Sequence-Based Prediction of Molecular Recognition Features in Disordered Proteins

Chun Fang and Hayato Yamana
Department of Computer Science and Engineering of Waseda University, Tokyo, Japan
Email: {fangchun, yamana}@yama.info.waseda.ac.jp

Tamotsu Noguchi
Pharmaceutical Education Research Center, Meiji Pharmaceutical University, Tokyo, Japan
Email: noguchi-tamotsu@aist.go.jp

***Abstract*—Molecular recognition features (MoRFs) act as molecular switches in molecular-interaction network of the cell, and assumed to have relationship with the causes of many diseases. The importance of identifying MoRFs in disordered proteins is becoming increasingly apparent. So far, only a limited number of experimentally validated MoRFs is known, and there are few specialized tools for identifying MoRFs. Existing methods used many predicted results, such as predicted disorder probabilities, solvent accessibility and B-factors as features for prediction, or used MoRFs database directly for alignment to assist the prediction; however, their design are complex, and the performance is also affected largely by other predictors. In this study, we proposed a novel method, named as MFPSSMPred (Masked and Filtered PSSM based Prediction), which adopts a masking method to extract high local conservative features, and a filtering method to filter out low local conservative scores in position-specific scoring matrix (PSSMs) for prediction. All features are extracted from the sequences only. We compared our method with a traditional PSSM-based method and 9 other existed methods on a same test dataset. The experimental results showed that, our method achieved the best performance with 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 revised PSSM was predictive features for identifying MoRFs.**

***Index Terms*—MoRFs prediction, disordered proteins, PSSM**

I. 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 [1]. The particular dynamic conformation of MoRFs allows them to interact with multiple targets. 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 ignoring some redundant features in standard PSSMs 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).

II. Research Method

A. Data Sets

Manuscript received January 13, 2013, revised March 20, 2013.

©2013 Engineering and Technology Publishing
doi: 10.12720/jomb.2.2.110-114

We prepared two datasets, the training dataset and the test dataset; both of them were extracted from the datasets used in the research of MoRFpred [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 dataset: We adopted the dataset TEST2012, which was a test dataset in the research of MoRFPred [1], to test our developed model. TEST2012 includes 45 MoRFs-contained chains which deposited in PDB from January 1 to March 11, 2012, and also includes those of UniProtKB released from February 22, 2012.

B. 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 Fig. 1. Fig. 1 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 1. Composition distribution of the 3 regions

C. 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 Fig. 2. Fig. 2 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.

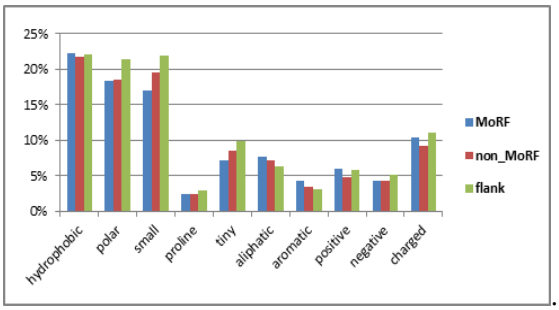

Figure 2. 10 physicochemical characteristics propensity.

D. 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 Fig. 3.

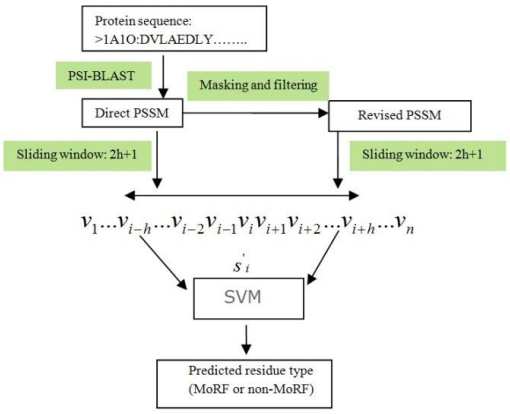

Figure 3. Prediction model. The length of sliding window is represented by 2h+1; *n* is the length of sequence.

E. 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.

F. 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

©2013 Engineering and Technology Publishing

scores while filtering out the low conservative scores, the positions below a mean conservation score were sited to 0 according to the formula (2).

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

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

$Masking\_C_i$ 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.

G. 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]. 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 Fig. 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.

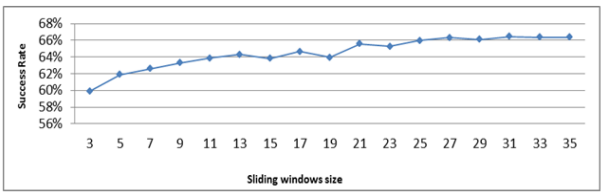

Figure 4. Success rate with different sliding window sizes.

III. 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. The true positive rate (TPR) and false negative rate (FPR) are defined as follows:

$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{TN+FP} \quad (3)$$

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

IV. Results and Discussion

A. 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 Fig. 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 Fig. 6 and 7 respectively. Fig. 6 demonstrates that the AUC of MFPSSMPred is 0.7578 which is higher than the AUC of SVM-PSSM (0.749) showed in Fig. 7.

