Protein Disordered Region Prediction by SVM with Post-Processing

Cheng-Wei Hsieh, Hui-Huang Hsu, Ming-Da Lu
Dept. of Computer Science and Information Engineering
Tamkang University
Taipei, Taiwan
E-Mail: 892190108@s92.tku.edu.tw, h_hsu@mail.tku.edu.tw, 490191771@s90.tku.edu.tw

Abstract

In proteomics, a protein's function is always strongly related to its structure. But, while some parts of a protein have a fixed definite structure, such as α-helix, β-sheet, or coil, other parts are not associated with well-defined conformations. Previously, these so-called disordered regions were not thought to have a specific function of their own. But, recent studies suggest that some disordered regions may have important signaling or regulatory functions. In addition, some critical diseases are strongly related to these disordered regions. Hence, prediction of these disordered regions is essential. In this paper, we try to use the support vector machine (SVM) to predict the disordered regions. Furthermore, this paper emphasizes post processing of the SVM prediction results. Two post-processing algorithms are introduced. These algorithms are used to smooth the primary results by SVM. Different from other studies, these smoothing steps are related to the neighbors' distance to the candidate node. The results show that these algorithms can improve the prediction accuracy further by 1%.

1. Introduction

According to the central dogma of structural biology, the function of a protein is determined by its three-dimensional structure. Proteins can adopt one of three states: fully folded, collapsed, or extended. Nevertheless, parts of proteins fail to self-fold into fixed 3D globular structure. Those which do not have a fixed conformation would be taken as functionless regions in the past views. Even so, recent studies proved that these regions have special functions as signal controlling or regulator roles and this kind of region is defined as "disordered region." Disordered regions often contain short linear peptide motifs, and these regions may cause disordered proteins partial or wholly unstructured. Moreover, various major protein conformational diseases are caused by disordered proteins such as synuclein, Tau, and prion protein.

The disordered regions are stored and displayed in several databases. The protein structure database, protein data bank [2] (PDB), records these structureless regions in its conformation files as remarks 465 which contain each amino acids' position and length. The database of protein disorder, DisProt [3], also collected more than 400 disordered proteins and about 1000 disordered regions.

Various instruments are used to identify the disordered regions, such as nuclear magnetic resonance (NMR) [4] spectroscopy, X-ray crystallography [1] and circular dichroism [5]. Nevertheless, in the experiments, it may take much more time, money and manpower than using a computerized method to predict.

The current trend is using machine learning technologies to discover the disordered regions of proteins. The most used models are neural networks, Bayesian network, and support vector machine [19]. By investigating the protein's sequences and functions relationship, these models can make predictions of disordered regions. However, some limitations of disordered regions prediction appeared while predicting with these models. First, the form of disordered regions various which cause the prediction rate decreased. Next, the sample of disordered regions is quite few. In 2006, besides the tertiary structure predictions and high resolution models, the Protein Structure Prediction Center's [6] CASP7 experiments also held a competition of disordered regions prediction. The predicting target proteins' distribution is that 6% disordered residues and 94% ordered residues. In addition, the situation of disordered regions on N or S terminal is common and this kind of disordered regions is very short which less than 30 residues. Hence, it's not easy to predict these regions.

In this paper we use the SVM to predict the disordered region of a protein. At the prediction part, more information is needed. Some particular amino acid properties are proved to be related to protein disordered region. We use the support vector machine with the above information to predict the disordered region of proteins. And finally, we use several post

0-7695-3109-1/08 $25.00 © 2008 IEEE
DOI 10.1109/CISIS.2008.75

processing algorithms as smoothing functions to improve the prediction accuracy.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the SVM training with smoothing algorithm. Section 4 presents the experimental results. And Section 5 draws the final conclusion.

2. Related work

In the past few years, several machine learning models were developed to solve problems in bioinformatics. Especially for structural biology, by analyzing the protein sequence composition, numerous issues could be solved. For example, protein structure prediction [7], solvent accessibility prediction [8], residue contact prediction [9], protein subcellular localization prediction [10],…etc. In all the above predictions, few post process were taken to improve the prediction accuracy. Smoothing process is the most used method to correct the short prediction errors regions which perform an easy step to smooth the discontinuous prediction results.

