

DisPSSMP2: A Two-Stage RBFN Classifier for protein disorder prediction

Chung-Tsai, Su and Chien-Yu, Chen

In the above sections, 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.

There are some observations for the adoption of cascading classification. 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. Peng *et al.* adopted a meta-classifier to predict the weights of the classifiers for short disordered regions and long disordered regions. Thus, in this section, 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.

1.1 The enlargement of training sets

In DisPSSMP, the training sets comprise PDB693 and D184. From Figure 1, there are partial results of DisPSSMP in the training set R80. The terminal regions of most proteins are predicted as disordered. However, the results comprise many false positives (light blue color) in N-/C- terminus of each protein. For example, 1b8z, 1n7v, and 3fit

have ordered regions in their N-/C- terminus but these regions are incorrectly predicted as disordered by DisPSSMP. It means that DisPSSMP prefers to predict residues in N-/C- terminus as disordered. The primary reason might be due to the window-based feature vectors of the residues of N-/C- terminus are padded with at most $(47-1)/2$ special spacer characters (set 0 in DisPSSMP). Since there are more than 60% of disordered residues in terminal regions of the training sets, it causes that the classifier has too much favour to over-prediction in N-/C- terminus as a result of the special spacer characters. Therefore, we collect a set of fully ordered proteins from PDB database to balance the frequency of disordered/ordered residues in N-/C- terminus.

