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

Article *in* Bioinformatics · September 2017

DOI: 10.1093/bioinformatics/btx614

Ghazaleh Taherzadeh
Wilkes University

Yaoqi Zhou
Shenzhen Bay Laboratory

Alan Wee-Chung Liew
Griffith University

Yuedong Yang
Sun Yat-Sen University

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Structural Bioinformatics

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[1,2,3 *]

[1]School of Information and Communication Technology, [2]Institue for Glycomics, Griffith University, Parklands Drive, Southport, Queensland 4215, Australia. [3]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein-peptide interactions are one of the most important biological interactions and play crucial role in many diseases including cancer. Therefore, knowledge of these interactions provides invaluable insights into all cellular processes, functional mechanisms, and drug discovery. Protein-peptide interactions can be analyzed by studying the structures of protein-peptide complexes. However, only a small portion has known complex structures and experimental determination of protein-peptide interaction is costly and inefficient. Thus, predicting peptide-binding sites computationally will be useful to improve 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 residues. These predicted binding residues are then employed to infer the peptide-binding site by a clustering algorithm.

Results: SPRINT-Str achieves robust and consistent results for prediction of protein-peptide binding regions in terms of residues and sites. Matthews' Correlation Coefficient (MCC) for 10-fold cross validation and independent test set are 0.27 and 0.293, respectively, as well as 0.775 and 0.782, respectively for Area Under the Curve (AUC). The prediction outperforms other state-of-the-art methods, including our previously developed sequence-based method. A further spatial neighbor clustering of predicted binding residues leads to prediction of binding sites at 20%-116% higher coverage than the next best method at all precision levels in the test set. The application of SPRINT-Str to protein binding with DNA, RNA, and carbohydrate confirms the method's capability of separating peptide-binding sites from other functional sites. More importantly, similar performance in prediction of binding residues and sites is obtained when experimentally determined structures are replaced by unbound structures or quality model structures built from homologs, indicating its wide applicability.

Availability: http://sparks-lab.org/server/SPRINT-Str.

Contact: yangyd25@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein-peptide interactions play vital roles in drug design because of their involvement in various cellular processes such as DNA repair, replication, gene expression and metabolism (Pawson and Nash, 2003; Rubinstein and Niv, 2009). In fact, 15-40% of protein-protein interactions are mediated by small peptides (Neduva, et al., 2005). Insight into this process has brought significant understanding of protein functions (Petsalaki and Russell, 2008). Recently, new functional roles of protein-peptide interaction were described and investigated (London, et al., 2012). These interactions were implicated in human diseases especially in cancers (Penna, et al., 2007; Tovar, et al., 2006) and viral infections (Clare and Clary, 2004). Therefore, there has been a growing interest to determine protein-peptide complex structures experimentally. Recent advance in structural biology has increased the number of these complexes significantly, and consequently provides better physical understanding of the interactions. The characteristics of the binding sites were studied in previous work (London, et al., 2013; Stanfield and Wilson, 1995). Like other protein-biomolecular interactions, peptides bind in the conserved region of the target protein (Ren, et al., 1993). In addition, peptides use hydrogen bonds to form interactions with their protein partner (London, et al., 2010). The peptide-binding regions in proteins appear to be dominated with large and flatter pockets (Olmez and Akbulut, 2012). Despite these efforts, it remains challenging to study them experimentally due to small peptide sizes (Vlieghe, et al., 2010), weak binding affinity (Dyson and Wright, 2005), and peptide flexibility (Bertolazzi, et al., 2014). Thus, it is desirable to have reliable computational methods to complement experimental studies. Most previous computational studies have focused on peptide binding sites of specific protein domains such as MHC, PDZ, SH2 and SH3 (Guo, et al., 2013; Hou, et al., 2009; Kundu, et al., 2013; Niv and Weinstein, 2005; Zhang, et al., 2009; Zhou, et al., 2005). A general method applicable to all protein domains would be useful.

Direct docking of peptides onto protein structures can predict binding residues by predicting protein-peptide complex structures. Examples of protein-peptide docking programs are Rosetta FlexPepDock (Raveh, et al., 2011), HADDOCK (De Vries, et al., 2010), Pep-SiteFinder (Saladin, et al., 2014), PepCrawler (Donsky and Wolfson, 2011), GalaxyPepDock (Lee, et al., 2015), MDockPeP (Yan, et al., 2016) and CABS-dock (Blaszczyk, et al., 2016). However, docking methods are less feasible for docking typical peptides of lengths between 5 and 10 residues onto proteins with unknown binding sites because of large search space for flexible peptide conformations (Trabuco, et al., 2012).

To avoid peptide conformational sampling, different strategies have been developed to predict putative binding sites. Pepsite (Trabuco, et al., 2012) employs spatial position specific scoring matrix derived from known protein-peptide complex structures to locate hot-spots on protein surfaces and determine binding sites based on distance constraints. FoldX (Verschueren, et al., 2013) attempts to infer peptide binding sites by employing interacting backbone fragment pairs.

However, the above methods are limited due to their requirement of binding peptide sequences that are not always known. To solve this problem, Peptimap (Lavi, et al., 2013) maps and clusters potential binding sites by docking small molecule probes. More recently, ACCLUSTER (Yan and Zou, 2015) scans 20 amino-acid probes on the protein surface to locate strong interactions, and determines putative binding sites by clustering.

