Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 414327, 8 pages
http://dx.doi.org/10.1155/2013/414327

Research Article

A Novel Method of Predicting Protein Disordered Regions Based on Sequence Features

Tong-Hui Zhao,[1,2] Min Jiang,[3] Tao Huang,[4] Bi-Qing Li,[5] Ning Zhang,[6] Hai-Peng Li,[7] and Yu-Dong Cai[1,2]

[1] *Institute of Systems Biology, Shanghai University, Shanghai 200444, China*
[2] *Department of Mathematics, College of Science, Shanghai University, Shanghai 200444, China*
[3] *State Key Laboratory of Medical Genomics, Institute of Health Sciences, Shanghai Jiaotong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200025, China*
[4] *Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA*
[5] *Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[6] *Department of Biomedical Engineering, Tianjin University, Tianjin Key Lab of BME Measurement, Tianjin 300072, China*
[7] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Hai-Peng Li; lihaipeng@picb.ac.cn and Yu-Dong Cai; cai_yud@yahoo.com.cn

Received 3 March 2013; Accepted 26 March 2013

Academic Editor: Bin Niu

With a large number of disordered proteins and their important functions discovered, it is highly desired to develop effective methods to computationally predict protein disordered regions. In this study, based on Random Forest (RF), Maximum Relevancy Minimum Redundancy (mRMR), and Incremental Feature Selection (IFS), we developed a new method to predict disordered regions in proteins. The mRMR criterion was used to rank the importance of all candidate features. Finally, top 128 features were selected from the ranked feature list to build the optimal model, including 92 Position Specific Scoring Matrix (PSSM) conservation score features and 36 secondary structure features. As a result, Matthews correlation coefficient (MCC) of 0.3895 was achieved on the training set by 10-fold cross-validation. On the basis of predicting results for each query sequence by using the method, we used the scanning and modification strategy to improve the performance. The accuracy (ACC) and MCC were increased by 4% and almost 0.2%, respectively, compared with other three popular predictors: DISOPRED, DISOclust, and OnD-CRF. The selected features may shed some light on the understanding of the formation mechanism of disordered structures, providing guidelines for experimental validation.

1. Introduction

The protein structure-function paradigm has been believed as a dogma in the 20th century. However, the discovery of intrinsically disordered proteins, which have regions devoid of stable secondary structures or have a large number of conformations [1], challenges the traditional view and calls for reassessment of the paradigm.

Eukaryotic proteins apparently have more intrinsic disordered regions than those of bacteria or archaea [2], suggesting also more important functions such as being involved in signaling and regulation of gene expression [3]. Lack of intrinsic structures could render protein additional functions, including binding to different targets [4], transcriptional regulation, translational regulation, and cellular signal transduction regulation [5].

Although there is a growing amount of disordered proteins discovered or shown to have disordered regions under physiological conditions [6], most of them were poorly detected by experimental approaches [2, 4, 5, 7–9]. Firstly, such experimental methods are often time consuming and expensive. Furthermore, it is believed, in X-ray crystallography, that regions missing electron density were related to disorder in many protein structures [6]. However, without additional experiments, it is not sure whether a low electron-density region is intrinsically disordered or is a wobbly domain, or just the result of technical difficulties [2]. NMR spectroscopy, one of the most readily suited techniques for detecting disordered proteins in solution, could also underrepresent a native molten globular domain, which is one of the types of disordered regions [2].

Generally speaking, intrinsically disordered proteins have a biased amino acid composition. Weathers and colleagues reported that amino acid composition was sufficient to be used to accurately recognize disorder [10]. Several algorithms for predicting intrinsically disordered proteins have been developed, such as DISOPRED [11], DISOclust [12], and OnD-CRF [13]. DISOPRED is a web service, which is trained on high resolution X-ray crystal structures and identifies disorder when the electron density map of a residue misses coordinates. It is initially generated sequence profile by a PSI BLAST [14] searching. After being trained using a support vector algorithm, the classifier can output a probability estimate. However, a limitation of this algorithm is that coordinates missing may be caused by the artifact of the crystallization process rather than disorder. DISOclust, based on analysis of three-dimensional structure models, identifies disorder when residues change or are consistently missing. OnD-CRF is a method for predicting the transition between structured and disordered regions. The approach uses conditional random fields relying on features derived from amino acid sequences and secondary structure prediction results.

In the present study, we developed a new strategy for analyzing and predicting protein disordered regions by means of Random Forest (RF), Maximum Relevancy Minimum Redundancy (mRMR), Incremental Feature Selection (IFS), and a scanning and modification strategy. Optimal feature set was selected from candidate features, containing Position Specific Scoring Matrix (PSSM) conservation score features and secondary structure features. Our method outperformed other three existing disorder predictors achieving the highest ACC and MCC values.

