Glycobiology, 2015, vol. 25, no. 2, 181–196
doi: 10.1093/glycob/cwu102
Advance Access Publication Date: 28 September 2014
Original Article

Original Article

Using crystallographic water properties for the analysis and prediction of lectin–carbohydrate complex structures

C Modenutti[2,3], D Gauto[4], L Radusky[2], J Blanco[2], A Turjanski[2,4], S Hajos[3], and MA Marti[1,2,4]

[2]Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. II (CE1428EHA), Buenos Aires, Argentina, [3]Departamento de Microbiología, Inmunología y Biotecnología, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, IDEHU-CONICET, Buenos Aires 1113, Argentina, and [4]Departamento de Química Inorgánica, Analítica y Química Física/INQUIMAE-CONICET, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pab. II (CE1428EHA), Buenos Aires, Argentina

[1]To whom correspondence should be addressed: marcelo@qi.fcen.uba.ar

Received 11 August 2014; Revised 19 September 2014; Accepted 20 September 2014

Abstract

Understanding protein–ligand interactions is a fundamental question in basic biochemistry, and the role played by the solvent along this process is not yet fully understood. This fact is particularly relevant in lectins, proteins that mediate a large variety of biological processes through the recognition of specific carbohydrates. In the present work, we have thoroughly analyzed a nonredundant and well-curated set of lectin structures looking for a potential relationship between the structural water properties in the apo-structures and the corresponding protein–ligand complex structures. Our results show that solvent structure adjacent to the binding sites mimics the ligand oxygen structural framework in the resulting protein–ligand complex, allowing us to develop a predictive method using a Naive Bayes classifier. We also show how these properties can be used to improve docking predictions of lectin–carbohydrate complex structures in terms of both accuracy and precision, thus developing a solid strategy for the rational design of glycomimetic drugs. Overall our results not only contribute to the understanding of protein–ligand complexes, but also underscore the role of the water solvent in the ligand recognition process. Finally, we discuss our findings in the context of lectin specificity and ligand recognition properties.

Key words: carbohydrate, docking, hydration, lectin, Naive Bayes classifier, sites, water sites

Introduction

The analysis of the structure and dynamic properties of the solvent around protein active sites, has gathered strong attention among the scientific community, showing that the water molecules that are displaced upon ligand binding are key players for determining the underlying thermodynamics of the process (Michel et al. 2009; Hummer 2010; Setny et al. 2010). As a result of the protein–solvent interactions, water molecules are not placed randomly on the protein surface, adopting thus a well-established structure, defined by regions of highly ordered water molecules. This solvent structure is particularly relevant in regions such as protein active- and ligand-binding/recognition sites (Barillari et al. 2007; Abel et al. 2008; Higgs et al. 2010; Beuming et al. 2012; Saraboji et al. 2012).

Lectins are one of the main categories of sugar-binding proteins, defined by the presence of carbohydrate recognition domain, with its corresponding carbohydrate binding site (CBS) and a lack of catalytic activity toward their ligands. They are present in all living organisms performing a wide variety of biological activities including cell

recognition, communication and cell growth, and are of wide interest in biotechnology and as potential therapeutic targets (Compagno et al. 2014). Thus, understanding and predicting protein–carbohydrate interactions if of paramount relevance in the field of glycobiology. One of the salient features of carbohydrate ligands, and their binding sites, lies in their hydrophilic nature, which accentuate or emphasize the effect of solvent reorganization throughout the biomolecular association process (Chervenak and Toone 1995).

Water molecules and carbohydrate hydroxyl groups usually participate in the same interactions with the protein. Thus, understanding solvent structure can provide significant insight into the carbohydrate binding and recognition processes. The relationships between the solvent structure in the CBS of several proteins, including many lectins, and the resulting protein–ligand complexes, have been steadily studied during the last decade using both computational (Frank and Schloissnig 2010; Kumar et al. 2013) and experimental methods (Kadirvelraj et al. 2008; Saraboji et al. 2012; von Schantz et al. 2012; Andres et al. 2013; Johal et al. 2013).

Explicit water molecular dynamics (MD) simulations, combined with a solid statistical thermodynamic analysis framework, as that provided by the inhomogeneous fluid solvation theory, have allowed to systematically characterize the solvent structure and dynamics at the protein surfaces, through the identification of the so-called water or hydration sites (Lazaridis 1998; Li and Lazaridis 2006). A water site corresponds to a confined region in the space adjacent to the protein surface, where the probability of finding a water molecule is significantly higher than that observed in the bulk solvent at the same density (Li and Lazaridis 2005; Abel et al. 2008; Di Lella et al. 2010). In previous works, using this strategy, we were able to show that the position and attribute of the water sites are good predictors of the hydroxyl groups positions in the resulting lectin–carbohydrate complex (Di Lella et al. 2007; Gauto et al. 2009), being able, even to predict the subtle selectivity of lectins between two epimers (Gauto et al. 2011). Given the relevance of the determination of an atomic resolution structure for any given protein–ligand complex (Fadda and Woods 2010), there is a widespread use of in silico strategies for their prediction (i.e., molecular docking methods) (Morris et al. 1998; Taylor et al. 2002; Brooijmans and Kuntz 2003; Friesner et al. 2004; Leach et al. 2006; Abel et al. 2008; Englebienne and Moitessier 2009; Yuriev et al. 2009; Wang et al. 2011). These, however, show a weak performance for lectin–carbohydrate complexes (Kerzmann et al. 2008; Agostino et al. 2009; Mishra et al. 2012; Gauto et al. 2013). In a previous work from our group, we used the above-mentioned MD-derived water sites, to significantly improve the docking of carbohydrates. Our results showed that by modifying the AutoDock4 scoring function favoring those ligand conformations where the carbohydrate–OH groups match the position of the water sites, the predictions show significant improvement in terms of both accuracy, measured as the potentiality to predict the complex structure closest to the one obtained by X-ray crystallography and also its capability for differentiating the correct complex among wrong predictions (Gauto et al. 2013).

To achieve a deeper understanding of the relationship between the solvent structure and lectin–carbohydrate complexes, we decided to study a large set of lectin structures available in the Protein Data Bank (PDB) and analyze the properties of the crystallographic water molecules. Analysis of crystallographic waters in the active or ligand-binding site of proteins has been extensively studied in the last decades and shown to provide useful information for the process of drug design. For instance, several properties such as hydrogen bond interactions, the mobility of the water molecules or the influence of the local site shape, have been extensively analyzed and correlated with the affinity of modified ligands designed to displace and/or keep those crystallographic waters (Li and Lazaridis 2005; Kadirvelraj et al. 2008; Barillari et al. 2011; Saraboji et al. 2012; Garcia-Sosa 2013). Also, the X-ray structures of protein–carbohydrate complexes as well as those of glycoproteins have also been extensively used to analyze and understand the complex structure of carbohydrates themselves, and many high-quality databases and web resources on the subject are available, like the carbohydrate structure suite (Lütteke et al. 2005) among others (Ranzinger et al. 2008; von der Lieth et al. 2011). However, to our knowledge, no work has thoroughly analyzed the relationship between the solvent structure described by the crystallographic waters in a ligand-free (or apo) lectin and the structure of the corresponding protein–carbohydrate complexes, and/or used solvent structure information to understand and predict lectin–saccharide complex properties.

In the present work, we have thoroughly analyzed a nonredundant and well-curated set of lectin structures looking for a potential relationship between the structural water properties in the apo-structures and the corresponding protein–ligand complex structures. Our results show that the position of crystallographic waters in the ligand-free structures tends to mimic the carbohydrate –OH structural framework, thus underscoring their role in the ligand recognition process. We also show how crystallographic water properties can be used to predict their likelihood of being replaced by carbohydrate ligand polar groups, and how this information can be used to improve docking predictions of lectin–carbohydrate complex structures. Therefore, the presented analysis results not only in a deeper understanding of the protein–ligand interactions, but also provides theoretical framework for the prediction of these complex and the rational design of glycomimetic drugs.

Results

The results are organized as follows: First, we describe and analyze general properties of protein–carbohydrate structures. Secondly, we analyze in detail several properties of the crystallographic water molecules and compare them with the protein–ligand complex structure, and use this information to develop a Bayesian predictive method to classify water sites in relation to their likelihood of being replaced by the ligand. Finally, we use the information derived from the crystallographic water analysis to improve the prediction of potential protein–carbohydrate complex using a molecular docking scheme.

Data set construction

The present lectin–carbohydrate data set was built retrieving first all available lectin structures from the PDB (date, April 2014). Structures were grouped in unique protein sets and only those sets containing at least one lectin–carbohydrate complex structure and one apo-structure were retained. A total of 19 unique lectin sets that could be grouped in 8 families, and consisting in 167 structures of lectin–carbohydrate complexes and 75 apo-structures were analyzed. All statistical analyses were performed considering (whenever possible) each individual structure, each individual unique protein set and an average value determined for each protein family, in order to avoid possible bias due to overrepresentation of a particular protein or protein family.

Analysis of protein–carbohydrate complexes

Figure 1A shows the fraction of available crystal structures that are bound either to mono-, di-, tri- or larger oligosaccharides (up to

