*Briefings in Bioinformatics*, 22(2), 2021, 2000–2011

**doi: 10.1093/bib/bbaa018**
Advance Access Publication Date: 28 February 2020
Problem Solving Protocol

# RFPR-IDP: reduce the false positive rates for intrinsically disordered protein and region prediction by incorporating both fully ordered proteins and disordered proteins

Yumeng Liu, Xiaolong Wang and Bin Liu

Corresponding author: Bin Liu, Harbin Institute of Technology, Shenzhen, HIT Campus Shenzhen University Town, Xili, Shenzhen, Guangdong 518055, China, and Beijing Institute of Technology, No. 5, South Zhongguancun Street, Haidian District, Beijing 100081, China. Tel.: (+86) 010-68911310; E-mail: bliu@bliulab.net

## Abstract

As an important type of proteins, intrinsically disordered proteins/regions (IDPs/IDRs) are related to many crucial biological functions. Accurate prediction of IDPs/IDRs is beneficial to the prediction of protein structures and functions. Most of the existing methods ignore the fully ordered proteins without IDRs during training and test processes. As a result, the corresponding predictors prefer to predict the fully ordered proteins as disordered proteins. Unfortunately, these methods were only evaluated on datasets consisting of disordered proteins without or with only a few fully ordered proteins, and therefore, this problem escapes the attention of the researchers. However, most of the newly sequenced proteins are fully ordered proteins in nature. These predictors fail to accurately predict the ordered and disordered proteins in real-world applications. In this regard, we propose a new method called RFPR-IDP trained with both fully ordered proteins and disordered proteins, which is constructed based on the combination of convolution neural network (CNN) and bidirectional long short-term memory (BiLSTM). The experimental results show that although the existing predictors perform well for predicting the disordered proteins, they tend to predict the fully ordered proteins as disordered proteins. In contrast, the RFPR-IDP predictor can correctly predict the fully ordered proteins and outperform the other 10 state-of-the-art methods when evaluated on a test dataset with both fully ordered proteins and disordered proteins. The web server and datasets of RFPR-IDP are freely available at http://bliulab.net/RFPR-IDP/server.

**Key words:** intrinsically disordered proteins and regions; fully ordered proteins; convolution neural network; bidirectional long short-term memory

**Yumeng Liu** is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. Her research areas include bioinformatics and machine learning.
**Xiaolong Wang**, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His research areas include nature language processing, bioinformatics and artificial intelligence.
**Bin Liu**, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, and School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China. His expertise is in bioinformatics, nature language processing and machine learning.

## Introduction

Intrinsically disordered proteins/regions (IDPs/IDRs) refer to those proteins/regions whose native state is intrinsically disordered without a stable three-dimensional structure [1, 2]. Despite the lack of stable three-dimensional structure, IDPs/IDRs have been confirmed to play important roles in many important biological functions, such as the folding of nucleic acids [3], cellular signaling and regulation [4] and molecular recognition and molecular assembly [5–8]. Besides, some diseases are also correlated with IDPs/IDRs, such as cancer [7] and Alzheimer's disease [9, 10]. Therefore, accurate identification of IDPs/IDRs is an important fundamental task for studying protein functions and drug design.

There are several traditional experimental techniques can be used to detect IDPs/IDRs [5, 11], such as NMR and X-ray crystallography. However, with the rapid growth of protein sequences, it is urged to propose fast and efficiently computational methods. The existing computational methods can be divided into four categories [1], including physicochemical-based methods [12, 13], machine-learning-based methods [14–16], template-based methods and meta-methods [17]. Machine-learning-based methods can be further divided into classification models and sequence labeling models.

Some studies have shown that the proportion of IDPs with long disordered regions (LDRs) is about 2 to 45% in different species [18–22], where LDRs represent those IDRs with more than 30 residues in length. Fully ordered proteins without IDRs are widespread in nature, but they are ignored by the existing predictors. The existing predictors are evaluated on test datasets consisting of disordered proteins without or with only a few fully ordered proteins; for example, the methods SPOT-disorder [14] and SPINE-D [15] are trained and tested with 3000 IDPs without fully ordered proteins, DISOPRED3 [17] is evaluated on 94 proteins with only 2 fully ordered proteins and AUCpreD [23] is evaluated on a test dataset with 94 proteins containing only 2 fully ordered proteins, and a test dataset with 117 proteins containing only 4 fully ordered proteins. However, a predictor trained and tested with disordered proteins without or with only a few fully ordered proteins will prefer to predict a newly sequenced protein as a disordered protein. However, as discussed above most of the proteins are fully ordered proteins without disordered regions in nature. This problem will prevent the real-world applications of these predictors.

In this regard, we first investigate the influence of the fully ordered proteins on the performance of various methods and then propose a predictor called RFPR-IDP using both the fully ordered proteins and intrinsically disordered proteins to construct the model. Deep learning technique has been successfully applied to bioinformatics, such as protein contact map prediction [24] and protein fold recognition [25]. Deep learning technique has also been applied to disordered protein and region prediction; for example, DeepCNF [26] and AUCpreD [23] are two predictors based on the combination of conditional neural fields (CNF) and deep convolutional neural networks (DCNN), and AUCpreD [23] shows better performance than DeepCNF [26] by adopting maximal-AUC training algorithm. SPOT-disorder [14] is a predictor for IDPs/IDRs and is constructed based on Bidirectional Long Short-Term Memory (BiLSTM) [27, 28]. Inspired by these methods, the proposed predictor RFPR-IDP is constructed based on the combination of convolution neural network (CNN) [29, 30] and BiLSTM [27, 28]. Different from SPOT-disorder [14], RFPR-IDP adopts CNN to capture local patterns of target residues from protein sequences. Based on the features obtained by CNN, BiLSTM is then performed to obtain the long-term dependence information of the proteins.

## Materials and methods

### Training dataset

The training dataset used in this study includes two parts: IDP set and ordered protein set. The ordered protein set contains 616 proteins collected from Protein Data Bank (PDB) [31] with following criteria: (i) the structure file for each protein contains only one chain, which ensures that no ordered regions are transformed from IDRs through binding with other proteins; (ii) the resolution of each protein is less than or equal to 2Å; (iii) the length of each protein is greater than or equal to 30; (iv) the similarity between sequences is less than 25%; (v) each residue has atomic co-ordinates recorded in PDB; and (vi) non-standard amino acids are removed. The training dataset can be formatted as

$$\mathbb{S}_{all}^{Train} = \mathbb{S}_{disorder}^{Train} \cup \mathbb{S}_{order}^{Train} \tag{1}$$

where $\mathbb{S}_{disorder}^{Train}$ contains 4229 IDPs constructed by Zhang *et al.* [15] with sequence similarity less than 25%; $\mathbb{S}_{all}^{Train}$ is the union of $\mathbb{S}_{disorder}^{Train}$ and $\mathbb{S}_{order}^{Train}$ to train RFPR-IDP.

### Test datasets

#### *DISORDER723 with only disordered proteins*

The DISORDER723 test dataset is reported in [32] with 723 disordered proteins without fully ordered proteins. DISORDER723 is a widely used test dataset.

#### $\mathbb{S}1_{order}^{Test}$ *with only fully ordered proteins*

$\mathbb{S}1_{order}^{Test}$ test dataset contains 329 fully ordered proteins selected by following the criteria for constructing $\mathbb{S}_{order}^{Train}$. The sequence similarities between $\mathbb{S}_{disorder}^{Train}$ and $\mathbb{S}1_{disorder}^{Test}$ are lower than 25% by using Blastclust algorithm [33].

