Sequence-Based Prediction of Protein–Peptide Binding Sites Using Support Vector Machine

Ghazaleh Taherzadeh,[a] Yuedong Yang,[a,b] Tuo Zhang,[c] Alan Wee-Chung Liew,[a] and Yaoqi Zhou*[a,b]

Protein–peptide interactions are essential for all cellular processes including DNA repair, replication, gene-expression, and metabolism. As most protein–peptide interactions are uncharacterized, it is cost effective to investigate them computationally as the first step. All existing approaches for predicting protein–peptide binding sites, however, are based on protein structures despite the fact that the structures for most proteins are not yet solved. This article proposes the first machine-learning method called SPRINT to make Sequence-based prediction of Protein–peptide Residue-level Interactions. SPRINT yields a robust and consistent performance for 10-fold cross validations and independent test. The most important feature is evolution-generated sequence profiles. For the test set (1056 binding and non-binding residues), it yields a Matthews' Correlation Coefficient of 0.326 with a sensitivity of 64% and a specificity of 68%. This sequence-based technique shows comparable or more accurate than structure-based methods for peptide-binding site prediction. SPRINT is available as an online server at: http://sparks-lab.org/. © 2016 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24314

Introduction

Up to 40% of protein–protein interactions are mediated by protein–peptide interactions[1] that are implicated in human diseases such as cancer and researched for drug discovery.[2,3] It is challenging, however, to identify protein–peptide interactions due to small peptide sizes,[4] weak binding affinity,[5] and short separated binding motifs.[6] Despite the challenges, there is a steady increase in protein–peptide complex structures determined by experimental techniques.

These complex structures permit an in-depth understanding of protein–peptide interactions and associated binding sites.[7] One third of protein-binding peptides have a helical conformation.[8] The remaining peptides typically bind to the gaps in an extended beta or polyproline category II helical conformation. Protein–peptide interaction sites are packed more tightly than protein–protein interaction sites and involve more hydrogen bonds especially in the peptide backbone.[7] Protein–peptide binding sites are flatter than ligand binding sites with larger pockets.[7] Other studies focused on putative peptide binding site on specific protein domains like MHC, PDZ, SH2, and SH3[9–13] but these findings are not generalizable.[14] Most of these studies were limited to protein–peptide interactions with known experimental complex structures. The majority of protein–peptide interactions remain unknown.

Protein–peptide complex structures can be predicted computationally by docking. Examples are Rosetta FlexPepDock,[15] HADDOCK,[16] Pep-SiteFinder,[17] and PepCrawler.[18] These docking predictions, however, are accurate only if peptide-binding sites are known reasonably well. As a result, several docking-based methods for predicting peptide-binding sites were developed. ACCLUSTER,[19] for example, employs 20 different amino acid types as a probe to scan the protein surface. The amino acids with strong chemical interactions are selected and clustered. The largest cluster is predicted as binding sites for the peptides. GalaxyPepDock,[2] on the other hand, builds peptide-binding sites based on known docked structures by structural similarity search. FoldX predicts peptide-binding sites by fragment interaction pattern.[20] Peptide-binding sites were also predicted by combining score-based hot-spot prediction with distance constraints in method called Pepsite[21] and by probe-based surface clustering in Peptimap method.[22] All the above computational methods, however, require that protein structures are known. Given the fact that the majority of protein structures are unknown, it is necessary to develop a method that can predict peptide-binding sites from sequence only.

This article established a machine-learning method called SPRINT (Sequence-based prediction of Protein–peptide

[a] G. Taherzadeh, Y. Yang, A. W.-C. Liew, Y. Zhou
School of Information and Communication Technology, Griffith University, Parklands Drive, Southport, Queensland 4215, Australia
E-mail: yaoqi.zhou@griffith.edu.au
[b] Y. Yang, Y. Zhou
Institute for Glycomics, Griffith University, Parklands Dr, Southport, Queensland 4215, Australia
[c] T. Zhang
Weill Cornell Medical College, 1300 York Avenue, New York, New York 10065

Contract grant sponsor: National Health and Medical Research Council of Australia; Contract grant numbers: 1059775 and 1083450; Contract grant sponsor: Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme; Contract grant number: LE150100161 (to Y.Z.); Contract grant sponsor: National Natural Science Foundation of China; Contract grant number: 61271378 (to Y.Y.); Contract grant sponsor: Microsoft Azure for Research Award (to Y.Y.); Contract grant sponsor: Queensland Cyber Infrastructure Foundation (QCIF)

Residue-level Interaction sites) that predicts peptide-binding sites directly from protein sequence. By using known protein–peptide complex structures for training and independent test, we showed that the sequence evolution profile is the most important feature that discriminates binding from non-binding residues. This feature combined with other features trained by support vector machine (SVM) led to a method that is comparable or more accurate than existing structure-based techniques for peptide-binding site prediction.