Previously, we have developed a sequence-based method called SPRINT (Taherzadeh, et al., 2016) to predict protein residues that bind to peptides by support vector machine technique. The method utilizes sequence profiles generated from multiple sequence alignment along with predicted local and global structural information from SPIDER2 (Heffernan, et al., 2016; Heffernan, et al., 2015). This sequence-based approach has accuracy comparable to other methods. It is expected that inclusion of information from experimental structures or accurate structural models will further boost the prediction accuracy of binding residues.

In this study, we developed a novel machine learning-based approach called SPRINT-Str (Structure-based Prediction of Residue-level INTeraction) to predict putative protein-peptide binding residues and binding sites. Both structural and sequence-based information were integrated by a Random Forest (RF) classifier for prediction of binding residues, which was then employed to infer binding sites using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester, et al., 1996). We then compared the performance of SPRINT-str to other state-of-the-art methods in prediction of binding residues and binding sites and discrimination of peptide-binding proteins against DNA, RNA, and carbohydrate-binding proteins.

2 Materials and Methods

2.1 Dataset

The initial dataset of protein-peptide complex structures was obtained from the BioLip, where peptides have been defined as chains containing less than 30 amino acid residues (Yang, et al., 2013). We excluded any peptide-binding proteins of length<30 or containing less than three binding residues, where a binding residue is defined if any of its heavy atoms is within 3.5Å from a heavy atom in the binding peptide (Taherzadeh, et al., 2016). These protein chains were further clustered by 30% sequence identity with “blastclust” in BLAST package (Altschul, et al., 1997), and one representative chain was randomly selected from each cluster. The final dataset consists of 1,241 protein-peptide complexes with 16,678 binding residues out of totally 297,598 residues. We randomly selected 10% complexes as an independent test set and the remaining as a training set, to ensure that there is no systematic difference between training and independent test data. The training set contains 1,116 proteins including 14,959 binding residues (TR1116) and 251,769 nonbinding residues, and the independent test set contains 125 proteins (TS125) including 1,719 binding and 29,151 nonbinding residues. The ratio of binding to non-binding residues in both sets is around 1:17.

Ten-fold cross validation was employed to train our method. That is, the training set (TR1116) was randomly divided into ten parts (folds) according to proteins. Nine folds were used for training and the remaining one fold was employed for test. A protein-based fold separation removes the possibility of the same protein in training and test sets and reduces the potential of over-training. Each fold was tested in turn and all ten folds were used for calculating the overall performance of cross validation. Finally, the whole training set was employed to train the final model that was tested over the independent test set. The comparison of performances in the cross validation and independent test provides a measure of the robustness of the developed method. These datasets are available online at: http://sparks-lab.org/server/SPRINT-Str.

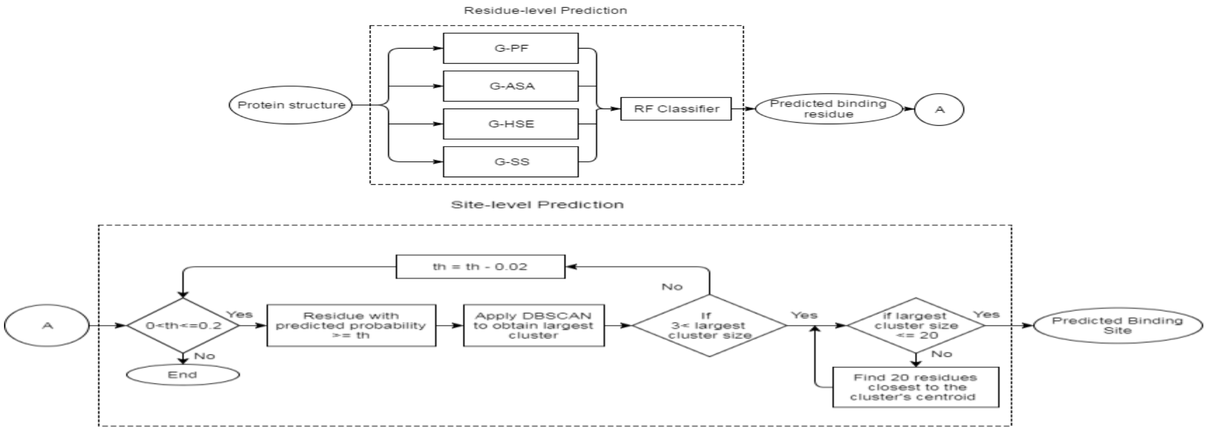
Residue-level Prediction
Protein structure
G-PF
G-ASA
G-HSE
G-SS
RF Classifier
Predicted binding residue
A
Site-level Prediction
A
th = th - 0.02
0<th<=0.2
Yes
No
End
Residue with predicted probability >= th
Apply DBSCAN to obtain largest cluster
No
If 3< largest cluster size
Yes
If largest cluster size <= 20
Yes
No
Find 20 residues closest to the cluster's centroid
Predicted Binding Site

Figure 1 The schematic diagram of SPRINT-Str.

2.2 Performance Evaluation Criteria

The performance of our method is measured by Matthews' Correlation Coefficient (MCC), F-measure, accuracy, sensitivity, and specificity that are defined as following:

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

$$F - measure = 2TP/(2TP + FP + FN)$$

$$Accuracy = \ (TP + TN)/(TP + TN + FP + FN)$$

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = \ TN/(FP + TN)$$

where TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively. In addition, we calculated the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC).

2.3 Binding region prediction

In this section, we described the steps towards developing SPRINT-Str. We first obtained sequence- and structure-based features. Next, we employed Random Forest as our classification technique to predict binding residues that were then clustered to predict binding sites by DBSCAN. Finally, we described parameter optimization and feature extraction techniques used in this study.