In this paper, the disordered proteins' issue follows the above studies' pattern. However, unlike protein structure prediction, normal physical features are not enough for disordered proteins prediction. The additional features of amino acids should also be provided for prediction. In the previous researches of disordered protein predictions, there are several successful works. For instance, the first study for disordered proteins prediction is by Williams et al. (1978) [11]. They noted the abnormally low charge/hydrophobic ratio for the two disordered proteins, and used this special property to predict. Uversky et al. [12] did the same analysis but on much larger set of proteins in 2000. And they made a list of disordered propensity for each amino acid. However, no post process of prediction was taken.

In 2006, Keith Dunker et al. developed the VSL2 [13] disordered region predictor which used an output smooth procedure for its prediction result. The smoothing algorithm is based on calculating the average of raw predictions for neighboring residues within an output window of an odd number length 61 to remove occasional misclassifications. In that work, the prediction accuracy exceeds 85%. Nevertheless, they did not show the smoothing effects. No description was made to explain the smoothing result. The only thing we know is that they use the average result within a sliding window as the central node's reference. In this paper, we modify the average idea to a weighted reference which will be described in Section 3.

Several researches also use smoothing procedure to remove the misclassified node. Such as IUPred [14], which is a web server presented a novel algorithm for predicting such regions from amino acid sequences by estimating their total pairwise interresidue interaction energy. In the end, the IUPred use a simple smoothing step to smooth some misclassified node over a window size of 21. Another predictor is PONDR [15] which measures proteins' complexity and encodes it with statistic information of amino acids. Then, PONDR trains the neural network with these features. The training error of it is about 17% based on the data set they collected. The smoothing procedure of it is by averaging over sliding windows of nine amino acids. In particular, the PONDR wouldn't smooth the first and last four sequence positions from the N- and C-terminal. Since some short disordered regions usually exist in the edges of a sequence.

The DisEMBL [16] predictor introduced a new idea to predict the protein disordered regions by a simple concept of "hot loops" which indicate the structure coils with high temperature factors. The DisEMBL also refer to the PONDR's smoothing algorithm. However, it did not improve the overall performance on their own datasets.

In the above prediction model, the smoothing algorithms are of few variations and few discussions were made. Hence, in this paper we present two other improved smoothing algorithms and make a complete discussion on them.

3. SVM training and smoothing algorithms

3.1. Feature sets

For disordered protein prediction, basic protein sequence composition information is needed. Therefore, in the protein prediction stage, the target proteins' composition should be analyzed. The first kind of composition information is target proteins' occurrence frequencies. We developed a program with encoding and predicting functions. Hence, while the target proteins' primary structure information is inputted, the 20 kinds of amino acids' frequencies would be calculated as the first 20 features.

In advance, only frequencies feature may satisfy protein secondary structure prediction, but not for disordered proteins' prediction. Because the frequencies features only provide the ensemble information, the individual position information is not concerned. Consequently, the position data also have to be added in the prediction program. However, only one target protein's position information is not sufficient. Since proteins would mutate in the nature environment, and we can find several closed family proteins. These proteins may have similar structure and functions. If all this kind of proteins could be collected and make an ensemble analysis, the position information is much more useful than only one protein's position information.

Nevertheless, to describe a set of proteins' position is not an easy work. In this paper, we used the position-specific scoring matrix [17] (PSSM) which is performed by several steps. Firstly, by querying the target protein against the selected database, the most closed family proteins would be searched out. Next, considering the blind position calculation for a set protein is not serviceable, to perform the multiple sequences alignment is needed. After multiple sequence alignment, the position of the selected protein set could be calculated. The PSSM also calculate the log-likelihoods of the substring under a product multinomial distribution which is very useful for disordered protein prediction.