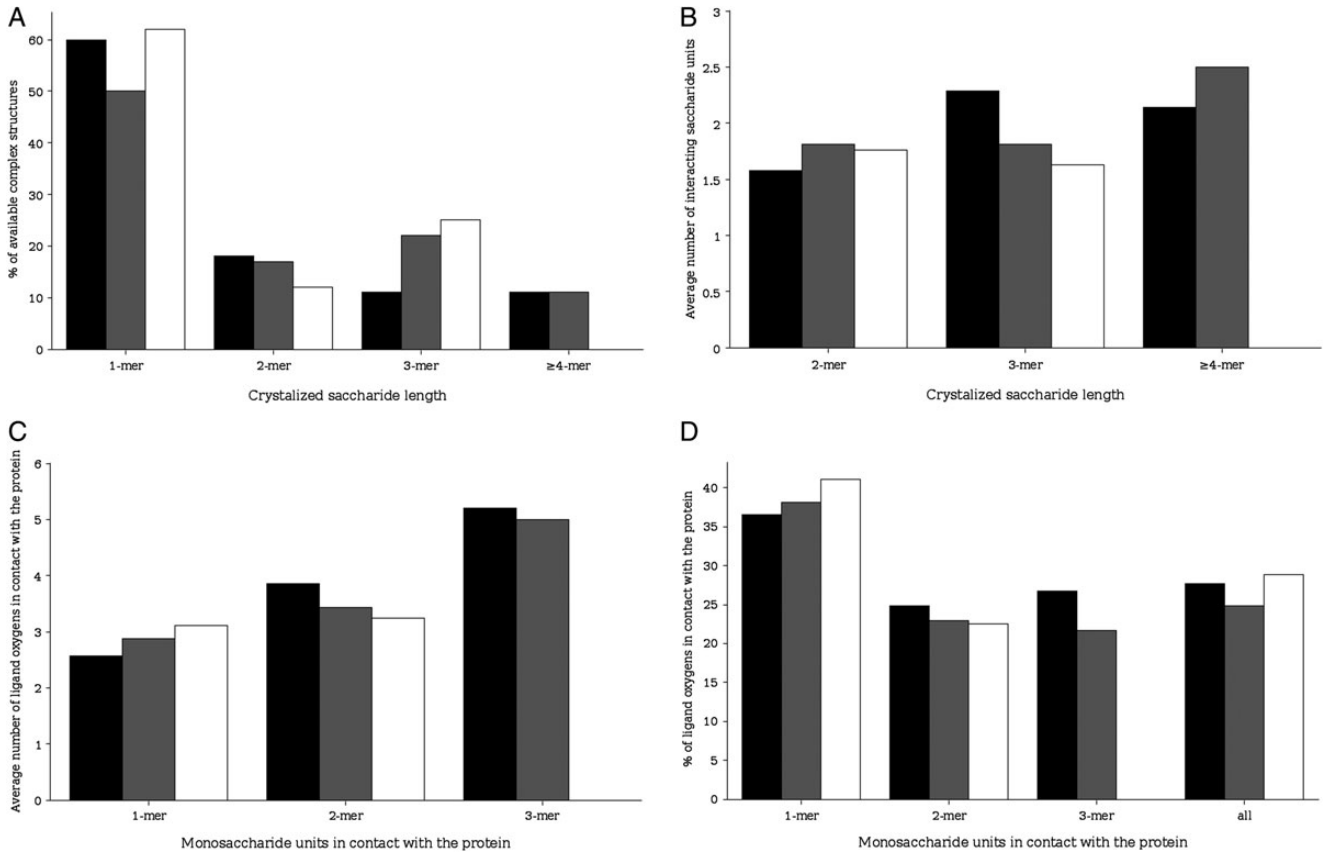


Fig. 1. (**A**) Percentage of Lectin–carbohydrate complex structures available with mono-, di-, tri- or larger oligosaccharides. (**B**) Average number of monosaccharide units found interacting with the protein in the corresponding mono-, di-, tri- and larger lectin carbohydrate complexes. (**C**) The number of carbohydrate ligand oxygen atoms in contact with the protein in lectin–carbohydrate complexes classified according to the number of contacting monosaccharide units. (**D**) The percentage of carbohydrate oxygen atoms in contact with the protein classified according to the number of contacting monosaccharide units. Black bars are determined considering each crystal structure, gray and white bars are the average results grouped by unique protein sets and families (when available), respectively.

heptasaccharides). The figure clearly shows that there is a predominance of structures crystallized with monosaccharaides (~60% of the total), followed by disaccharides (~20%), and so on. The same analysis but considering each unique protein set, or even family shows similar results. Also, it is interesting to note that for several proteins that can be crystallized with disaccharides and trisaccharides, also crystals with monosaccharide are available. Figure 1B shows how many monomers (i.e., saccharide units) of the crystallized ligand in complex with the protein are actually making contact with the protein (see the "Computational methods" section for definition of a contacting monomer). Interestingly, for disaccharides average number is <2 (close to 1.5), showing that in several cases only one monomer binds to the protein. Moreover, tri- and larger ligands show that the number of contacting units is between 1.5 and 2.5, thus most of the binding and therefore the affinity and specificity in lectin–carbohydrate interactions seems to be given by 1–2 saccharides per units and no more.

Since hydroxyl (and to a lesser extent carbonyls, acid and amide) functional groups are mainly responsible for protein–carbohydrate interactions, we analyzed how many of these polar groups make interactions with the protein surface, in relation to the number of units in contact. The results presented in Figure 1C show that when the ligand is a monosaccharide ~3 (sometimes 2) interactions are established, which can be rationalized considering that common biological monosaccharides have between 6 and 7 potential polar interactions and they bind with one face facing the protein and the other facing the solvent. What is interesting that even when a disaccharide is in contact, the number of interactions increases only slightly to approximately four interactions, and even for ligands contacting the protein by a trisaccharide no more than five interactions are established. Overall, as shown in Figure 1D, ~1 every three to four hydroxyl oxygens of the ligand are in contact with the protein, with higher numbers (slightly below half of them) for the monomers. In so far, the results thus show that for lectins, the binding core of carbohydrate ligands is determined by one or two monomers, presenting no more than two to three interactions per subunit.

Amino acid composition analysis of the carbohydrate recognition domain and CBS

We now focus our analysis on the carbohydrate recognition domain (CRD), by looking which residues are found preferentially in the CBS. For this sake, we first defined all unique protein CBS residues as those residues that are in contact with the ligand in any of the available complex structures, and computed the relative frequency of appearance for each of the 20 amino acids in the CBS or as being part of the whole CRD. The results presented in Figure 2A show that while several residues are preferentially found in the CBS, others clearly tend to be excluded. There is an enrichment of negatively charged residues (Asp and Glu) as well as positively charged Arg and His. Polar Ans is also preferentially found in the CBS, as well as polar aromatic Trp and Tyr. On the other hand, nonpolar residues such as Val, Ile, Leu, Phe, Met and Pro as well as Cys tend to be excluded from the CBS. To analyze the relative impact of aromatic–carbohydrate interactions, we determined how many monosaccharide units in contact with the protein present these types of interactions. The results show that ~70% of bound monosaccharide units show an aromatic residue interacting with the carbohydrate aliphatic core, with Trp being the predominant (43% of the cases) followed by Tyr (33%) and Phe (24%). Overall these results are consistent with previous observations from our group and others (Asensio et al. 2000; Lütteke et al. 2005; Guardia et al. 2011), which show that lectins bind their ligands combining polar interactions with the ligand –OH group and an aromatic (nonpolar) interaction with the ligand aliphatic core (Guan et al. 2003; Sujatha et al. 2004; Terraneo et al. 2007; Laughrey et al. 2008; Nishio et al. 2014).

Analysis of crystallographic water sites in relation with CBS and protein–carbohydrate complex

To analyze the solvent structure, we begin looking at which residues of the CBS (or the whole protein domain) are preferentially found in contact with a crystallographic water site (CWS), which is defined as the average position where crystallographic water oxygens are found across the several available apo structures of each unique protein. In the case where only one crystal structure is available, CWS corresponds unequivocally to crystallographic waters. The resulting data, presented in Figure 2B, show as expected, that water molecules are mostly found close to hydrophilic residues. Moreover, there is a clear preference for finding water molecules associated to those polar residues found in the CBS of the proteins (Arg, Asn, Asp and Glu). Thus, and as expected, the results are similar to those observed above, but excepting the preference for aromatic residues.

To analyze now the relation between the CWS and the corresponding lectin–carbohydrate complex, we first determined how many of the –OH groups that are interacting with the protein in any of the studied unique protein sets, show the presence of a CWS in the corresponding apo-structure. To assess that a CWS is replaced by a carbohydrate polar group, the apo- and ligand-bound structures are first aligned and the minimum distance between the water oxygen and any ligand oxygen atom is determined, this value is what we call R_{min}. If it is <1.2 Å, the crystallographic water is classified as replaced. Other CWSs found inside the CBS, which are lost upon ligand binding, but which are not replaced by ligand hydroxyl (or other polar group) are termed as displaced CWSs. The results presented in Figure 3A, for the whole protein set, show that ~70% of all unique protein–carbohydrate–OH interactions replace a CWS, and the results are similar for different size ligands. Moreover, Figure 3B shows that for the whole unique proteins analyzed, ~40% of CWS inside the binding site will be replaced by the ligand OH group. Interestingly, the correspondence between CWS and ligand –OH groups is slightly higher for proteins contacting monosaccharides, while in both analyses, the results are similar if unique proteins or protein families are considered, showing the absence of any bias.

As a particular example, Figure 4 shows the CBS of the carbohydrate recognition domain of hSPD and Gal-9 (pdbids 3G83 (Crouch et al. 2009) and 3NV4 (Yoshida et al. 2010), respectively), superimposing both the apo-structure showing the CWS and the complex structure showing the bound ligand. The figure shows how some of the CWSs (shown in red) are replaced by the ligand –OH groups while others still close to the ligand (i.e., inside the CBS) are not, being classified as displaced (shown in blue). In hSP-D bound to a monosaccharide (Figure 4A), there are three replaced CWSs that perfectly superimpose on the ligand hydroxyls, while the displaced CWS is in the middle of the carbohydrate ring. For Gal-9 (Figure 4B) bound to a trisaccharide (Sialyllactose), there are three replaced CWSs that match three ligand hydroxyls in the first two monomers, while there is a fourth CWS, which is displaced by the third monomer. In summary, it is clear that CWS mimics to some significant extent the structural framework of the carbohydrate ligand hydroxyl groups, which are responsible for ligand affinity and selectivity.