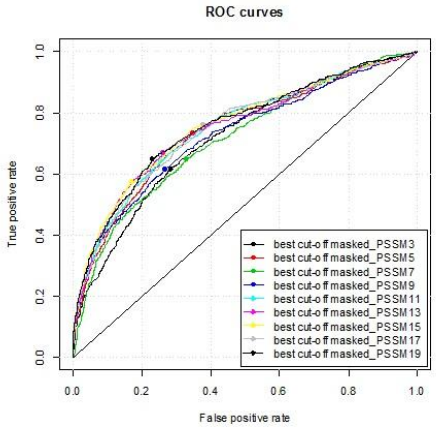

Figure 5. ROC Plots of the MFPSSMPred at different masking window size

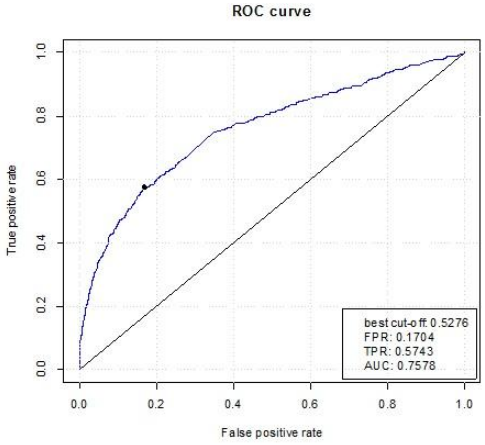

Figure 6. The best ROC Plot of the MFPSSMPred model.

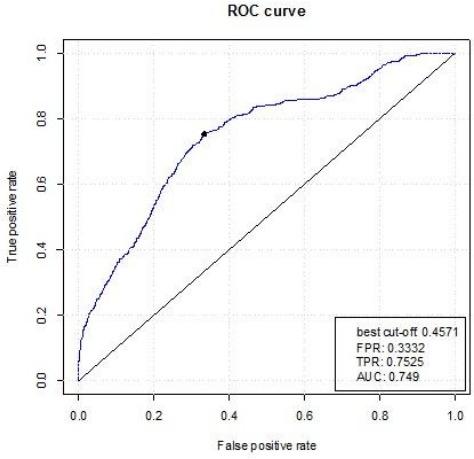

Figure 7. ROC Plot of the SVM-PSSM model

©2013 Engineering and Technology Publishing

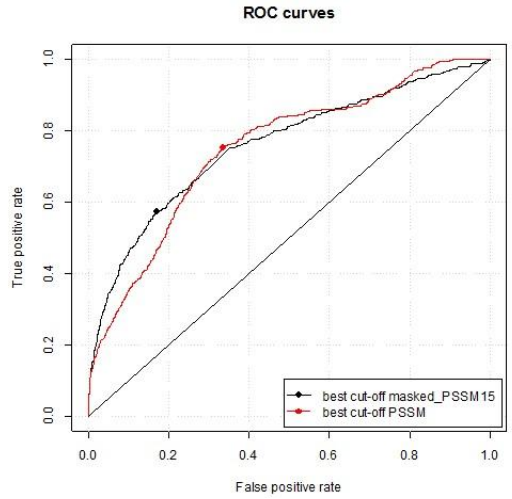

Figure 8. Integrated ROC Plots of the SVM-PSSM model and MFPSSMPred model

For facilitating comparison, the two ROC plots are also integrated into Fig. 8. Fig. 8 demonstrates that, besides its bigger area under the ROC curve, the ROC shape of the MFPSSMPred is more perfect than the one of SVM-PSSM.

B. 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 I. Table I demonstrates that our MFPSSMPred predictor achieves the best AUC.

TABLE I. PERFORMANCE OF THE MFPSSMPRED AT DIFFERENT THRESHOLD

Methods	TPR	FPR	AUC
MFPSSMPred (our method)	0.574	0.170	**0.758**
MoRFpred (Fatemeh M.D,2012)	0.236	0.045	**0.697**
MD (Schlessinger et al., 2009)	0.613	0.436	**0.679**
ANCHOR (Dosztányi et al., 2009)	0.433	0.236	**0.638**
IUPredS (Dosztányi et al., 2005)	0.449	0.287	**0.634**
IUPredL (Dosztányi et al., 2005)	0.572	0.382	**0.62**
MFDp (Mizianty et al., 2010)	0.752	0.556	**0.62**
Spine-D (Faraggu et al., 2009)	0.72	0.522	**0.605**
DISOPRED2 (Ward et al., 2004)	0.543	0.455	**0.548**
DISOclust (McGuffin,2008)	0.653	0.593	**0.512**

C. 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 Fig. 9. Fig. 9 demonstrates that, the difference between the results trained on 1:1 ratio and 2:1 ratio is not very obvious.