2.3.1 Feature vector

Sequence-based information:

In the sequence-based prediction of protein-peptide binding site (Taherzadeh, et al., 2016), we showed that features derived from sequence profile provided significant discriminative information of the binding sites. The sequence-profiles were obtained from PSI-BLAST (Altschul, et al., 1997) using E-value threshold of 0.001 in three iterations to extract the 20-dimensional Position Specific Scoring Matrix (PSSM) for each amino acid in the protein and entropy was calculated by $S_E = \sum_{j=1}^{j=20} P_{i,j} \times \ln(P_{i,j})$, where $P_{i,j}$ is the substitution probability of a given residue in the PSSM matrix with other 20 amino acids. We named this group of features as G-PF.

Structural Information

Previous studies showed the importance of solvent accessibility of binding residues (London, et al., 2010), helical conformation (Diella, et al., 2008), and flexible regions (London, et al., 2010) for peptide binding. We calculated Accessible Surface Area (ASA), Secondary Structure (SS), Half Sphere Exposure (HSE), and flexibility (Normal Mode) from protein structures detailed below:

Accessible Surface Area (ASA): We obtained ASA values for all residues through DSSP (Kabsch and Sander, 1983), which were normalized into the relative ASA (rASA) values. The rASA was further averaged over the neighbor residues with window sizes ranging from one to the to-be-optimized value (Average-rASA). We named this group of features as G-ASA.

Half Sphere Exposure (HSE): Solvent exposure can also be described by the number of structural neighbors of a given amino acid (contact number). The contact numbers in upward and downward hemispheres along with pseudo Cβ-Cα bond (upper half sphere and lower half sphere) (Hamelryck, 2005) were also obtained. We named this feature group as G-HSE.

Secondary Structure: We obtained secondary structure of each residue from DSSP (Kabsch and Sander, 1983). We further derived the fraction of each SS type in a window size in addition to features derived from a segment of SS. A segment was defined as continuous residues with the same type of SS. Length of segment and the position of given residue within the segment were used as features. We named this feature group as G-SS.

Normal mode: Intra-molecular dynamic motion of a protein is a good indicator of conformational flexibility in protein interactions (Dobbins, et al., 2008). We used iModeS (López-Blanco, et al., 2014) server to extract the normal mode information to describe the functional motions of proteins (Dykeman and Sankey, 2010) and flexibility of each amino acid. This feature, however, was not selected during feature selection as described below.

2.4 Random Forest

We employed the Random Forest (RF)(Breiman, 2001) classifier to predict peptide-binding residues. RF is an ensemble-based algorithm containing multiple decision trees that is widely utilized in classification and regression problems (Liaw and Wiener, 2002). It is well-known for its ability to deal with unbalanced datasets (Chen, et al., 2004) that we have here. RF builds bootstrap samples of the dataset using random selection with replacement from the training dataset. It grows a classifi-

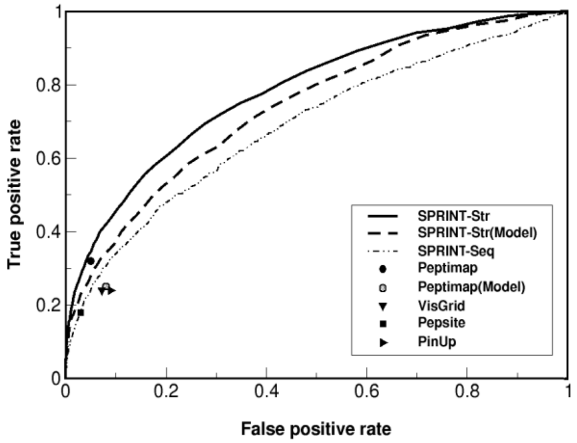

Figure 2 ROC curves given by various methods including sequence-based method SPRINT-Seq, and structure-based methods, VisGrid, PinUP, PepSite, Peptimap, and this work that employed actual structures and homology modelling structures, respectively.

cation tree for each subsample (bagging). It constructs the classification model by employing balanced subsamples of a dataset to fit multiple trees and classify an object from an input vector. The samples that are not involved in tree growth state are named 'out-of-bag' samples. Each tree classifies the 'out-of-bag' samples. Then the prediction output is calculated by combining results of all trees using majority votes.

2.5 Density-based Spatial clustering of applications with noise

The predicted binding residues are clustered to identify the largest binding region on the protein surface using Density-based Spatial Clustering of Applications with Noise (DBSCAN) (Ester, et al., 1996). The DBSCAN is a density-based clustering algorithm that requires a minimum number of points in a cluster (MinPts) and a parameter ε as the distance threshold to find the point within the distance ε of a given point. It starts from initial point from the input data and looks for the points within the specified distance and radius. This process repeats until no neighbour point can be found. DBSCAN outputs the cluster number and specifies the points as core (1), border (0) or noise (-1).

2.6 Method development

Initial models were selected by employing different window sizes with a number of trees ranging from 50 to 400 incrementing by 50 trees in each step. We found the highest AUC value in ten-fold cross validation when using a window size of 3 for all features, ([R_{i-3}, R_{i-2}, R_{i-1}, R_i, R_{i+1}, R_{i+2}, R_{i+3}], where *i* is the given residue). The optimal number of trees is 200, using bootstrapping samples to grow the trees and Gini function to measure the node impurity. Further increase of the window size or the number of trees did not significantly influence the prediction performance.

