**Open Access Article**
*The authors, the publisher, and the right holders grant the right to use, reproduce, and disseminate the work in digital form to all users.*

*Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102*
*Volume 29, Issue Number 4, (2012)*
*©Adenine Press (2012)*

# SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method

http://www.jbsdonline.com

**Tuo Zhang[1,2,5]**
**Eshel Faraggi[1,2,5]**
**Bin Xue[3]**
**A. Keith Dunker[2]**
**Vladimir N. Uversky[3,4]**
**Yaoqi Zhou[1,2]***

[1]School of Informatics, Indiana University Purdue University, Indianapolis, IN 46202, USA
[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA
[3]Department of Molecular Medicine, University of South Florida, Tampa, FL 33612, USA
[4]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia
[5]Equal contribution.

***Abstract***

Short and long disordered regions of proteins have different preference for different amino acid residues. Different methods often have to be trained to predict them separately. In this study, we developed a single neural-network-based technique called SPINE-D that makes a three-state prediction first (ordered residues and disordered residues in short and long disordered regions) and reduces it into a two-state prediction afterwards. SPINE-D was tested on various sets composed of different combinations of Disprot annotated proteins and proteins directly from the PDB annotated for disorder by missing coordinates in X-ray determined structures. While disorder annotations are different according to Disprot and X-ray approaches, SPINE-D's prediction accuracy and ability to predict disorder are relatively independent of how the method was trained and what type of annotation was employed but strongly depend on the balance in the relative populations of ordered and disordered residues in short and long disordered regions in the test set. With greater than 85% overall specificity for detecting residues in both short and long disordered regions, the residues in long disordered regions are easier to predict at 81% sensitivity in a balanced test dataset with 56.5% ordered residues but more challenging (at 65% sensitivity) in a test dataset with 90% ordered residues. Compared to eleven other methods, SPINE-D yields the highest area under the curve (AUC), the highest Mathews correlation coefficient for residue-based prediction, and the lowest mean square error in predicting disorder contents of proteins for an independent test set with 329 proteins. In particular, SPINE-D is comparable to a meta predictor in predicting disordered residues in long disordered regions and superior in short disordered regions. SPINE-D participated in CASP 9 blind prediction and is one of the top servers according to the official ranking. In addition, SPINE-D was examined for prediction of functional molecular recognition motifs in several case studies. The server and databases are available at http://sparks.informatics.iupui.edu/.

## *Introduction*

Some proteins do not have a well-defined three-dimensional structure and others contain unstructured regions. These intrinsically disordered (or unstructured) proteins (IDPs) or regions in proteins (IDRs) play crucial functional roles in many biological processes, including transcriptional regulation, translation and cellular signal transduction (1-7). They are found often in eukaryotic organisms in particular (8-10), and are involved in various human diseases such as cancer, cardiovascular disease, and genetic diseases (11-13). The functional importance of IDPs and IDRs led to numerous theoretical and experimental studies in recent years (9, 12, 14-18).

IDPs and IDRs are commonly characterized by different theoretical and experimental approaches. Typically, IDRs are identified from the residues that are invisible

*Corresponding author:
Yaoqi Zhou
E-mail: yqzhou@iupui.edu

with X-ray crystallography and highly flexible regions based on Nuclear Magnetic Resonance (NMR) while fully disordered proteins are characterized by other experimental techniques such as NMR, circular dichroism (CD), and small angle X-ray scattering (19). Theoretical analysis of short ($\leq$30) and long ($>$30 residues) disordered regions further reveals different preferences for amino acid residues (20-23). For example, compared to structured proteins, long disordered regions are enriched in K, E, and P but depleted in G and N. On the other hand, compared to structured proteins, short disordered regions are enriched in G and D but depleted in I, V and L (22, 23). As a result, separate theoretical methods are often built for short and long disordered regions, and their combination is often used for a balanced prediction (20, 24-32). The first attempts to deal with differences between long and short regions of disorder were by PONDR VSL1 (29) and VSL2 (21). They were constructed from two first-level predictors, one for short regions of disorder ($\leq$15 residues), one for long regions of disorder ($\geq$30 residues) and a second-level, meta-predictor was used to assign weights for combining the two first-level predictors into a single predictor. Alternatively, for example, the meta predictor MFDp (28) employs several disorder predictors and other input features to make three separate predictors of short, long, and generic disordered regions, respectively, and combines them into a two-state predictor based on the maximal probability from three separate predictions.

In this paper we introduce a neural-network-based technique that makes an initial three-state, rather than the commonly used two-state, prediction of disorder. For continuity we call it SPINE-D (Sequence based Prediction with Integrated NEural network for Disordered residues). The method's input features include predicted torsion angle fluctuations (33), predicted secondary structure (34, 35) and solvent accessibility (ASA) (36). We found that the resulting method not only provides simultaneous training for detecting short and long disordered regions but also leads to a technique that has consistently high accuracy in predicting both short and long disordered regions. Another advantage of this technique is its insensitivity to the training database. Previous studies suggested that disorder annotations derived from X-ray structures in the PDB or from the Disprot database are not always consistent (37-39). The former database contains more short disordered regions while the latter one has more long disordered regions and fully disordered proteins verified by experiments (40). Here, we report that the method trained by X-ray annotations works equally well for predicting disordered regions annotated by the Disprot database (40) with improved annotations for ordered regions (38). A comparison to 11 different methods was made. Our single-method predictor is found to be comparable in predicting long disordered regions to and more accurate in predicting short disordered regions than the consensus meta-predictors [MFDp (28) and MD(41)] which were top performing in our testing. Moreover, SPINE-D was officially assessed to be among the best performing methods in the latest Critical Assessment of Structure Prediction techniques (CASP 9) (42).

## *Methods*

### *Neural Network*

The method follows our previously developed neural-network techniques (34, 36, 43) for sequence-based continuous-value prediction of backbone torsion angles and residue solvent accessibility. SPINE-D contains a two-hidden-layer neural network with an additional one-layer filter for smoothing the predictions. Each of the two hidden layers contains 51 hidden neurons and one bias and the filter layer contains 11 hidden neurons. SPINE-D employs a hyperbolic activation function and a guided learning technique that assigns a lower weight for residues further apart in sequence distance within a pre-defined sliding window (36). More importantly, to balance the need for predicting both short and long disordered regions, the method makes a three-state prediction on the residue level: ordered residues, residues in

short disordered regions ($\leq$30 residues) and residues in long disordered regions ($>$30 residues). The three-state prediction is reduced to a two-state prediction by simply adding the probabilities in short and long disordered regions. To reduce fluctuations caused by the random selection of initial weights, we trained five independent predictors and the final prediction is based on the average result of the five predictors. The back propagation algorithm with momentum was applied to optimize the weights (44). The learning rate and momentum were set to 0.01 and 0.4, respectively.

