

Enhancing Protein Disorder Detection by Refined Secondary Structure Prediction

Chung-Tsai Su¹, Tong-Ming Hsu¹, Chien-Yu Chen^{*2}, Yu-Yen Ou^{3,4}, and Yen-Jen Oyang¹

¹Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C, ²Department of Bio-industrial Mechatronics Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C, ³Graduate School of Biotechnology and Bioinformatics, Yuan Ze University, Chung-Li, 320, Taiwan, R.O.C., and ⁴Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, 320, 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. Previous studies largely use evolutionary information and combinations of physicochemical properties of amino acids to detect disordered regions from primary sequences. 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 investigate how the information of secondary structure can be incorporated in DisPSSMP to enhance the predicting power. We propose a new representation of secondary structure information and compare it with three naïve representations that have been discussed or employed in some related works. The experimental results reveal that the refined information from secondary structure prediction is of benefit to this problem.

Key words: protein disorder prediction; secondary structure; Radial Basis Function Network

1 Introduction

Intrinsically disordered proteins or protein regions exhibit unstable and changeable three dimensional structures under physiological conditions [1, 2, 3]. Although lacking fixed structures, many unfolded disordered proteins or partial protein regions have been identified to participate in many biological processes and carry out

* Corresponding author.

important biological functions [2, 4]. Also, it has been observed that the absence of a rigid structure allows disordered binding regions to interact with several different targets [5, 6]. 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 [6, 7, 8, 9]. Disordered regions are distinguished from ordered regions by its low sequence complexity [10], amino acid compositional bias [7], high evolutionary tendencies [11], or high flexibility [12]. In our recent work DisPSSMP, we investigated the predicting power of a condensed position specific scoring matrix with respect to physicochemical properties (PSSMP) on the prediction accuracy, where the PSSMP is derived by merging several amino acid columns of a PSSM belonging to a certain property into a single column [13]. Additionally, DisPSSMP decomposes each conventional physicochemical property of amino acids into two disjoint groups which have a propensity for order and disorder respectively. It outperforms the existing packages in predicting protein disorder by employing some new properties that perform better than their parent properties [13].

Several studies have attempted to incorporate the predicted information of secondary structure elements (SSE) in predicting protein disorder [3, 6, 14, 15, 16, 17, 18, 19]. NORSp aims to identify the regions with no sufficient regular secondary structure as disorder, by means of merging predictions of secondary structure from PROFphd, transmembrane helices from PHDhtm, and coiled-coil regions from COILS [14, 15]. The GlobPlot service detects sequence regions of globularity or disorder by calculating a running sum of the propensity for random coils and secondary structures [16]. Meanwhile, some other approaches employed secondary structure information as parts of their features. DISOPRED2 employs the predicted secondary structures from PSIPred to refine its prediction of disordered residues [3, 6]. DISpro combines evolutionary information from PSI-BLAST, secondary structures from SSpro, and relative solvent accessibility from ACCpro for protein disorder prediction [17]. Similarly, VLS2 adopts various features in predicting protein disorder, including amino acid frequencies, spacer frequency, sequence complexity, charge-hydrophobicity ratios, averaged flexibility indices, and averaged PSI-BLAST profiles, as well as the averaged PHD and PSIPred secondary structure predictions [18, 19].

Although employing the predicted information of secondary structure is not new to this problem, it is not clear how much this feature contributes when employed with other important features such as amino acid physicochemical properties. There are also some challenges when retrieving secondary structure information from existing packages of secondary structure prediction. Here we use two examples to illustrate the difficulties. Fig. 1(a) shows the partial results of six famous secondary structure predictors for the protein *platelet membrane glycoprotein IIIa beta subunit* (chain B of PDB structure 1JV2) and Fig. 1(b) for *DNA-directed RNA polymerase II largest subunit* (chain A of 1I3Q). It is observed that the categories of secondary structure predicted from various predictors are sometimes inconsistent and the boundaries they detected are largely unlike, especially for the short segments. Thus, it is of great interest to develop a refined description of SSE that benefits the problem of protein disorder prediction.