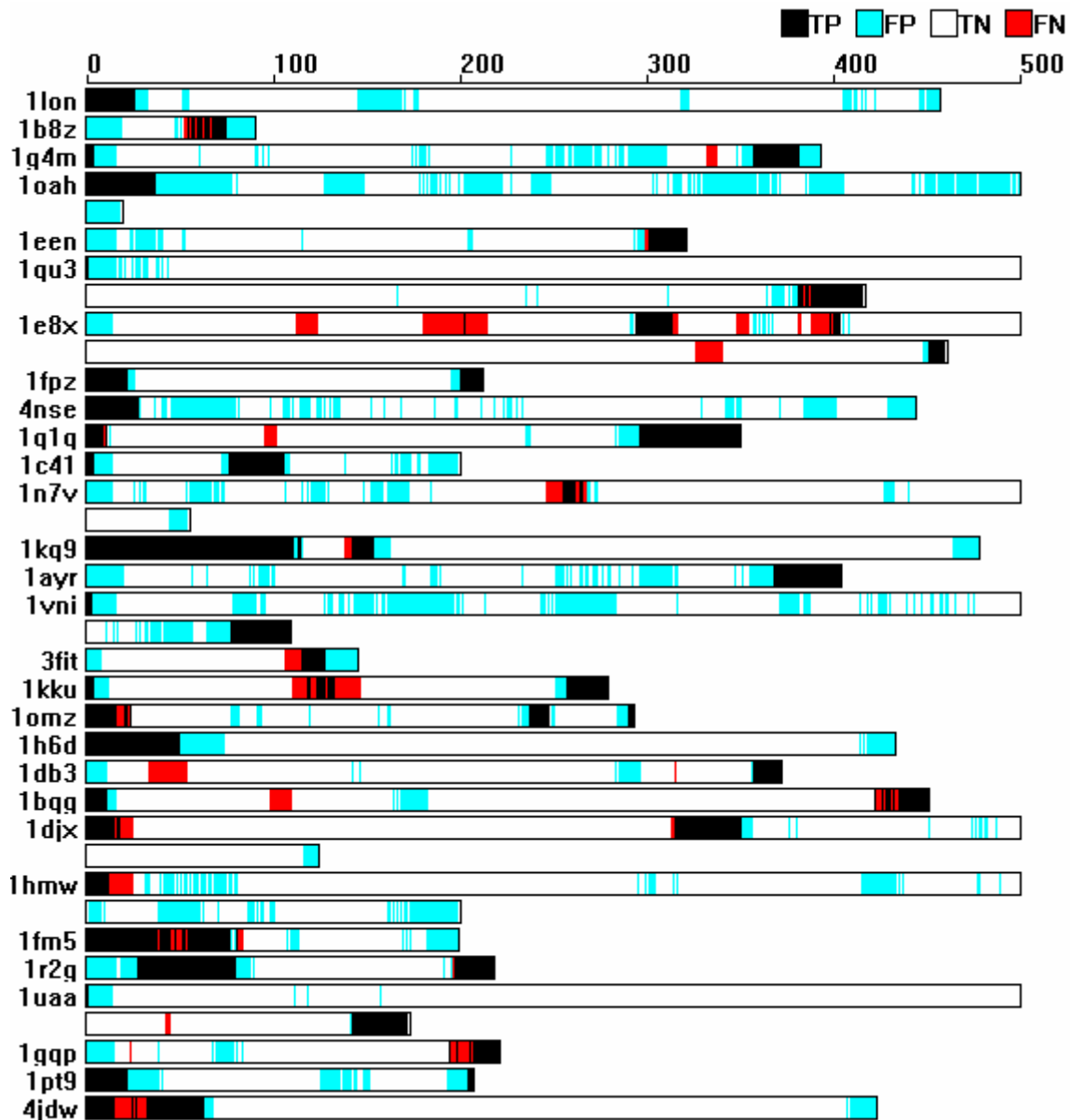


Figure 1 The partial results of DisPSSMP in the training set R80

According to Table 1, the first training set PDB652 contains 652 partially disordered proteins from the PDB database, each of which contains at least one disordered regions with more than 30 consecutive residues. The second training set D184 is derived from DisProt database, a curated database that provides information about proteins that wholly or partially lack a stable three-dimensional structure under putatively native conditions. For the details of the procedures in preparing datasets PDB652 and D184,

the readers can refer to our recent work. Different from the set PDB693 in, PDB652 excludes the sequences with similarity identity of more than 70% against any protein sequence in the other training sets by running Cd-Hit, resulting 652 proteins.

Table 1 The training sets for the two-stage RBFN classifier

Number of :	Training data		
	PDB652	D184	G200
Chains	652	184	200
Ordered regions	1281	257	200
Disordered regions	1613	274	0
Residues in ordered regions	190936	55164	37959
Residues in disordered regions	49365	27116	0
Total residues in the dataset	240301	82280	37959

Since PDB652 and D184 contain more than 60% of disordered residues in terminal regions of the proteins, which causes the window-based classifiers to over-predict the terminal residues as disorder, this work collects an additional training set G200 from the PDB database (there are 35579 proteins structures with 85233 chains in the PDB release of 13-May-2006). After removal of the DNA chains and the protein chains shorter than 80 residues or with disordered regions, only 1847 fully ordered chains remain. Among the completely ordered proteins, 200 of them are randomly selected as the dataset G200. Similarly, we use the same criterion described above to handle the redundancy issue.

1.2 The modification of the Classifier

DisPSSMP adopts the QuickRBF package 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 thesis, 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 the number of the centers in the network be c , the probability distribution functions for the classes order and disorder are represented as follows:

$$\begin{bmatrix} \text{Order}(R_i) \\ \text{Disorder}(R_i) \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,c} \\ w_{2,1} & w_{2,2} & \dots & w_{2,c} \end{bmatrix} \begin{bmatrix} \phi_1(R_i) \\ \phi_2(R_i) \\ \dots \\ \phi_c(R_i) \end{bmatrix}. \quad (5)$$

, where R_i is the feature vector of a residue in the data sets, $\phi_j(R_i)$ is the j -th kernel function employed, and $w_{1,j}$ and $w_{2,j}$ are the weights on the edges connecting the hidden nodes and output nodes of the RBF network. Since $\text{Disorder}(R_i)$ and $\text{Order}(R_i)$ are supposed to be in between 0 and 1, they are set to 1 and 0 when the values generated by QuickRBF are larger than 1 or smaller than 0, respectively. Then, the formula for calculating the disorder propensity of residue R_i is represented as follows:

$$\text{Propensity}_D(R_i) = (\text{Disorder}(R_i) - \text{Order}(R_i) + 1)/2. \quad (6)$$

Next, the alternative decision function $\text{Classifier}(R_i)$ is shown in Equation (7). 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.