Table 1. Performance of SPRINT-Str on the 10-fold cross validation (CV) and independent test set (ACC: Accuracy; SEN: Sensitivity; SP: Specificity)

Methods	Datasets	MCC	AUC	ACC	SEN	SP
SVM (All Features)	CV	0.242	0.735	0.920	0.181	0.948
	Ind. Test	0.254	0.746	0.938	0.201	0.977
Random Forest (All features)	CV	0.259	0.757	**0.934**	0.235	**0.976**
	Ind. Test	0.281	0.763	0.939	0.199	**0.987**
Random Forest (Selected features)	CV	**0.271**	**0.775**	0.923	**0.30**	0.960
	Ind. Test	**0.293**	**0.782**	**0.941**	**0.241**	0.982

For comparison, we also trained a model by SVM implemented in the libsvm (Chang and Lin, 2011) as it was used in our previous study of sequence-based prediction, SPRINT (Taherzadeh, et al., 2016). However, by employing all features we found RF has better performance (details in results section). Thus, we selected random forest model for further study.

To further select the most discriminative features, we re-optimized the window size and applied sequential forward feature selection (SFFS) (Kudo and Sklansky, 2000) for each feature group at the same time. This was done by starting with an empty feature set and applying RF on each feature group to re-optimize the window size. The feature group with the highest AUC was selected. Then, we added each remaining feature group with different window sizes, and trained new RF model by selecting the feature group and window size for the highest AUC. This process was repeated until there is no improvement for the AUC value. As a result, G-PF, G-ASA, G-HSE, and G-SS were selected with optimized window size of 5, 4, 4 and 2, respectively.

Using the selected features, RF classifier outputs predicted probability for each amino acid. Based on the predicted binding probability for each residue, the binding site was determined as below:

1- Selecting predicted binding residues with predicted binding probability greater than a threshold (initially 0.2).
2- Applying DBSCAN algorithm to cluster spatially neighbouring residues.
3- Selecting the largest cluster from the output of DBSCAN algorithm.
4- Control of cluster size:
    a. If the cluster contains less than 3 residues, decrease the predicted binding residues probability threshold by 0.02 and repeat steps 1-3.
    b. If the cluster contains more than 20 residues, select the closest residues to the centroid.

In the DBSCAN algorithm, pairwise distances are calculated between all predicted binding residues, and two residues are defined as neighbour if their Cα atom-Cα atom distance is less than 7 Å, a commonly used cut off distance (Atilgan, et al., 2004). The residues are considered as core residues if they have at least three neighbouring residues, while other residues neighbouring with these core residues are border residues. Core and border residues are clustered together, and the biggest cluster is selected as the predicted binding site. Reduction of threshold is undertaken because the initial threshold of 0.2 was optimized for maximum MCC for predicting binding residues of all proteins, which may cause very few predicted binding residues in some proteins. The threshold of cluster size was chosen to maintain a balance of precision and sensitivity of predicted binding sites. Based on binding probability predicted by the above RF model, the predicted binding site is assessed according to the average probability on its included residues, namely the reliability score.

Table 2. Performance of Random Forest models by employing individual feature group or by removing each feature group from the final model.

Feature groups [a]	MCC	AUC	Feature groups [b]	MCC	AUC
-	-	-	SPRINT-Str	0.293	0.782
G-PF[c]	0.268	0.711	- G-PF	0.226	0.744
G-HSE[d]	0.175	0.66	- G-HSE	0.271	0.773
G-ASA[e]	0.113	0.675	- G-ASA	0.284	0.747
G-SS[f]	0.026	0.53	- G-SS	0.277	0.78

[a] Performances based on individual feature groups
[b] Performances by removing each feature group from the final model
[c] G-PF: Sequence profile group from PSSM
[d] G-HSE: Half Sphere Exposure group
[e] G-ASA: Accessible Surface Area group
[f] G-SS: Secondary Structure group

This procedure was implemented in python using scikit-learn package (Pedregosa, et al., 2011). A schematic diagram of our method is illustrated in Figure 1.

3 Results and Discussion

3.1 Prediction of binding residues

The performance of models trained by RF and SVM using all features as well as the final model in ten-fold cross validation and independent test is shown in Table 1. In the ten-fold cross validation, SVM model obtained an AUC value of 0.73 that is less than the AUC value of 0.757 achieved by RF. After feature selection for RF model, the AUCs for the ten-fold cross validation and independent test set increase to 0.775 and 0.782, with an MCC of 0.27 and 0.293, respectively. Similar performance of MCC and AUC values between ten-fold cross validation and independent test set indicates the robustness of the final model for predicting protein-peptide binding residues. To further confirm the consistency of our RF model, we have randomly re-sampled 30 sets of training and independent test datasets. These datasets yielded essentially the same results for all independent test sets with an average MCC and AUC values of 0.27±0.02 and 0.77±0.01, respectively.

Due to the unbalanced dataset, SPRINT-Str has chosen a high threshold (0.2) for achieving the maximum MCC value, which led to a low sensitivity of 0.24 to ensure high specificity (98%). If a lower cut off value such as 0.11 is employed, SPRINT-str yields a sensitivity of 0.506 with a specificity of 0.873. For completeness, the test performance in term of AUC, MCC, sensitivity, specificity and precision for all individual proteins is shown in Supplementary Table S1.

Table 3. Comparison of different methods on the TS125 test set.