#### $\mathbb{S}1$ *with both IDPs and fully ordered proteins*

In order to simulate the real-world application situation, the test dataset $\mathbb{S}1$ is constructed with both disordered proteins and fully ordered proteins, which can be defined as

$$\mathbb{S}1 = \mathbb{S}1_{disorder}^{Test} \cup \mathbb{S}1_{order}^{Test} \tag{2}$$

where $\mathbb{S}1_{disorder}^{Test}$ contains 329 disordered proteins constructed by Sirota *et al.* [34], which is widely used by many studies [14]; $\mathbb{S}1_{order}^{Test}$ contains 329 fully ordered proteins; and $\mathbb{S}1$ is the union of $\mathbb{S}1_{disorder}^{Test}$ and $\mathbb{S}1_{order}^{Test}$. The sequence similarities between $\mathbb{S}_{disorder}^{Train}$ and $\mathbb{S}1_{disorder}^{Test}$ as well as those between $\mathbb{S}_{order}^{Train}$ and $\mathbb{S}1_{order}^{Test}$ are less than 25% by using the Blastclust algorithm [33].

The statistical information of the three test datasets is shown in Table 1.

### Features

Feature extraction is a key step in building machine-learning-based predictors [35–40]. In this study, evolutionary information and physicochemical properties are employed to represent each residue in proteins. The evolutionary information is more

**Table 1.** The statistical information of the three test datasets

| Test dataset | #D (percent)[a] | #O (percent)[b] | #IDPs (percent) | #Ordered proteins (percent) |
|---|---|---|---|---|
| DISORDER723 | 13 526 (6.3%) | 201 703 (93.7%) | 723 (100%) | 0 (0%) |
| $\mathbb{S}1^{Test}_{order}$ | 0 (0%) | 79 091 (100%) | 0 (0%) | 329 (100%) |
| $\mathbb{S}1$ | 39 544 (23.3%) | 130 383(76.7%) | 329 (50%) | 329 (50%) |

[a]The number of disordered residues, and the percent represents the ratio of disordered residues to all residues.
[b]The number of ordered residues, and the percent represents the ratio of ordered residues to all residues.

discriminative than the sequence information [41–44], which is acquired from position-specific scoring matrices (PSSMs) obtained from PSI-BLAST [33] by searching against nrdb90 database [45]. For PSI-BLAST [33], the numbers of iterations and parameter E-value are set as 3 and 0.001, respectively. Besides, seven widely used physicochemical properties are adopted [46]. Therefore, the dimension of feature vector for each residue is 27.

## Neural network architecture of RFPR-IDP and implementations

The neural network was employed in bioinformatics for many predicting problems [47–50]. The neural network architecture of RFPR-IDP is shown in Figure 1. As shown in Figure 1, RFPR-IDP contains five layers: input layer, CNN layer, FC (fully connected) layer, BiLSTM layer and output layer. In the input layer, proteins are transformed into feature matrices, which are constructed by PSSMs and seven physicochemical properties described in Section "Features". The feature matrices are padded with zero vectors at both ends and then used as the input of CNN so as to ensure that the output of CNN has the same length as the input. Then, the CNN layer with rectified linear units (ReLU) [51] activation function scans on the feature matrices by using multiple one-dimensional convolution filters to capture protein local information or motifs. The FC layer is used to weight the output of the CNN layer to capture the effective features of the CNN output, and the dimension of the feature vector for each residue can be reduced. For a target residue in a protein sequence, the information on its left and right sides is asymmetric. Therefore, BiLSTM is adopted to capture the long-term dependence information in both directions of proteins. The final output layer is composed of a fully connected layer and a Softmax layer, which is used as a classifier to generate the prediction of residues.

The network of RFPR-IDP is implemented by using TensorFlow 1.4.1 [52], and RFPR-IDP is trained with ADAM optimization algorithm [53]. To avoid overfitting during training, dropout algorithm [54] is adopted with a 70% dropout rate at the outputs of CNN and BiLSTM layers. Due to the extremely imbalanced ratio of positive and negative samples in the training set, weighted cross-entropy loss function [55] is used during training.

## Criteria for performance evaluation

Several measures are adopted in this study [2, 56, 57], including sensitivity (Sn), specificity (Sp), balanced accuracy (BACC) and Matthew's correlation coefficient (MCC). Besides, the false positive rates (FPRs) of different predictors are estimated at both protein level and residue level [20]. In particular, FPR_R and FPR_P$_l$ are used to distinguish the FPR at residue level from the FPR at protein level defined as [58]

$$\begin{cases} \mathrm{Sn} = \frac{TP}{TP+FN} \\ \mathrm{Sp} = \frac{TN}{TN+FP} \\ \mathrm{BACC} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \\ \mathrm{MCC} = \frac{\left(TP\times TN\right)-\left(FP\times FN\right)}{\sqrt{\left(TP+FP\right)\left(TP+FN\right)\left(TN+FP\right)\left(TN+FN\right)}} \\ \mathrm{FPR\_R} = \frac{FP}{TN+FP} \\ \mathrm{FPR\_P}_l = \frac{FP_l}{TN+FP_l}, \left(l = 1, 2, 3 \ldots, L\right) \end{cases} \quad (3)$$

where $TP$, $FP$, $TN$ and $FN$ are the numbers of true positive, false positive, true negative and false negative samples, respectively. $L$ represents the length of the protein. For ordered proteins, $FP_l$ represents the number of false positive proteins with at least $l$ consecutive disordered residues. Some studies have suggested that IDRs less than 4 in length may be caused by experimental uncertainties [32, 59, 60]. Therefore, FPR_P$_4$ is regarded as the FPR standard at the protein level in this study. However, FPR_P$_1$ is still retained as the most stringent evaluation criterion at the protein level.

In addition, several widely used measures are also used in this study, including AUC (area under the receiver operating characteristics curve) [14, 23, 61], AULC (area under the receiver operating characteristics curve with lower FPRs) and AUC$^{PR}$ (area under the precision recall curve) [23].

## Method comparison

In order to fairly compare the performance of various methods on test datasets, the standalone packages of several predictors have been downloaded, including AUCpreD [23] (http://raptorx2.uchicago.edu/StructurePropertyPred/predict/), DISOPRED 3.16 [17] (URL: http://bioinf.cs.ucl.ac.uk/web_servers/), SPOT-disorder [14] (URL: http://sparks-lab.org/server/SPOT-disorder/index.php), IUPred 1.0 [12] (URL: http://iupred.enzim.hu/), SPINE-D 2.0 [15] (URL: http://sparks-lab.org/), DisEMBL 1.4 (URL: http://dis.embl.de) [16] and GlobPlot 2.3 [13] (URL: http://globplot.embl.de/). These packages are all run with default parameters.

# Results and discussion

## Training and parameter optimization

The parameters of RFPR-IDP are optimized on the training dataset $\mathbb{S}^{Train}_{all}$ by using 2-fold cross-validation according to AUC. The optimization range and the optimal value of each parameter are shown in Table 2. The optimized parameter values on the training dataset are used to predict the samples in the three test datasets.