$$\text{Classifier}(R_i) = \begin{cases} 1 & \text{if } (\text{Propensity}_D(R_i) \geq \text{Threshold}); \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

1.3 The architecture of the two-stage classifier

According to DisPSSMP, we present a two-stage RBFN classifier to further enhance the predicting power. Figure 2 shows the procedure of generating feature vectors from a protein sequence into FS-PSSMP-4 in ONE-STAGE and extracting the feature vectors of TWO-STAGE from the intermediary results of ONE-STAGE. Among the complicated procedure, FS-PSSMP-4 has been introduced in our previous study. The first classifier outputs $\text{Disorder}(R_i)$ and $\text{Order}(R_i)$, and the first-level prediction is determined by Equation (7) 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 (7) with Threshold_2 . In this thesis, the classifier of the single-stage is denoted as ONE-STAGE, and the classifier including both stages is denoted as TWO-STAGE.

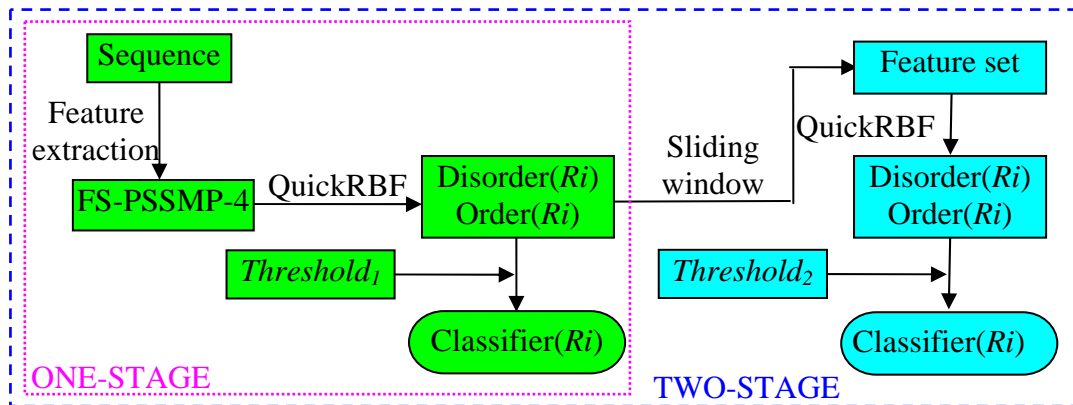


Figure 2 The architecture of the proposed two-stage classifier.

1.4 Cross-Validation of the two-stage classifier

In order to conduct a five-fold cross validation, all the chains in the training sets PDB652, D184, and G200 are randomly split into five subsets of approximately equal sizes. Before that, we would like to introduce another popular measure to determine the threshold in Equation (7). The *Receiver Operating Characteristic (ROC)* curve has been considered as with the discriminating ability when comparing disorder prediction methods. Indeed, the *area under the ROC curve (AUC)* in Equation (8) represents the performance of each method fairly. In Equation (8), p is the threshold of the classifier in Equation (7) from 0.01 to 1.00; $Spec(p)$ and $Sens(p)$ are specificity and sensitivity when the threshold in Equation (7) is p . Therefore, the evaluation in cross validation is also based on this measure.

$$AUC = \sum_{p=0.01}^1 [Spec(p) - Spec(p - 0.01)] \times [Sens(p) + Sens(p - 0.01)] / 2 \quad (8)$$

To determine the window size of the feature sets of both ONE-STAGE and TWO-STAGE, a range of window size from 11 to 59 has been evaluated. 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. Then, the parameter *Threshold* in Equation (7) is determined when the probability excess is maximal. From Table 2, the parameters $Threshold_1$ for ONE-STAGE and $Threshold_2$ for TWO-STAGE are 0.23 and 0.36.

Table 2 Cross-Validation for the two-stage RBFN classifier

Method	TP	FP	TN	FN	Sens.	Spec.	Prob.	AUC.	Threshold
--------	----	----	----	----	-------	-------	-------	------	-----------

								<i>Excess</i>		
ONE-STAGE	48587	52609	231450	27894	63.53	81.48	45.01	78.23	0.23	
TWO-STAGE	49763	54497	229562	26718	65.07	80.81	45.88	78.12	0.36	

1.5 Results and Discussions

When compared with DisPSSMP, both ONE-STAGE and TWO-STAGE improve the performance of protein disorder prediction on all the testing sets. As shown in Table 3 and Table 4, 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.