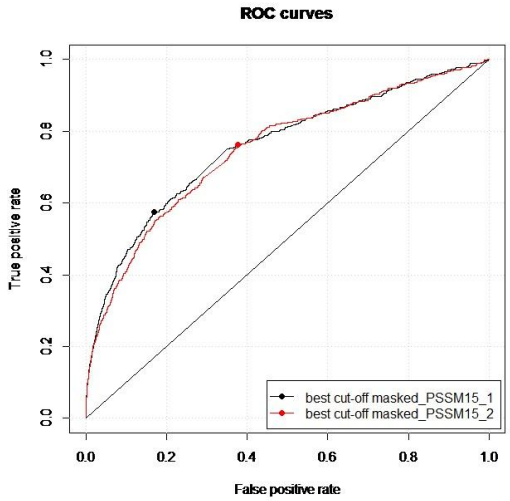

Figure 9. ROC Plots of MFPSSMPred based on unbalanced training samples

V. CONCLUSION

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.

ACKNOWLEDGMENT

We would like to thank Prof. TOMINAGA of Computational Biological Research Center (CBRC) providing a lot of convenience for our study. We also thank an anonymous reviewer for his/her helpful comments, which improved the manuscript.

REFERENCES

[1] M. D. Fatemeh, H. Wei-Lun, J. M. Marcin, *et al*., "MoRFpred, a computational tool for sequence-based prediction and

©2013 Engineering and Technology Publishing

characterization of short disorder-to-order transitioning binding regions in proteins," *Bioinformatics*, vol. 28, no. 12, pp. 75-83, Dec 2012.
[2] E. D. Norman, V. R. Kim, and J. W. Robert, "Attributes of short linear motifs," *Molecular Bio Systems*, vol. 8, pp. 268-281, Jan 2012.
[3] F. Monika, T. Peter and S. Istva n, "Local structural disorder imparts plasticity on linear motifs," *Bioinformatics*, vol. 23, no. 8, pp. 950-956, Aug 2007.
[4] Z. Dosztanyi, M ész áros, and I. Simon, "ANCHOR: Web server for predicting protein binding regions in disordered proteins," *Bioinformatics*, vol. 25, no. 20, pp. 2745-2746, Aug 2009.
[5] E. D. Norman, C. S. Denis and J. E. Richard, "Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery," *Bioinformatics*, vol. 25, no. 4, pp. 443-450, Jan 2009.
[6] E. D. Norman, L. C. Joanne, C. S. Denis, J. G. Toby, J. C. Mark and J. E.Richard, "SLiMPrints: Conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions," *Nucleic Acids Research*, vol. 40, no. 21, pp. 1-14, Nov 2012.
[7] N. J. Haslam and C. S. Denis, "Profile-based short linear protein motif discovery," *BMC Bioinformatics*, vol. 13, pp. 104-113, May 2012.
[8] F. Chun, N. Tamotsu, and Y. Hayato, "Using physicochemical features to condense the position-specific scoring matrix for flavin adenine dinucleotide-binding prediction," *International Journal of Data Mining and Bioinformatics*, in assessing.
[9] W. Li, G. Adam, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, May 2006.
[10] F. A. Stephen, L. M. Thomas, A. S. Alejandro, Z. Jinghui, and Z. Zheng, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389-3402, Sep 1997.
[11] NR. [Online]. Available: ftp://ftp.ncbi.nih.gov/blast/db/fasta/nr.gz
[12] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 1-27, April 2011.

Chun Fang is a Ph.D. student at Waseda University in Japan. She received her M.S. degree in Computer Science from Huazhong Normal University, Wuhan, China, in 2009. Her research interests include data mining, artificial intelligence and bioinformatics.

Tamotsu Noguchi is a principal research scientist of the Computational Biology Research Center (CBRC), AIST in Japan, from 2001. He is also a visiting professor of Research Institute of IT Biology and Mining of Waseda University from 2005. He received his Doctor of Engineering degree at Osaka University, in 2001. His research interests include prediction of protein functions, prediction of secondary and tertiary structures in proteins, mechanism of protein folding and disorder regions in a protein.

Hayato Yamana is a member of the IEEE and the IEEE Computer Society.He received his Doctor of Engineering degree at Waseda University in 1993. He began his career at the Electro technical Laboratory (ETL) of the former Ministry of International Trade and Industry (MITI), and was seconded to MITI's Machinery and Information Industries Bureau for a year in 1996. He was subsequently appointed Associate Professor of Computer Science at Waseda University in 2000, and has been a professor in that department (as well as visiting professor at the National Institute of Informatics) since 2005. Professor Yamana has received a number of awards, including the IPSJ (Information Processing Society of Japan) Yamashita Memorial Research Award in 1995, the IPSJ Best Author Award in 2002 and the ITE (Institute of Image Information and Television Engineers) Best Author Award in 2003. He has written, co-written and translated a number of books including Google Hacks (translation supervisor), Google Pocket Guide (translation), Com Series: An Introduction to Super Parallel Computers (co-author), How to Search the World Wide Web-A Guide to Search Engines (co-author) and Objects that Evolve the Internet (author)

©2013 Engineering and Technology Publishing