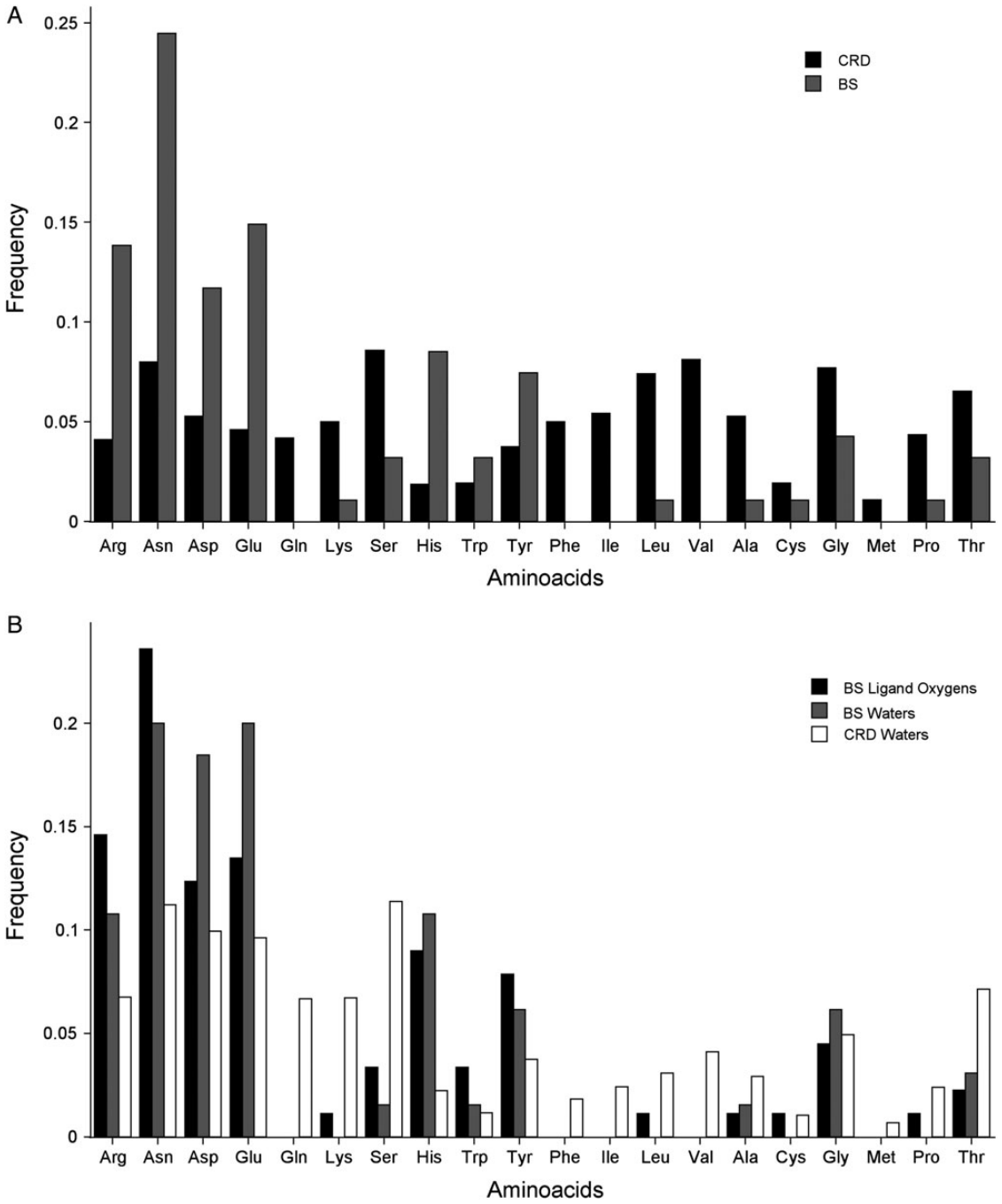


Fig. 2. (**A**) Observed frequencies for each of the 20 different amino acid residues in: the whole carbohydrate recognition domain (black columns) and only the CBS (gray columns) for all analyzed lectin–carbohydrate complexes. (**B**) Observed frequencies for each of the 20 amino acid residues in contact with: a CWS in the whole carbohydrate recognition domain (white columns), a CWS in the CBS (gray columns) a carbohydrate ligand oxygen atom (black columns).

Determination of CWS structural parameters

The data from the previous section show that CWS can be classified into three groups according to whether they will be replaced or displaced from the CBS when the ligand binds, and those outside the CBS. We will now analyze several CWS properties in order to see if there are any significant differences between the mentioned groups. We analyzed six properties, which are (i) the average number of polar interactions, i.e., hydrogen bonds, that each CWS establishes with the protein; (ii) the closeness to protein, which is defined as the nearest distance of the CWS to any protein heavy atom; (iii) the contact surface between the crystallographic water oxygen and the protein; (iv) the mobility of the water molecules of the CWS, which is related to the average B-factor of all oxygen atoms that define the CWS in the different apo-structures if available; (v) the occupancy of the water molecules, which is computed as the ratio of apo-crystal structures where a crystallographic water is found, times the total number of available apo-structures (or protein is an oligomer); (vi) the occupancy of the water molecules, which is computed as the

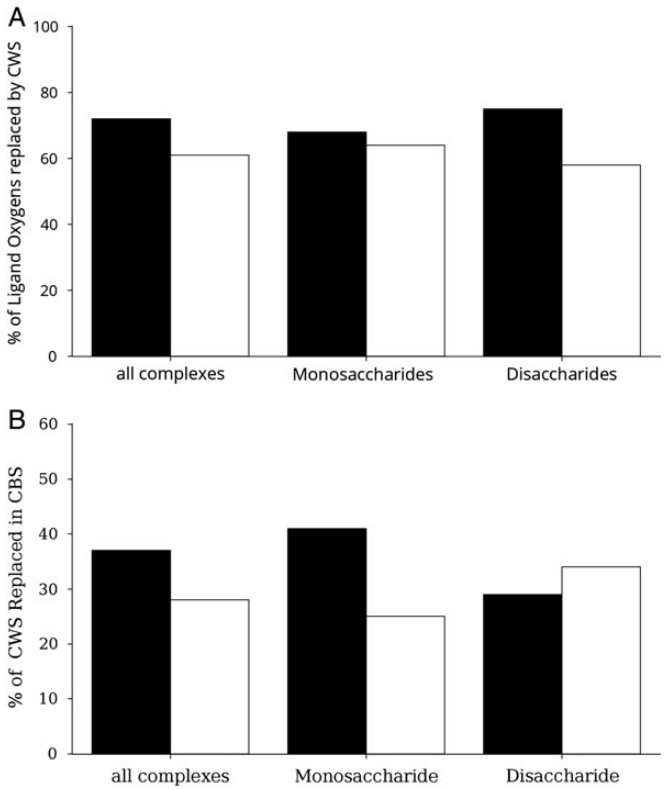


Fig. 3. (**A**) The percentage of ligand oxygens interacting with the protein that are replaced by CWS in unique proteins, and grouped by protein families. Results are shown for all complexes and for monosaccharide and disaccharide units in contact with the protein separately. (**B**) The percentage of CWSs inside the CBS that are replaced in the Protein–carbohydrate complex in unique proteins and grouped by protein families. Results are shown for all complexes and for mono-, and disaccharides units in contact with the protein separately.

ratio of apo-crystal structures where a crystallographic water is found, times the total number of available apo-structures (or protein is an oligomer). The resulting histograms for the probability distribution of all six properties for replaced, displaced and those CWS outside the binding site are shown in Figure 5A–F.

First sight of Figure 5A clearly shows that replaced CWSs have on average more polar (mostly hydrogen bond) interactions with the protein (usually 1 or 2) than the other CWSs that are only displaced by the ligand from the CBS (0–1 contact with the protein). Interestingly, those CWSs out of CBS, also show mostly none or one hydrogen bond. Thus, the number of contacts with the protein, not only potentially distinguishes those CWSs inside CBS that are in place where ligand polar groups will be found from the others, but also could point to the CBS itself. Complementary analysis that helps to understand this behavior is presented in Figure 5B, where the minimum distance of the CWS to any protein atom is measured. The results clearly show that while most of the replaced CWSs are in the first solvation shell (protein CWS distance of ~3 Å), those in the displaced group are farther from the protein surface (4–5 Å) usually in the second solvation shell. A third property related to the CWS interactions is the contact surfaces analyzed in Figure 5C, where it is shown that replaced CWSs present a more narrow distribution with a slight tendency to enrichment in higher values. Concerning CWS mobility, the distribution of average B-factors presented in Figure 5D shows that replaced CWSs have significantly smaller B-factors than all the others. The out of CBS CWS and displaced CWS shows broader distributions, probably corresponding to background or random distribution. The next computed property of the CWS is related to the variability observed for the presence of a given crystallographic water in different apo-structures (occupancy). The resulting values presented in Figure 5E again show that displaced and out of CBS CWS show similar distributions with peaks at 0.2, 0.5 (possibly reflecting cases where two crystals are available and the CWS is only present in one) and 1. On the other hand, the replaced CWSs show all values close to 1, highlighting again their differential properties. The last parameter, which is the number of CWS neighbors presented in Figure 5F, shows similar distributions for all types of CWSs, and thus bears little predictive value. In summary, the above analyzed parameters, particularly the number of contacts with the protein, the distance to the protein, the B-factor and the occupancy, are able to distinguish those CWSs from the CBSs that will be replaced by ligand polar groups, from those that not and also from other CWSs out of CBSs. Most important, except for the occupancy, the other properties can be obtained from only one apo-structure, thus adding significance to the resolution of lectin crystal structures even in the absence of their ligands.

Predicting the replaced CWS using a Naive Bayes classifier

Given that significant differences are observed between replaced CWS properties and those of the displaced and out of CBS sites, we used the data to build a Naive Bayes classifier (Maruyama 2013) that could allow to determine the likelihood of any CWS of being replaced. The method takes as input the six CWS characteristic values described above and determines the probability of being replaced given those values (PR), and that of not being replaced given those values (Pnot), if

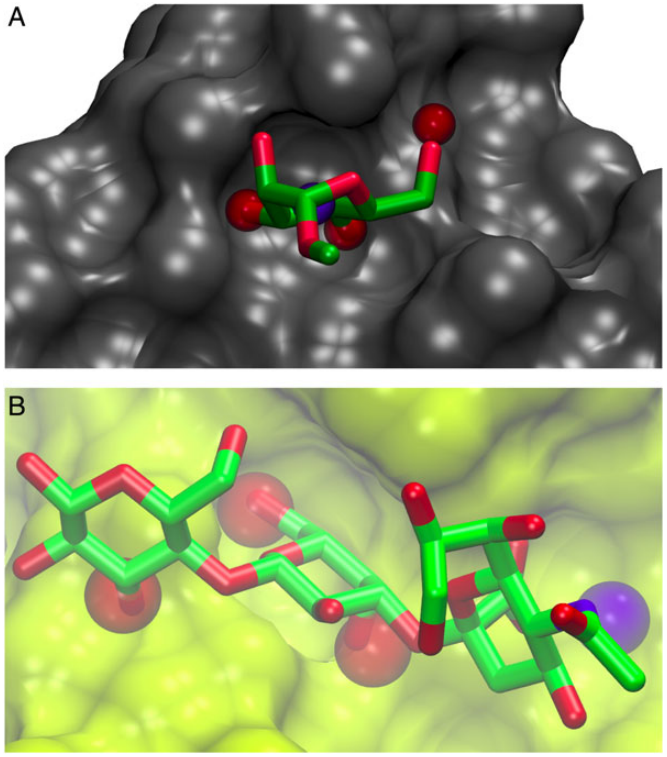


Fig. 4. The structure of the hSP-D mannose complex (**A**) and structure of the Gal-9 C terminal sialyllactose (**B**) complex superimposed on the CWS positions relative to the protein structure. Replaced CWSs are shown as red spheres and Displaced CWSs as blue spheres.