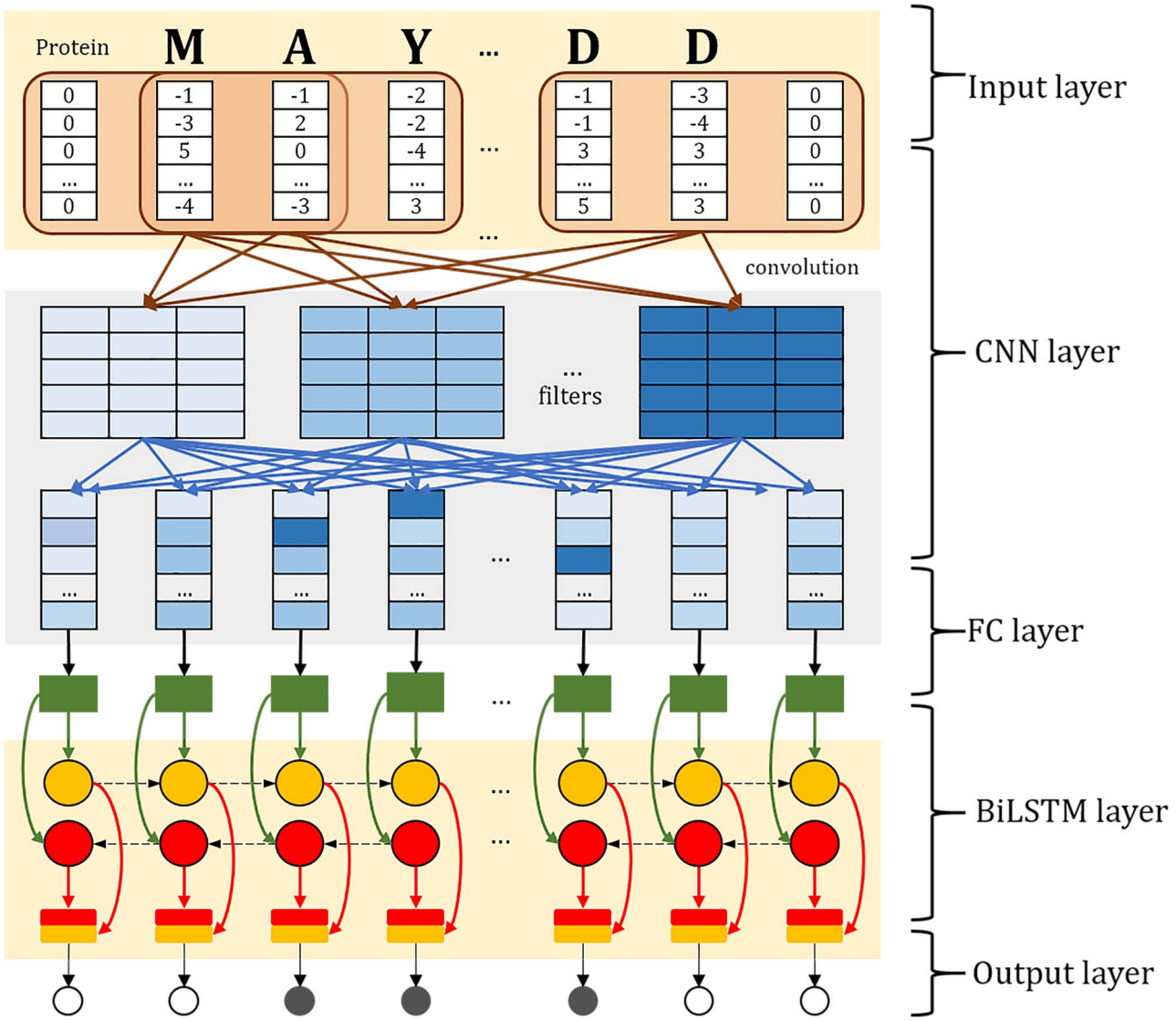


**Figure 1. The framework of RFPR-IDP.** RFPR-IDP contains five layers, including (i) input layer, which transforms proteins into feature vectors; (ii) CNN layer, which captures the local information or motifs of proteins by using a series of one-dimensional convolution filters; (iii) FC layer, which captures the effective features of CNN output; (iv) BiLSTM layer, which captures the long-term dependence information in both directions of proteins; and (v) output layer, which generates the final prediction category for each residue.

**Table 2.** The optimization range and the optimal value of each parameter for RFPR-IDP

| Parameters | Range | Increment | Optimal value |
|---|---|---|---|
| Number of CNN layer | [1, 2] | 1 | 1 |
| Number of BiLSTM layer | [1, 2] | 1 | 1 |
| Length of filters[a] | [5, 13] | 2 | 7 |
| Number of filters[b] | [50, 400] | 50 | 200 |
| Number of units[c] | [100, 400] | 50 | 300 |
| Weight[d] | [1, 5] | 1 | 4 |

[a]The length of one-dimensional convolution filters.
[b]The number of one-dimensional convolution filters.
[c]The number of units in the LSTM cell in each direction.
[d]The weight of positive samples.

**Table 3.** Performance of various methods on the test dataset DISORDER723

| Predictor[a] | Residue level | | | | | | | Protein level | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sn | Sp | MCC | BACC | $AUC^{PR}$ | AULC[b] | AUC | Sn | Sp | MCC | BACC | *P* value[c] |
| RFPR-IDP | 0.522 | 0.974 | 0.519 | 0.748 | 0.556 | 0.032 | 0.898 | 0.526 | 0.973 | 0.517 | 0.750 | NA[d] |
| AUCpreD [23] | 0.580 | 0.974 | **0.564** | 0.777 | **0.621** | **0.036** | **0.914** | 0.590 | 0.974 | **0.563** | 0.782 | 1.2E-02 |
| DISOPRED3 [17] | 0.452 | **0.986** | 0.536 | 0.719 | 0.597 | 0.032 | 0.899 | 0.455 | **0.986** | 0.533 | 0.720 | 1.5E-02 |
| SPOT-disorder [14] | 0.470 | 0.983 | 0.531 | 0.726 | 0.574 | 0.033 | 0.898 | 0.483 | 0.983 | 0.533 | 0.733 | 1.8E-01 |
| SPINE-D [15] | 0.779 | 0.840 | 0.376 | **0.810** | 0.560 | 0.032 | 0.891 | **0.791** | 0.841 | 0.384 | **0.816** | 4.4E-08 |
| IUPred-short [12] | 0.495 | 0.943 | 0.382 | 0.719 | 0.423 | 0.023 | 0.810 | 0.508 | 0.943 | 0.388 | 0.726 | 3.8E-02 |
| IUPred-long [12] | 0.298 | 0.949 | 0.240 | 0.623 | 0.247 | 0.014 | 0.721 | 0.289 | 0.950 | 0.236 | 0.619 | 9.9E-21 |
| DisEMBL-C [16] | **0.699** | 0.449 | 0.073 | 0.574 | NA | NA | NA | 0.700 | 0.452 | 0.074 | 0.576 | 5.6E-32 |
| DisEMBL-R [16] | 0.296 | 0.983 | 0.374 | 0.640 | NA | NA | NA | 0.293 | 0.983 | 0.362 | 0.638 | 8.7E-17 |
| DisEMBL-H [16] | 0.560 | 0.805 | 0.214 | 0.682 | NA | NA | NA | 0.566 | 0.804 | 0.216 | 0.685 | 1.9E-07 |
| Globplot [13] | 0.304 | 0.883 | 0.136 | 0.594 | NA | NA | NA | 0.307 | 0.885 | 0.140 | 0.596 | 1.4E-27 |

Best value for each measure is shown in bold.
[a]The parameters of the proposed method RFPR-IDP are described in Section "Training and parameter optimization". The threshold of SPOT-disorder is set as 0.5, and the parameters of other related methods are set as default values.
[b]The AULC is computed based on lower FPRs. The threshold of FPR is set as 6.3%, which is equal to the ratio of positive samples in the test dataset DISORDER723.
[c]The *P* value is calculated between RFPR-IDP and the other methods in terms of BACC per 10 proteins.
[d]NA represents not available.

**Table 4.** False positive rates of various methods on the test dataset $\mathbb{S}1^{Test}_{order}$

| Predictors[a] | FPR_R | $FPR\_P_1$ | $FPR\_P_4$ | Rank | | |
|---|---|---|---|---|---|---|
| | | | | FPR_R | $FPR\_P_1$ | $FPR\_P_4$ |
| **RFPR-IDP** | **0.53%** | **24.3%** | **10.9%** | **1** | **1** | **1** |
| DISOPRED3 [17] | 1.56% | 78.4% | 25.8% | 2 | 2 | 2 |
| SPOT-disorder [14] | 4.16% | 90.9% | 65.1% | 3 | 4 | 3 |
| AUCpreD [23] | 4.91% | 99.7% | 74.2% | 4 | 5 | 5 |
| SPINE-D [15] | 6.95% | 100.0% | 82.7% | 5 | 6 | 6 |
| IUPred-long [12] | 11.23% | 85.4% | 66.0% | 6 | 3 | 4 |
| IUPred-short [12] | 24.95% | 100.0% | 99.4% | 7 | 6 | 7 |

The FPRs are calculated at the same sensitivity of 0.7 acquired by evaluating test dataset $\mathbb{S}1$.
[a]The parameters of RFPR-IDP are listed in Section "Training and parameter optimization", and the parameters of other related methods are set as the default values.

