Vol. 17, pp. 147–154 (2020)
doi: 10.2142/biophysico.BSJ-2020026

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

*Database and Computer Program*

# NeProc predicts binding segments in intrinsically disordered regions without learning binding region sequences

Hiroto Anbo, Hiroki Amagai and Satoshi Fukuchi

*Department of Life Science and Informatics, Faculty of Engineering, Maebashi Institute of Technology, Maebashi, Gunma 371-0816, Japan*

Received September 15, 2020; accepted October 29, 2020; Released online in J-STAGE as advance publication November 3, 2020

**Intrinsically disordered proteins are those proteins with intrinsically disordered regions. One of the unique characteristics of intrinsically disordered proteins is the existence of functional segments in intrinsically disordered regions. These segments are involved in binding to partner molecules, such as protein and DNA, and play important roles in signaling pathways and/or transcriptional regulation. Although there are databases that gather information on such disordered binding regions, data remain limited. Therefore, it is desirable to develop programs to predict the disordered binding regions without using data for the binding regions. We developed a program, NeProc, to predict the disordered binding regions, which can be regarded as intrinsically disordered regions with a structural propensity. We only used data for the structural domains and intrinsically disordered regions to detect such regions. NeProc accepts a query amino acid sequence converted into a position specific score matrix, and uses two neural networks that employ different window sizes, a neural network of short windows, and a neural network of long windows. The performance of NeProc was comparable to that of existing programs of the disordered binding region prediction. This result presents the possibility to overcome the shortage of the disordered binding region data in the development of the prediction programs for these binding regions. NeProc is available at http://flab.neproc.org/neproc/index.html**

**Key words:** intrinsically disordered protein, binding regions, structure prediction, neural network

## Introduction

Intrinsically disordered proteins (IDPs) contain regions that lack unique three-dimensional (3D) structures under physiological conditions [1–3]. Computer programs can be used to predict intrinsically disordered regions (IDRs) using amino acid sequences as input data. *In silico* analysis has disclosed the nature of IDPs, and has revealed that they are abundant in eukaryotes, in residues with frequent post-translational modifications, such as phosphorylation, and that nuclear proteins contain many IDRs [4,5]. Many programs have been developed to predict IDRs, and their performance has improved [6–9].

A unique feature of IDPs is their ability to bind to partner proteins and/or other biological molecules. The regions involved in binding are short segments that range

Corresponding author: Satoshi Fukuchi, Department of Life Science and Informatics, Faculty of Engineering, Maebashi Institute of Technology, Kamisadori 460-1, Maebashi, Gunma 371-0816, Japan. e-mail: sfukuchi@maebashi-it.ac.jp

◀ ***Significance*** ▶

The protein binding regions found in the intrinsically disordered regions play pivotal roles in important biological processes. We developed NeProc to predict such binding regions without using experimental data for binding regions. NeProc demonstrated a performance comparable to state-of-the-art programs. Since data for such binding regions remains limited, out findings highlight a possible method for overcoming the shortage of binding region data in developing prediction programs for disordered binding regions.

from a few to tens of residues, and some can adopt local two-dimensional structures based on binding. This phenomenon is referred to as the coupled folding and binding mechanism [10], and through this mechanism, IDRs play crucial roles in various biological processes, such as signal transduction and transcriptional regulation [2,3,11,12]. Interactions via IDRs are involved in liquid-liquid phase separation, which is important in many biological processes and diseases [13–16].

Although some programs can predict both IDRs and binding segments in IDRs [7,17], there is scope for improvement in predicting disordered binding regions. However, one of the difficulties in predicting these regions is the lack of defined examples of disordered binding regions. Many programs predicting IDRs employ machine learning approaches, in which increased training data enables better performance. Some databases collect such disordered binding regions together with their structures in the binding complexes. These binding regions are termed molecular recognition features (MORFs) [18], short linear motifs (SLiMs) [19], and disordered binding sites (DIBSs) [20]. We also developed the IDP database IDEAL [21,22], that collects IDRs and experimentally verified disordered binding regions, which are termed protean segments (ProSs). Despite an effort to experimentally determine such disordered binding regions, the number of examples is insufficient; IDEAL has number of 146,276 ordered residues, 33,053 disordered residues, and 9,444 ProS residues. Considering their importance, there must be more interactions via IDRs that require accurate programs to predict disordered binding regions.

