Molecular Therapy
Nucleic Acids
Original Article

Identification of Intrinsically Disordered Proteins and Regions by Length-Dependent Predictors Based on Conditional Random Fields

Yumeng Liu,[1] Shengyu Chen,[2] Xiaolong Wang,[1] and Bin Liu[1,3,4]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China; [2]School of Informatics, Computing and Engineering, Indiana University Bloomington, Bloomington, IN 47408, USA; [3]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China; [4]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China

Accurate identification of intrinsically disordered proteins/regions (IDPs/IDRs) is critical for predicting protein structure and function. Previous studies have shown that IDRs of different lengths have different characteristics, and several classification-based predictors have been proposed for predicting different types of IDRs. Compared with these classification-based predictors, the previously proposed predictor IDP-CRF exhibits state-of-the-art performance for predicting IDPs/IDRs, which is a sequence labeling model based on conditional random fields (CRFs). Motivated by these methods, we propose a predictor called IDP-FSP, which is an ensemble of three CRF-based predictors called IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G. These three predictors are specially designed to predict long, short, and generic disordered regions, respectively, and they are constructed based on different features. To the best of our knowledge, IDP-FSP is the first predictor that combines a sequence labeling algorithm with IDRs of different lengths. Experimental results using two independent test datasets show that IDP-FSP achieves better or at least comparable predictive performance with 26 existing state-of-the-art methods in this field, proving the effectiveness of IDP-FSP.

INTRODUCTION

Proteins/regions whose native states are intrinsically disordered without a stable 3D structure are called intrinsically disordered proteins/regions (IDPs/IDRs).[1,2] IDPs/IDRs are abundant in all species, especially in eukaryotes.[3] IDPs/IDRs are associated with many biological functions,[4–6] such as regulation of transcription and translation, storage of small molecules, cellular signal transduction, and protein phosphorylation. They execute functions mainly through a disordered state or induced folding when binding to a partner molecule.[2] IDPs/IDRs are associated with many diseases,[7] such as cardiovascular disease,[8] cancer,[5] and genetic diseases.[9] Therefore, accurate identification of IDPs/IDRs is important for drug design and a better understanding of biological processes.

With the help of artificial intelligence and machine learning techniques,[10] some computational predictors have been constructed,[1,11–13] including physiochemically based predictors,[14,15] machine learning-based predictors,[16,17] template-based predictors, and meta-predictors.[18] More information regarding these methods can be found in a recent review paper.[1]

Among these predictors, some predictors are constructed to identify IDRs of different lengths based on the assumption that IDRs with different lengths have different characteristics. In general, the intrinsically disordered regions are divided into long disordered regions (LDRs) and short disordered regions (SDRs). LDRs are defined as regions with more than 30 residues, and SDRs are defined as regions with 30 residues or less. These predictors can be divided into two categories: (1) predictors designed for LDRs or SDRs only, which do not work well for predicting LDRs and SDRs, such as POODLE-L,[19] POODLE-S,[20] and Spritz,[21] and (2) predictors designed for both LDRs and SDRs. Compared with the first-category predictor, these predictors can achieve better performance when predicting both LDRs and SDRs, such as VSL1,[22] VSL2,[23] and SPINE-D.[24] The superior performance of these predictors indicates that length-dependent predictors can capture the different characteristics of IDRs with vary lengths. Furthermore, the better performance of the second-category predictors shows that these length-dependent predictors are complementary. According to the comparison results presented in a recent review paper,[1] the sequence labeling methods outperform the classification methods, and the latest proposed IDP-conditional random field (CRF) predictor[25] based on CRFs achieves state-of-the-art performance.

Inspired by these methods, we propose a predictor called IDP-FSP based on CRFs. IDP-FSP is a fusion of three CRF-based predictors—IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G—that are specifically

Received 25 February 2019; accepted 7 June 2019;
https://doi.org/10.1016/j.omtn.2019.06.004.

Correspondence: Xiaolong Wang, School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China.
E-mail: wangxl@insun.hit.edu.cn
Correspondence: Bin Liu, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China.
E-mail: bliu@bliulab.net

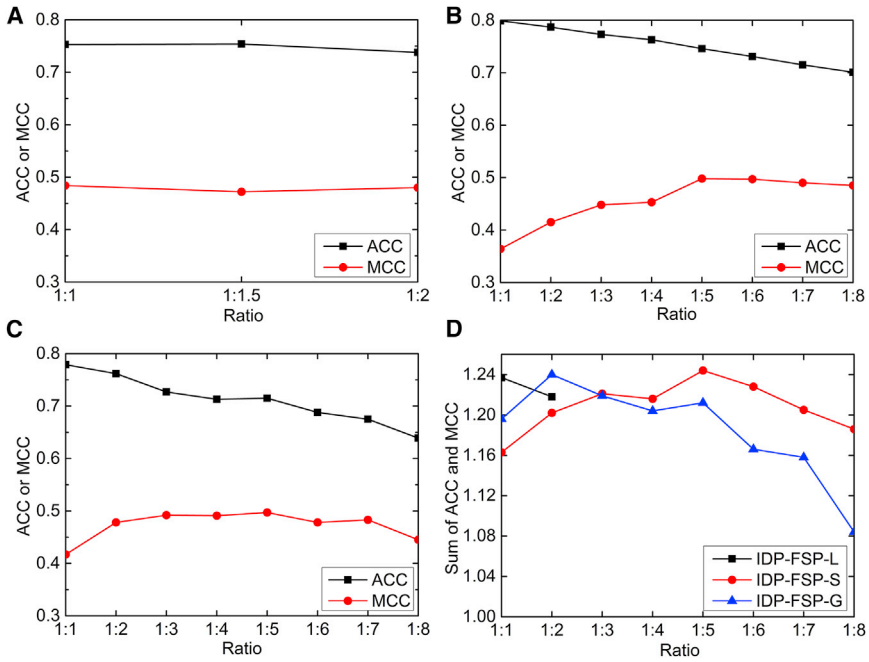

Figure 1. The Influence of Different Ratios of Positive and Negative Samples on the Performance of the Proposed Three Predictors

(A–C) IDP-FSP-L (A), IDP-FSP-S (B), and IDP-FSP-G (C). (D) The sum of the ACC and MCC of the proposed three predictors in different ratios, used as the performance measure for selecting the final optimal ratio.

designed to predict long, short, and generic disordered regions, respectively. To the best of our knowledge, IDP-FSP is the first predictor that combines a sequence labeling algorithm with IDRs of different lengths. Experimental results using two independent test datasets show that IDP-FSP achieves better or at least comparable predictive performance with 26 highly related state-of-the-art predictors in this field.