Methods

Datasets

Our dataset of protein–peptide complex structures was extracted from the BioLip protein–ligand database with peptide as the ligand[23] derived from the Protein Data Bank (PDB).[24] The initial peptide-binding set was reduced by blastclust[25] to remove redundant proteins with more than 30% sequence identity. This dataset was further divided into training and independent test sets. We randomly selected 80 proteins to use as the independent test set and employed the remaining proteins (1199) as the training/cross validation set.

Peptide-binding residues in a protein were defined as residues that contain at least one atom with distance less than 3.5 Å from any atom in the peptide.[26]

The above protein–peptide dataset has a total of 307,683 residues in which 16,744 residues are binding residues and the rest are non-binding. To maximize learning and avoid bias caused by large number of non-binding residues, a balanced dataset was made using under sampling approach[27] by randomly selecting a number of non-binding residues that is equal to the number of binding residues. The final training/cross validation set has 15,688 binding and 15,688 non-binding residues and the independent test set has 1056 binding, and 1056 non-binding residues. These two data sets are available at: http://sparks-lab.org/yueyang/server/SPRINT/. We also built 10 variants of test sets by randomly choosing different 1056 non-binding residues to examine the robustness of our testing results.

Input features

Sequence. Each sequence position is represented by a 20-dimensional binary vector. The type of the amino acid residue at the sequence position is represented by "1" and the rests are filled by "0." This sequence feature is denoted as G-SEQ.

Sequence Profile from PSSM. Conserved residues of the proteins have more tendencies to interact with the peptides.[28] Thus, we obtained a 20-dimensional Position Specific Scoring Matrix (PSSM) and the other 20-dimensional probability matrix from PSI-BLAST that iteratively searches and aligns homologous sequences to produce a substitution score for each amino acid in the protein sequence at its sequence position by other amino acids.[29] Three iterations and an *E*-value cut off of 0.001 were used. In addition to 20 features per position directly from PSSM, based on the probability matrix $\{P_{ij}\}$ *where* $j = 1, 2, \ldots, 20$ and $i = 1, \ldots, L$ (L is the length of protein), we also calculated the information entropy $[S_E = \sum_{j=1}^{j=20} P_{i,j} \times \ln(P_{i,j})]$, and Close Neighbor Correlation Coefficient (CNCC).[30]

$$\mathrm{CNCC}_{ik} = \frac{\sum_{j=1}^{20} P_{ij} P_{kj}}{\sqrt{\sum_{j=1}^{20} (P_{ij})^2 \sum_{j=1}^{20} (P_{kj})^2}}, \quad (1)$$

where CNCC_{ik} is the Pearson correlation coefficient between the sequence profile of the given residue i and that of its adjacent residue k in a sliding window of w ($k = i - w, \ldots, i + w$). This group of features is referred as G-PF.

Predicted Accessible Surface Area and SS. We obtained predicted solvent accessible surface area (ASA) and secondary structure (SS) by SPIDER 2.0,[31] which is one of the most accurate methods for predicting ASA and SS (0.74 for the correlation between predicted and actual ASA and 82% for the accuracy of secondary structure prediction).

Accessible Surface Area. SPIDER 2.0 provided an ASA value of each residue and we normalized the values[32] to obtain rASA of each residue. We further calculated the average rASA of adjacent amino acids within different window sizes from one to the size of the applied window (rASA_avg). This group of features is referred as ASA-based features (G-ASA).

Secondary Structure. From SPIDER 2.0, we obtained the predicted probabilities of three secondary structure types (SSProb) of the sequence position. We also calculated fraction of each SS type over number of residues within the window (SSCont). In addition, the secondary structure triplet of the center residue and two nearest neighbors (SSTri) is encoded by a 27-dimensional vector. Furthermore, we extracted two additional features from the continuous SS segments enclosing the given residue. We obtained SegLen that records the number of residues in the segment and SegDB that specifies the minimum and maximum distances of the given residue from both ends of its continuous SS segment. Similar features were employed previously for predicting RNA-binding residues.[33] This group of features is referred as SS-based features (G-SS).

Mixed SS and ASA Features. Predicted rASA and SS values are simplified into two-dimensional and three-dimensional vectors, separately, to represent exposed and buried states for ASA or coil, helix and sheet states for SS. Nine different cutoffs, continuously from 0.1 to 0.9, were employed to defined exposed or buried states. That is, each position has 54 features (2*3*9). This mixing feature group is referred as G-MIX.