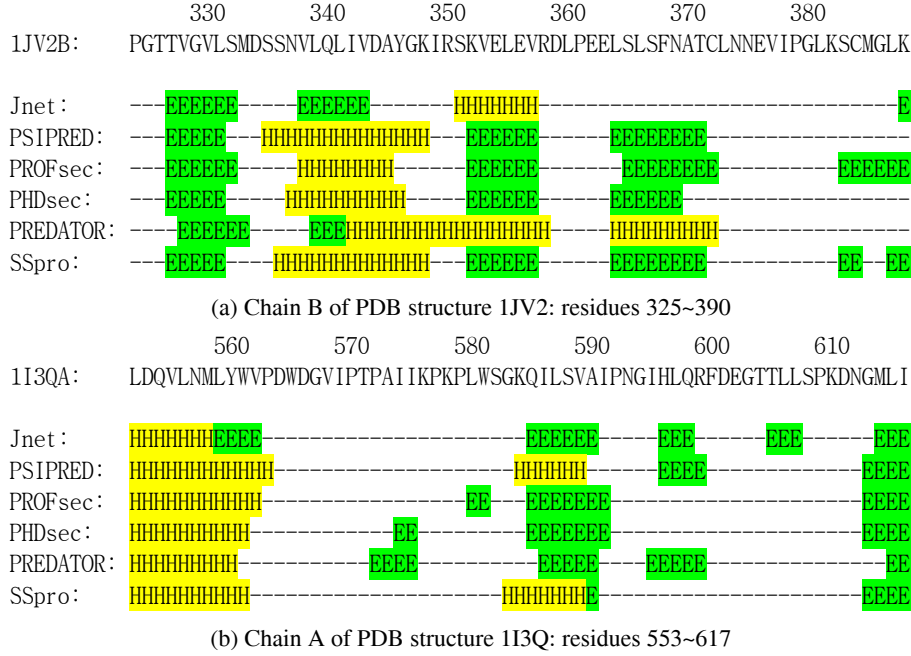


Fig. 1. The partial results of several secondary structure predictors on the chain B of PDB structure 1JV2 and chain A of 1I3Q. Helices are denoted as ‘H’, strands as ‘E’, and coils as ‘-’.

In this study, we propose a new representation to refine the secondary structure information and integrate it with the condensed PSSM features developed in our recent study [13]. The new representation transforms the predicted secondary structure elements (SSE) into a distance-based feature. The proposed idea is compared with three naïve representations that have been discussed or employed in some related works. The experimental results reveal that after the refinement procedure, the influence of potential errors from secondary structure prediction can be effectively reduced.

2 Materials

For training and validation processes, six datasets have been extracted from different databases. The number of chains, ordered/disordered regions, and residues in ordered/disordered regions of each dataset are provided in Table 1. The training data is composed of datasets PDB652, D184, and G200, which are based on the procedures described in the following paragraphs. On the other hand, three independent datasets, which are named R80, U79, and P80, are employed as validation benchmarks as in related studies [20, 21]. These blind testing sets serve as a platform for comparing the performance of different SSE representations.

Table 1. Summary of the datasets employed in this study

Number of :	Training data			Testing data		
	PDB652	D184	G200	R80	U79	P80
Chains	652	184	200	80	79	80
Ordered regions	1281	257	200	151	0	80
Disordered regions	1613	274	0	183	79	0
Residues in ordered regions	190936	55164	37959	29909	0	16568
Residues in disordered regions	49365	27116	0	3649	14462	0
Total residues in the dataset	240301	82280	37959	33558	14462	16568

The first training set PDB652 contains 652 partially disordered proteins from the PDB database [22], 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 [23], 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 [13]. Different from the set PDB693 in [13], PDB652 excludes the sequences with similarity identity of more than 70% against any protein sequence in the other training sets by running Cd-Hit [24], resulting 652 proteins.

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 [22] (there are 35579 proteins structures containing 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.

There are three independent datasets for evaluation in this study. The first set R80, prepared by Yang *et al.* [20], contains 80 protein chains from PDB database. The second set U79, provided by Uversky *et al.* in 2000 [21], contains 79 wholly disordered proteins. Finally, the third testing set P80, organized by PONDR (retrieved in February 2003), includes 80 completely ordered proteins. Like Yang *et al.* did in their study [20], these testing sets were employed in some recent related studies as a platform in comparison of different approaches in protein disorder prediction. Particularly, the testing sets U79 (wholly disordered proteins) and P80 (entirely globular proteins) examine whether the proposed method is under- or over-predicting protein disorder.

3 Methods

In this section, we first introduce the disorder predictor DisPSSMP, which was proposed in our recent work based on condensed PSSMs considering propensity for order or disorder [13]. Next, four representations of summarizing the local information of secondary structure are presented. Finally, the procedures of constructing the Radial Basis Function Networks (RBFN) classifier and some widely used evaluation measures are described in details.

