Identifying functional site of disordered proteins

Chun Fang † Tamotsu Noguchi ‡ and Hayato Yamana †

† School of Fundamental Science and Engineering, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555, Japan

‡ Pharmaceutical Education Research Center, Meiji Pharmaceutical University, 2-522-1, Noshio, Kiyose, Tokyo, 204-8588, Japan

E-mail: † {fangchun, yamana}@yama.info.waseda.ac.jp, ‡ noguchit@my-pharm.ac.jp

Abstract

Molecular recognition features (MoRFs) has been proved to play critical role in molecular-interaction network of the cell, especially in eukaryotes. They act as molecular switches in the processes of cell signaling and regulation, and have relationship with the cause of many diseases. The importance of identifying MoRFs in disordered proteins is becoming increasingly apparent. However, due to unique structural and particular physiochemical properties of disordered proteins, traditional methods used for identifying functional sites of ordered proteins are ineffective for predicting MoRFs in disordered proteins. In this study, we propose a novel method, using a masking encoding scheme which extracts the relatively conservative information of residues for prediction. We compared our method with 9 other existing methods on a same test-dataset; the experimental results showed that, our method achieved the best performance, which got an AUC of 0.758. This study demonstrated that: 1) the flanking regions of MoRFs affected the plasticity of MoRFs; 2) MoRFs were flanked by less conserved residues; and 3) the masked and filtered PSSM was predictive features for identifying MoRFs.

Keyword MoRFs, disordered proteins, PSSM

1. Introduction

With breaking of the traditional concept on protein structure and function, the functional importance of disordered regions has become more and more apparent. MoRFs are short binding regions located in intrinsically disordered protein regions. They have no well-defined three-dimensional structure in natural state, but are easy to undergo a disorder-to-order transition upon binding to partner proteins. The particular dynamic conformation of MoRFs allows them to interact with multiple targets. [1]. MoRFs play critical role in various cellular functions, such as signaling and regulation; they act as molecular switches in molecular-interaction network of the cell, and assumed to have relationship with the causes of many diseases. Thus, identification of MoRFs is a key step to annotate protein functions and to find applications in drug design.

MoRFs have attracted the interest of many researchers. Norman E. D [2] analyzed the attributes of MoRFs and found that, several strong physicochemical preferences were shown in all MoRFs types compared to the disordered regions in general. Fuxreiter M., et al [3] demonstrated that both amino acid composition and charge/hydropathy properties of MoRFs exhibit a mixture characteristic of folded and disordered proteins. AHCHOR [4] applied biophysical principles to identify MoRFs.

Evolutionary information has been proved to be a predictive feature in identifying protein functional site, because in order to maintain certain function, the functional sites of proteins must maintain a high degree of conservation. MoRFs have also been found to be more conserved than surrounding residues; however, disordered proteins evolve rapidly compared to ordered proteins. The standard PSSM which incorporates the conservation information of proteins is ineffective when used directly. Relative local conservation has been proved to be a good feature for motifs discovery in disordered protein regions, and has been used by researches [5-7].

In our previous study [8], we found that there are some redundant features in standard PSSMs, and

ignoring some noise features could improve the functional site prediction for ordered proteins. In order to evaluate whether it is also suitable for disordered proteins, in this study, we adopt a masking and filtering encoding scheme to develop the MoRFs predictor. This method could strengthen the high relative local conservative information while filtering out the low relative local conservative scores in PSSMs. A traditional PSSM-based method was also developed for a comparison. Both of the models used the support vector machines (SVM).

2. Research method

2.1 Data Sets

We prepared two datasets, training-dataset and test-dataset, both of them were extracted from the datasets used in the research [1].

Training-datasets:

The research [1] used 421 MoRFs-contained chains for training and 419 MoRFs-contained chains for test. Since we found that some sequences among the 840 chains are sharing a similarity higher than 40%, we used CD-HIT [9] to cluster them, and the chains with similarity >40% were discarded. After removed, 447 chains that contain 5,601positive samples (MoRFs) and 262,732 negative samples (non-MoRFs) were remained. All the positive samples, and the same amount of negative samples selected randomly from the 262,732 non-Morphs were used to construct the training dataset for our experiment.

