Sequence-Based Prediction of Protein-Protein Binding Residues in Alpha-Helical Membrane Proteins

Feng Xiao and Hong-Bin Shen[*]

Institute of Image Processing & Pattern Recognition, Shanghai Jiao Tong University,
800 Dongchuan Road, Shanghai, 200240, China
hbshen@sjtu.edu.cn

Abstract. A specific number of chains form alpha-helical membrane protein complexes in order to realize the biochemical function, i.e. as gateways to decide whether specific substances can be transported across the membrane or not. However, few structures of membrane proteins have been solved. The knowledge of protein-protein binding residues can help biologists figure out how the function works and solve the 3D structures.

We present a novel, sequence-based method to predict protein-protein binding residues from primary protein sequences by machine learning classifiers. We use a support vector regression model to predict relative solvent accessibility by features based on sequences, including position specific scoring matrix, conserved score, z-coordinate prediction, second structure prediction, physical parameter and sequence length. Afterwards, combining features mentioned above with the predicted solvent accessibility, we use ensemble support vector machines to predict protein-protein binding residues. To the best of our knowledge, there is no method to predict protein-protein binding residues in alpha-helical membrane proteins. Our method outperforms MAdaBoost successfully used in predicting protein-ligand binding residues and random forest used in protein-protein binding residues from surface residues. We also assess the importance of each individual type of features. PSSM profile and conserved score are shown to be more effective to predict protein-protein binding residues in alpha-helical membrane proteins.

Keywords: Relative solvent accessibility, binding residues, alpha-helical membrane proteins.

1 Introduction

Alpha-helical transmembrane proteins (TMPs) are mostly present in the inner membranes of bacterial cells and the plasma membrane of eukaryotes. They constitute the majority of all TMPs, especially in humans. They are estimated to account for 27% of all proteins [1]. Moreover alpha-helical TMPs are often regarded as the important drug targets, i.e. G protein-coupled receptor (GPCR). Hence many efforts have been made to solve the three-dimensional structures and to understand the functions of

[*] Corresponding author.

S. Li et al. (Eds.): CCPR 2014, Part II, CCIS 484, pp. 419–427, 2014.

TMPs. However little progress has been made during past two decades, from statistical data in PDBTM database [2] by the end of 2014/05/16, there were only 2131 solved TMP structures of which 1840 are alpha-helical and 283 beta-barrel TMPs. Because of this difficulty, computational methods (template-based or ab initio) have been developed for single chain structure prediction such as Membrane-Rosetta [3] and FILM3 [4] and for several easy multi-chain complexes such as BCL::MP-Fold [5]. Accurate protein-protein binding residue prediction in membrane can help membrane complex structure prediction.

Although there are many computational methods, generally speaking, structure-based, sequenced-based methods and hybrid methods, for predicting protein-ligand binding site [6][7], only little progress have been made in protein-protein binding residue prediction in TMPs. To our knowledge the existing method proposed by Andrew J Bordner employed a Random Forest with sequence-based and structure-based features to predict the binding residues from surface residues in membrane proteins and reported the AUC of 0.75 [8]. The definition of binding residues, also using structure information, is that the surface residue has contact with another chain in the complex structure (< 4 Å non-H atom separation). The definition of surface residues include: (1) relative solvent accessibility surface area (RSA) ≥ 0.2, (2) within the hydrophobic core of the membrane, in other words, the absolute number of the z-coordinates predicted from the real structures are no more than $15\overset{\circ}{A}$. All the surface residues are included in the training dataset.

With the development of machine learning methods, there have been many sequence-based methods using artificial neural networks (ANNS) and support vector machines (SVMs) to predict membrane protein structure information, i.e. relative solvent accessibility (RSA) [9][10]. Although structure-based method has proven effective in protein-protein binding residues prediction, there still exists several problems needed to solve:

First, by the end of 2014/05/16, there were 101245 structures in PDB database, of which 2131 are TMPs and 1840 are alpha-helical TMPs. However the number of sequences grow rapidly contrast to the real structure considering the homology influence. So given a sequence of membrane proteins, if its real structure is not available, this structure-based method is not able to do the prediction.