*Input Features*

The input nodes incorporate residue-level and window-level information, as well as one terminal tag. The residue-level information includes: a) seven representative physical parameters (45); b) a 20-dimension position-specific substitution matrix (PSSM) vector derived from the PSI-BLAST profiles (46) by searching the given sequence with three iterations against the NCBI's non-redundant protein sequence database; c) predicted secondary structure (3 dimensions) from SPINE-X (34) and predicted solvent accessibility (1 dimension) (36, 43); and d) predicted torsion-angle fluctuation (2 dimensions) (33). Predicted secondary structures are encoded as a 3-dimensional probability vector, *i.e.* the probabilities of coil, strand and helix predictions. Predicted solvent accessibility is normalized by the solvent accessible surface area (ASA) of an extended conformation (Ala-X-Ala) (36). We further employ the above information from sequence neighbors with a sliding window of 21 residues, 10 on each side from the residue to be predicted. A smaller window (5 on each side from the center residue) is employed for the filter layer. In addition to the residue-level information, the window-level information is generated by the current residue plus 15 residues on either side. It contains a) amino acid composition (20 dimensions); b) local compositional complexity (1 dimension) (47); and c) predicted secondary structure content (3 dimensions). Furthermore, five residues in N-terminus are encoded as $-1.0$, $-0.8$, $-0.6$, $-0.4$, $-0.2$ and five residues on C-terminus as 0.2, 0.4, 0.6, 0.8, 1.0 and the rest as 0.0. The majority of the dimensions in the input vector are in the range $[-1, 1]$. For those that are not, we linearly transformed them to this range.

*Datasets*

We prepared our datasets from two different sources: 1) X-ray structures in the protein databank and 2) disorder annotated proteins in the Disprot database (40). In our own database, we did not employ any NMR structures directly because there is a lack of a well-defined criterion to separate ordered and disordered residues. For the database from X-ray structures, we started with the original Disprot database that contained the annotations for X-ray structures solved before August 05, 2003 (48). We then expanded the dataset by retrieving X-ray structures from the PDB released after that date with the following criteria: a) resolution $\leq 2$ Å; b) size $\geq 60$ residues; c) having residues without coordinates (recorded as missing residues in REMARK465 section of PDB file, those missing residues are considered as disordered); and d) sequence identity cutoff of 90%. These two X-ray structure sets were combined and chains with unusual amino acid types were removed. We further removed HIS-tags in protein sequences because they were introduced for protein purification. The remaining proteins were clustered at 25% sequence identity using blastclust (46). One representative was chosen from each cluster according to the following prioritized criteria: a) a protein with the largest number of disordered residues; b) protein with the smallest number of disordered regions (*i.e.* more contiguous regions of disordered residues); and c) the largest protein. This led to a total of 4178 protein chains. We combined those chains with 91 fully disordered proteins from the Disprot database v5.0 (40), and performed another round of clustering at 25% sequence identity using blastclust. The representative of each cluster was chosen from a) a fully disordered protein if any or b) the protein with the largest number of

amino acid residues. The final dataset contains 4229 non-redundant chains (referred to as DM4229) that mix 4157 chains from PDB and 72 chains from Disprot. It consists of 1036634 residues, of which 103252 (about 10%) are annotated as disordered residues. We randomly selected 3000 chains (referred to as DM3000) from the DM4229 dataset to design our neural-network predictor and to perform a 10 fold cross validation test. The remaining 1229 chains (referred to as DM1229) were used as an independent test set.

In addition, we downloaded the benchmark SL dataset (38). The SL dataset was built by re-annotating Disprot to include reliable disorder and order contents. We used blastclust to filter the SL dataset and obtained a set of 477 chains, named SL477, in which the sequence identity between each pair is below 25%. The SL477 dataset was further filtered to remove sequences with sequence identity greater than 25% to chains in the DM4229 dataset, obtaining a set of 329 chains. We named it SL329 and employed it as an additional independent test set for comparison of our method with existing methods. Note that the residues without any annotation in SL datasets were not used in the evaluations reported here. Another distinction from the DM4229 dataset is that the SL dataset contains disordered regions annotated according to NMR structures. In particular, there are 24 chains with such NMR determined disordered residues in SL329.

To examine the dependence of our technique on training databases, we further filtered the DM4229 dataset by removing chains from Disprot and chains that are similar to the benchmark SL477 set. The remaining 4080 X-ray structures (DX4080) have low sequence identity ($<25\%$) with the SL477 dataset.

*Performance Evaluation*

The performance of disorder predictors is assessed by the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC). In addition, for a two-state prediction, we also calculate, as in the CASP competition (42, 49), sensitivity [TP/(TP + FN)], specificity [TN/(TN + FP)], the accuracy [ACC = (Sensitivity + Specificity)/2], the weighted score $S_w$,

$$S_W = \frac{W_d \times TP - W_o \times FP + W_o \times TN - W_d \times FN}{W_d \times N_d + W_o \times N_o}$$

and Mathews correlation coefficient (MCC),

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, FP, TN, and FN denote true positives (correctly predicted disordered residues), false positives (ordered residues that were incorrectly predicted as disordered), true negatives (correctly predicted ordered residues) and false negatives (disordered residues that were incorrectly predicted as ordered), respectively; $N_o$ and $N_d$ are the total numbers of ordered and disordered residues, respectively; $W_o$ and $W_d$ are the total percentages of disordered and ordered residues, respectively. It was found that there is a linear relation between the weighted score and ACC ($S_w = 2ACC - 1$) (42). That is, these two measures are essentially equivalent. Since $S_w$ was used in previous CASPs, we will use it for the rest of our analysis. $S_w$ and MCC range between $-1$ and 1 where 1 represents a perfect prediction and $-1$ a case where all predictions are incorrect. The higher the MCC and $S_w$ values, the better the prediction. To estimate the accuracy for predicting long and short disordered regions, we also define $Q_L$ and $Q_S$, the fractions of correctly predicted residues in long ($>30$) and short ($\leq 30$) disordered regions, respectively. During these calculations, short disordered regions with less than four residues were removed.

Ten-fold cross validation tests were performed on the DM3000 and SL477 datasets, separately. Specifically, we randomly divided the DM3000 or SL477 dataset into ten subsets with equal numbers of chains. Each subset was in turn chosen as the testing set, while the remaining nine subsets were merged to form a training set. We further performed an independent test by training on the entire DM3000 and then testing on the DM1229 dataset. Ten-fold cross validations in the DM3000 set were used to optimize the input window and neural network size based on AUC values. To remedy imbalanced populations of disordered and ordered residues, we employed repeated training (3 times) for disordered residues during the training process. To be specific, three duplicates of disordered samples were fed into the neural network. The purpose is to reinforce the information contained in disordered residues. We found that repeated training up to three times yields improved results and speeds up the convergence. The final SPINE-D web server, used in CASP9, was trained on the whole DM4229 dataset and tested on the independent SL329 dataset. For CASP 9, we employed a disorder threshold (0.06) that was optimized to yield the highest $S_w$ score for the CASP 8 dataset.