When IDRs were predicted in an amino acid sequence, we observed that disordered binding regions demonstrated some structural propensity. These binding regions are reported to have greater similarity to structural domains (SDs) than IDRs in terms of sequence conservation [23], and exhibit a mixture characteristics of SDs and IDRs in the amino acid composition [24]. Thus, disordered binding regions can be defined as short segments with structural properties located in long IDRs. If a segment with structural propensity is identified in a long IDR, the binding regions in IDRs can be predicted. Considering the limited examples of disordered binding regions and the abundance of SD data, we developed a new program, NeProc (Next ProSs Classifier), to predict binding regions in IDRs without using data for disordered binding regions.

## Materials and Methods

We aimed to predict the binding regions in IDRs without using the disordered binding region data. We developed NeProc by employing similar methods used in the reference programs in order to validate the impact of the training data. ANCHOR2 [8], DISOPRED3 [7], and MoRFchibi-Web [25] were selected as the references. ANCHOR2 uses the statistical potential of the program IUpred2 [8] to predict the disordered binding regions. The potential estimates residue pair contact using the SD. In this sense, ANCHOR2 is similar to NeProc in that both use the SD data, but ANCHOR2 does not use machine learning techniques. The other programs, DISOPRED3 and MoRFchibi-Web, employ the traditional neural network and/or the support vector machine with the use of the disordered binding regions as the training data. We developed NeProc using similar methods to these two programs; thus, these two programs are similar to NeProc in terms of the methods used, but are dissimilar in terms of training data. NeProc predicts disordered binding regions by identifying segments with structural propensity in IDRs. NeProc uses both short and long window length models (Smodel and Lmodel, respectively) to achieve this. The Lmodel was used to predict IDRs and the Smodel identified short segments with structural propensity within the IDRs predicted by the Lmodel.

### Sequence data

NeProc uses amino acid sequences of IDRs and SDs since it detects regions with structural propensity in a predicted IDR. We used DM4229, which is a training dataset for the IDR prediction program, SPINE-D [26]. DM4229 contains 4,229 sequences selected from PDB and DisProt. In the procedure to create DM4229, PDB structures, with resolution < 2 Å and length > 60 residues, were clustered at 25% sequence identity to select representative proteins with long contiguous IDRs from each cluster. These representative sequences were combined with fully disordered proteins from the IDP database, DisProt [27], and sequence redundancy was reduced again. The sequences identified in the test dataset described below were excluded. The procedure produced 4,189 sequences consisting of 925,412 ordered and 100,284 disordered residues. Among these, 842 sequences were used to validate the hyperparameters of the neural networks, and the remaining 3,347 sequences were used to optimize the biases and weights.

IDEAL provides annotations for ProSs, which are disordered binding regions. These ProSs, collected via manual annotation, possess evidence of disorder in an isolated state and one or more structures with one or more binding partners in PDB. Since ProSs possessed both experimental evidence of disordered in an isolated state and ordered in a bound state, we used the IDEAL data as the test dataset for the performance of disordered binding region predictions.

### The NeProc model construction

NeProc is designed to accept an amino acid sequence, and the query sequence is input into the position-specific