PR > Pnot then the CWS is predicted as being replaced. Table I shows the results of the prediction for all CWSs analyzed. The data show that the method correctly predicts over 90% of those CWSs that are going to be replaced, while wrongly assigning as potentially replaced about one-fourth (28%) of the CWSs inside the binding site that are not. Although this value may seem a little high, it is important to note that replaced CWSs are determined depending on the available complex structures. Thus there may well be other ligands that displace those, thus bringing the percentage of wrong positives down. Also interestingly, the classifier shows that most CWSs out of the CBSs are not predicted as replaced, an observation that is consistent with the idea that CWSs inside the active site and specially those that are mimicking protein–ligand interactions have distinctive properties. The Naive Bayes classifier yields along the mentioned probabilities for each CWS, Gaussian-like distribution parameters (mean ± SD), for each property in each set. As expected, and confirming the analysis of the previous paragraph, those properties that show significant differences in their mean distribution values between replaced and displaced CWSs are those previously identified as potentially important for differentiating the CWSs. Encouraged by our results we decided to test also another Bayesian analysis, called Bayesian category interference (Green 1995; Patil et al. 2010), which is similar to the Classifier but does not assume that elements (i.e., the CWSs) are known to belong to different categories and instead automatically separates them. The results of this strategy show, using all CWSs inside the CBS, that the method correctly assigns 92% of Replaced CWSs and 72% of those not, and yields for each property probability distributions that are similar to those obtained with the NBC. These results confirm once again that replaced CWSs have distinct properties in relation to their likelihood of being replaced or not by the ligand polar groups in the corresponding lectin–carbohydrate complex.

Last but not least, to highlight the utility of the predictive method as shown in Figure 6 a lectin with all their CWSs are colored according to their derived relative probabilities (PR/Pnot), computed with the Naive Bayes classifier, and shown superimposed on the corresponding complex structure. The figure nicely shows how the predicted replaced CWSs (red spheres) tend to cluster in the CBS and are superimposed on the ligand structure. Moreover, there are few waters predicted not to be replaced (blue spheres) inside the CBS, thus highlighting the predictive capacity.

Improved protein–carbohydrate complex prediction combining CWS information in a molecular docking scheme

In a previous study from our group, we showed that using the water sites, determined from explicit water MD simulations, resulted in significant improvement of carbohydrate docking predictions (Gauto et al. 2013). The use of the water site-derived information for docking studies was rationalized based on still previous studies showing that water sites found in the ligand-free receptor in the CBS, mimic the positions of the carbohydrate ligand hydroxyl groups in the corresponding complex (Di Lella et al. 2007; Gauto et al. 2009). So far, in the present work, we have shown that CWSs also, to some extent mimic the carbohydrate hydroxyl positions, thus we decided to test whether the information computed for the CWS could be used to improve carbohydrate docking calculations. For this sake, we compared the performance of the conventional (or unmodified) Autodock4 docking method (CADM), with the same docking scheme and protocol but using the information from the CWS, which we will call water site biased docking method (WSBDM), see the "Computational methods" section for details. We used as test cases 18 different

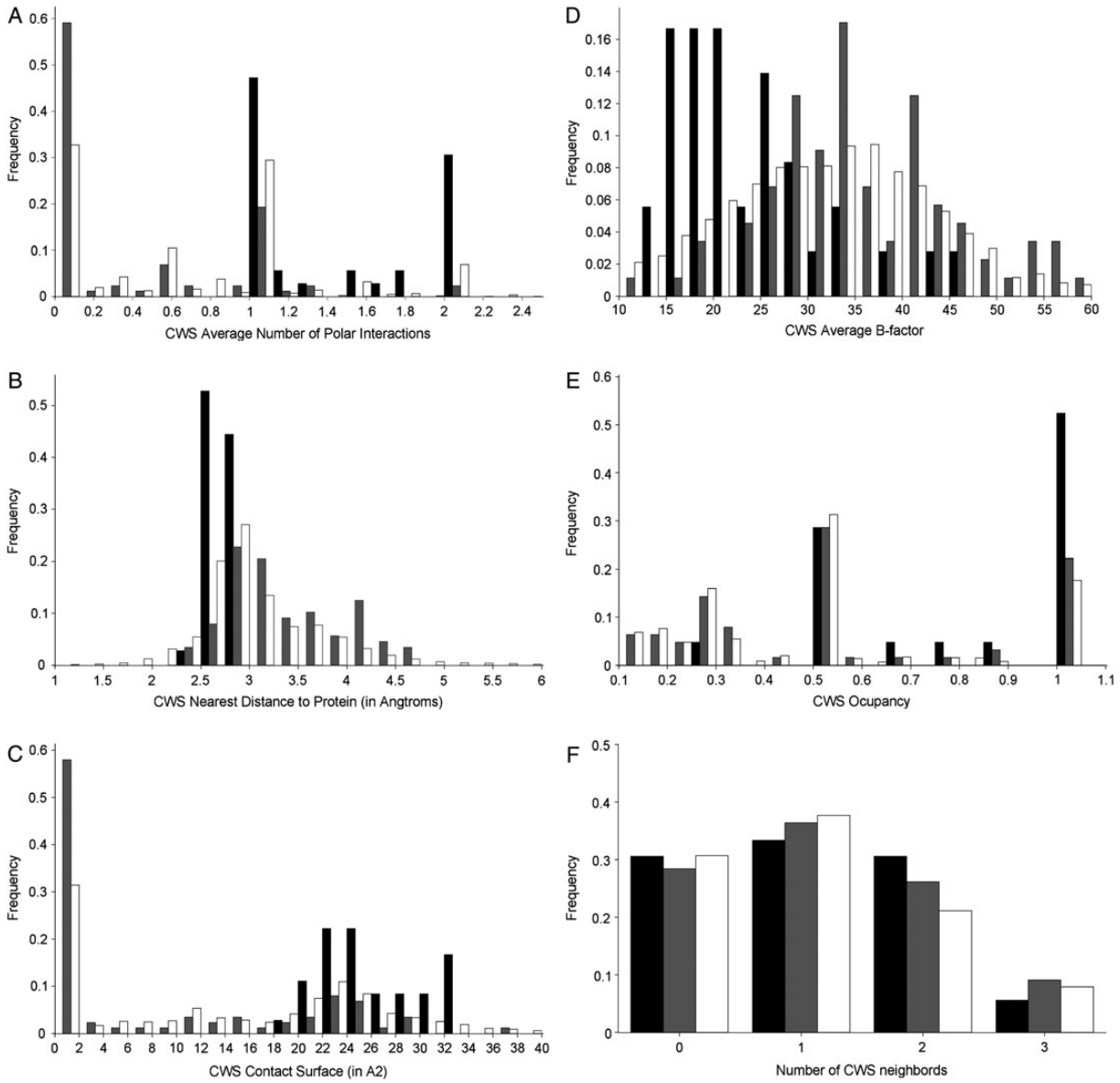


Fig. 5. Distribution histograms for: (**A**)The number of hydrogen bonds formed by each CWS, (**B**) Nearest distance to protein, (**C**) CWS contact surface, (**D**) Average β-factor of the CWS, (**E**) CWS occupancy and (**F**) number of CWS neighbors. Replaced, displaced and out of CBS are shown in black, gray and white columns, respectively.

lectin–carbohydrate complexes (derived from previous analysis), which cover a wide range of protein and ligand types. In all cases, the WSBDM was performed with only those CWSs that are replaced by the ligand –OH groups (best case scenario) and for all CWSs found inside the CBSs (no selection/filter scenario). The results for two selected examples are shown in Figure 7.

The first case corresponds to docking of a monosaccharide to hSPD. The population vs energy plot (Figure 7A) for the conventional method shows that lowest energy-high population cluster has a very high RMSD (3.8 Å), the ligand is placed upside down (Figure 7B) and it is close in energy/population to many other clusters. Moreover, the best cluster still has an RMSD of 2.9 Å and shows that the ligand is wrongly placed. Thus, it can be considered a clear failure. On the other hand both biased methods, that where only replaced (black) or all (gray) CWSs are used to bias the scoring function show a clear outlier, displaying high population and more negative-binding energy. Both predictions are fairly accurate (Figure 7A), as shown by the very small RMSD (<0.7 Å) and perfect fit compared with the reference structure (Figure 7C). The second case corresponds to docking of a trimannoside to Concanavalin A. The population vs energy plot (Figure 7D) for all obtained clusters with the conventional method (white dots) shows no clear outlier in the upper-left corner (i.e., a good candidate). Also both best energy and best population clusters show huge (>7 Å) RMSD against the reference, thus correspond to completely wrong predictions. This is clearly seen in Figure 7E, where the best docked structure is superimposed on the crystal structure complex, and shows that only one of the monomers is correctly placed, while the other two are not. The biased methods on the contrary show clear outliers, with small (<2 Å) RMSD against the reference and thus can be considered fairly good predictions. Indeed the

docked structure is quite well superimposed on the reference complex as shown in Figure 7F, particularly in the two terminal monomers, which are contacting the protein.

Overall analysis: To perform an overall analysis of the comparative docking results, we first characterized each docking calculation, defined by the structure-ligand pair and chosen method, according to the following parameters: (i) The RMSD to the reference (i.e., crystal structure) of the lowest energy (or highest population) complex, and (ii) the predicted complex with the lowest RMSD to the reference structure, together with its ranking, binding energy and population. To visualize this global analysis in Figure 8, we plotted the computed RMSD against the reference complex for the highest ranked (i.e., lowest energy) prediction.

The results from Figure 8 show some interesting trends and highlight improved performance of the CWS-biased docking. For those complexes where the CADM (white columns) fails to correctly predict the right complex, i.e., those cases where the first-ranked complex displays a high RMSD (>3 Å), the CWBDM (dark gray and black bars) dramatically improves the prediction yielding a correct structure (RMSD <2 Å) in most cases and always reducing the RMSD. For those cases where the CADM yields moderate (RMSD ~3 Å) or good results (RMSD <2 Å), both methods show similar performance, with some cases showing lower RMSD with the biased method. Results also show that there is only a slight improvement selecting only those CWSs that are replaced (compare dark gray and black bars). As was observed previously (Gauto et al. 2013) the inclusion of additional CWS that are known not to be replaced by the particular ligand, does not affect the predicting capacity in a significant way.

Table I. Prediction of replaced CWSs using Naive Bayes classifier

CWS type	CWS Total number in set	Predicted to be replaced	Predicted not to be replaced	% TP	%FP
Replaced	39	36	3	92%	–
Displaced	88	21	67	–	28%
Out of CBS	4078	1515	2563	–	38%

%TP is the % of true positive predictions, i.e., those CWSs predicted to be replaced that are effectively replaced. %FP is the % of false-negative predictions, i.e., those CWSs predicted to be replaced, which are not.