In our program, we calculated the PSSM by using the PSI-BLAST model which is download from the NCBI [18] website. The setting of PSI-BLAST in our program is listed as follows:I. Iterations:3. II. Database: NCBI nonredundant database. III. E-value: 10.

According to the past studies, only sequence's physical information is not sufficient. Some specific amino acids side chain properties should also be included in the prediction system. In this paper, because several amino acids' side chain properties would affect a protein's conformation. In this paper, we have collected several amino acids' side chain properties. They are aliphatic, tiny, small, aromatic, hydropathy index (Kyte-Doolittle), polar, charged, and hydrophobic.

Finally, we use the above features to train the SVM. There are totally 48 features which includes frequencies features (20), PSSM position features (20), and side chain properties (8).

3.2. Support Vector Machine

In this paper, we use the support vector machine to solve the disordered problem. In the previous section, we performed one node with 48 features, and that means we can draw these nodes in a 48 dimension space. So that if we could find a hyperplane to separate these nodes into ordered and disordered classes, the nodes are training successfully. The SVM also could project these nodes into a higher dimension, and in that newer space another hyperplane could always be found. Moreover, SVM is based on the SV (support vector) learning. That means the SVM would not always compare the prediction target to all the existing training nodes. In contrast, the SVM selects several nodes as its SVs, and use these SVs to judge the prediction target to be ordered or disordered.

In the testing stage, the SVM model would use the SVs to do the prediction. And also, these SVs would locate on the maximum margin of separation. The SVM is also rated an excellent classifier in practical applications. The SVM can handle more complex nonlinear problems. Fig. 1 demonstrates the maximum margin between two classes which are separated by the hyperplane in the SVM model. The H_1 and H_2 are the boundaries. And the nodes which are located on these two lines would be support vectors.

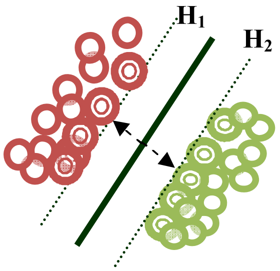

Figure 1. The SVM could find out the maximum margin and use the SVs to predict the prediction targets. The line H1 and H2 are located on these SVs. (doubled circles are SVs).

The "maximum margin" or "optimal separation" idea comes true with several steps. Assume that there are l nodes and each node contains two parts (x_i , y_i) where x_i represents the node's feature vector, and the y_i shows its class (1 or -1). H_1 and H_2 are the two boundaries of the "maximum margin" in (1)

$$\begin{aligned} &\mathrm{H}_1 : \omega \cdot x_i - b \geq +1 \text{ for } y_\mathrm{i} = +1 \\ &\mathrm{H}_2 : \omega \cdot x_i - b \leq -1 \text{ for } y_\mathrm{i} = -1 \end{aligned} \qquad (1)$$

The SVM uses the Lagrangian to solve this constrained optimization problem. Eq. (2) demonstrates it. The α_i is the Lagrangian multiplier. To adjust each node's α value for optimization is the main task. In the end, those nodes with nonzero α_i are the support vectors. To identify the class of each node, a decision function is used.

Maximize

$$\sum_{i=1}^{l} \alpha_i - \tfrac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

Subject to

$$\alpha_i \geq 0, \ i = 1, \ldots, l, \ \ and \ \sum_{i=1}^{l} \alpha_i y_i = 0 \qquad (2)$$

Decision function

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i \cdot (x \cdot x_i) + b\right)$$

3.3. Datasets

Our ordered and disordered sequence is collected from the DisProt [3] and PDB [2] database. The proteins in DisProt are all with disordered regions. The number of ordered regions is very few. If we use the SVM with these disordered data, this may cause the unbalanced training problem. Hence we also collect proteins from the PDB which contains much more ordered regions. Those data selected from DisProt are taken as positive training data, and the negative training data are derived from PDB_Select_25 [17] which is a nonredundant dataset of the Protein Data Bank (PDB). Finally, 119 protein sequences are collected and there are totally 21676 residues.