3.1 Organizing feature sets

The disorder predictor DisPSSMP constructs its predicting model based on the condensed position specific scoring matrix with respect to physicochemical properties (PSSMP), which are shown to exhibit better performance on protein disorder prediction than the original PSSM features [13]. The success of DisPSSMP thanks to its invention of considering the disorder propensity of amino acids when searching for an optimized feature combination of PSSMPs. The selected condensed PSSM properties include: *Aliphatic*, *Aromatic_o*, *Polar*, and *Small_D* [13]. The derived feature set improves the performance of a classifier built with RBFN in comparison with the feature set constructed with PSSMs or PSSMPs that adopt simply the conventional physicochemical properties. The original feature set employed by DisPSSMP is named PSSMP-4 in the rest of this study.

In this study, we aim to investigate how the predicting power of DisPSSMP can be improved when incorporating secondary structure information with PSSMP-4. We propose a new representation named SSE-DIS, and compare it with other representations listed in Table 2, named SSE-BIN, SSE-PRO, and SSE-DEN respectively. Before extracting the features from the results of a secondary structure predictor, a SSE with less than five successive secondary structure residues are removed. We expect the remaining secondary structure segments to provide more reliable information than the original predictions. As summarized in Table 2, SSE-BIN comprises three binary features which correspond to the predicted secondary structure classes (helices, strands, and coils), respectively. Like DISpro [17], for each residue, only one of the three features is set to 1 according to its SSE class. Another representation, named SSE-PRO, includes three real values which represent the probabilities for each class. Next, SSE-DEN calculates the density of secondary structures within a specific window size. Some pilot experiments based on training data show that the performance of SSE-DEN using various window sizes from 11 to 61 is almost the same. In this regard, a window size of 41 is employed in the experiments reported in this study. The feature SSE-DEN is derived from the concept of NORS (no regular secondary structure) [14]. Finally, the proposed representation SSE-DIS takes the distance of a residue to its nearest secondary structure element. This feature aims to emphasize the locations which are far from the regions consisted of regular secondary structures. The procedures of generating these four representations are exemplified by Fig. 2.

Table 2. The definition of four representations of secondary structure

Name	Description	Number of features*	Parameters associated with the representation
SSE-BIN	Binary values decoding helixes, strands, and coils	3	None
SSE-PRO	Probability values for helixes, strands, and coils	3	None
SSE-DEN	The density of secondary structures	1	A window size for calculating densities. It is set as 41 after some pilot experiments were conducted.
SSE-DIS	The distance from the nearest secondary structure element	1	None

*Each feature is normalized into the interval [0, 1] before they are employed in prediction.

In Fig. 2, column (a) stands for the original protein sequence, and column (b) is the predicted secondary structure element. In this study, we employ Jnet [25] as the secondary structure predictor, which is a neural network secondary structure predictor based on multiple sequence alignment profiles. Next, the predicted information is refined by removal of secondary structure elements with less than five residues, resulting column (c). The refined information is next transformed into features SSE-BIN (d1), SSE-DEN (d3), and SSE-DIS (d4) respectively. At the same time, the feature SSE-PRO (d2) is transformed from the original results of Jnet. The example used in Fig. 2 is from the protein *cys regulon transcriptional activator cysb* (PDB structure 1AL3). For each residue, the feature values falling in a window size of l centered at the given residue are extracted as its feature vector and the experiments of considering different window sizes are shown in the next section.

3.2 Classifier

DisPSSMP adopts the QuickRBF package [26] in constructing RBFN for classification. In this study, we in particular tackle the problem of handling skewed datasets, which stands for the problems with unbalanced numbers of positive (disorder) and negative (order) samples. In the previous implementation of DisPSSMP, equal quantity of residues from ordered and disordered regions was used in constructing the classifier. However, a large volume of ordered regions was lost after randomly removal of unwanted ordered residues. Thus in this study, all of ordered and disordered residues in the training datasets are included to construct the classifier without 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:

