**Carbohydrate-binding protein identification by coupling structural similarity searching with binding affinity prediction**

Huiying Zhao\textsuperscript{a,b}, Yuedong Yang\textsuperscript{a,b,c,*}, Mark von Itzstein\textsuperscript{c} and Yaoqi Zhou\textsuperscript{a,b,c,d,*}

\textsuperscript{a}Indiana University School of Informatics, Indiana University Purdue University, Indianapolis, 719 Indiana Ave, Suite 319, Indianapolis, IN 46202

\textsuperscript{b}Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202

\textsuperscript{c}Institute for Glycomics, Griffith University, Southport, QLD 4222, Australia

\textsuperscript{d}School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia

*Co-corresponding authors (yuedong.yang@griffith.edu.au and yaoqi.zhou@griffith.edu.au).

**ABSTRACT**

Carbohydrate-binding proteins (CBPs) are potential biomarkers and drug targets. However, the interactions between carbohydrates and proteins are challenging to study experimentally and computationally because of their low binding affinity, high flexibility, and the lack of a linear sequence in carbohydrates as exists in RNA, DNA, and proteins. Here, we describe a structure-based function-prediction technique called SPOT-Struc that identifies carbohydrate-recognizing proteins and their binding amino acid residues by structural alignment program SPalign and binding affinity scoring according to a knowledge-based statistical potential for binding affinity scoring based on the distance-scaled finite-ideal gas reference state (DFIRE). The leave-one-out cross-validation of the method on 113 carbohydrate-binding domains and 3442 non-carbohydrate binding proteins yields a Matthews correlation coefficient of 0.56 for SPalign alone and 0.63 for SPOT-Struc (SPalign + binding affinity scoring) for CBP prediction. SPOT-Struc is a technique with high positive predictive value (79\% correct predictions in all positive CBP predictions) with a reasonable sensitivity (52\% positive predictions in all CBPs). The sensitivity of the method was changed slightly when applied to 31 APO (unbound) structures found in the protein databank (14/31 for APO versus 15/31 for HOLO). The result of SPOT-Struc will not change significantly if highly homologous templates were used. SPOT-Struc predicted 19 out of 2076 structural genome targets as CBPs.
In particular, one uncharacterized protein in *Bacillus subtilis* (1oq1A) was matched to galectin-9 from *Mus musculus*. Thus, SPOT-Struc is useful for uncovering novel carbohydrate-binding proteins. SPOT-Struc is available at [http://sparks-lab.org](http://sparks-lab.org).

### Introduction

Carbohydrates are associated with many biological processes and perform essential roles at the cellular level in living organisms. These roles are typically through engagement with proteins by both non-covalent (carbohydrate-protein binding) and covalent (glycosylation) interactions. Glycoproteins and glycolipids decorate the surfaces of all living cells and, consequently, tissues with a variety of carbohydrates to form the cell’s glycocalyx. The spatial patterns of such carbohydrate decorations alter during cell development\(^1\) and in tumor progression and metastasis\(^2,3\). Thus, recognition of cell-surface carbohydrates by carbohydrate-binding proteins (CBPs) is the subject of intensive studies particular in the area of biomarker discovery and inhibitor design\(^2,4\). Moreover, cell surface resident carbohydrates are also exploited by carbohydrate-binding proteins associated with clinically significant pathogens, including various bacteria, viruses and parasites, for cell invasion and immune detection avoidance\(^5\). As a result, many CBPs in pathogens have been conceived as potential drug targets\(^6\).

However, not all CBPs and their binding mechanisms are known. Moreover, more and more proteins have been sequenced without knowledge of their function(s) due to increasingly inexpensive sequencing techniques. Although there is one high-throughput technique (glycan arrays) for detecting novel CBPs and investigating their binding specificity\(^7-11\), it is challenging to construct a sizeable, diverse glycan array because of difficulty in synthesis and isolation of carbohydrates. Here, we focus on a complementary approach: prediction of CBPs and their binding amino acid residues by computational techniques.