Methods	MCC	AUC	F-measure	ACC	SEN	SPE
SPRINT-Str	0.29	0.78	0.309	0.941	0.24	0.98
w/ model[1]	0.25	0.74	0.24	0.94	0.212	0.98
SPRINT-Seq[a]	0.20	0.68	0.221	0.92	0.21	0.96
Peptimap	0.27	0.63	0.294	0.92	0.32	0.95
Pepsite	0.20	0.61	0.219	0.929	0.18	0.97
PinUp	0.13	0.58	0.18	0.88	0.24	0.91
VisGrid	0.15	0.63	0.19	0.89	0.24	0.928

[1] Prediction of SPRINT-Str by using structure models built by SPARKS X based on templates with sequence identity <60%

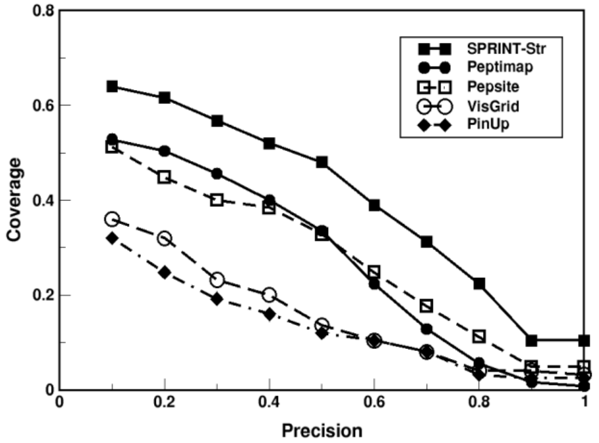

Figure 3 Performance comparison on coverage and precision by several methods as labelled for binding-site prediction.

Table 2 examines the contribution of individual feature group by using single feature group only or by removing it from our final model. The sequence profile group (G-PF) yields the highest MCC and AUC as a single feature group and has the largest drop in MCC and AUC when removing it. This is followed by orientation-dependent contact number (G-HSE), solvent accessible area (G-ASA), and secondary structure (G-SS). Adding G-ASA, G-HSE and G-SS to G-PF improves AUC from 0.71 to 0.782.

It is of interest to compare the structure-based method to the method based on sequence only (SPRINT-Seq) (Taherzadeh, et al., 2016). To avoid having overlap between the independent test of this work and the training set of SPRINT-Seq, we compared the results on 80 proteins (TS80) from the independent test set of SPRINT-Seq, which is a subset of TS125. Our structure-based method has similar performance on TS80 and TS125 with MCC and AUC of 0.28 and 0.77, respectively, for TS80, compared to 0.29 and 0.78 for TS125. SPRINT-Seq achieves MCC of 0.19 and AUC of 0.687, respectively. The improvement of our method over the sequence-based technique demonstrates the importance of employing accurate structural information.

In the above prediction, both buried and exposed residues are predicted. Obviously buried residues are unlikely to participate in binding. Removing those residues may decrease the performance of our method. To examine the effect, we have removed all amino acids with ASA equal to zero from proteins in the test set TS125. We found that the impact is small: the MCC value decreases slightly from 0.293 to 0.288. This happens because only 11% residues (including 8 actual binding residues) have been excluded. If we further remove mostly buried amino acids, the overall accuracy will decrease further but remains reasonable (MCC~0.25) even if we have treated all residues with relative ASA<80% as buried and excluded them from binding residues.

Because most proteins do not have experimentally determined structures, we tested how modeled structures affect the performance. Here, we employed the fold recognition technique SPARK-X (Yang, et al., 2011) to generate model protein structure by using homology (sequence identity < 60% to the query sequence). As shown in Figure 2, the use of model structures leads to a slight decrease in the performance (MCC changing from 0.293 to 0.251) with a median of GDT score of 0.695 for model structures. Here the structural accuracy of a model is measured by the global distance test (GDT) score, which is 1 for a perfect match with

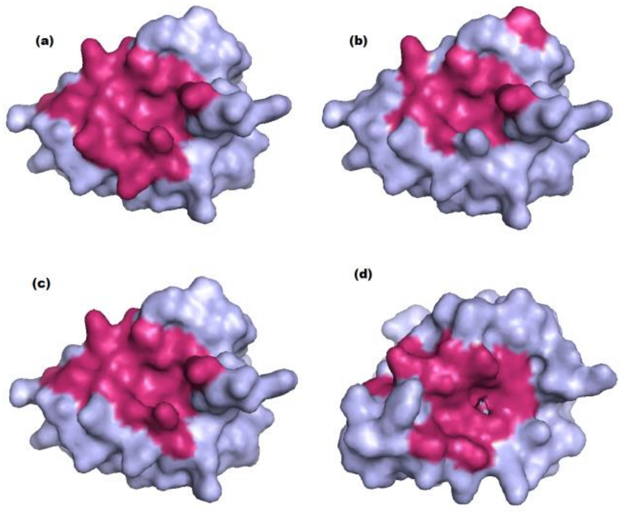

Figure 4 (a) Actual binding residues, (b) Predicted binding residues, (c) Predicted binding site from the actual protein structure, and (d) Predicted binding sites based on the homology model for the PTPN4 PDZ domain (pdbID: 3nfkA).

experimental structure. By comparison, the performance of Peptimap depends more strongly on model structural accuracy. Its MCC value decreases from 0.265 to 0.195 when model structures are used.

To further assess the dependence of method performance on model structure quality, we made predictions for all top 10 models predicted by the template-based structure modelling method SPARKS X. As shown in Supplementary Fig. S1, the accuracy of predicted binding sites improves when the GDT score increases. It reaches a plateau for 0.7<GDT≤1.0. In other words, using a model with GDT>0.7 is sufficient to yield a prediction of binding sites as accurate as using an experimental structure. Details of model structure quality, template structure and prediction accuracy for each protein is shown in Table S2.