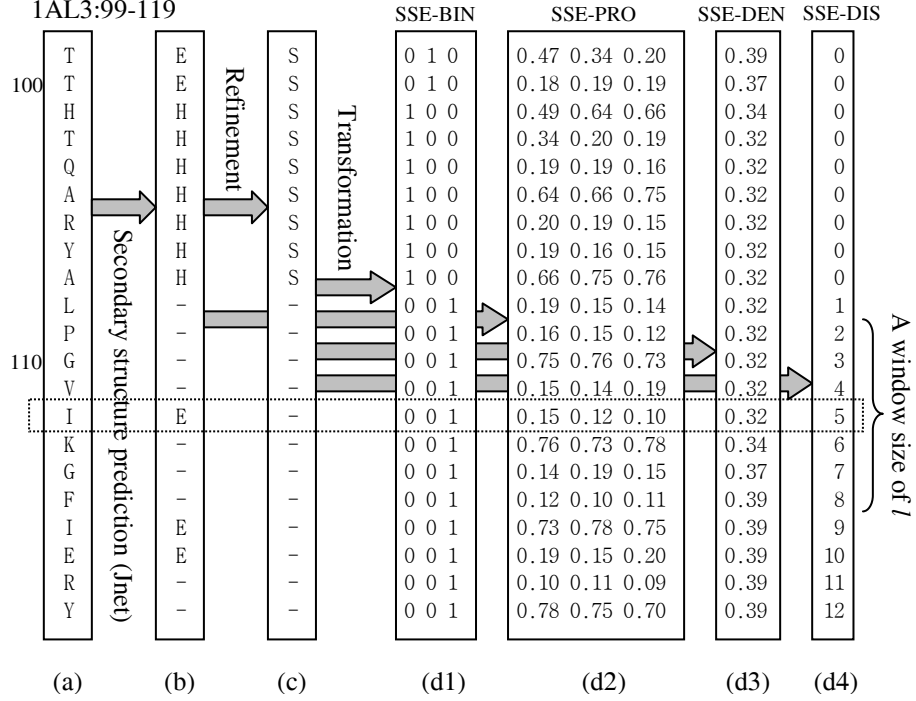


Fig. 2. The procedures of generating different representations (d1~d4) for the predicted secondary structures.

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

, 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 of 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:

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

Finally, the alternative classification function $\text{Classifier}(R_i)$ is shown in Equation (3). 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.

$$\text{Classifier}(R_i) = \begin{cases} 1 & \text{if } (\text{Propensity}_D(R_i) \geq \text{Threshold}); \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

3.3 Evaluation measures

Protein disorder prediction is a binary classification task, and many validation measures have been introduced to this problem [27, 28, 29]. Since the disordered residues of proteins in PDB database are so rare that a skewed dataset is considered here, we employ four measures that are considered proper in this problem together to evaluate the performance of different feature sets. As listed in Table 3, *sensitivity* and *specificity* represent the fraction of disordered and ordered residues correctly identified, respectively. Another commonly used evaluation measure *probability excess*, recommended by CASP6 [28] and Yang *et al.* [20], is employed here, too. Last, the Receiver Operating Characteristic (ROC) curve has been considered as with the most discriminating ability when comparing disorder prediction methods [17, 19, 28]. Indeed, the *area under the ROC curve* (AUC) represents the performance of each method fairly. Accordingly, the evaluation in cross validation is also based these measures.

3.4 Constructing predicting models with cross validation

In order to conduct a five-fold cross validation, all the chains in datasets PDB652, D184, and G200 are randomly split into five subsets of approximately equal sizes. As suggested in DisPSSMP [13], the feature set PSSMP-4 with the window size set as 47 is employed in cross validation to determine the parameter *Threshold* of Equation 3. In general, *sensitivity* increases when *specificity* decreases, and vice versa. Therefore, in this study, *Threshold* is determined by maximizing the probability excess. From Table 4, *probability excess* achieves maximal when *Threshold* is 0.22.

Table 3. The equations of the evaluation measures

Measure	Abbreviation	Equation
Sensitivity	<i>Sens.</i>	TP/(TP+FN)
Specificity	<i>Spec.</i>	TN/(TN+FP)
Probability excess	<i>Prob. Excess</i>	(TP×TN−FP×FN)/((TP+FN)×(TN+FP))
Area under ROC curve	<i>AUC</i>	$\sum_{p=1}^{100} [\text{Spec}(p/100) - \text{Spec}((p-1)/100)] * [\text{Sens}(p) + \text{Sens}((p-1)/100)] / 2$

TP: the number of correctly classified disordered residues; FP: the number of ordered residues incorrectly classified as disordered; TN: the number of correctly classified ordered residues; FN: the number of disordered residues incorrectly classified as ordered; *p*: the threshold employed in Equation (3) (*p* = 1, 2, ..., 100); *Spec*(*p*/100): specificity when the threshold in Equation (3) is *p*/100; *Sens*(*p*/100): sensitivity when the threshold in Equation (3) is *p*/100.