Second, the existing method only predicts binding residues from surface residues in membrane proteins, in other words, before predicting protein-protein binding residues in TMPs we need to know whether the residues are surface ones or not.

In view of the above-mentioned two problems, we proposed a sequence-based protein-protein binding residue predictor for entire membrane proteins. First, we constructed a relative solvent accessibility predictor for TMP complexes with support vector regression (SVR) models. Second, protein conserved matrix (both PSSM and rate4site), predicted secondary structure matrix, predicted z-coordinate matrix, and predicted relative solvent accessibility matrix consist of the final feature set; considering the imbalance of positive (unbinding residues) and negative (binding residues) samples in our experiments, under-sampling technique was used to balance the dataset, afterwards, ensemble SVM was chosen to train the final model.

2 Material and Methods

2.1 Benchmark Dataset of Alpha-Helical Membrane Protein Complex Structures

In order to predict solvent accessibility of both single- or multiple-chain in membrane proteins, we used the same dataset as originally used in MPRAP [10]. In this dataset, the sequence identity cutoff was set to 20% and length cutoff 0.9, fragments, low-resolution structures and structures with second structure or membrane boundary problems were excluded. Thus there are 52 complexes including 80 chains in the final dataset. In order to avoid high homology in different folds, chains from the same super family were put in the same fold. The dataset was finally divided into 5 folds in advance. This dataset was also used as a benchmark dataset to predict protein-protein binding residues. It is available at http://mprap.cbr.su.se/dataset_MPRAP_feb2010.fa. All the results showed in this paper are calculated after 5-fold cross-validation.

2.2 Calculation of Relative Solvent Accessibility

In this study, the RSA of each residue was calculated by Naccess 2.1.1 [11]. In our experiments we set the probe size $1.4\overset{\circ}{A}$ for that 2.0 did not perform well and the combination would bring error when calculating RSA intramembrane and outside separately. During the calculation for RSA of complexes, all chains in the complex were included.

2.3 Definition of Binding Residues

In Andrew J Bordner's work, the definition of binding residues was that (1) relative solvent accessibility surface area (SASA) ≥ 0.2 and the residues lied in the membrane core; (2) residues in one chain had contact with another chain in the complex structure. Contacts were defined that the atom-atom (except H-atom) distance between different residues was less than $4\overset{\circ}{A}$. However Arne Elofsson's definition was that (1) relative solvent accessibility surface area (rSASA) was lower than a certain cutoff in protein complexes; (2) SASA was not lower than that certain cutoff in single-chain protein. The cutoff was set to 0.25 in Arne's work.

In our work, we define that residues in one chain contact with another chain in the complex structure are binding residues.

2.4 Feature Extraction

In previous work, several common features have been used successfully in the field of either solvent accessibility prediction or binding residue prediction. In this paper, we extracted 6 types of sequence-based features, including position specific scoring matrix (PSSM), conserved score by rate4site (R4S), z coordinate predictions, predicted secondary structure (SS) information, representative physical parameters (PP) and sequence length.

Position Specific Scoring Matrix. (PSSM) is generated by PSI-BLAST [12] to search against the UniRef90 database with 3 iterations and an E-value cutoff of 0.00001. All elements in PSSM are normalized by the following logistic function.

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

where x is the original score.

Conserved Score is generated by Rate4Site [13] from the multiple sequence alignment (MSA). According to Arne Elofsson et.al's work, exposed residues evolved slowly and were considered to be conserved, and buried residues evolved rapidly and were considered to be active. Thus a conclusion that the relative substitution rate was almost linearly related to the solvent accessibility in membrane protein complexes was obtained. The conserved score of each residue is normalized by subtracting the average score and dividing by the standard deviation.

Z-coordinate Prediction is generated by Zpred [14]. Zpred predicted the absolute z-coordinate and few predicted numbers were no more than 25 $\overset{o}{A}$ based on our statistics, so we normalized the predictions by dividing 25. Then the normalized numbers were added into the final feature set.

Second Structure Prediction is generated by PSIPRED [15]. Each residue in the sequence got the possibilities of three classes (coil (C), helix (H) and strand (E)). In our experiments, we take the three possibilities directly as the input features rather than converting to binary numbers.