RESULTS AND DISCUSSION

The Influence of Different Ratios of Positive and Negative Samples on Three Length-Dependent Predictors

In this study, a series of training datasets with different ratios of positive and negative samples is generated by randomly removing negative samples from the origin training datasets. Different ratios would affect the performance of the computational predictors, and both accuracy (ACC) and Matthew's correlation coefficient (MCC) are two important metrics in this field. The ACC, MCC, and the sum of ACC and MCC changing curves of the proposed three specialized predictors with different ratios are shown in Figure 1. We can see that different ratios have significant effects on the performance of the predictors IDP-FSP-S and IDP-FSP-G compared with IDP-FSP-L because the positive and negative ratios of the training datasets of IDP-FSP-S and IDP-FSP-G are extremely imbalanced. In particular, IDP-FSP-S and IDP-FSP-G achieve the best performance when the training datasets are imbalanced, which helps CRFs to capture the imbalanced information between positive and negative samples. In Figure 1D, we can see that IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G achieve the best performance at 1:1, 1:5, and 1:2, respectively. This shows that LDRs, SDRs, and generic disordered regions have different characteristics.

The Influence of Different Window Sizes on Three Length-Dependent Predictors

The window size represents the length of the subsequence centered on the target residue, which is a parameter in the process of feature extraction. To explore its influence on the performance of the proposed models, the ACC, MCC, and the sum of ACC and MCC changing curves with different window sizes of our proposed three specialized predictors are shown in Figure 2. We can see that different window sizes have little influence on the performance of these three specialized predictors. In Figure 2D, we can see that IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G achieve the best performance with window sizes of 13, 11, and 13, respectively.

Fusion of Length-Dependent Predictors Can Improve Predictive Performance

IDP-FSP is an ensemble of three length-dependent predictors, including IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G. This fusion approach has been successfully applied to solve many important tasks in bioinformatics.[26–28] These three specialized predictors are trained on their respective types of training datasets, and their parameters are adjusted separately by using their corresponding benchmark test datasets. The performance of IDP-FSP-L, IDP-FSP-S, IDP-FSP-G, and IDP-FSP on different types of test datasets is shown in Table 1 and Figure 3, from which the following conclusions can be drawn. (1) IDP-FSP-L and IDP-FSP-S have better predictive performance on their corresponding datasets than on other datasets, which fully illustrates that LDRs and SDRs have different characteristics. (2) For a certain type of dataset, the corresponding predictor obtains better or comparable performance in terms of ACC. IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G are constructed based on different types of datasets and use different positive and negative ratios of training datasets, which are 1:1, 1:5, and 1:2, respectively. The more negative samples are in the training dataset, the more negative sample information is included, leading to higher predictive performance for negative samples. Therefore, IDP-FSP-S and IDP-FSP-G outperform IDP-FSP-L for predicting the negative samples. For test datasets $\mathbb{S}_{all}^{Test}$ and $\mathbb{S}_{short}^{Test}$, the proportion of negative samples is much higher than that of $\mathbb{S}_{long}^{Test}$, and, therefore, the MCC is more dependent on the predictive performance of negative samples. As a result, the MCC values obtained by IDP-FSP-S and IDP-FSP-G are higher than that

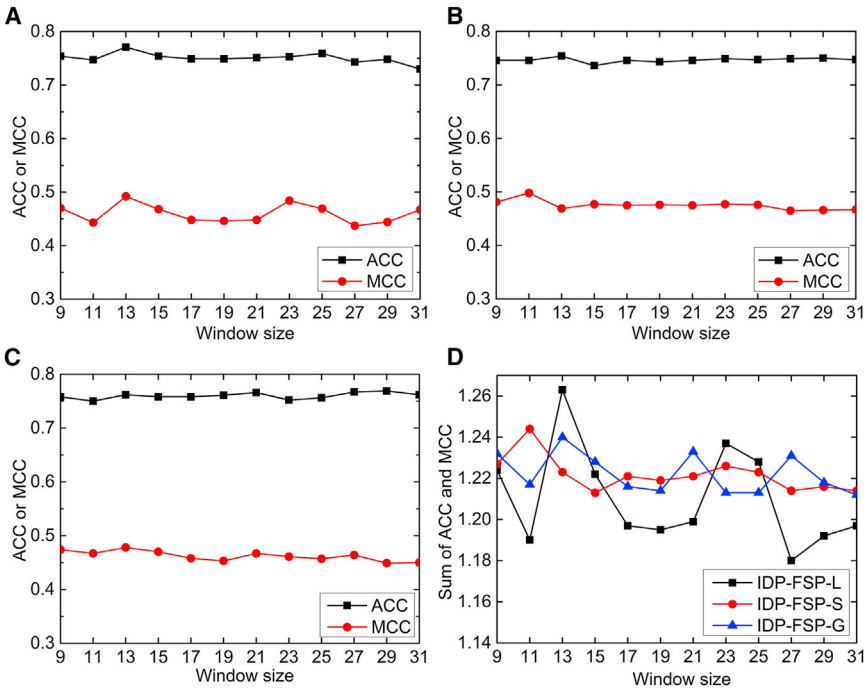

Figure 2. The Influence of Different Window Sizes on the Performance of the Proposed Three Predictors

(A–C) IDP-FSP-L (A), IDP-FSP-S (B), and IDP-FSP-G (C). (D) The sum of the ACC and MCC of the proposed three predictors in different window sizes, used as the performance measure for selecting the final optimal window size.

of IDP-FSP-L. (3) IDP-FSP outperforms IDP-FSP-G on every dataset, indicating that fusing these specialized predictors can effectively improve predictive performance.

