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

Ghazaleh Taherzadeh,† Yaoqi Zhou,†,‡ Alan Wee-Chung Liew,† and Yuedong Yang*†,‡

†School of Information and Communication Technology and ‡Institute for Glycomics, Griffith University, Parklands Drive, Southport, Queensland 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 using support vector machines (SVMs). We found that integrating evolution-derived sequence profiles with additional information on sequence and predicted solvent accessible surface area leads to a reasonably accurate, robust, and 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 10-fold cross validation and independent test without balancing binding and nonbinding residues. The quality of the method is further demonstrated by having statistically significantly more binding residues predicted for carbohydrate-binding proteins than presumptive nonbinding proteins in the human proteome, and by the bias of rare alleles toward predicted carbohydrate-binding sites for nonsynonymous mutations from the 1000 genome project. SPRINT-CBH is available as an online server at http://sparks-lab.org/server/SPRINT-CBH.

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. 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. It is also the first layer of protection for pathogens. 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. Protein−carbohydrate interactions, however, are challenging to study experimentally because of weak binding affinity and synthetic complexity of specific carbohydrates. 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. 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 data set 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. Agarwal et al. also developed a similar method for predicting mannose-binding sites by using PSSM. Thus, for predicting nonspecific carbohydrate-binding sites, there exists only one sequence-based method. The method was subjected to a limited test (leave-one-out) and relied on sequence profiles from PSSM only. Moreover, lacking an online 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, protein–DNA, protein–RNA, protein–ligand, and protein–peptide interactions. Many sequence-based techniques above have shown that integration of PSSM with physicochemical 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 noncovalent 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 (SVMs). 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 data sets contain significantly more nonbinding residues than binding residues, we found that the model developed by direct training on unbalanced full data sets improved over the methods trained on more balanced data sets by employing under- and oversampling techniques. The quality of the method was further confirmed by similar performance of cross validation and independent test on the full data set, the application to a protein–peptide binding data set as a control, 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 using its corresponding protein–carbohydrate complex structure. We defined a residue as a carbohydrate-binding site if any atom in the residue is within 3.5 Å from any carbohydrate atom. For the 152 carbohydrate-binding proteins we obtained 1530 binding and 39 484 nonbinding residues. Because this data set has 26 times more nonbinding residues than binding residues, we have balanced the TR102 data set using undersampling and SMOTE techniques and compared the methods trained on more balanced data sets with the method trained on the full data set.

Undersampling was done by randomly selecting a portion of nonbinding residues so that its number is the same as the number of binding sites for each protein. This balanced data set with the ratio of 1:1 contains 1042 binding and 1042 nonbinding residues in TR102. The SMOTE technique, on the other hand, undersampling the majority class and oversampling the minority class, to balance the TR102 data set. The balanced data set produced by SMOTE has 20 816 residues with the ratio of ≈1:4 (4168 binding and 16 648 nonbinding residues). We would like to emphasize that all balanced data sets are only utilized for training and the full data set 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 3.5 Å for defining binding residues is somewhat arbitrary. Thus, we also employed a cutoff of 6 Å. This nearly doubled the number of binding residues (3044 binding and 37 970 nonbinding residues) for 152 carbohydrate-binding proteins. The ratio (binding:nonbinding) 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, we have also employed a one-dimensional vector as terminus indicator:terminal residues (the first and last three residues) are represented by 1 and other residues are by 0.

Evolutionary Information. Evolutionary conserved residues may have functional roles such as binding. Here we employed the position specific scoring matrix (PSSM) which is a 20L dimensional matrix (where L is protein length) generated from PSI-BLAST using E-value threshold of 0.001 in three iterations. In addition, we calculated information entropy from PSSM, $S_\theta = \sum_{i=1}^{20} P_{ij} \times \ln(P_{ij})$, where $P_{ij}$ 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.

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, 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. 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 by induced fit. We employed SPINE-D 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.

