Intrinsically Disorder Protein Prediction using Undersampling Feedforward Neural Networks and Predicted Amino Acid Features

Qiaoyi Li[1], Steven Pascal[2], and Yaohang Li[3]

[1]Ocean Lakes High School Math and Science Academy, Virginia Beach VA
[2]Department of Chemistry, Old Dominion University, Norfolk VA
[3]Department of Computer Science, Old Dominion University, Norfolk VA

Keywords: Intrinsically disorder protein, prediction, neural network, balanced training set

Abstract

In recent years, intrinsically disorder proteins have been found to be related to different types of diseases and serve important biological functions in organisms. Thus, disorder prediction has become ever more important in the understanding and modeling of these proteins. This disorder region prediction model is a neural network using around one million data samples with 525 context-based features. One problem in the use of neural network prediction methods is an unbalance of data between outputs (much lower ratio of disorder). Such data results in high discrepancies between sensitivity and specificity in order and disorder classes. By utilizing undersampling techniques to train a neural network, both sensitivity and specificity of predictions in disorder class are improved. Neural networks trained with original data reach only 25.0% sensitivity in disorder class while the ones trained with 1:1 ratio undersampling achieve 70.6% sensitivity and 76.4% specificity. Our predictions on Spinach Thylakoid Soluble Phosphoprotein (TSP9) and Prostate apoptosis response factor-4 (Par-4) agree with the NMR experimental results.

1. INTRODUCTION

A challenge in protein structure prediction is the recognition and modeling of intrinsically disorder proteins (IDPs), also called intrinsically unstructured or naturally unfolded proteins. Disorder regions of proteins do not display a stable tertiary structure when the polypeptide is isolated in vitro. Their dynamic structure results from frequent variation of phi and psi angles in the protein backbone. Due to the unique, flexible nature of intrinsically disorder proteins, they can serve important biological functions and play a role in neurological disease.

Theoretically, IDPs contain a high proportion of polar and charged amino acid residues, which prevent the protein from establishing a stable globular structure. Disorder proteins are often low-complexity, featuring repetition of a few amino acids. However, not all low-complexity protein sequences are intrinsically disorder. Another characteristic of IDPs is low contents of stable secondary structure such as α-helices and β-sheets. All disorder regions are found within segments of coiled secondary structures (loops). Some of these structures are called "hot loops" [Linding et al., 2003] due to their mobility from high C-α temperature factors. While structured proteins do move continuously due to kinetic and thermal energy, disorder polypeptides have much higher B-factors signifying a more dynamic structure.

The greatest advantage of IDPs is their flexibility, which facilitates binding to target molecules of various sizes and shapes. Intrinsic disorder occurs more frequently in proteins involved in cell signaling, transcription, and chromatin remodeling. Some disorder proteins link together two globular or trans-membrane domains. IDPs also form fuzzy complexes where their structural uncertainty is essential for function. In coupled folding and binding, disorder proteins fold into a more stable structure after attachment to another macromolecule. The combination of folding and binding enhance selective abilities of the molecules and speed up the binding process.

There are many existing methods for the prediction of disorder regions in proteins. Most methods of prediction mostly consider amino acid sequencing information (*ab initio* methods). An early ab initio method SEG used areas composed of low complexity of sequencing with some success with limited success since not all low complexity sequences are disorder. Some early sequence based predictors (e.g. HCA, PreLink, FoldIndex) use mainly hydrophobicity and hydropathy to predict disorder based on their low content of hydrophobicity. However, this method is more suited to predicting shorter segments of disorder

[Deng et al., 2012]. GlobPlot takes a simple approach by using a running sum of the amino acid's propensity to be disorder or ordered (from Russell/Linding scale of disorder) to predict the segment's propensity for globularity. The propensities are then plotted to show different areas of disorder/order. The method's simplicity only allows it to provide a general estimate [Linding et al., 2003]. PONDR, a collection of three predictors (VL-XT, XLT, and CaN) incorporated and combined other features of disorder such as flexibility as well as sequence complexity and hydropathy in the analysis of amino acid sequencing through a feedforward neural network [Li et al., 1999]. DisEMBL uses a neural network trained on structure data from X-ray crystallography that can predict regions of loops and "hot loops" with 64% sensitivity [Linding et al., 2003]. The PSI-BLAST algorithm has contributed to the improvement of predictions. Based on CASP10 results in 2012, Diopred3 was ranked second in Disorder Protein Prediction. It uses PSI-BLAST to create sequence profiles for each protein target which then trains linear support vector machines, reaching accuracies of 70%. Prdos-CNF [Eickholt and Cheng, 2013], the highest ranked overall disorder protein predictor based on the CASP10 report, uses a recursive neural network with inputs of PSI-BLAST arranged profile, predicted secondary structure, and solvent accessibility. CASP10 results showed an accuracy of 71% and a precision of 70%.