In order to test whether a post-scoring selection of the CADM obtained poses using the solvent structure constraints (i.e., the modified scoring function) also leads to better predictions, we re-scored for all cases, all the obtained poses using the modified function. The resulting RMSD against the reference for all cases for the re-scored best pose is also shown in Figure 8 as light gray bars. The results show that post-re-scoring of the CADM poses also significantly improves the docking prediction (compare light gray and with white bars). However, it seems that including the modified function in the conformational search (dark gray and black bars) is able to find better solutions in some cases (3G81 and 1ONA). In summary, these results show that while AutoDock4 conformational search is able to find correct poses, the main problem lies in the scoring function. A similar result was obtained by (Nivedha et al. 2014) considering carbohydrate ligand conformations showing that adding better score for the glycosidic torsion angles, also improves docking results.

We now turn our attention to the method discriminating capability, which can be thought of a measure of its precision. For this sake, we determined the differences in the predicted binding free energy ($\Delta\Delta G_B$) and in the cluster population (ΔPop) of the best complex (that with the lowest RMSD) and the best ranked of the remaining complexes. A negative $\Delta\Delta G_B$ value implies that best obtained

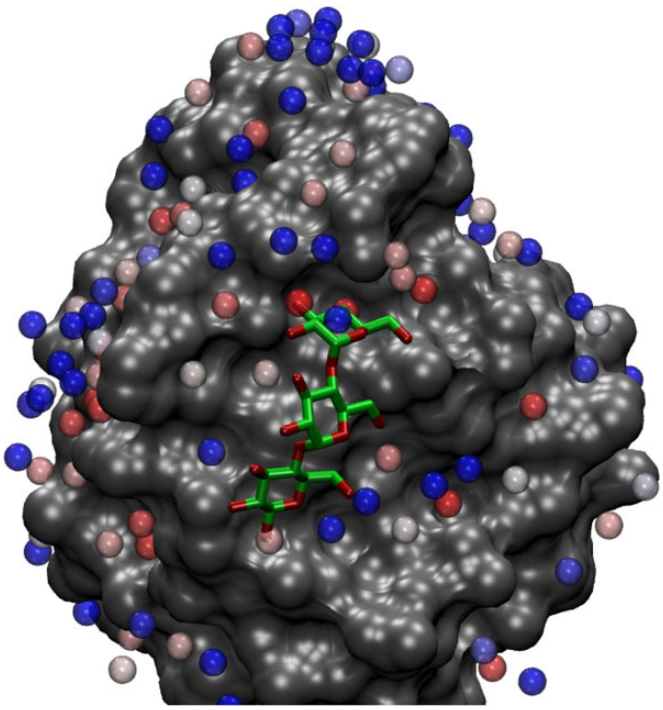


Fig. 6. Protein CWSs colored according on their relative likelihood of being replaced (PR/Pnot) superimposed on the Protein–ligand complex structure. Color code goes from Red PR >> Pnot to Blue Pnot >> PR in a log scale.

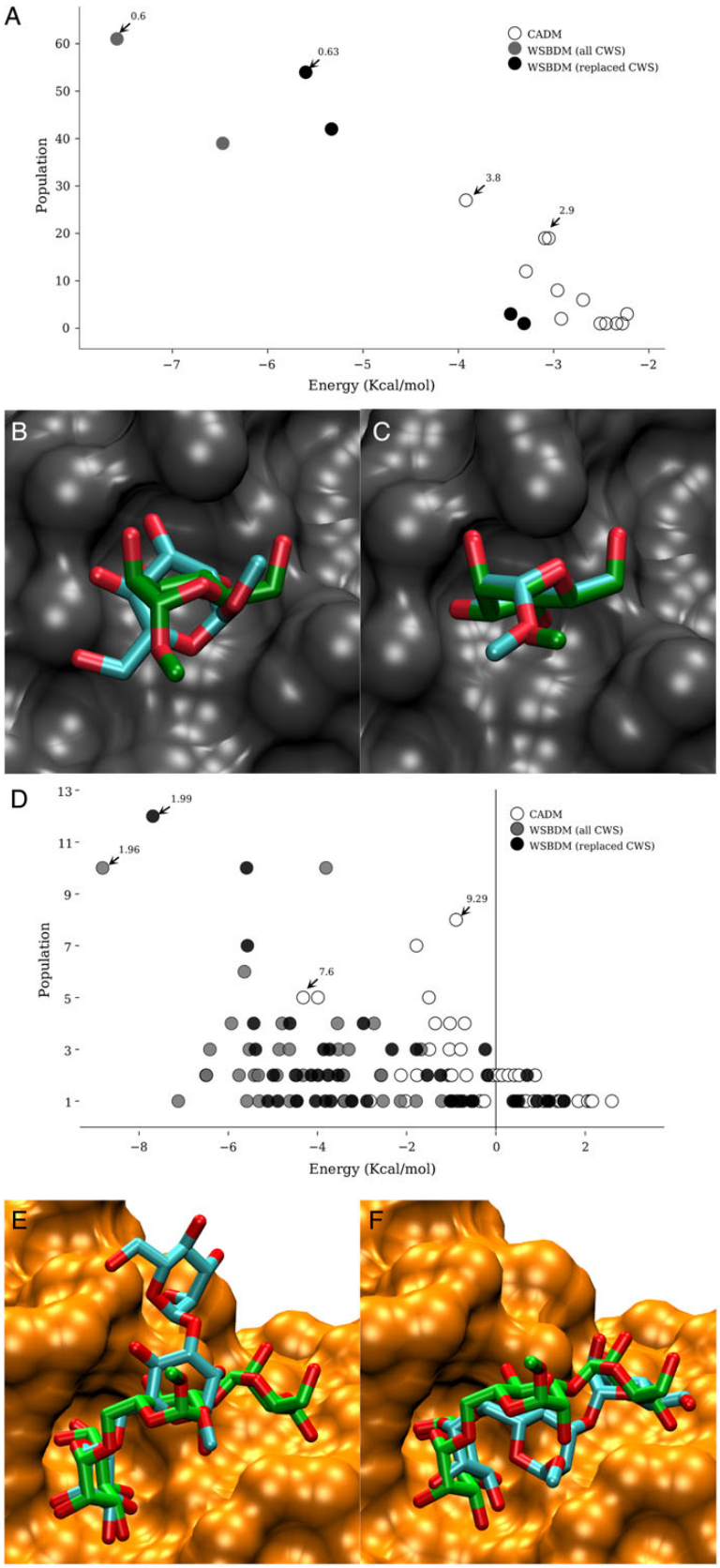


Fig. 7. Population vs binding energy plots for the docking with CADM and WSBDM. (**A**) Results for re-docking of mannose to the crystal structure of hSP-D (PDB ID 3G81) using CADM (**B**) or WSBDM (**C**). Population vs binding energy plots for the docking with CDM, CWSDM with replaced CWS and CWSDM with all CWSs in the CBS (**D**). Results for re-docking of trimannoside to the crystal structure of ConA (PDB ID 1ONA) using CADM (**E**) or WSBDM (**F**). The values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure.

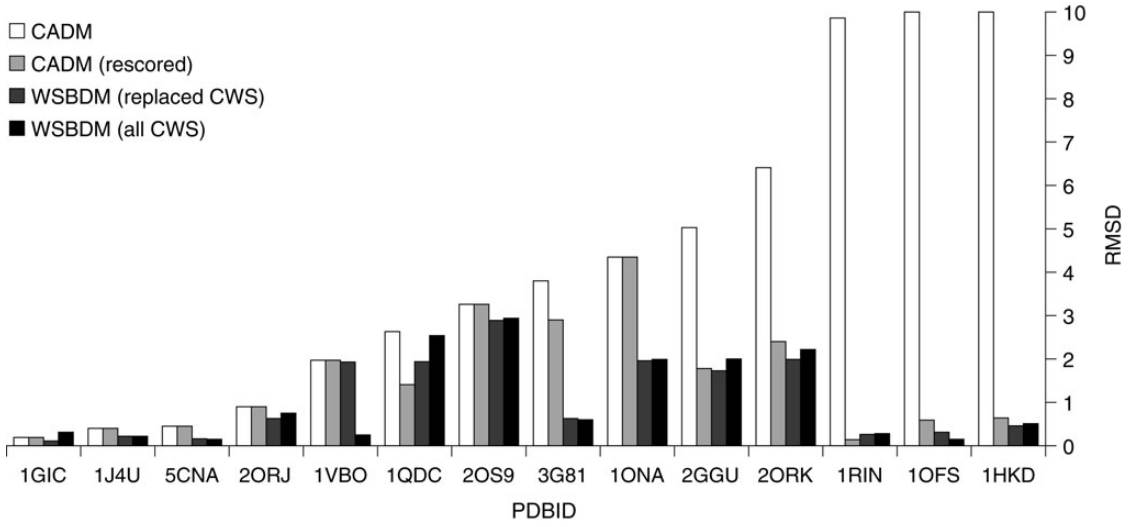


Fig. 8. Heavy atom ligand RMSD for the first-ranked complex against the reference complex structure for the docking results obtained with the CADM (white bars), WSBDM using all CBS CWSs (gray bars) and WSBDM using only the replaced CWSs (best case, black bars).

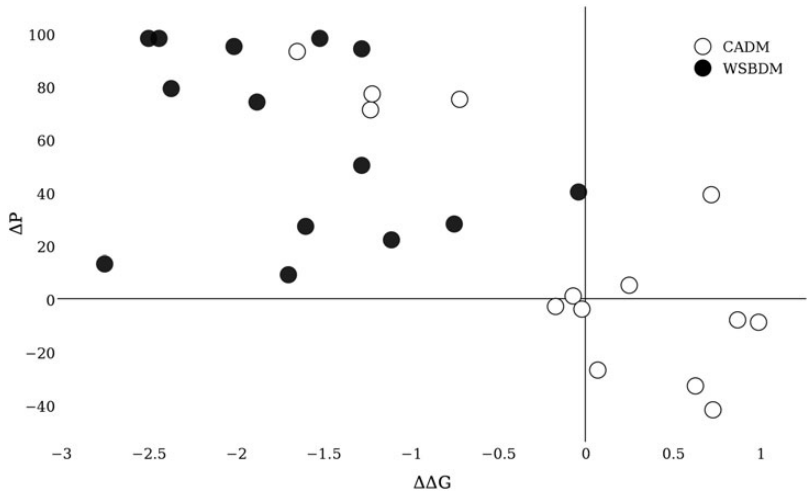


Fig. 9. $\Delta\Delta G_B$ vs ΔPop plot for the docking calculations performed with the CADM (white dots) and WSBDM (black dots). For the definition of $\Delta\Delta G_B$ and ΔPop see text.

complex has better binding energy than any other predicted complex, and the magnitude of $\Delta\Delta G_B$ measures the difference in energy between the best obtained prediction and the first-ranked prediction, while a positive $\Delta\Delta G_B$ means that best complex is among false-positives. Similar reasoning applies for ΔPop. Results located in the upper-left corner of the plot correspond to those cases where the best obtained complex is correctly ranked and if its RMSD against reference is small (<2 Å) it thus corresponds to successful prediction. The corresponding results for all tested cases with both conventional and biased docking methods, presented in Figure 9, show that, there is a clear preference for the biased method results to be located in the upper-left quadrant, while most results obtained with the CADM are closer to the center or even in the lower right quadrant. This means that the biased method has a significant higher discriminating capacity between correct complex and false-positives, i.e., predictions with lower energy or higher population that place the ligand in a wrong way.