Table 3 Results on the testing set R80

	TP	FP	TN	FN	<i>Sens.</i>	<i>Spec.</i>	<i>Prec.</i>	<i>Accu.</i>	<i>MCC</i>	<i>CASP S score</i>	<i>Prod.</i>	<i>Prob. Excess</i>	<i>AUC</i>
DisPSSMP	2800	4550	25359	849	0.767	0.848	0.381	0.839	0.463	0.119	0.651	0.615	0.884
ONE-STAGE	2720	3535	26374	929	0.745	0.882	0.435	0.867	0.501	0.122	0.657	0.627	0.896
TWO-STAGE	2553	2476	27433	1096	0.700	0.917	0.508	0.894	0.538	0.120	0.642	0.617	0.897

Table 4 Results on the testing set U79 and P80

	TP	FP	TN	FN	<i>Sens.</i>	<i>Spec.</i>	<i>Prec.</i>	<i>Accu.</i>	<i>MCC</i>	<i>CASP S score</i>	<i>Prod.</i>	<i>Prob. Excess</i>	<i>AUC</i>
DisPSSMP	11934	3896	12672	2528	0.825	0.765	0.754	0.793	0.589	0.294	0.631	0.590	0.869
ONE-STAGE	11777	2538	14030	2685	0.814	0.847	0.823	0.832	0.662	0.329	0.690	0.661	0.899
TWO-STAGE	11737	1737	14831	2725	0.812	0.895	0.871	0.856	0.711	0.352	0.726	0.707	0.922

First, we focus on between DisPSSMP and ONE-STAGE. Since the number of ordered residues which is introduced by ONE-STAGE is four times more than DisPSSMP, the number of the false positives in the testing set R80 decreases from 4550 to 3535 (more than 20%) without explicit decrement of the true positives in Table 3. According to Table 4, although the number of the true positives decreases slightly from 11934 in

DisPSSMP to 11777 in ONE-STAGE (about 1.3%), the number of the false positives is reduced from 3896 in DisPSSMP to 2538 in ONE-STAGE (about 34.8%). To compare ONE-STAGE with DisPSSMP in Table 5, *Sensitivity* of ONE-STAGE decreases about 1.4%, but *Specificity* of ONE-STAGE increases about 5.1%. Therefore, the enlargement of the training data and the alternative decision function of the RBFN classifier benefit the predicting power of protein disorder.