Currently, predicting CBPs and determining their carbohydrate-binding amino acid residues are treated as two separate problems\(^12-16\). Someya *et al.*\(^12\) predicted carbohydrate-binding proteins by combining protein sequence information with support vector machines (SVM). This approach employed sequence patterns and frequencies of grouped amino acids as input and has achieved a Matthews correlation coefficient of 0.67 by leave-one-out cross-validation based on a dataset of 345 CBPs and non-CBPs. This method is limited to CBP prediction. Many more methods have been developed for predicting carbohydrate-binding amino acid residues based on their three-dimensional structures. For example, Shionyu-Mitsuyama *et al.*
predicted binding residues by building empirical interaction rules\textsuperscript{13}. Tsai \textit{et al.} utilized 3D probability density maps\textsuperscript{15}. Machine-learning techniques have also been established by utilizing binding propensity and solvent accessibility\textsuperscript{17} or selected geometric and chemical features\textsuperscript{18}. These methods, however, cannot distinguish CBPs from non-CBPs.

Here, we will introduce a template-based method that can simultaneously predict CBPs and their carbohydrate-binding amino acid residues. This work is inspired by our template-based technique named SPOT-Struc for structure-based prediction of DNA-/RNA- binding proteins and their binding sites\textsuperscript{19,20}. In this approach, a target structure is first structurally aligned to the proteins with known protein-RNA/DNA complex structures. Significantly aligned structures are then employed for building model complex structures between target structure and template RNA/DNA and for predicting binding affinities. Binding affinity calculations are employed as a filter to recognize true positives and remove false positives.

In this work, we extend SPOT-Struc to CBPs. Such an extension is possible because hundreds of protein-carbohydrate complex structures have been deposited in the protein databank\textsuperscript{21} despite experimental challenges as a consequence of the low binding affinity and the highly flexible structures of carbohydrates. This dataset of CBP complex structures is employed as templates to predict CBPs based on structural similarity between a template and a target structure by SPalign\textsuperscript{22}. Structural similarity is further combined with binding affinity scoring by a knowledge-based energy function proved useful for proteins\textsuperscript{23,24}, protein-DNA/RNA\textsuperscript{19,20} and protein-ligand interactions\textsuperscript{25}. Our method was tested by leave-homolog-out on 113 non-redundant CBPs and 3442 non-CBPs and achieved the Matthews correlation coefficients of 0.63 and 0.58 for prediction of CBPs and carbohydrate-binding amino acid residues, respectively. The sensitivity and positive predictive value (PPV) of CBP prediction are 52\% and 79\%, respectively. A similar level of sensitivity was achieved for APO and HOLO structures. Application of this method to structural genomics targets revealed several new putative CBPs.

\textbf{Methods and Materials}

\textbf{Template library of carbohydrate-binding proteins (T523).} A template library was built based on the PROCARB database that contains 604 protein-carbohydrate complex structures\textsuperscript{26}. We then removed all glycoproteins and selected only those proteins with more than 5 amino acid residues binding with carbohydrates. Here, a residue is defined as a carbohydrate-binding amino acid residue if it has one or more heavy atoms that are within 4.5
distance from any heavy atoms of carbohydrates as used for defining DNA and RNA-binding residues. We further divided selected proteins into domains by the automatic domain parser DDomain. Both domains and their corresponding chains are included in the final template library that has 523 CBPs as templates. We have included both domains and chains in the template library so as to increase the possibility of locating a suitable template.

**Positive binding-domain dataset (BD113):** We built a positive database of carbohydrate-binding domains for training and cross-validation by firstly excluding the chains in T523. We further removed the redundant proteins by using BLASTClust with a sequence identity cutoff of 30%. The final dataset contains 113 CBPs.

**Negative (non-binding) dataset (NB3442):** We built the negative dataset by querying the PDB database and removing all PDB files containing carbohydrates. The protein chains were parsed into domains by DDomain. All redundant domains were removed by BLASTClust with a sequence identity cutoff of 30%. One representative domain was randomly selected from each cluster. The final dataset contains 3442 protein domains.

**APO31/HOLO31 dataset:** To examine the effect of binding-induced change of protein conformations on PPV and sensitivity of CBP detection, we built a dataset with both bound (HOLO) and unbound (APO) structures of CBPs. We located APO structures by selecting homologous sequences of proteins in BD113. All APO chains are divided into domains according to SCOP or by DDomain. Only HOLO and APO domains with sequence identity ≥ 50% were selected. Here, the pairwise sequence identity was calculated by ALIGN0 program from FASTA2 package. We found 31 APO-HOLO domain pairs. The majority of the pairs (27 out of 31) have sequence identity more than 80%.