Table 4. Cross validation for the training sets with window size of 47 for PSSMP-4

TP	FP	TN	FN	<i>AUC</i>	<i>Threshold</i>	<i>Sens.</i>	<i>Spec.</i>	<i>Prob. Excess</i>
49672	57826	226233	26809	0.781	0.22	0.650	0.796	0.446

4 Results and Discussions

In this section, we first evaluate how DisPSSMP performs when incorporating four representations of the secondary structure respectively. After that, we show some statistics from the secondary structures predicted by Jnet and discuss how SSE-DIS benefits the protein disorder prediction.

4.1 Results on testing data

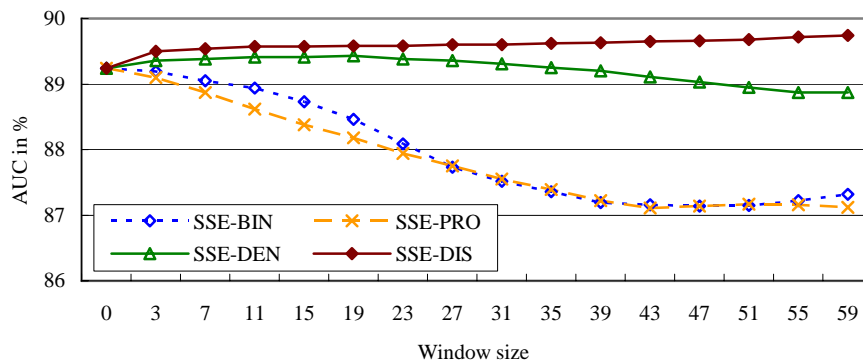
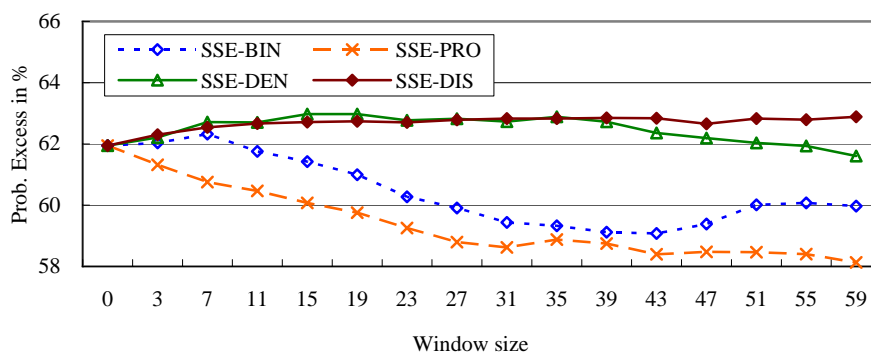
For the blind testing sets, the comparison of the performance of PSSMP-4 with four representations of secondary structure is performed by a range of the window size l from 0 to 59, while zero means that the feature set comprises only PSSMP-4.

In Fig. 3, the results on the testing set R80 are shown. According to *AUC* in Fig. 3(a), the performance of SSE-DIS is improved when the window size increases. Meanwhile, the performance of SSE-DEN increases slightly when the window size is smaller than 19 and decreases when larger window sizes are considered. On the other hand, the performance of SSE-BIN and SSE-PRO decrease apparently when they are incorporated with PSSMP-4. It is concluded that SSE-DIS performs consistently better than the other representations when different window sizes are considered. Fig. 3(b) shows the comparison based on another measure *probability excess*. From this point of view, the predicting powers of SSE-DIS and SSE-DEN are comparable when the window size is not large. Combining the results of all the testing sets, it reveals that the representations SSE-BIN and SSE-PRO fail to improve the accuracy of DisPSSMP when they are incorporated with the original feature set PSSMP-4.

For wholly ordered or disordered proteins, the comparison is conducted on the testing sets U79 and P80. It can be observed in Fig. 4(a) and (b) that the difference between SSE-DIS and SSE-DEN is more significant in this comparison. Like in Fig. 3(b), Fig. 4(b) shows that SSE-BIN has a better *probability excess* than PSSMP-4 when l is smaller than 19. It seems that SSE-BIN provides some useful information when the sliding window is small.

4.2 Discussions

We observed that the classifier trained with PSSMP-4+SSE-DIS predicts more disordered residues than the classifier trained from PSSMP-4 alone. Here we use two examples to explain the difference between these two classifiers. The experimental results of the protein *methionyl-tRNA synthetase* (PDB structure 1PG2 in the dataset R80) and the protein *Heat shock transcription factor, N-terminal activation domain*

(a) Comparison based the measure *AUC*.(b) Comparison based on the measure *probability excess*.**Fig. 3.** The results on the testing set R80.

