Sequence-based prediction of protein-peptide binding sites using support vector machine

Sequence-based Prediction of Protein-Carbohydrate Binding Sites Using Support Vector Machines

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] Institute for Glycomics, Griffith University, Parklands Dr. Southport, QLD 4215, Australia.

ABSTRACT. Carbohydrate-binding proteins play significant roles in many diseases including cancer. Here, we established a machine-learning-based method (called Sequence-based Prediction of Residue-level INTeraction sites of carbohydrates, SPRINT-CBH) to predict carbohydrate-binding sites in proteins by using Support Vector Machines (SVM). We found that integrating evolution-derived sequence profiles with additional information of sequence and predicted solvent accessible surface area leads to a reasonably accurate, robust, predictive method, with area under receiver operating characteristic curve (AUC) of 0.78 and 0.77, and Matthew's correlation coefficient of 0.34 and 0.29, respectively for ten-fold cross validation and independent test without balancing binding and non-binding residues. The quality of the method is further demonstrated by having statistically significantly more binding residues predicted for carbohydrate-binding proteins than presumptive non-binding proteins in the human proteome, and by the bias of rare alleles toward predicted carbohydrate-binding sites for non-synonymous mutations from the 1000 genome project. SPRINT-CBH is available as an online server at: http://sparks-lab.org/server/SPRINT-CBH.

Keywords: Protein–carbohydrate, Binding site, Sequence-based, Prediction, Features, Machine Learning, Support Vector Machines.

1. INTRODUCTION

The essential molecules of life are nucleic acids, lipids, proteins, and carbohydrates (or glycans). The interactions between proteins and carbohydrates mediate diverse biological functions, including cellular adhesion, cellular recognition, and signal transduction.[1] In fact all surfaces of living cells are decorated by glycoproteins and glycolipids. These exposed carbohydrates serve as key components in cell-cell communication. It is the first line of defense for human cells against pathogens.[2] It is also the first layer of protection for pathogens.[3] Carbohydrate-binding proteins (or glycan-binding proteins), which play a central role in recognizing these cell-surface carbohydrates, are useful as biomarkers and/or drug targets.[4-6] Protein-carbohydrate interactions, however, are challenging to study experimentally because of weak binding affinity and synthetic complexity of specific carbohydrates.[7] As a result, computational prediction becomes an important complementary tool.

One important aspect of studies in protein-carbohydrate interactions is to locate the sites of proteins that bind to carbohydrates. The first method for predicting protein-carbohydrate binding sites from a known protein structure was proposed by Taroni et al.[8] They evaluated six attributes of amino acids (solvation potential, residue propensity, hydrophobicity, planarity, protrusion and relative accessible surface area) and found that a simple combination of three parameters (residue propensity, protrusion index, and solvent accessibility) can be employed for predicting binding sites with an overall accuracy of 65% for a set of 40 protein-carbohydrate complexes. Sujatha and Balaji developed another structure-based method called COTRAN for predicting protein-galactose binding sites.[9] They

employed a combination of geometrical and structural features that allow detection of potential galactose-binding sites with a very high specificity and sensitivity based on known galactose-binding proteins in the same structural fold. Kulharia et al. developed InCa-SiteFinder for predicting inositol and carbohydrate binding sites on the protein surface.[10] The method was based on amino acid propensities and the van der Waals interaction energy between protein and a probe. A continuous surface pocket interacting with probes was predicted as binding sites. Nassif et al. employed random forests for feature selections and selected chemical and residue features, such as charges, hydrophobicity, and hydrogen bonding, and input them into support vector machines for predicting protein-glucose binding sites.[11] More recently, Tsai et al. predicted binding sites by using neural networks and support-vector-machines with probability distributions of interacting atoms in protein surfaces as input.[12]

The above-mentioned structure-based methods for binding site prediction rely on protein structures that are often not available. In 2007, Malik and Ahmad developed the first sequence-based method.[13] They used a simple neural network with the Position Specific Scoring Matrix (PSSM) as their input features. The method was tested by leave-one-out and achieved the average of 87% sensitivity and 23% specificity for a dataset of 40 protein-carbohydrate complexes. Pai and Mondal further developed a method called MOWGLI specific for mannose binding sites by using an ensemble of classifiers with PSSM as input.[14] Agarwal et al also developed a similar method for predicting mannose-binding sites by using PSSM.[15] Thus, for predicting non-specific carbohydrate-binding sites, there exists only one sequence-based method.[13] The method was subjected to a limited test (leave-one-out) and relied on sequence profiles from PSSM only. Moreover, lacking an on-line server or a standalone downloadable version further limits the usefulness of the method developed.

While only a few methods were dedicated to carbohydrate-binding sites, many other methods have been established for binding site prediction in protein-protein,[16-20] protein-DNA,[21-24] protein-RNA,[25-27] protein-ligand,[28-30] and protein-peptide[31-33] interactions. Many sequence-based techniques above have shown that integration of PSSM with physico-chemical properties of amino acid residues and predicted structural properties such as predicted secondary structures and solvent accessible surface area will significantly improve the overall performance of sequence-based techniques.

The objective of this paper is to develop an accurate sequence-based method by integrating sequence and predicted structural features for prediction of non-covalent carbohydrate-binding sites. We investigated the effectiveness of various feature groups for protein-carbohydrates binding site prediction. Effective features were selected to build a classifier based on Support Vector Machines (SVM). The new method, called SPRINT–CBH (Sequence-based Prediction of Residue-level INTeraction sites of carbohydrates), was trained and cross-validated by 102 carbohydrate-binding proteins and independently tested by 50 proteins with known high-resolution protein-carbohydrate complex structures. Although the datasets contain significantly more non-binding residues than binding residues, we found that the model developed by direct training on unbalanced full datasets improved over the methods trained on more balanced datasets by employing under-sampling and oversampling techniques. The quality of the method was further confirmed by similar performance of cross validation and independent test on the full dataset, the application to a protein-peptide binding dataset as a control, identification of the number of predicted binding residues within annotated carbohydrate-binding and non-binding proteins in the human proteome, and examination of the frequency of rare single nucleotide variants from the 1000 Genomes Project in predicted carbohydrate-binding sites.

