Glycobiology Advance Access published October 22, 2012

Solvent structure improves docking prediction in Lectin-Carbohydrate complexes

Diego F. Gauto,[2, 3] Ariel A. Petruk,[2] Carlos P. Modenutti,[1] Juan I. Blanco,[1] Santiago Di Lella[1] and Marcelo A. Martí.[1,2,3]*

[1] Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.

[2] Departamento de Química Inorgánica, Analítica y Química Física, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.

[3] Instituto de Química de los Materiales, Medio Ambiente y Energía (INQUIMAE), CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires, C1428EHA, Argentina.

*) To whom correspondence should be sent: marcelo@qi.fcen.uba.ar

Keywords: Docking, Lectin, Carbohydrate, Saccharide, Solvent Structure, Water Site, Hydration Site, Proteins, AutoDock4, Complex, Galectins,

ABSTRACT

Recognition and complex formation between protein and carbohydrates is a key issue for many important biological processes. Determination of the three-dimensional structure of such complexes is thus most relevant, but particularly challenging due to their usually low binding affinity. *In-silico* docking methods have a long standing tradition in predicting protein-ligand complexes, and allow the potentially fast exploration of a number of possible protein-carbohydrate complex structures. However, determining which of these predicted complexes represents the correct structure is not always straightforward.

In this work, we present a modification of the scoring function provided by Autodock4, a widely used docking software, based on the analysis of the solvent structure adjacent to the protein surface, as derived from Molecular Dynamics simulations, that allows the definition and characterization of regions with higher water occupancy than the bulk solvent, called Water Sites. They mimic the interaction held between the carbohydrate -OH groups and the protein. We used this information for an improved docking method in relation to its capacity to correctly predict the protein-carbohydrate complexes for a number of tested proteins, whose ligands range from the mono- to the tetrasaccharide size. Our results show that the presented method significantly improves the docking predictions. The resulting solvent-structure-biased docking protocol therefore appears as a powerful tool for the design and optimization of glycomimetic drugs development, while providing a new insight for the basic understanding of protein carbohydrate interactions. Moreover, the achieved improvement also underscores the relevance of the solvent structure for the protein carbohydrate recognition process.

1. Introduction

Formation of protein-ligand complexes is one of the most fundamental processes in biochemistry. For a given protein, identifying with high precision which ligands should be bound –and which should not– is thus a crucial requisite to accomplish a variety of tasks such as enzyme catalysis, cell communication, signaling, adhesion and differentiation. In a more applied field, the rational design of new and more effective drugs also depends on our knowledge about the specific protein-ligand complexes that can be established.[1-3] In this context, determination of the atomic resolution structure for any given protein-ligand complex, is of fundamental relevance for understanding and characterizing its interactions, with a potential strong impact in both, basic and applied biochemistry.[4,5]

In-silico strategies for predicting the structure of a given protein-ligand (or protein-protein) complex are usually referred as docking methods.[6-9] Widely used in the last decade, they are currently an essential part of many protein biochemical characterization studies, and rational drug design programs programs.[10-12] The potential and reliability of any docking method lies in its capability to correctly predict the complex structure, and therefore the interactions held between the units, taking as starting point the protein and ligand structures separately. Nevertheless, given the approximations involved in the theoretical developments employed, results are not always successfully achieved.[7-9,13-18]

Presently, there are several docking softwares available,[6,19-21] and although several works have compared different docking programs and versions,[15,22,23] there is still no clear best choice. In particular, for sugar docking, the work by Mishra et. al. showed that AutoDock3 performs better than version 4, Vina and DOCK,[24,25] but still yields many false positives.[22] Agostino et. al. on

the other hand, showed that Glide[26] and AutoDock4[27] performed better, but the results were strongly dependent on the particular ligand receptor pair. In any case, one of the most popular, widely used and more important free under the GNU General Public License, docking programs is the AutoDock4. The method combines a genetic algorithm[27] to explore possible binding conformations of the ligand and an empirical function, including electrostatic, hydrophobic and solvation effects, to compute the ligand binding free energy (ΔG_B), and thus ranking the resulting complex structure predictions.[19,28]

During the ligand binding process, significant solvent reorganization is produced along the contact surface. Several works in this area have shown that this reorganization contributes to the ligand binding free energy.[29-34] From a structural viewpoint, and as a result of the interactions held between the protein and the solvent, water molecules are not placed randomly on the macromolecule surface, but instead tend to occupy specific positions and orientations. The latter results in a well defined solvent structure associated to the protein surface, characterized by regions of highly ordered water molecules.[29,35,36] This is especially evident in regions such as protein active sites or ligand binding regions,[37] and together with the fact that displacing these ordered water molecules has been shown to improve and correlate with the experimentally determined binding free energy,[31,34]_ENREF_34 underscores the relevance of such well-defined solvent structure.[29,30,37,38]