Discussion

The aim of the present work was to analyze the relationship between the solvent structure adjacent to the protein surface as characterized by the crystallographic waters, which define the CWS, and the structure of the corresponding lectin–carbohydrate complexes. Our results started from the analysis of lectin–ligand interactions focusing in the sugar polar (mainly hydroxyl) groups, which establish tight hydrogen bonds with the protein, followed by a characterization of the CWS structure and their relation with the ligand polar groups. These results clearly showed that the solvent structure mimics the structural framework of the carbohydrate ligand hydroxyl groups, a fact that is in line with several experimental crystallographic and binding studies (Li and Lazaridis 2005; Kadirvelraj et al. 2008; Saraboji et al. 2012; Garcia-Sosa 2013; Grant and Woods 2014) as well as previous works from our group on the subject (Di Lella et al. 2007; Gauto et al. 2009;2013). Garcia-Sosa, for example, (Garcia-Sosa 2013) analyzed

over 2000 hydrated and nonhydrated protein–ligand complexes and found that although tightly bound, bridging water molecules may in some cases be replaced and targeted as a strategy, sometimes keeping them as bridges may be a better strategy.

Novel to the present study is our further characterization of each CWS using several properties that can be derived directly from the structure, like the CWS B-factor and its number of polar (hydrogen bonds) contacts with the protein, whose results showed significant differences in the displayed properties for those CWSs that will be replaced by a ligand polar group and those that not. These differences were coded in a Naive Bayes classifier, which allows prediction of which CWSs have highest probability of being replaced, and allows when looking at all CWSs from a crystal structure, identify the potential ligand-binding site. The use of crystallographic water properties to infer potential ligand-binding sites is well-documented in the literature with a varied degree of success (García-Sosa et al. 2003; Barillari et al. 2007; Beuming et al. 2012) and the present results encourage the use of these strategies. The novelty of the present approach in this context lies in the specificity for lectins and the further use of the derived CWS key properties in molecular docking calculations.

The use of solvent structure has been widely used for structural predictions (Rarey et al. 1999; Forli and Olson 2012; Garcia-Sosa 2013). For example, Huang and Schoichet (2008) and Forli and Olson (2012) developed force field for docking, which allows including displaceable waters that bridge protein–ligand interactions(while Lie et al. 2011) included discrete waters as part of the ligand for the same aim. Other works analyzed as in the present work, the relation of solvent structure adjacent to the protein surface to the protein's ligand binding properties (García-Sosa et al. 2003; Barillari et al. 2007; Beuming et al. 2012). In this context, it is important to remark that the present improved docking method, is similar to our previous developed method using MD derived WS (Gauto et al. 2013), and is based on the use of a biasing potential to guide the ligand –OH groups to the position of those CWSs more likely to be replaced and with predicted tighter binding. The results clearly show that carbohydrate docking is significantly improved both in its accuracy, measured as the capacity to predict the complex structure close to the one obtained by X-ray crystallography, and also its capacity for differentiating the correct complex among wrong predictions (precision). The main difference between present and previous method is related to how water sites are obtained. In the previous work, water sites were computed using moderately long MD simulations and applying a statistic thermodynamic-based analysis that requires expertise and is time-consuming. On the other hand, present analysis based on the definition of CWS is straight forward once crystal structure is available. Most important is that both methodologies show significant and similar improvement over conventional docking method. The improved docking underscores the first conclusion of the present work that CWSs mimic ligand hydroxyl group framework and provide a fast and easily implemented methodology that allows better predictions of protein–carbohydrate complexes.

Taking the results altogether, a nice picture emerges not only concerning the predictive role of the solvent structure for the resulting complex, but also in terms of how lectins recognize and select their ligands (Dam & Brewer 2010; Gabius et al. 2011). Our results show that for monosaccharides, usually no more than three OH groups are in contact with the protein, which is consistent with a binding from "one side"; that most lectins recognize only one or two monosaccharide units, even if larger ligands can be accommodated and that the total of contact groups is hardly more than 4. For instance, when talking of the sugar code, for a monosaccharide hexose recognition lectin, 16 possibilities are usually considered (alpha or beta, Glu, Man, Gal, Ido, Tal, Alt and Gul), but if only three oxygens participate and one of them is not quiral (i.e., C6–OH) half the possibilities remain for discrimination. Similarly, if we consider β(1–4)-linked disaccharides, usually the number of possibilities can be estimated as $N \times 8^2$, where N is the number of stable different conformations that the glycosidic bond can adopt. However, if only four oxygens (two from each unit) are contacting with the protein, the number reduces four times and even more if one of the bound –OH groups is not chiral. Thus, although the sugar code is still theoretically huge, the boundary conditions defined by the nature of the recognition process, in terms of the number of monosaccharide units and –OH groups that can be in contact with the protein, significantly reduce this number. Moreover, given the increasing evidence, as that presented here among other works (Li and Lazaridis 2005; Kadirvelraj et al. 2008; Gauto et al. 2009; Saraboji et al. 2012), of the tight relationship between solvent structure and lectin–saccharide complex, a potential rationalization and predictive method emerges for trimming down the sugar code complexity.

Conclusion

Analysis of the solvent structure adjacent to the binding sites of carbohydrate-binding proteins as derived from the crystal structures shows that (i) CWSs mimic the ligand hydroxyl structural framework in the resulting protein–ligand complex, (ii) CWS properties allow predicting their likelihood of being replaced by a ligand polar group and (iii) CWS properties can be used to bias and improve carbohydrate docking calculations. The presented analysis framework thus provides a powerful tool for the advancement in our basic understanding of protein–carbohydrate complexes, and their interactions as well as for the development of glycomimetic drugs. The results also highlight key properties of lectin specificity and ligand recognition properties.

Computational methods

Lectin–carbohydrate complex data set

The working data set was built starting with all available structures in the PDB (updated at April 2014), (Bernstein et al. 1977). We built a curated list of all possible natural and nonmodified carbohydrate ligands and retained only those proteins, where a structure was found bound to any of these. Proteins were compared on sequence basis and those with over 95% identity and >80% coverage of the carbohydrate recognition domain were considered as "the same". Thus, all structures whose sequences are this similar comprise, and are joined in a unique protein set. We further trimmed down the sets by keeping only those corresponding to lectins and sets with at least one structure in the presence, and one in the absence (i.e., apo-structure) of any carbohydrate ligand. These filters result in a total of 19 unique lectin sets, with a total of 167 structures of lectin–carbohydrate complexes and 75 apo-structures. We also grouped unique sets in families, resulting in eight different families. All statistical analyses were performed considering (whenever possible) each individual structure, each individual set and an average value determined for each protein family.

Definition of the CBS

Each carbohydrate ligand was classified by its saccharide number (monosaccharide, disaccharide etc.) computed by linking the chEBI Ontologies database (Hastings et al. 2013) with each ligand name in the protein data bank. To define the CBS of each individual protein set, we decided that a residue is defined as forming the CBS if any of each

heavy atoms, is found at less than 3 Å from any ligand heavy atom. The combination of all the residues assigned to the CBS as derived from all structures in the set, which are also present in all the structures, thus conforms to the arrangement of CBS residues. For each protein set, we finally aligned structurally all monomeric carbohydrate recognition domains, using all the heavy atoms of the residues forming the above defined CBS. If any structure(s) shows a >2 Å RMSD deviation to most of the other structures, it is discarded. This is performed to eliminate structures that show large conformational changes upon ligand binding. The resulting aligned set of structures were used to compute and analyze the following properties.

Determination of carbohydrate parameters

To analyze the protein–carbohydrate interactions we determined the following structural parameters for each ligand. First we determined the actual number of monomers (or saccharide units) that are in contact with the protein. A monomer was defined as in contact with the protein if there is at least one interaction, either hydrogen bond or nonpolar. For all ligand polar (N and O) atoms the number and nature of hydrogen bonds with the protein were determined. A hydrogen bond was defined as present whenever donor and acceptor heavy atoms were closer than 3.5 Å. A nonpolar interaction was defined as present whenever a carbon atom from the saccharide framework was found at <5 Å than a protein nonpolar atom.

Determination of crystallographic water structural parameters

To analyze the properties of the crystallographic waters, we first defined the presence of specific CWS, in a similar way as that used in our previous works, based on explicit water MD simulations (Di Lella et al. 2007; Gauto et al. 2009;2013). CWSs are defined by the presence of crystallographic waters in the available aligned structures. To defined CWSs across several structures, if any crystallographic waters from different structure are closer than 1.4 Å, they are combined and then, CWS position is then defined by the center of mass of all resulting oxygen atoms that define it.