2. Materials and Methods

2.1. Benchmark Dataset. In this study, disordered proteins were downloaded from the Database of Protein Disorder (DisProt) (version 4.9) [8], which is constructed based on literature description, providing structured and functional information for intrinsically disordered proteins. Ordered proteins were collected from DisProt database and PDB-Select-25 (the October 2008 version) [15]. PDB-Select-25 is a representative subset of the Protein Data Bank (PDB), containing protein families less than 25% sequence identity [16]. Data was preprocessed according to the following criteria. (i) Only disordered protein chains having more than 50 residues and only proteins with low resolution (≥2 Å) were retained. (ii) Only chains having no missing backbones or side chain coordinates were retained. Finally, 960 protein chains containing 293,780 residues were obtained, in which 55,637 residues were in disordered regions. All protein chains were divided randomly into training set and test set.

A 21-residue sliding window approach was employed along each of the protein sequence, containing the center ordered or disordered residue and 10 residues upstream and downstream of the center residue. Since the dataset used in this study was an unbalanced dataset with much more ordered samples than disordered ones, for the training set, we randomly selected the equal number of ordered samples to match the disordered ones. Finally, 43,903 ordered samples and 43,903 disordered samples from 753 proteins were obtained in the training set, which can be found in Online Supporting Information S1 available online at http://dx.doi.org/10.1155/2013/414327. The test set contained 54,582 ordered samples and 11,734 disordered samples from 192 proteins, which were given in Online Supporting Information S2.

2.2. Feature Extraction

2.2.1. Feature of PSSM Conservation Scores. Evolutionary conservation is considered important in biological sequence analysis. A more conserved residue within a protein sequence indicates that it is under stronger selective pressure and hence more important for the protein function. Mutations on such residues may cause significant changes of the protein. In view of this, we used conservation scores to encode peptides.

Herein, the Position Specific Iterative BLAST (PSI BLAST) was employed to measure the conservation status for a specific residue. For each residue, a 20-dimensional vector was calculated to denote the conservation probabilities of mutations to 20 basic amino acids. For a given protein sequence, a Position Specific Scoring Matrix (PSSM [17]) was obtained, which was constructed by all vectors of all residues in the sequence.

2.2.2. Feature of Secondary Structures. Intrinsically disordered proteins are devoid of well-defined tertiary structures under physiological conditions; however, generally speaking, they often display signs of local secondary structures [18, 19]. After statistical analysis of complex of 24 intrinsically disordered proteins, Fuxreiter et al. [20] found that some regions in disordered proteins had strong preference for helical structures. Therefore, in this study, each amino acid was encoded as three types of secondary structures: helix, strand, or coil, as predicted by SSPro [21]. Helix, strand, and coil are the three major kinds of protein secondary structures. Helix is the protein region with spiral conformation. Strand is a protein structural unit of twisted, pleated sheet. The coil region is the region that does not belong to helix or strand. SSPro predicts the protein secondary structures based on PSI-BLAST profiles with an ensemble neural network model [21].

Thus, each 21-residue peptide was encoded into a vector containing $(20+3)\times 21 = 483$ features. The features are named with following rules: first, the amino acid position ("AA" with position), then, feature types ("PSSM" and "SS"), and last, detail information. For PSSM features, it is the amino acid type. For secondary structure (SS) features, it is the secondary structure code. In secondary structure code, H, E, and C strand for helix, strand, and coil, respectively.

2.3. Maximum Relevancy Minimum Redundancy (mRMR). A classification model containing more features may not have more discriminating power. Additional features may have detrimental effects on the classification such as slowing down the learning process and causing overfitting the training data. It is believed that feature selection is an effective way of reducing the dimension of the feature space to improve the prediction performance.

The Maximum Relevancy Minimum Redundancy (mRMR) method was used in this study to select an optimal feature subset. The mRMR was originally developed by Peng et al. [22] to deal with the microarray data processing. If a feature had better tradeoff between maximum relevance to the target and minimum redundancy among other features, it was deemed as a better feature and would be ranked first (with a smaller index) in the final ordered list. The algorithm is described briefly below.

To determine the relevance properties of the feature space, the mutual information (MI), denoted as I, is defined as

$$I(x,y) = \iint P(x,y)\log\frac{P(x,y)}{P(x)P(y)}dx\,dy, \quad (1)$$

where x and y are two random variables. $P(x, \mathrm{y})$ is the joint probabilistic density function of x and y. $P(x)$ and $P(y)$ are the margin probabilistic density functions of x and y, respectively. To calculate MI, the joint probabilistic density function $P(x, y)$ and the margin probabilistic density functions $P(x)$ and $P(y)$ should be given in advance.

