

A Two-Stage RBFN Classifier for Protein Disorder Prediction

Chung-Tsai Su¹, Chien-Yu Chen²

¹Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C., and ²Department of Bio-industrial Mechatronics Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C.
sbb@mars.csie.ntu.edu.tw; cychen@mars.csie.ntu.edu.tw

Abstract. More and more proteins have been observed to display functions through intrinsic disorder. Such structurally flexible regions are shown to play important roles in biological processes and are estimated to be abundant in eukaryotic proteomes. In our recent work DisPSSMP, it is demonstrated that the accuracy of protein disorder prediction is greatly improved if the disorder propensity of amino acids is considered when generating the condensed PSSM features. This work aims to present a two-stage classifier of Radial Basis Function Networks (RBFN) to further enhance the predicting power of DisPSSMP. In this study, an alternative determining function is adopted when delivering predictions based on the function values generated by the RBFN classifier. The experimental results reveal that the proposed two-stage mechanism is of benefit to this problem.

1 Introduction

Intrinsically disordered proteins or protein regions exhibit unstable and changeable three dimensional structures under physiological conditions [1]. Although lacking fixed structures, many unfolded disordered proteins or partial protein regions have been identified to participate in many biological processes and carry out important biological functions [1, 2]. Also, it has been observed that the absence of a rigid structure allows disordered binding regions to interact with several different targets [3]. Therefore, the automated prediction of disordered regions is a necessary preliminary procedure in high-throughput methods for understanding protein function.

Many studies have demonstrated that the disordered regions can be detected by examining the amino acid sequences [3]. Disordered regions are distinguished from ordered regions by its low sequence complexity, amino acid compositional bias, or high flexibility. In our recent work DisPSSMP [4], we investigated a condensed position specific scoring matrix with respect to physicochemical properties (PSSMP) on the prediction accuracy. Additionally, DisPSSMP decomposes each conventional physicochemical property of amino acids into two disjoint groups which have a propensity for order and disorder respectively.

Ward *et al.* showed in their paper that the accuracy of disorder prediction can be improved by employing a smoothing classifier after the first-level SVM classifier [5]. Thus,

in this study, we hope to investigate how the predicting power of DisPSSMP can be enhanced by a two-stage classification architecture. When compared with the original DisPSSMP, the experimental results reveal that the two-stage RBFN classifier indeed outperforms the single-stage classifier and is considered useful to the problem of protein disorder prediction.

2 Material and Method

2.1 Datasets

For training and validation processes, six datasets have been extracted from different databases in our recent report [6]. These blind testing sets serve as a platform for comparing the performance of our proposed approach and DisPSSMP [4].

2.2 Classifier

DisPSSMP adopts the QuickRBF package [7] in constructing Radial Basis Function Networks (RBFN) for classification. To handle the problem of skewed datasets, DisPSSMP takes equal quantity of residues from ordered and disordered regions in constructing the classifier. In this regard, a large volume of ordered regions was lost after the random removal of unwanted ordered residues. Thus in this study, all of ordered and disordered residues in the training datasets are included to reduce the loss of information. In order to not over-predict residues as ordered, we adopt an alternative function in determining the outputs based on the function values generated by the RBF network. Let R_i be the feature vector of a residue in the data sets, $\text{Disorder}(R_i)$ and $\text{Order}(R_i)$ are output values of the RBF networks. Then, the formula for calculating the disorder propensity of residue R_i is represented as follows:

$$\text{Propensity}_D(R_i) = (\text{Disorder}(R_i) - \text{Order}(R_i) + 1)/2. \quad (1)$$

Next, the alternative classification function $\text{Classifier}(R_i)$ is shown in Equation (2). It means that the residue R_i is predicted as disordered if $(\text{Propensity}_D(R_i) \geq \text{Threshold})$, where the parameter Threshold is determined by conducting cross-validation procedure on the training dataset.

$$\text{Classifier}(R_i) = \begin{cases} 1 & \text{if } (\text{Propensity}_D(R_i) \geq \text{Threshold}); \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

2.3 Two-stage classification architecture

Fig. 1 shows the procedure of generating feature vectors from a protein sequence into FS-PSSMP-4 introduced in [4]. The first classifier outputs $\text{Disorder}(R_i)$ and $\text{Order}(R_i)$, and the first-level prediction is determined by equation (2) with Threshold_1 . After that, the feature set of the second stage is generated based on $\text{Disorder}(R_i)$ and $\text{Order}(R_i)$ with a sliding window, as what we did in generating FS-PSSMP-4. Finally, the prediction decision is made by equation (2) with Threshold_2 . In this paper, the classifier of the single-stage is denoted as ONE-STAGE, and the classifier including both stages is denoted as TWO-STAGE.