Test-datasets:

We used the same test dataset with the research of MoRFPred [1], named as TEST2012, which includes 45 MoRFs-contained chains deposited in PDB from January 1 to March 11, 2012, and also include those of UniProtKB released from February 22, 2012.

2.2 Length distribution of MoRFs

The length distributions of MoRFs in the 447 proteins were analyzed. As shown in Figure 1, most of them had length between 5 to25 residues.

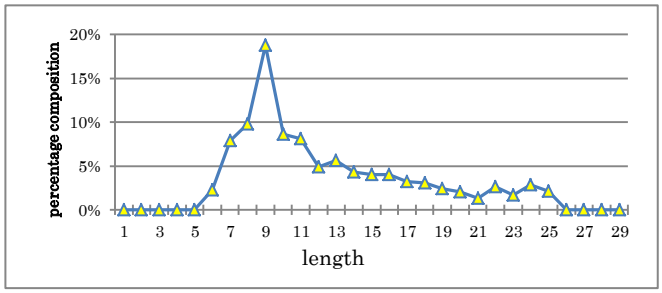

Figure1. Length distribution of MoRFs

2.3 Composition analysis

We calculated the composition of the 447 proteins. The sequences are divided into 3 regions: the MoRFs, the flanking regions of MoRFs within 5 residues, and the non-MoRFs in the general disordered region. The composition percentage of the 20 amino acid residues is shown in Figure 2 (a). Figures 2 (a) demonstrates that Ile, Leu, Phe, Tyr, Lys, Arg residues are overrepresented in MoRFs, most of them are hydrophobic amino acids. While amino acids Ala, Gly, Lys, Ser, Pro are overrepresented in flanking-MoRFs regions, most of them are small and tiny amino acids.

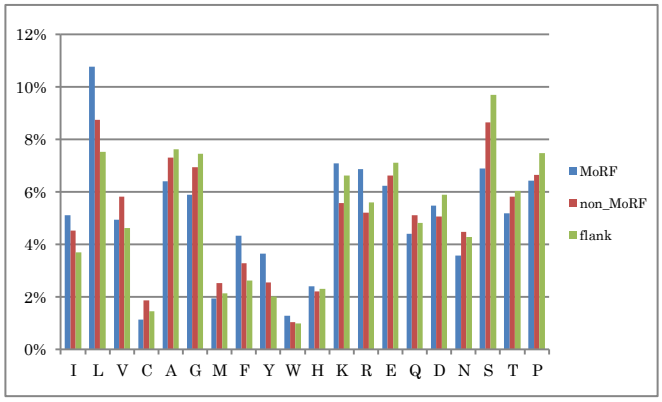

Figure 2(a). Composition distribution of the 20 amino acids for the 3 kinds of regions

2.4 Physicochemical properties preference

We also calculated the physicochemical properties propensity for each kind of regions. 10 discriminative physicochemical features of residue were considered in our study. They were hydrophobic, polar, small, proline, tiny, aliphatic, aromatic, positive, negative, and charged. The distribution of 10 physicochemical properties is shown in Figure 2 (b). Figure 2 (b) demonstrates that physicochemical characteristics propensity of flanking regions seems to vary widely compared to the flanking regions and general disordered regions. Properties such as polar,

small, tiny and charged are over-represented in flanking regions.