Suppose G denotes the entire feature space; we aim to find a subset S of the features to satisfy both maximum relevance and minimum redundancy.

Based on MI, the following mRMR function is constructed:

$$\max_{f_j\in\Omega_t}\left[I\left(f_j,c\right)-\frac{1}{m}\sum_{f_i\in\Omega_s}I\left(f_{j,}f_i\right)\right] \quad (j=1,2,\ldots,n), \quad (2)$$

where Ω_s is the already selected feature set and Ω_t is the to-be-selected feature set, and m and n are the sizes of these two feature sets, respectively. The higher the ordered rank is, the more important the feature is.

A parameter is introduced here to deal with the continuous variables. In our study, t was assigned to be 1. Finally, an ordered feature list was obtained in which better features had smaller indexes.

The mRMR software could be obtained from http://penglab.janelia.org/proj/mRMR/.

2.4. The Random Forest (RF) Method. The Random Forest (RF) algorithm, firstly introduced by Svetnik [23] in 2003, is a combining ensemble tree-structured classifier. The individual decision tree in the forest depends on a random vector and has independent identically distribution. The Random Forest has been widely used in various fields such as economics and medical and text categorization. It has been also successfully employed in biological prediction problems [24–26] and even can efficiently handle large-scale dataset.

In our research, we use the Random Forest (RF) algorithm to construct a prediction model to predict whether an amino acid is in disordered region or not. The method is briefly introduced as follows.

Firstly, 10 decision trees are grown according to the following criteria.

(a) Suppose the number of cases in the training data is M; sample M cases randomly with replacement from the original data to keep the size of the original data not changing.

(b) When dealing with each note, n predictors are selected randomly in terms of N features (where $n \ll N$). The split on the n predictors is also implemented to split the corresponding note. The m value is set to constant.

(c) Each tree is grown as large as possible and unnecessarily pruned. Then each tree gives the queried input a classification. Finally, the forest will choose the one that has the most votes among the trees.

In this study, the Random Forest classifier in Weka was employed with default parameters. The WEKA program is available at http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.

2.5. The Cross-Validation Method. In the literature, cross-validation methods are used to evaluate the stability of a predictor. The independent dataset test, subsampling (k-fold cross-validation), and jackknife analysis are the three methods generally used [27]. For a given benchmark dataset, the jackknife test generates a unique outcome and is deemed as the most objective one compared to other two methods, as elucidated in [28, 29] and demonstrated by [30, equations (28)–(32)] in. However, to reduce the computational time, in this study, 10-fold cross-validation test was used instead of jackknife test. During the 10-fold cross-validation, the whole dataset is divided into 10 equal parts. Each part is in turn used as test set and the remaining 9 parts as training set. We introduced prediction accuracy (ACC), specificity (SP), sensitivity (SN), and Matthews correlation coefficient (MCC) to evaluate the performance of the predictor, which are calculated as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},$$

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}}, \quad (3)$$

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

2.6. Incremental Feature Selection (IFS). The incremental feature selection (IFS) [31–33] procedure was used to find an optimal subset from the mRMR feature list generated above. Suppose the total number of the features is N; we can obtain N feature subsets which are initiated from a subset containing one feature and generated by adding them one by one from the mRMR feature list.

The ith subset is denoted by

$$S_i = \{f_1, f_2, \ldots, f_i\} \quad (1 \le i \le N). \tag{4}$$

Based on the N feature subsets, N Random Forest predictors were constructed with 10-fold cross-validation evaluating its performance. Then the IFS curve of MCC to the feature subset index i was plotted, in which the peak point was noted as h. Finally an optimal feature subset was obtained with which the corresponding predictor yields the best MCC.

3. Results and Discussion

3.1. Feature Reduction. We calculated the Cramer's V coefficient [34, 35] between features and targets. The Cramer's V coefficient is a statistical measurement derived from the Pearson chi-square test [36]. It ranges from 0 to 1 with smaller value indicating weaker association. Features with Cramer's V coefficient less than 0.1 were removed. After this procedure, 175 features remained containing 112 PSSM conservation features and 63 secondary structure features, which can be found in Online Supporting Information S3.

3.2. The mRMR Result. Two kinds of outcomes were obtained after executing the mRMR program. One was called "MaxRel feature list" that ranked the features according to the relevance to the target; the other was named "mRMR feature list" that ranked the features based on the criteria of maximum relevance and minimum redundancy. In our research, only the "mRMR feature list" was used to select optimal feature subset in the IFS procedure. It was listed in Online Supporting Information S4.