(u56 from dataset U79) are drawn in Fig. 5(a) and Fig. 5(b), respectively. For 1PG2, there are three disordered regions which are residues of 1-3, 126-184, and 551. The first and third disordered regions are predicted correctly by using PSSMP-4 alone (the blue line) as well as PSSMP-4+SSE-DIS (the pink line). However, for the second disordered region, there are only two residues predicted as disordered by using PSSMP-4, while there are twenty residues predicted as disordered by using PSSMP-4+SSE-DIS. It is shown in the Fig. 5(a) that there are five short segments of secondary structure which are predicted as beta strands by Jnet but have been removed after refinement step. This procedure increases the values of the SSE-DIS near this region and then enlarges the disorder propensity of these residues. Similarly, the accuracy of protein disorder prediction in the first 100 residues of u56, a totally disordered protein, is improved explicitly due to the removal of eight short segments of secondary structure, including seven beta-strands and one helix.

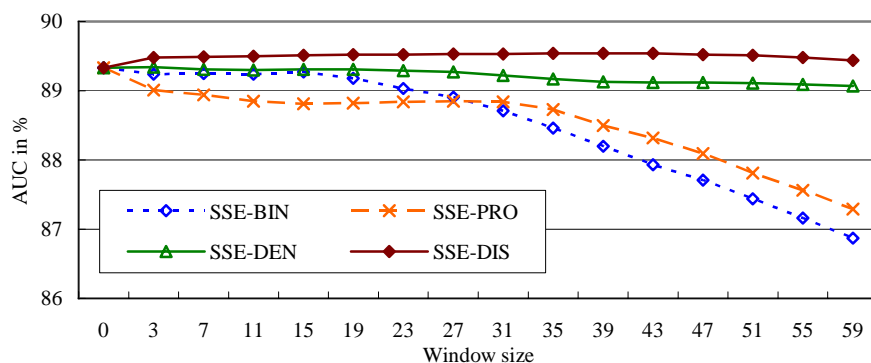
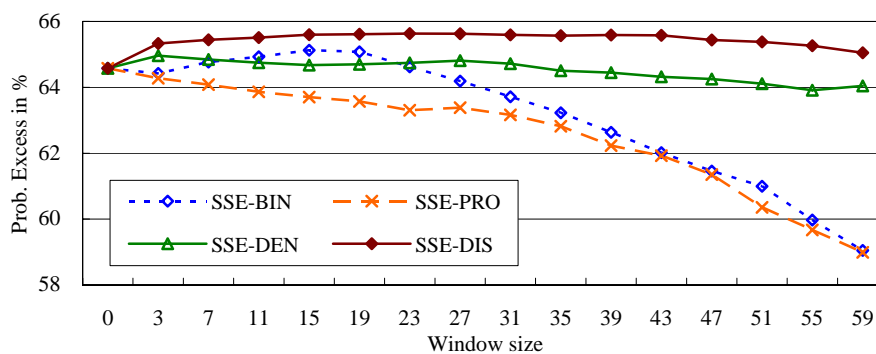
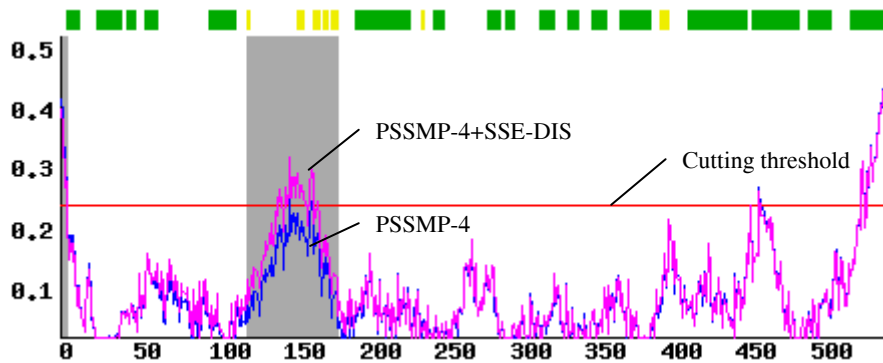

 (a) Comparison based on the measure *AUC*.

 (b) Comparison based on the measure *probability excess*.