Physiochemical Properties. We employed seven representative physiochemical attributes of the amino acids for feature extraction (steric parameter, hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability).[34] This feature group is referred as the physicochemical-based feature group in this study (G-PP7).

All above feature values were transformed to the range of [−1, 1] as input for the training and test.

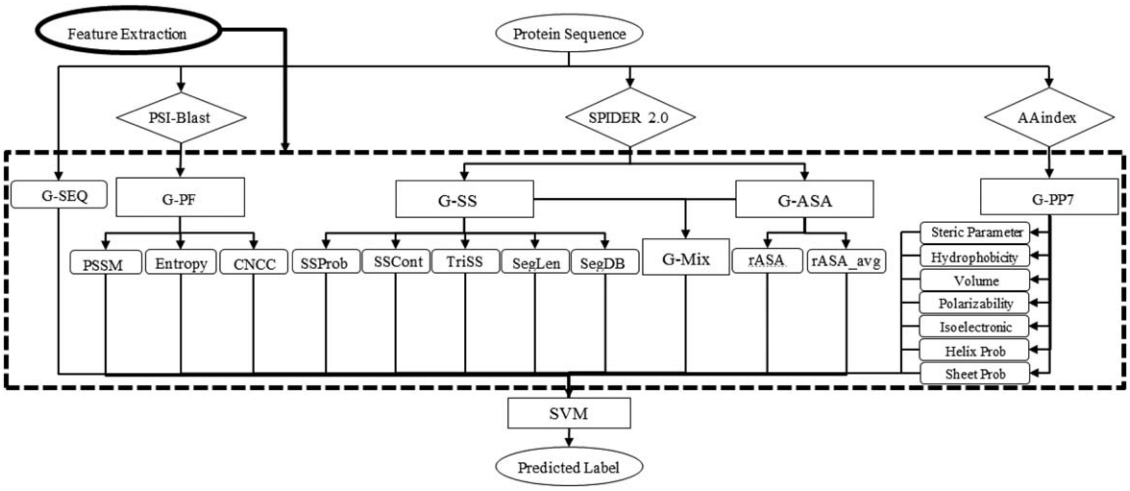

Figure 1. The general architecture of the SPRINT.

Support vector machine

In this article, we employed support vector machine (SVM)[35] as our classification technique. SVM is one of the state-of-the-art algorithms in machine learning. It has been broadly employed to address the classification and regression problems.[36] The objective of SVM is to find the hyper-plane with the largest margin to decrease the misclassification rate. If we define the classification value of the points in the input space x_i as y_i and recognize as either $+1$ or -1, then we can define the SVM algorithm in the eq. (2).

$$y' = \text{sign}\left(\sum_{i=1}^{n} a_i y_i K(x_i, x') + b\right).$$

where y' is the predicted class of x'. K defines the type of kernel, n is the number of support vectors, α_i is the adjustable weight, and b is the bias. SVM uses a kernel to map the input data into its specific feature space.[37] In this work, we use LibSVM[38] for training and testing with the radial basis function (RBF) kernel. The RBF kernel function is defined as,

$$K(z_i,\ z_j) = \exp\left(-\gamma * \| z_i - z_j \|^2\right). \quad (2)$$

A schematic flow diagram for our SVM model is shown in Figure 1.

Performance evaluation

Method performance can be measured by the number of correct classified or misclassified instances using the terms below:

- TP: number of actual binding residues predicted as binding sites.
- TN: number of actual non-binding residues predicted as non-binding sites.
- FP: number of actual non-binding residues incorrectly predicted as binding sites.
- FN: number of actual binding residues incorrectly predicted as non-binding sites.

We evaluated the performance of our proposed prediction method in terms of the Matthews' Correlation Coefficient (MCC), accuracy, sensitivity, specificity, Receiver Operating Characteristic (ROC) curve, and the Area Under the Curve (AUC). Equations are given below.

$$\text{MCC} = \frac{(\text{TP}\times\text{TN} - \text{FP}\times\text{FN})}{\sqrt{(\text{TP}+\text{FP})\times(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}}. \quad (3)$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}. \quad (4)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}). \quad (5)$$

$$\text{Specificity} = \text{TN}/(\text{FP}+\text{TN}). \quad (6)$$

MCC varies between -1 and 1, where -1 represents exactly opposite correlation, 0 represents no correlation, and 1 represents perfectly positive correlation.