3.3. IFS and Optimal Feature Subset. 175 predictors were constructed based on the 175 feature subsets in the IFS procedure. Prediction results of the predictors were listed in Online Supporting Information S5 and the IFS curve was plotted in Figure 1 in which the MCC reached the topmost 0.3895 with 128 features on the training set. Thus, the top 128 features were considered as the optimal feature subset and were used to construct the final predictor. The 128 features were given in Online Supporting Information S6. The MCC of the predictor on independent test set was 0.2791.

3.4. Feature Analysis. The distribution of the feature types in the final optimal feature set was shown in Figure 2. In the 128 optimal features, 92 were from PSSM conservation scores and 36 from secondary structure features (Figure 2(a)). The two types of features contributed to the prediction. It can be seen from the site-specific distribution of the optimal feature set (Figure 2(b)) that features at sites 8–14 played important roles. In addition, features at sites 1-2, sites 5-6, and sites 15–17, 19, and 21 also had considerable impacts on the prediction of disordered protein.

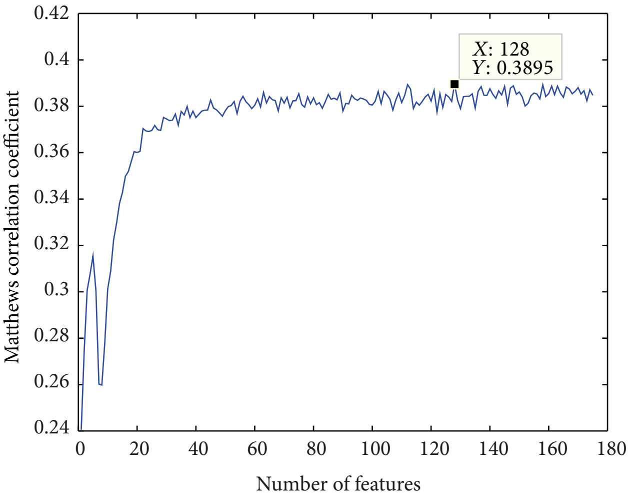

Figure 1: The IFS curve showing the Matthews correlation coefficient (MCC) against the number of features. The details were given in Online Supporting Information S5. With the top 128 features, the MCC on training set by 10-fold cross-validation takes the peak 0.3895.

3.5. PSSM Conservation Score Feature Analysis. As mentioned above, among the 128 optimal features, 92 belonged to the PSSM conservation scores, accounting for the most. Mutations to 20 different amino acids could have different impacts in determining the disordered regions. It can be clearly seen from Figure 3(a) that only 8 out of 20 amino acid mutations were affected. In this regard, the amino acid P (Proline) or S (Serine) could impact most, successively followed by K (Lysine), Q (Glutamine), and so forth. Interestingly, it has been reported that Q was overrepresented in protein interaction domains [37]. It was recently reported that the Ure2p prion and other Q/N-rich yeast prion proteins, which were completely disordered, were driven to format amyloid primarily by intermolecular interactions [38]. Meanwhile, as shown in Figure 3(b), for the 21-length peptides, PSSM conservation scores at sites 8–15 played the most important role. Furthermore, 6 out of the top 10 features in the optimal feature list were PSSM conservation features. The first one was the conservation feature against residue K (Lys) at site 6 (index 2, "AA6_PSSM-12-K"). The other 5 were conservation features against residues E, P, and K at sites 7, 1, and 8, respectively (index 5, "AA7_PSSM-7-E", index 6, and index 7, "AA1_PSSM-15-P" and "AA8_PSSM-12-K") and conservation features against residue E, D at site 21 and site 15 (index 6 and index 7, "AA21_PSSM-7-E" and "AA15_PSSM-4-D").

3.6. Secondary Structure Feature Analysis. The feature subtypes and site-specific distributions of the secondary structure features in the optimal feature set were plotted in Figure 4. From Figure 4, it can be seen that features of "coil"