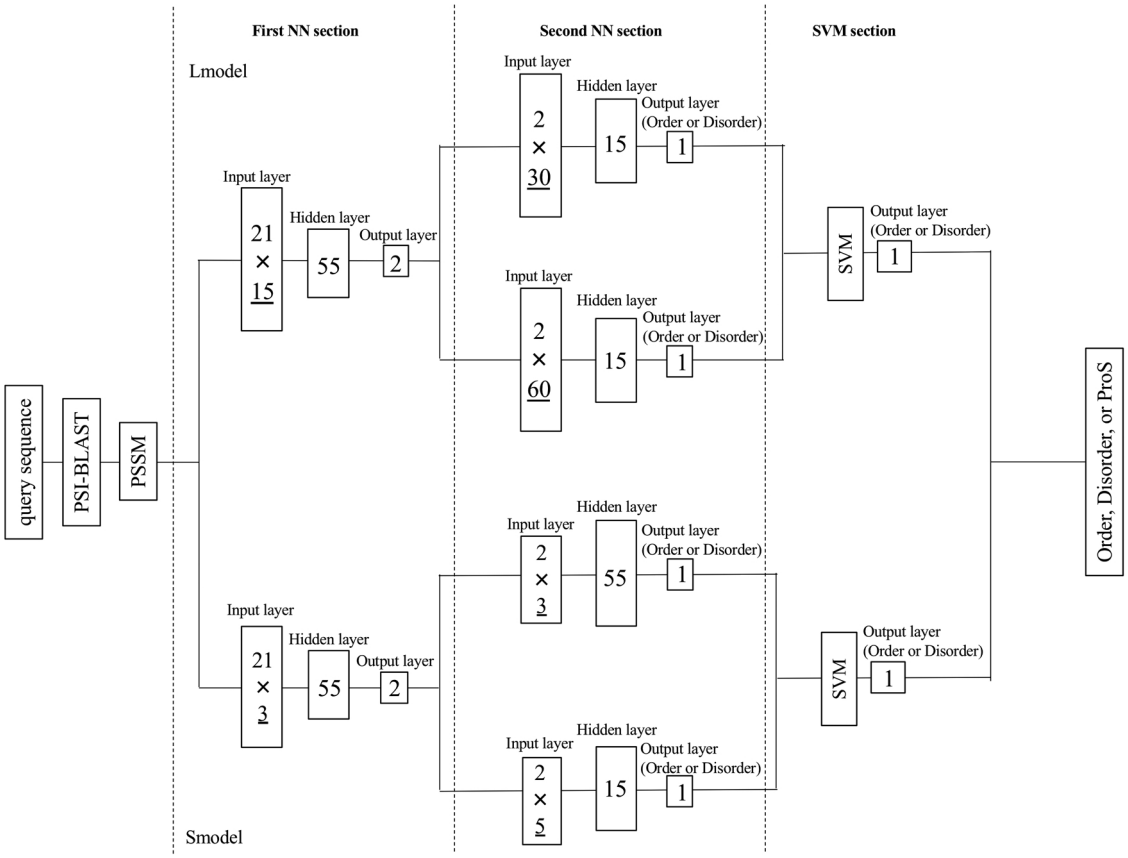


**Figure 1** Structure of the NeProc model. The digits in the layers represent the numbers of nodes. The underlined digits indicate the window sizes of the layers. ProS represents the disordered binding region.

iteration blast (PSI-BLAST) [28] to obtain a position-specific scoring matrix (PSSM). The PSI-BLAST searches were conducted against the UniRef90 database with three iterations and an E-value threshold of 0.001. From a PSSM, we extracted the 21-dimensional vector for a site, including scores for each of the amino acid residues and one for the information per position, as with DISOPRED3. For the training and the predictions, all the information from each PSSM for an entire sequence was used.

We referred to the DISOPRED3 model (Supplementary Fig. S1) as the starting point to construct the NeProc model. The DISOPRED3 model has two neural networks tandemly connected, and each network contains a single hidden layer. The NeProc model employed this structure frame to have two network models (Fig. 1). As previously mentioned, NeProc contains two models, Lmodel and Smodel. The first networks of the NeProc model took over the DISOPRED3 model with the modification of window size. The first network of the Lmodel used that of DISOPRED3, and the Smodel network employed a shorter window size of three residues so that the difference in the Lmodel and Smodel windows was enlarged. The number of hidden layers and nodes was tested using the combinations listed in Supplementary Table S1. These combinations were constructed based on the DISOPRED3 model.

We considered different window sizes in the construction of the second network. Window sizes were selected based on DISOPRED3 of 15 residues. The window sizes for the Lmodel were selected as 15, 30, 40, 50, and 60, where 30 and 60 are multiples of 15, and 40 and 50 filled the gap between 30 and 60. The window sizes for the Smodel were shorter than those in the Lmodel, and were 3, 5, and 10 residues. In the second network, we tested all the combinations of the networks of different window size that were placed parallel, and tested the neural network and the support vector machine as the following unit of the combined output of the second network. The number of hidden layers and nodes were also tested using the combinations listed in Supplementary Table S1.

The hyperparameters were determined using the subset of the training dataset of 842 sequences. The parameters, weights and biases of the first and the second networks, were optimized using the subset of the remaining 3,347 sequences. The biases were initialized using the values

reported by He, *et al.* [29], and the adaptive moment estimation (Adam) [30] was employed as the optimizer. 0.001, 0.9, and 0.999 were used in Adam for the learning rate, exponential decay rate for the first moment estimation, and exponential decay rate for the second estimation, respectively. The rectified linear activation (ReLU) function was used for the activation function. We used the linearSVC of Sikit-learn by changing the cost parameters, from 0.1 to 10, and used the default values for the other parameters.