Parameter optimization and feature selections

We first optimized kernel parameters gamma and C for SVM and also a fixed window size for all features by a grid search for the highest MCC value when using randomly selected 10% of the training set. The optimal parameters for gamma and C were 0.05 and 1, respectively. The optimal window size for all

Table 1. Method performance on the 10-fold cross validation and independent test.

Dataset	MCC	AUC	Accuracy	Sensitivity	Specificity
Cross validation	0.333	0.748	0.666	0.646	0.687
Std[a]	0.01	0.01	0.01	0.02	0.01
Independent test	0.326	0.711	0.662	0.642	0.683
Std[b]	0.005	0.013	0.009	0.015	0.016

[a] Standard deviation of 10-fold cross validation results. [b] Standard deviation of 10 sets of alternatively balanced datasets.

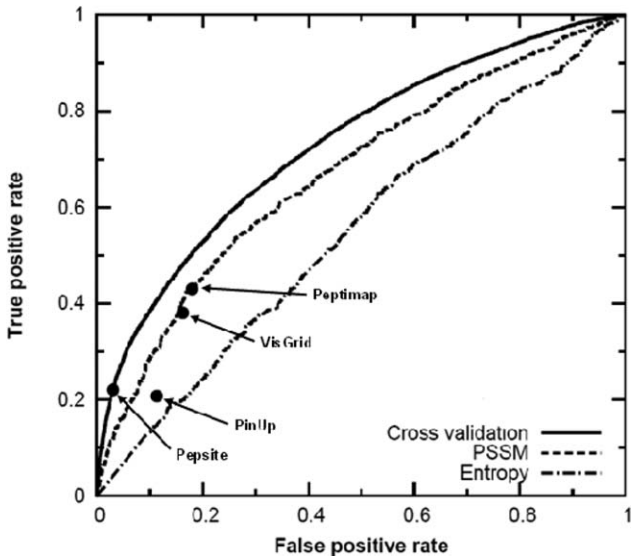

Figure 2. The ROC curve for SPRINT on 10-fold cross validation, trained by entropy and PSSM along with structure-based Peptimap, Pepite, PinUp, VisGrid.

features was 4 (totally nine residues). With gamma = 0.05 and $C = 1$, the window size for each feature group was reoptimized by using that feature group only for the SVM model. Window sizes of 9, 2, 1, 2, 1, and 1 were found for G-PF, G-SS, G-SEQ, G-ASA, G-PP7, and G-MIX, respectively. To reduce the possibility of over-training, we reduced the number of features by the greedy sequential forward selection. Only G-PF, G-SS, G-SEQ, and G-ASA feature groups were finally selected. The feature set and all other parameters were then used for 10-fold cross validation and independent test as described below.

Cross validation and independent test

As it was mentioned, we divided our dataset into training and independent test sets of 1199 and 80 proteins, respectively. We conducted the protein-based 10-fold cross validation on the training set. That is, the training set was divided into 10 parts according to proteins, not according to residues. This protein-based separation removes the possibility of the same protein in training and test sets and reduces the potential of over-training. In the 10-fold cross validation, nine folds were used for training and the remaining one fold was employed for test and each fold was tested in turn. We also used the whole training set to train the SVM model and test it on the independent test set.

Results and Discussion

Table 1 compares the result of 10-fold cross validation, the standard deviation of all folds, and that of independent test. The MCC value and AUC for 10-fold cross validation are 0.333 and 0.748, respectively, compared to 0.326 and 0.711 for the independent test set.

In comparison, the test on the unbalanced independent test set gives a similar AUC value 0.68. Essentially the same values of accuracy, sensitivity, and specificity were obtained for the cross validation and independent test set.

The performance of SPRINT on the cross validation is compared to those of two best features (PSSM and entropy) in Figure 2. It is clear that PSSM is substantially more accurate than entropy in discriminating binding from non-binding residues. The performance is further improved by combining different features.

Table 2. Evaluation of the importance of each feature group individually as well as by adding or deleting them.

Feature groups[a]	MCC	AUC	Accuracy	Feature groups[b]	MCC	AUC	Accuracy	Feature groups[c]	MCC	AUC	Accuracy
G-SEQ	0.123	0.571	0.560	−G-SEQ	0.317	0.701	0.658	G-PF	0.295	0.69	0.647
G-PF	0.295	0.69	0.647	−G-PF	0.228	0.61	0.614	+G-SS	0.310	0.701	0.655
G-SS	0.13	0.586	0.554	−G-SS	0.305	0.696	0.652	+G-SEQ	0.315	0.707	0.657
G-ASA	0.115	0.565	0.556	−G-ASA	0.315	0.707	0.657	+G-ASA	0.326	0.711	0.662

[a] Performance of individual feature group. [b] The effect from removing one feature group. [c] Adding the ranked feature group one by one.