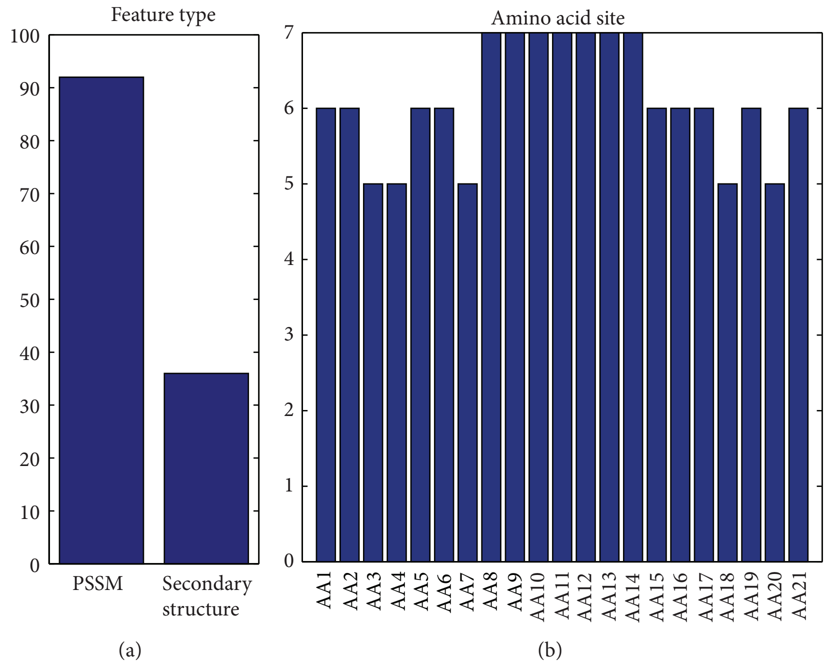

Figure 2: The distribution of feature types and amino acid sites in optimal feature subset. The histograms show the number of each type and each site of features in optimal feature subset. In (a), there are 92 PSSM features and 36 secondary structure features. (b) provides the site distributions of the features in the optimal feature set.

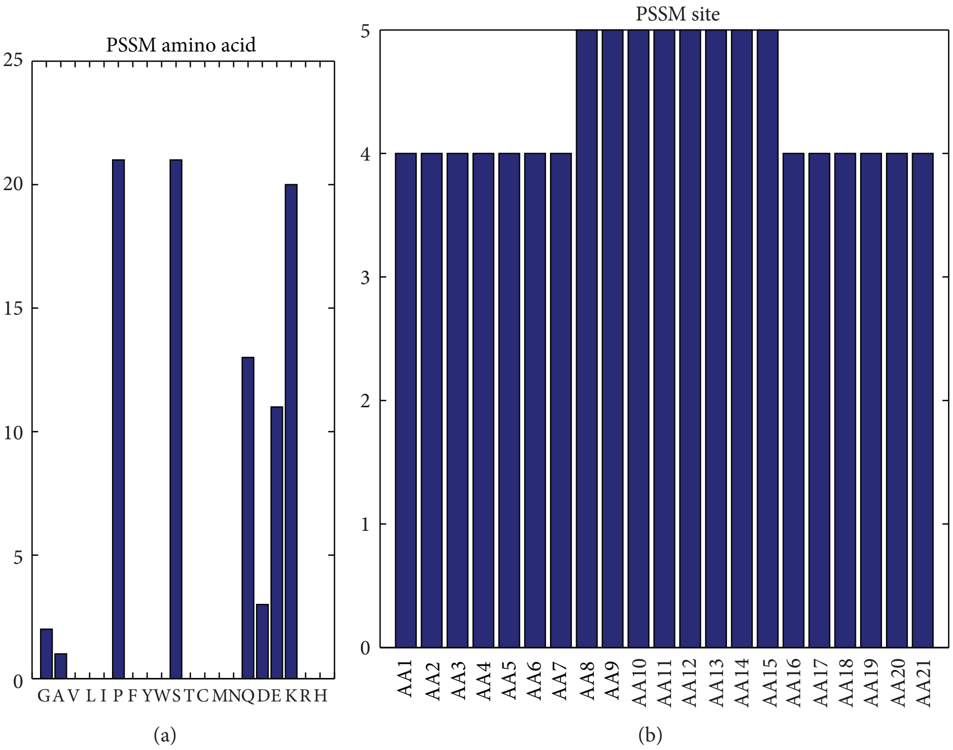

Figure 3: The distribution of amino acid compositions and sites on PSSM conservation feature. The histograms reveal the types and site distributions of PSSM features in the optimal feature set. (a) indicates the effects on prediction of mutations to 20 different amino acids. (b) provides the site distributions of the PSSM features in the optimal feature set.