2. MATERIALS AND METHODS

Dataset

The dataset for protein-carbohydrate complex structures was obtained by combining the dataset curated from the previous work[34] and the structures collected in Glyco3d.[35] We removed all glycosylated proteins because we are interested in non-covalent binding only. The dataset in the previous work was obtained from the PROCARB database.[36] After removing low resolution (>3Å) complexes and redundant proteins with a sequence identity cut-off of 30%, the dataset contains 113 protein-carbohydrate complexes. In addition, Glyco3d database (available from http://glyco3d.cermav.cnrs.fr/home.php) contains more than 1000 three-dimensional structures of protein-carbohydrate complexes. The combined set was filtered by removing proteins of low-resolution structures (>3Å) and homologous proteins (>30% sequence similarity according to BLAST-CLUST[37]). The final set has 152 protein-carbohydrate complexes, one third of which (50 proteins) was randomly chosen as the independent test set (TS50) and the remaining (102 proteins) as the training and cross-validation set (TR102). These proteins are listed in our website.

For each protein, binding residues were defined by using its corresponding protein-carbohydrate complex structure. We defined a residue as a carbohydrate-binding site if any atom in the residue is within 3.5Å [13, 38] from any carbohydrate atom. For the 152 carbohydrate-binding proteins we obtained 1,530 binding and 39,484 non-binding residues. Because this dataset has 26 times more non-binding residues than binding residues, we have balanced the TR102 dataset using under-sampling[39] and SMOTE[40] techniques and compared the methods trained on more balanced datasets with the method trained on the full dataset.

Under-sampling was done by randomly selecting a portion of non-binding residues so that its number is the same as the number of binding sites for each protein. This balanced dataset

with the ratio of 1:1 contains 1,042 binding and 1,042 non-binding residues in TR102. The SMOTE technique[40], on the other hand, under-sampling the majority class and over-sampling the minority class, to balance the TR102 dataset. The balanced dataset produced by SMOTE has 20,816 residues with the ratio of ≈ 1:4 (4,168 binding and 16,648 non-binding residues). We would like to emphasize that all balanced datasets are only utilized for training and the full dataset is employed in cross-validation and independent test because in the real-world situation the number of binding residues involving in carbohydrate-binding is only a small portion of all residues in a protein.

A cutoff of 3.5Å for defining binding residues is somewhat arbitrary. Thus, we also employed a cutoff of 6Å. This nearly doubled the number of binding residues (3044 binding and 37970 non-binding residues) for 152 carbohydrate-binding proteins. The ratio (binding: non-binding) is ≈1:12.

Input features

Sequence information. Each amino acid residue is represented by 1 for its residue type in a 20-dimensional vector (0 for all other rows). Similar to previous work,[41] we have also employed a one-dimensional vector as terminus indicator: terminal residues (the first and last 3 residues) are represented by 1 and other residues are by 0.

Evolutionary information. Evolutionary conserved residues may have functional roles such as binding.[42] Here we employed the Position Specific Scoring Matrix (PSSM) which is a 20*L dimensional matrix (where L is protein length) generated from PSI-BLAST using E-value threshold of 0.001 in three iterations.[43] In addition, we calculated information entropy from PSSM, $S_E = \sum_{j=1}^{j=20} P_{i,j} \times \ln(P_{i,j})$, where $P_{i,j}$ is the probability matrix at residue i, and j represents 20 standard amino acids. We also evaluated close neighbor correlation coefficient (CNCC), the Pearson's correlation coefficient (PCC) between P_{ij} of the query residue and those of its neighbors within a selected window.

We investigated seven sequence-based and sequence-derived structural features that have been used successfully for predicting protein-peptide binding sites.[33]

Sequence-derived Structural information. Solvent Accessible Surface Area (ASA) and Secondary Structure (SS) are two structural features that are highly related to binding. We used SPIDER 2.0,[44] a newly developed tool, to obtain predicted ASA at a residue level that is normalized by its residue's maximal possible ASA (rASA). We also calculated a window-averaged rASA. In addition, SPIDER 2.0 provides prediction of SS and three-state (helix, coil, and sheet) probability (*SSprob*). Based on the predicted residue-specific secondary structure, we evaluated the following segment-based features: the fraction of each SS type in a slide window (*SScont*), three-residue (the query and two nearest neighbors) 27-dimensional secondary-structural binary vector (*SStri*), the number of continuous residues containing the query residue in the same SS type segment (*SegLen*), and the position of the query residue in the SS segment from both ends.

Physicochemical properties. We utilized seven representative physicochemical features of amino acids.[45] These features are, namely, steric parameter, hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability.

Protein disorder region. Intrinsically Disordered Regions (IDRs) in proteins may be involved in binding to the target partner[46] by induced fit. We employed SPINE-D[47] to obtain predicted probability of being disordered for a given residue along the protein sequence.

Protein Length. We employed protein lengths as the only global feature in our feature vector.[48]

All the above seven feature groups for a query residue along with its neighboring residues within a sliding window were examined for their usefulness for carbohydrate-binding

prediction. Only some of these features will be selected for the final model as described below.

Support vector machines (SVM).