**Structural genomics targets (SG2076).** Our method was applied to 2076 structural genomics targets that were collected by us from previous study on structure-based prediction of DNA-binding proteins. This dataset was obtained by querying structural genomics targets in the protein databank. All structures were divided into domains by the automatic domain parser DDomain. Redundancy was removed by using BLASTClust with a sequence identity cutoff of 30%.

**DFIRE-based statistical potential for protein-carbohydrate interactions**

We employed the same equation as the DFIRE-based interaction for proteins as below
\[
\begin{align*}
    u_{ij}(r) &= \begin{cases} 
    -\ln \frac{N_{\text{obs}}(i,j,r)}{r^\alpha r_{\text{cut}}^{\Delta r_{\text{cut}}}} \frac{N_{\text{obs}}(i,j,r_{\text{cut}})}{r_{\text{cut}}^{\Delta r_{\text{cut}}}}, & r < r_{\text{cut}} \\
    0, & r \geq r_{\text{cut}}
    \end{cases}
    
    \end{align*}
\]

where \(N_{\text{obs}}(i,j,r)\) is the number of pairs of atoms \(i\) and \(j\) within the spherical shell at distance \(r\) observed in a given complex-structure database; \(r_{\text{cut}}\) is the interaction cutoff distance; \(\Delta r_{\text{cut}}\) is the bin width at \(r_{\text{cut}}\); the value of \(\alpha\) (1.61) was determined by the best fit of \(r^\alpha\) to the actual distance-dependent number of ideal-gas points in finite protein-size spheres\(^23\).

Due to lack of a large non-redundant dataset for protein-carbohydrate complex structures and in order to avoid the possibility of over training, we built our statistical potential based on a non-redundant dataset of 3574 protein structures\(^31\) but with simplified atom types so that they can be utilized for most protein-ligand (including protein-carbohydrate) interactions. We employed 11 standard sybyl atom types in mol2 format as defined by openbabel (C.2, C.3, C.ar, C.cat, O.2, O.3, O.co2, N.4, N.am, N.pl3, S.3, N.2, N.ar) for proteins. As before, we set \(r_{\text{cut}}=15\text{Å}\) and \(\Delta r_{\text{cut}}=0.5\text{Å}\)\(^31\). To ensure the usefulness of the protein-based statistical potential for scoring protein-ligand binding affinity, we have applied it to the PDBbind dataset (v2013-core)\(^32\). This dataset contains 195 protein-ligand complexes with known binding affinity. The observed correlation coefficient between predicted and actual binding affinity of 0.52 confirms the usefulness of the statistical potential for scoring binding affinity.

**Prediction protocol**

The protocol for CBP prediction is as follows. First, the target structure is aligned against those templates with sequence identity < 30% from the template library T523 by structure alignment tool SPalign\(^22\). Here, templates with >30% sequence identity to the target are excluded to examine the ability of our method to go beyond sequence homologs. SP-score is employed to measure the structural similarity between template and query structures. If an SP-score is greater than a threshold, the model for the complex structure between the query protein and the template carbohydrate is constructed by replacing the template protein structure with the query structure in the template complex structure. The model complex structure will be utilized to calculate the binding affinity by the DFIRE statistical potential. The binding affinity is calculated by using only \(\text{C}_\alpha\) and \(\text{C}_\beta\) atoms on the interface of proteins and all atoms in carbohydrates to avoid possible steric clashes between un-optimized side chains of amino acid residues and template carbohydrates. The interface atoms of proteins are atoms that have distance less than 7Å with at least one atom in carbohydrate. This interface
cutoff for binding affinity scoring was optimized to achieve the best result for predicting CBPs. We do not perform side-chain optimization because of the lack of efficient and accurate methods for side-chain optimization in the presence of carbohydrates. If the binding affinity is stronger than a threshold, the query is predicted as the CBPs. If the binding affinity does not satisfy the threshold or the structural similarity SP-Score is lower than a structural similarity threshold, the query is predicted as a non-carbohydrate binding protein.

**Performance evaluation**

The performance of the proposed method is measured by sensitivity \[SN = \frac{TP}{(TP+FN)}\], PPV \[PPV = \frac{TP}{(TP+FP)}\] and MCC.