*Other Methods*

We compare SPINE-D with 11 relevant disorder prediction methods on the SL329 dataset. The eleven existing methods cover the four categories of disorder prediction methods, including: methods that only use amino acid propensity associated with disorder, *e.g.* IUPred short/long disorder predictor (25) and UCON (37); methods that incorporate 3D structure predictions, *e.g.* DISOCLUST (50); methods based on machine learning approaches, *e.g.* PONDR®VLXT (31, 51), Dispro (52), Disopred2 (53) and NORSnet (54); and meta servers that combine multiple disorder predictors, *e.g.* MD (41), PONDR-FIT (55), and MFDp (28). Predictions from these methods were generated on the SL329 dataset using either standalone implementations or their web servers. We note that chains in the SL329 dataset have low sequence identity ($<$25%) to the chains in the DM4229 dataset that was used to train the SPINE-D web server. This may decrease the accuracy difference between SPINE-D and other methods because proteins similar to those in SL329 were possibly contained in the training sets of other methods.

## *Results*

*Ten-Fold Cross Validation and Independent Tests*

We first performed ten fold cross validation on DM3000 and achieved an AUC of 0.858. Training on DM3000 and applying to the independent test set of DM1229 leads to a consistent AUC of 0.860. Table I further compares the results from the two tests. For DM3000, the threshold was optimized for achieving the highest $S_w$ score. To facilitate the comparison of the results between training and testing, we chose the threshold so that the specificity in DM1229 is the same as the specificity (87%) in DM3000. Table I shows the performance of SPINE-D for training and testing to be essentially the same. The consistency of the two tests indicates robust training.

The SPINE-D server was trained on the DM4229 and its disorder probability threshold was optimized for maximal $S_w$ score. We have designed SL329 as a set of proteins with low similarity to proteins in the DM4229 set to evaluate the server. Table I shows the results of this analysis. The robustness of training was also tested by comparing results on SL477 from ten-fold cross validated training and from training on independent PDB X-ray structures only (DX4080). Because $Q_L$ and $Q_S$ are based only on disordered residues, one can get perfect $Q_L$ or $Q_S$ by trivially predicting everything as disordered. Thus, we used the same specificity (85%) obtained

**Table I**

Performance comparison in different cross-validation and test sets.

| Test | AUC | SE[a] | SP[a] | MCC | $S_w$ | $Q_L$ | $Q_S$ |
|---|---|---|---|---|---|---|---|
| DM3000[b] | 0.858 | 0.70 | 0.87 | 0.45 | 0.58 | 0.66 | 0.74 |
| DM1229[c] | 0.860 | 0.70 | 0.87 | 0.43 | 0.57 | 0.65 | 0.73 |
| SL329[d] | 0.886 | 0.78 | 0.85 | 0.63 | 0.63 | 0.81 | 0.64 |
| SL477-T[e] | 0.870 | 0.77 | 0.85 | 0.63 | 0.62 | 0.80 | 0.57 |
| SL477-X[f] | 0.882 | 0.77 | 0.85 | 0.63 | 0.62 | 0.79 | 0.63 |

[a]SE: Sensitivity, SP: Specificity.
[b]Ten-fold cross validation. For the binary prediction, we adjusted threshold (0.08) of probability prediction to achieve the highest $S_w$.
[c]The threshold was adjusted so that comparison was done at same specificity as DM3000 (87%).
[d]Trained by DM4229 and independently tested on SL329.
[e]Ten-fold cross validation on SL477 only. Threshold was adjusted to yield the same specificity as SL329 (85%).
[f]Trained by X-ray structures only (DX4080) and independently tested on SL477. Threshold was adjusted to yield the same specificity as SL329 (85%).

for the SL329 set to set the thresholds for ten-fold cross validation on SL477 and independent test on SL477 by training on DX4080 in order to facilitate comparison among training and testing in SL datasets. Table I shows that the sensitivity, MCC, $S_w$ and $Q_L$ are essentially the same. The consistency between training on X-ray structures only and training on Disprot annotated disorder in SL477 indicates consistent assignment of disordered residues in the two datasets and in general between the PDB and the Disprot databases. Table I also shows a lower $Q_S$ when SPINE-D is trained and cross-validated on SL477. This result comes about because the DX4080 and SL477 datasets have very different compositions of short and long disordered regions. The number of short/long disordered regions for the SL477 and DX4080 sets are 568/383 and 5349/412 respectively. There are less training of short disordered regions in the SL477 set than in the DX4080 set.

Interestingly, Table I reveals that $Q_L < Q_S$ for DM3000 and DM1229 and $Q_L > Q_S$ for SL477 and SL329, regardless of how the method was trained. To understand this discrepancy, we compare the fraction of correctly predicted disordered residues (sensitivity) of SPINE-D on four different datasets (SL329, SL477, DM1229, and DM3000) as a function of region length in Figure 1. It is clear that the results on DM1229 and DM3000 are consistent with each other regardless if an independent or a ten-fold cross validation test is used. The same consistency is also evident between results on the SL329 and SL477 datasets. However, we found differences between the DM and SL datasets irrespective of the testing procedure. This suggests that the difference between them is mainly due to intrinsic difference among the databases. For DM1229 and DM3000, the sensitivity is high ($>$70%) in short regions ($\leq$60 residues) and very long regions ($>$180), and low (~55%) in long regions (60-180). By contrast, the overall trend for SL329 and SL477 is that longer disordered regions (30 residues or longer) are easier to predict (sensitivity $>$70%) than the shorter ones (3-15 residues, sensitivity ~50%). This behavior is caused by the different ratio between the number of long and short disordered regions in the different databases as mentioned earlier. Statistically speaking it is easier to find a more populated state. In addition, many more ordered residues (90%) in DM1229 and DM3000 as in CASP9 (42), makes prediction of disordered residues in long disordered regions more challenging because of more potential false positives. In other words, the prediction accuracy for short and long disordered regions can depend on the characteristics of the dataset.

*The SPINE-D Server (DM4229) and CASP 9*

To prepare our web-server for the blind predictions in CASP 9, we trained our SPINE-D predictor on the DM4229 dataset. We generated the binary prediction

by optimizing the probability threshold (0.06) to achieve the highest $S_w$ score (0.68) based on the CASP 8 dataset. According to the official assessment (42), our SPINE-D server is among the top 10 best groups in CASP 9, with AUC ranked in the 4th place and ACC ranked in the 6th place, as shown in Table I of the official assessment (42).

The two-state threshold in SPINE-D was trained based on the highest $S_w$ score in the CASP 8 dataset. If it was set by giving the highest MCC value in CASP 8 dataset, it would produce a higher MCC value for the CASP 9 and rank the 3rd in MCC, instead of the 9th (42). Increasing the MCC value is accompanied by a decrease in the $S_w$ score. This suggests that the performance of a given method in CASP strongly depends on how the threshold was trained and which quality measure was used.