## Performance of various methods for predicting disordered proteins on the test dataset DISORDER723 with only disordered proteins

In this section, the predictive results of various predictors for predicting the disordered proteins on the test dataset DISORDER723 are listed in Table 3. In order to avoid overestimating the proposed method, proteins sharing >25% sequence similarities with any protein in the DISORDER723 test dataset are removed from the training dataset $\mathbb{S}^{Train}_{disorder}$ (cf. Eq. 1) by using the Blastclust algorithm [33]. RFPR-IDP is re-trained with the non-redundant training dataset to predict the samples in the DISORDER723 test dataset so as to give the final results. DISORDER723 is a widely used test dataset for evaluating the performance of predictors. However, this test dataset only contains 723 disordered proteins without fully ordered proteins. In other words, this test dataset can only evaluate the performance for predicting disordered proteins. We can see that RFPR-IDP outperforms all the other methods except for AUCpreD [23] and SPINE-D [15].

## Performance of various methods for predicting fully ordered proteins on the test dataset $\mathbb{S}1^{Test}_{order}$ with only fully ordered proteins

As discussed above, the DISORDER723 test dataset can only evaluate the performance for predicting the disordered proteins. In this section, we will investigate the performance of various methods for predicting the fully ordered proteins, and the results are listed in Table 4. As expected, RFPR-IDP achieves the lowest FPRs at both residue level and protein level for predicting the fully ordered proteins. All the other methods prefer to predict the fully ordered proteins as disordered proteins because of the ignorance of the fully ordered protein in their training processes.

## Incorporation of fully ordered proteins into the training process can reduce FPRs and improve the predictive performance

Ordered proteins are widespread in nature, but they are ignored by most of existing predictors. RFPR-IDP is trained with both fully ordered proteins and disordered proteins, which can effectively reduce the false positive rates as discussed in Table 4. In order to further analyze the impact of the fully ordered proteins on the performance of RFPR-IDP, RFPR-IDP is trained with different training datasets with different ratios of fully ordered proteins and disordered proteins. These training datasets are constructed by randomly removing ordered proteins from the $\mathbb{S}^{Train}_{all}$. The FPRs with the same sensitivity of these models on test dataset $\mathbb{S}1^{Test}_{order}$ are compared (see Figure 2). From this figure, we can see that the FPR obviously decreases when increasing the number of fully ordered proteins. The

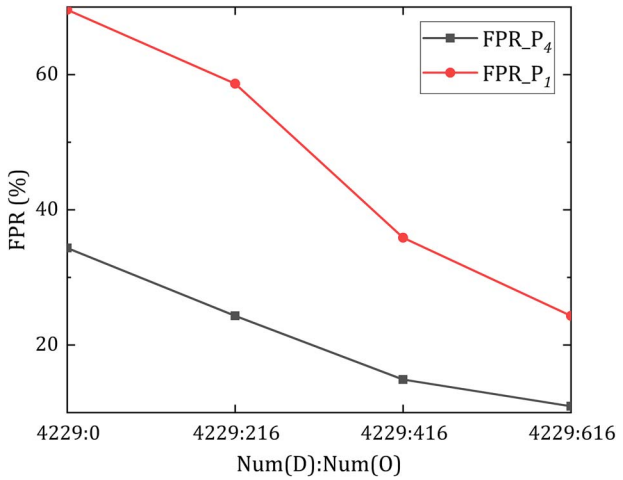


**Figure 2**. The false positive rates on test dataset $\mathbb{S}1^{Test}_{order}$ achieved by different RFPR-IDP models trained with different ratios of IDPs and fully ordered proteins. Num(D) represents the number of IDPs, and Num(O) represents the number of fully ordered proteins. The FPRs are calculated at the same sensitivity of 0.7 acquired by evaluating test dataset $\mathbb{S}1$.

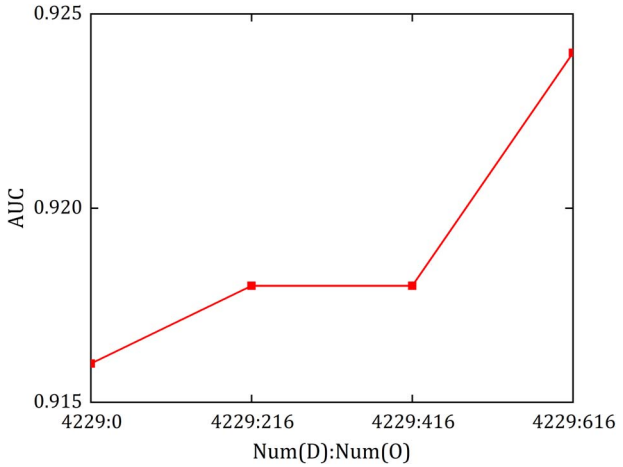


**Figure 3**. The performance comparison on test dataset $\mathbb{S}1$ achieved by different RFPR-IDP models trained with different ratios of IDPs and fully ordered proteins. Num(D) represents the number of IDPs, and Num(O) represents the number of fully ordered proteins.

performance of these models on the test dataset $\mathbb{S}1$ is compared (see Figure 3), from which we can see that the performance of RFPR-IDP can be obviously improved by adding the fully ordered proteins into the training dataset. These observations are fully consistent with our assumption that the fully ordered proteins should be considered during the training process of RFPR-IDP, and this approach will make it more suitable for real-world applications.

## Performance of various methods on test datasets with different ratios of disordered proteins and fully ordered proteins

Some studies have shown that proteins in different species contain about 2% to 45% proteins with LDRs [18–22]. To estimate the performance of RFPR-IDP for predicting disordered proteins from ordered proteins in nature, the test dataset $\mathbb{S}1$ is adopted

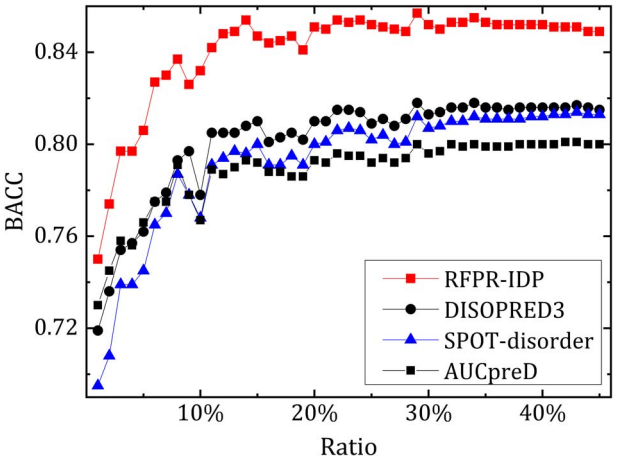


**Figure 4**. Performance comparison of RFPR-IDP, DISOPRED3, SPOT-disorder, and AUCpreD for predicting proteins with different ratios of disordered proteins on test dataset $\mathbb{S}1$.

to simulate the ratio of disordered proteins and ordered proteins in nature. Proteins are randomly selected from $\mathbb{S}1^{Test}_{disorder}$ and $\mathbb{S}1^{Test}_{order}$ respectively in the ratios between 1% and 45%. For a given ratio, proteins are randomly selected for 10 times, and the average predictive performance on these 10 subsets is taken as the final performance. Figure 4 shows the performance of the top four predictors according to the results listed in Table 3 on these random selected datasets. As can be seen from this figure, RFPR-IDP obviously outperforms other three predictors on the test dataset $\mathbb{S}1$ with different ratios of disordered proteins and fully ordered proteins, which further confirms that RFPR-IDP is useful for real-world applications.