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

where TP, TN, FP and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

**Results**

**SPalign for CBP prediction**

We first examine the possibility of using SP-score alone from SPalign for CBP prediction. SP-score is a structural-alignment score that is independent of the sizes of proteins in comparison. SP-score ranges from 0 to 2. A higher SP-score indicates higher structural similarity. An SP-score at about 0.5 indicates the same structural fold is likely shared by the two structures in comparison. Figure 1 compares the distributions of SP-scores obtained by comparing template structures to the structures in BD113 (filled bars) to those in NB3442 (open bars). The comparison is made after removing any templates with sequence identity more than 30% to the positive query structure. The result shows that only 5.3% non-binding targets from NB3442 have an SP-score of more than 0.7 with a template structure. In contrast, 98% of binding targets can find a template with SP-score > 0.7. It is clear that a structure-alignment program alone can provide a reasonable prediction of CBPs. We found that SPalign can achieve an MCC value of 0.56 with a sensitivity of 43% and a PPV of 75% for the SP-score threshold of 0.778.

**Combining SPalign with the DFIRE-based statistical potential**
To go beyond the one-step prediction of CBPs by the structural alignment program SPalign, we combined SPalign with binding affinities estimated by the extended DFIRE statistical potential, [Equation(1)]. Two thresholds, SP-score and binding affinity were optimized by using the leave-one-out scheme on BD113/NB3442. The grid for SP-score was 0.01. For a given SP-score, we located the binding affinity that yields the highest MCC value. The final MCC value is 0.63 with 0.72 and -0.43 as the thresholds for SP-score and binding, respectively. The corresponding sensitivity and PPV are 52% and 79%, respectively. This result indicates that combining SPalign and binding affinity can significantly improve over SPalign (12% for the MCC value, 20% for sensitivity, and 4% for PPV) as shown in Table 1.

For a baseline comparison, we also predict CBPs by using PSI-BLAST\textsuperscript{28} - a commonly used tool for sequence-to-profile homolog search. We made four iterations of search by PSI-BLAST utilizing the NCBI non-redundant protein sequence library. It predicts a target as CBP if the most significant template from T523 has an E-value smaller than a threshold. As with SPalign-based techniques, the templates are removed if their sequence identities with a target are higher than 30%. The highest MCC value of PSI-BLAST is 0.53 with a PPV of 86% and a sensitivity of 34%. As shown in Table 1, the MCC value is 6% lower than SPalign and 19% lower than SPalign combining with binding affinity scoring. Thus, the combination of SPalign with binding affinity is the most effective method for detecting CBPs. The Receiver Operating Characteristic (ROC) curves for PSI-BLAST, SPalign and SPalign + Binding affinity scoring (SPOT-Struc) are shown in Figure 2.

**The effect of bound/unbound structures on CBP prediction (APO/HOLO dataset)**

We examined the effect of bound/unbound structures on CBP prediction. A target protein will be predicted as a CBP if its SP-score and binding affinity value satisfy the above-established thresholds. The numbers of positive predictions for HOLO and APO sets are 15 and 14, respectively. Only the APO structure 1e4lA was not recognized as a CBP because its SP-score with template 2d24A1 is 0.719 (slightly less than the threshold 0.72) while the HOLO structure 1e55B has a SP-score of 0.724. These results suggest that using APO structures does not lead to a large reduction of the sensitivity of our method, at least for the dataset we were able to obtain.

**The effect of homology models on CBP prediction (APO/HOLO dataset)**
To examine the effect of use of homology-modeled structures on CBP prediction, we employed the SPARKS X\textsuperscript{33} method that performs sequence-to-template structure profile-profile alignment followed by modeller for model building\textsuperscript{34}. If we allow homologous templates in the SPARKS X template library (sequence identity>30%), there is no change to the sensitivity of the prediction: the same 15 proteins were predicted as CBPs. On the other hand, if only templates with a sequence identity of <30% are used as the templates for target proteins, SPOT-struc predicted 11 proteins as CBPs. Three of the four mis-predicted proteins did match to complex templates with structural similarity that is close to but lower than the threshold of 0.72. Another protein failed to be predicted as a CBP because its predicted binding affinity, -0.426, is slightly higher than the threshold of -0.43.

