laboratory structural characterization experiments to determine their actual order/disorder status.

Accuracy variation and prediction confidence. Although the overall accuracy of VL3 reached over 83%, the accuracy varied substantially among proteins: while the accuracies were more than 80% on about 3/4 of all 442 proteins, VL3 completely misclassified disorder on a number of proteins (Fig. 2). However, this is not a phenomenon unique to protein disorder prediction, but seems to be intrinsic to all aspects of protein structure and function predictions[59]. Given an out-of-sample protein, the problem is how to assess the prediction quality without the actual structural information. A partial solution is to use the absolute difference between the real-value prediction and 0.5 as a measure of prediction confidence. As suggested in Fig. 3, the larger the absolute difference, i.e. the more stringent thresholds, the higher the accuracy or confidence of prediction, although at the expense of decreasing prediction coverage.

Sequence representation and window length selection. Using attributes that represent frequencies over a window was motivated by results known as the incompressibility of protein sequence[60] showing that very little information can be extracted from protein sequences beyond first order statistics. In the case of secondary structure formation, interactions between non-local residues are known to be important[61]. In support of this view, protein engineering was used to confirm the context dependence of secondary structure formation for example segments[62]. In the native protein α-lactoglobulin, context-suppressed helical tendencies were relieved to yield non-native helical structure when specific non-local interactions were lost upon conversion to the molten globule form[63]. The difficulty in inferring such non-local interactions from protein sequence is a limiting factor for secondary structure prediction accuracy[53,64]. Because disordered regions lack strong non-local interactions, this problem is probably not as important for disorder predictions.

The optimal window lengths for secondary structure predictors are typically shorter than 20 based on the fact that α-helices/β-strands are typically of 5-40/5-10 consecutive residues[39-41,59]. Since the new predictors were developed to predict long disorder using a set of disordered regions longer than 30 residues (and 81 on average), we expected larger values for appropriate W_{in} and W_{out}. Accordingly, a predictor[28] specific in disordered

regions of 10 residues or shorter achieved its maximum accuracy with much smaller window lengths of W_{in}= 9 and W_{out} = 7.

Incorporation of evolutionary information. By incorporating evolutionary information, VL3H and VL3P predictors achieved accuracy improvements of 1.7/1.6% over VL3. Furthermore, the VL3E predictor which combined VL3H and VL3P resulted in an additional accuracy improvement of more than 3.1% compared to VL3. These results are in accord with improvements achieved in secondary structure predictions[42].

The success of VL3H may be due to using homologous sequences that diversify the training data and thus achieve better coverage of the attribute space of disordered proteins. On the other hand, the VL3P predictor adopted a popular strategy of using PSI-BLAST profiles for attribute construction. Due to the small amount of available data, we used a moving-average approach for attribute construction. It is expected that more efficient ways of utilizing information in the profile would result in further improvement, e.g. using all elements of the profile within the input window as attributes and performing attribute selection/extraction to reduce dimensionality. Although the correlation between the predictions of VL3H and VL3P was relatively high at 0.86, the majority-voting mechanism successfully utilized their differences to achieve additional improvements towards a respectable accuracy of 87%.

Performance on short disordered regions. Poor performance on short disordered regions was also observed during our participation in the CASP5 experiment [51]: all 3 predictors successfully predicted both long disordered regions (43 residues and 216 residues) in the target proteins with accuracy higher than 80%, while they were less successful on short disordered regions. In addition, a predictor[28] trained on a similar set of short disordered regions achieved only 65.9% accuracy on the long disordered protein dataset used in Vucetic *et al.*[20]. A detailed analysis revealed that short disordered regions did exhibit significantly different amino acid compositions and were more similar to flexible ordered regions in terms of flexibility index, hydropathy and net charge [28]. This calls for further research to better characterize short disordered regions and their biological functions.