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. As shown in Figure 1, we use SVM with RBF kernel implemented in LibSVM 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). 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 10-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 nonbinding residues while the training set (9 folds) can be the full set or more balanced sets by undersampling 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.

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

(1)

\[
\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]  

(2)

\[
\text{sensitivity} = \frac{TP}{TP + FN}
\]  

(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 nonbinding residues predicted as nonbinding sites (true negative), FP is the number of actual nonbinding residues incorrectly predicted as binding sites (false positive), and FN is the number of actual binding residues incorrectly predicted as nonbinding 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 data sets.

3. RESULTS AND DISCUSSION

Table 1 and Figure 2A compare the performance of three approaches for handling the unbalanced data sets: undersampling of the nonbinding set, undersampling of nonbinding, and oversampling of binding sets (SMOTE), and the direct use of unbalanced sets. It shows that SMOTE improves over undersampling in MCC and AUC values for the independent test while training on the unbalanced full data sets has the best performance. MCC values for the independent test set increase from 0.195 by undersampling, 0.223 by SMOTE to 0.270 by nonsampling (the use of the full data set). 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 data sets, indicating the robustness of the training. Table 1 also shows consistent high accuracy for all methods because the test data set is dominated by nonbinding 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 data sets. 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 undersampling, respectively.

Above results indicate that both undersampling and the combination of under- and oversampling do not perform as well as the direct training on the unbalanced data set. This is largely because the model based on the undersampling technique is not sufficiently trained against wide variety of negative samples. Thus, training on a data set dominated by negative samples did not bias the method to overpredict nonbinding residues. In addition to the undersampling 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 data set 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 10-fold cross-validation as well as in the independent test, while the differences between 10-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 Figure 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

Figure 2. (a) Receiver operating characteristic curves on the unbalanced, full test set by using the trained data set produced by undersampling, SMOTE as well as the unbalanced, full data set as labeled. (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 data set.
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

Table 2. Performance of the 10-Fold Cross-Validation and Independent Test on the Unbalanced Dataset by SVM Models before and after Feature Selections

<table>
<thead>
<tr>
<th>no. of features</th>
<th>MCC</th>
<th>AUC</th>
<th>accuracy</th>
<th>sensitivity</th>
<th>specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>538 CV&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.330</td>
<td>0.765</td>
<td>0.966</td>
<td>0.234 (0.226&lt;sup&gt;b&lt;/sup&gt;)</td>
<td>0.985 (0.988)</td>
</tr>
<tr>
<td>400 Test&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.270</td>
<td>0.767</td>
<td>0.954</td>
<td>0.266 (0.197&lt;sup&gt;c&lt;/sup&gt;)</td>
<td>0.979 (0.988)</td>
</tr>
<tr>
<td>400 CV&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.335</td>
<td>0.777</td>
<td>0.965</td>
<td>0.188 (0.238&lt;sup&gt;c&lt;/sup&gt;)</td>
<td>0.996 (0.988)</td>
</tr>
<tr>
<td>400 Test&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.285</td>
<td>0.772</td>
<td>0.961</td>
<td>0.223</td>
<td>0.988</td>
</tr>
</tbody>
</table>

<sup>a</sup>Cross-validated on the full training set. <sup>b</sup>Sensitivity when thresholds are set by fixing specificity at 97.9%. <sup>c</sup>Results on the full independent test set.

Table 3. 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

<table>
<thead>
<tr>
<th>features group (final model)</th>
<th>individual feature group performance</th>
<th>remove the feature group from the reduced model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MCC</td>
<td>AUC</td>
</tr>
<tr>
<td>PSSM-based</td>
<td>0.241</td>
<td>0.734</td>
</tr>
<tr>
<td>sequence-based</td>
<td>0.134</td>
<td>0.677</td>
</tr>
<tr>
<td>ASA-based</td>
<td>0.053</td>
<td>0.58</td>
</tr>
<tr>
<td>protein length</td>
<td>0.03</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 4. Results of the Reduced-Feature Model on the Unbalanced Dataset in Different Types of Carbohydrates