As was noted previously (42), there was a large difference in performance between the CASP 8 and CASP 9 competitions. For SPINE-D the AUC dropped from 0.908 in CASP 8 to 0.832 in CASP 9. In contrast, the results from DM3000 and DM1229 shown in Table I, are essentially the same (AUC = 0.858 vs. 0.860). Figure 1 further demonstrates that the CASP 9 set is a difficult subset of X-ray structures even for short disordered regions. More importantly, there are very few long disordered regions. For example, the longest disordered region in CASP 9 has only 59 residues. This suggests that the CASP 9 dataset (2417 disordered residues and 23658 ordered residues) is not large enough to produce statistically consistent results for different methods, for long disordered regions, in particular.

*Independent Test on SL329 and Comparison to other Methods*

Because the CASP 9 set is too small and lacks long disordered regions, we compare our method to other methods by employing the independent test set SL329 that has a more balanced number of ordered and disordered residues (39544 disordered residues; 51292 ordered residues). Figure 2 compares the ROC curves given by 12

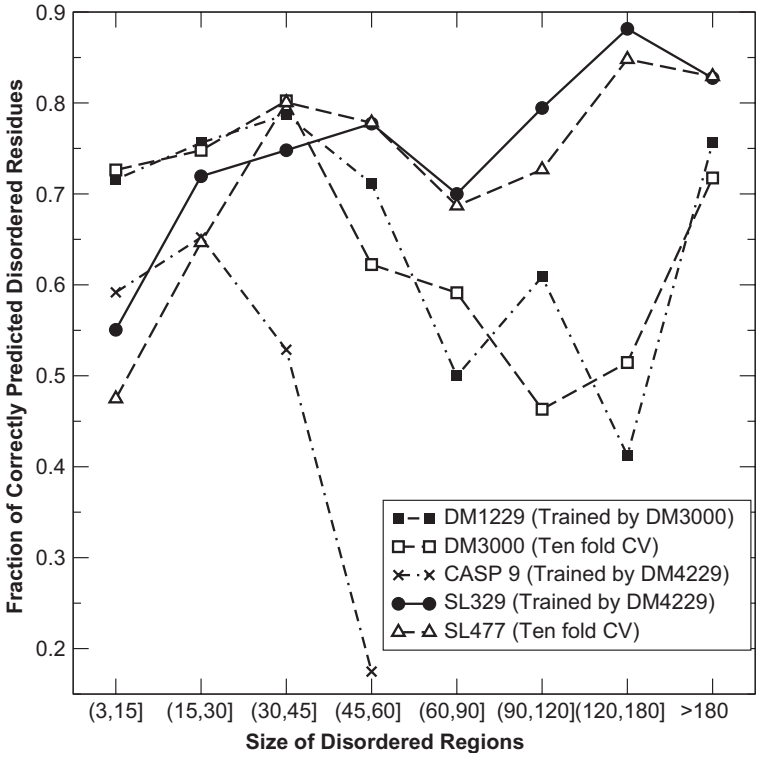


**Figure 1:** The fraction of correctly predicted disordered residues on DM1229, DM3000, CASP 9, SL329 and SL477 datasets as labeled.

different methods. The ROC curve given by SPINE-D at <50% false positive rate is consistently above the curves of other methods. SPINE-D achieves the highest AUC of 0.886. The next two highest AUC values are 0.873 and 0.864, given by two meta predictors MFDp and MD, respectively.

Comparison among different methods is also made quantitatively in Table II in term of sensitivity, specificity, MCC, $S_w$ score, $Q_L$ and $Q_S$ by employing default thresholds. SPINE-D gives the highest MCC and $S_w$ while sensitivity and specificity vary significantly among different methods. This variance is likely because different methods were optimized for different objectives and optimizing for sensitivity of an inexact method, for example, will lead to lower specificity. To facilitate the comparison, we adjusted the probability cutoff of each predictor so that it has the same specificity (85%) as SPINE-D. In this case, the sensitivity of SPINE-D along with MCC, $S_w$, $Q_L$ and $Q_S$ are the highest in all methods compared (See Table II). In particular, SPINE-D is 12% more accurate in making correct prediction of residues in short disordered regions ($Q_S$) and achieves comparable $Q_L$ for residues in long disordered regions when compared to the meta server MFDp with the second highest AUC. This indicates that training with an unbalanced dataset (DM4229 with 90% ordered residues) does not prohibit superior performance in a balanced dataset (56.5% ordered residues in SL329).

Figure 3 compares the accuracy according to the size of disordered regions. SPINE-D can correctly predict more disordered residues than all methods except MFDp and Dispro. MFDp can yield comparable predictions in disordered regions that are longer than 60 residues but gives significantly worse predictions in shorter disordered regions (≤60 residues), while Dispro can provide comparable predictions in shorter disordered regions (≤60 residues) but worse predictions in longer disordered regions (>60 residues). SPINE-D is the only predictor that generates greater or comparable number of correct predictions in disordered regions of all sizes.

The above results are for the residue-level accuracy. Figure 4 compares correct prediction of short and long disordered regions given a coverage cutoff. That is, a disordered region is considered as correctly predicted (successful) if more than

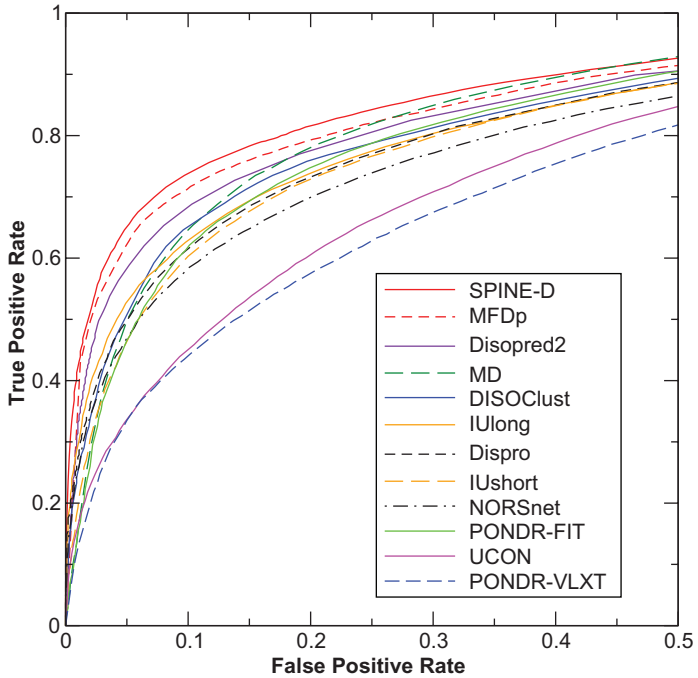


**Figure 2:** ROC curves for disorder prediction on the SL329 dataset given by twelve methods as indicated.

a particular fraction of residues in that region is predicted as disorder (coverage cutoff). As a shown in Figure 4A, SPINE-D gives the highest success rate if the coverage cutoff is 60% or more for short disordered regions. For long disordered regions (Figure 4B), SPINE-D gives the highest success rate for coverage cutoff between 40% and 90%. Thus, as in the residue-level prediction, SPINE-D has consistent performance in both short and long disordered regions.