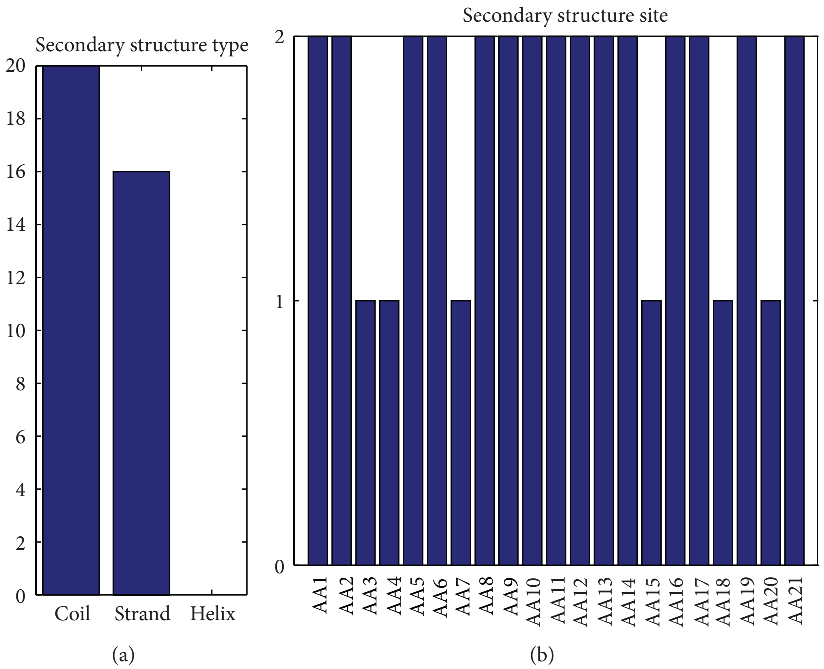

FIGURE 4: The distribution of secondary structure types and amino acid sites on secondary structure feature. The histograms give the types and site distributions of secondary structure features in the final optimal feature set. (a) indicates the effects on prediction of three different types of secondary structures: coil, strand, and helix. (b) provides the site distributions of the secondary structure features in the optimal feature set.

and "strand" did affect the disorder (Figure 4(a)). And the "coil" feature was affected the most, followed by the "strand" feature. The secondary structure features at 15 out of 21 sites had relatively more impact than the left 5 (Figure 4(b)). Intrinsically disordered and aggregation-prone domains exist within the very diverse set of human extracellular matrix protein [39]. Recently Evans reported that the aragonite extracellular matrix proteins (AECMPs) had evolved signature molecular traits of intrinsically disordered and aggregation-prone "interactive" sequences that enabled matrix assembly [40]. It was also reported that cyclization of the skeletal DHPR II-III loop affected the secondary structures and the dynamic properties of the helical A/B region as well as the critical C region. These structural effects were correlated with a change in vitro activation profile of the RyR1 and with an interaction with DHPR II-III loop α-helical recognition sites in the SPRY2 domain of RyR1 [41]. So it is believed the sequence location and number of intrinsically disordered and secondary motifs may be important for aggregation, protein orientation, and assembly stability and may also play a role in the recognition and interaction between proteins with other specific component(s).

3.7. Scan the Entire Protein Sequence to Refine the Disordered Region Prediction. The prediction result of the predictor constructed based on the 128 optimal features on the independent test was shown in Online Supporting Information S7. The third column, *predicted*, was the prediction result where "1" indicated the residue was in ordered region while "2" denoted the residue was in disordered region. It can be seen that many ordered (1) sites were wrong predicted as disordered (2), resulting in short disordered segments (2) being inserted in an ordered segment (1), and vice versa. Therefore, we used a scanning method to refine the prediction results according to the following criteria [42]. (i) Any predicted disordered sites (2) were refined to ordered (1) if there were more than 4 continuous "1s" upstream of the site but less than 4 continuous "2s" downstream of it. (ii) Any predicted ordered sites (1) were changed to disordered (2) if there were more than 4 continuous "2s" upstream of the site but less than 3 continuous "1s" downstream of it. After the refinement procedure, the performance improved much as shown in Table 1. The scanning results can be found in the last column, *scanning*, also can be found in Online Supporting Information S7.

3.8. Comparison with the Existing Methods. Our method was compared with three other existing methods, DISOPRED, DISOclust, and OnD-CRF. The DISOPRED server allows users to submit a protein sequence and returns a probability estimate of being disordered of each residue in the sequence [11]. In the prediction results by DISOPRED, disordered residues were marked with asterisks (∗) and ordered residues were marked with dots (·). The prediction results by DISOclust were formulated by a series of "D" and "O," denoting the residues being in disordered region and ordered region, respectively. The predicting result by Ond-CRF only delivers users the information of disordered regions. As a result, our method outperformed the other three existing methods. As shown in Table 1, the ACC and MCC were improved to

Table 1: The evaluation of prediction result on independent test set by different methods.