As shown in Figure 1, we use SVM with RBF kernel implemented in LibSVM[49] to build our predictive model. The performance of SVM with the RBF kernel depends on two parameters: *gamma* and *C*. We optimized these two parameters using a grid search implemented in LibSVM and chose the parameters that resulted in the highest MCC for our cross-validation set. The optimal values for *gamma* and *C* parameters along with the input window size were found to be 0.05, 1, and 4, respectively. For a window size of four residues at each side, the total number of features is 538. We further reduced the number of features by sequential forward feature selection (SFFS).[50] SFFS starts from the empty feature set and adds a feature or a feature group that yields the highest performance in each iteration until no further improvement can be made.

Cross validation and independent test.

We performed the protein-based ten-fold cross-validation on the training set (TR102). That is, proteins (not residues) in the training set were separated into 10 parts (folds). In each round nine folds were employed for training and one fold as test. The test fold consists of the unbalanced, full list of binding and non-binding residues while the training set (9 folds) can be the full set or more balanced sets by under-sampling or the SMOTE sampling technique depending on the methods. This process was repeated 10 times. The trained model was further tested on the independent test set (TS50) to confirm the generality of the developed method. This independent test set also has the full list of binding and nonbinding residues.

Performance Evaluation Criteria.

The overall performance of the method is assessed by Matthews Correlation Coefficient (MCC), accuracy, and sensitivity, specificity that are defined as below.

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (1)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivity = \frac{TP}{(TP + FN)} \quad (3)$$

$$Specificity = \frac{TN}{(FP + TN)} \quad (4)$$

where *TP* is the number of actual binding residues predicted as binding sites (True Positive), *TN* is the number of actual non-binding residues predicted as non-binding sites (True Negative), *FP* is the number of actual non-binding residues incorrectly predicted as binding sites (False Positive), and *FN* is the number of actual binding residues incorrectly predicted as non-binding sites (False Negative). In addition, we also employed the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). The MCC and AUC are balanced measures for unbalanced datasets.

3. RESULTS AND DISCUSSION

Table 1 and Figure 2A compare the performance of three approaches for handling the unbalanced datasets: under-sampling of the non-binding set, under-sampling of nonbinding and oversampling of binding sets (SMOTE), and the direct use of unbalanced sets. It shows that SMOTE improves over under-sampling in MCC and AUC values for the independent test while training on the unbalanced full datasets has the best performance. MCC values for the independent test set increase from 0.195 by under-sampling, 0.223 by SMOTE to 0.270 by non-sampling (the use of the full dataset). AUC values increase from 0.755, 0.762 to

0.767, respectively. We note that the change of AUC values between the cross validation and independent test is the smallest for training on the unbalanced full datasets, indicating the robustness of the training. Table 1 also shows consistent high accuracy for all methods because the test dataset is dominated by non-binding residues. Results of sensitivity and specificity are based on the threshold determined by maximizing MCC values. Low but reasonable sensitivity (~27%) was obtained at high specificity (98%) for the independent test set by training on unbalanced datasets. By comparison, if other approaches also set a specificity at 98% for a threshold, the sensitivities for the independent test are much lower at 0.166 for SMOTE and 0.174 for under-sampling, respectively.

Above results indicate that both under-sampling and the combination of under-sampling and oversampling do not perform as well as the direct training on the unbalanced dataset. This is largely because the model based on the under-sampling technique is not sufficiently trained against wide variety of negative samples. Thus, training on a dataset dominated by negative samples did not bias the method to over-predict non-binding residues. In addition to the under-sampling technique and the SMOTE method mentioned above, we have also examined the possibility of applying different weights for the minor class (binding residues). This technique, however, did not improve over our current model with the same weight to minor and major classes.

The above results were obtained by the full-feature set (538 features). To reduce possible overtraining, feature selection was performed by starting from the best feature group and then adding one feature group at a time. Here we focus on models trained on the unbalanced dataset only as they have the best performance. Four feature groups were selected as PSSM-based, Sequence-based, ASA-based, and protein length. As shown in Table 2 and Figure 2B,

the reduced feature model yields a slightly better performance in the ten-fold cross validation as well as in the independent test, while the differences between ten-fold cross validation and independent test set are essentially the same despite different number of selected features. The small difference in AUC between the cross-validation and independent tests further indicates the robustness of the method developed.

Table 3 compares the importance of four different feature groups in the final reduced-feature model by examining them individually and by removing them from the final model. The PSSM-based feature group has the best performance with MCC=0.241 as a single feature group (See Fig 2B) while removing the PSSM-based feature group will decrease MCC from 0.335 to 0.18. This performance is followed by the sequence-based feature, ASA-based and protein length. It is interesting that secondary-structure-based, physio-chemical properties, and predicted protein disordered region were removed during feature selection.

It is of interest to examine the performance on proteins binding with different types of carbohydrates. There are 15 mannose, 25 glucose, 32 galactose, 29 glucosamine, 20 amino, 6 sialic acids and 3 sulfated carbohydrates in the training/cross-validation set and 10 mannose, 12 glucose, 13 galactose, 14 glucosamine, 8 amino, 4 sialic acids and 3 sulfated carbohydrates for the independent test set, respectively. Because of the small sets and stable performance between training and independent test sets, we combined the proteins from two sets for analysis.

As shown in Table 4, all carbohydrate types except glucosamine, amino and sulfated carbohydrates have an AUC around 0.75 that is very close to the overall performance (0.77). The differences between individual and overall ROC curves from the independent test set are not statistically significant ($p > 0.05$) for mannose, glucose, galactose and sialic acid and

statistically significant for glucosamine, amino, and sulfated carbohydrates according to the significance of the difference between the areas under two independent ROC curves test.[51] Except sialic acids, the performance for three charged carbohydrates (glucosamine, amino, and sulfated carbohydrates) is lower, suggesting that it may be more difficult to predict the binding sites of charged molecules. However, the binding datasets for these specific carbohydrates are all too small to make a conclusive assessment. It may be beneficial to train binding sites of charged and uncharged carbohydrates separately when more data become available.