Other predictor methods include clustering and meta methods. Clustering methods including DISOclust [McGuffin, 2008] predict a tertiary structure which is then layered over the target protein to calculate a probability of disorder. These meta methods include metaPrDOS2, which acquired the best accuracy of 77.8% in CASP10 with a sensitivity of 64.73% and a specificity of 89.4% [Monastyrskyy et al., 2014].

2. METHODS

2.1. Dataset

Overall, 121,298 disorder residue samples are obtained from the disorder protein database Disprot [Sikmeier et al., 2007]. 1,010,750 structured residue samples are extracted from a set of chains with maximum 25% pair-wise sequence identity, 1.8A resolution, and 0.25 R-factor generated by the PISCES server [Wang and Dunbrack, 2003]. These data samples are mixed together and then are randomly split into two disjoint sets. Using a preprogrammed feedforward neural network training application in MATLAB written for pattern recognition, the first fold and second fold data sets are applied alternatively for training and validating the neural network of 200 neurons (i.e. first fold is used in training and second fold for testing and vice versa).

2.2. Neural Network Encoding

For each data sample, a sliding window of 21 residues is selected, where the feedforward neural network is trained to predict if the centered residue in the window is disorder. Each residue is represented by 20 PSSM values, 1 boundary value indicating C- or N-terminals overlap, 3 values of secondary structure probabilities predicted by SCORPION [Ashraf and Li, 2014], and 1 value of solvent accessibility probability predicted by CASA [Ashraf and Li, 2014]. Putting every feature together, there are totally 525 features associated with each residue sample. Figure 1 shows the neural network encoding scheme.

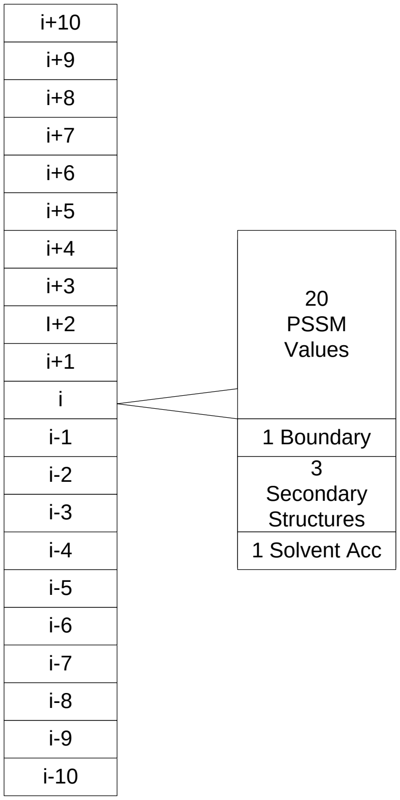

Figure 1: Neural Network Encoding Scheme for Intrinsically Disorder Region Prediction

2.3. Balancing the Training Set

Due to proportionally low occurrence of IDPs, the training was unbalanced. A low ratio of one output class results in higher identification accuracy for the majority output and lower accuracy of the minority output. The method undersampling alleviates unbalanced training by utilizing fewer structured data samples in the training data set. Undersampling ratios of ordered to disorder data samples used to train the neural networks are 1:1, 3:1, and 6:1 respectively. All three undersampling ratios are applied to both data sets yielding six unique sets. Each set is used to train the feedforward neural network and non-undersampled sets are utilized for cross validation.

3. RESULTS

3.1. Prediction Accuracy and Sensitivity

Table 1 compares sensitivity and specificity of the trained neural networks using various undersampling strategies. In the original training data, the neural network predictions biases to the class of “ordered” residue because the ordered residue samples significantly outnumber the disorder ones. When the undersampling strategies are employed, such sampling biases are reduced. Hence both sensitivity and specificity of disorder residue predictions increase with the price of relatively small reduction of sensitivity and specificity of ordered residue predictions. When the numbers of ordered and disorder samples are approximately equal (1:1 undersampling), the sensitivity and specificity of both disorder and ordered predictions are above 70%.