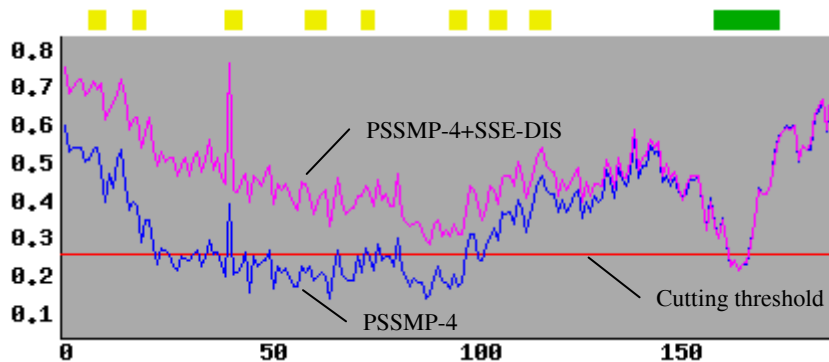
Fig. 4. The results on the testing sets U79 and P80.

At the end of this section, it is of interest to see the statistics of the predicted secondary structures present in ordered regions and different groups of disordered regions. With all of the datasets used in this study, the ratio of each secondary structure type in different groups of protein regions is exhibited in Table 5. In Table 5(a), the statistics are calculated from the original results from Jnet, whereas the statistics in Table 5(b) are calculated from the refined results by removing short SSE segments. Here are some observations. First, the ratios of coils in ordered region in Table 5(a) among all datasets are about 50%, which are lower than in short disordered regions (~90%), middle disordered regions (~80%), and long disordered regions (~65%). When comparing Table 5(b) with Table 5(a), it is observed that many beta strands predicted by Jnet are shorter than five successive residues and are not used in constructing the feature sets. Furthermore, it is attractive that the ratios of long disordered regions are more similar to the ratios of ordered regions than the other groups of disordered regions. This phenomenon might correspond to some long

disordered segments of proteins with specific functions [2, 4]. Since the disorder-to-order transition upon binding occurs in some of long disordered segments of proteins, the segments might comprise secondary structure to stabilize the interfaces or binding domains between a protein and its ligand. This observation needs more investigations and discussions in future studies.



(a) The predicting result (propensity for disorder) of 1PG2



(b) The predicting result (propensity for disorder) of u56

Fig. 5. The comparison of protein disorder prediction with PSSMP-4 and PSSMP-4+SSE-DIS. The figure plots the disorder propensity of a residue (*y-axis*) versus the position of a protein sequence (*x-axis*), where shaded areas are annotated disordered regions (*gray blocks*), the blue line shows the results by using PSSMP-4 alone, the pink line shows the results of using PSSMP-4+SSE-DIS, and the red line presents the cutting threshold of the classification function. The refined secondary structure segments are marked as green blocks (the darker ones), and the segments that contain less than five successive residues and thus have been removed are shown as yellow blocks (the lighter ones).

Table 5. The statistics of secondary structure in ordered regions and three groups of disordered regions with different lengths.

(a) The original results from Jnet

Dataset	Ordered region			Short disordered region(1~4)			Middle disordered region(5~9)			Long disordered region(10~)		
	Helix	Strand	Coil	Helix	Strand	Coil	Helix	Strand	Coil	Helix	Strand	Coil
PDB652	35.2	17.9	46.9	8.7	5.5	85.7	13.8	5.3	80.9	27.7	11.3	61.0
D184	32.0	15.9	52.1	8.2	1.4	90.4	22.0	8.4	69.6	24.5	9.6	65.9
G200	31.3	18.7	49.9									
R80	33.0	17.6	49.5	1.1	5.7	93.1	7.7	5.1	87.2	20.8	9.0	70.2
U79										28.3	9.4	62.3
P80	32.2	17.9	50.0									

(b) The refined results after removal of secondary structure segments with length of less than five successive residues

Dataset	Ordered region			Short disordered region(1~4)			Middle disordered region(5~9)			Long disordered region(10~)		
	Helix	Strand	Coil	Helix	Strand	Coil	Helix	Strand	Coil	Helix	Strand	Coil
PDB652	34.8	14.1	51.1	8.4	3.9	87.7	13.6	3.4	83.0	27.3	8.1	64.6
D184	31.6	12.2	56.2	8.2	0.0	91.8	22.0	6.5	71.5	24.0	6.8	69.2
G200	31.0	14.5	54.4									
R80	32.6	13.7	53.7	0.0	4.6	95.4	7.7	3.8	88.5	20.7	5.8	73.5
U79										27.7	6.5	65.8
P80	31.9	13.3	54.8									

A value in this table is the percentage of a certain type of secondary structure in ordered regions or disordered regions. The empty squares mean that there is no such region in that dataset.