Representative Physical Parameters and Sequence Length are residue-based features from statistical data. Representative physical parameters included a steric parameter, hydrophobicity, volume, polarity, isoelectric point, helix probability, strand probability, average accessible surface area (ASA), charge, acidity, occurrence, and average mass of twenty common amino acids. In addition to features mentioned above, sequence length was added in the feature set.

2.5 Using Sliding Windows to Include Neighborhood Information into Feature Set

Previous studies have indicated that the use of sliding windows can include more useful information and thus improve the prediction accuracy, i.e. second structure and relative solvent accessibility prediction [10]. In this study, we used sliding windows to cover neighborhood information in 4 types of features: (1) position specific scoring matrix (PSSM), (2) evolution rate, (3) z-coordinate prediction, (4) predicted second structure information. In our work, we found that the window size set to 9 seemed to be optimal.

2.6 Prediction of Solvent Accessibility

In this section, in order to predict relative solvent accessibility a simple SVR model was used with the combination of 6 types of features. These features include: (1) PSSM; (2) Second structure prediction; (3) conserved score calculated by Rate4Site; (4) z-coordinate prediction calculated by Zpred; (5) representative physical parameters; (6) sequence length. From (1) to (4), these 4 types of features are extracted using a sliding window of length 9. (5) and (6) these two are residue-based that the sliding window is not necessary.

Afterwards, the predicted real value-RSA was also added into the binding residue-specific feature set using a sliding window, the length is set to 9.

2.7 Ensemble Classifier Approach to Predict Protein-Protein Binding Residues with Support Vector Machines

In order to predict binding residues, we used ensemble classifiers with support vector machines to predict membrane protein-protein binding residues from sequence information only. In our training dataset, 4629 binding residues were defined as negative samples and 16789 unbinding residues as positive samples. The ratio of positive and negative samples is about 3.6. To balance the dataset the under-sampling approach is used that positive samples with the same number of negative samples were randomly selected. Afterwards, in order to reduce the impact of under-sampling, we introduced ensemble classifiers. L different models were generated using SVMs followed by each samples, the prediction results are probabilities rather than binaries. Thus we added all the L predictions together and then divided by L.

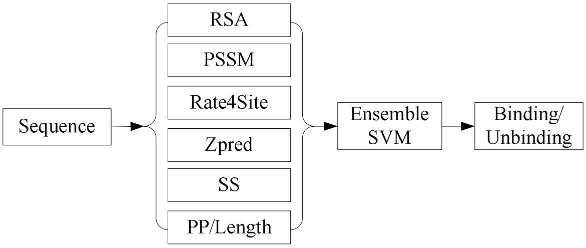

Fig. 1. The flowchart of protein-protein binding residue prediction

Figure 1 illustrates the flowchart. There are 7types of features used in training dataset: (1) PSSM; (2) second structure prediction; (3) conserved score; (4) solvent accessibility prediction; (5) Z-coordinate prediction; (6) physical parameters; (7) sequence length. For a given sequence, we combined all the features together as the input to the ensemble SVM models. Finally, we got the predicted probability of binding and unbinding. In the training procedure, a certain cutoff was select to maximize the Matthems correlation coefficient. By using that certain threshold, real-valued probability was transformed to binary states (binding and unbinding).

3 Results and Discussions

3.1 Performance of Relative Solvent Accessibility Prediction

In section 2, the predicted RSA was added to the feature set so that the performance of RSA prediction is very important. The mean absolute error (MAE) and Pearson correlation coefficient (CC) of our predictor outperforms MPRAP.

Table 1. Performance of different input types of features

Features	MAE	CC	MCC
MPRAP	18.202	0.583	0.470
MPRAP+SS	18.159	0.589	0.467
MPRAP+SS+length	18.136	0.593	0.472
MPRAP+SS+length+para	18.013	**0.599**	**0.478**

In Table 1, MPRAP represents three features (PSSM rate4site and Zpred) used in that method, the results (MAE: 18.202 and CC: 0.583) run in local is comparable to that reported in literature (MAE: 18.4 and CC: 0.58). Matthems correlation coefficient (MCC) was calculated by transforming the predicted real values into binary states using a cutoff. The cutoffs were optimized to maximize MCC. We found that our predictor only added there features (SS length and parameters) improved MAE CC and MCC by 0.187 0.016 and 0.008 respectively when compared with MPRAP.