Binding induced conformational changes may also affect the performance of our method. In order to assess the impact, we have built a list of 87 unbound chains that are the same or highly homologous to the sequences in TS125 (sequence identity > 70%). SPRINT-Str has achieved an average AUC of 0.75 by using these unbound structures, which is only slightly lower than 0.77 based on bound structures. The performance of SPRINT-Str in term of AUC, MCC, sensitivity and specificity for all individual proteins (both bound and unbound) is shown in Supplementary Table S3.

Table 3 further compares this work with two other available structure-based approaches for protein-peptide binding sites in TS125: Peptimap (Lavi, et al., 2013) and Pepsite(Petsalaki, et al., 2009; Trabuco, et al., 2012). For Pepsite, the native sequence of the binding peptide was employed. In addition, we compare our method with two other structure-based methods, PinUp (Liang, et al., 2006) (protein-protein binding site) and VisGrid (Li, et al., 2008) (protein-ligand binding site). The positions of these methods on AUC curves are shown in Figure 2. Our method achieved the highest MCC (0.29), AUC (0.78) and F-measure (0.31), followed by Peptimap (0.26, 0.63, and 0.29, respectively). The difference in AUC between Peptimap and our method is significant (P-value=2E-08) in statistics (Hanley and McNeil, 1982). When computing MCC and AUC on individual proteins, their average values are 0.411 and 0.77 for SPINT-Str, and 0.224 and 0.645 for the second-best method Peptimap. The t-test indicates a significant difference with P-value = 2e-08 and 7e-13. Here, we did not compare with docking methods because they usually require peptide sequences and specification of binding regions.

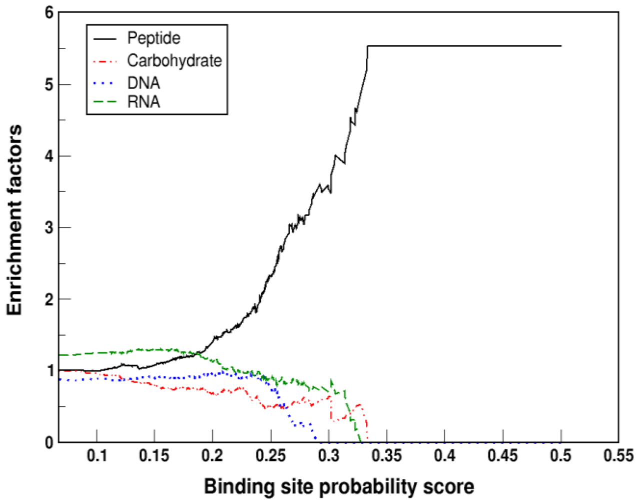

Figure 5 The ratio of predicted binding sites for proteins binding with peptide or other molecular types by SPRINT-Str (enrichment factors) as a function of average probability (Reliability Score).

3.2 Prediction of binding sites

Correct prediction of binding regions (binding sites) sometimes is more important than identification of specific residues. DBSCAN is employed to cluster and detect binding sites based on predicted binding probabilities (see methods).

Figure 3 plots coverage as a function of precision. $Precision = S_i / B_i$, where S_i is the number of true positive residues in the predicted binding site and B_i is the number of residues in the predicted binding site. *Coverage* is the fraction of correctly predicted peptide-binding sites in all actual binding sites at a given precision value cut-off. Our method consistently outperforms other methods in binding site prediction. At all precision levels (10-100%), SPRINT-Str has a coverage of between 20% and 116% higher than the next best methods (Peptimap at low precision and Pepsite at high precision).

As an example, we demonstrate the prediction of the human tyrosine phosphatase protein PTPN4 PDZ domain (pdbID: 3nfkA). In Figure 4, there is one false positive residue predicted by the residue-level prediction that was far from other predicted binding residues and filtered in the predicted binding site after clustering. Finally, we predicted 17 binding residues where 15 residues are truly binding. For the homologous modeling, SPARKS X has detected the template 2q9vA, which is an unbound structure (not binding with any peptide or ligand). Despite of only 26% sequence identity with the template, the homology model has a SP-score of 0.82 with an RMSD of 2.2Å over all 92 residues by SP-align (Yang, et al., 2012). Based on this model structure, we predicted 15 binding residues, where 12 residues are truly binding. The total accuracy is 95%, slightly lower than 97% by using the actual structure.

3.3. Discrimination from other binding types

It is interesting to know if our peptide-binding predictor is specific for peptide-binding. Here, we have made the assumption that proteins do not use the same site to bind chemically different ligands (carbohydrate, DNA, RNA, and peptides) as in previous studies (Miao and Westhof, 2015; Yan, et al., 2015; Zhang and Kurgan, 2017) although it is possible that in some cases, same binding sites could bind to different ligands (Jeffery, 2003).

We employed structural datasets of protein-carbohydrate complexes (*Zhao, et al., 2014; Taherzadeh, et al., 2016), protein-DNA complexes (Zhao, et al., 2014) and protein-RNA complexes (*Zhao, et al., 2011) that include 152, 179, and 212 complex structures, respectively. At the pre-determined threshold of 0.2, 2.4%, 1.1%, and 1.4% of total residues were predicted as peptide-binding residues for carbohydrate, DNA, and RNA binding proteins, respectively. By comparison, peptide-binding proteins have predicted 3.5% of total residues as binding residues, 41% of which are predicted correctly.