**Table II**

Performance of various methods on the SL329 dataset.

| Method | AUC | Sensitivity | Specificity | MCC | $S_w$ | $Q_L^a$ | $Q_S^b$ | MSE[c] |
|---|---|---|---|---|---|---|---|---|
| SPINE-D | **0.886** | 0.78 (**0.78**[d]) | 0.85 (0.85) | **0.63** (**0.63**) | **0.63** (**0.63**) | −(**0.81**) | −(**0.64**) | −(**0.070**) |
| MFDp | 0.873 | **0.88** (0.76) | 0.62 (0.85) | 0.51 (0.61) | 0.50 (0.61) | −(0.80) | −(0.52) | −(0.113) |
| Disopred2 | 0.858 | 0.69 (0.74) | 0.90 (0.85) | 0.59 (0.60) | 0.61 (0.59) | −(0.76) | −(0.60) | −(0.093) |
| MD | 0.864 | 0.66 (0.72) | 0.89 (0.85) | 0.58 (0.59) | 0.55 (0.58) | −(0.75) | −(0.57) | −(0.110) |
| DISOClust | 0.846 | 0.81 (0.72) | 0.70 (0.85) | 0.51 (0.57) | 0.51 (0.57) | −(0.73) | −(0.60) | −(0.093) |
| PONDR-FIT | 0.843 | 0.61 (0.69) | 0.91 (0.85) | 0.55 (0.55) | 0.51 (0.54) | −(0.72) | −(0.54) | −(0.086) |
| IUlong | 0.839 | 0.60 (0.69) | 0.92 (0.85) | 0.55 (0.55) | 0.52 (0.54) | −(0.74) | −(0.41) | −(0.118) |
| Dispro | 0.837 | 0.28 (0.69) | **0.99** (0.85) | 0.40 (0.55) | 0.27 (0.54) | −(0.70) | −(0.63) | −(0.077) |
| IUshort | 0.829 | 0.50 (0.67) | 0.94 (0.85) | 0.50 (0.54) | 0.44 (0.53) | −(0.71) | −(0.47) | −(0.096) |
| NORSnet | 0.815 | 0.54 (0.65) | 0.92 (0.85) | 0.51 (0.51) | 0.46 (0.50) | −(0.69) | −(0.42) | −(0.144) |
| UCON | 0.779 | 0.59 (0.54) | 0.81 (0.85) | 0.42 (0.41) | 0.40 (0.39) | −(0.57) | −(0.38) | −(0.148) |
| PONDR®VLXT | 0.755 | 0.59 (0.51) | 0.78 (0.85) | 0.38 (0.39) | 0.38 (0.36) | −(0.53) | −(0.44) | −(0.140) |

Predictors are ranked according to $S_w$. Highest number in each column is shown in bold. We did not calculate $Q_L$, $Q_S$ and MSE of disorder content based on default binary predictions because without fixing specificity, those two quality measures are strongly depending on how the methods were trained.

[a]Fraction of correctly predicted residues in long disordered regions (>30 residues).

[b]Fraction of correctly predicted residues in short disordered regions (≤30 residues).

[c]Mean square errors on predicting disorder content.

[d]The number in parentheses is when all thresholds are adjusted to a specificity of 85%. (Thresholds are 0.670 for MFDp, 0.039 for Disopred2, 0.476 for MD, 0.658 for DISOClust, 0.434 for IUlong, 0.061 for Dispro, 0.389 for IUshort, 0.400 for NORSnet, 0.621 for UCON, 0.620 for PONDR®VLXT, and 0.390 for PONDR-FIT, respectively.).

We also tested our method on the protein level. Disorder content is the fraction of disordered residues for a given protein. Sometimes, disorder content, rather than the exact location of disordered residues, is needed (56). We calculated disorder content predicted for each protein and obtained mean square error (MSE) (56, 57) between predicted and annotated disorder contents. Annotated disorder content is calculated based on annotated ordered and disordered regions only. Table II shows that at a constant specificity SPINE-D gives the lowest MSE (0.070) and the next lowest is 10% higher (0.077) by Dispro.

*Disorder and Function: Case Studies*

Several classes of disorder-based binding events are mediated by specific, short (around 20 residues) structural elements, known as molecular recognition features (MoRFs, because such regions "morph" from disorder to order upon binding) (59-61). These short binding fragments are located within long disordered regions and gain structural order upon binding to their specific partners. In disorder probability plots produced by disorder predictors, they are often seen as characteristic downward spikes within the disordered regions (57-61). Earlier, based on the comparison of the outputs of several disorder predictors for three proteins with known disorder-based binding sites it has been concluded that PONDR®VLXT was more sensitive for features associated with regions potentially undergoing disorder-to-order transition than other predictors (61). Here, we compare the capability of SPINE-D to find potential binding sites with PONDR®VLXT (31, 51) and a meta-predictor PONDR-FIT (55). Specifically, we examined whether or not each predictor produced "dips" – or short regions of predicted order within longer regions of predicted disorder – corresponding to known binding regions.

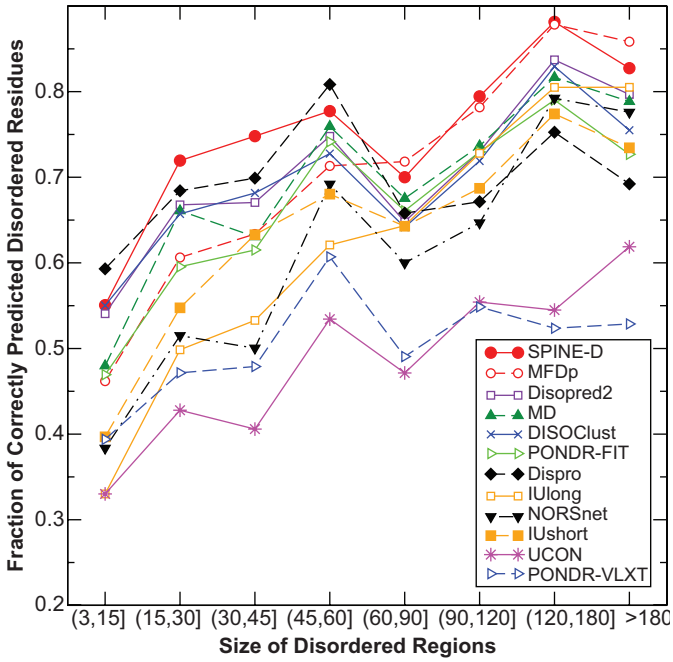


**Figure 3:** Comparison of fraction of correctly predicted disordered residues as a function of the size of disordered regions for different methods as labeled on the SL329 dataset.