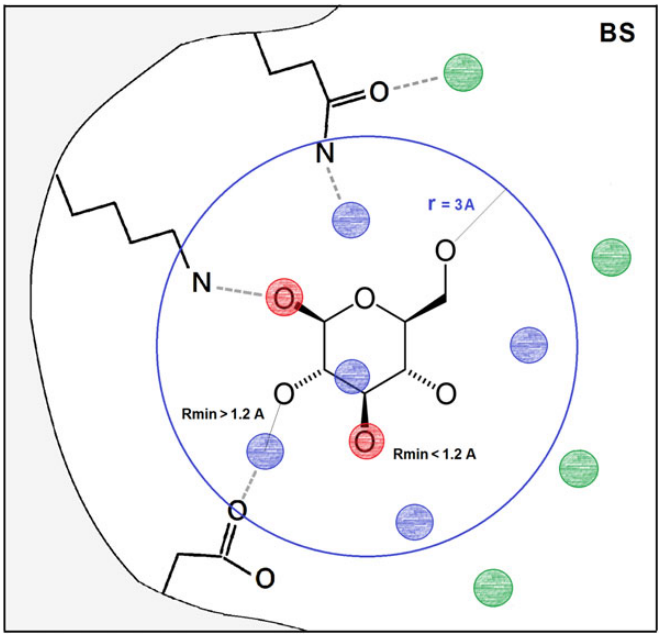
For each identified CWS we computed the following parameters: (i) The average number of polar interactions, as the number of hydrogen bonds that the CWS establishes with the protein (the presence of a hydrogen bond was defined whenever the CWS was closer than 3.5 Å from any protein hydrogen bond donor or acceptor atom); (ii) the mobility of the water molecule computed as the average of the reported B-factor from all oxygen atoms from crystallographic data belonging to water molecules that define the CWS; (iii) the occupancy of water molecules, which is what we can call the crystallographic water finding probability, and is computed as the ratio between the number of structures from the set where a crystallographic water is found and the total number of ligand-free structures in the set; (iv) the closeness between the oxygen's water molecules and the protein, which is defined as the distance between the CWS and the nearest protein heavy atom; (v) the contact of surface between the crystallographic water oxygen and the protein and (vi) the average number of neighboring crystallographic water molecules to the CWS, looking how many crystallographic waters are closer than 2.8 Å of the selected CWS); (vii) finally, we computed the R_{min} parameter, which compares the position of the CWS with that of the ligand in the corresponding lectin–carbohydrate complex. The R_{min} is defined by the minimum distance between the CWS position and any ligand heavy atom. This distance is determined after structurally aligning all the apo-structures that define the CWS position and the complex structures that define the ligand position. Based on the value of R_{min}, we classified all the identified CWS in three categories. Those CWSs, which have R_{min} values smaller than 1.2 Å to any polar heavy atom of the ligand, will be classified as replaced. In other words, these are the CWSs that are replaced by ligand polar groups in the complex. The remaining CWSs inside the CBS, which are missing in any of the complex structure, are classified as displaced CWSs. This CWSs are those that either are displaced by the ligand due to steric interactions, but are not replaced by a ligand polar group. This definition also includes those CWSs that may be missing in the complex crystal not due to directly steric hindrance of the ligand but due to loss of key interactions. The remaining CWSs, which are not affected by the presence of the ligand and are not in contact with any of the residues defining the CBS are referred as out of CBS CWS. A scheme showing the three types of CWS is presented below (Scheme 1).

Conventional AutoDock4 docking method

To analyze the improvement on the carbohydrate docking process using the information derived from the CWSs we compared, as in our previous work, the performance of the Autodock4 program using its usual parameters, which we will call the Conventional AutoDock4 docking method (CADM), and a modified protocol that includes information derived from the CWSs, which we will call the WSBDM. The CADM was performed with the AutoDock4.2 program (Morris et al. 2009) using the same parameters and strategy as in our previous work. Briefly, based solely on the protein receptor structure, the energy grids for each ligand atom type, are computed. The grid size and position were chosen so that they include the whole CBS. This was achieved by placing the grid center in the geometric center of the CBS, and extending its size 20 Å (for the mono- and disaccharide-binding proteins) and 25 Å (for the tri- and tetrasaccharide-binding proteins) in each direction using a spacing of 0.375 Å. For each complex, 100 different docking runs were performed and the results were clustered according to the ligand-heavy atom RMSD using a cut-off of 1.5 Å. The genetic algorithm parameters for each conformational search run were kept at their default values (150 for initial population size, 2.5×10^6 as the maximum number of energy evaluations and 2.7×10^4 as the maximum number of generations).

Water sites biased docking method

The present Water sites biased docking method (WSBDM) is based on the method developed and thoroughly tested in our previous work (Gauto et al. 2013). There, in order to take advantage of the fact that carbohydrate –OH groups tend to occupy or replace the positions of the MD derived water sites, and that there is a positive correlation between the water site's water finding probability and its possibility of being replaced, and a negative correlation with its R_{90}, we modified the AutoDock4 energy function, adding an additional energy term for each carbohydrate–ligand oxygen (OA atom type) to the original function, whose deepness is proportional to the probability and size to the R_{90}. Analyses of many water sites derived from MD simulations show that water finding probabilities lie in the 1–20 range, thus resulting in a 0–1.7 kcal/mol energy scale for the deepness of the OA atom type energy well, while determined R_{90} values are in the range of 0.4–4.4. In the present work, and in order to perform a modification of the scoring function in the same spirit, we used as key parameters the number of polar interactions performed by the CWS with the protein, and the CWS β-factor, which is a measure of its dispersion. The resulting modified scoring function is described then by the



Scheme 1. CWS's categories. Red for "O-replaced", blue for "displaced", green for "Outside CRD". Dotted lines represent H-bonds (H atoms have been omitted for clarity). All distances in angstroms.

following equation:

$$\Delta G_{\mathrm{O}}^{\mathrm{M}} = \Delta G_{\mathrm{O}}^{\mathrm{AD}} - RT\sum_{i=1}^{N}\ln(PCP_i)\; e^{\frac{-\left((x-x_{\mathrm{WS},i})^2+(y-y_{\mathrm{WS},i})^2+(z-z_{\mathrm{WS},i})^2\right)^{1/2}}{\mathrm{CWSD},i}}. \quad (1)$$

where ΔG_{M} corresponds to the resulting "modified" scoring function, ΔG_{AD} corresponds to the original AutoDock4 scoring function, PCP is the polar contact probability, x, y and z are grid point coordinates, X_{WS}, Y_{WS} and Z_{WS} are the corresponding CWS position coordinates and CWSD is the crystallographic water site dispersion. $\mathrm{PCP_i}$ is computed as number of possible polar interaction with the protein multiplied by scaling factor (11), which relates the number of polar interactions derived from the analysis of the crystal structure, with the water finding probability derived from explicit water MD, which was used in our previous work. The CWSD is computed as $\sqrt{(\text{B-factor}/4\Pi^2)}$ and similarly to the scaling factor applied to $\mathrm{PCP_i}$ allows obtaining values that are in the same scale as those found for R_{90} values in the MD derived water sites.

In this modified function, each CWS considered provides thus an interaction energy between the center of the CWS position and every OA atom type (i.e., any carbohydrate oxygen), with a magnitude that is proportional to the number of hydrogen bonds it can establish with the protein and an amplitude that is related to the dispersion of the CWS as measured by its β-factor. To the modified function, first the receptor grid for all OA atom type is built using the Autogrid4 program using the standard scoring function. Then, the OA atom type grid is modified with a homemade script to include the bias potential. The WSBDM is then employed in the same manner as the CADM but by introducing the modified function computed with all (or a selected number of) CWS and their corresponding parameters. For strict comparison purposes, all other docking parameters, were the same as those used in the CADM. All scripts to analyze crystal structures, determine and characterize the CWS and modify the Autodock4 grids are available under request.

CADM rescoring analysis based on CWS position

To re-score the results obtained with the CADM, we simply took all poses obtained with the CADM and recomputed for all of them the binding energy using the modified scoring function described above. In this sense, the function is not used for the conformational search, and only to sort the results.

Docking data analysis

To compare the conventional and biased docking methods, we considered two main issues: first, how close to the reference complex structure the method docks the corresponding ligand, thus resulting in a measure of the method accuracy. This is computed as the ligand heavy atoms RMSD of each predicted complex using each method, with respect to the position of the ligand in the corresponding complex crystal reference structure; secondly, what is the method capability to distinguish the right complex from wrong predictions, a parameter that may be thought of as the method precision. This in Autodock4 is done by looking at two parameters, the predicted binding free energy (ΔG_{B}) derived from the original or modified scoring function, and the population (%Pop), which is the percentage of individual docking runs that resulted in the same binding mode for a particular receptor structure ligand pair. Best possible results should give a high population and low binding energy conformation, which also significantly differs in both parameters from the others (assigned as false-positives). As shown in the results section, this can be easily analyzed by plotting population vs binding energy for all obtained predicted complexes in the given docking calculation.

Funding

Computer power was provided by Centro de Computación de Alto Rendimiento (C.E.C.A.R.) at the FCEN-UBA and by the cluster MCG PME no. 2006-01581 at the Universidad Nacional de Cordoba. Research was funded by grants PICT-2010-416, UBACyT2012, PICTO-GSK-2012-0057 and PIP 112 201101 00850 awarded to M.A.M. M.A.M. is member of CONICET.

Conflict of interest

None declared.

Abbreviations

CADM, conventional AutoDock4 docking method; CBS, carbohydrate binding site; CWS, crystallographic water site; CWSD, crystallographic water site dispersion; MD, molecular dynamics; PCP, polar contact probability; PCP_i, polar interaction with the protein; PDB, Protein Data Bank; WSBDM, water site biased docking method; ΔPop, cluster population; $\Delta\Delta G_B$, predicted binding free energy.

References

Abel R, Young T, Farid R, Berne BJ, Friesner RA. 2008. Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J Am Chem Soc*. 130(9):2817–2831.

Agostino M, Jene C, Boyle T, Ramsland P, Yuriev E. 2009. Molecualr docking of carbohydrate ligands to antibodies: Structural validation against crystal structures. *J Chem Inf Model*. 49:2749–2760.

Andres D, Gohlke U, Broeker NK, Schulze S, Rabsch W, Heinemann U, Barbirz S, et al. 2013. An essential serotype recognition pocket on phage P22 tailspike protein forces Salmonella enterica serovar Paratyphi A O-antigen fragments to bind as nonsolution conformers. *Glycobiology*. 23(4):486–494.

Asensio JL, Cañada FJ, Siebert HC, Laynez J, Poveda A, Nieto PM, Soedjanaamadja UM, et al. 2000. Structural basis for chitin recognition by defense proteins: GlcNAc residues are bound in a multivalent fashion by extended binding sites in hevein domains. *Chem Biol*. 7(7):529–543.

Barillari C, Duncan AL, Westwood IM, Blagg J, van Montfort RLM. 2011. Analysis of water patterns in protein kinase binding sites. *Proteins*. 79 (7):2109–2121.

Barillari C, Taylor J, Viner R, Essex JW. 2007. Classification of water molecules in protein binding sites. *J Am Chem Soc*. 129(9):2577–2587.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, et al. 1977. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur Biochem/FEBS*. 80(2):319–324.

Beuming T, Che Y, Abel R, Kim B, Shanmugasundaram V, Sherman W. 2012. Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization. *Proteins*. 80 (3):871–883.

Brooijmans N, Kuntz ID. 2003. Molecular recognition and docking algorithms. *Annu Rev Biophy Biomol Struct*. 32:335–373.