IDP-FSP achieves good performance mainly because the following two reasons: (1) IDP-FSP is a fusion of three specializes predictors constructed based on the LDR dataset, SDR dataset, and generic dataset, respectively. Therefore, IDP-FSP can capture the characteristics of different IDRs. (4) IDP-FSP is able to capture the complementarity of the three specialized predictors.

Visualization of Predicted Proteins

In this section, the true structure and the predicted structure of three proteins are visualized to show the advantages of our proposed method. These proteins are PDB: 1MSVA, 3KC2B, and 1O0BA.[29] For these proteins, the PyMOL (https://pymol.org/2/) software is used to generate the 3D structure of their ordered regions, and the 3D structure of their disordered regions is drawn manually.

The schematic diagrams of PDB: 1MSVA, 3KC2B, and 1O0BA are shown in Figures 4, 5, and 6, respectively. According to these figures, we can observe the following. (1) IDP-FSP-L and IDP-FSP-S can correctly identify some IDRs incorrectly predicted by IDP-FSP-G. For example, for the IDR {22, 27} of protein PDB: 1MSVA in Figure 4 and the IDR {268, 288} of protein PDB: 3KC2B in Figure 5, IDP-FSP-G fails to identify them. However, IDP-FSP is able to identify them. (2) IDP-FSP-L and IDP-FSP-S can correct some erroneous IDRs predicted by IDP-FSP-G. For example, for the ordered region {134, 171} of protein PDB: 1O0BA in Figure 6, IDP-FSP-G predicts it as an IDR. However, both IDP-FSP-L and IDP-FSP-S predict it as an ordered region, correcting the predictive results of IDP-FSP-G. (3) For some regions, IDP-FSP is more accurate than IDP-FSP-G. For example, for the IDR {443, 453} of protein PDB: 1O0BA in Figure 6, IDP-FSP-G predicts {438, 454} as an IDR, and IDP-FSP predicts {443, 453} as an IDR. For the IDR {329, 354} of protein PDB: 1MSVA in Figure 4, IDP-FSP-G predicts {342, 345} as an IDR, and IDP-FSP predicts {328, 354} as an IDR, which corrects 13 false negatives predicted by IDP-FSP-G. These observations indicate that IDP-FSP-L and IDP-FSP-S can capture the characteristics of LDRs and SDRs, respectively, and can predict some IDRs that IDP-FSP-G fails to predict. Besides, it is further proven that IDP-FSP can capture the complementarity of these three length-dependent predictors.

Comparison with the Existing Methods

The proposed method is compared with 26 highly related methods using two widely used independent test datasets (MxD494 and SL329). As shown in Tables 2 and 3, IDP-FSP achieves a predictive performance comparable with two state-of-the-art predictors and outperforms 24 other highly related predictors. In particular,

Table 1. Performance Comparison of IDP-FSP-L, IDP-FSP-S, IDP-FSP-G, and IDP-FSP on the LDR Test Dataset, SDR Test Dataset, and General Test Dataset, Respectively

Test Datasets[a]	Predictors	Sn	Sp	ACC	MCC
$\mathbb{S}_{long}^{Test}$	IDP-FSP-L	0.788	0.754	0.771	0.492
	IDP-FSP-S	0.446	0.958	0.702	0.501
	IDP-FSP-G	0.582	0.916	0.749	0.533
	IDP-FSP	0.585	0.923	0.754	0.550
$\mathbb{S}_{short}^{Test}$	IDP-FSP-L	0.662	0.753	0.708	0.228
	IDP-FSP-S	0.552	0.961	0.757	0.487
	IDP-FSP-G	0.597	0.935	0.766	0.437
	IDP-FSP	0.599	0.940	0.770	0.451
$\mathbb{S}_{all}^{Test}$	IDP-FSP-L	0.716	0.753	0.735	0.303
	IDP-FSP-S	0.507	0.961	0.734	0.495
	IDP-FSP-G	0.590	0.933	0.762	0.478
	IDP-FSP	0.593	0.938	0.766	0.493

[a]These datasets are described in Equation 1.

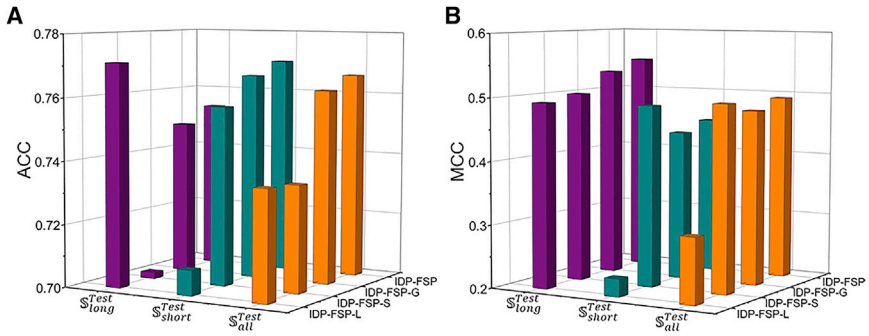

Figure 3. Performance Comparison of IDP-FSP-L, IDP-FSP-S, IDP-FSP-G, and IDP-FSP on Different Types of Test Datasets

(A) ACC comparison of these four predictors on different test datasets. (B) MCC comparison of these four predictors on different test datasets. In each subgraph, the three axes represent datasets, predictors, and performance metrics, respectively. The datasets $\mathbb{S}_{long}^{Test}$, $\mathbb{S}_{short}^{Test}$, and $\mathbb{S}_{all}^{Test}$ are described in Benchmark Datasets.

IDP-FSP outperforms IDP-CRF in predicting MxD494 and SL329. IDP-CRF is also a CRF-based model. Different from IDP-CRF, IDP-FSP is a fusion of three CRF-based predictors that are constructed based on different types of IDRs. This fully demonstrates the effectiveness of constructing predictors for LDRs, SDRs, and generic disordered regions. In addition to IDP-CRF, IDP-FSP achieves performance comparable with MFDp[18] on dataset MxD494 and SPOT-disorder[17] on dataset SL329 and outperforms all other related methods using these two datasets.