Figure 3 demonstrates two successful examples of actual versus predicted binding sites. Figure 3A shows the result of D-mannose-binding protein FimH of *E. coli* in the test set. From the structure (pdbid: 4cst), there are 11 binding residues over a total of 159 residues. Our prediction predicts eight binding residues that are all correct. Figure 3B illustrates another case of glucose-binding protein epithelial adhesin 1 A domain (Epa1A) from *Candida glabrata* (pdbid: 4af9). This 229-residue protein has seven binding sites. SPRINT-CBH predicts all binding sites plus one misclassified residue that is separated from the main actual binding region.

Our method for carbohydrate binding-site prediction employed features, such as PSSM and predicted ASA that are commonly employed for predicting other binding sites such as protein-peptide interactions by SPRINT-peptide.[33] Thus, it is of interest to know whether or not they are predicting the same binding sites. We found that applying SPRINT-peptide to our independent test set leads to 0.668 for AUC and 0.118 for MCC, compared to 0.772 and 0.285 by this work. On the other hand, applying this method to 50 protein-peptide complexes achieved 0.614 for AUC and 0.07 for MCC, compared to 0.687 for AUC and 0.182 for MCC

by SPRINT-peptide. This indicates that developing dedicated methods for binding sites specific for target molecules are necessary despite similar features were involved.

To further test our method, we obtained all human proteins that were annotated as lectin or carbohydrate binding from UniProt.[52] After mapping them to non-redundant (less than 30% sequence identity) CCDS protein set, 462 CCDS proteins are considered as annotated carbohydrate binding proteins, and the remaining 16946 proteins are considered as non-carbohydrate binding proteins. It should be noted that the set of non-carbohydrate binding proteins might contain proteins binding with carbohydrates because proteins usually have multiple functions. By the pre-determined threshold of 0.18, 1,568 binding residues were predicted out of 356,737 total numbers of residues for the carbohydrate-binding proteins (0.43%), which is 2.7 times more than those in presumed non-carbohydrate-binding proteins where 18,211 predicted binding residues out of a total number of 9,172,180 residues (0.19%). The difference is significant with P-value = 2.2e-16 by binominal test. The ability to predict more binding residues for actual binding proteins indicates the reliability of our method considering the fact that the method was not trained on non-binding proteins.

Mutation of carbohydrate-binding residues in a carbohydrate binding protein will likely affect its carbohydrate-binding capability and potentially have phenotypic implication. To examine this possibility, we investigate human mutations due to single-nucleotide variation (SNV). It is known that a frequently occurred SNV in a human population (high minor allele frequency) would be more fitted to its biological function and less likely to be associated with a disease than a rare allele[52-55]. Thus, if carbohydrate-binding residues were predicted correctly, we would expect that carbohydrate-binding residues are less likely mutated in frequent alleles for satisfying the requirement of functional fitness. We obtained single-

nucleotide variants (SNVs) along with their minor allele frequencies (MAF) collected by the 1000 Genomes Project.[56] We found that the odd of locating non-synonymous mutations at predicted binding sites in rare alleles (MAF<0.003) is 2.15 times more than in frequent alleles (MAF>0.003) (P-value=0.013 from fisher's exact test). As a control, we also examined mutation occurrence of synonymous mutations that do not change coded amino acid residues (i.e. less likely to have a functional impact). Indeed, the odd ratio is near 1 (1.1) for synonymous mutations occurred at predicted binding sites in rare alleles to that in frequent alleles (P-value= 0.79), confirming the functional impact of mutations in predicted carbohydrate-binding sites.

All above results were obtained by using 3.5 Å as a cutoff for defining binding residues. To examine the effect of the cutoff, we also built the datasets based on the cutoff of 6.0 Å (see Methods). We trained and tested our model with the same selected features based on unbalanced newly defined binding residues. The independent test of the new model yields 0.30 for MCC and 0.768 for AUC. This performance, which is similar to the performance with the cutoff of 3.5 Å (0.29 for MCC and 0.772 for AUC), supporting the robustness of our model for predicting carbohydrate-binding sites.

In summary, we have developed the first sequence-based method for predicting carbohydrate-binding sites that goes beyond evolution-derived sequence profiles. We have shown that incorporating additional sequence information, predicted solvent accessible surface area, and protein length by SVM leads to a method applicable directly to proteins (unbalanced data) with reasonable accuracy according to ten-fold cross validation and independent test. The quality of the method is further confirmed by its application to the human proteome and 1000 Genomes Project. Our method and datasets are available online and by standalone package.

However, it should be noted that the method is only useful to predict binding sites when carbohydrates bind to proteins non-covalently. Prediction of covalent-bound, glycosylation sites require separate methods (e.g. [57, 58])

ACKNOWLEDGMENTS

This work was supported by National Health and Medical Research Council (1059775 and 1083450) of Australia to Y.Z. The authors thank the Australian Research Council grant LE150100161 for infrastructure support. We also gratefully 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).

AUTHOR INFORMATION

Corresponding Author

* **Contact:** yuedong.yang@griffith.edu.au

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. All authors contributed equally.

REFERENCES

1. Shin, I.; Park, S.; Lee, M. r., Carbohydrate Microarrays: An Advanced Technology for Functional Studies of Glycans. *Chem. Eur. J.* **2005**, 11, 2894-2901.
2. McKinley, M. P.; O'Loughlin, V. D.; Pennefather-O'Brien, E.; Harris, R. T., *Human Anatomy*. Fourth edition ed.; McGraw-Hill Education: New York, NY, 2015.
3. Costerton, J.; Irvin, R.; Cheng, K., The Bacterial Glycocalyx in Nature and Disease. *Annu. Rev. Biol.* **1981**, 35, 299-324.