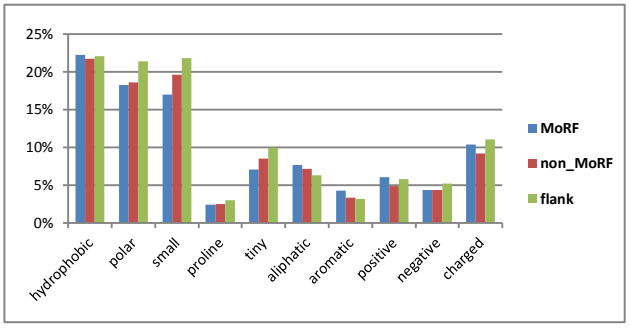

Figure 2(b). 10 physicochemical characteristics propensity for the 3 regions

2.5 Prediction model

Based on the above analysis, we considered to use a masked and filtered PSSM, which incorporates the information of amino acid position, relative evolutionary information, and dependency on neighboring residues, to design our prediction model. The prediction model is shown in Figure 3.

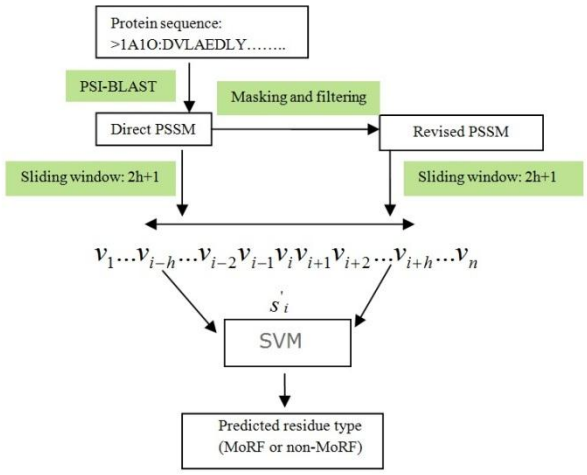

Figure 3. Prediction model. The length of sliding window is represented by 2h+1; n is the length of sequence.

2.6 Evolutionary information (PSSM)

Evolutionary information were obtained from PSSMs, which were generated by PSI-BLAST [10], searching against NCBI non-redundant (nr) database [11] by three times iteration with an e-value of 0.001. Evolutionary information for each amino acid was encapsulated in a vector of 20 dimensions; the size of PSSM of a protein with N residues is 20 × N. 20 dimensions were considered as a standard amino acid size. N is the length of a protein.

2.7 Masking and filtering the PSSM

The masked PSSM is used to describe the relative evolutionary information of each residue in a protein; it was converted from a standard PSSM according to the formula (1).

Firstly, a masking sliding window with appropriate size was used to calculate the mean conservation score for each residue in a standard PSSM, and then the relative conservation scores were calculated. After that, in order to strengthen the high conservative scores while filtering out the low conservative scores, the positions below a mean conservation score were sited to 0 according to the formula (2).

$$Masking\_C_i = C_i - \frac{1}{2n+1}\sum_{j=i-n}^{j=i+n} C_j \qquad (1)$$

$$Filtering\_C_i = \begin{cases} Masking\_C_i, (Masking\_C_i > 0) \\ 0, (Masking\_C_i \le 0) \end{cases} \qquad (2)$$

$Masking\_C_i$ is the mean conservation score of residue i, C_i is the standard conservation score in PSSM, $2*n+1$ is the masking window size.

2.8 SVM

Prediction of MoRFs can be addressed as a two-classification problem; determining whether a given residue belong to MoRFs or not. Our prediction model was trained by the LIBSVM software package which was written by in Chih-Jen [12-13]. The Radial Basis Function (RBF kernel) was adopted to construct the SVM classifiers. In order to obtain the optimal sliding window size, the standard PSSM based model was analyzed as an example. Results of the success rate according to different sliding window sizes were listed in Figure 4. The performance tends to be stable from the window sizes 21. Thus, we chose the relatively best size 27 as the sliding window size for our models.