Method	Accuracy (ACC)	Matthews correlation coefficient (MCC)	Sensitivity (SN)	Specificity (SP)
Before scanning	0.7028	0.2791	0.7189	0.6281
After scanning	0.7508	0.3304	0.7806	0.6118
DISOPRED	0.7173	0.3285	0.7239	0.6864
DISOclust	0.6650	0.3105	0.6453	0.7570
OnD-CRF	0.6562	0.3228	0.6265	0.7941

a certain extent, at least 4% increase on ACC and almost 0.2% increase on MCC. It is suggested that our method is pretty more effective than other methods on prediction of intrinsically disordered protein region.

3.9. Useful Insights for Guiding Experiments or Being Validated by Experiments. About 50% of human proteins were previously predicted to contain at least one larger disordered region, and it was shown that the main reason for the existence of such regions was to harbor binding sites [43]. In this study, the selected features at different sites could provide insights for researchers to find or validate new disordered protein or disordered regions, as can be seen from the following two aspects. (i) PSSM feature: it was found from the results that the PSSM conservation score that mutates to amino acid P or S had the most impact. Besides, mutations to amino acids K, Q, and E also had more impacts than others. However mutations to other 12 amino acids were affected little. For example, phosphorylation of Ser66 in the intrinsically disordered N-terminal region of AtREM1.3 weakened the interaction strength with importin α proteins, indicating a regulatory domain in the N-terminal region stabilizing the interactions [44]. (ii) Secondary structure feature: it was found in our optimal feature set that the second structure feature at site 11 had the ranking index 1, implying the most important role to the prediction. Interestingly, it has been reported that disorder regions often correlated domain boundaries where usually harbor some coil structures [45]. Accordingly, other features in the optimal feature set are certainly worth being further investigated by future experiments.

4. Conclusion

The plasticity of disordered regions provides interaction capacity. In this study, we investigated important features for predicting disordered protein regions. As a result, the PSSM conservation scores and the second structures are two types of important features, which play key important roles in determining disordered regions. Among these, only 8 amino acids play major roles. The coil and strand structures also affected the prediction. These may provide additional insight into disordered proteins.

Authors' Contribution

T.-H. Zhao, M. Jiang and T. Huang contributed equally to this work.

Acknowledgments

This work was supported by Grants from the National Basic Research Program of China (2011CB510102, 2011CB510101), and the Innovation Program of Shanghai Municipal Education Commission (12ZZ087) and the grant of "The First-class Discipline of Universities in Shanghai."

References

[1] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure*, vol. 11, no. 11, pp. 1453–1459, 2003.

[2] A. K. Dunker, J. D. Lawson, C. J. Brown et al., "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26–59, 2001.

[3] K. Nishikawa, "Natively unfolded proteins: an overview," *Biophysics*, vol. 5, pp. 53–58, 2009.

[4] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm," *Journal of Molecular Biology*, vol. 293, no. 2, pp. 321–331, 1999.

[5] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 3, pp. 197–208, 2005.

[6] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Science*, vol. 11, no. 4, pp. 739–756, 2002.

[7] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradović, "Intrinsic disorder and protein function," *Biochemistry*, vol. 41, no. 21, pp. 6573–6582, 2002.

[8] M. Sickmeier, J. A. Hamilton, T. LeGall et al., "DisProt: the database of disordered proteins," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D786–D793, 2007.

[9] D. Eliezer, "Biophysical characterization of intrinsically disordered proteins," *Current Opinion in Structural Biology*, vol. 19, no. 1, pp. 23–30, 2009.

[10] E. A. Weathers, M. E. Paulaitis, T. B. Woolf, and J. H. Hoh, "Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein," *FEBS Letters*, vol. 576, no. 3, pp. 348–352, 2004.

[11] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder," *Bioinformatics*, vol. 20, no. 13, pp. 2138–2139, 2004.

[12] L. J. Mcguffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models," *Bioinformatics*, vol. 24, no. 16, pp. 1798–1804, 2008.

[13] L. Wang and U. H. Sauer, "OnD-CRF: predicting order and disorder in proteins conditional random fields," *Bioinformatics*, vol. 24, no. 11, pp. 1401–1402, 2008.

[14] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[15] U. Hobohm and C. Sander, "Enlarged representative set of protein structures," *Protein Science*, vol. 3, no. 3, pp. 522–524, 1994.

[16] T. Huang, Z. S. He, W. R. Cui et al., "A sequence-based approach for predicting protein disordered regions," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 243–248, 2013.

[17] S. Ahmad and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins," *BMC Bioinformatics*, vol. 6, no. 1, article 33, 2005.