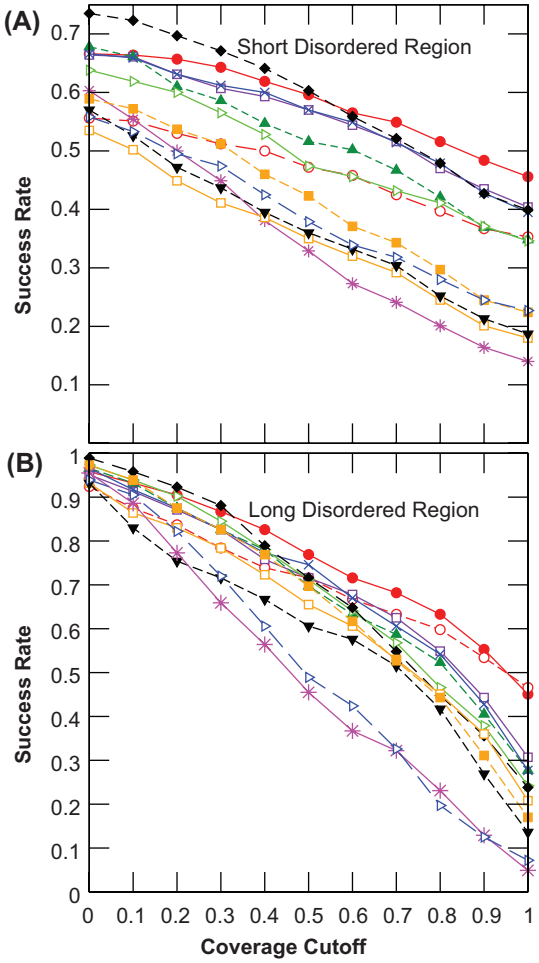


**Figure 4:** The fraction of correctly predicted disordered regions (success rate) as a function of the minimum coverage that defines a region as correctly predicted. (**A**) Short disordered regions (**B**) Long disordered regions. Various methods as labeled in Figure 3.

Figure 5 presents the results of such analysis for several illustrative examples of proteins with experimentally validated MoRF regions (eukaryotic translation initiation factor 4E-binding protein 1 (4E-BP1), human p53, *E. coli* RNase E, measles virus nucleoprotein, and axin, UniProt IDs: Q13541, P04637, P21513, P10050, and O15169, respectively). In all plots, disorder scores (the probability of a given residue being disordered) are shown as a function of residue index. Red, blue and green curves present the results of PONDR®VLXT, SPINE-D, and PONDR-FIT, respectively. SPINE-D and PONDR-FIT have a comparable performance for 4E-BP1 (Q13541, Figure 5A), human p53 (P04637, Figure 5B) and Axin (O15169, Figure 5E). For 4E-BP1 and axin, the performances of SPINE-D and PONDR-FIT are similar to that of PONDR®VLXT. For p53, both SPINE-D and PONDR-FIT can see the N-terminal binding site and miss the C-terminal binding motif. Figure 5C shows that PONDR-FIT outperformed SPINE-D in the case RNase E (P21513) and was similar to the PONDR®VLXT outputs for this protein. On the other hand, Figure 5D illustrates that SPINE-D can detect a C-terminal binding site in the measles virus nucleoprotein (P10050), whereas PONDR-FIT missed this binding site. Therefore, these data suggest that SPINE-D and PONDR-FIT have comparable sensitivity for potential binding sites and both of them are still less sensitive than PONDR®VLXT, subjected to limitation of a small sample. It should also be mentioned that PONDR®VLXT has the largest fluctuations in terms of disorder probability along the sequence. Hence, though we find it produces the most true positives in our test-cases, it will also most probably produce the most false positives as well. Thus, a more careful comparison will be possible only with a larger dataset.

## *Discussion*

We presented a sequence-based disorder predictor SPINE-D that utilized a single-neural-network-based technique. Multiple independent tests (DM1229, CASP 9, and SL329) confirm the robustness in its performance

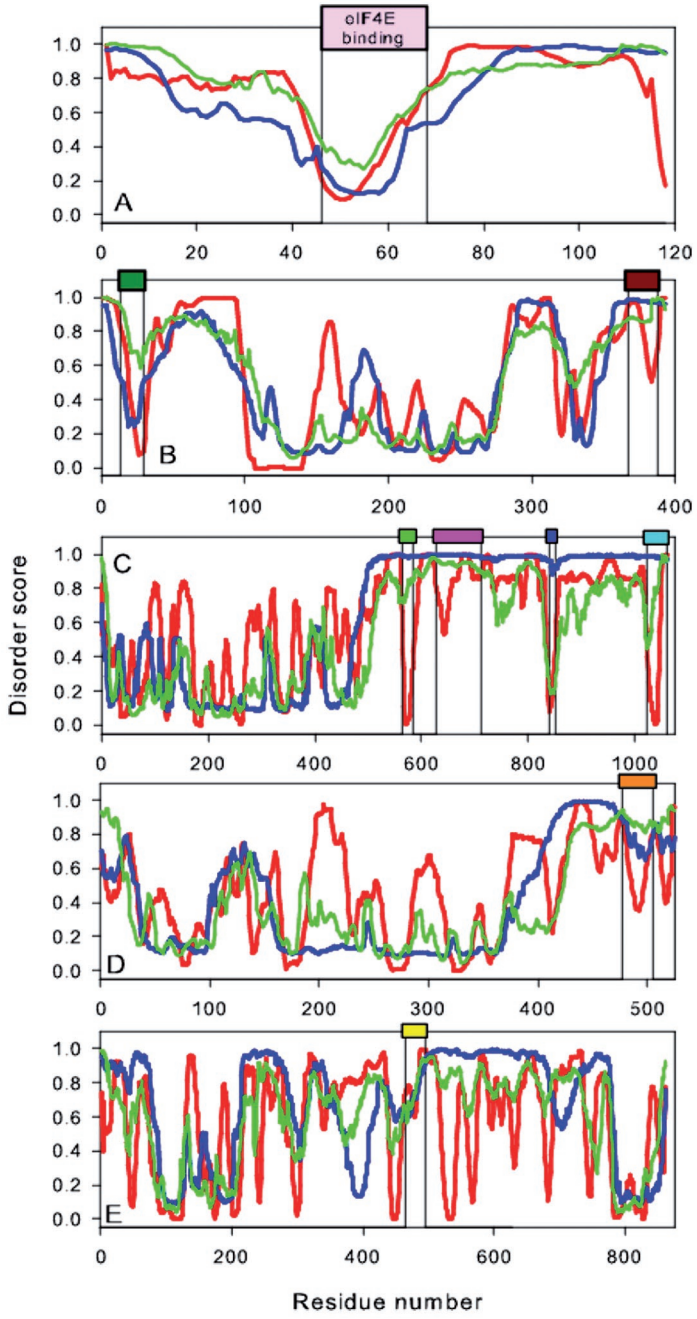


**Figure 5:** Analysis of the disorder propensities of 4E-BP1 (Q13541) (**A**), human p53 (P04637) (**B**), *E. coli* RNase E (P21513) (**C**), measles virus nucleoprotein (P10050) (**D**), and axin (O15169) (**E**), by different predictors of intrinsic disorder, PONDR®VLXT (red curve), SPINE-D (blue curve) and PONDR-FIT (green curve). Bars at the top of each panel indicate the experimentally validated regions of corresponding disordered proteins involved in interaction with specific binding partners: **A**. Pink bar shows 4E-BP1 region (residues 45-65) involved in binding of the eukaryotic initiation factor 4E (eIF4E); **B**. Dark green and dark red bars indicate the p53 regions (residues 13-29 and 367-388) involved in binding of Mdm2 and S100B(ββ), respectively; **C**. Bars indicate regions responsible for the RNase E interaction with different binding partners: green bar (residues 565-585), protein-RNA interaction site; pink bar (residues 633-712), self-recognition region; blue bar (residues 839-850), enolase binding site; and cyan bar (residues 1021-1061), PNPase binding site; **D**. Orange bar shows the nucleoprotein region (residues 477-505) responsible for the interaction with phsosphoprotein; **E**. Yellow bar corresponds to the axin region (residues 464-495) involved in interaction with the β-catenin.

as a single method that matches or exceeds meta (or consensus) predictors. In particular, we found that its superior performance is due to its ability to make highly accurate prediction in both short and long disordered regions at the same time.