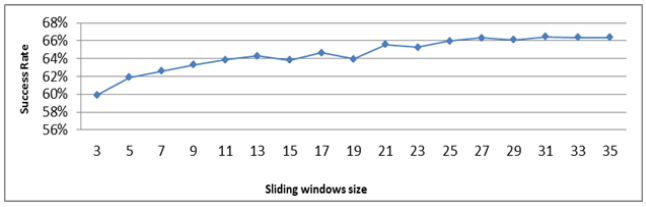

Figure 4. Success rate of the standard PSSM based model at different sliding window sizes

3. Evaluation criteria

The area under the corresponding receiver operating characteristic (ROC) curve (AUC) was adopted to evaluate the performance of the classifiers. the ROC plots with the AUC values were created by using the R statistical package [14]. The accuracy, true positive rate (TPR), false negative rate (FPR) are defined as follows:

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$FPR = \frac{FP}{TN + FP} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Where TP, TN, FP and FN represents true positive, true negative, false positive and false negative respectively.

4. Results and discussion

4.1 Performance comparison with SVM-PSSM

We used the test dataset TEST2012 to evaluate the performance of our developed models. The standard PSSM based model (SVM-PSSM) was developed for a comparison. Both the MFPSSMPred model and the SVM-PSSM model were used a same training set, with a same sliding window size of 27. Since MFPSSMPred method requires a particular masking window size, we compared the ROCs calculated with different masking window size, as shown in Figure 5. Among them, the mode with a masking size of 15 got the best ROC (yellow ROC). Thus, 15 was chosen as the masking window size for the MFPSSMPred. The detail ROC plots of the MFPSSMPred and SVM-PSSM are shown in Figure 6 and 7 respectively. Figure 6 demonstrates that the AUC of MFPSSMPred is 0.7578 which is higher than the AUC of SVM-PSSM (0.749) showed in Figure 7.

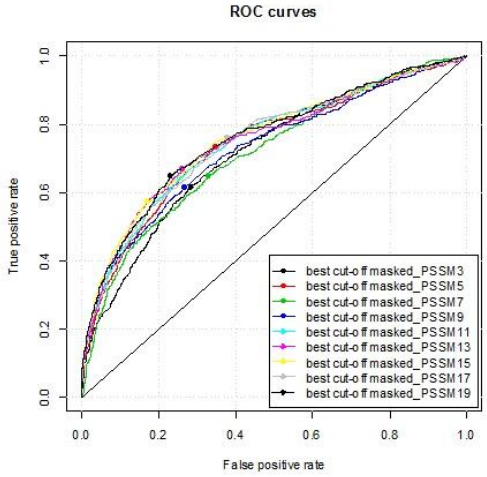

Figure 5. ROC Plots of the MFPSSMPred at different masking window size

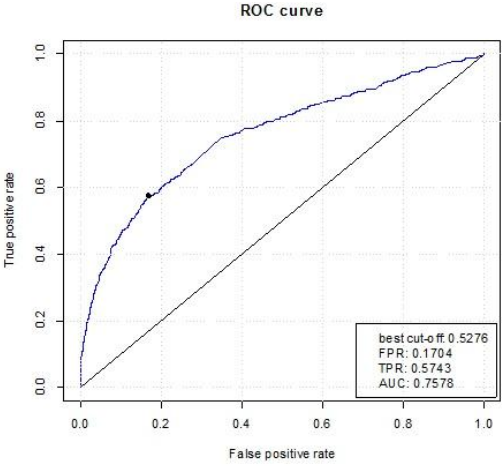

Figure 6. ROC Plot of the MFPSSMPred model at masking window size 15.

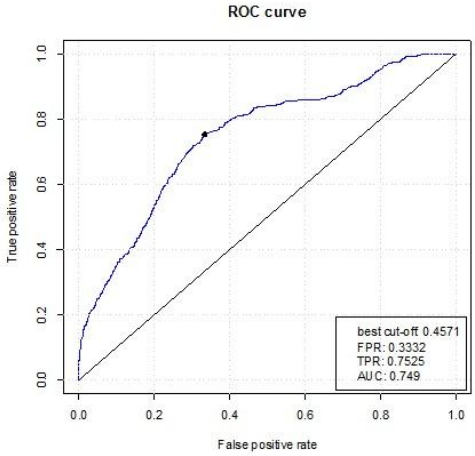

Figure 7. ROC Plot of the SVM-PSSM model

For facilitating comparison, the two ROC plots are also integrated into Figure 8. From Figure 8, we can see that, the ROC shape of the MFPSSMPred is more perfect besides its bigger area under the ROC curve

compared to the ROC curve of SVM-PSSM.