4. Brown, A.; Higgins, M. K., Carbohydrate Binding Molecules in Malaria Pathology. *Curr. Opin. Struct. Biol.* **2010**, 20, 560-566.
5. François, K.; Balzarini, J., Potential of Carbohydrate-Binding Agents as Therapeutics against Enveloped Viruses. *Med. Res. Rev.* **2012**, 32, 349-387.
6. Nakahara, S.; Raz, A., Biological Modulation by Lectins and Their Ligands in Tumor Progression and Metastasis. *Anti-Cancer Agents Med. Chem.* **2008**, 8, 22.
7. Ng, S.; Lin, E.; Kitov, P. I.; Tjhung, K. F.; Gerlits, O. O.; Deng, L.; Kasper, B.; Sood, A.; Paschal, B. M.; Zhang, P., Genetically Encoded Fragment-Based Discovery of Glycopeptide Ligands for Carbohydrate-Binding Proteins. *J. Am. Chem. Soc.* **2015**, 137, 5248-5251.
8. Taroni, C.; Jones, S.; Thornton, J. M., Analysis and Prediction of Carbohydrate Binding Sites. *Protein Eng.* **2000**, 13, 89-98.
9. Sujatha, M.; Balaji, P. V., Identification of Common Structural Features of Binding Sites in Galactose-Specific Proteins. *Proteins: Struct., Funct., Bioinf.* **2004**, 55, 44-65.
10. Kulharia, M.; Bridgett, S. J.; Goody, R. S.; Jackson, R. M., Inca-Sitefinder: A Method for Structure-Based Prediction of Inositol and Carbohydrate Binding Sites on Proteins. *J. Mol. Graphics Modell.* **2009**, 28, 297-303.
11. Nassif, H.; Al-Ali, H.; Khuri, S.; Keirouz, W., Prediction of Protein-Glucose Binding Sites Using Support Vector Machines. *Proteins: Struct., Funct., Bioinf.* **2009**, 77, 121-132.
12. Tsai, K.-C.; Jian, J.-W.; Yang, E.-W.; Hsu, P.-C.; Peng, H.-P.; Chen, C.-T.; Chen, J.-B.; Chang, J.-Y.; Hsu, W.-L.; Yang, A.-S., Prediction of Carbohydrate Binding Sites on Protein Surfaces with 3-Dimensional Probability Density Distributions of Interacting Atoms. *PloS one* **2012**, 7, e40846.
13. Malik, A.; Ahmad, S., Sequence and Structural Features of Carbohydrate Binding in Proteins and Assessment of Predictability Using a Neural Network. *BMC Struct. Biol.* **2007**, 7, 1.
14. Pai, P. P.; Mondal, S., Mowgli: Prediction of Protein–Mannose Interacting Residues with Ensemble Classifiers Using Evolutionary Information. *J. Biomol. Struct. Dyn.* **2015**, 1-15.
15. Agarwal, S.; Mishra, N. K.; Singh, H.; Raghava, G. P., Identification of Mannose Interacting Residues Using Local Composition. *PLoS One* **2011**, 6, e24039.
16. Deng, L.; Guan, J.; Wei, X.; Yi, Y.; Zhang, Q. C.; Zhou, S., Boosting Prediction Performance of Protein–Protein Interaction Hot Spots by Using Structural Neighborhood Properties. *J. Comput. Biol.* **2013**, 20, 878-891.
17. Lei, C.; Ruan, J., A Novel Link Prediction Algorithm for Reconstructing Protein–Protein Interaction Networks by Topological Similarity. *Bioinformatics* **2013**, 29, 355-364.
18. Pierce, B. G.; Wiehe, K.; Hwang, H.; Kim, B.-H.; Vreven, T.; Weng, Z., Zdock Server: Interactive Docking Prediction of Protein–Protein Complexes and Symmetric Multimers. *Bioinformatics* **2014**, 30, 1771-1773.
19. Rao, V. S.; Srinivas, K.; Sujini, G.; Kumar, G., Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics* **2014**, 2014.
20. Liang, S.; Zhang, C.; Liu, S.; Zhou, Y., Protein Binding Site Prediction Using an Empirical Scoring Function. *Nucleic Acids Res.* **2006**, 34, 3698-3707.
21. Lin, C.-K.; Chen, C.-Y., Pidna: Predicting Protein–DNA Interactions with Structural Models. *Nucleic Acids Res.* **2013**, 41, W523-W530.
22. Si, J.; Zhang, Z.; Lin, B.; Schroeder, M.; Huang, B., Metadbsite: A Meta Approach to Improve Protein DNA-Binding Sites Prediction. *BMC Syst. Biol.* **2011**, 5, S7.
23. Zhou, W.; Yan, H., A Discriminatory Function for Prediction of Protein–DNA Interactions Based on Alpha Shape Modeling. *Bioinformatics* **2010**, 26, 2541-2548.

