Structure-based prediction of protein-peptide binding regions using Random Forest

Ghazaleh Taherzadeh[1], Yaoqi Zhou[1, 2], Alan Wee-Chung Liew[1] and Yuedong Yang[1, 2 *]
[1]School of Information and Communication Technology, Griffith University, Parklands Drive, Southport, Queensland 4215, Australia, [2]Institue for Glycomics, Griffith University, Parklands Drive, Southport, Queens-land 4215, Australia.
*To whom correspondence should be addressed: yuedong.yang@griffith.edu.au

1. INTRODUCTION

Protein-peptide interactions are one of the most important biological interactions and play crucial role in many diseases including cancer (1). Therefore, knowledge of these interactions provides invaluable insights into all cellular processes, functional mechanisms, and drug discovery (2). Protein-peptide interactions can be analyzed by studying the structures of protein-peptide complexes. Thus, predicting peptide-binding sites computationally will be useful to increase efficiency and cost effectiveness of experimental studies. Here, we established a machine learning method called SPRINT-Str (Structure-based prediction of protein-Peptide Residue-level Interaction) to use structural information for predicting protein-peptide binding regions.

2. METHOD

The initial dataset of protein-peptide complex structures was obtained from the BioLip (3). After removing redundant chains with sequence identity more than 30%, the final dataset consists of 1,242 protein-peptide complexes, which is divided into training set and independent test set containing 1,116 and 125 proteins, respectively. Several structural-based features and the most discriminative sequence-based features reported in the SPRINT (4) were extracted and integrated by a Random Forest (RF) classifier (5) for prediction of binding residues. Predicted binding residues were employed to infer binding sites using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (6). The largest binding site of each protein was then selected by setting some restrictions on the predicted binding sites.

3- RESULT AND DISCUSSION

SPRINT-Str achieves robust and consistent results for predictions of protein-peptide binding regions in terms of residues and sites. It achieves consistent Matthews' Correlation Coefficient (MCC) of 0.27 and 0.293 as well as Area Under the Curve (AUC) of 0.775 and 0.782 for 10-fold cross validation and independent test set, respectively. The clustering of the predicted binding residues lead to a dramatically accurate prediction of the binding site. The obtained accuracy of 46.4% in the test set is more than >1.7 times higher than other methods. The method was found to improve over sequence-based and other structure-based techniques (Table 1). The application to

Table 1. Comparison of different methods on the test set. [a]Binding residue prediction [b] Binding site prediction

Methods	[a]MCC	[a]Accuracy	[a]Sensitivity	[a]Specificity	[b]Accuracy
SPRINT-Str	0.293	0.941	0.24	0.98	0.464
SPRINT-Seq	0.198	0.92	0.21	0.96	--
Peptimap	0.26	0.92	0.32	0.95	0.264
Pepsite	0.198	0.929	0.18	0.97	0.112
PinUp	0.13	0.89	0.22	0.90	0.18
VisGrid	0.145	0.89	0.24	0.928	0.256

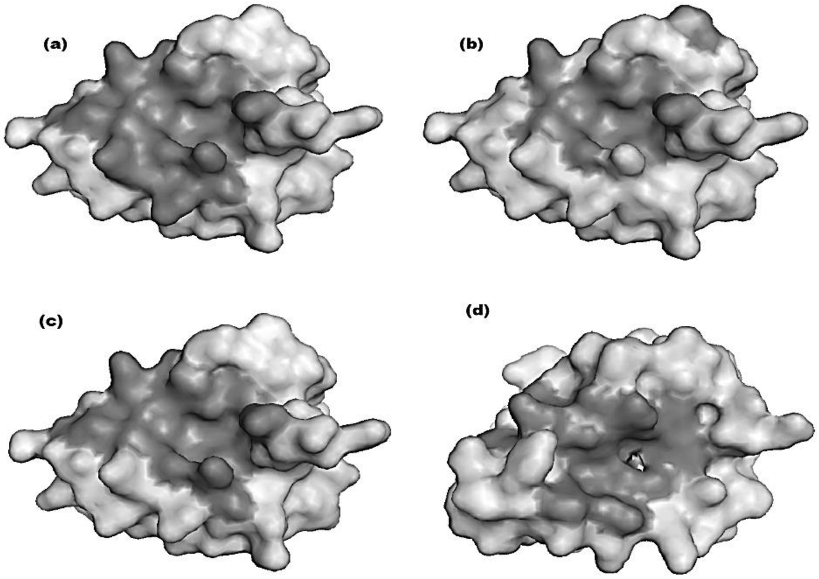

Figure 1 (a) Actual binding residues, (b) Predicted binding residues from the actual protein structure, (c) Predicted binding sites and (d) Predicted binding sites based on the homology model for the PTPN4 PDZ domain (pdbID: 3nfkA).

the proteins binding with DNA, RNA, and carbohydrate indicates that the clustering can correct the falsely predicted binding residues so that the predicted sites are significantly enriched in peptide-binding proteins. At the certain reliability score cutoff, the percentage of predicted binding residue in peptide-binding proteins is 44%, 1% for DNA-binding proteins, 6% for RNA-binding proteins, 6.5% for carbohydrate-binding proteins and 8.2% for all proteins. Thus, the method can discriminate peptide-binding proteins from others. Meanwhile, a similar performance by using homologous models indicates its wide applicability.

4. REFERENCES

1. Rubinstein, M. and Niv, M.Y. 2009. Peptidic modulators of protein-protein interactions: progress and challenges in computational design, *Biopolymers*, 91, 505-513.

2. Petsalaki, E. and Russell, R.B. 2008. Peptide-mediated interactions in biological systems: new discoveries and applications, *Curr. Opin. Biotechnol.*, 19, 344-350.

3. Yang, J., Roy, A. and Zhang, Y. 2013. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, *Nucleic Acids Res.*, 41, D1096-D1103.

4. Taherzadeh, G., *et al.* 2016. Sequence-based prediction of protein–peptide binding sites using support vector machine, *J. Comput. Chem.*, 37, 1223-1229.

5. Breiman, L. 2001. Random forests, *Machine learning*, 45, 5-32.

6. Ester, M., *et al.* 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*. pp. 226-231.