Table 3. Importance of individual features revealed from single feature discrimination and by removing it from the SVM model.

Features[a]	MCC	AUC	Accuracy	Sensitivity	Specificity	Features[b]	MCC	AUC	Accuracy	Sensitivity	Specificity
SEQ	0.123	0.571	0.560	0.451	0.668	−SEQ	0.317	0.701	0.658	0.642	0.675
PSSM	0.253	0.651	0.610	0.618	0.651	−PSSM	0.25	0.671	0.625	0.627	0.622
Entropy	0.122	0.57	0.561	0.552	0.570	−Entropy	0.312	0.69	0.657	0.633	0.681
CNCC	0.091	0.558	0.54	0.635	0.454	−CNCC	0.317	0.7	0.662	0.647	0.677
SSProb	0.092	0.561	0.556	0.536	0.575	−SSProb	0.311	0.69	0.656	0.647	0.664
SScont	0.073	0.534	0.536	0.465	0.608	−SScont	0.316	0.701	0.661	0.64	0.680
TriSS	0.069	0.53	0.532	0.363	0.701	−TriSS	0.315	0.7	0.661	0.64	0.679
SegLen	0.042	0.52	0.521	0.515	0.527	−SegLen	0.314	0.69	0.660	0.641	0.679
SegDB	0.048	0.52	0.522	0.335	0.709	−SegDB	0.319	0.694	0.662	0.643	0.680
ASA	0.104	0.561	0.556	0.646	0.465	−ASA	0.320	0.708	0.66	0.640	0.679
rASA_avg	0.090	0.544	0.545	0.596	0.493	−rASA_avg	0.322	0.71	0.662	0.642	0.682

[a] Performance of individual features. [b] Performance of SPRINT by excluding individual feature.

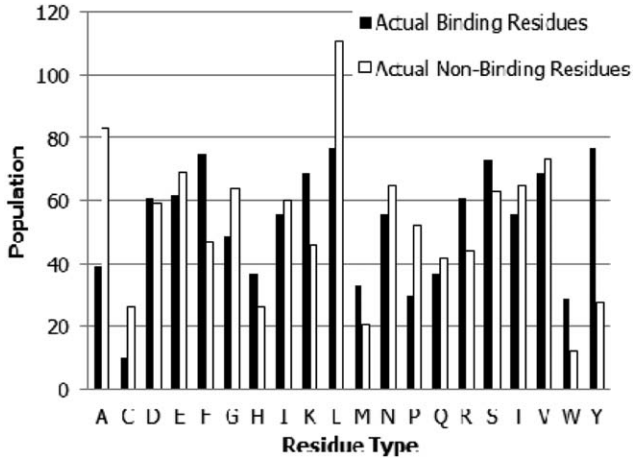

Figure 3. Population distribution of actual binding and non-binding residues according to residue types.

To further confirm the robustness of the overall performance of the SVM model, we generated 10 sets of alternatively balanced datasets by randomly selecting different non-binding residues for each protein in the independent test set. The average MCC is 0.331 with a small standard deviation of 0.005.

It is of interest to know the contribution of each feature group to the overall prediction performance of SPRINT. Table 2 compares the performance of each feature group individually by removing individual feature group and by sequentially adding ranked feature groups. All indicate that the G-PF feature group makes the highest contribution. The contributions from other feature groups are similar. Combining the four feature groups leads to a better prediction performance than using the PF feature group alone.

We further analyzed the contribution of each individual feature. As Table 3 shows, PSSM in the PF feature group is the most important feature whether it is assessed as a single feature or it is removed from the SVM model. All features make statistically significant contribution to the final SVM model except rASA_avg.

Figure 3 compares the population of actual binding and non-binding amino acid residue types in the independent test dataset. Binding residues are enriched in Phenylalanine (F), Lysine (K), Arginine (R), and Tyrosine (Y) whereas non-binding residues are enriched in Alanine (A) and Leucine (L).

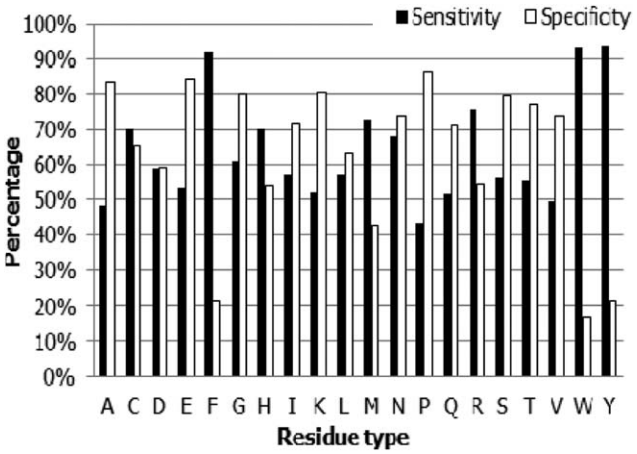

Figure 4. Sensitivity and specificity for 20 amino acid residue types.