The Smodel and Lmodel output a binary decision of "ordered" or "disordered". The final prediction was made by combining these outputs from the Smodel and Lmodel, in which a simple decision rule was employed. The input was one of the following four states: disordered/disordered (D/D), disordered/ordered (D/O), ordered/disordered (O/D) or ordered/ordered (O/O) (Smodel output/Lmodel output). The D/D state was disordered, the O/O state was ordered, the O/D state was the disorder binding region, and the D/O state was unknown. Taken together, NeProc accepts a query amino acid sequence, and outputs the three states labels of binding regions, disordered, and ordered as a result.

### Performance evaluation

Evaluation of performance was not straightforward for the binding region prediction. Since it was expected that cryptic binding regions in IDRs would remain hidden, the possibility that residues labeled as disordered may be residues in a disordered binding region could not be excluded. We used the method employed in the ANCHOR2 evaluation [8]. In the disordered binding region prediction, predictions of "disordered" and "disordered binding regions" could not be separated, because of the possibility of cryptic binding regions. On the contrary, "ordered" and "disordered binding regions" did not have to be mixed. Thus, only "disordered binding regions" and "ordered" were considered in the evaluation of the binding region prediction.

Four measures, namely, sensitivity, precision, F-score, and Matthews Correlation Coefficient (MCC), were used, as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$F - score = 2\frac{Sensitivity \times Precision}{Sensitivity + Precision} \tag{3}$$

and

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

We also assessed the statistical significance of the disorder binding region predictions to the other predictors. First, we performed 50,00 resampling experiments, each randomly sampling 80% of the proteins from the test dataset and calculated four measures of each of the considered predictors. We recorded the values of the paired differences (i.e., the NeProc result compared with the results of another predictor). Next, if the differences were normally distributed, as calculated using the Shapiro-Wilk test [31] with 0.05 significance, paired t-test was used, otherwise Wilcoxon signed-rank test was used [32].

ANCHOR2 [8], DISOPRED3 [7] and MoRFchibi-Web [25] were selected as the reference programs. The performance of these programs was evaluated using the data of disordered binding regions obtained from the IDEAL database. We previously inferred disordered binding regions by combining the IDR predictions and UniProt annotation [33]. We found 1,518 binding regions (17,148 residues) of UniProt annotations in the predicted IDRs, called these regions putative binding regions, and used them as another test dataset for the predictions.

## Results and Discussion

### Performance evaluation of prediction of binding regions in IDRs

The NeProc structure is shown in Figure 1. We compared the NeProc results with those of DISOPRED3, ANCHOR2 and MoRFchibi-Web (Table 1). Among the four programs, NeProc showed the highest performance in terms of MCC, precision, and F-score, although ANCHOR2 showed the highest sensitivity. The sensitivity of NeProc was lower than that of ANCHOR2 by 0.010, with greater precision than that of ANCHOR2 by 0.013, suggesting that NeProc provided slightly more false negatives and fewer false positives. It is intriguing that ANCHOR2 uses the DIBS

**Table 1** Performance of disordered binding region predictions of NePorc, ANCHOR2, MoRFchibi-Web, and DISOPRED3

| | MCC | Sensitivity | Precision | F-score |
|---|---|---|---|---|
| NeProc | 0.388 | 0.487 | 0.358 | 0.413 |
| ANCHOR2 | 0.381* | 0.497 | 0.345* | 0.408* |
| MoRFchibi-Web | 0.196* | 0.221* | 0.249* | 0.234* |
| DISOPRED3 | 0.175* | 0.171* | 0.233* | 0.198* |

MCC, Matthews correlation coefficient. * p-value $<1.0\times10^{-15}$ in the comparison with NeProc.

database [20] as a training dataset, as the DIBS database integrates disordered binding regions from IDEAL. This performance test was conducted using the IDEAL dataset, and the ANCHOR2 training data may contain some of the proteins used in this performance test. Nevertheless, the performance of NeProc was comparable to that of ANCHOR2 without using data for disordered binding regions.

IDEAL contained 7,253 residues of binding region in IDRs comprising 321 binding regions. These regions included various length of sequences and secondary structures involved in binding to their partners. First, we analyzed the length dependence of the NeProc binding region prediction. Figure 2 shows the median and mean values of the sensitivity, together with the density of sensitivity by length. Although the sensitivity was 0.487 (Table 1), the mean 10–50 residue range was close to 0.6. On the other hand, very short and very long binding regions showed low sensitivities. The bins from 10 to 50 residues showed higher medians than the mean values, suggesting that many samples had high sensitivity and few samples had low sensitivity. The density plot (grey area) also showed this trend. This result reflected the NeProc method, which detected short segments with structural propensity in long IDRs. The high median values in the short regions indicated that NeProc detected most of the short binding regions in IDRs. As more than 90% of the binding regions in the IDEAL database were shorter than 50 residues (Supplementary Fig. S2), this feature could be practically useful in the prediction of binding regions in IDRs.