3.4. Smoothing Algorithm

After predicting by SVM model, the result would be saved in our program. In Fig. 2, one can easily find that some predicted disordered region is not continues. According to the disordered regions' characteristic, it is not possible for only one or two residues to form a disordered region except for the edge regions. Hence, we develop two smoothing algorithms to recover these discontinued regions.

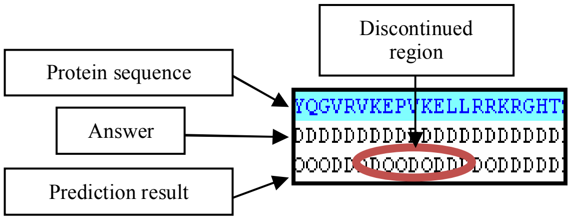

Figure 2. The prediction result contains discontinued regions which should be smoothed.

Algorithm 1: the disordered-density algorithm

Steps: 1. Calculate the neighbors' disordered frequencies of the central node. 2. Divide the above frequencies value by the window size. 3. If the quotient is great than the threshold, then the central node should be set to disordered, and vice versa.

The above algorithm is based on the idea that discontinued disordered/ordered regions should be smoothed. This idea is realized by inspecting the neighborhood's disordered/ordered residues distribution. In other words, by using the sliding window with a fixed size, the central residue's disordered/ordered state should reference to its neighbors' disordered/ordered density. If the disordered density is greater than the threshold which is predefined by the user (default 0.5), the central residue of the current window should change its states from ordered to disordered, and vice versa. Fig. 3 shows the smoothing process. Because the candidate node in Fig. 3 is 0 (ordered) while all its neighbors are 1 (disordered), the calculation should be (number of disordered) / (window size). In this example, the trend score of disorder is 8/9 (win size=9) which is greater than 0.5 (default: 0.5). Therefore, the central node should be set to 1 (ordered).

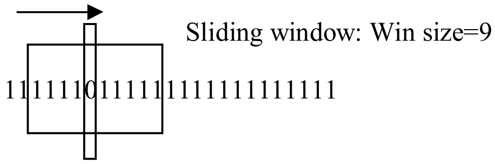

Figure 3. The central node of the sliding window would be changed from 0 to 1, because of the vote result within the window. (window size=9, number of 1=8, number of 0=1)

Algorithm 2: the disordered-distance algorithm

Steps: 1. Calculate disordered nodes' distance value from the central node within the sliding window. 2. The voting weight of each disordered node is one over its distance value. 3. Sum up the voting weights in Step 2. If the value is great than the threshold, then the central node should be set to disordered, and vice versa.

In the disordered-density algorithm, the residues within the windows have rights to vote the central residue's state. When the majority is on the disordered side, the central residue should be disordered. Nevertheless, it is not fair for the residues near the center. Each residue within the window should not share a equal voting right. The residue which is farer away from the central residue should have a smaller voting weight instead of taking the same weight as the residues near the central residue. Therefore, this algorithm is named disordered-distance algorithm that the voting weight is inversely related to its distance to the central residue. Fig. 4 gives an example.

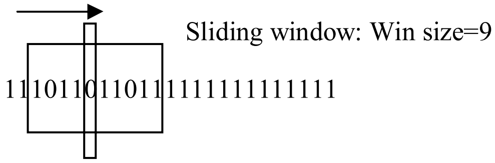

Figure 4. The vote result of 1 is great than 0.5, hence the central node should be set to 1. The calculation would multiple each node's distance to the central node.

4. Experimental results

4.1. Program & Experiment environment

We use the C# to develop our system, and it is run on a PC with 1.83 core duo CPU and 2048 MB memory. The software that we combined in our system is LIBSVM [19] for SVM implementation.

4.2. Results

Our SVM program is trained with the above 48 features which contains frequencies features (20), PSSM position features (20), and side chain properties (8). At first, we trained and predicted the target dataset with the following model setting:

Kernel: RBF kernel

Penalty(C): 10

Accuracy calculation: five-fold cross-validation

The prediction result is 84.66%. Next, the different smoothing algorithms are performed. At first, we apply the disordered-density algorithm to the previous prediction. The table 1 lists different setting of smoothing window size and the smoothing result.