Table 5 Results on the testing set R80, U79, and P80

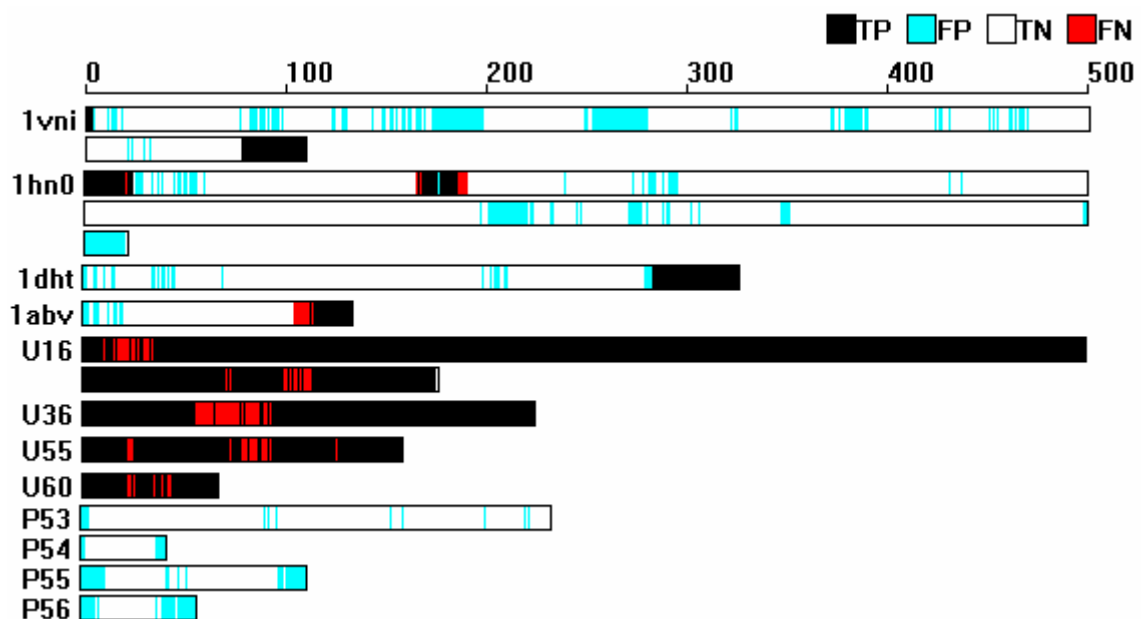
	TP	FP	TN	FN	<i>Sens.</i>	<i>Spec.</i>	<i>Prec.</i>	<i>Accu.</i>	<i>MCC</i>	<i>CASP S score</i>	<i>Prod.</i>	<i>Prob. Excess</i>	<i>AUC</i>
DisPSSMP	14734	8446	38031	3377	0.814	0.818	0.636	0.817	0.592	0.255	0.666	0.632	0.887
ONE-STAGE	14497	6073	40404	3614	0.800	0.869	0.705	0.850	0.646	0.270	0.696	0.670	0.904
TWO-STAGE	14290	4213	42264	3821	0.789	0.909	0.772	0.876	0.694	0.282	0.718	0.698	0.920

Next, we discuss about ONE-STAGE and TWO-STAGE. From Table 3, the number of the false positives in TWO-STAGE decreases about 30% although the number of the true positives decreases about 6.1%. Especially, the number of the false positives in TWO-STAGE decreases more than 31%, but the number of the true positives decreases less than 0.4% in Table 4. Since the technique of the two-stage RBFN classifier contributes the smoothing effect on prediction, the *probability excess* is improved apparently from 66.1% to 70.7% in the testing sets U79 and P80, which are totally disordered and fully ordered, respectively. However, TWO-STAGE doesn't bring a better performance which evaluated by the *probability excess* and *AUC* in R80. For further observation, several proteins in the testing sets are shown in Figure 3. Comparing with ONE-STAGE, TWO-STAGE not only removes most of sparse short ordered segments predicted improperly as disordered, but also integrates several dense short disordered segments predicted incorrectly as ordered into a long disordered segment. Like 1vni in Figure 3, each short ordered segment predicted incorrectly as

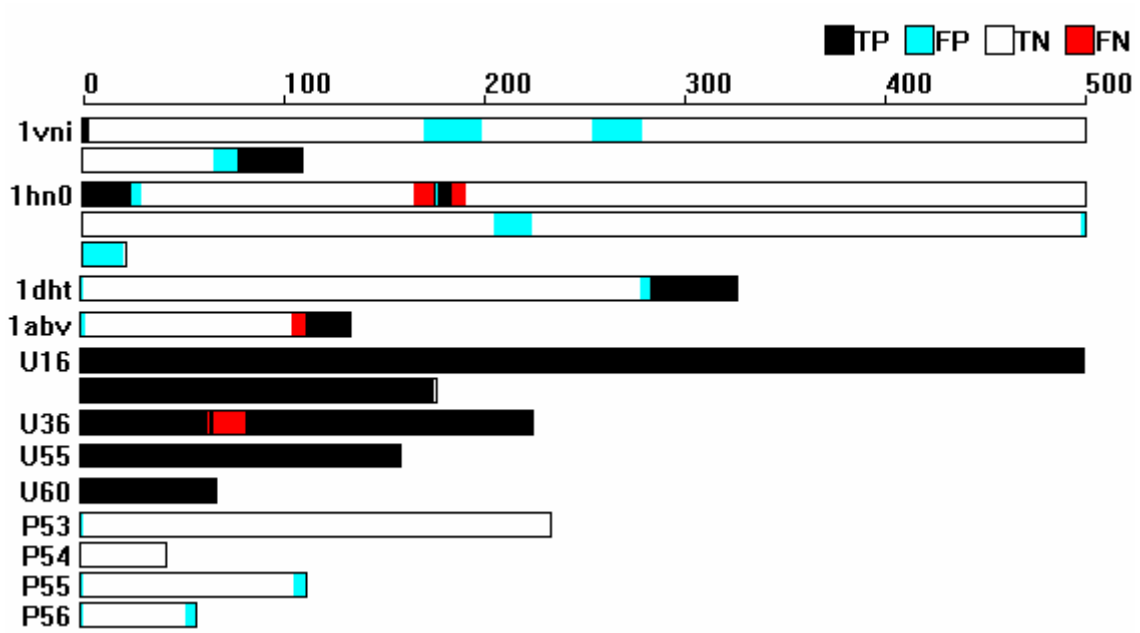
disordered is removed in TWO-STAGE although the two long ordered segments predicted incorrectly as disordered are still kept. According to Figure 3, all ordered residues predicted improperly as disordered (false positives) of u16 are corrected as disordered (true positives). In particular, the false positives from p53 to p56 are markedly reduced in TWO-STAGE, including those in their N-/C- terminus. In addition to improving the performance by the smoothing effect, TWO-STAGE also reduces the effect on over-prediction on N-/C- terminus of proteins. Indeed, TWO-STAGE has a better ability to characterize a propensity for disorder and order of protein segments than ONE-STAGE. Besides, there are four examples shown in Figure 4 for visualization.