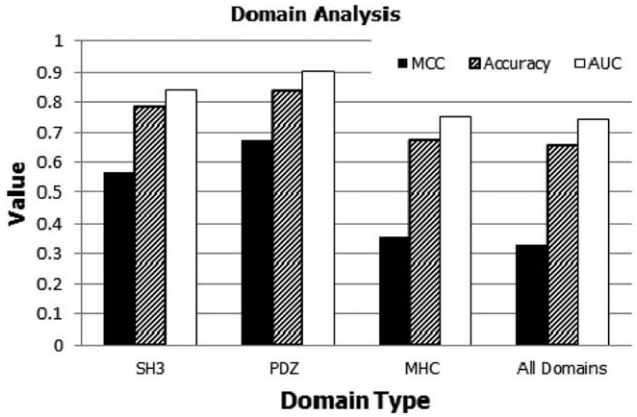

Figure 5. Domain dependent performance based on MCC, accuracy and AUC as labeled.

The performance of our method for each amino acid residue type is shown in Figure 4. SPRINT can achieve high sensitivity for the prediction of Phenylalanine (F), Tryptophan (W), and Tyrosine (Y) with a sensitivity of 92%, 93%, and 93%, respectively, but with low specificity. More balanced accurate prediction (both sensitivity and specificity $> 50\%$) was made for the majority of amino acids (13/20). No correlation was found between the population and the overall accuracy of the prediction for a given residue type.

Figure 5 examines the dependence of method performance on specific protein domains based on 10-fold cross validation results. Cross validation data is employed because its overall accuracy is similar to the independent test and the large dataset allows a more reliable statistics. In this dataset, there are 32 proteins in SH3, 36 proteins PDZ, and 34 proteins in MHC domains.

Figure 5 shows that peptide binding sites of SH3 and PDZ domains are more accurately predicted (MCC > 0.5, Accuracy $> 70\%$, and AUC > 0.7) whereas those sites of MHC domains are less accurately predicted (comparable to the overall accuracy). This domain-dependent performance suggests that the enrichment of one structural fold does not always lead to superior performance. This supports that our method is not over-trained.

It is of interest to compare the accuracy of our sequence-based technique with two recently developed structure-based methods (Peptimap[22] and PepSite[21]). The results were

Table 4. Comparison of SPRINT with the previous study on protein–peptide biding site prediction on the independent test set.

Methods	MCC	Accuracy	Sensitivity	Specificity
This work	0.326	0.662	0.642	0.683
Peptimap	0.262	0.629	0.436	0.823
PepSite	0.141	0.598	0.221	0.975
PinUp	0.071	0.547	0.207	0.887
VisGrid	0.224	0.609	0.38	0.838