Table 1. Comparison of original and undersampling methods in neural network training and testing

		Sensitivity (%)		**Specificity (%)**		Total Acc.
		Disorder	Ordered	Disorder	Ordered	
Original	Train	26.0	98.7	68.9	92.3	91.4
	Test	25.3	98.6	66.8	92.3	91.3
Undersampling 6:1	Train	32.3	98.0	72.8	89.6	88.5
	Test	31.7	97.9	70.8	89.9	88.7
Undersampling 3:1	Train	46.3	94.5	73.8	84.0	82.4
	Test	45.2	94.3	72.4	83.9	82.1
Undersampling 1:1	Train	71.0	78.3	76.6	73.0	74.7
	Test	70.6	77.8	76.4	72.3	74.2

3.2. Prediction on Spinach Thylakoid Soluble Phosphoprotein (TSP9)

Thylakoid Soluble Phosphoprotein (TSP9) (PDB ID: 2FFT) is a plant-specific protein in the photosynthetic thylakoid membrane, which is disorder under aqueous conditions discovered by NMR spectroscopy [Song et al., 2006]. Figure 2 depicts a 3D NMR model of TSP9. The 15-23 residues form a small α-helix, which is a potential binding site and the rest of the protein chain is disorder. Figure 3 shows the secondary structure assignment by NMR and the disorder regions predicted by our neural network trained with 1:1 undersampling. One can find that our neural network is able to correctly identify the region of the small α-helix as ordered and the majority of the disorder loops. There are a few mispredictions in the disorder loops, which may be corrected by an additional trained neural network for refinement.

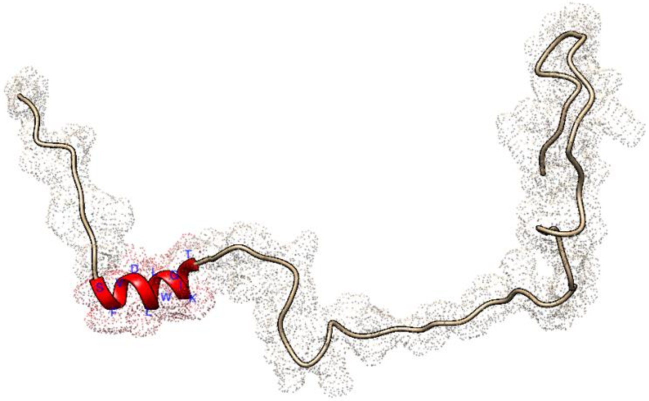

Figure 2: 3D NMR Model of Spinach Thylakoid Soluble Phosphoprotein (TSP9)

```
1       SAAKGTAETK QEKSFVDWLL GKITKEDQFY
                      SHHHHHH HHHS    S
        DDDDDDDDDO OOOOOOOOOO OOODDDDDDO

31      ETDPILRGGD VKSSGSTSGK KGGTTSGKKG
         S  SSS S      S        S S S  S
        DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

61      TVSIPSKKKN GNGGVFGGLF AKKD
        SSSS SS S   S  SSSSSS
        ODDDDDDDDD DDDDOOOOOO ODDD
```

Figure 3: Secondary Structure Assignment and Predicted Disorder Regions in Spinach Thylakoid Soluble Phosphoprotein (TSP9)

3.3. Prediction on Prostate apoptosis response factor-4 (Par-4)

Prostate apoptosis response factor-4 (Par-4) is a proapoptotic protein coding for tumor-suppression. We apply our trained neural network for disorder region prediction with 1:1 undersampling to Par-4, which is shown in Figure 4. Our neural network predicts that residues from