Chervenak MC, Toone EJ. 1995. Calorimetric analysis of the binding of lectins with overlapping carbohydrate-binding ligand specificities. *Biochem*. 34 (16):5685–5695.

Compagno D, Jaworski FM, Gentilini L, Contrufo G, Gonzalez Perez I, Elola MT, Pregi N, et al. 2014. Galectins: Major signaling modulators inside and outside the cell. *Curr Mol Med*. 14(5):630–651.

Crouch E, Hartshorn K, Horlacher T, McDonald B, Smith K, Cafarella T, Seaton B, et al. 2009. Recognition of mannosylated ligands and influenza A virus by human surfactant protein D: Contributions of an extended site and residue 343. *Biochem*. 48(15):3335–3345.

Dam TK, Brewer CF. 2010. Lectins as pattern recognition molecules: The effects of epitope density in innate immunity. *Glycobiology*. 20(3):270–279.

Di Lella S, Martí MA, Alvarez RMS, Estrin DA, Ricci JCD. 2007. Characterization of the galectin-1 carbohydrate recognition domain in terms of solvent occupancy. *J Phys Chem B*. 111(25):7360–7366.

Di Lella S, Martí MA, Croci DO, Guardia CM, Díaz-Ricci JC, Rabinovich GA, Caramelo JJ, et al. 2010. Linking the structure and thermal stability of beta-galactoside-binding protein galectin-1 to ligand binding and dimerization equilibria. *Biochemistry*. 49(35):7652–7658.

Englebienne P, Moitessier N. 2009. Docking ligands into flexible and solvated macromolecules. 5. Force-field-based prediction of binding affinities of ligands to proteins. *J Chem Inform Model*. 49(11):2564–2571.

Fadda E, Woods RJ. 2010. Molecular simulations of carbohydrates and protein-carbohydrate interactions: Motivation, issues and prospects. *Drug Discov Today*. 15(15–16):596–609.

Forli S, Olson AJ. 2012. A force field with discrete displaceable waters and desolvation entropy for hydrated ligand docking. *J Med Chem*. 55(2):623–638.

Frank M, Schloissnig S. 2010. Bioinformatics and molecular modeling in glycobiology. *Cel Mol Life Sci*. 67(16):2749–2772.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, et al. 2004. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 47(7):1739–1749.

Gabius H-J, André S, Jiménez-Barbero J, Romero A, Solís D. 2011. From lectin structure to functional glycomics: Principles of the sugar code. *Trends Biochem Sci*. 36(6):298–313.

Garcia-Sosa AT. 2013. Hydration properties of ligands and drugs in protein binding sites: Tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies. *J Chem Inform Model*. 0:0–00.

García-Sosa AT, Mancera RL, Dean PM. 2003. WaterScore: A novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J Mol Model*. 9 (3):172–182.

Gauto DF, Di Lella S, Estrin DA, Monaco HL, Martí MA. 2011. Structural basis for ligand recognition in a mushroom lectin: Solvent structure as specificity predictor. *Carbohy Res*. 346(7):939–948.

Gauto DF, Di Lella S, Guardia CM, Estrin DA, Martí MA. 2009. Carbohydrate-binding proteins: Dissecting ligand structures through solvent environment occupancy. *J Phys Chem B*. 113(25):8717–8724.

Gauto DF, Petruk AA, Modenutti CP, Blanco JI, Di Lella S, Martí MA. 2013. Solvent structure improves docking prediction in lectin-carbohydrate complexes. *Glycobiology*. 23(2):241–258.

Grant OC, Woods RJ. 2014. Recent advances in employing molecular modelling to determine the specificity of glycan-binding proteins. *Curr Opin Struct Bio*. 28C:47–55.

Green PJ. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 82(4):711–732.

Guan L, Hu Y, Kaback HR. 2003. Aromatic stacking in the sugar binding site of the lactose permease. *Biochemistry*. 42(6):1377–1382.

Guardia CM, Gauto DF, Di Lella S, Rabinovich GA, Martí MA, Estrin DA. 2011. An integrated computational analysis of the structure, dynamics, and ligand binding interactions of the human galectin network. *J Chem Inform Model*. 51(8):1918–1930.

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, et al. 2013. The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013. *Nucleic Acids Rese*. 41(database issue):D456–D463.

Higgs C, Beuming T, Sherman W. 2010. Hydration site thermodynamics explain SARs for triazolylpurines analogues binding to the A2A receptor. *ACS Med Chem Lett*. 1(4):160–164.

Huang N, Schoichet BR. 2008. Exploiting ordered waters in molecular docking. *J Med Chem*. 52:4862–4865.

Hummer G. 2010. Molecular binding: Under water's influence. *Nat Chem*. 2 (11):906–907.

Johal AR, Jarrell HC, Letts JA, Khieu NH, Landry RC, Jachymek W, Yang Q, et al. 2013. The antigen-binding site of an N-propionylated polysialic acid-specific antibody protective against group B meningococci is consistent with extended epitopes. *Glycobio*. 23(8):946–954.

Kadirvelraj R, Foley BL, Dyekjaer JD, Woods RJ. 2008. Involvement of water in carbohydrate-protein binding: Concanavalin A revisited. *J Am Chem Soc*. 130(50):16933–16942.

Kerzmann A, Fuhrmann J, Kohlbacher O, Neumann D. 2008. BALLDock/SLICK: A new method for protein-carbohydrate docking. *J Chem Inform Model*. 48(8):1616–1625.

Kumar S, Frank M, Schwartz-Albiez R. 2013. Understanding the specificity of human Galectin-8C domain interactions with its glycan ligands based on molecular dynamics simulations. *PloS One*. 8(3):e59761.

Laughrey ZR, Kiehna SE, Riemen AJ, Waters ML. 2008. Carbohydrate-pi interactions: What are they worth? *J Am Chem Soc*. 130(44):14625–14633.

Lazaridis T. 1998. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *J Phys Chem B*. 102(18):3531–3541.

Leach AR, Shoichet BK, Peishoff CE. 2006. Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. *J Med Chem*. 49 (20):5851–5855.

Li Z, Lazaridis T. 2005. The effect of water displacement on binding thermodynamics: Concanavalin A. *J Phys Chem B*. 109(1):662–670.

Li Z, Lazaridis T. 2006. Thermodynamics of buried water clusters at a protein-ligand binding interface. *J Phys Chem B*. 110(3):1464–1475.

Lie MA, Thomsen R, Pedersen CNS, Schiøtt B, Christensen MH. 2011. Molecular docking with ligand attached water molecules. *J Chem Inform Model*. 51(4):909–917.

Lütteke T, Frank M, von der Lieth C-W. 2005. Carbohydrate structure suite (CSS): Analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res*. 33(database issue):D242–D246.

Maruyama O. 2013. Heterodimeric protein complex identification by Naïve Bayes classifiers. *BMC Bioinform*. 14:347.

Michel J, Tirado-rives J, Jorgensen WL. 2009. Energetics of displacing water molecules from protein binding sites: Consequences for ligand optimization. *J Am Chem Soc*. 9(9):15403–15411.

Mishra SK, Adam J, Wimmerová M, Koča J. 2012. In silico mutagenesis and docking study of Ralstonia solanacearum RSL lectin: Performance of docking software to predict saccharide binding. *J Chem Inform Model*. 52 (5):1250–1261.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*. 19 (14):1639–1662.

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*. 30:2785–2791.

Nishio M, Umezawa Y, Fantini J, Weiss MS, Chakrabarti P. 2014. CH-π hydrogen bonds in biological macromolecules. *Phy Chem Phys*. 16(25): 12648–12683. The Royal Society of Chemistry.

Nivedha AK, Makeneni S, Foley BL, Tessier MB, Woods RJ. 2014. Importance of ligand conformational energies in carbohydrate docking: Sorting the wheat from the chaff. *J Comput Chem*. 35(7):526–539.

Patil A, Huard D, Fonnesbeck CJ. 2010. PyMC: Bayesian stochastic modelling in Python. *J Stat Software*. 35(4):1–81.

Ranzinger R, Herget S, Wetter T, von der Lieth C-W. 2008. GlycomeDB—Integration of open-access carbohydrate structure databases. *BMC Bioinform*. 9:384.

Rarey M, Kramer B, Lengauer T. 1999. The particle concept: Placing discrete water molecules during protein-ligand docking predictions. *Proteins*. 34 (1):17–28.

Saraboji K, Håkansson M, Genheden S, Diehl C, Qvist J, Weininger U, Nilsson UJ, et al. 2012. The carbohydrate-binding site in galectin-3 is preorganized to recognize a sugarlike framework of oxygens: Ultra-high-resolution structures and water dynamics. *Biochemistry*. 51(1): 296–306.

Setny P, Baron R, McCammon J. 2010. How can hydrophobic association be enthalpy driven? *J Chem Theory Comput*. 6(9):2866–2871.

Sujatha MS, Sasidhar YU, Balaji PV. 2004. Energetics of galactose- and glucose-aromatic amino acid interactions: Implications for binding in galactose-specific proteins. *Protein Sci Publ Protein Soc*. 13(9):2502–2514.

Taylor RD, Jewsbury PJ, Essex JW. 2002. A review of protein-small molecule docking methods. *J Comput Aided Mol Des*. 16(3):151–166.

Terraneo G, Potenza D, Canales A, Jiménez-Barbero J, Baldridge KK, Bernardi A. 2007. A simple model system for the study of carbohydrate—Aromatic interactions. *J Am Chem Soc*. 129(10):2890–2900.

Von der Lieth C-W, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, et al. 2011. EUROCarbDB: An open-access platform for glycoinformatics. *Glycobiology*. 21(4):493–502.

Von Schantz L, Håkansson M, Logan DT, Walse B, Osterlin J, Nordberg-Karlsson E, Ohlin M. 2012. Structural basis for carbohydrate-binding specificity—A comparative assessment of two engineered carbohydrate-binding modules. *Glycobiology*. 22(7):948–961.

Wang J-C, Lin J-H, Chen C-M, Perryman AL, Olson AJ. 2011. Robust scoring functions for protein-ligand interactions with quantum chemical charge models. *J Chem Inform Mode*. 51(10):2528–2537.

Yoshida H, Teraoka M, Nishi N, Nakakita S, Nakamura T, Hirashima M, Kamitori S. 2010. X-ray structures of human galectin-9 C-terminal domain in complexes with a biantennary oligosaccharide and sialyllactose. *J Biol Chem*. 285(47):36969–36976.

Yuriev E, Agostino M, Ramsland PA. 2009. Challenges and advances in computational docking: 2009 in review. *J Recogn*. 24(2):149–164.