24. Zhao, H.; Wang, J.; Zhou, Y.; Yang, Y., Predicting DNA-Binding Proteins and Binding Residues by Complex Structure Prediction and Application to Human Proteome. *PloS one* **2014**, 9, e96694.
25. Murakami, Y.; Spriggs, R. V.; Nakamura, H.; Jones, S., Piranha: A Server for the Computational Prediction of Rna-Binding Residues in Protein Sequences. *Nucleic Acids Res.* **2010**, 38, W412-W416.
26. Zhang, T.; Zhang, H.; Chen, K.; Ruan, J.; Shen, S.; Kurgan, L., Analysis and Prediction of Rna-Binding Residues Using Sequence, Evolutionary Conservation, and Predicted Secondary Structure and Solvent Accessibility. *Curr. Protein Pept. Sci.* **2010**, 11, 609-628.
27. Zhao, H.; Yang, Y.; Zhou, Y., Highly Accurate and High-Resolution Function Prediction of Rna Binding Proteins by Fold Recognition and Binding Affinity Prediction. *RNA Biol.* **2011**, 8, 988-996.
28. Bolia, A.; Gerek, Z. N.; Ozkan, S. B., Bp-Dock: A Flexible Docking Scheme for Exploring Protein–Ligand Interactions Based on Unbound Structures. *J. Chem. Inf. Model.* **2014**, 54, 913-925.
29. Komiyama, Y.; Banno, M.; Ueki, K.; Saad, G.; Shimizu, K., Automatic Generation of Bioinformatics Tools for Predicting Protein–Ligand Binding Sites. *Bioinformatics* **2015**, 32, 901-907.
30. Yang, Y.; Zhan, J.; Zhou, Y., Spot-Ligand: Fast and Effective Structure-Based Virtual Screening by Binding Homology Search According to Ligand and Receptor Similarity. *J. Comput. Chem.* **2016**, 37, 1734-1739.
31. Lavi, A.; Ngan, C. H.; Movshovitz-Attias, D.; Bohnuud, T.; Yueh, C.; Beglov, D.; Schueler-Furman, O.; Kozakov, D., Detection of Peptide-Binding Sites on Protein Surfaces: The First Step toward the Modeling and Targeting of Peptide-Mediated Interactions. *Proteins: Struct., Funct., Bioinf.* **2013**, 81, 2096-2105.
32. Petsalaki, E.; Stark, A.; García-Urdiales, E.; Russell, R. B., Accurate Prediction of Peptide Binding Sites on Protein Surfaces. *PLoS Comput. Biol.* **2009**, 5, e1000335.
33. Taherzadeh, G.; Yang, Y.; Zhang, T.; Liew, A. W. C.; Zhou, Y., Sequence-Based Prediction of Protein–Peptide Binding Sites Using Support Vector Machine. *J. Comput. Chem.* **2016**, 37, 1223-1229.
34. Zhao, H.; Yang, Y.; von Itzstein, M.; Zhou, Y., Carbohydrate-Binding Protein Identification by Coupling Structural Similarity Searching with Binding Affinity Prediction. *J. Comput. Chem.* **2014**, 35, 2177-2183.
35. Pérez, S.; Sarkar, A.; Rivet, A.; Breton, C.; Imberty, A., Glyco3d: A Portal for Structural Glycosciences. *Glycoinformatics* **2015**, 241-258.
36. Malik, A.; Firoz, A.; Jha, V.; Ahmad, S., Procarb: A Database of Known and Modelled Carbohydrate-Binding Protein Structures with Sequence-Based Prediction Tools. *Adv. Bioinf.* **2010**, 436036.
37. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L., Ncbi Blast: A Better Web Interface. *Nucleic Acids Res.* **2008**, 36, W5-W9.
38. Miao, Z.; Westhof, E., Prediction of Nucleic Acid Binding Probability in Proteins: A Neighboring Residue Network Based Score. *Nucleic Acids Res.* **2015**, gkv446.
39. Yen, S.-J.; Lee, Y.-S. Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset. In *Intelligent Control and Automation*; Springer Berlin Heidelberg: 2006, pp 731-740.
40. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P., Smote: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, 16, 321-357.

41. Chen, K.; Mizianty, M. J.; Kurgan, L., Prediction and Analysis of Nucleotide-Binding Residues Using Sequence and Sequence-Derived Structural Descriptors. *Bioinformatics* **2012**, 28, 331-341.
42. Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N., Consurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics* **2003**, 19, 163-164.
43. 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.* **1997**, 25, 3389-3402.
44. Heffernan, R.; Paliwal, K.; Lyons, J.; Dehzangi, A.; Sharma, A.; Wang, J.; Sattar, A.; Yang, Y.; Zhou, Y., Improving Prediction of Secondary Structure, Local Backbone Angles, and Solvent Accessible Surface Area of Proteins by Iterative Deep Learning. *Sci. Rep.* **2015**, 5, 11476.
45. Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F., Generation and Evaluation of Dimension-Reduced Amino Acid Parameter Representations by Artificial Neural Networks. *J. Mol. Model.* **2001**, 7, 360-369.
46. Hsu, W. L.; Oldfield, C. J.; Xue, B.; Meng, J.; Huang, F.; Romero, P.; Uversky, V. N.; Dunker, A. K., Exploring the Binding Diversity of Intrinsically Disordered Proteins Involved in One-to-Many Binding. *Protein Sci.* **2013**, 22, 258-273.
47. Zhang, T.; Faraggi, E.; Xue, B.; Dunker, A. K.; Uversky, V. N.; Zhou, Y., Spine-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method. *J. Biomol. Struct. Dyn.* **2012**, 29, 799-813.
48. Wang, K.; Gao, J.; Shen, S.; Tuszynski, J. A.; Ruan, J.; Hu, G., An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein Surface Using Svm and Statistical Depth Function. *BioMed Res. Int.* **2013**, 2013, 409658.
49. Chang, C.-C.; Lin, C.-J., Libsvm: A Library for Support Vector Machines. *ACM. TIST.* **2011**, 2, 27.
50. Kudo, M.; Sklansky, J., Comparison of Algorithms That Select Features for Pattern Classifiers. *Pattern Recognit.* **2000**, 33, 25-41.
51. DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L., Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, 44, 837-845.
52. Consortium, U., Uniprot: A Hub for Protein Information. *Nucleic Acids Res.* **2014**, 43, 204-212.
53. Hu, J.; Ng, P. C., Predicting the Effects of Frameshifting Indels. *Genome Biol* **2012**, 13, R9.
54. Zhao, H.; Yang, Y.; Lin, H.; Zhang, X.; Mort, M.; Cooper, D. N.; Liu, Y.; Zhou, Y., Ddig-In: Discriminating between Disease-Associated and Neutral Non-Frameshifting Micro-Indels. *Genome Biol.* **2013**, 14, R23.
55. Folkman, L.; Yang, Y.; Li, Z.; Stantic, B.; Sattar, A.; Mort, M.; Cooper, D. N.; Liu, Y.; Zhou, Y., Ddig-In: Detecting Disease-Causing Genetic Variations Due to Frameshifting Indels and Nonsense Mutations Employing Sequence and Structural Properties at Nucleotide and Protein Levels. *Bioinformatics* **2015**, 31, 1599-1606.
56. Consortium, G. P., A Global Reference for Human Genetic Variation. *Nature* **2015**, 526, 68-74.
57. Hamby, S. E.; Hirst, J. D., Prediction of Glycosylation Sites Using Random Forests. *BMC Bioinf.* **2008**, 9, 1.
58. Caragea, C.; Sinapov, J.; Silvescu, A.; Dobbs, D.; Honavar, V., Glycosylation Site Prediction Using Ensembles of Support Vector Machine Classifiers. *BMC Bioinf.* **2007**, 8, 1.