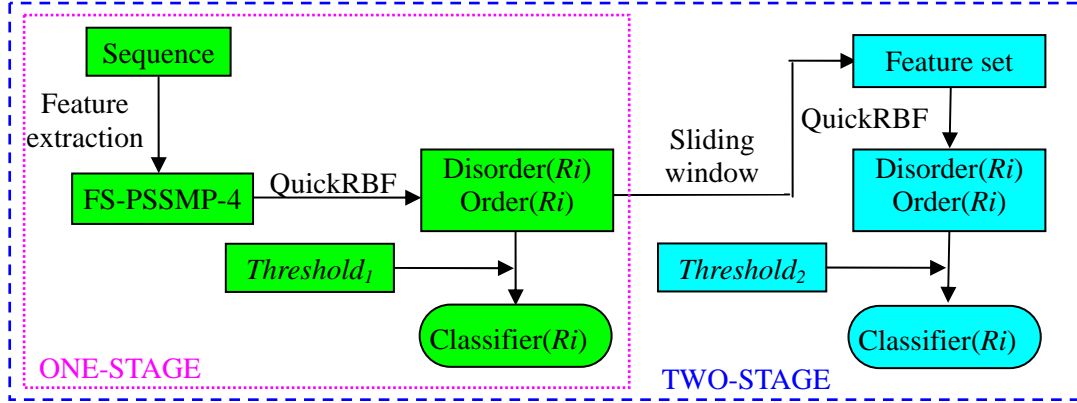


Fig. 1. The architecture of the proposed two-stage classifier.

3 Results and Discussions

In this section, the window size of feature set and the parameter *Threshold* are determined by conducting cross validation. To evaluate the proposed approach, the performance of ONE-STAGE and TWO-STAGE will be compared with DisPSSMP. Then, the improvement delivered by TWO-STAGE is analyzed and discussed.

3.1 Cross-validation

In order to conduct a five-fold cross validation, the chains in the training sets PDB652, D184, and G200 are randomly split into five subsets of approximately equal residues. To determine the window size of the feature sets, a range of window size from 11 to 59 has been evaluated for both ONE-STAGE and TWO-STAGE. For both cases, the results show that the performance of the classifier can not be explicitly improved when the window size is larger than 31. Thus, the window size of 47 is selected to be consistent with DisPSSMP. Next, the parameter *Threshold* in Equation (2) is determined when the probability excess is maximal. For ONE-STAGE and TWO-STAGE, $Threshold_1$ and $Threshold_2$ are 0.22 and 0.25, respectively.

3.2 Results on testing data

When compared with DisPSSMP, both ONE-STAGE and TWO-STAGE improve the performance of protein disorder prediction on all the testing data. The definition of the evaluating measures employed can be found at [6]. As shown in Table I, the number of false positives in DisPSSMP is largely reduced by ONE-STAGE and TWO-STAGE, which thanks to the fully ordered dataset G200 included in the training sets. For this dataset, all of ordered residues are adopted as the training instances. Furthermore, the number of false positive in TWO-STAGE is much less than in ONE-STAGE, which is considered as the contribution of the refinement by the second stage. With the measures of accuracy, Matthews' correlation coefficient (MCC), probability excess (Prob. Excess), and the area under the ROC curve (AUC), TWO-STAGE strikingly has the best performance in the testing set R80.

Table I. Results on the testing data R80

	TP	FP	TN	FN	Sens.	Spec.	Prec.	Accu.	MCC	Prob. Excess	AUC
DisPSSMP	2800	4550	25359	849	0.767	0.848	0.381	0.839	0.463	0.615	0.884
ONE-STAGE	2763	4057	25852	886	0.757	0.864	0.405	0.853	0.481	0.622	0.892
TWO-STAGE	2731	3084	26825	918	0.748	0.897	0.470	0.881	0.531	0.645	0.896

Table II. Results on the testing data U79 and P80

	TP	FP	TN	FN	Sens.	Spec.	Prec.	Accu.	MCC	Prob. Excess	AUC
DisPSSMP	11934	3896	12672	2528	0.825	0.765	0.754	0.793	0.589	0.590	0.869
ONE-STAGE	11969	2919	13649	2493	0.828	0.824	0.804	0.826	0.650	0.651	0.895
TWO-STAGE	12033	2781	13787	2429	0.832	0.832	0.812	0.832	0.663	0.664	0.900

According to Table II, we have similar results on the testing sets U79 and P80. On these sets, TWO-STAGE performs the best on each measure. It is noticed in Table II that the number of true positives is increased and the number of false positives is decreased simultaneously by ONE-STAGE and even more by TWO-STAGE.

4 Conclusion

In this study we provide a two-stage RBFN classifier to improve the predicting power of protein disorder. The detected disorder information is expected to be useful in protein structure prediction and functional analysis. In the future, more predicted information from primary sequences and more machine learning skills to deal with skew datasets can be merged to enhance the strength of the classifiers.

References

1. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm*. J Mol Biol, 1999. **293**(2): p. 321-31.
2. Fink, A.L., *Natively unfolded proteins*. Curr Opin Struct Biol, 2005. **15**(1): p. 35-41.
3. Jones, D.T. and J.J. Ward, *Prediction of disordered regions in proteins from position specific score matrices*. Proteins, 2003. **53 Suppl 6**: p. 573-8.
4. Su, C.T., C.Y. Chen, and Y.Y. Ou, *Protein disorder prediction by condensed PSSM considering propensity for order or disorder*. BMC Bioinformatics, 2006. **7**: p. 319.
5. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life*. J Mol Biol, 2004. **337**(3): p. 635-45.
6. Su, C.T., C.Y. Chen, and T.M. Hsu, *Enhancing Protein Disorder Detection by Refined Secondary Structure Prediction*. <http://biominer.bime.ntu.edu.tw/disorderps/> (Technical report).
7. QuickRBF, <http://muse.csie.ntu.edu.tw/~yien/quickrbf/index.php>.