Table 2 shows the secondary structure dependence of the prediction of binding regions in IDRs. Table 2A shows the sensitivity of the secondary structures, in which the coil regions have the highest sensitivity followed by the helix and sheet regions. Table 2B shows the sensitivity by classifying the binding regions based on the secondary structure content. The sensitivity values are the mean of the regions in the class. The values in Table 2B are mostly greater than those in Table 2A. Since the statistics in Table 2A were evaluated according to residues whereas Table 2B were analyzed according to regions, the discrepancy between Table 2A and 2B suggests that there were many short binding regions with high sensitivity and a low

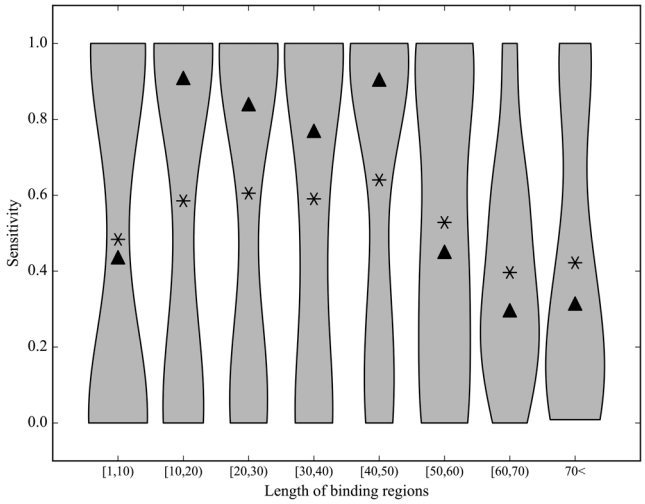


**Figure 2** Relationship between prediction accuracy and binding region lengths. The horizontal axis represents the lengths of the binding regions, and the vertical axis represents sensitivity. Gray areas represent the distribution of sensitivity at each length of the binding regions. Triangles and asterisks represent the median and mean values of sensitivity, respectively.

**Table 2** Secondary structure dependence of the prediction of the disordered binding regions

A) Sensitivity according to secondary structure

| | Helix | Sheet | Coil |
|---|---|---|---|
| Sensitivity | 0.434 | 0.341 | 0.530 |

B) Sensitivity according to disordered binding region secondary structure class

| | H | S | C | H&S |
|---|---|---|---|---|
| Sensitivity | 0.587 | 0.643 | 0.481 | 0.421 |

C) Fractions of judgments according to disordered binding region secondary structure class

| | Structured | Disorderd | Binding regions | Unknown | Average length |
|---|---|---|---|---|---|
| H | 0.340 | 0.174 | 0.483 | 0.003 | 20.8 |
| S | 0.292 | 0.184 | 0.524 | 0.000 | 13.2 |
| C | 0.402 | 0.221 | 0.373 | 0.004 | 18.5 |
| H&S | 0.485 | 0.161 | 0.352 | 0.001 | 38.2 |

H: α helix/helices; S: β sheet/sheets; C: coil/coils; H&S: both α-helices and β-sheets.

number of long binding regions with low sensitivity. Table 2C shows the misjudge trend by the secondary structure class together with the mean length of each of these classes. The H&S class, which contains both α-helices and β-sheets, has the lowest sensitivity in Table 2B, and is largely misjudged as "structured". The average length of this class is the longest. Although the reason for the low sensitivity of the H&S class was not clear, the length of the binding regions may be a factor.

NeProc predicted disordered binding regions without using the binding region data by only learning the SD and IDR data. This result suggested that NeProc identified shared features between binding regions in IDRs and SDs. We then assessed how disordered binding region sequences were similar to SD sequences and/or how binding region sequences were different from IDR sequences. The distances from the binding region data to the SD and IDR data are plotted in Figure 3. In this plot, a circle placed around the diagonal indicates the intermediate amino acid composition between IDRs and SDs. In the shorter windows, the gray circle, which represents all binding regions in the IDRs, is located near the diagonal, while it shifts to the bottom in the longer windows. This trend was observed for all prediction classes, successfully predicted sites (red), unsuccessfully predicted as ordered (blue), and unsuccessfully predicted as disordered (light blue). Although the successfully predicted sites were located near the diagonal in the shorter windows, they shifted to near the horizontal line in the longer windows, where the successfully predicted sites overlaps with the sites unsuccessfully predicted as disordered in the two longest windows. The sites unsuccessfully predicted as ordered did not move far from the shorter windows to the longer windows, and were consistently located as the nearest points to the vertical axis. Thus, the sites unsuccessfully predicted as ordered had similar features to the ordered residues and were not predicted as disordered, even in the longer windows. Similarly, the sites unsuccessfully