Conclusions

In this study, an ensemble predictor, IDP-FSP, is proposed that fuses three length-dependent predictors specially designed for the prediction of long, short, and generic disordered regions. The experimental results using different types of test datasets show that LDRs and SDRs have different characteristics, and there is complementarity between these three proposed specialized predictors. The experimental results using two independent test datasets show that IDP-FSP achieves better or at least comparable predictive performance with 26 currently existing state-of-the-art methods. IDP-FSP achieves good performance mainly for the following reasons. The proposed three length-dependent predictors can capture the different characteristics of different types of IDRs. Therefore, IDP-FSP fusing these specialized predictors can capture the characteristics of different types of IDRs and the complementarity among the three specialized predictors.

MATERIALS AND METHODS

Benchmark Datasets

The benchmark dataset used in this study was constructed by Zhang et al.[24] and contains 4,229 proteins, and the similarity between sequences is less than 25%. The benchmark dataset is divided into 3,000 proteins for training and 1,229 proteins for testing. In this study, according to different types of IDRs, the training dataset and test dataset are divided into two datasets. One is the LDR dataset, in which each protein contains at least one LDR, and the other is the SDR dataset, in which each protein contains only SDRs. Therefore, the benchmark dataset can be formatted as

$$\begin{cases} \mathbb{S}_{all}^{Train} = \mathbb{S}_{long}^{Train} \cup \mathbb{S}_{short}^{Train} \\ \mathbb{S}_{all}^{Test} = \mathbb{S}_{long}^{Test} \cup \mathbb{S}_{short}^{Test} \end{cases}, \qquad \text{(Equation 1)}$$

where $\mathbb{S}_{long}^{Train}$ represents the LDR dataset in the training dataset containing 342 proteins, which is used to train IDP-FSP-L; $\mathbb{S}_{short}^{Train}$ represents the SDR dataset in the training dataset containing 2,658 proteins, which is used to train IDP-FSP-S; and $\mathbb{S}_{all}^{Train}$ is the union of

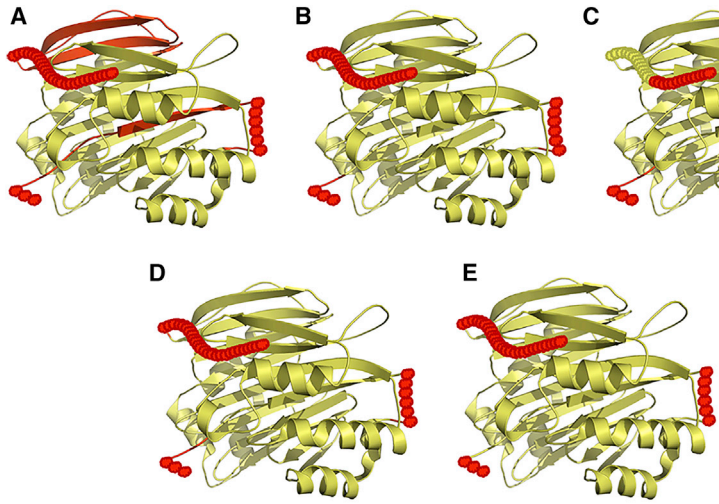

Figure 4. A Schematic Diagram of Protein PDB: 1MSVA with IDRs Predicted by IDP-FSP-L, IDP-FSP-S, IDP-FSP-G, and IDP-FSP

Yellow residues represent ordered residues, and red residues represent disordered residues. (A) IDRs predicted by IDP-FSP-L: {1, 31} and {286, 354}, (B) IDRs predicted by IDP-FSP-S: {1, 6}, {21, 30}, and {328, 354}. (C) IDRs predicted by IDP-FSP-G: {1, 7} and {342, 354}. (D) IDRs predicted by IDP-FSP: {1, 7}, {21, 30}, and {328, 354}. (E) True IDRs: {1, 3}, {22, 27}, and {329, 354}. The curly braces represent the position intervals of IDRs in the protein sequence.

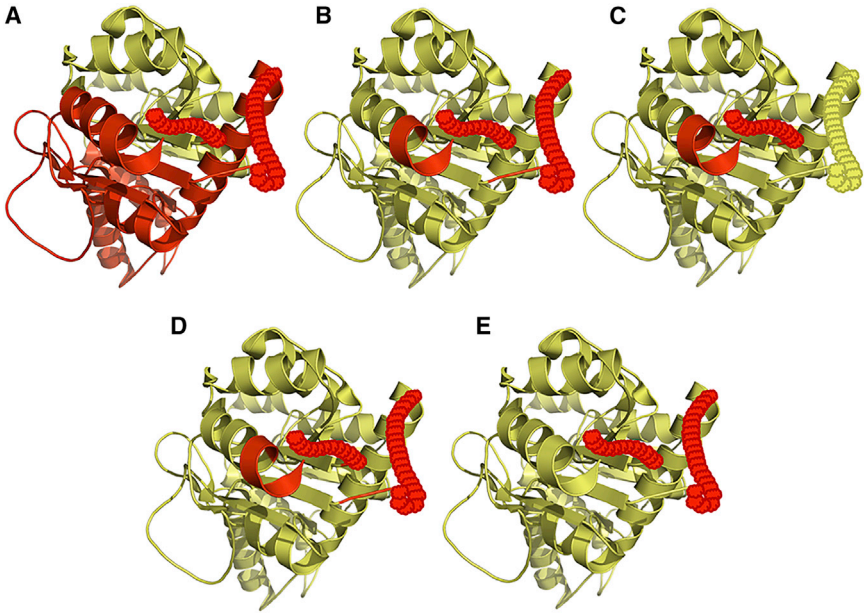

Figure 5. A Schematic Diagram of Protein PDB: 3KC2B with IDRs Predicted by IDP-FSP-L, IDP-FSP-S, IDP-FSP-G, and IDP-FSP

Yellow residues represent ordered residues, and red residues represent disordered residues. (A) IDRs predicted by IDP-FSP-L: {1, 12} and {268, 288}. (B) IDRs predicted by IDP-FSP-S: {1, 12}, {268, 291}, and {348, 352}. (C) IDRs predicted by IDP-FSP-G: {1, 13} and {347, 352}. (D) IDRs predicted by IDP-FSP: {1, 13}, {268, 291}, and {348, 352}. (E) True IDRs: {1, 12} and {268, 288}. The curly braces represent the position intervals of IDRs in the protein sequence.