The overall performance of SPINE-D is built on our previous accurate prediction of secondary structure (82% ten-fold cross-validated accuracy for three-state prediction) (34, 43), solvent accessibility (correlation coefficient of 0.74 between predicted and actual values) (36), and fluctuations of backbone torsion angles (33) with the same neural network architecture. Moreover, it makes a three-state prediction that successfully captures the difference between short and long disordered regions. In addition, we found that repeated training of disordered residues improves our results. It increases AUC from 0.837 to 0.853 for ten fold cross validation of DM3000 and reduces training time from 5.5 hours to 4 hours because of faster convergence. Furthermore, an average over five independent runs leads to a slight increase of AUC (from 0.853 to 0.858).

It is of interest to know how predicted residues in short and long disordered regions contribute separately to the overall accuracy of SPINE-D. For DM1229, $Q_S = 73\%$ with 59% and 14% coming from residues predicted to be in short and long disordered regions, respectively. $Q_L = 65\%$ with 17% and 48% coming from residues predicted to be in short and long disordered regions, respectively. That is, the contribution to $Q_S$ ($Q_L$) mainly comes from residues predicted to be in short (long) disordered regions as expected. The clearer correspondence is seen for long disordered regions in SL329 where $Q_L = 81\%$ with 9% and 72% coming from residues predicted to be in short and long disordered regions, respectively. For short disordered regions in SL329 we found $Q_S = 64\%$ with 34% and 30% coming from residues predicted to be in short and long disordered regions, respectively. The increased confusion in this case is probably due to the limited number of short disordered regions in SL329. Overall, these results are consistent with our previous observations showing that the prediction accuracy for long and short disordered regions depends on the ratio between the numbers of long and short disordered regions in different databases.

Some previous studies suggested that short disordered regions are more difficult to predict than long disordered regions (20, 62) while CASP 9 assessment suggests that capability of disorder prediction decreases with increasing disordered segment length (42). Our results (Figure 1 and Table I) indicate that this behavior depends on the balance in the populations of disordered and ordered residues in the dataset and the size of dataset. To further confirm the ability of predicting both short and long disordered regions, we calculated predicted and actual compositions of amino acid residues for short and long disordered regions in SL329. We further evaluated the root mean square difference (rmsf) between the two compositions by $\text{rmsf} = \sqrt{\sum \left(f_i^p - f_i^a\right)^2 / 20}$ where the summation is over 20 amino acid types, $f_i^a$ and $f_i^p$ are the fractions of amino acid residue type $i$, in all annotated and predicted disordered residues, respectively. We found that the root mean square difference between predicted and actual compositions for short and long disordered regions are 0.0052 and 0.0027, respectively, significantly smaller than 0.0106 between predicted short and long disordered regions and 0.0079 between actual short and long disordered regions (p values $< 10^{-15}$). That is, SPINE-D is capable of capturing the composition difference between short and long disordered regions automatically. However, the low probability of having a long disordered region in X-ray structure (63) makes the assessment of method accuracy at this region unreliable for a small dataset such as CASP 9.

We would also like to note that despite having only 10% disordered residues in DM4229 or DX4080, a neural network trained on it can perform well on a balanced dataset ($>$40% disordered residues) extracted from Disprot (SL329 and SL477). This indicates that it is not necessary to have a balanced dataset for training. In fact cross-validated training on SL477 has similar performance as training on imbalanced data set of DX4080 (Table I). Training on a much larger number of short

disordered regions in DX4080 improves prediction of short disordered regions in SL477 over use of SL477 in training.

Another interesting result from our work is that disordered residues derived from X-ray structures are sufficient to make accurate prediction of manually annotated disorder in the Disprot database including fully disordered proteins. This indicates that there is a level of consistency for annotations from X-ray structures and from Disprot. We found that overall compositions of amino acid residues between X-ray derived disordered residues (DX4080) and Disprot disordered residues (SL447) are highly correlated with a correlation coefficient of 0.92. Only Pro ( $f^{a}_{pro} = 0.057$ in DX4080 vs. 0.080 in SL477), His (0.043 vs. 0.019), Glu (0.079 vs. 0.099), Gly (0.090 vs. 0.073) can be considered as moderate outliers. This further indicates the similarity between the annotation used in the two different databases.

Finally, our SPINE-D server was designed for identifying short and long disordered regions. It was not trained to predict short ordered regions in long disordered regions, the MoRF regions recently found to be of functional significance (57-59). An examination of five proteins suggests that our predicted probabilities for disorder are smoother than PONDR®VLXT (Figure 5) as a result of training to predict entire long disordered regions as disordered. Thus, it is not surprising that SPINE-D seems less sensitive in detecting these MoRF regions. On the other hand, higher sensitivity (as in PONDR®VLXT) would potentially lead to lower specificity. Obviously, more systematic studies and larger datasets are required in this important area.

***Acknowledgements***

We thank Yuedong Yang for many helpful suggestions and Marcin J. Mizianty for running their MFDp predictor for us. This work was supported by the National Institutes of Health grants GM R01 085003 and GM R01 067168 (Co-PI) to Y. Z., the National Natural Science Foundation of China (grant no. 61003187, 61170099) to T. Z., the National Science Foundation [EF 0849803 to A. K. D and V. N. U.], and the Russian Academy of Sciences ["Molecular and Cellular Biology" Program to V. N. U.].