**Binding site prediction**

Predicted structures from SPOT-Struc can be employed to predict binding residues. A residue is defined as a binding site if any heavy atom in that residue is <4.5Å away from any heavy atom of carbohydrate. All other residues are defined as non-binding residues, regardless of if they are on the surface or in the protein core. Predicted binding sites are evaluated against actual binding sites by using the MCC value, sensitivity, and PPV. For 59 correctly predicted CBPs from BD113, an average MCC value of 0.51 with standard deviation of 0.30 was achieved with a sensitivity of 57% and a PPV of 57%. The MCC values of the predicted binding residues in these 59 proteins are shown in Figure 3 as a function of the structural similarity score SP-score between predicted and actual protein structures. In general, a higher SP-score tends to achieve a higher MCC value although the overall correlation coefficient between MCC and SP-score is only 0.30.

Figure 4A shows one successful example by comparing predicted CBP binding sites with native binding sites for Fucolectin-related protein in *Streptococcus pneumoniae* serotype 4 (target 2j1uA). For this target, the SPOT-Struc prediction achieved an MCC of 0.88 although the sequence identity between this target and template 2j7mA (a carbohydrate-binding module from a *Clostridium perfringens* N-acetyl-β-hexosaminidase) is only 18%. The SP-score (the structural similarity) between predicted and native structures is 0.81. There are 125 residues aligned within 2.2 Å root-mean-squared distance.

For some proteins, their binding-site prediction is poor despite accurate structure predictions (Figure 3). An example is shown in Figure 4B, The SP-score between the predicted structure based on the template 1uxxX and the native structure 1uyzB is 0.94. There are 120 residues
aligned with a RSMD of 1.6Å. However, the predicted binding region and native binding region do not overlap with each other as shown in the figure. We found that both template (1uxxX for ctcmB 6) and native proteins (1uyzB for cmcBM6-2) belong to the same family 6 carbohydrate-binding module. This binding module contains two carbohydrate-binding clefts, cleft A (ctcBM6) and cleft B (cmcBM6-2)\textsuperscript{35}. Cleft A is located in the loop region connecting the inner and outer β-sheet, and cleft B is located at the concave surface of two β-sheets\textsuperscript{35}. The predicted binding region of the target is cleft A, in good agreement with experimentally validated binding sites (Trp-92, Tyr-33, Glu-20 and Thr-23). The native complex structure, on the other hand, reveals binding in cleft B. That is, the predicted complex structure captures another binding region of the target.

**Binding site prediction for APO/HOLO/homology modeling structures**

For 14 predicted CBPs shared by APO and HOLO targets, we achieved an average MCC value of 0.44 for predicted binding sites with a sensitivity of 54% and a PPV of 49% for HOLO structures, compared to the MCC value of 0.45, a sensitivity of 55%, and a PPV of 49% for APO structures, and the MCC value of 0.43, a sensitivity of 53%, and a PPV of 48% for model structures based on highly homologous templates. If only remote-homology templates (sequence identity <30%) were used, the average MCC value is reduced to 0.36 for predicted binding sites of 11 predicted CBPs. Thus, binding-site prediction does not strongly depend on HOLO (bound), APO (unbound), or close homology models for this set of proteins.

**Application to structural genomics targets**

We further applied our method to 2076 domains determined by structure genome projects. The trained thresholds (0.72 for SP-score and -0.42 for the binding scoring) were employed. Nineteen targets from 2076 domains were predicted as CBPs. Among them, 11 out of 19 (57%) were annotated as CBPs or putative CBPs by Uniprot\textsuperscript{36}. Four of the eight remaining predicted CBPs are listed with other known functions – a putative SusD homolog of *Bacteroides fragilis* (3ejnA), an amidohydrolase from *Burkholderia phytofirmans* (3mkv), a CMRF35-like molecule 1 from mouse (1zoxA), and a calcium-binding protein from *Agrobacterium tumefaciens* (2ghsA). The remaining proteins are uncharacterized proteins from *Bacillus subtilis* (1oq1A), *Caenorhabditis elegans* (1t9fA), *Nostoc punctiforme* (2p4oA) and *Bacteroides thetaiotaomicron* (2p0vA). For example, a putative susD homolog of *Bacteroides fragilis* (3ejnA) was matched to susD, a starch-binding protein in *Bacteroides thetaiotaomicron* VPI-5482 (3ck7B)\textsuperscript{37}. CMRF35-like molecule 1 of mouse (1zoxA) was
matched to a sugar-binding antibody FAB fragment (1mfH1)\textsuperscript{38}. The uncharacterized protein from *Bacillus subtilis* (1oq1A) was matched to galectin-9 from *Mus musculus* (2d6oX)\textsuperscript{39}. In the latter case, the identification of a putative galectin in *Bacillus subtilis* is important, if confirmed, because the protein could serve as a drug target for an antibacterial agent.