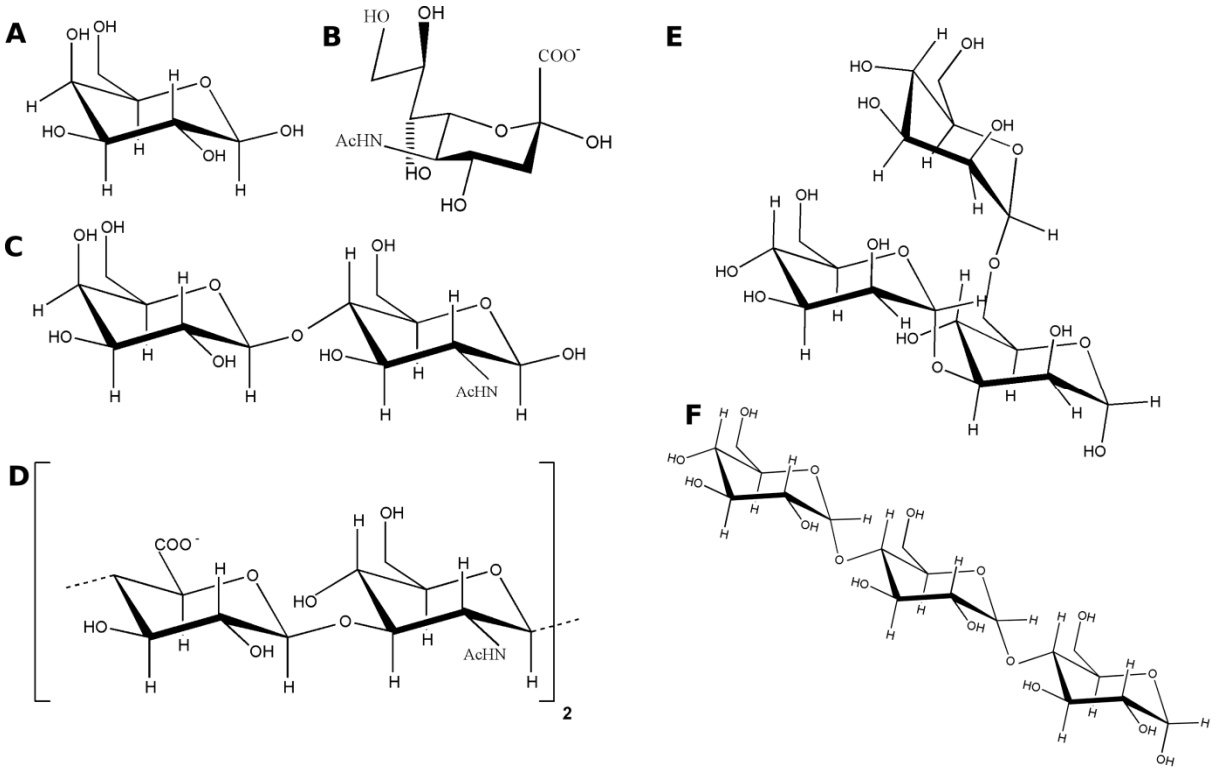
Carbohydrate binding proteins are a large and diverse group of biomolecules that harbors enzymes as well as non catalytic members. Lectins for example, are multivalent carbohydrate-binding proteins present in all living organisms displaying a wide variety of biological activities including cell recognition, communication and cell growth. [39-41] Some of them, like the well known and thoroughly studied galectins, have also recently emerged as key components for the

development of drug targets in several diseases, including cancer. [42] [43-48] In this context, understanding protein-carbohydrate interactions with atomic resolution, (i.e determining the structure of the corresponding complexes) is of fundamental importance for basic and applied glycobiology. [49,15,23,50-58] A common, but usually overlooked feature of carbohydrates, is the fact that their polar -OH groups, quite frequently bind to hydrophilic patches of the protein surface, resulting in significant solvent displacement and reorganization. [29,30,35,36,44,55,59,60] Water molecules and carbohydrate -OH groups can participate in similar hydrogen bond networks when establishing contacts with protein surfaces. This has been recently evidenced and characterized by our group and others, for several carbohydrate binding sites (CBS) of a diverse set of proteins. [35,36,59,60] Therefore, it is expected that the corresponding solvent structure, would result useful for the *in-silico* prediction of protein-carbohydrate complex structures, with higher accuracy than conventional docking methods.