Table 1 Performance of the ten-fold cross-validation and independent test on the full and balanced datasets by SVM models.

Training Set		MCC	AUC	Accuracy	Sensitivity	Specificity
Under-sampling	CV[a]	0.250	0.773	0.950	0.180(0.229[b])	0.989(0.979)
	Test[c]	0.195	0.755	0.906	0.389(0.174[b])	0.925(0.979)
SMOTE	CV[a]	0.260	0.752	0.964	0.150(0.265[b])	0.999(0.979)
	Test[c]	0.223	0.762	0.958	0.18(0.166[b])	0.986(0.979)
Unbalanced	CV[a]	0.330	0.765	0.966	0.234(0.274[b])	0.985(0.979)
	Test[c]	0.270	0.767	0.954	0.266	0.979

[a]Cross validated on the full training set; [b]sensitivity when thresholds are set by fixing specificity at 97.9%; [c]Results on the full independent test set.

Table 2. Performance of the ten-fold cross-validation and independent test on the unbalanced dataset by SVM models before and after feature selections.

# Features		MCC	AUC	Accuracy	Sensitivity	Specificity
538	CV[a]	0.330	0.765	0.966	0.234(0.226[b])	0.985(0.988)
	Test[c]	0.270	0.767	0.954	0.266(0.197[b])	0.979(0.988)
400	CV[a]	0.335	0.777	0.965	0.188(0.238[b])	0.996(0.988)
	Test[c]	0.285	0.772	0.961	0.223	0.988

[a]Cross validated on the full training set; [b]sensitivity when thresholds are set by fixing specificity at 97.9%; [c]Results on the full independent test set.

Table 3. The performance of four individual feature groups in the final reduced model for the unbalanced data set along with the result of removing the feature group for the reduced model.

Features Group (Final model)	Individual Feature Group Performance				Remove the feature group from the reduced model			
	MCC	AUC	Sensitivity	Specificity	MCC	AUC	Sensitivity	Specificity
PSSM-based	0.241	0.734	0.209	0.984	0.183	0.713	0.082	0.996
Sequence-based	0.134	0.677	0.129	0.981	0.243	0.752	0.166	0.991
ASA-based	0.053	0.58	0.836	0.295	0.281	0.771	0.215	0.989
Protein Length	0.03	0.52	0.90	0.22	0.281	0.771	0.215	0.989

Table 4. The results of the reduced-feature model on the unbalanced dataset in different types of carbohydrates.

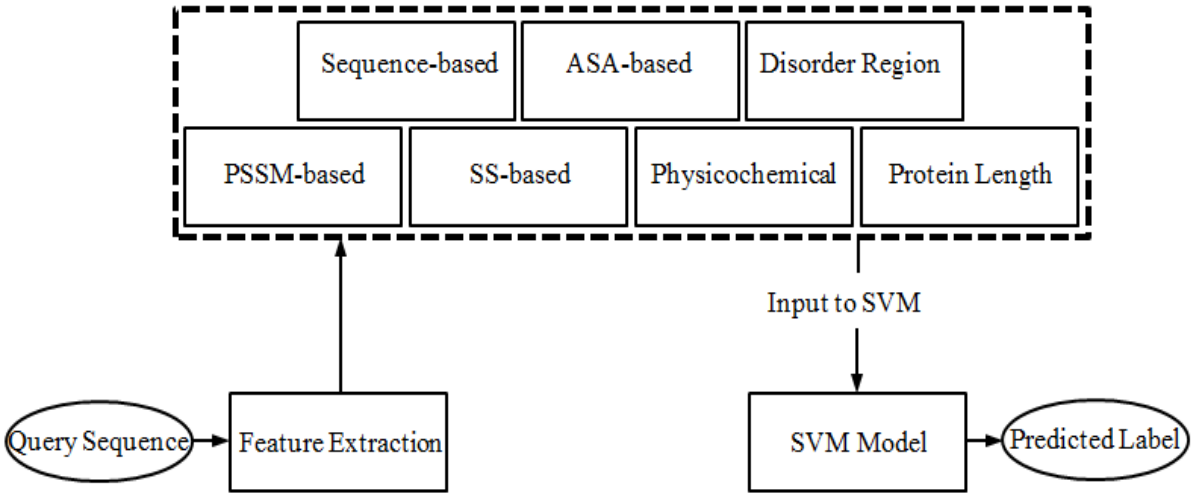
Types	#	P-value	AUC	MCC	Accuracy	Sensitivity	Specificity
Test Only	50	NA	0.772	0.285	0.961	0.223	0.988
Glucose	37	0.958	0.773	0.209	0.940	0.305	0.960
Mannose	25	0.516	0.758	0.344	0.956	0.227	0.992
Galactose	45	0.326	0.753	0.306	0.969	0.167	0.996
Sialic acids	10	0.4255	0.745	0.24	0.972	0.095	0.99
Sulfated	6	0.0184	0.697	0.2	0.904	0.365	0.926
Glucosamine	43	0.000049	0.694	0.214	0.956	0.133	0.991
Amino	28	0.000029	0.682	0.184	0.95	0.144	0.985