**DISCUSSION**

In this study, we have developed a method for simultaneous prediction of CBPs and their binding residues. This is the first template-based method to detect CBPs based on their known protein structures. The leave-one-out validation reveals its high PPV (79\%) with reasonable sensitivity (52\%) for a dataset of 113 binding and 3442 non-binding domains. Its application to 31 APO structures leads to only a small reduction in sensitivity. Using highly homologous templates did not reduce the accuracy of CBP prediction, while using models based on remote homology (<30\% sequence identity) somewhat decreased sensitivity of the method. Our method predicted eight structural genome targets that are putative CBPs.

Two thresholds, SP-score and binding affinity, were trained as two independent variables to optimize the MCC of prediction. It is possible that a less restrictive binding affinity may be applicable for a highly similar structure. Such structure-alignment-dependent binding affinity was found useful in improving structure-based prediction of DNA-binding proteins\textsuperscript{19,40} but not for RNA-binding proteins\textsuperscript{20}. Here we found that introducing different thresholds for different SP-scores doesn’t lead to increase in MCC values.

In this work, non-binding residues were defined as all other residues including residues in the protein core. This definition was made to facilitate future comparison with a sequence-based technique we are developing. Such a definition will improve the MCC value because protein core residues can be easily predicted as non-binding residues. If binding and non-binding residues were limited to surface residues defined as 20\% or more residue surface areas accessible, the MCC value did reduce somewhat from 0.51 to 0.48 for 59 predicted CBPs.

One limitation of the current method is its dependence on complex template structures. That is, it will not be able to predict CBPs with novel carbohydrate-binding structures. The fact that we can achieve 52\% sensitivity suggests that binding structural types are likely limited\textsuperscript{41,42}. As more and more complex structures are solved, our method will likely further improve in its sensitivity by incorporating newly solved structures in its template library.
Acknowledgement. This work is supported in part by National Health and Medical Research Council (1059775) of Australia to Y.Z. M.v.I gratefully acknowledges the Australian Research Council (DP130102945) and the National Health and Medical Research Council Project Grant (1006618) for financial support. We 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. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

References

33. Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. Bioinformatics 2011, 27, 2076-2082.

Table 1. Leave-homolog-out cross-validation for PSI-BLAST, SPalign, and SPOT-Struc with positive BD110 and negative NB3442 datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>Positive Predictive Value</th>
<th>Sensitivity</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSI-BLAST</td>
<td>86%</td>
<td>34%</td>
<td>0.53</td>
</tr>
<tr>
<td>Method</td>
<td>Accuracy</td>
<td>Specificity</td>
<td>Matthews Correlation Coefficient</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>----------</td>
<td>-------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>SPalign</td>
<td>75%</td>
<td>43%</td>
<td>0.56</td>
</tr>
<tr>
<td>SPalign+binding scoring (SPOT-Struc)</td>
<td>79%</td>
<td>52%</td>
<td>0.63</td>
</tr>
</tbody>
</table>

*Matthews Correlation Coefficient.
Table 2. Structural genome targets predicted as CBPs