$\mathbb{S}_{long}^{Train}$ and $\mathbb{S}_{short}^{Train}$, which is used to train IDP-FSP-G. Similarly, $\mathbb{S}_{long}^{Test}$ represents the LDR dataset in the test dataset containing 144 proteins, which is used to test IDP-FSP-L, and $\mathbb{S}_{short}^{Test}$ represents the SDR dataset in the test dataset containing 1,085 proteins, which is used to test IDP-FSP-S. $\mathbb{S}_{all}^{Test}$ is the union of $\mathbb{S}_{long}^{Test}$ and $\mathbb{S}_{short}^{Test}$, which is used to test IDP-FSP-G. These different types of datasets are given in Data S1.

MxD494[18,30] and SL329[17,31] are two widely used independent test datasets adopted for this study to test different methods. To test our method fairly, sequences with a similarity of more than 25% between the benchmark dataset and the two test datasets were removed from the benchmark

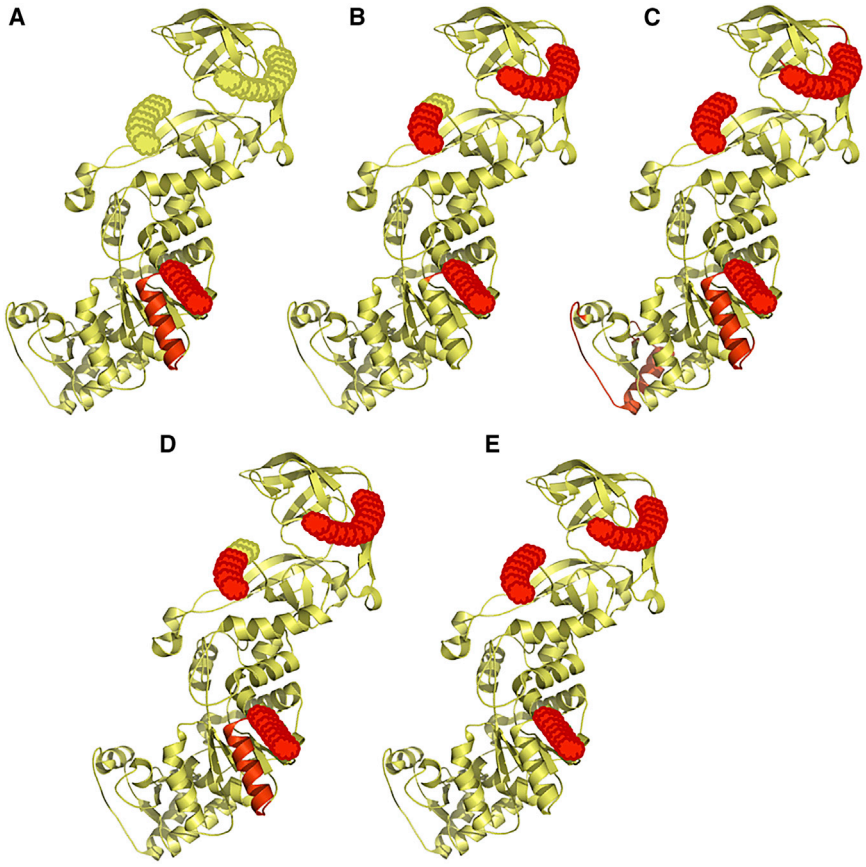

Figure 6. A Schematic Diagram of Protein PDB: 1O0BA with IDRs Predicted by IDP-FSP-L, IDP-FSP-S, IDP-FSP-G, and IDP-FSP

Yellow residues represent ordered residues, and red residues represent disordered residues. (A) IDR predicted by IDP-FSP-L: {1, 24}. (B) IDRs predicted by IDP-FSP-S: {1, 9}, {443, 453}, and {550, 554}. (C) IDRs predicted by IDP-FSP-G: {1, 26}, {134, 171}, {438, 454}, and {548, 554}. (D) IDRs predicted by IDP-FSP: {1, 24}, {443, 453}, and {550, 554}. (E) True IDRs: {1, 7}, {443, 453}, and {548, 554}. The curly braces represent the position intervals of IDRs in the protein sequence.

Table 2. Comparison of Different Predictors on Independent Test Dataset MxD494

Predictor[a]	Sn	Sp	ACC	MCC	Rank ACC	Rank MCC
IDP-FSP[b]	0.670	0.831	0.751	0.465	2	1
IDP-CRF[25]	0.680	0.821	0.750	0.460	3	2
MFDp[18]	0.746	0.768	0.757	0.451	1	3
MD[53]	0.673	0.813	0.743	0.444	4	4
PONDR-FIT[54]	0.631	0.821	0.726	0.419	7	5
DISOPRED2[55]	0.647	0.800	0.724	0.406	8	6
IUPred-long[50]	0.581	0.841	0.711	0.405	9	7
PONDR VSL2B[23]	0.774	0.698	0.736	0.401	5	8
OnD-CRF[16,c]	0.752	0.711	0.732	0.396	6	9
IUPred-short[50]	0.522	0.866	0.694	0.389	11	10
RONN[56]	0.664	0.754	0.709	0.368	10	11
NORSnet[57]	0.532	0.829	0.681	0.347	12	12
DisEMBL-R[58]	0.316	0.936	0.626	0.323	16	13
DISpro[59,60]	0.303	0.940	0.622	0.318	17	14
Ucon[61]	0.554	0.787	0.671	0.313	13	15
Spritz[21]	0.494	0.812	0.653	0.293	15	16
FoldIndex[14]	0.602	0.717	0.660	0.278	14	17
DisEMBL-H[58]	0.435	0.792	0.614	0.216	18	18
PROFbval[62]	0.835	0.387	0.611	0.196	19	19
GlobPlot[15]	0.353	0.826	0.590	0.182	20	20
DisEMBL-C[58]	0.760	0.414	0.587	0.150	21	21

[a]In addition to OnD-CRF, the results of 19 other compared predictors were obtained from Liu et al.[25] and Peng and Kurgan.[30]
[b]IDP-FSP-L with the parameters ratio = 1:1 and window size = 13, IDP-FSP-S with the parameters ratio = 1:5 and window size = 11, and IDP-FSP-G with the parameters ratio = 1:2 and window size = 9.
[c]The results of OnD-CRF were obtained from the web server.