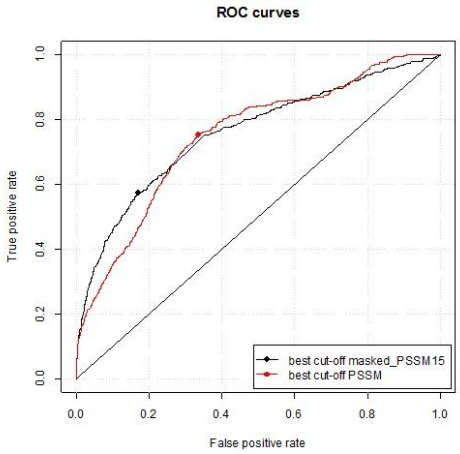

Figure 8. Integrated ROC Plots of the SVM-PSSM model and MFPSSMPred model

Details of the accuracy, TPR, FPR of the FPSSMPred according to different evaluation thresholds were shown in table 1.

Table 1. Accuracy, TPR and FPR of the MFPSSMPred at different threshold

Threshold	Accuracy	TPR	FPR
0	0.949	0.000	0.000
0.1	0.949	0.000	0.000
0.2	0.850	0.323	0.130
0.3	0.766	0.435	0.226
0.4	0.667	0.530	0.334
0.5	0.565	0.644	0.446
0.6	0.465	0.720	0.555
0.7	0.358	0.804	0.672
0.8	0.247	0.900	0.790
0.9	0.127	0.962	0.920
1	0.051	1.000	1.000

4.2 Performance comparison with 9 existing methods

There exist some tools for MoRFs predication. Here we also list them out for a comparison. The results of other classifiers are quoted from research [1]. All the methods were tested on the TEST2012 dataset. Results are shown in Table 2. Table 2 demonstrates that our MFPSSMPred predictor achieves the best AUC.

Table 2. Performance of MFPSSMPred and 9 other predictors on dataset TEST2012

Methods	ACC	TPR	FPR	AUC
MFPSSMPred (our method)	0.897	0.267	0.078	**0.758**
MoRFpred (Fatemeh M.D,2012)	0.943	0.236	0.045	**0.697**
MD (Schlessinger et al., 2009)	0.565	0.613	0.436	**0.679**
ANCHOR (Dosztányi et al., 2009)	0.759	0.433	0.236	**0.638**
IUPredS (Dosztányi et al., 2005)	0.708	0.449	0.287	**0.634**
IUPredL (Dosztányi et al., 2005)	0.618	0.572	0.382	**0.62**
MFDp (Mizianty et al., 2010)	0.45	0.752	0.556	**0.62**
Spine-D (Faraggu et al., 2009)	0.482	0.72	0.522	**0.605**
DISOPRED2 (Ward et al., 2004)	0.545	0.543	0.455	**0.548**
DISOclust (McGuffin,2008)	0.411	0.653	0.593	**0.512**

4.3 Performance on unbalanced training samples

There are 5,601positive samples and 262,732 negative samples in our dataset; the ratio between them was 1:46.9, in order to analyze whether this imbalance bias the prediction method, we also developed the training model with a 2:1 ratio between the non-MoRF and Morph residues, that is, 5,601 MoRFs with 112,02 non-Morphs residues. Results tested on the TEST2012 are shown in Figure 9. Figure 9 demonstrates that, the difference between the results trained on 1:1 ratio and 2:1 ratio is not very obvious.