[18] I. Radhakrishnan, G. C. Pérez-Alvarado, H. J. Dyson, and P. E. Wright, "Conformational preferences in the Ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of CREB," *FEBS Letters*, vol. 430, no. 3, pp. 317–322, 1998.

[19] M. K. Yoon, V. Venkatachalam, A. Huang, B. S. Choi, C. M. Stultz, and J. J. Chou, "Residual structure within the disordered C-terminal segment of p21 Waf1/Cip1/Sdi1 and its implications for molecular recognition," *Protein Science*, vol. 18, no. 2, pp. 337–347, 2009.

[20] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa, "Preformed structural elements feature in partner recognition by intrinsically unstructured proteins," *Journal of Molecular Biology*, vol. 338, no. 5, pp. 1015–1026, 2004.

[21] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins: Structure, Function and Genetics*, vol. 47, no. 2, pp. 228–235, 2002.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[23] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[24] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Using random forest for protein fold prediction problem: an empirical study," *Journal of Information Science and Engineering*, vol. 26, no. 6, pp. 1941–1956, 2010.

[25] K. K. Kandaswamy, K. C. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[26] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, no. 13, pp. 1616–1622, 2010.

[27] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[28] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[29] K. C. Chou and H. B. Shen, "Cell-PLoc: a package of web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, 2008.

[30] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[31] Z. He, J. Zhang, X. H. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS ONE*, vol. 5, no. 3, Article ID e9603, 2010.

[32] B. Q. Li, L. L. Hu, S. Niu, Y. D. Cai, and K. C. Chou, "Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches," *Journal of Proteomics*, vol. 75, no. 5, pp. 1654–1665, 2011.

[33] T. Huang, L. Chen, Y. D. Cai, and K. C. Chou, "Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property," *PloS ONE*, vol. 6, no. 9, Article ID e25297, 2011.

[34] H. Cramér, *Methods of Mathematical Statistics*, vol. 23, Princeton University Press, Princeton, NJ, USA, 1946.

[35] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics. Volume 2: Inference and: Relationship*, Griffin Printing, Santa Ana, Calif, USA, 1973.

[36] K. M. D. Harrison, T. Kajese, H. I. Hall, and R. Song, "Risk factor redistribution of the national HIV/AIDS surveillance data: an alternative approach," *Public Health Reports*, vol. 123, no. 5, pp. 618–627, 2008.

[37] S. Roy, D. Martinez, H. Platero, T. Lane, and M. Werner-Washburne, "Exploiting amino acid composition for predicting protein-protein interactions," *PLoS ONE*, vol. 4, no. 11, Article ID e7813, 2009.

[38] S. Ngo, V. Chiang, E. Ho, L. Le, and Z. Guo, "Prion domain of yeast Ure2 protein adopts a completely disordered structure: a solid-support EPR study," *PloS ONE*, vol. 7, no. 10, Article ID e47248, 2012.

[39] F. Peysselon, B. Xue, V. N. Uversky, and S. Ricard-Blum, "Intrinsic disorder of the extracellular matrix," *Molecular BioSystems*, vol. 7, no. 12, pp. 3353–3365, 2011.

[40] J. S. Evans, "Aragonite-associated biomineralization proteins are disordered and contain interactive motifs," *Bioinformatics*, vol. 28, no. 24, pp. 3182–3185, 2012.

[41] H. S. Tae, Y. Cui, Y. Karunasekara, P. G. Board, A. F. Dulhunty, and M. G. Casarotto, "Cyclization of the intrinsically disordered α1S dihydropyridine receptor II-III loop enhances secondary structure and in vitro function," *The Journal of Biological Chemistry*, vol. 286, no. 25, pp. 22589–22599, 2011.

[42] B. Q. Li, L. L. Hu, L. Chen, K. Y. Feng, Y. D. Cai, and K. C. Chou, "Prediction of protein domain with mRMR feature selection and analysis," *PloS One*, vol. 7, no. 6, Article ID e39308, 2012.

[43] B. Meszaros, I. Simon, and Z. Dosztanyi, "Prediction of protein binding regions in disordered proteins," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000376, 2009.

[44] M. Marin, V. Thallmair, and T. Ott, "The intrinsically disordered N-terminal region of AtREM1. 3 remorin protein mediates protein-protein interactions," *The Journal of Biological Chemistry*, vol. 287, no. 47, pp. 39982–39991, 2012.

[45] N. Nagarajan and G. Yona, "Automatic prediction of protein domains from sequence information using a hybrid learning system," *Bioinformatics*, vol. 20, no. 9, pp. 1335–1360, 2004.