Table 1. Disordered-density smoothing results
(Threshold: 0.5 Total residues: 21676)

Window size	Accuracy	Numbers of change	
		Ordered to Disordered	Disordered to Ordered
7	84.74%	794	1590
15	84.72%	166	331
21	84.73%	974	2000
31	84.76%	1240	2505
41	84.71%	719	1478

Then, we changed the threshold of disordered-density smoothing algorithm. The result is presented in table 2.

Table 2. Disordered-density smoothing results
(Window size: 15 Total residues: 21676)

Threshold	Accuracy	Numbers of change	
		Ordered to Disordered	Disordered to Ordered
0.1	84.90%	1690	3217
0.3	84.99%	1892	3636
0.5	84.73%	974	2000
0.7	84.84%	2140	4212
0.9	84.57%	2454	5123

Next, we use the second smoothing algorithm, "disordered-distance." Similarly, we change the window size first. The results show in the table 3.

Table 3. Disordered-distance smoothing results
(Threshold: 0.5 Total residues: 21676)

Window size	Accuracy	Numbers of change	
		Ordered to Disordered	Disordered to Ordered
7	84.65%	2747	5643
15	84.68%	3356	6517
21	84.70%	4153	7557
31	84.69%	5165	8808
41	84.65%	6354	10184

Next, we changed the threshold of disordered-distance smoothing algorithm. The result is presented in Table 4.

Table 4. Disordered-distance smoothing results
(Window size: 15 Total residues: 21676)

Threshold	Accuracy	Numbers of change	
		Ordered to Disordered	Disordered to Ordered
0.1	85.27%	3024	4364
0.3	85.31%	609	887
0.5	85.23%	1218	1761
0.7	85.16%	1819	2630
0.9	85.11%	2418	3493

Finally, we summarized the above parameters' setting, and found the optimal threshold and window size should be set to 0.3 and 31. The Table 5 shows the experiment result.

Table 5. Optimal parameter setting results
(Threshold: 0.3 Window size: 31 Total residues: 21676)

Smoothing Algorithms	Accuracy	Numbers of change	
		Ordered to Disordered	Disordered to Ordered
Disordered-density	85.76%	331	666
Disordered-distance	85.29%	1773	2582
Without smoothing	84.66%	-	-

4.3. Discussion

In the smoothing steps, two algorithms are performed with different parameter settings. In the first disordered-density smoothing algorithm, we can observe that no matter how we change the window size, the accuracy is still around 84.7%. It is only better than prediction without smoothing 0.1%. Then we change the threshold from 0.1 to 0.9, and the result is better than only change the window size. The best setting of the disordered-density smoothing algorithm is that the window size set to 31 and threshold set to 0.3. However, this is not a satisfied result. Therefore, we develop a second smoothing algorithm, disordered-distance smoothing.

As the parameter setting with disordered-density smoothing algorithm, we changed the window size of it in the order of 7, 15, 21, 31and 41. The result shows that it works the same as the previous disordered-distance smoothing algorithm. Nevertheless, when we change the threshold value from 0.1 to 0.9, the accuracy increased to 85%. The best parameter setting of disordered- distance smoothing algorithm is that window size set to 21 and threshold set to 0.3.

However, in the disordered-distance smoothing experiment with different window size, we can observe that the accuracy of window size 21 and 31 is almost the same. Therefore, we integrate the two algorithm's best parameter setting. We found that the best parameter setting for these two algorithms is the same that window size set to 31 and threshold set to 0.3. The Table 5 displays the optimal parameter setting experiment results. Specially, the disordered-density algorithm improves accuracy from previous experiment result of 84.76% to 85.76%.

After these smoothing steps, we can improve the prediction accuracies by 1%. This is benefit to the final prediction result. In the disordered protein prediction studies, it's not a easy work to improve the prediction result. For instance, the VSL2 predictor of DisProt only improves 2.1% accuracy since it uses 22 additional features derived from the computationally expensive PSI-BLAST profiles (PSSM). Thus it can be seen that our work profit the prediction result.