<table>
<thead>
<tr>
<th>Target</th>
<th>Template</th>
<th>SP-score</th>
<th>Binding</th>
<th>Seq ID</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1y89B</td>
<td>2ri1A</td>
<td>0.95</td>
<td>-0.66</td>
<td>24.3</td>
<td>CBP</td>
</tr>
<tr>
<td>3cbwA</td>
<td>2cipA</td>
<td>0.89</td>
<td>-1.51</td>
<td>22.7</td>
<td>CBP</td>
</tr>
<tr>
<td>1vdwA</td>
<td>2qvrA</td>
<td>0.88</td>
<td>-0.87</td>
<td>19.8</td>
<td>CBP</td>
</tr>
<tr>
<td>3hnmA</td>
<td>2j7mA</td>
<td>0.87</td>
<td>-1.04</td>
<td>19.6</td>
<td>CBP</td>
</tr>
<tr>
<td>3gg1A</td>
<td>2j7mA</td>
<td>0.87</td>
<td>-0.74</td>
<td>28.7</td>
<td>CBP</td>
</tr>
<tr>
<td>2p0vA</td>
<td>2vn4A1</td>
<td>0.85</td>
<td>-0.74</td>
<td>6.8</td>
<td>UNK</td>
</tr>
<tr>
<td>3ebvA</td>
<td>2dsvA</td>
<td>0.84</td>
<td>-4.16</td>
<td>23.5</td>
<td>CBP</td>
</tr>
<tr>
<td>3eypA2</td>
<td>2j1uA</td>
<td>0.83</td>
<td>-0.82</td>
<td>19.9</td>
<td>CBP</td>
</tr>
<tr>
<td>3ejnA</td>
<td>3ck7B</td>
<td>0.82</td>
<td>-0.57</td>
<td>20.4</td>
<td>Other</td>
</tr>
<tr>
<td>3mkvA</td>
<td>2vhlA</td>
<td>0.82</td>
<td>-0.66</td>
<td>22.7</td>
<td>Other</td>
</tr>
<tr>
<td>1zoxA</td>
<td>1mfbH1</td>
<td>0.82</td>
<td>-0.71</td>
<td>23.6</td>
<td>Other</td>
</tr>
<tr>
<td>3c5yE</td>
<td>2vvpA</td>
<td>0.8</td>
<td>-1.4</td>
<td>22.9</td>
<td>CBP</td>
</tr>
<tr>
<td>1oq1A</td>
<td>2e6vA</td>
<td>0.8</td>
<td>-0.82</td>
<td>16.6</td>
<td>UNK</td>
</tr>
<tr>
<td>1t9fA</td>
<td>1v6wA2</td>
<td>0.79</td>
<td>-1.58</td>
<td>14.9</td>
<td>UNK</td>
</tr>
<tr>
<td>1xpwA</td>
<td>2vmdA</td>
<td>0.77</td>
<td>-0.61</td>
<td>4.2</td>
<td>CBP</td>
</tr>
<tr>
<td>2nlyA</td>
<td>1w1a1</td>
<td>0.76</td>
<td>-0.55</td>
<td>15.7</td>
<td>CBP</td>
</tr>
<tr>
<td>2ghsA</td>
<td>1ms1A1</td>
<td>0.73</td>
<td>-0.63</td>
<td>6.8</td>
<td>Other</td>
</tr>
<tr>
<td>3e5zA</td>
<td>1ms1A1</td>
<td>0.73</td>
<td>-1.71</td>
<td>16.6</td>
<td>CBP</td>
</tr>
<tr>
<td>2p4oA</td>
<td>1ms8A1</td>
<td>0.73</td>
<td>-0.6</td>
<td>7.9</td>
<td>UNK</td>
</tr>
</tbody>
</table>

“a” Targets were annotated as having putative function related to carbohydrate-binding.
“b” Unknown functions.
“c” Annotated with other functions.
Figure 1. Distributions of the best SP-scores of the templates in the template library for the protein domains in the positive (BD113, filled bars) and negative (NB3442, open bars) datasets, respectively. The templates with more than 30% sequence identity to the query sequence are excluded.
Figure 2. Sensitivity versus false positive rate given by PSI-BLAST, SPalign and SPOT-Struc (SPalign + Binding affinity Scoring).
Figure 3 The MCC value for predicted carbohydrate-binding residues as a function of the structural similarity between the query and template according to the SP-score.
Figure 4. Comparison between predicted and native binding regions for protein (red) and carbohydrate (magenta) of target, and protein (green) and carbohydrate (cyan) in template. The red and green colors represent predicted and native structures, respectively. (A) Query 2j1uA on template 2j7mA with sequence identity of 18% and SP-score of 0.81 (125 residues aligned with a RMSD of 2.16 Å). (B) Query 1uyzB on template 1uxxX with sequence identity of 25% and SP-score of 0.94 (120 residues aligned with a RMSD of 1.58 Å).