To further examine whether our method can discriminate peptide-binding proteins from other types of binding proteins, we calculated enrichment factors of predicted peptide-, carbohydrate-, RNA-, or DNA-binding proteins. Here, $EF_i = (B_i/\sum B_i)/(N_i/\sum N_i)$, where B_i is the number of proteins having binding site with the reliability score above a cutoff, and N_i is the total number of proteins for a given binding type *i* (peptide, carbohydrate, RNA, and DNA). As shown in Figure 5, for DNA, RNA, and carbohydrate-binding proteins, the enrichment factor is around 1 (random prediction) and significantly below 1 after reliability score is above 0.2. Known peptide-binding proteins can have an enrichment factor significantly above 1, and the enrichment factor is above 4 when the reliability is >0.25. Specifically, at a threshold of 0.31, the percentage of predicted binding sites in peptide-binding proteins is 40%, 5.5% for DNA-binding proteins, 8% for RNA-binding proteins and carbohydrate-binding proteins, and 12.5% for all proteins, which leads to an enrichment factor of 3.2 for peptide binding proteins.

To quantitatively compare against other methods, Supplementary Figure S2 plots an AUC curve for predicting peptide-binding proteins based on a threshold for predicting binding residues. We have employed TS125 dataset of peptide-binding proteins (positive) along with 30 DNA, 30 RNA, and 30 carbohydrate binding proteins (negatives). We found that our method can discriminate peptide-binding proteins from other binding types significantly better than Peptimap, with AUC of 0.697 for SPRINT-str and 0.54 for Peptimap, respectively. The difference is significant (P=0.002).

4 Conclusion

In this work, we developed a machine-learning based method called SPRINT-Str to predict protein-peptide binding residues from protein 3D structure. We used the structural information in addition to the most effective sequence-based features used in our previous work (Taherzadeh, et al., 2016). We confirmed that evolutionary information is an important discriminative feature to predict protein-peptide binding residues. Using information based on actual structures improves over the sequence-based method. The clustering algorithm DBSCAN improves prediction of binding sites by removing prediction noises. The performance is similar when quality homology models are used instead of experimental structures. The SPRINT-Str web-server is available at: http://sparks-lab.org/server/SPRINT-Str.

Acknowledgements

We would like to acknowledge the use of the High Performance Computing Cluster "Gowonda" to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF). We also would like to thank the Peptimap and Pepsite authors for providing the webservers.

Funding

This work is supported by the Australian Government Research Training Program Scholarship for G. T. This work is also in part supported by NHMRC (1059775 and 1083450) and ARC (LE150100161) of Australia to Y.Z, and the National Natural Science Foundation of China (U1611261, 61772566) and the program for Guangdong Introducing Innovative and Enterpreneurial Teams (2016ZT06D211) to YY.

Conflict of Interest: none declared.

References

Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research*, **25**, 3389-3402.

Atilgan, A.R., Akan, P. and Baysal, C. (2004) Small-world communication of residues and significance for protein dynamics, *Biophys. J.*, **86**, 85-91.

Bertolazzi, P., Guerra, C. and Liuzzi, G. (2014) Predicting protein-ligand and protein-peptide interfaces, *The European Physical Journal Plus*, **129**, 1-10.

Blaszczyk, M., *et al.* (2016) Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking, *Methods*, **93**, 72-83.

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

Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: a library for support vector machines, *ACM. TIST.*, **2**, 27.

Chen, C., Liaw, A. and Breiman, L. (2004) Using random forest to learn imbalanced data, *University of California, Berkeley*, 1-12.

Clare, D.F. and Clary, D.C. (2004) Computational studies of protein–peptide interactions with systematic mutation of residues, *Mol. Phys.*, **102**, 939-951.

De Vries, S.J., van Dijk, M. and Bonvin, A.M. (2010) The HADDOCK web server for data-driven biomolecular docking, *Nature protocols*, **5**, 883-897.

Diella, F., *et al.* (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation, *Front Biosci*, **13**, 6580-6603.

Dobbins, S.E., Lesk, V.I. and Sternberg, M.J. (2008) Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–protein docking, *Proceedings of the National Academy of Sciences*, **105**, 10390-10395.

Donsky, E. and Wolfson, H.J. (2011) PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors, *Bioinformatics*, **27**, 2836-2842.

Dykeman, E.C. and Sankey, O.F. (2010) Normal mode analysis and applications in biological physics, *J. Phys.: Condens. Matter*, **22**, 423202.

Dyson, H.J. and Wright, P.E. (2005) Intrinsically unstructured proteins and their functions, *Nature reviews Molecular cell biology*, **6**, 197-208.

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

Guo, L., Luo, C. and Zhu, S. (2013) MHC2SKpan: a novel kernel based approach for pan-specific MHC class II peptide binding prediction, *BMC genomics*, **14**, 1.

Hamelryck, T. (2005) An amino acid has two sides: a new 2D measure provides a different view of solvent exposure, *Proteins: Structure, Function, and Bioinformatics*, **59**, 38-48.

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, **143**, 29-36.

Heffernan, R., *et al.* (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins, *Bioinformatics*, **32**, 843-849.

Heffernan, R., *et al.* (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning, *Sci. Rep.*, **5**.

Hou, T., *et al.* (2009) Characterization of domain-peptide interaction interface a generic structure-based model to decipher the binding specificity of SH3 domains, *Molecular & Cellular Proteomics*, **8**, 639-649.

Jeffery, C.J. (2003) Moonlighting proteins: old proteins learning new tricks, *TRENDS in Genetics*, **19**, 415-417.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen‐bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.

Kudo, M. and Sklansky, J. (2000) Comparison of algorithms that select features for pattern classifiers, *Pattern Recognit.*, **33**, 25-41.

Kundu, K., *et al.* (2013) Semi-supervised prediction of SH2-peptide interactions from imbalanced high-throughput data, *PloS one*, **8**, e62732.