As a pre-requisite for further calculations, we need to provide a simple methodology to analyze and characterize the mentioned solvent structure. In principle, interactions between water molecules and a protein can be thoroughly studied by means of Molecular Dynamics (MD) simulations in an explicit water environment. However, specialized methodologies are required to estimate accurately the structural and thermodynamic properties of the surface bound water molecules. [29,30,35,36] One of the most potent methods for achieving this task is based on the inhomogeneous fluid solvation theory (IFST) as developed by Li and Lazaridis.[29,30] Using this methodology, we were recently able to show that solvent structure and dynamics at protein surfaces involved in carbohydrate binding proteins are very different from those of the bulk solvent, allowing the identification of the so called water sites (WS) or hydration sites. A WS corresponds to a definite 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. A further thermodynamic and structural characterization can be achieved employing the IFST.[35,36,61,62]

In the present work, we used the information provided by the identification and characterization of the WS in the CBS of several carbohydrate binding proteins, to modify the scoring function of the docking program AutoDock4,[19,20,28,63] in order to perform the *in-silico* prediction of their corresponding protein-ligand complex structures. To test the performance of the presented implementation, we chose six protein-carbohydrate complexes with known crystal structures, whose ligands range from the mono to the tetrasaccharide (as depicted in Scheme 1).



Scheme 1: Schematic representation of the carbohydrate ligands used in the present work. Receptor protein–ligand pairs are: A) CBM32-Galactose; B) CBM40-Sialic Acid; C) Gal3-LAcNAc; E) ConA-Trimannoside; D) CD44-Hialuronan tetrasaccharide F) SPD-Maltotriose.

The systems are: The carbohydrate binding domains of a large multimodular sialidase from *Clostridium perfringes*, belonging to the carbohydrate binding modules number 32 and 40, from now on referred as CBM32 and CBM40. Both modules display a β-sandwich fold with a single monosaccharide binding site that binds Galactose (β-D-Galactose) and Sialic Acid (α-D-N-acetylneuraminic acid) respectively (Scheme 1A and B). [64] This type of modules are widely used to engineer lectins with a desired interaction.[65] The well studied human Galectin-3 (Gal-3), which belongs to the widespread family of animal lectins. Within several Gal-3-ligand complexes available, we have chosen the ligand with higher reported affinity, LacNAc (β-D-Galactosyl-1,4-N-Acetyl-D-Glucosamine) (Scheme 1C).[59,66] Concanavalin-A (Con-A), a trimmanoside (Scheme 1E) binding lectin frequently employed for analyzing protein-carbohydrate interactions (specifically, the trisaccharide chosen for this study was 3,6-Di-O-(α-D-Mannopyranosyl)-α-D-Mannopyranoside, usually referred as the trimmanoside).[44,67] The soluble domain of the CD44, a key cell surface receptor implicated in cancer biology and inflammation, that binds hyaluronic acid, a polymer of disaccharides, themselves composed of β-D-Glucuronic acid and β-D-N-Acetylglucosamine linked via alternating β-1,4 and β-1,3 glycosidic bonds. The structure of the CD44:Hialuronan octasaccharide complex was solved showing an octasaccharide ligand for each CD44 monomer, of which four monosaccharides (the hyalunoran tetrasaccharide, β-D-Glucuronyl(1-3)-2-Acetamido-β-D-Glucopyranosyl(1-4)-β-D-Glucuronyl(1-3)-N-Acetyl-D-Glucosamine) are tightly bound to the protein and thus selected as the ligand in the present work (Scheme 1D).[53] The Surfactant Protein D (SPD), an immune effector related to antimicrobial host defense and immune regulation through the recognition of the carbohydrate patterns from several microorganisms or apoptotic cells. SPD has a C-type lectin Carbohydrate Domain, which binds the trisaccharide maltotriose (α-D-Glucopyranosyl(l-4)-α-D-Glucopyranosyl(l-4)-α-D-Glucopyranose) (Scheme 1F). [68]

Our results clearly show that the modified function significantly improves the quality and accuracy of the results, both in terms of how close the predicted complex structure resembles the real one (i.e. obtained by crystallography), and in the differentiation between good and bad predictions. The resulting solvent structure biased docking protocol, thus results in a powerful tool for the design and optimization of glycomimetic drugs, and for the basic understanding of

protein carbohydrate interactions. Moreover, the achieved improvement also underscores the relevance of the solvent structure for the protein carbohydrate recognition process.