Furthermore, the FPRs of different methods under different evaluation metrics at protein level are compared (see Figure 5**A**). From the figure, we can see that (i) the FPRs of RFPR-IDP are comparable with DISOPRED3 [17] and are lower than those of other methods; (ii) From $FPR\_P_{30}$ to $FPR\_P_1$, the stricter the judgement of false positive proteins is, the more obvious the advantages of our method are. Figure 5**B** shows a comparison of FPRs of four methods with the best performance under different evaluation criteria. From this figure, we can see that RFPR-IDP outperforms SPOT-disorder [14] and AUCpreD [23] under different evaluation metrics and is comparable with DISOPRED3 [17]. From $FPR\_P_{30}$ to $FPR\_P_{22}$, RFPR-IDP is highly comparable with DISOPRED3 [17] and obviously outperforms DISOPRED3 [17] from $FPR\_P_{16}$ to $FPR\_P_1$.

## Performance comparison of various methods on the test dataset $\mathbb{S}1$ with both fully ordered proteins and disordered proteins

Various methods are further evaluated on the test dataset $\mathbb{S}1$ (see Table 5) to objectively evaluate their performance for predicting both disordered proteins and ordered proteins. Compared with the results listed in Table 3, although AUCpreD [23] and SPINE-D [15] outperform RFPR-IDP for predicting the disordered proteins (see Table 3), RFPR-IDP achieves the best performance for predicting both fully ordered proteins and disordered proteins (see Table 5). These results are not surprising because RFPR-IDP considers both the fully ordered proteins and disordered proteins in the training and test processes, making it more suitable for predicting newly sequenced proteins, most of which are in fact fully ordered proteins.

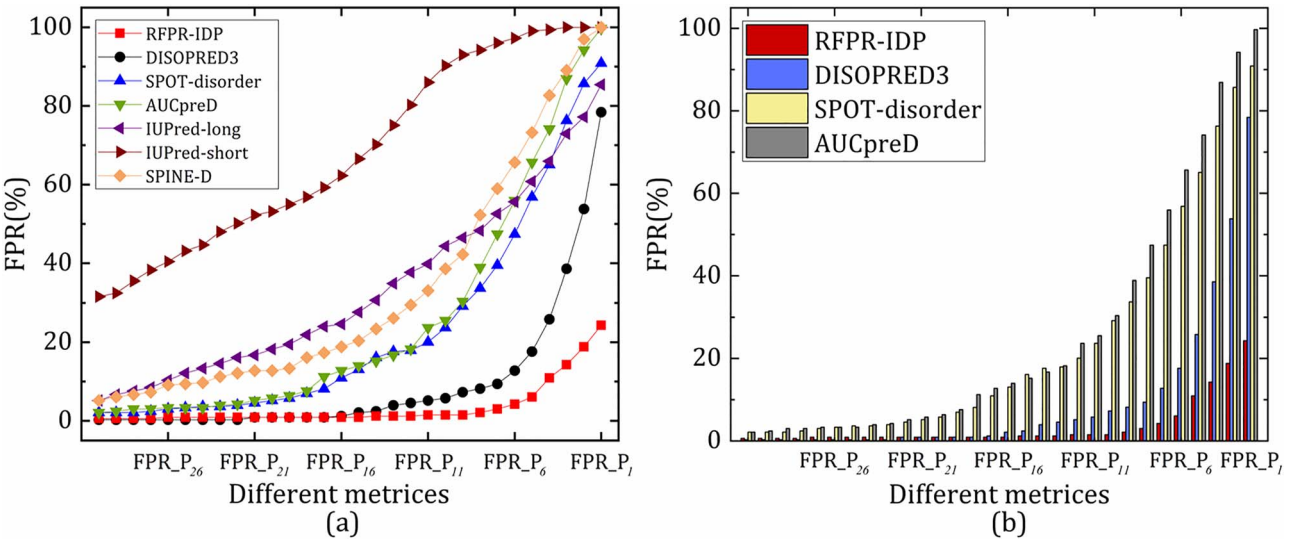


**Figure 5**. The false positive rates of different predictors under different evaluation metrics at protein level on test dataset $\mathbb{S}1^{Test}_{order}$. (**A**) The FPRs' comparison of seven related methods; (**B**) the FPRs' comparison of four best performance methods, including RFPR-IDP, DISOPRED3, SPOT-disorder and AUCpreD. The FPRs are calculated at the same sensitivity of 0.7 acquired by evaluating test dataset $\mathbb{S}1$.

**Table 5.** Performance of various methods on test dataset $\mathbb{S}1$

| Predictor[a] | Residue level | | | | | | | Protein level | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sn | Sp | MCC | BACC | $AUC^{PR}$ | AULC[b] | AUC | Sn | Sp | MCC | BACC | *P* value[c] |
| RFPR-IDP | 0.782 | 0.923 | **0.697** | **0.853** | **0.852** | **0.183** | **0.924** | 0.749 | 0.918 | **0.656** | **0.834** | NA[d] |
| DISOPRED3 [17] | 0.673 | 0.961 | 0.688 | 0.817 | 0.848 | 0.172 | 0.922 | 0.633 | 0.961 | 0.640 | 0.797 | 2.3E-02 |
| SPOT-disorder [14] | 0.653 | 0.969 | 0.689 | 0.811 | 0.838 | 0.176 | 0.908 | 0.606 | 0.966 | 0.628 | 0.786 | 5.3E-03 |
| AUCpreD [23] | 0.633 | 0.966 | 0.668 | 0.799 | 0.817 | 0.161 | 0.897 | 0.612 | 0.964 | 0.623 | 0.788 | 4.9E-03 |
| SPINE-D [15] | **0.819** | 0.804 | 0.555 | 0.811 | 0.806 | 0.168 | 0.890 | **0.795** | 0.801 | 0.510 | 0.798 | 2.1E-02 |
| IUPred-long [12] | 0.598 | 0.939 | 0.584 | 0.769 | 0.749 | 0.148 | 0.854 | 0.565 | 0.937 | 0.530 | 0.751 | 1.0E-06 |
| IUPred-short [12] | 0.502 | 0.932 | 0.489 | 0.717 | 0.621 | 0.125 | 0.822 | 0.497 | 0.930 | 0.451 | 0.713 | 1.7E-12 |
| DisEMBL-C [16] | 0.775 | 0.425 | 0.174 | 0.600 | NA | NA | NA | 0.768 | 0.425 | 0.159 | 0.597 | 1.1E-34 |
| DisEMBL-R [16] | 0.305 | **0.976** | 0.417 | 0.641 | NA | NA | NA | 0.300 | **0.976** | 0.381 | 0.638 | 9.3E-28 |
| DisEMBL-H [16] | 0.441 | 0.795 | 0.228 | 0.618 | NA | NA | NA | 0.434 | 0.792 | 0.197 | 0.613 | 1.6E-30 |
| Globplot [13] | 0.361 | 0.859 | 0.236 | 0.610 | NA | NA | NA | 0.359 | 0.858 | 0.217 | 0.609 | 1.7E-32 |

Best value for each measure is shown in bold.

[a]The parameters of RFPR-IDP are described in Section "Training and parameter optimization". The threshold of SPOT-disorder is set as 0.5, and the parameters of other related methods are set as default values.
[b]The AULC is computed based on lower FPRs. The threshold of FPR is set as 23.3%, which is equal to the ratio of positive samples in the test dataset $\mathbb{S}1$.
[c]The *P* value is calculated between RFPR-IDP and the other methods in terms of BACC per 10 proteins. Because the test dataset $\mathbb{S}1$ contains fully ordered proteins, the BACC of fully ordered proteins cannot be calculated for per protein.
[d]NA represents not available.

## Examples of predicted proteins

In this section, the results of four proteins predicted by RFPR-IDP, DISOPRED3 [17], SPOT-disorder [14] and AUCpreD [23] are visualized (see Figures 6–9), including two fully ordered proteins (3CSZA and 5OSWA) and two disordered proteins (1GVEB and 1J3WA). Their structures are obtained from PDB database [31]. PyMOL (https://pymol.org/2/) software is adopted to generate 3D structures of these proteins.