Table 3. Comparison of Different Predictors on Independent Test Dataset SL329

Predictor[a]	Sn	Sp	ACC	MCC	Rank ACC	Rank MCC
IDP-FSP[b]	0.75	0.89	0.821	0.65	1	2
IDP-CRF[25]	0.75	0.88	0.817	0.64	2	3
SPOT-disorder[17]	0.67	0.96	0.815	0.67	3	1
SPINE-D[24]	0.78	0.85	0.815	0.63	3	4
DISOPRED3[63]	–	–	0.795	0.61	5	5
DISOPRED2[55]	0.69	0.90	0.795	0.59	5	6
OnD-CRF[16,c]	0.79	0.80	0.793	0.58	7	7
MD[53]	0.66	0.89	0.775	0.58	8	7
PONDR-FIT[54]	0.61	0.91	0.760	0.55	9	9
IUPred-long[50]	0.60	0.92	0.760	0.55	9	9
MFDp[18]	0.88	0.62	0.750	0.51	12	11
DISOClust[64]	0.81	0.70	0.755	0.51	11	11
NORSnet[57]	0.54	0.92	0.730	0.51	13	11
IUPred-short[50]	0.50	0.94	0.720	0.50	14	14
Ucon[61]	0.59	0.81	0.700	0.42	15	15
DisEMBL[58]	–	–	0.660	0.40	17	16
Dispro[59,60]	0.28	0.99	0.635	0.40	19	16
PONDR VL-XT[49]	0.59	0.78	0.685	0.38	16	18
Espritz[65]	–	–	0.605	0.35	20	19
PROFbval[62]	–	–	0.648	0.30	18	20

[a]In addition to OnD-CRF, the results of 18 other compared predictors were obtained from Hanson et al.,[17] Zhang et al.,[24] and Liu et al.[25]
[b]IDP-FSP-L with the parameters ratio = 1:1 and window size = 13, IDP-FSP-S with the parameters ratio = 1:5 and window size = 11, and IDP-FSP-G with the parameters ratio = 1:2 and window size = 13.
[c]The results of OnD-CRF were obtained from the web server.

dataset[32] by using the Blastclust algorithm[33], and the filtered benchmark dataset was used to retrain IDP-FSP.

Features

Feature extraction plays an important role in machine learning-based predictors.[34–36] The previously proposed predictor IDP-CRF[25] has proven that CRFs combined with PSSMs, kmer, secondary structure (SS), and relative solvent accessibility (RSA) are effective for predicting IDPs/IDRs. Therefore, these four state features were adopted for this study. The features of IDP-FSP-L, IDP-FSP-S and IDP-FSP-G we used are described in The Framework of IDP-FSP.

Transition Feature

Suppose $\varnothing = \{O, D\}$ is the label set of residues, where O and D represent ordered and disordered residues, respectively. The transition feature for each label pair $(l, l' \in \varnothing)$ is defined as[37]

$$t_{l,l'}(y_{i-1}, y_i, x, i) = \begin{cases} 1 & if\ y_{i-1} = l\ and\ y_i = l' \\ 0 & otherwise \end{cases}, \quad \text{(Equation 2)}$$

where y_{i-1} and y_i are defined as labels for position i-1 and position i residues, respectively.

State Features

The PSSMs are generated by running PSI-BLAST[33] searching against the nrdb90 database.[38] For PSI-BLAST, the parameters E-value and iteration time are set to 0.001 and 3, respectively, and other parameters are set to default. The final features are obtained by using the formula[39]

$$\mathrm{norm}(x) = \begin{cases} 0.0 & if\ x \le -5 \\ 0.5 + 0.1x & if\ -5 < x < 5 \\ 1.0 & if\ x \ge 5 \end{cases}, \quad \text{(Equation 3)}$$

where x is the value of PSSMs. kmer defines the occurrence frequencies of k neighboring residues. The PSSM and kmer (k set to 1) features of target residue are constructed based on the subsequence centered on the target residue.

The SS features are obtained by using the profile-based PSIPRED v.4.01 package.[40] If the profile of a protein is not generated by

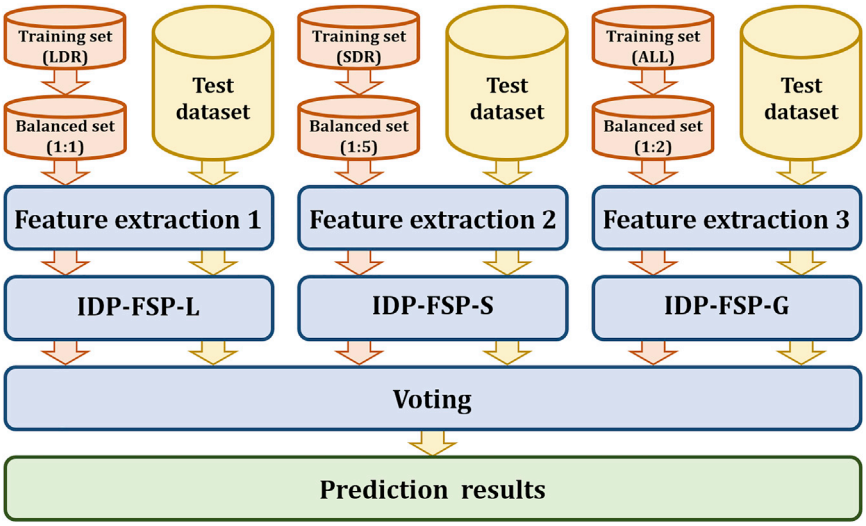

Figure 7. Flowchart of IDP-FSP
The training process and test process are indicated by orange arrows and yellow arrows, respectively.

searching against the nrdb90 database,[38] then sequence-based PSIPRED is used instead of profile-based PSIPRED. The RSA information is generated by using the Sable v.2 package.[41,42] For the Sable package, two parameters, SA_ACTION and SA_OUT, are set to SVR and RELATIVE, respectively, and other parameters are set to default. For each target residue, both SS and RSA are one-dimensional features.