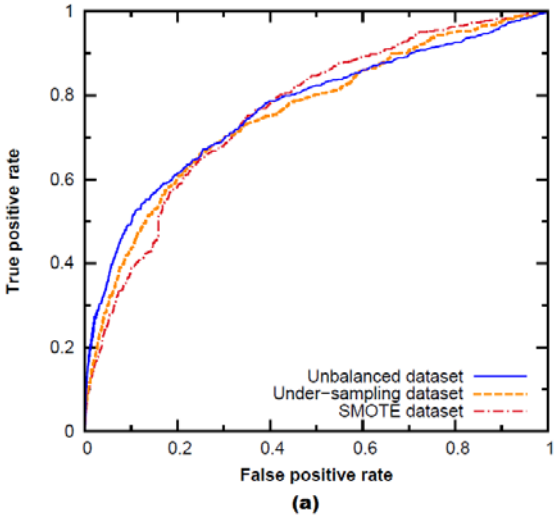
Figure Captions

Figure 1 The flowchart of SPRINT-CBH. Over a query sequence, seven feature groups are extracted, and input to the trained SVM model for prediction.

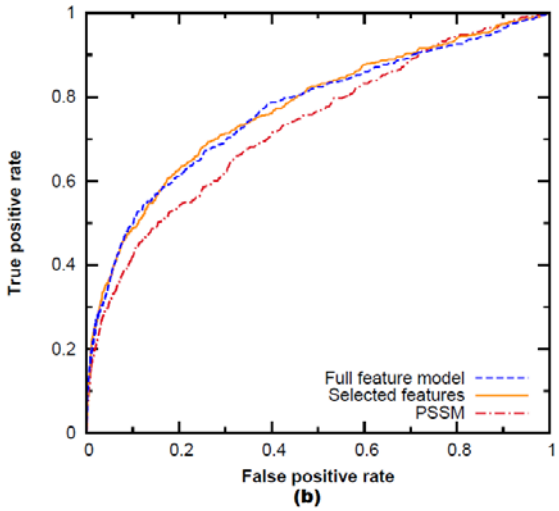
Figure 1 a) Receiver operating characteristic curves on the unbalanced, full test set by using the trained dataset produced by under-sampling, SMOTE as well as the unbalanced, full dataset as labelled. b) Receiver operating characteristic curves on the unbalanced, full test set by using PSSM only, the full-feature model, and the reduced-feature model all trained on the unbalanced, full dataset.

Figure 3 a) The actual (Left) versus predicted (Right) binding sites (in red) of a) D-mannose-binding protein FimH of *E. coli* (PDBID: 4cst) in the test set and b) Epithelial Adhesin from *Candida glabrata* (PDBID: 4af9) in the test set.


Sequence-based
ASA-based
Disorder Region
PSSM-based
SS-based
Physicochemical
Protein Length
Input to SVM
Query Sequence
Feature Extraction
SVM Model
Predicted Label

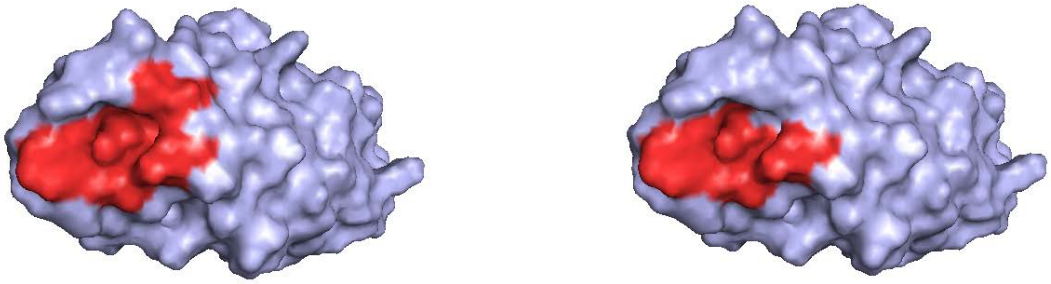

True positive rate
False positive rate
Unbalanced dataset
Under−sampling dataset
SMOTE dataset
0
0.2
0.4
0.6
0.8
1

(a)

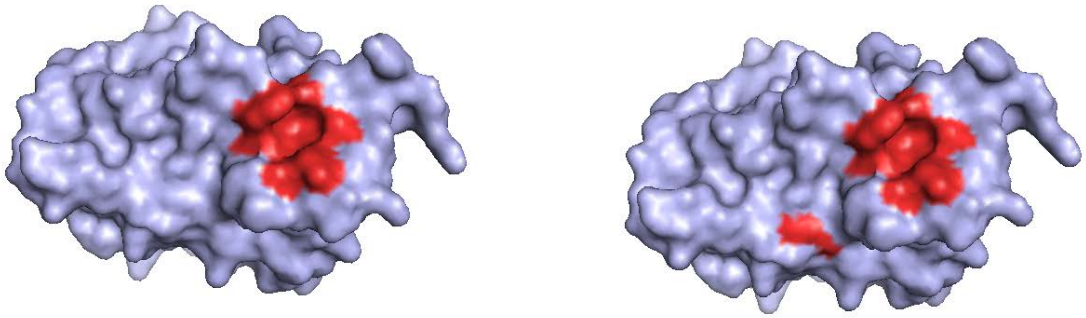

True positive rate
False positive rate
Full feature model
Selected features
PSSM
0
0.2
0.4
0.6
0.8
1

(b)

a) 4cst



b)4af9



For Table of Contents Use Only