***References***

1. V. N. Uversky, C. J. Oldfield, and A. K. Dunker. *J Mol Recognit 18*, 343-384 (2005).
2. J. Liu, N. B. Perumal, C. J. Oldfield, E. W. Su, V. N. Uversky, and A. K. Dunker. *Biochemistry-Us 45*, 6873-6888 (2006).
3. C. A. Galea, Y. Wang, S. G. Sivakolundu, and R. W. Kriwacki. *Biochemistry-Us 47*, 7598-7609 (2008).
4. M. Fuxreiter, P. Tompa, I. Simon, V. N. Uversky, J. C. Hansen, and F. J. Asturias. *Nature Chemical Biology 4*, 728-737 (2008).
5. A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky. *The FEBS Journal 272*, 5129-5148 (2005).
6. P. E. Wright and H. J. Dyson. *J Mol Biol 293*, 321-331 (1999).
7. H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, and Z. Obradovic. *J Proteome Res 6*, 1882-1898 (2007).
8. A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown. *Genome informatics. Workshop on Genome Informatics 11*, 161-171 (2000).
9. J. L. Sussman, A. K. Dunker, I. Silman, and V. N. Uversky. *Curr Opin Struc Biol 18*, 756-764 (2008).
10. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. *J Mol Biol 337*, 635-645 (2004).
11. L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, and A. K. Dunker. *J Mol Biol 323*, 573-584 (2002).
12. S. Raychaudhuri, S. Dey, N. P. Bhattacharyya, and D. Mukhopadhyay. *PLoS One 4*, e5566 (2009).
13. V. N. Uversky, C. J. Oldfield, and A. K. Dunker. *Ann Rev Biophys 37*, 215-246 (2008).
14. D. Eliezer. *Curr Opin Struc Biol 19*, 23-30 (2009).
15. S. Longhi, J. M. Bourhis, and B. Canard. *Curr Protein Pept Sc 8*, 135-149 (2007).

16. K. K. Turoverov, I. M. Kuznetsova, and V. N. Uversky. *Prog Biophys Mol Bio 102*, 73-84 (2010).
17. V. N. Uversky and A. K. Dunker. *Bba-Proteins Proteom 1804*, 1231-1264 (2010).
18. P. E. Wright and H. J. Dyson. *Curr Opin Struc Biol 19*, 31-38 (2009).
19. V. Receveur-Brechot, J. M. Bourhis, V. N. Uversky, B. Canard, and S. Longhi. *Proteins 62*, 24-45 (2006).
20. B. He, K. J. Wang, Y. L. Liu, B. Xue, V. N. Uversky, and A. K. Dunker. *Cell Res 19*, 929-949 (2009).
21. K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic. *BMC Bioinformatics 7*, 208 (2006).
22. P. Radivojac, Z. Obradovic, D. K. Smith, G. Zhu, S. Vucetic, C. J. Brown, J. D. Lawson, and A. K. Dunker. *Protein Sci 13*, 71-80 (2004).
23. P. Romero, Z. Obradovic, and K. Dunker. *Genome Informatics. Workshop on Genome Informatics 8*, 110-124 (1997).
24. J. L. Cheng, M. J. Sweredoski, and P. Baldi. *Data Min Knowl Disc 11*, 213-222 (2005).
25. Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon. *Bioinformatics 21*, 3433-3434 (2005).
26. Z. P. Feng, P. F. Han, and X. Z. Zhang. *BMC Bioinformatics 10* (2009).
27. S. Hirose, K. Shimizu, S. Kanai, Y. Kuroda, and T. Noguchi. *Bioinformatics 23*, 2046-2053 (2007).
28. M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani, and L. Kurgan. *Bioinformatics 26*, i489-i496 (2010).
29. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, and A. K. Dunker. *Proteins 61*(*Suppl. 7*), 176-182 (2005).
30. G. Pollastri, A. Vullo, O. Bortolami, and S. C. E. Tosatto. *Nucleic Acids Research 34*, W164-W168 (2006).
31. P. Romero, Z. Obradovic, X. H. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. *Proteins-Structure Function and Genetics 42*, 38-48 (2001).
32. K. Shimizu, S. Hirose, and T. Noguchi. *Bioinformatics 23*, 2337-2338 (2007).
33. T. Zhang, E. Faraggi, and Y. Q. Zhou. *Proteins 78*, 3353-3362 (2010).
34. E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou. *J Computational Chemistry*, doi: 10.1002/jcc.21968 (2012). (Epub ahead of print)
35. O. Dor and Y. Zhou. *Proteins 66*, 838-845 (2007).
36. E. Faraggi, B. Xue, and Y. Zhou. *Proteins 74*, 847-856 (2009).
37. A. Schlessinger, M. Punta, and B. Rost. *Bioinformatics 23*, 2376-2384 (2007).
38. F. L. Sirota, H. S. Ooi, T. Gattermayer, G. Schneider, F. Eisenhaber, and S. Maurer-Stroh. *BMC Genomics 11*(*Suppl. 1*), S15 (2010).
39. S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic. *Proteins 52*, 573-584 (2003).
40. M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and A. K. Dunker. *Nucleic Acids Research 35*, D786-793 (2007).
41. A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost. *PLoS One 4* (2009).
42. B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, and A. Kryshtafovych. *Proteins 79* (*Suppl. 10*), 107-18 (2011).
43. E. Faraggi, Y. D. Yang, S. S. Zhang, and Y. Zhou. *Structure 17*, 1515-1527 (2009).
44. J. Zupan. *Acta Chimica Slovenica 41*, 327-354 (1994).
45. J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke. *J Mol Model 7*, 360-369 (2001).
46. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. *Nucleic Acids Research 25*, 3389-3402 (1997).
47. J. C. Wootton. *Curr Opin Struc Biol 4*, 413-421 (1994).
48. 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. *Bioinformatics 21*, 137-140 (2005).
49. O. Noivirt-Brik, J. Prilusky, and J. L. Sussman. *Proteins 77* (*Suppl. 9*), 210-216 (2009).
50. L. J. McGuffin. *Bioinformatics 24*, 1798-1804 (2008).
51. X. Li, P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic. *Genome Informatics. Workshop on Genome Informatics 10*, 30-40 (1999).
52. J. Hecker, J. Y. Yang, and J. L. Cheng. *BMC Genomics 9* (2007).
53. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. *J Mol Biol 337*, 635-645 (2004).
54. A. Schlessinger, J. Liu, and B. Rost. *PLoS Computational Biology 3*, e140 (2007).
55. B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker, and V. N. Uversky. *Biochimica et Biophysica Acta 1804*, 996-1010 (2010).
56. M. J. Mizianty, T. Zhang, B. Xue, Y. Q. Zhou, A. K. Dunker, V. N. Uversky, and L. Kurgan. *BMC Bioinformatics 12* (2011).
57. A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky. *J Mol Biol 362*, 1043-1059 (2006).
58. C. J. Oldfield, Y. G. Cheng, M. S. Cortese, P. Romero, V. N. Uversky, and A. K. Dunker. *Biochemistry-Us 44*, 12454-12470 (2005).

59. V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky, and A. K. Dunker. *J Proteome Res 6*, 2351-2366 (2007).
60. Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky, and A. K. Dunker. *Biochemistry-Us 46*, 13468-13477 (2007).
61. E. Garner, P. Romero, A. K. Dunker, C. Brown, and Z. Obradovic. *Genome Inform Ser Workshop Genome Inform 10*, 41-50 (1999).
62. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, and A. K. Dunker. *Proteins-Structure Function and Genetics 53*, 566-572 (2003).
63. T. Le Gall, P. R. Romero, M. S. Cortese, V. N. Uversky, and A. K. Dunker. *J Biomol Struct Dyn 24*, 325-341 (2007).

*Date Received: June 14, 2011*

***Communicated by the Editor Ramaswamy H. Sarma***