Unification of various disorder definitions. While it is commonly accepted that *disorder* means unstructured, unfolded, or without stable 3-D structure in native state, there is still no universally accepted definition. Existing definitions such as equating disorder with random coils[26,27], high Cα atom temperature factor (B-factor)[26-28], missing coordinates for backbone atoms[26,27,50], and no regular secondary structure[24] *etc.* do not exactly match our concept of disorder, but rather reflect different features of protein disorder or the fact that disorder may have several flavors [20] (types). This situation means that comparisons of disorder predictors based on different definitions should be carried out only to inform potential users of possible limitations, not to imply that one predictor is "better" than another. For example, the GlobPlot predictors based on different propensity sets (disorder definitions) could perform very differently on the same test data. Unifying various definitions is needed to achieve better understanding of protein disorder and biological functions relating to it. Before such unification can be undertaken, however, a very large,

well annotated and organized set of disordered regions needs to be assembled; ideally, these disordered regions and proteins should be characterized by multiple biophysical methods so that the types or flavors of disorder can be established and identified for each disordered protein example.

Acknowledgments

Support from N.I.H. R01 LM007688-01 to A.K.D. and Z.O is gratefully acknowledged. A.K.D. thanks the Indiana Genomics Initiative (INGEN) for its generous support; INGEN is funded in part by the Lilly Endowment, Inc. We thank the anonymous reviewers for their very helpful comments and detailed suggestions. We also thank Vladimir Vacic for the construction of disorder database and predictor website.

References

1. A.P. Demchenko, "Recognition between flexible protein molecules: Induced and assisted folding," *J. Mol. Recognit.* **14**(1), 42-61 (2001).
2. A.K. Dunker and Z. Obradovic, "The protein trinity - linking function and disorder," *Nat. Biotechnol.* **19**, 805-806 (2001).
3. A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva and Z. Obradovic, "Intrinsic disorder and protein function," *Biochemistry* **41**(21), 6573 – 6582 (2002).
4. V.N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Sci.* **11**(4), 739-756 (2002).
5. V.N. Uversky, "What does it mean to be natively unfolded," *Eur. J. Biochem.* **269**, 2-12 (2002).
6. P.E. Wright, and H.J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm ," *J. Mol. Biol.* **293**(2), 321-331 (1999).
7. A.K. Dunker, J.D. Lawson, C.J. Brown, P. Romero, J. Oh, C.J. Oldfield, A.M. Campen, Ratlif, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C.H. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner and Z. Obradovic, "Intrinsically disordered proteins," *J. Mol. Graph. Model.* **19**, 28-61 (2001).
8. A.K. Dunker, C.J. Brown and Z. Obradovic, "Identification and functions of usefully disordered proteins," *Adv. Protein Chem.***62**, 25-49 (2002).
9. C.B. Anfinsen, "Principles that govern the folding of protein chains," *Science* **181**, 223-230 (1973).
10. O.B. Ptitsyn and V.N. Uversky, "The molten globule is a third thermodynamical state of protein molecules," *FEBS Lett.* **341**, 15-18 (1994).
11. D.A. Dolgikh, R.I. Gilmanshin, E.V. Brazhnikov, V.E. Bychkova, G.V. Semisotnov, S.Y. Venyaminov and O.B. Ptitsyn, "Alpha-Lactalbumin: compact state with fluctuating tertiary structure? "*FEBS Lett.* **136**(2), 311-315 (1981).
12. C.J. Brown, S. Takayama, A. Campen, P. Vise, T. Marshall, C.J. Oldfield, C.J. Williams and A.K. Dunker, "Eveolutionary rate heterogeneity in proteins with long disordered regions," *J. Mol. Evol.* **55**, 102-107 (2002).
13. P. Radivojac, Z. Obradovic, C.J. Brown and A.K. Dunker, "Improving sequence alignments for intrinsically disordered proteins," In *Proc. of 7th Pacific Symposium on Biocomputing*, Lihue, Hawaii, pp. 589-600, 2002.
14. G.W. Daughdrill, M.S. Chadsey, J.E. Karlinsey, K.T. Hughes and F.W. Dahlquist, "The C-terminal half of *the* anti-sigma factor, FlgM, becomes structured when bound to its target,