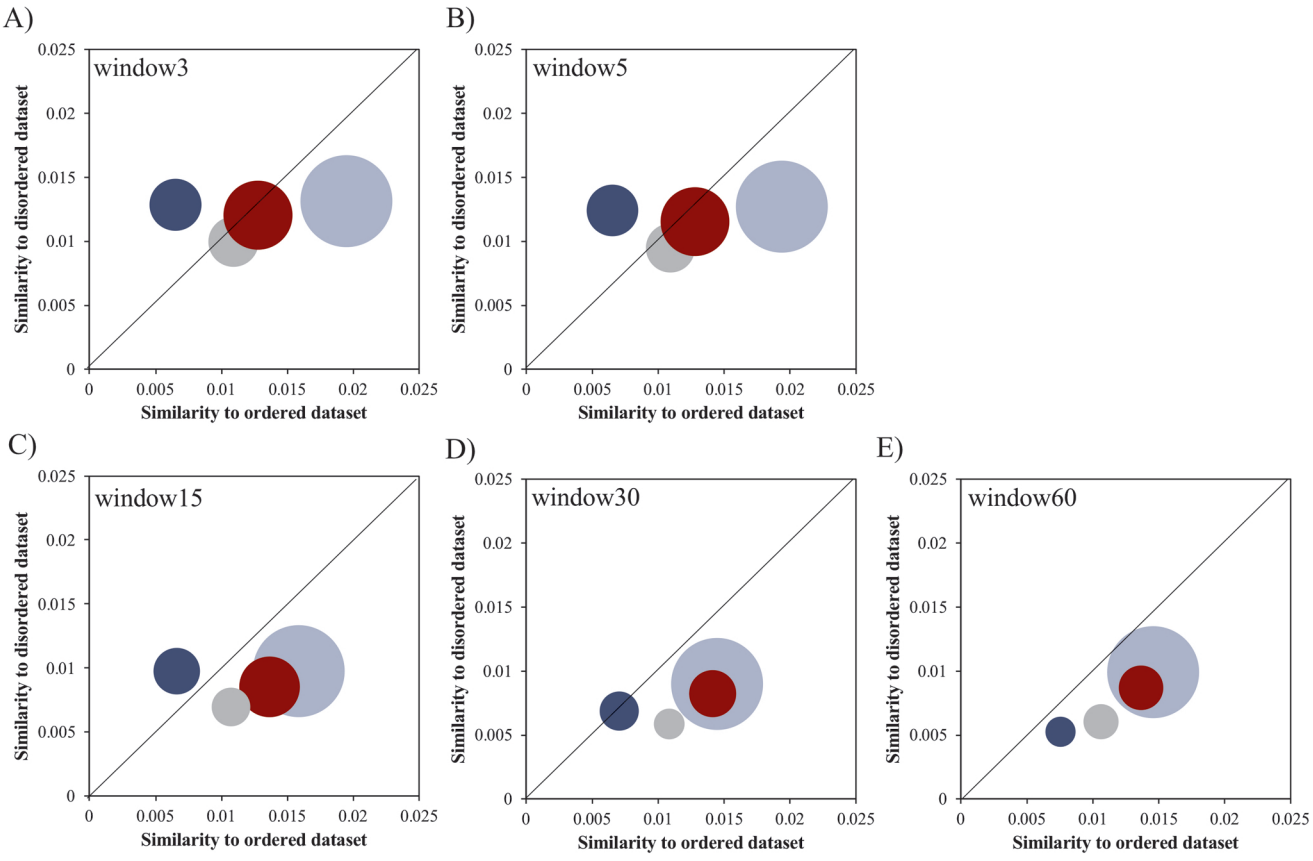


**Figure 3** Similarity in the amino acid composition of the binding regions in IDRs to that of ordered and disordered regions. The amino acid composition of binding regions in IDRs were compared with that of the ordered and disordered data in the training dataset. The horizontal axis showed similarity to the ordered dataset, while the vertical axis showed similarity to the disordered dataset. The binding regions were divided into the sites that were successfully predicted (red), unsuccessfully predicted as ordered (blue), and unsuccessfully predicted as disordered (light blue). The gray circle in each panel represents all the binding regions. The circular centers represent the distance of the average amino acid composition to the ordered and disordered data. The circle areas are proportional to the variances. The compositions were calculated using window sizes of 3, 5, 15, 30, and 60 and are presented in panels A), B), C), D), and E). Details of the calculations are found in Supplementary Text S1.