31 to 138 belong to a long disorder region, which agrees with the experimental results [Libich et al., 2009] observed by NMR spectroscopy.

```
1       MATGGYRSSG STTDFLEEWK AKREKMRAKQ
        DDDDDODDDD DDOOOOOOOO OOOOOOOOOO

31      NPVGPGSSGG DPAAKSPAGP LAQTTAAGTS
        DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

61      ELNHGPAGAA APAAPGPGAL NCAHGSSALP
        DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

91      RGAPGSRRPE DECPIAAGAA GAPASRGDEE
        DDDDDDDDDD DDDDDDDDDD DDDDDDDDDD

121     EPDSAPEKGR SSGPSARKGK GQIEKRKLRE
        DDDDDDDDDD DDDDDDDDOO OOOOOOOOOO

151     KRRSTGVVNI PAAECLDEYE DDEAGQKERK
        OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

181     REDAITQQNT IQNEAASLPD PGTSYLPQDP
        OOOOOOOOOO OOOOOOOODD DOOOOODDOO

211     SRTVPGRYKS TISAPEEEIL NRYPRTDRSG
        OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

241     FSRHNRDTSA PANFASSSTL EKRIEDLEKE
        OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

271     VLRERQENLR LTRLMQDKEE MIGKLKEEID
        OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

301     LLNRDLDDME DENEQLKQEN KTLLKVVGQL
        OOOOOOOOOO OOOOOOOOOO OOOOOOOOOO

331     TR
        OO
```

Figure 4: Disorder Region Prediction in Par-4

4. CONCLUSIONS

Using a pre-established, unspecialized MATLAB pattern recognition application for neural network training, the network has achieved a best sensitivity of 70.6% and a best total accuracy of 76.4% in predicting disorder residues. Best results originated from using 1:1 ratio under sampled data. By balancing neural network training through undersampling, sensitivity and specificity became closer to each other thus improving the sensitivity where networks would usually be skewed toward structured proteins. Our predictions show good agreements with the NMR experimental results on Spinach Thylakoid Soluble Phosphoprotein (TSP9) and Prostate apoptosis response factor-4 (Par-4).

There are a lot of space for this method to be improved. For example, using oversampling instead of undersampling on a powerful computer may lead to further prediction accuracy improvements. Incorporation of additional predicted features such as B-factor profiles, inter-residue contacts, and disulfide bonding states may be helpful for training a more effective neural network. Advanced machine learning algorithms such as deep learning, if properly used, may also lead to advancement in disorder region prediction. All these will be our future research directions.

REFERENCES

1. B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, K. Fidelis, “Assessment of Protein Disorder Region Predictions in CASP 10.” *Proteins: Structure, Function and Bioinformatics*, 82: 127-137, 2014.
2. X. Li, P. Romero, M. Rani, A. K. Dunker, Z. Obradovic, “Predicting Protein Disorder for n-, c-, and Internal Regions,” Genome informatics, 10: 30-40, 1999.
3. R. Linding, L. J. Jenson, F. Diella, P. Bork, T. J. Gibson, R. Russell, “Protein Disorder Prediction: Implications for Structural Promteomics,” Structure, 11: 1453-1459, 2003.
4. R. Linding, R. Russell, V. Neduva, T. Gibson, “Globplot: Exploring Protein Sequences for Globularity and Disorder,” Nucleic Acid Research, 31: 3701-3708, 2003.
5. X. Deng, J. Eickholt, J. Cheng, “A Comprhensive Overview of Computational Protein Disorder Prediction Methods,” Mol. Biosyst., 1: 114-121, 2012.
6. A. Yaseen, Y. Li, “Context-based Features Enhance Protein Secondary Structure Prediction Accuracy,” Journal of Chemical Information and Modeling, in press, 2014.
7. A. Yaseen, Y. Li, “CASA: A Protein Solvent Accessibility Prediction Server using Context-based

Features to Enhance Prediction Accuracy," BMC Bioinformatics, in press, 2014.

8. J. Song, M. S. Lee, I. Carlberg, A. V. Vener, J. L. Markley, "Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications," Biochemistry, 45: 15633-15643, 2006.
9. D. S. Libich, M. Schwalbe, S. Kate, H. Venugopal, J. K. Claridge, P. J. B. Edwards, K. Dutta, S. M. Pascal, "Intrinsic disorder and coiled-coil formation in prostate apoptosis response factor 4," the FEBS Journal, 276: 3710-3728, 2009.
10. 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, A. K. Dunker, "DisProt: the Database of Disorder Proteins," Nucleic Acids Res., 35: D786-793, 2007.
11. G. L. Wang, R. L. Dunbrack, "PISCES: a protein sequence culling server," Bioinformatics, 19: 1589-1591, 2003.
12. J. Eickholt, J. Cheng, "DNdisorder: predicting protein disorder using boosting and deep networks," BMC Bioinformatics, 14: 88, 2013.
13. L. J. McGuffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models," Bioinformatics, 24: 1798-804, 2008.