sigma 28," *Nat. Struct. Biol.* **4**, 285-291 (1997).
15. G.W. Daughdrill, L.J. Hanely and F.W. Dahlquist, "The C-terminal half of the anti-sigma factor *FlgM* contains a dynamic equilibrium solution structure favoring helical conformations," *Biochemistry* **37**, 1076-1082 (1998).
16. M.M. Dedmon, C.N. Patel, G.B. Young, G.J. Pielak, "FlgM gains structure in living cells," In *Proc. Natl. Acad. Sci. USA* **99**, 12681-12684 (2002).
17. P. Romero, Z. Obradovic, C.R. Kissinger, J.E. Villafranca and A.K. Dunker, "Identifying disordered regions in proteins from amino acid sequences," In *Proc. of IEEE International Conference on Neural Networks*, Houston, Texas, pp. 90-95, 1997.
18. P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown and A.K. Dunker, "Sequence complexity and disordered protein," *Proteins.* **42**, 38-48 (2001).
19. X. Li, P. Romero, M. Rani, A.K. Dunker and Z. Obradovic, "Predicting protein disorder for N-, C-, and internal regions," In *Proc. of Genome Informatics 10*, Tokyo, Japan, pp. 30-40, 1999.
20. S. Vucetic, C. Brown, A.K. Dunker and Z. Obradovic, "Flavors of protein disorder," *Proteins.* **52**, 573-584 (2003).
21. R.J. Williams, "The conformational mobility of proteins and its functional significance," *Biochem. Soc. Trans.* **6,** 1123-1126 (1978).
22. V.N. Uversky, J.R. Gillespie and A.L. Fink, "Why are 'natively unfolded' proteins unstructured under the physiological conditions?" *Proteins.* **42**(3), 415-427 (2000).
23. J.C. Wootton and S. Federhen, "Statistic of local complexity in amino acid sequences and sequence databases," *Comput. Chem.* **17**, 149-163 (1993).
24. J. Liu and B. Rost, "NORSp: predictions of long regions without regular secondary structure," *Nucleic Acids Res.* **31**(13), 3833-3835 (2003).
25. D.T. Jones and J. Ward, "Prediction of disordered regions in proteins from position specific score matrices," *Proteins, Special CASP5 Edition* **53**(S6), 573-578 (2003).
26. R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson and R.B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure* **11**, 1453-1459 (2003)
27. R. Linding, R.B. Russell, V. Neduva and T.J. Gibson, "GlobPlot: exploring protein sequences for globularity and disorder," *Nucleic Acids Res.* **31**(13), 3701-3708 (2003).
28. P. Radivojac, Z. Obradovic, D.K. Smith, G. Zhu, S. Vucetic, C.J. Brown, J. David Lawson and A.K. Dunker, "Protein flexibility and intrinsic disorder," *Protein Sci.* **13**(1), 71-80 (2004).
29. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.* **28**, 235-242 (2000).
30. D.K. Smith, P. Radivojac, Z. Obradovic, A.K. Dunker and G. Zhu, "Improved amino acid flexibility parameters," *Protein Sci.* **12**, 1060-1072 (2003).
31. M. Vihinen, E. Torkkila and P. Riikonen, "Accuracy of protein flexibility predictions," *Proteins.* **19**, 141-149 (1994).
32. J. Kyte and R.F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.* **157**, 105-132 (1982).
33. S.G. Galaktionov and G.R. Marshall, "Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D," Technical Report, Center for Molecular Design, Washington University, St. Louis, 1996.
34. P. Romero, Z. Obradovic and A.K. Dunker, "Intelligent data analysis for identifying protein disorder," *Artificial Intelligence Review* **14**, 447-484 (2000).
35. R. Davidson and J. Mackinnon, "Estimator and inference in econometrics," *Estimator and Inference in Econometrics*, Oxford University Press, New York, 1993.
36. L. Breiman, "Bagging predictors," *Mach. Learning* **24**(2), 123-140 (1996).
37. L. Breiman, *Bias, variance, and arcing classifiers*, Technical Report 460, University of California at Berkeley, 1996.