**Table 3** Performance of disordered binding region predictions using the putative binding region dataset

| | MCC | Sensitivity | Precision | F-score |
|---|---|---|---|---|
| NeProc | 0.588 | 0.896 | 0.421 | 0.572 |
| ANCHOR2 | 0.569 | 0.755 | 0.481 | 0.587 |
| DISOPRED3 | 0.418 | 0.626 | 0.307 | 0.411 |
| MoRFchibi-Web | 0.365 | 0.282 | 0.565 | 0.376 |

predicted as disordered showed features similar to those of the disordered sites, even in the shorter windows. Therefore, we could improve the performance of the prediction of disordered binding regions by NeProc if we could predict the residues in the sites unsuccessfully predicted as ordered as disordered in the long window models and the sites unsuccessfully predicted as disordered as ordered in the short window models.

NeProc and ANCHOR2 showed similar performance in terms of predicting the disordered binding regions with MCCs of 0.388 and 0.381, respectively. The IDEAL database, which provided the binding region samples for this study, focuses mainly on nuclear proteins in the annotation and likely resulting in the biased sampling of the binding regions. We repeated the testing of the prediction of the binding regions using the putative binding region dataset (see Materials and Methods section). As shown in Table 3, NeProc and ANCHOR2 showed comparable performances of 0.588 and 0.569. All programs showed higher performances than the binding region predictions in Table 1. The putative binding regions were defined as the regions with UniProt annotations associated with protein binding found in predicted IDRs. Thus, IDRs containing putative binding regions were considered to have features easily detected by IDR prediction programs. The improved performances in this test may be due to the correct IDR predictions, and improved IDR predictions around disordered binding regions may improve the prediction. As a reference, the performance of IDR predictions was shown in Supplementary Table S2.

Although MoRFchibi-Web showed low performance in the binding site predictions for our test datasets, the framework to develop MoRFchibi-Web differed from NeProc and ANCHOR2. MoRFchibi-Web separates amino acid sequences into binding regions in IDRs and "others", which contain SDs and IDRs. These IDRs could contain binding regions that have not yet been discovered (cryptic binding regions). As described in the Materials and Methods section, NeProc and ANCHOR2 use the disordered binding regions and SDs in the performance tests, since the possibility of cryptic binding regions in IDRs cannot be ruled out. The strategy of MoRFchibi-Web may yield a small number of predicted binding regions as it intends to exclude cryptic binding regions. In contrast, NeProc and ANCHOR2 may yield a large number of predicted binding regions, since the cryptic binding regions they provide cannot be evaluated. The balance between sensitivity and precision in Table 3 may reflect this difference in the development frameworks.

## Conclusions

NeProc was developed without using data for the disordered binding regions, and showed good performance in the predictions of disordered binding regions. These findings highlight the possibility of overcoming the shortage of binding region data in the development of prediction programs.

## Acknowledgements

We are grateful to the IDEAL development team for their efforts to maintain the database. We also thank Shigetaka Sakamoto, Kazuo Hosoda, and the laboratory members for their support implementing NeProc.

## Conflicts of Interest

The authors have no conflicts of interest to declare.

## Authors contributions

Conceived and designed the experiments: HAN SF. Developed the models and the web interface: HAN HAM. Analyzed the data: HAN SF. Wrote the manuscript: HAN SF.