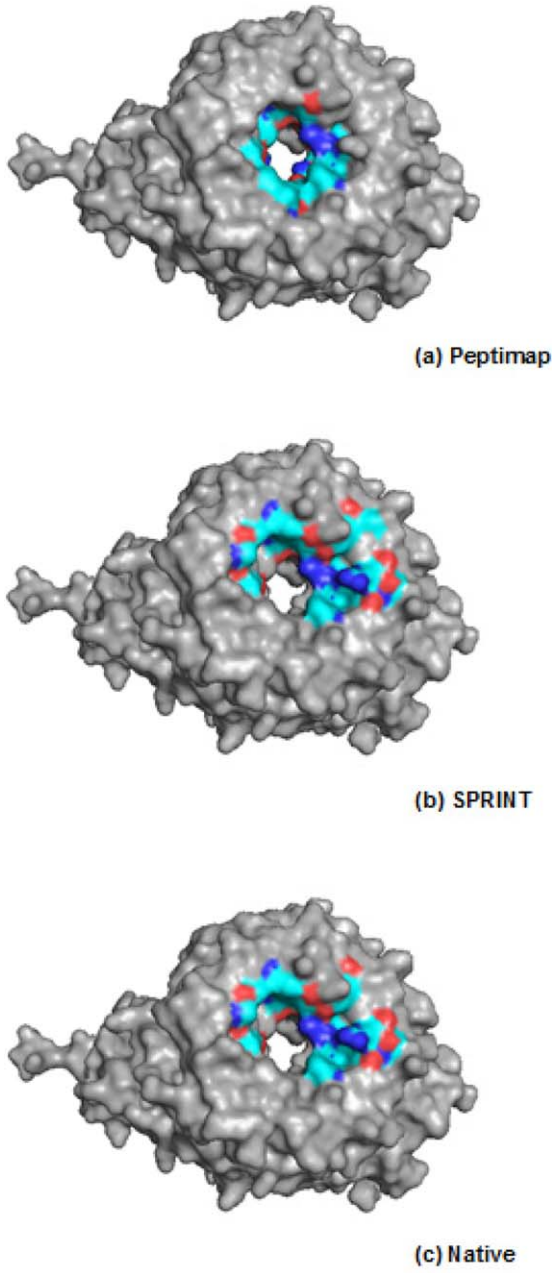

Figure 6. The predicted binding sites by Peptimap a) and this work b) as compared to the actual peptide binding site c) for ubiquitin ligase (cell division control protein 4, PDB 3v7d domain B) protein. The binding sites are colored based on the element's type.

obtained by submitting protein sequences (peptide sequences for PepSite as well) directly to the provided webserver. As Table 4 shows, PepSite and Peptimap produced low MCC values (0.14 and 0.26, respectively) and low accuracy (59% and 62%, respectively). That is, our method, despite relying on sequence information only, is the first method with MCC > 0.3 for peptide binding prediction. The comparison to other methods is also shown in Figure 2 where each method was shown as one point because of preset thresholds. It should be noted that these methods achieved high specificity at the expense of low sensitivity. If the threshold for SPRINT is set so that the specificity is the same as Peptimap (82.3%), SPRINT will achieve the sensitivity of 49.5%, which is 6% higher than Peptimap. On the other hand, if we make the specificity to 97.5%, SPRINT will have a sensitivity of 22.1%, the same as Pepsite. That is, SPRINT is comparable to Pepsite based on this measure.

As one example, Figure 6 compared actual binding sites with predicted binding sites of ubiquitin ligase (cell division control protein 4, PDB 3v7d, chain B) by SPRINT and by Peptimap. A nearly perfect agreement is made by this work with MCC $= 0.92$ and accuracy $= 0.96$. By comparison, the MCC and accuracy for Peptimap are 0.43 and 0.69, respectively.

Table 4 also compared PINUP[39] that was developed for predicting protein–protein interaction sites and Vis-Grid[40] that was dedicated to ligand binding site prediction. Both are structure-based techniques. The results were obtained from their respective servers. PINUP's performance on predicting peptide-binding sites is poor, confirming significant difference between peptide-binding sites and protein-binding sites. On the other hand, the performance of Vis-Grid is close to that of PepSite and Peptimap, indicating there exists stronger similarity between ligand-binding sites and peptide binding sites.

Conclusions

We have developed the first sequence-based method for predicting protein–peptide binding sites. We showed that using SVM model to combine sequence evolution and predicted structural information leads to a method that is more accurate than existing structure-based techniques. The consistence between cross validation and independent test suggests its ability to predict binding sites of unseen proteins. The method called SPRINT is implemented as an online server available at: http://sparks-lab.org/.

Acknowledgments

The authors would like to thank Abdollah Dehzangi and Lukas Folkman for helpful discussions. They would like to thank the Peptimap and Pepsite for providing the webservers. They also gratefully acknowledge the support of the Griffith University eResearch Services Team and the use of the High Performance Computing Cluster "Gowonda" to complete this research.

Keywords: protein–peptide · binding site · sequence-based · prediction · features · machine learning · support vector machine

How to cite this article: G. Taherzadeh, Y. Yang, T. Zhang, A. W.-C. Liew, Y. Zhou. *J. Comput. Chem.* **2016**, *37*, 1223–1229. DOI: 10.1002/jcc.24314

[1] E. Petsalaki, R. B. Russell, *Curr. Opin. Biotechnol.* **2008**, *19*, 344.
[2] H. Lee, L. Heo, M. S. Lee, C. Seok, *Nucl. Acids Res.* **2015**, *43*, 431.
[3] S. Dutta, T. S. Chen, A. E. Keating, *ACS Chem. Biol.* **2013**, *8*, 778.
[4] P. Vlieghe, V. Lisowski, J. Martinez, M. Khrestchatisky, *Drug Discov. Today* **2010**, *15*, 40.
[5] H. J. Dyson, P. E. Wright, *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197.
[6] M. Fuxreiter, P. Tompa, I. Simon, *Bioinformatics* **2007**, *23*, 950.