Lavi, A., *et al.* (2013) Detection of peptide‐binding sites on protein surfaces: The first step toward the modeling and targeting of peptide‐mediated interactions, *Proteins: Struct., Funct., Bioinf.*, **81**, 2096-2105.

Lee, H., *et al.* (2015) GalaxyPepDock: a protein–peptide docking tool based on interaction similarity and energy optimization, *Nucleic Acids Res.*, gkv495.

Li, B., *et al.* (2008) Characterization of local geometry of protein surfaces with the visibility criterion, *Proteins: Structure, Function, and Bioinformatics*, **71**, 670-683.

Liang, S., *et al.* (2006) Protein binding site prediction using an empirical scoring function, *Nucleic Acids Res.*, **34**, 3698-3707.

Liaw, A. and Wiener, M. (2002) Classification and regression by randomForest, *R news*, **2**, 18-22.

London, N., Movshovitz-Attias, D. and Schueler-Furman, O. (2010) The structural basis of peptide-protein binding strategies, *Structure*, **18**, 188-199.

London, N., Raveh, B. and Schueler-Furman, O. (2012) Modeling peptide–protein interactions, *Homology Modeling: Methods and Protocols*, 375-398.

London, N., Raveh, B. and Schueler-Furman, O. (2013) Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how, *Current opinion in structural biology*, **23**, 894-902.

López-Blanco, J.R., *et al.* (2014) iMODS: internal coordinates normal mode analysis server, *Nucleic Acids Res.*, **42**, W271-W276.

Miao, Z. and Westhof, E. (2015) A large-scale assessment of nucleic acids binding site prediction programs, *PloS Comput Biol*, **11**, e1004639.

Neduva, V., *et al.* (2005) Systematic discovery of new recognition peptides mediating protein interaction networks, *PLoS Biol*, **3**, e405.

Niv, M.Y. and Weinstein, H. (2005) A flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains, *J. Am. Chem. Soc.*, **127**, 14072-14079.

Olmez, E.O. and Akbulut, B.S. (2012) *Protein-peptide interactions revolutionize drug development*. chapter.

Pawson, T. and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains, *science*, **300**, 445-452.

Pedregosa, F., *et al.* (2011) Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**, 2825-2830.

Penna, G., *et al.* (2007) Spontaneous and prostatic steroid binding protein peptide-induced autoimmune prostatitis in the nonobese diabetic mouse, *The Journal of Immunology*, **179**, 1559-1567.

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

Petsalaki, E., *et al.* (2009) Accurate prediction of peptide binding sites on protein surfaces, *PLoS Comput. Biol.*, **5**, e1000335.

Raveh, B., *et al.* (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors, *PLoS One*, **6**, e18934.

Ren, R., *et al.* (1993) Identification of a ten-amino acid proline-rich SH3 binding site, *SCIENCE-NEW YORK THEN WASHINGTON-*, **259**, 1157-1157.

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

Saladin, A., *et al.* (2014) PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces, *Nucleic Acids Res.*, **42**, W221-W226.

Stanfield, R.L. and Wilson, I.A. (1995) Protein-peptide interactions, *Current opinion in structural biology*, **5**, 103-113.

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

Taherzadeh, G., *et al.* (2016) Sequence-Based Prediction of Protein–Carbohydrate Binding Sites Using Support Vector Machines, *J. Chem. Inf. Model.*, **56**, 2115-2122.

Tovar, C., *et al.* (2006) Small-molecule MDM2 antagonists reveal aberrant p53 signaling in cancer: implications for therapy, *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 1888-1893.

Trabuco, L.G., *et al.* (2012) PepSite: prediction of peptide-binding sites from protein surfaces, *Nucleic acids research*, **40**, W423-W427.

Verschueren, E., *et al.* (2013) Protein-peptide complex prediction through fragment interaction patterns, *Structure*, **21**, 789-797.

Vlieghe, P., *et al.* (2010) Synthetic therapeutic peptides: science and market, *Drug Discovery Today*, **15**, 40-56.

Yan, C., Xu, X. and Zou, X. (2016) Fully Blind Docking at the Atomic Level for Protein-Peptide Complex Structure Prediction, *Structure*, **24**, 1842-1853.

Yan, C. and Zou, X. (2015) Predicting peptide binding sites on protein surfaces by clustering chemical interactions, *J. Comput. Chem.*, **36**, 49-61.

Yan, J., Friedrich, S. and Kurgan, L. (2015) A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues, *Briefings in bioinformatics*, **17**, 88-105.

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.

Yang, Y., *et al.* (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates, *Bioinformatics*, **27**, 2076-2082.

Yang, Y., *et al.* (2012) A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction, *Proteins*, **80**, 2080-2088.

Zhang, H., Lund, O. and Nielsen, M. (2009) The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding, *Bioinformatics*, **25**, 1293-1299.

Zhang, J. and Kurgan, L. (2017) Review and comparative assessment of sequence-based predictors of protein-binding residues, *Briefings in Bioinformatics*, bbx022.

Zhao, H., Yang, Y. and Zhou, Y. (2011) Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets, *Nucleic Acids Res*, **39**, 3017-3025.

Zhao, H., *et al.* (2014) Predicting DNA-binding proteins and binding residues by complex structure prediction and application to human proteome, *PloS one*, **9**, e96694.

Zhao, H., *et al.* (2014) Carbohydrate-binding protein identification by coupling structural similarity searching with binding affinity prediction, *J Comput Chem*, **35**, 2177-2183.

Zhou, H., *et al.* (2005) Solution structure of AF-6 PDZ domain and its interaction with the C-terminal peptides from Neurexin and Bcr, *The Journal of biological chemistry*, **280**, 13841-13847.