38. J.M. Chandonia and M. Karplus, "Neural networks for secondary structure and structural class prediction," *Protein Sci.* **4**, 275-285 (1995).
39. N. Qian and T.J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.* **202**, 865-884 (1988).
40. R.D. King and M.J. Sternberg, "Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, " *Protein Sci.* **5**(11), 2298-2310 (1996).
41. D.T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.* **292**, 195-202 (1999).
42. D. Przybylski, and B. Rost, "Alignments grow, secondary structure prediction improves," *Proteins***46**, 197-205 (2002).
43. B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *J. Mol. Biol.* **232**(2), 584-599 (1993).
44. S.F. Altschul, T.L. Madden, A.A. Schiffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.* **25**, 3389-3402 (1997).
45. B. Efron, and R.J. Tibshirani, *An introduction to the Bootstrap*, Chapman & Hall, New York, 1993.
46. M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," In *Proc. of the IEEE International Conference on Neural Networks*, San Francisco, CA, 1993.
47. S. Henikoff and J.G. Henikoff, "Amino acid substitution matrices from protein blocks," In *Proc. of the National Academic Science USA* **89,** 10915-10919 (1992).
48. S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.* **215**, 403-410 (1990).
49. C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.
50. E. Melamud and J. Moult, "Evaluation of disorder predictions in CASP5," *Proteins, Special CASP5 Edition* **53**(S6), 561-565 (2003).
51. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C.J. Brown and A.K. Dunker, "Predicting intrinsic disorder from amino acid sequence," *Proteins, Special CASP5 Edition* **53**(S6), 566-572 (2003).
52. E. Garner, P. Cannon, P. Romero, Z. Obradovic and A.K. Dunker, "Predicting disordered regions from amino sequence: common theme despite differing structural characterization," *Genome Inform. Ser. Workshop Genome Inform.***9**, 201-214 (1998).
53. D. Frishman and P. Argos, "Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence," *Protein Eng.* **9**, 133-142 (1996).
54. E. Garner, P. Romero, A.K. Dunker, C.J. Brown and Z. Obradovic, "Predicting binding regions within disordered proteins," *Genome Inform.* **10**, 41-50 (1999).
55. A. Bottger, V. Bottger, C. Garcia-Echeverria, P. Chene, H.K. Hochkeppel, W. Sampson, K. Ang, S.F. Howard, S.M. Picksley and D.P. Lane, "Molecular characterization of the hdm2-p53 interaction," *J. Mol. Biol.* **269**, 744-756 (1997).
56. B. Han, S. Vucetic and Z. Obradovic, "Reranking medline citations by relevance to a difficult biological query," In *Proc. IASTED Int'l Conf. Neural Networks and Computational Intelligence*, Cancun, Mexico, ***1***, 38-43 (2003).
57. K. Peng, S. Vucetic, B. Han, H. Xie and Z. Obradovic, "Exploiting unlabeled data for improving accuracy of predictive data mining," In *Proc. Third IEEE Int'l Conf. Data Mining*, Melbourne, FL, 267-274 (2003).
58. S. Vucetic, D. Pokrajac, H. Xie and Z. Obradovic, "Detection of underrepresented biological sequences using class-conditional distribution models," In *Proc. Third SIAM Int'l Conf. on Data Mining*, San Francisco, CA, pp. 279-283, 2003.

59. B. Rost, "Protein secondary structure prediction continues to rise," *J. Struct. Biol.* **134**, 204-218 (2001).
60. C.G. Nevill-Manning and I.H. Witten, "Protein is incompressible," *In Proc. of Data Compression Conference*, Snowbird, Utah, pp. 257-266, 1999.
61. W. Kabsch and C. Sander, "On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations," In *Proc. of the National Academy of Science USA* **81**, 1075-1078 (1984).
62. D.L. Minor, Jr. and P.S. Kim, "Context-dependent secondary structure formation of a designed protein sequence," *Nature* **380**, 730-734 (1996).
63. D. Hamada and Y. Goto, "The equilibrium intermediate of beta-lactoglobulin with non-native alpha-helical structure," *J. Mol. Biol.* **269**, 479-487 (1997).
64. S. Sudarsanam, "Structural diversity of sequentially identical subsequences of proteins: identical octapeptides can have different conformations," *Proteins* **30**, 228-231 (1998).