The schematic diagrams of two fully ordered proteins are shown in Figures 6 and 7, and the schematic diagrams of two IDPs are shown in Figures 8 and 9. As shown in Figures 6 and 7, RFPR-IDP can correctly predict the IDRs or predict fewer false positive residues than those of other methods, such as the regions {1, 8}, {11, 11} and {159, 159} in protein 3CSZA (see Figure 6), the regions {31, 31}, {58, 58}, {60, 60}, {62, 62}, {64, 65}, {68, 69}, {89, 89}, {91, 104}, {266, 276}, {278, 278}, {558, 578} and {582, 583} in protein 5OSWA (see Figure 7). As shown in Figures 8 and 9, RFPR-IDP can more accurately predict these two IDPs, such as the region {140, 163} of protein 1J3WA in Figure 9. Furthermore, RFPR-IDP is able to identify IDRs that cannot be identified by other methods, such as the region {210, 216} of 1GVEB (see Figure 8). These examples further demonstrate that RFPR-IDP can effectively reduce the FPRs and can accurately predict both fully ordered proteins and disordered proteins.

## Conclusion

This study investigates the influence of the fully disordered proteins on training and testing the computational predictors for intrinsically disordered protein and region prediction. Based on

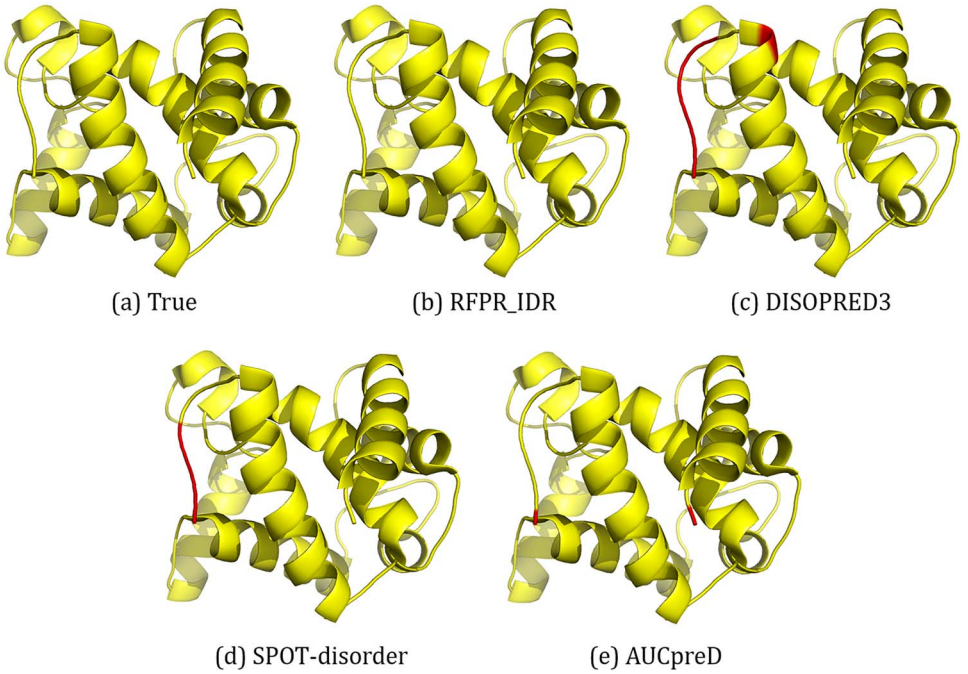


**Figure 6**. A schematic diagram of ordered protein 3CSZA predicted by RFPR_IDR, DISOPRED3, SPOT-disorder and AUCpreD, where yellow residues and red residues represent ordered and disordered residues, respectively. (**A**) The real structure of protein 3CSZA without IDRs; (**B**) no false IDR predicted by RFPR_IDP; (**C**) false IDRs predicted by DISOPRED3 are: {1, 8} and {11, 11}; (**D**) false IDR predicted by SPOT-disorder is: {1, 5}; (**E**) false IDRs predicted by AUCpreD are: {1, 1} and {159, 159}. The curly brace indicates the positional interval of region in the protein.

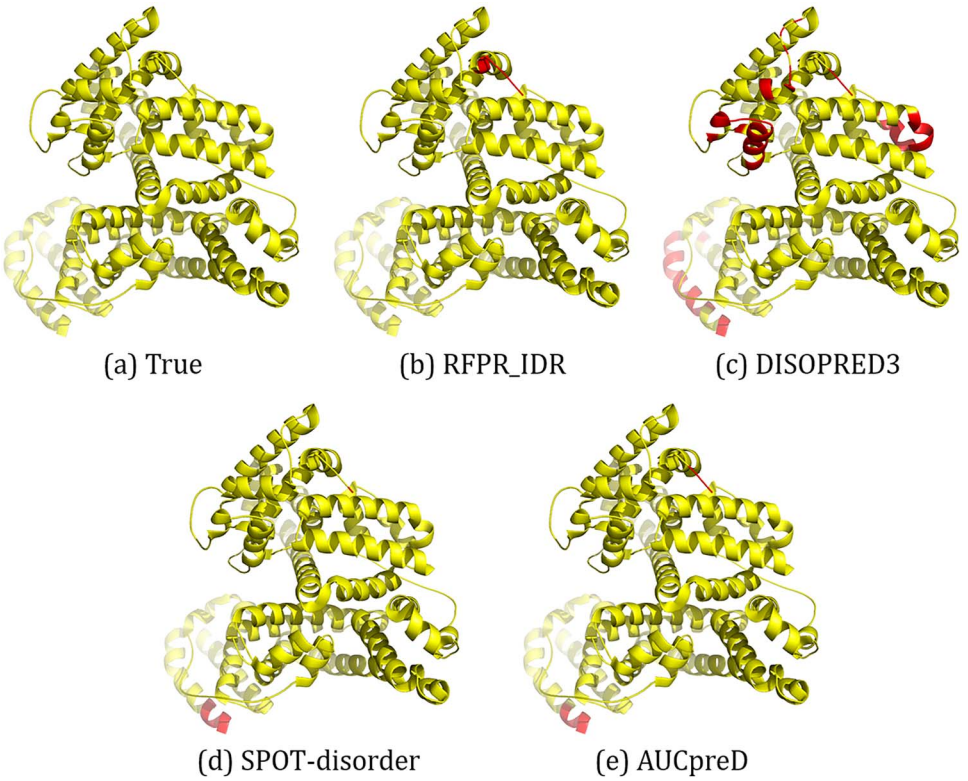


**Figure 7**. **A schematic diagram of ordered protein 5OSWA predicted by RFPR_IDR, DISOPRED3, SPOT-disorder and AUCpreD, where yellow residues and red residues represent ordered and disordered residues, respectively.** (**A**) The real structure of protein 5OSWA without IDRs; (**B**) false IDR predicted by RFPR_IDP is: {1, 6}; (**C**) false IDRs predicted by DISOPRED3 are: {1, 3}, {31, 31}, {58, 58}, {60, 60}, {62, 62}, {64, 65}, {68, 69}, {89, 89}, {91, 104}, {266, 276}, {278, 278}, {558, 578} and {582, 583}; (**D**) false IDRs predicted by SPOT-disorder are: {1, 1} and {578, 583}; (**E**) false IDRs predicted by AUCpreD are: {1, 3} and {579, 583}. The curly brace indicates the positional interval of region in the protein.