[7] N. London, D. Movshovitz-Attias, O. Schueler-Furman, *Structure* **2010**, *18*, 188.
[8] F. Diella, N. Haslam, C. Chica, A. Budd, S. Michael, N. P. Brown, G. Travé, T. J. Gibson, *Front Biosci.* **2008**, *13*, 6580.
[9] L. Guo, C. Luo, S. Zhu, *BMC Genomics* **2013**, *14*, S11.
[10] K. Kundu, F. Costa, M. Huber, M. Reth, R. Backofen, *PLoS One* **2013**, *8*, e62732.
[11] H. Zhang, O. Lund, M. Nielsen, *Bioinformatics* **2009**, *25*, 1293.
[12] T. Hou, Z. Xu, W. Zhang, W. A. McLaughlin, D. A. Case, Y. Xu, W. Wang, *Cell Proteomics* **2009**, *8*, 639.
[13] M. Y. Niv, H. Weinstein, *J. Am. Chem. Soc.* **2005**, *127*, 14072.
[14] D. Gfeller, *FEBS Lett.* **2012**, *586*, 2764.
[15] B. Raveh, N. London, L. Zimmerman, O. Schueler-Furman, *PLoS One* **2011**, *6*, e18934.
[16] S. J. De Vries, M. van Dijk, A. M. Bonvin, *Nat. Protoc.* **2010**, *5*, 883.
[17] A. Saladin, J. Rey, P. Thévenet, M. Zacharias, G. Moroy, P. Tufféry, *Nucl. Acids Res.* **2014**, *42*, W221.
[18] E. Donsky, H. J. Wolfson, *Bioinformatics* **2011**, *27*, 2836.
[19] C. Yan, X. Zou, *J. Comput. Chem.* **2015**, *36*, 49.
[20] E. Verschueren, P. Vanhee, F. Rousseau, J. Schymkowitz, L. Serrano, *Structure* **2013**, *21*, 789.
[21] E. Petsalaki, A. Stark, E. García-Urdiales, R. B. Russell, *PLoS Comput. Biol.* **2009**, *5*, e1000335.
[22] A. Lavi, C. H. Ngan, D. Movshovitz-Attias, T. Bohnuud, C. Yueh, D. Beglov, O. Schueler-Furman, D. Kozakov, *Proteins: Struct., Funct., Bioinformatics* **2013**, *81*, 2096.
[23] J. Yang, A. Roy, Y. Zhang, *Nucl. Acids Res.* **2013**, *41*, D1096.
[24] H. Berman, K. Henrick, H. Nakamura, *Nat. Struct. Mol. Biol.* **2003**, *10*, 980–980.
[25] A. Biegert, C. Mayer, M. Remmert, J. Söding, A. N. Lupas, *Nucl. Acids Res.* **2006**, *34*, W335.
[26] V. Bianchi, P. F. Gherardini, M. Helmer-Citterich, G. Ausiello, *BMC Bioinformatics* **2012**, *13*, S17.
[27] S. J. Yen, Y. S. Lee, In Intelligent Control and Automation; Huang, D.-S.; Li, K; Irwin, G.W. (eds). Springer, **2006**; Vol. *344*, pp. 731–740.
[28] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, N. Ben-Tal, *Bioinformatics* **2003**, *19*, 163.
[29] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, *Nucl. Acids Res.* **1997**, *25*, 3389.
[30] J. Cheng, P. Baldi, *BMC Bioinformatics* **2007**, *8*, 113.
[31] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, *Sci. Rep.* **2015**, 5.
[32] E. Faraggi, B. Xue, Y. Zhou, *Proteins: Struct. Funct. Bioinformatics* **2009**, *74*, 847.
[33] T. Zhang, H. Zhang, K. Chen, J. Ruan, S. Shen, L. Kurgan, *Curr. Protein Pept. Sci.* **2010**, *11*, 609.
[34] J. Meiler, M. Müller, A. Zeidler, F. Schmäschke, *J. Mol. Model.* **2001**, *7*, 360.
[35] V. Vapnik, The Nature of Statistical Learning Theory; Springer-Verlag: New York, **2000**.
[36] C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273.
[37] C. M. Bishop, Pattern Recognition and Machine Learning; Springer, New York, **2006**.
[38] C. C. Chang, C. J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27.
[39] S. Liang, C. Zhang, S. Liu, Y. Zhou, *Nucl. Acids Res.* **2006**, *34*, 3698.
[40] B. Li, S. Turuvekere, M. Agrawal, D. La, K. Ramani, D. Kihara, *Proteins: Struct. Funct. Bioinformatics* **2008**, *71*, 670.

Received: 14 November 2015
Accepted: 6 January 2016
Published online on 2 February 2016