There are two distinct observations from Figure 4 :

- (1) TWO-STAGE provides a more smoothing effect for its prediction than ONE-STAGE;
- (2) TWO-STAGE is a classifier of greater ability to enhancing its predicting confidence than ONE-STAGE. In other words, TWO-STAGE can predict correctly disordered residues with higher propensity for disorder, and visa versa.



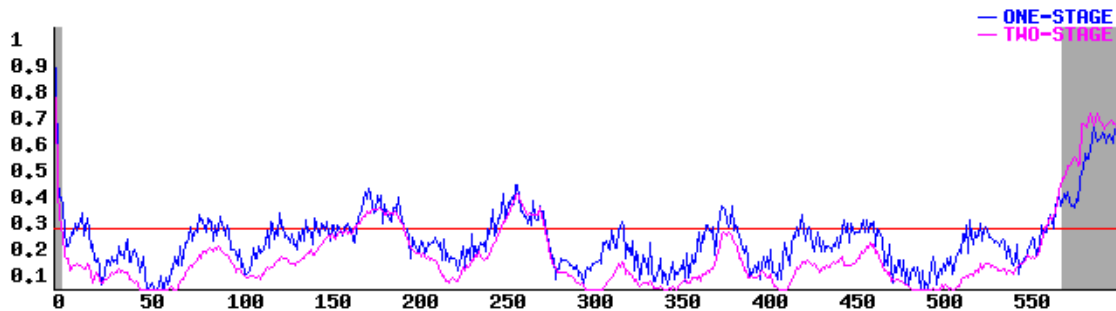
(a) ONE-STAGE



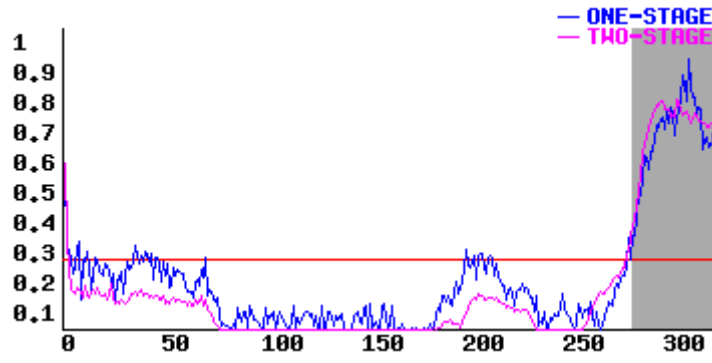
(b) TWO-STAGE

Figure 3 Examples for ONE-STAGE and TWO-STAGE

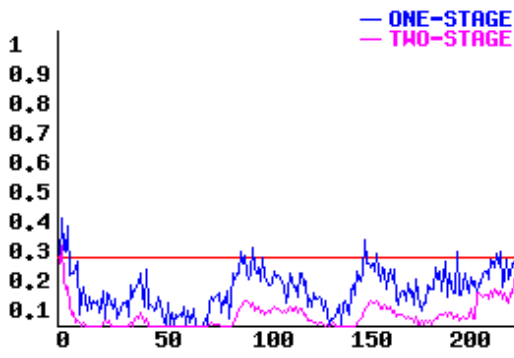
For example, the false positives in all short segments of 1vni in Figure 4 (a) are corrected by using TWO-STAGE, but two long segments with false positives are still remained. In addition, the false positives in P80 and the false negatives in U79 predicted from ONE-STAGE are almost corrected by TWO-STAGE according to Figure 3 and Figure 4. Moreover, the faults of over-prediction in N-/C- terminus among the testing set P80 are corrected by TWO-STAGE.



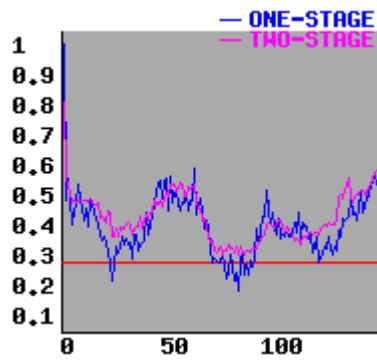
(a) 1vni



(b) 1dht



(c) p53



(d) u55

Figure 4 Plots of ONE-STAGE and TWO-STAGE

Furthermore, we compare the performance of ONE-STAGE and TWO-STAGE with those existing packages shown in all testing sets. Then there are twenty disorder prediction packages illustrated with Figure 5, Figure 6, and Figure 7 which refer to R80, U79, and P80, respectively. In conclusion, the two-stage RBFN classifier we propose has a benefit to protein disorder prediction.

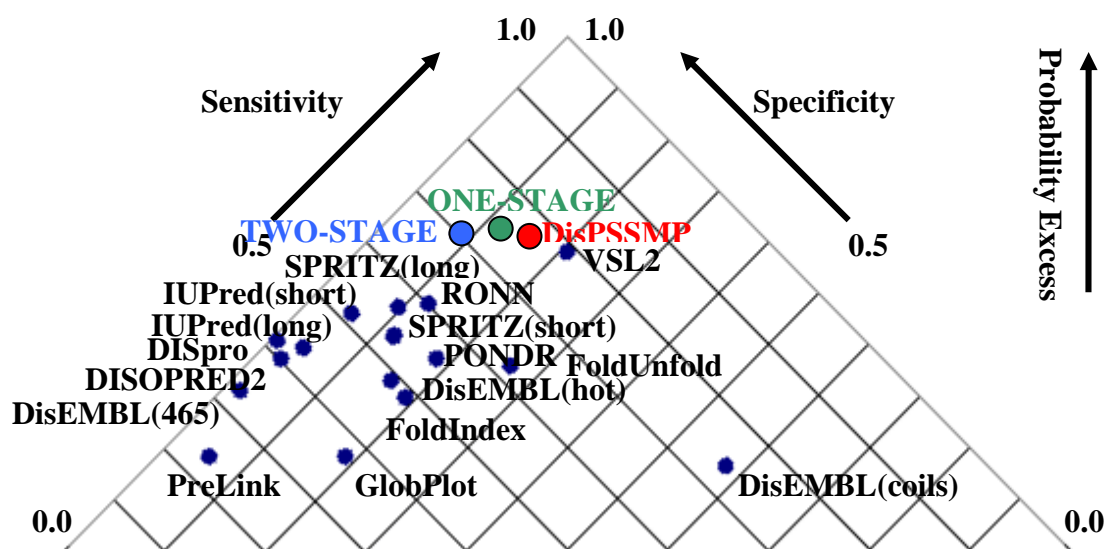


Figure 5 Comparing the performance of twenty disorder prediction packages on testing data R80