<table>
<thead>
<tr>
<th>types</th>
<th>no.</th>
<th>P-value</th>
<th>AUC</th>
<th>MCC</th>
<th>accuracy</th>
<th>sensitivity</th>
<th>specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>test only</td>
<td>50</td>
<td>NA</td>
<td>0.772</td>
<td>0.285</td>
<td>0.961</td>
<td>0.223</td>
<td>0.988</td>
</tr>
<tr>
<td>glucose</td>
<td>37</td>
<td>0.958</td>
<td>0.773</td>
<td>0.209</td>
<td>0.940</td>
<td>0.305</td>
<td>0.960</td>
</tr>
<tr>
<td>mannose</td>
<td>25</td>
<td>0.516</td>
<td>0.758</td>
<td>0.344</td>
<td>0.956</td>
<td>0.227</td>
<td>0.992</td>
</tr>
<tr>
<td>galactose</td>
<td>45</td>
<td>0.326</td>
<td>0.753</td>
<td>0.306</td>
<td>0.969</td>
<td>0.167</td>
<td>0.996</td>
</tr>
<tr>
<td>sialic acids</td>
<td>10</td>
<td>0.4255</td>
<td>0.745</td>
<td>0.24</td>
<td>0.972</td>
<td>0.095</td>
<td>0.99</td>
</tr>
<tr>
<td>sulfated</td>
<td>6</td>
<td>0.0184</td>
<td>0.697</td>
<td>0.2</td>
<td>0.904</td>
<td>0.365</td>
<td>0.926</td>
</tr>
<tr>
<td>glucosamine</td>
<td>43</td>
<td>0.000049</td>
<td>0.694</td>
<td>0.214</td>
<td>0.956</td>
<td>0.133</td>
<td>0.991</td>
</tr>
<tr>
<td>amino</td>
<td>28</td>
<td>0.000029</td>
<td>0.682</td>
<td>0.184</td>
<td>0.95</td>
<td>0.144</td>
<td>0.985</td>
</tr>
</tbody>
</table>

Figure 3. (a) Actual (left) versus predicted (right) binding sites (in red) of (a) D-mannose-binding protein FimH of *E. coli* (pdb ID: 4cst) in the test set and (b) epithelial adhesin from *Candida glabrata* (pdb ID: 4af9) in the test set.
statistically significant for glucosamine, amino, and sulfated carbohydrates according to the significance of the difference between the areas under two independent ROC curves test.\textsuperscript{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 data sets 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 Dmannose-binding protein FimH of \textit{E. coli} in the test set. From the structure (pdb ID: 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 (Ep1A) from \textit{Candida glabrata} (pdb ID: 4a9). 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.\textsuperscript{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.\textsuperscript{52} After mapping them to nonredundant (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 noncarbohydrate binding proteins. It should be noted that the set of noncarbohydrate binding proteins might contain proteins binding with carbohydrates because proteins usually have multiple functions. By the predetermined threshold of 0.18, 1568 binding residues were predicted out of 356737 total numbers of residues for the carbohydrate-binding proteins (0.43\%), which is 2.7 times more than those in presumed noncarbohydrate-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 \textit{P}-value = 2.2 × 10^{-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 nonbinding 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.\textsuperscript{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.\textsuperscript{56} We found that the odd of locating nonsynonymous mutations at predicted binding sites in rare alleles (MAF < 0.003) is 2.15 times more than in frequent alleles (MAF > 0.003) (\textit{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 (\textit{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 data sets based on the cutoff of 6.0 Å (see Materials and 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 10-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 data sets 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 noncovalently. Prediction of covalent-bound, glycosylation sites require separate methods (e.g., refs 57 and 58).

\section*{AUTHOR INFORMATION}

\textbf{Corresponding Author}
\textsuperscript{*}E-mail: yuedong.yang@griffith.edu.au.

\textbf{Notes}
The authors declare no competing financial interest.

\section*{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).
REFERENCES