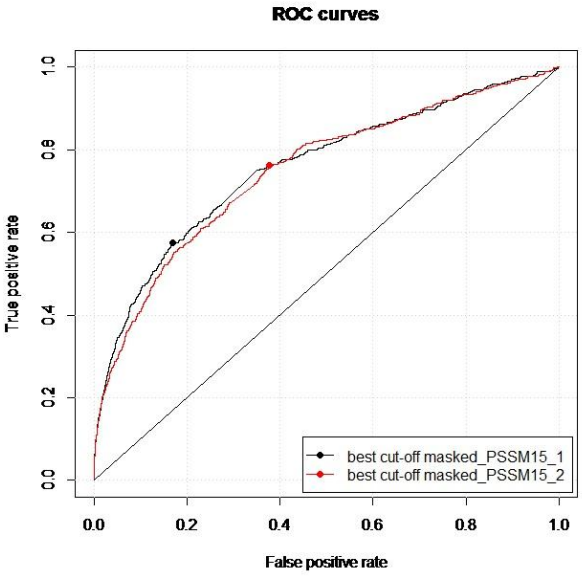

Figure 9. ROC Plots of MFPSSMPred based on unbalanced training samples

5. Conclusions

In this paper, we proposed a Masked and filtered PSSM based method, which used the masking and filtering skills to extract the relative high evolutionary information, while filtering out the low

conservative information of residues for MoRFs prediction. The traditional PSSM based method and 9 other existing methods were analyzed for comparisons. We tested all the methods on the same dataset TEST2012. Experimental results showed that, when comparing with the traditional PSSM based method, the MFPSSMPred method performed better than the SVM-PSSM method, which not only obtained a bigger AUC but also got a more perfect ROC curved shape. When comparing with the 9 other existing methods, our method demonstrated the best AUC of 0.758, which was 0.109~0.246 higher than others.

In summary, this study demonstrated that the masked and filtered PSSM, which incorporated relative evolutionary information and filtered the noise features of residues, was predictive for identifying MoRFs. It also revealed some hallmarks of MoRFs: the flanking regions of MoRFs affected the plasticity of MoRFs; MoRFs were flanked by less conserved residues. Though the performance of our method is still not satisfactory due to the complexity of disorder proteins, we shall try our best to further improve the performance of our model.

References

[1] Fatemeh M.D Wei-Lun H, Marcin J.M, Christopher J.O, Bin X, A. Keith D, Vladimir N.U, Lukasz K. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins, Bioinformatics (2012) 28 (12): i75-i83.

[2] Norman E. D, Kim V. R, Robert J. W. Attributes of short linear motifs. Molecular BioSystems. 2012, 8, 268-281.

[3] Monika F, Peter T and Istva´n S. Local structural disorder imparts plasticity on linear motifs. Bioinformatics (2007) 23 (8): 950-956.

[4] Dosztanyi Z, Mészáros, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics (2009) 25 (20): 2745-2746.

[5] Norman E. D, Denis C. S and Richard J. E. Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics (2009) 25 (4): 443-450.

[6] Norman E. Davey1, Joanne L. C, Denis C. S, Toby J. G. Mark J. C. and Richard J. E. SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Research, 2012, 1–14

[7] Niall J Haslam and Denis C Shields. Profile-based short linear protein motif discovery, BMC Bioinformatics 2012, 13:104

[8] Chun F, Tamotsu N, and Hayato Y: Using physicochemical features to condense the position-specific scoring matrix for flavin adenine dinucleotide-binding prediction. BioDataMing, in assessing.

[9] Weizhong Li, Adam G. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (2006) 22 (13): 1658-1659.

[10] Stephen F.A, Thomas L. M, Alejandro A. S, Jinghui Z, Zheng Z. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997, 25 (17): 3389-3402.

[11] NR [ftp://ftp.ncbi.nih.gov/blast/db/fasta/nr.gz].

[12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011, 2:27:1-27:27

[13] A library for support vector machines: [http://www.csie.ntu.edu.tw/~cjlin/libsvm/]

[14] http://www.r-project.org/