CRFs and Implementation

Being widely used in the field of bioinformatics,[43–46] CRFs are a probabilistic model proposed by Lafferty et al.[37] for labeling sequence data. In this study, protein sequences and their corresponding label sequences are used to train the CRF model, which is a conditional probability model to annotate unlabeled protein sequences. In particular, CRFs have been adopted and proven to be effective for predicting IDPs/IDRs.[25]

FlexCRFs[47] are a widely used tool of CRFs. In this study, the modified FlexCRFs as described in a previous paper[48] are adopted, which can handle real value features. For FlexCRFs, the first-order Markov CRFs are used, and two parameters, num_iterations and init_lambda_val, are set to 50 and 0.05, respectively.

The Framework of IDP-FSP

Previous studies have shown that LDRs and SDRs have different characteristics,[21,23,49,50] so constructing specialized predictors for LDRs and SDRs can effectively improve the predictive performance of IDPs/IDRs. Therefore, we construct three specialized predictors—IDP-FSP-L, IDP-FSP-S, and IDP-FSP-G—for predicting long, short, and generic disordered regions, respectively. These three specialized predictors are built based on CRFs, and the parameters of CRFs are described in CRFs and Implementation. The features and parameters adopted in these three predictors are optimized separately by using their corresponding test datasets. IDP-FSP-L is constructed based on PSSMs and kmer. IDP-FSP-S and IDP-FSP-G are constructed based on all of the features described in Features. Furthermore, the window sizes and ratios of the three predictors were optimized separately and are discussed in The Influence of Different Ratios of Positive and Negative Samples on Three Length-Dependent Predictors and The Influence of Different Window Sizes on Three Length-Dependent Predictors, respectively. However, for unlabeled proteins, it is unknown whether they contain LDRs or SDRs or both types of IDRs. A predictor designed for one type of IDRs cannot achieve good performance for other types. To solve this problem, these three specialized predictors are combined into IDP-FSP via voting. To illustrate this more intuitively, a flowchart of IDP-FSP is shown in Figure 7.

Criteria for Performance Evaluation

Two commonly used metrics, sensitivity (Sn) and specificity (Sp), are used in this study. Because of the imbalance of positive samples (disordered residues) and negative samples (ordered residues) in the IDP/IDR datasets, balanced ACC and MCC are also adopted. They are defined as[51,52]

$$\begin{cases} Sn = \dfrac{TP}{TP+FN} \\ Sp = \dfrac{TN}{TN+FP} \\ ACC = \dfrac{1}{2}\left(\dfrac{TP}{TP+FN} + \dfrac{TN}{TN+FP}\right) \\ MCC = \dfrac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \end{cases}, \quad \text{(Equation 4)}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative samples, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.omtn.2019.06.004.

AUTHOR CONTRIBUTIONS

B.L. and Y.L. designed the experiments and participated in drafting the manuscript and performing the statistical analysis. Y.L. performed the experiments. X.W. and S.C. participated in revising the manuscript.

CONFLICTS OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

The authors are very much indebted to the three anonymous reviewers, whose constructive comments were very helpful in strengthening the presentation of this article. This work was supported by the National Natural Science Foundation of China (61672184, 61732012, 61822306, and 61573118), the Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China (161063), the Shenzhen Overseas High Level Talents Innovation Foundation (KQJSCX20170327161949608), and the Scientific Research Foundation in Shenzhen (JCYJ20180306172207178, JCYJ20170307150528934, and JCYJ20170811153836555).