3.2 Comparison with Other Methods

To the best of our knowledge, no work has been done to predict protein-protein binding residues in a-helical membrane proteins. Andrew J Bordner used random forests (RFs) to predict binding sites from surface residues in both a-helical and b-barrel membrane proteins. In order to do comparison, we use different methods, including (1) the MAdaBoost method used in TargetS to predict protein-ligand binding sites, (2) a simple SVM model to validate the effectiveness of the under-sampling method, (3) random forests used in Bordner's method.

In Table 2, specificity sensitivity accuracy and MCC are threshold-based, so we select the optimal threshold to maximize the MCC value. AUC is used to examine the predicted probabilities. Our method is shown to outperform other methods that it improves the AUC and MCC by 0.006 and 0.006 when comparing with the MAdaBoost method used in TargetS. Also our method achieves the best sensitivity and accuracy among all the methods. However the specificity performs not very well comparing with simple and random forests. MAdaBoost performs very in protein-ligand binding sites prediction. In our experiments, this method performs slightly worse than our method. In MAdaBoost, the base classifier is SVM. In order to evaluate the error of each base classifier, an independent dataset extracted from the training dataset is set as the evaluation dataset. So the number of samples used to train the model in MAdaBoost is less than our predictor. Considering that the number of samples is not very large, MAdaBoost is expected to achieve not very accurate result. Simple SVM and random forest achieve the high specificity and low sensitivity. These two methods trained models in the original dataset directly. The imbalanced dataset could affect the performance of the prediction. The predictions prefer to the major class. In order to maximize the MCC value, the threshold is adjusted near the major class. This explains the high specificity and low sensitivity. Figure 2 shows the receiver operating characteristic curves of these four methods.

Table 2. Comparison between different methods

Methods	AUC	SPE	SEN	ACC	MCC
Ensemble SVM	0.705	0.450	0.812	0.734	0.251
MAdaBoost	0.699	0.467	0.794	0.723	0.245
Simple SVM	0.668	0.690	0.573	0.598	0.217
Random Forest	0.617	0.597	0.572	0.577	0.139

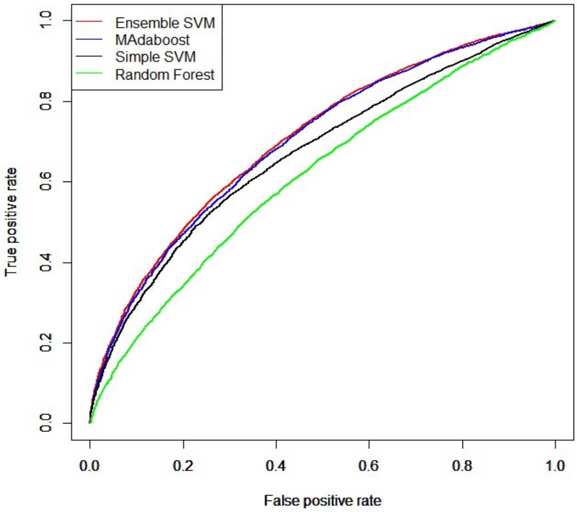

Fig. 2. ROC curves for different methods

3.3 Effectiveness of Individual Types of Input Features

In this section, we will describe the effectiveness of different types of input features in our experiments. Table 3 shows the performance of different inputs and the corresponding ROC curves are shown in Figure 3. Among the listed 5 types of features, PSSM outperforms others for that it achieves the highest AUC and MCC values, followed by Rate4Site and SS. It is expected that PSSM has proved to be the most important feature to predict protein-ligand binding residues and solvent accessibility and so on by using sequence-based methods. Rate4Site is used to calculate conserved score and generated from the multiple sequence alignment (MSA), it performs slightly worse than PSSM and better than SS. These three features make a majority of contribution to the final prediction. RSA is obtained by our predictor directly, achieves the AUC value of 0.52 and MCC value of 0.033. By adding RSA prediction and z-coordinate prediction into feature dataset, the final results improve a little.