5. Conclusion

In this paper, we use two different smoothing algorithms to improve the prediction accuracy of SVM disordered region prediction. The SVM is trained with 48 features that we combined in our program. The accuracy of it is 84.66%. By performing the smoothing algorithm, almost 1% accuracy is improved. This could be helpful for the discontinued prediction result for different prediction models. In addition, the smoothing steps can be processed in linear time. No additional computationally expensive steps are taken. In the future, we will also apply structure features to improve the smoothing result. For example, the protein secondary structure information within the sliding node to smooth the central node.

Acknowledgement

This research is supported by the National Science Council under grant number NSC 96-2221-E-032-051-MY2.

References

[1] C.R. Cantor, *Principles of Protein X-Ray Crystallography*, Springer Verlag, New York, 1996, pp. 1-341.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E Bourne, "The Protein Data Bank," *Nucleic Acids Resource*, Vol. 28, 2000, pp. 235-242.

[3] S, Vucetic, Z. Obradovic, V. Vacic, P. Radivojac, K. Peng, L.M. Iakoucheva, M.S. Cortese, J.D. Lawson, C.J. Brown, J.G. Sikes, C.D. Newton, and A.K. Dunker, "DisProt: A Database of Protein Disorder," *Bioinformatics*, Vol. 21, 2005, pp. 137-140.

[4] C. Bracken, "NMR spin relaxation methods for characterization of disorder and folding in proteins," *J Mol Graph Model*, Vol. 19, 2001, pp. 3-12.

[5] G.D. Fasman, *Circular Dichroism and the Conformational Analysis of Biomolecules*, Plenum Press, New York, 1996.

[6] Protein Structure Prediction Center, http://predictioncenter.gc.ucdavis.edu/ (last accessed August 3, 2007).

[7] C. Bracken, L.M. Iakoucheva, P.R. Romero, and A.K. Dunker, "Combining prediction, computation and experiment for the characterization of protein disorder," *Curr. Opin. Struct. Biol*, Vol. 14, 2004, pp. 570-576.

[8] Jaehyun Sim, Seung-Yeon Kim and Julian Lee, "Prediction of protein solvent accessibility using fuzzy k – nearest neighbor method," *Bioinformatics*, Vol. 21, 2005, pp. 2844-2849.

[9] J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *Bioinformatics*, Vol. 8, 2007.

[10] J.L. Gardy, M.R. Laird, F. Chen, S. Rey, C.J. Walsh, M. Ester, and F.S.L. Brinkman, "Psortb V.2.0: Expanded Prediction Of Bacterial Protein Subcellular Localization And Insights Gained From Comparative Proteome Analysis," *Bioinformatics*, Vol. 21, 2005, pp. 617-623.

[11] R.J. Williams, "The conformational mobility of proteins and its functional significance," *Biochem. Soc. Trans*, Vol. 6, 1978, pp. 1123-1126.

[12] V.N. Uversky, J.R. Gillespie, and A.L. Fink, "Why are "natively unfolded" proteins unstructured under physiologic conditions?" *Proteins*, Vol. 41, 2000, pp. 415-27.

[13] K. Peng, P. Radovojac, S. Vucetic, A.K. Dunker and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *Bioinformatics*, Vol. 7, 2006.

[14] Zsuzsanna Dosztányi, Veronika Csizmók, Péter Tompa and István Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content," Bioinformatics, Vol.21, 2005, pp. 3433-3434.

[15] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, and A.K. Dunker, "Sequence complexity of disordered protein," *Proteins*, Vol. 42, 2001, pp. 38-48.

[16] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson and R.B. Russell, "Protein disorder prediction: implications for structural proteomics," *Structure*, Vol. 11, 2003.

[17] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, Vol. 215, 1990, pp. 403-410.

[18] National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/ (last accessed August 3, 2007).

[19] LIBSVM - A Library for Support Vector Machines,http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (last accessed August 3, 2007).