REFERENCES

1. Liu, Y., Wang, X., and Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Brief. Bioinform. *20*, 330–346.
2. Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. Trends Biochem. Sci. *37*, 509–516.
3. Pancsa, R., and Tompa, P. (2012). Structural disorder in eukaryotes. PLoS ONE *7*, e34687.
4. Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C.J., Aspromonte, M.C., Davey, N.E., Davidović, R., Dosztányi, Z., et al. (2017). DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. *45* (D1), D1123–D1124.
5. Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradović, Z., and Dunker, A.K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. *323*, 573–584.
6. Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.
7. Uversky, V.N., Davé, V., Iakoucheva, L.M., Malaney, P., Metallo, S.J., Pathak, R.R., and Joerger, A.C. (2014). Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases. Chem. Rev. *114*, 6844–6879.
8. Cheng, Y., LeGall, T., Oldfield, C.J., Dunker, A.K., and Uversky, V.N. (2006). Abundance of intrinsic disorder in protein associated with cardiovascular disease. Biochemistry *45*, 10448–10460.
9. Midic, U., Oldfield, C.J., Dunker, A.K., Obradovic, Z., and Uversky, V.N. (2009). Protein disorder in the human diseasome: unfoldomics of human genetic diseases. BMC Genomics *10* (*Suppl 1*), S12.
10. Liu, B., Jiang, S., and Zou, Q. (2018). HITS-PR-HHblits: protein remote homology detection by combining PageRank and Hyperlink-Induced Topic Search. Brief. Bioinform. Published online November 7, 2018. https://doi.org/10.1093/bib/bby104.
11. Deng, X., Eickholt, J., and Cheng, J. (2012). A comprehensive overview of computational protein disorder prediction methods. Mol. Biosyst. *8*, 114–121.
12. Deng, X., Gumm, J., Karki, S., Eickholt, J., and Cheng, J. (2015). An Overview of Practical Applications of Protein Disorder Prediction and Drive for Faster, More Accurate Predictions. Int. J. Mol. Sci. *16*, 15384–15404.
13. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V.N., and Dunker, A.K. (2009). Predicting intrinsic disorder in proteins: an overview. Cell Res. *19*, 929–949.
14. Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. (2005). FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics *21*, 3435–3438.
15. Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. *31*, 3701–3708.
16. Wang, L., and Sauer, U.H. (2008). OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. Bioinformatics *24*, 1401–1402.
17. Hanson, J., Yang, Y., Paliwal, K., and Zhou, Y. (2017). Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics *33*, 685–692.
18. Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., and Kurgan, L. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics *26*, i489–i496.
19. Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., and Noguchi, T. (2007). POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. Bioinformatics *23*, 2046–2053.
20. Shimizu, K., Hirose, S., and Noguchi, T. (2007). POODLE-S: web application for predicting protein disorder by using physicochemical features and reduced amino acid set of a position-specific scoring matrix. Bioinformatics *23*, 2337–2338.
21. Vullo, A., Bortolami, O., Pollastri, G., and Tosatto, S.C. (2006). Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. Nucleic Acids Res. *34*, W164-8.
22. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A.K. (2005). Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins *61* (*Suppl 7*), 176–182.
23. Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. (2006). Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics *7*, 208.
24. Zhang, T., Faraggi, E., Xue, B., Dunker, A.K., Uversky, V.N., and Zhou, Y. (2012). SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. J. Biomol. Struct. Dyn. *29*, 799–813.
25. Liu, Y., Wang, X., and Liu, B. (2018). IDP–CRF: Intrinsically Disordered Protein/Region Identification Based on Conditional Random Fields. Int. J. Mol. Sci. *19*, E2483.
26. Yan, K., Fang, X., Xu, Y., and Liu, B. (2019). Protein Fold Recognition based on Multi-view Modeling (Bioinformatics). Published online January 21, 2019. https://doi.org/10.1093/bioinformatics/btz040.
27. Ding, H., Feng, P.-M., Chen, W., and Lin, H. (2014). Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol. Biosyst. *10*, 2229–2235.
28. Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: An Ensemble Multi-Label Classifier for Human Protein Subcellular Location Prediction with Imbalanced Source. Proteomics *17*, 1700262.
29. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
30. Peng, Z.L., and Kurgan, L. (2012). Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr. Protein Pept. Sci. *13*, 6–18.
31. Sirota, F.L., Ooi, H.S., Gattermayer, T., Schneider, G., Eisenhaber, F., and Maurer-Stroh, S. (2010). Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. BMC Genomics *11* (*Suppl 1*), S15.
32. Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018). Sequence clustering in bioinformatics: an empirical study. Brief. Bioinform. Published online September 18, 2018. https://doi.org/10.1093/bib/bby1090.
33. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.
34. Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Brief. Bioinform. Published online December 19, 2017. https://doi.org/10.1093/bib/bbx165.
35. Liu, B., Yang, F., Huang, D.S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.
36. Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing *173*, 346–354.
37. Lafferty, J.D., Mccallum, A., and Pereira, F.C.N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, C.E. Brodley and A.P. Danyluk, eds. (Morgan Kaufmann Publishers), pp. 282–289.

38. Holm, L., and Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics *14*, 423–429.

39. Kim, H., and Park, H. (2003). Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng. *16*, 553–560.

40. Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

41. Adamczak, R., Porollo, A., and Meller, J. (2004). Accurate prediction of solvent accessibility using neural networks-based regression. Proteins *56*, 753–767.

42. Wagner, M., Adamczak, R., Porollo, A., and Meller, J. (2005). Linear regression models for solvent accessibility prediction in proteins. J. Comput. Biol *12*, 355–369.

43. Dong, Z., Wang, K., Dang, T.K., Gültas, M., Welter, M., Wierschin, T., Stanke, M., and Waack, S. (2014). CRF-based models of protein surfaces improve protein-protein interaction site predictions. BMC Bioinformatics *15*, 277.

44. Hayashida, M., Kamada, M., Song, J., and Akutsu, T. (2011). Conditional random field approach to prediction of protein-protein interactions using domain information. BMC Syst. Biol. *5* (*Suppl 1*), S8.

45. Dang, T.H., Van Leemput, K., Verschoren, A., and Laukens, K. (2008). Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics *24*, 2857–2864.

46. Meysman, P., Dang, T.H., Laukens, K., De Smet, R., Wu, Y., Marchal, K., and Engelen, K. (2011). Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. Nucleic Acids Res. *39*, e6.

47. Xuan, H., and Nguyen, M.L. (2004). FlexCRFs: Flexible Conditional Random Fields, http://flexcrfs.sourceforge.net/.

48. Li, M.H., Lin, L., Wang, X.L., and Liu, T. (2007). Protein-protein interaction site prediction based on conditional random fields. Bioinformatics *23*, 597–604.

49. Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. Proteins *42*, 38–48.

50. Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

51. Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., and Fidelis, K. (2014). Assessment of protein disorder region predictions in CASP10. Proteins *82* (*Suppl 2*), 127–137.

52. Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics.

53. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., and Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. PLoS ONE *4*, e4433.

54. Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., and Uversky, V.N. (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim. Biophys. Acta *1804*, 996–1010.

55. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. *337*, 635–645.

56. Yang, Z.R., Thomson, R., McNeil, P., and Esnouf, R.M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics *21*, 3369–3376.

57. Schlessinger, A., Liu, J., and Rost, B. (2007). Natively unstructured loops differ from other loops. PLoS Comput. Biol. *3*, e140.

58. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. (2003). Protein disorder prediction: implications for structural proteomics. Structure *11*, 1453–1459.

59. Cheng, J., Sweredoski, M.J., and Baldi, P. (2005). Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. Data Min. Knowl. Discov. *11*, 213–222.

60. Hecker, J., Yang, J.Y., and Cheng, J. (2008). Protein disorder prediction at multiple levels of sensitivity and specificity. BMC Genomics *9* (*Suppl 1*), S9.

61. Schlessinger, A., Punta, M., and Rost, B. (2007). Natively unstructured regions in proteins identified from contact predictions. Bioinformatics *23*, 2376–2384.

62. Schlessinger, A., Yachdav, G., and Rost, B. (2006). PROFbval: predict flexible and rigid residues in proteins. Bioinformatics *22*, 891–893.

63. Jones, D.T., and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics *31*, 857–863.

64. McGuffin, L.J. (2008). Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. Bioinformatics *24*, 1798–1804.

65. Walsh, I., Martin, A.J., Di Domenico, T., and Tosatto, S.C. (2012). ESpritz: accurate and fast prediction of protein disorder. Bioinformatics *28*, 503–509.