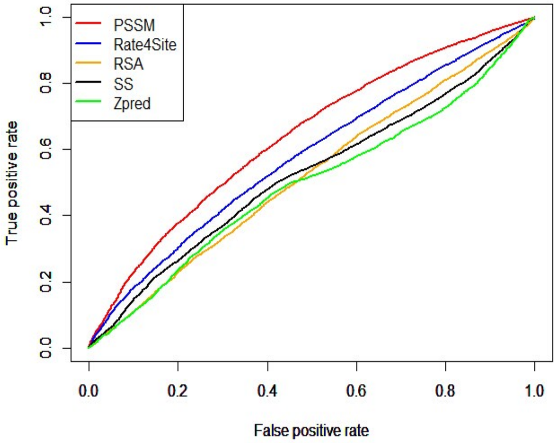

Fig. 3. ROC curves for individual types of input features

Table 3. Performance for individual type of features

Features	AUC	SPE	SEN	ACC	MCC
PSSM	0.641	0.478	0.719	0.667	0.174
Rate4Site	0.580	0.680	0.439	0.491	0.100
RSA	0.520	0.387	0.651	0.594	0.033
SS	0.525	0.629	0.451	0.490	0.067
Zpred	0.497	0.695	0.360	0.432	0.048

4 Conclusion

We developed a novel sequence-based predictor to predict protein-protein binding residues in a-helical membrane proteins. Our predictor used under-sampling methods to balance the dataset and ensemble SVMs to get the final model and outperforms other methods. We hope that our predictor would have application in guiding the experiments of solving 3-dimensional structures in membrane protein complexes.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (No. 61222306, 91130033, 61175024), Shanghai Science and Technology Commission (No. 11JC1404800), a Foundation for the Author of National

Excellent Doctoral Dissertation of PR China (No. 201048), and Program for New Century Excellent Talents in University (NCET-11-0330).

References

1. Almén, M.S., Nordström, K.J.V., Fredriksson, R., et al.: Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biology 7(1), 50 (2009)
2. Kozma, D., Simon, I., Tusnády, G.E.: PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Research 41(D1), D524–D529 (2013)
3. Yarov-Yarovoy, V., Schonbrun, J., Baker, D.: Multipass membrane protein structure prediction using Rosetta. Proteins: Structure, Function, and Bioinformatics 62(4), 1010–1025 (2006)
4. Nugent, T., Jones, D.T.: Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis [J]. Proceedings of the National Academy of Sciences 109(24), E1540–E1547 (2012)
5. Weiner, B.E., Woetzel, N., Karakaş, M., et al.: BCL: MP-fold: folding membrane proteins through assembly of transmembrane helices. Structure 21(7), 1107–1117 (2013)
6. Chen, K., Mizianty, M.J., Kurgan, L.: ATPsite: sequence-based prediction of ATP-binding residues. Proteome Sci. 9(suppl. 1), S4 (2011)
7. Yu, D., Hu, J., Yang, J., et al.: Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering (2013)
8. Bordner, A.J.: Predicting protein-protein binding sites in membrane proteins. BMC Bioinformatics 10(1), 312 (2009)
9. Adamczak, R., Porollo, A., Meller, J.: Accurate prediction of solvent accessibility using neural networks–based regression. Proteins: Structure, Function, and Bioinformatics 56(4), 753–767 (2004)
10. Illergård, K., Callegari, S., Elofsson, A.: MPRAP: An accessibility predictor for a-helical transmem-brane proteins that performs well inside and outside the membrane. BMC Bioinformatics 11(1), 333 (2010)
11. Hubbard, S.J.T.J.: NACCESS, Computer program. Department of Biochemistry and Molecular Biology 1, 1–2 (1993), http://wolf.bi.umist.ac.uk/unix/naccess.html
12. McGuffin, L.J., Bryson, K., Jones, D.T.: PSIPRED protein structure prediction server. Bioinformatics 16, 404–405 (2000)
13. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997)
14. Granseth, E., Viklund, H., Elofsson, A.: ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. Bioinformatics 22(14), e191-e196 (2006)
15. Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T.: Comparison of site-specific rate-inference methods: Bayesian methods are superior. Mol. Biol. Evol. 21, 1781–1791 (2004)