5 Conclusions

In this study we investigate how the predicting power of condensed PSSM features in recognizing protein disorder can be enhanced by secondary structure information. This work compares four kinds of representations in depicting secondary structures detected by secondary structure predictors. The results suggest that the proposed representation achieves more coverage of disordered residues without increasing the false positives obviously. The detected disorder information is expected to be useful in protein structure prediction and functional analysis. In the future, there are several directions for further improving the performance of protein disorder prediction. More predicted information from primary sequences can be merged to enhance the predicting power of the classifiers. On the other hand, incorporating more machine learning skills for handling skewed datasets in this problem also deserves further studies.

References:

1. Dunker, A.K., Obradovic, Z., Romero, P., Kissinger, C., Villafrance, E.: On the importance of being disordered. *PDB Newsletter* (1997) 81:3-5
2. Wright, P.E., Dyson, H.J.: Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* (1999) 293(2):321-331
3. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., Jones, D.T.: Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* (2004) 337:635-645
4. Fink, A.L.: Natively unfolded proteins. *Current Opinion in Structural Biology* (2005) 15:35-41
5. Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albercht, K., Hart, J., Obradovic, Z., Kissinger, C., Villafranca, J.E.: Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput* (1998) 3:473-484
6. Jones, D.T., Ward, J.J.: Prediction of disordered regions in proteins from position specific scoring matrices. *Proteins* (2003) 53:573-578
7. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., Dunker, A.K.: Identifying disordered regions in proteins from amino acid sequence. *Proc. IEEE Int. Conf. Neural Networks.* (1997) 1:90-95
8. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., Garner, E., Guilliot, S., Dunker, A.K.: Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* (1998) 3:437-448
9. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., Dunker, A.K.: Predicting intrinsic disorder from amino acid sequence. *Proteins* (2003) 53:566-572
10. Wotton, J.C., Federhen, S.: Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* (1993) 17:149-163
11. Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J., Dunker, A.K.: Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* (July 2002) 55:104-110
12. Vihinen, M., Torkkila, E., Riikonen, P.: Accuracy of protein flexibility predictions. *Proteins* (1994) 19:141-149
13. Su, C.T., Chen C.Y., Ou, Y.Y.: Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics* (2006) 7:319
14. Liu, J., Rost, B.: NORSp: predictions of long regions without regular secondary structure. *Nucl. Acids Res.* (2003) 31(13):3833-3835
15. Liu, J., Tan, H., Rost, B.: Loopy proteins appear conserved in evolution. *J. Mol. Biol.* (2002) 322:53-64
16. Linding, R., Russell, R.B., Neduva, V., Gibson, T.J.: GlobPlot: exploring protein sequences for globularity and disorder. *Nucl. Acids Res.* (2003) 31:3701-3708
17. Cheng, J., Sweredoski, M.J., Baldi, P.: Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* (2005) 11:213-222
18. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Dunker, A.K.: Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* (2005) Suppl 7:176-182

19. Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., Obradovic, Z.: Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* (2006) 7:208
20. Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M.: RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics Advance Access Published June 9, 2005*
21. Uversky, V.N., Gillespie, J.R., Fink, A.L.: Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* (2000) 41:415-427
22. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. *Nucl. Acids Res.* (2000) 28:235-242
23. Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Lakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., Newton, C.D., Dunker, A.K.: DisProt: a database of protein disorder. *Bioinformatics* (2005) 21:137-140
24. Li, W., Jaroszewski, L., Godzik, A.: Tolerating some redundancy significantly speeds up clustering of large proteins databases. *Bioinformatics* (2002) 18:77-82
25. Cuff, J.A., Barton, G.J.: Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins* (2000) 40:502-511
26. Ou, Y.Y., Chen, C.Y., Oyang, Y.J.: A Novel Radial Basis Function Network Classifier with Centers Set by Hierarchical Clustering, *IJCNN '05. Proceedings.* (2005) 3:1383-1388.
27. Melamud, E., Moult, J.: Evaluation of disorder predictions in CASP5. *Proteins* (2003) 53:561-565
28. Jin, Y., Dunbrack, R.L.: Assessment of disorder predictions in CASP6. *Proteins* (2005) Early View
29. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., Jones, D.T.: The DISOPRED server for the prediction of protein disorder. *Bioinformatics* (2004) 20:2138-2139