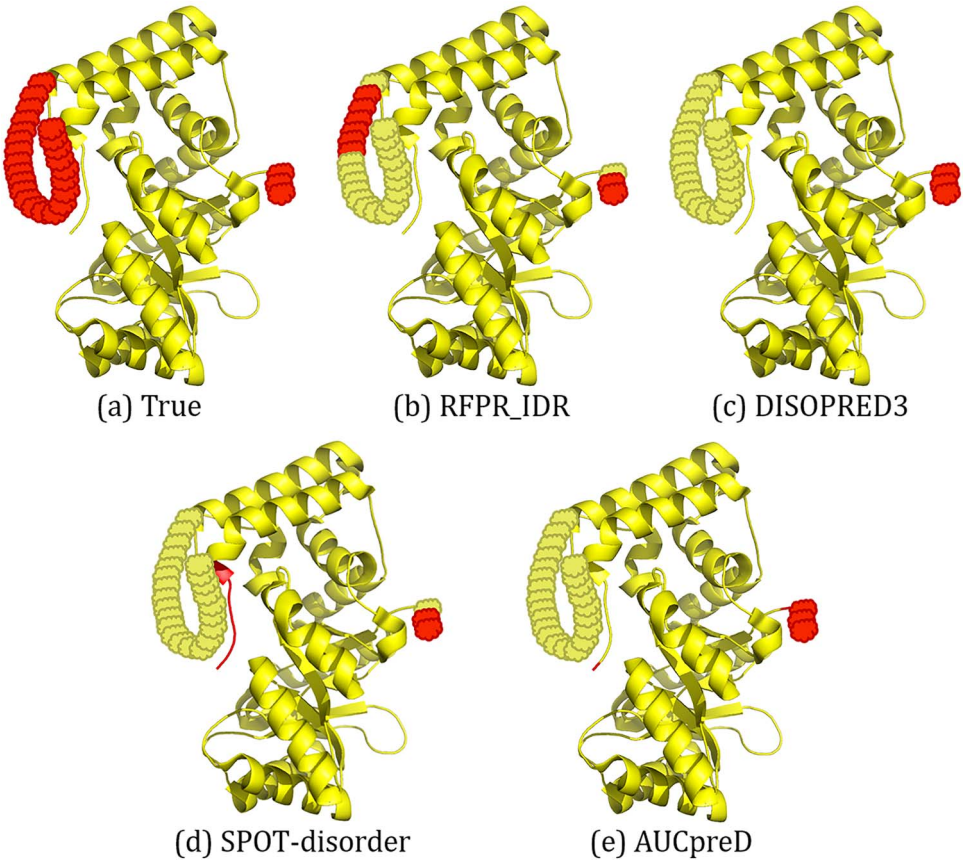


**Figure 8**. A schematic diagram of protein 1GVEB with IDRs predicted by RFPR_IDR, DISOPRED3, SPOT-disorder and AUCpreD, where yellow residues and red residues represent ordered and disordered residues, respectively. **(A)** True IDRs: {1, 3} and {209, 231}; **(B)** IDRs predicted by RFPR_IDR are: {1, 2} and {210, 216}; **(C)** IDR predicted by DISOPRED3 is: {1, 3}; **(D)** IDRs predicted by SPOT-disorder are: {1, 2} and {321, 327}; **(E)** IDRs predicted by AUCpreD are: {1, 4} and {327, 327}. The curly brace indicates the positional interval of region in the protein.

the results, we conclude that the predictors in this field should consider both the fully disordered proteins and ordered proteins during their training and test processes. Otherwise, a predictor will prefer to predict the fully ordered proteins as disordered proteins. This will prevent their real-world applications, because most of the proteins are fully ordered proteins in nature. Based on these findings, we make an attempt to propose the RFPR-IDP predictor to use both the fully ordered proteins and disordered proteins and show that this method is able to overcome the aforementioned disadvantage. The performance improvement of the RFPR-IDP is mainly benefited from the incorporation of the fully ordered proteins into the training processes. It should be noted that although RFPR-IDP shows better performance for predicting fully ordered proteins, AUCpreD [23] and SPINE-D [15] outperform the RFPR-IDP for predicting the disordered proteins (see Table 3). Therefore, we believe that these two methods and other approaches will be benefited from the findings of this study and will be improved by considering both the disordered proteins and fully ordered proteins.

## Key points

- As reported in previous studies, most of the proteins are fully ordered proteins in nature. However, most of the existing methods are only evaluated on datasets consisting of disordered proteins without or with only a few fully ordered proteins. As a result, these predictors prefer to predict a fully ordered protein as a disordered protein, preventing their real-world applications.
- In order to solve this problem, we propose a new method called RFPR-IDP trained and tested with both fully ordered proteins and disordered proteins. RFPR-IDP is constructed based on the combination of convolution neural network (CNN) and bidirectional long short-term memory (BiLSTM), in which CNN can capture the local information or motifs of proteins and BiLSTM can learn the long-term dependence information in both directions of proteins.
- RFPR-IDP outperforms 10 existing state-of-the-art methods in this field with the lowest FPRs, especially for predicting the fully ordered proteins. We believe that other methods will be benefitted from the findings of this study, and they can be further improved by considering both the fully ordered proteins and disordered proteins.

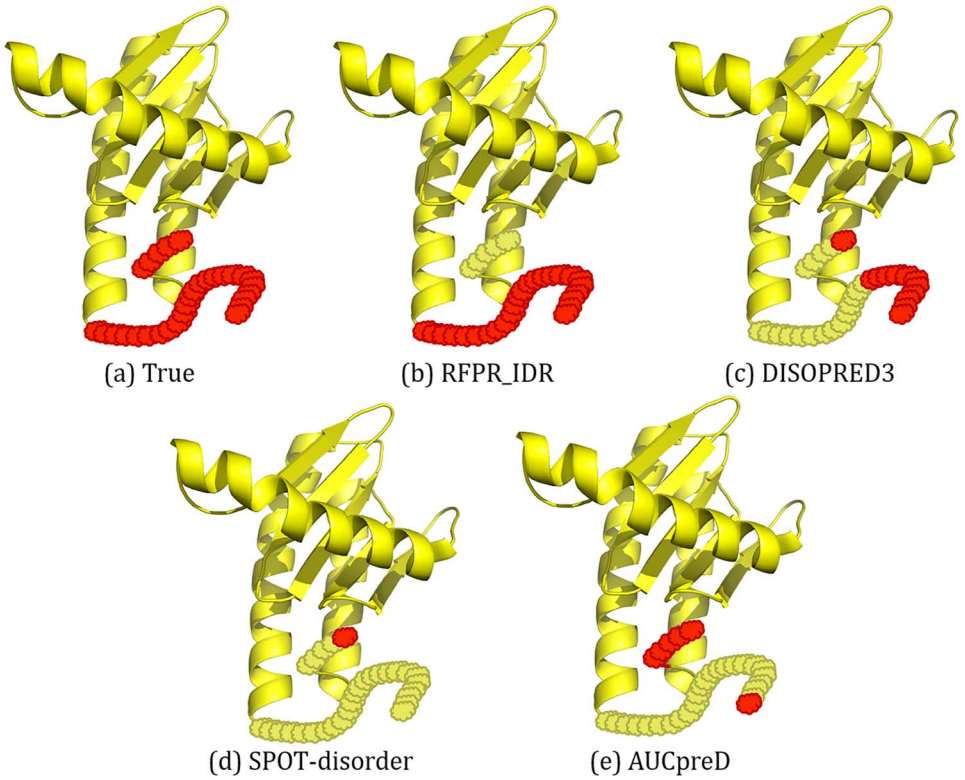


**Figure 9**. A schematic diagram of protein 1J3WA with IDRs predicted by RFPR_IDR, DISOPRED3, SPOT-disorder and AUCpreD, where yellow residues and red residues represent ordered and disordered residues, respectively. **(A)** True IDRs: {1, 5} and {140, 163}; **(B)** IDR predicted by RFPR_IDR is: {140, 163}; **(C)** IDRs predicted by DISOPRED3 are: {1, 1} and {153, 163}; **(D)** IDR predicted by SPOT-disorder is: {1, 1}; **(E)** IDRs predicted by AUCpreD are {1, 6} and {163, 163}. The curly brace indicates the positional interval of region in the protein.