## Reference

[1] Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999). DOI: 10.1006/jmbi.1999.3110

[2] Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. Intrinsic disorder and protein function. *Biochemistry* **41**, 6573–6582 (2002). DOI: 10.1021/bi012159+

[3] Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579**, 3346–3354 (2005). DOI: 10.1016/j.febslet.2005.03.072

[4] Fukuchi, S., Hosoda, K., Homma, K., Gojobori, T. & Nishikawa, K. Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.* **11**, 29 (2011). DOI: 10.1186/1472-6807-11-29

[5] Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**, 635–645 (2004). DOI: 10.1016/j.jmb.2004.02.002

[6] Hanson, J., Yang, Y., Paliwal, K. & Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **33**, 685–692 (2017). DOI: 10.1093/bioinformatics/btw678

[7] Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 (2015). DOI: 10.1093/bioinformatics/btu744
[8] Meszaros, B., Erdos, G. & Dosztanyi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018). DOI: 10.1093/nar/gky384
[9] Necci, M., Piovesan, D., Dosztanyi, Z. & Tosatto, S. C. E. MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* **33**, 1402–1404 (2017). DOI: 10.1093/bioinformatics/btx015
[10] Dyson, H. J. & Wright, P. E. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60 (2002). DOI: 10.1016/s0959-440x(02)00289-0
[11] Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005). DOI: 10.1038/nrm1589
[12] Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584 (2002). DOI: 10.1016/s0022-2836(02)00969-5
[13] Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C. C., Eckmann, C. R., Myong, S., *et al.* The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc. Natl. Acad. Sci. USA* **112**, 7189–7194 (2015). DOI: 10.1073/pnas.1504822112
[14] Nott, T. J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., *et al.* Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015). DOI: 10.1016/j.molcel.2015.01.013
[15] Kato, M., Han, T. W., Xie, S., Shi, K., Du, X., Wu, L. C., *et al.* Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**, 753–767 (2012). DOI: 10.1016/j.cell.2012.04.017
[16] Kwon, I., Kato, M., Xiang, S., Wu, L., Theodoropoulos, P., Mirzaei, H., *et al.* Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell* **155**, 1049–1060 (2013). DOI: 10.1016/j.cell.2013.10.033
[17] Meszaros, B., Simon, I. & Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **5**, e1000376 (2009). DOI: 10.1371/journal.pcbi.1000376
[18] Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., *et al.* Analysis of molecular recognition features (MoRFs). *J. Mol. Biol.* **362**, 1043–1059 (2006). DOI: 10.1016/j.jmb.2006.07.087
[19] Ren, S., Uversky, V. N., Chen, Z., Dunker, A. K. & Obradovic, Z. Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. *BMC Genomics* **9**, S26 (2008). DOI: 10.1186/1471-2164-9-s2-s26
[20] Schad, E., Ficho, E., Pancsa, R., Simon, I., Dosztanyi, Z. & Meszaros, B. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* **34**, 535–537 (2018). DOI: 10.1093/bioinformatics/btx640
[21] Fukuchi, S., Sakamoto, S., Nobe, Y., Murakami, S. D., Amemiya, T., Hosoda, K., *et al.* IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Res.* **40**, D507–D511 (2012). DOI: 10.1093/nar/gkr884
[22] Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., *et al.* IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.* **42**, D320–D325 (2014). DOI: 10.1093/nar/gkt1010
[23] Ota, H. & Fukuchi, S. Sequence conservation of protein binding segments in intrinsically disordered regions. *Biochem. Biophys. Res. Commun.* **494**, 602–607 (2017). DOI: 10.1016/j.bbrc.2017.10.099
[24] Fuxreiter, M., Tompa, P. & Simon, I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* **23**, 950–956 (2007). DOI: 10.1093/bioinformatics/btm035
[25] Malhis, N., Jacobson, M. & Gsponer, J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. *Nucleic Acids Res.* **44**, W488–W493 (2016). DOI: 10.1093/nar/gkw409
[26] Zhang, T., Faraggi, E., Xue, B., Dunker, A. K., Uversky, V. N. & Zhou, Y. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J. Biomol. Struct. Dyn.* **29**, 799–813 (2012). DOI: 10.1080/073911012010525022
[27] Hatos, A., Hajdu-Soltesz, B., Monzon, A. M., Palopoli, N., Alvarez, L., Aykac-Fas, B., *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276 (2020). DOI: 10.1093/nar/gkz975
[28] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997). DOI: 10.1093/nar/25.17.3389
[29] He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034 (2015).
[30] Kingma., D. P. & Ba., J. L. ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION. arXiv: 1412.6980v9 (2015).
[31] Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611 (1965). DOI: 10.2307/2333709
[32] Wilcoxon, F. Individual comparisons of grouped data by ranking methods. *J. Econ. Entomol.* **39**, 269–270 (1946). DOI: 10.1093/jee/39.2.269
[33] Anbo, H., Sato, M., Okoshi, A. & Fukuchi, S. Functional Segments on Intrinsically Disordered Regions in Disease-Related Proteins. *Biomolecules* **9**, 88 (2019). DOI: 10.3390/biom9030088

(*Edited by Motonori Ota*)