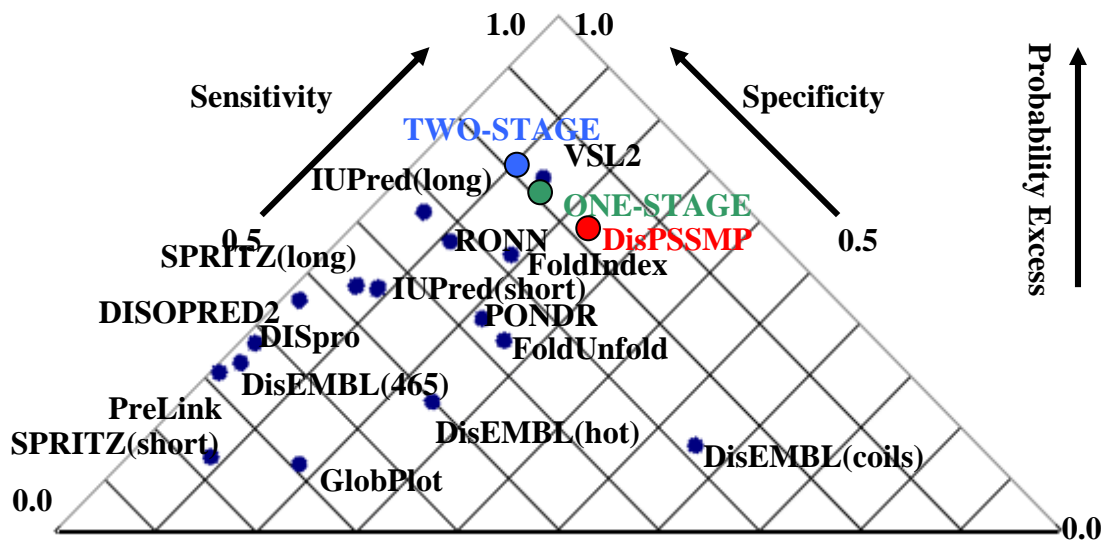


Figure 6 Comparing the performance of twenty disorder prediction packages on testing data U79 and P80

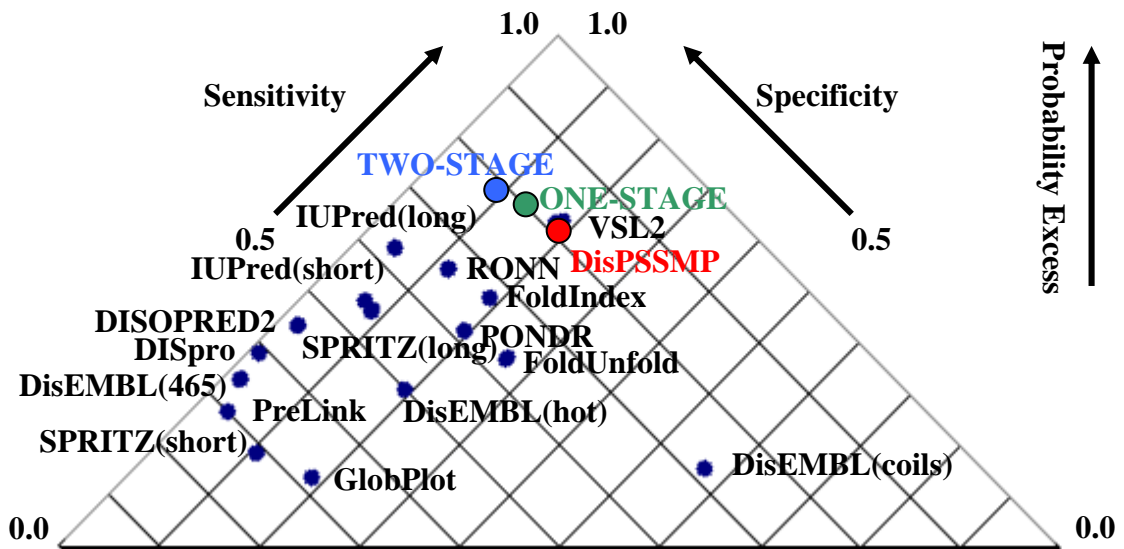


Figure 7 Comparing the performance of twenty disorder prediction packages on testing data U79 and P80