## Acknowledgements

The authors are very much indebted to the four anonymous reviewers, whose constructive comments are very helpful in strengthening the presentation of this article.

## Funding

National Natural Science Foundation of China (61822306, 61672184 and 61732012 61573118); Beijing Natural Science Foundation (JQ19019); Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China (161063); Scientific Research Foundation in Shenzhen (JCYJ20180306172207178, JCYJ20170307150528934 and JCYJ20170811153836555).

## Conflicts of interest

The authors declare no competing interests.

## References

1. Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform* 2019;**20**:330–46.
2. Liu Y, Wang X, Liu B. IDP–CRF: intrinsically disordered protein/region identification based on conditional random fields. *Int J Mol Sci* 2018;**19**: 2483.
3. Holmstrom ED, Liu Z, Nettels D, *et al*. Disordered RNA chaperones can enhance nucleic acid folding via local charge screening. *Nat Commun* 2019;**10**:1–11.
4. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* 2015;**16**:18–29.
5. van der Lee R, Buljan M, Lang B, *et al*. Classification of intrinsically disordered regions and proteins. *Chem Rev* 2014;**114**:6589–631.
6. Piovesan D, Tabaro F, Mičetić I, *et al*. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res* 2017;**45**:D219–27.
7. Iakoucheva LM, Brown CJ, Lawson JD, *et al*. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;**323**:573–84.
8. H Jane D, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 2005;**6**:197–208.
9. Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008;**37**:215–46.
10. Uversky VN, Oldfield CJ, Midic U, *et al*. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* 2009;**10**(Suppl 1):S7.
11. Receveur-Brechot V, Bourhis JM, Uversky VN, *et al*. Assessing protein disorder and induced folding. *Proteins* 2006;**62**: 24–45.
12. Dosztányi Z, Csizmok V, Tompa P, *et al*. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;**21**:3433–4.

13. Linding R, Russell RB, Neduva V, *et al*. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 2003;**31**:3701–8.
14. Hanson J, Yang Y, Paliwal K, *et al*. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* 2017, **33**: 685–92.
15. Zhang T, Faraggi E, Xue B, *et al*. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn* 2012;**29**: 799–813.
16. Linding R, Jensen LJ, Diella F, *et al*. Protein disorder prediction: implications for structural proteomics. *Structure* 2003;**11**:1453–9.
17. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 2015;**31**:857–63.
18. Tompa P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* 2012;**37**:509–16.
19. Peng Z, Mizianty MJ, Kurgan L. Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* 2014;**82**:145–58.
20. Ward JJ, Sodhi JS, McGuffin LJ, *et al*. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;**337**:635–45.
21. Pentony MM, Jones DT. Modularity of intrinsic disorder in the human proteome, proteins-structure function. *Bioinformatics* 2010;**78**:212–21.
22. Rita P, Peter T. Structural disorder in eukaryotes. *PLoS ONE* 2012;**7**:e34687.
23. Wang S, Ma J, Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. 2016;**32**:i672–9.
24. Wang S, Sun S, Li Z, *et al*. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**:e1005324.
25. Li C-C, Liu B. MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief Bioinform*. doi: 10.1093/bib/bbz133.
26. Sheng W, Shunyan W, Jianzhu M, *et al*. DeepCNF-D: predicting protein order/disorder regions by weighted deep convolutional neural fields. *Int J Mol Sci* **16**:17315–30.
27. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
28. Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition. In: *Artificial Neural Networks: Formal Models & Their Applications-icann*. Warsaw: International Conference, 2005: 799–804.
29. Liu B, Li S. ProtDet-CCH: Protein remote homology detection by combining Long Short-Term Memory and ranking methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2019;**16**:1203–1210.
30. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 2012: 1097–1105.
31. Berman HM, Westbrook J, Feng Z, *et al*. The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.
32. Cheng J, Sweredoski MJ, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Min Knowl Disc* 2005;**11**:213–22.
33. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
34. Sirota FL, Ooi HS, Gattermayer T, *et al*. Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics* 2010;**11**:S15–5.
35. Liu B, Li K. iPromoter-2L2.0: identifying promoters and their types by combining Smoothing Cutting Window algorithm and sequence-based features. *Molecular Therapy-Nucleic Acids* 2019;**18**:80–87.
36. Liu B, Chen J, Guo M, *et al*. Protein remote homology detection and fold recognition based on Sequence-Order Frequency Matrix. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2019;**16**:292–300.
37. Chen Z, Zhao P, Li F, *et al*. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief Bioinform* doi: 10.1093/bib/bbz041.
38. Liu B, Gao X, Zhang H. BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res* 2019;**47**:e127.
39. Yan K, Fang X, Xu Y, *et al*. Protein fold recognition based on multi-view modeling. *Bioinformatics*, 2019;**35**: 2982–2990.
40. Chen Z, Zhao P, Li F, *et al*. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 2018;**34**:2499–2502.
41. Bao W, Huang Z, Yuan C-A, *et al*. Pupylation sites prediction with ensemble classification model. *International Journal of Data Mining and Bioinformatics (IJDMB)* 2017;**18**:91–104.
42. Wei L, Tang J, Zou Q. Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform Sci* 2017;**384**:135–44.
43. Liu B, Jiang S, Zou Q. HITS-PR-HHblits: protein remote homology detection by combining PageRank and hyperlink-induced topic search. *Brief Bioinform* 2020, **21**:298–308.
44. Yan K, Wen J, Liu J-X, *et al*. Protein fold recognition by combining support vector machines and pairwise sequence similarity scores. *IEEE ACM T Comput Biol Bioinf*. doi: 10.1109/TCBB.2020.2966450.
45. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics* 1998;**14**:423–9.
46. Meiler J, Müller M, Zeidler A, *et al*. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J Mol Model* 2001;**7**: 360–9.
47. Zou Q, Xing P, Wei L, *et al*. Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 2019;**25**:205–18.
48. Zeng XX, Wang W, Deng GS, *et al*. Prediction of potential disease-associated microRNAs by using neural networks. *Mol Ther Nucleic Acids* 2019;**16**:566–75.
49. Liu B, Li C, Yan K. DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief Bioinform*. doi: 10.1093/bib/bbz098.
50. Zhang J, Chen Q, Liu B. DeepDRBP-2L: a new genome annotation predictor for identifying DNA-binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory. *IEEE/ACM Trans Comput Biol Bioinform*. doi: 10.1109/TCBB.2019.2952338.
51. Glorot X, Bordes A, Bengio Y. *Deep Sparse Rectifier Neural Networks*. International Conference on Artificial Intelligence & Statistics, 2011: 315–23.

52. Abadi M, Agarwal A, Barham P, *et al*. TensorFlow: large-scale machine learning on heterogeneous distributed systems. 2015.
53. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* 2015.
54. Srivastava N, Hinton G, Krizhevsky A, *et al*. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
55. Aurelio YS, de Almeida GM, de Castro CL, *et al*. Learning from imbalanced data sets with weighted cross-entropy function. *Neural Process Lett* 2019; **50**: 1937–49.
56. Liu Y, Chen S, *Wang X et al*. Identification of intrinsically disordered proteins and regions by length-dependent predictors based on conditional random fields, *Mol Ther Nucleic Acids* 2019; **17**: 396–404.
57. Zhao X, Zou Q, Liu B, *et al*. Exploratory predicting protein folding model with random forest and hybrid features. *Current Proteom* 2014;**11**:289–99.
58. Monastyrskyy B, Kryshtafovych A, Moult J, *et al*. Assessment of protein disorder region predictions in CASP10. *Proteins* 2014;**82**(Suppl 2):127–37.
59. Bordoli L, Kiefer F, Schwede T. Assessment of disorder predictions in CASP7. *Proteins* 2007;**69**(Suppl 8): 129–36.
60. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. *Proteins* 2009;**77**(Suppl 9): 210–6.
61. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res* 2017;**45**: e84.