2. Results

The results are organized as follows. For each tested case we firstly describe briefly the protein CBS and the characterized WS (a summary of the WSs characteristics used to build the modified grids, is presented in Table 1). Then, we present the results obtained with both the Conventional AutoDock4 docking method (CADM) and the Water Site Biased Docking Method (WSBDM), always using the corresponding complex crystallographic structure as a reference. The cases are organized in the order of increased ligand size (mono- to tetrasaccharide), and for selected cases the results varying key parameters such as the number of WSs used to build the grid, receptor structure, and ligand size are presented and briefly discussed. At the beginning of the discussion, a final comparative summary of all the obtained results is presented, analyzed and discussed.

2.1 *Monosaccharide Docking to CBM32 and CBM40 modules*

We begin our study by comparing the performance of the CADM and WSBDM for docking the lactose monosaccharide into the CBS of the above described CBM32. This CBS harbors four WSs, as characterized in our previous work and shown in Table 1, three of which (WS1, WS3 and WS4) are displaced by lactose upon binding. Figure 1 (left panel) shows the population vs. binding energy plots for lactose docking to CBM32 crystal structure using the CADM. The results show that the method is unable to correctly predict the complex structure, since the lowest energy cluster is very far from the reference structure. The highest population cluster is closer to the reference, but still clearly unacceptable (See Figure 1C). Moreover, the docked configuration

closest to the reference complex predicted structure has a RMSD of 2.9 Å, and may result very difficult to identify among other predictions.

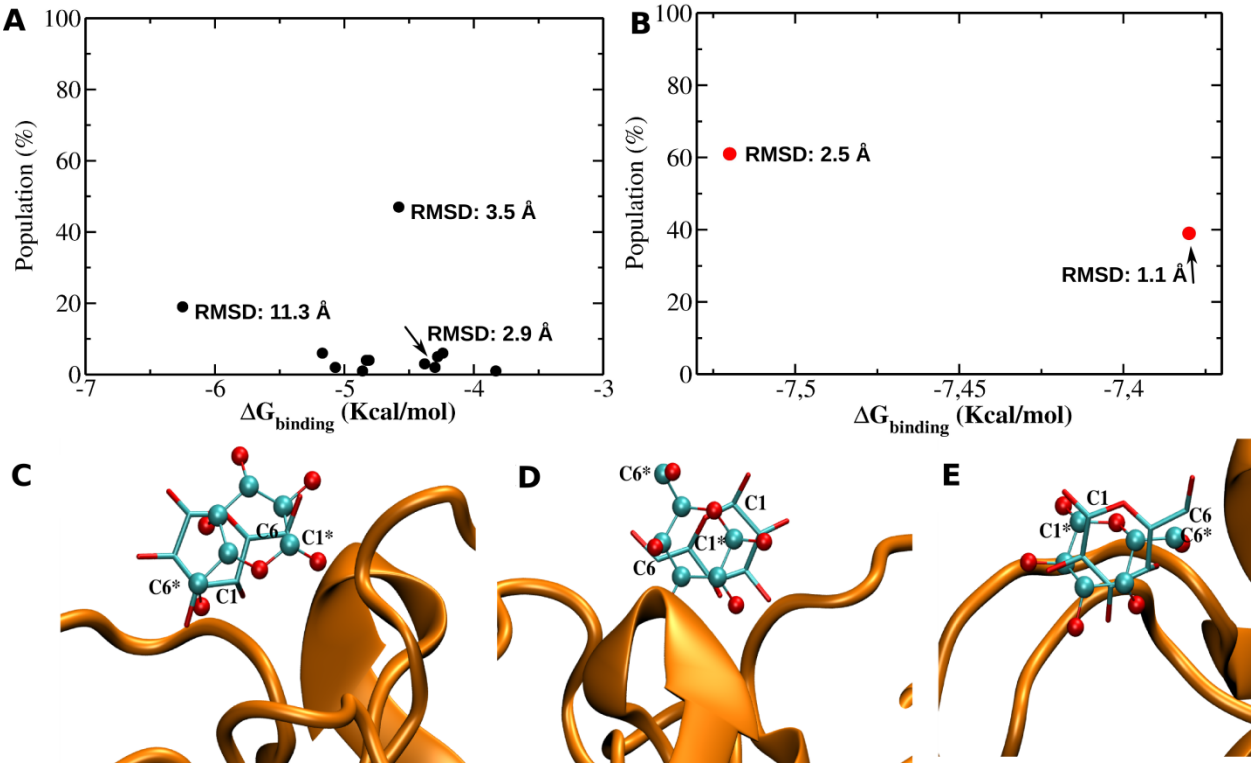


Figure 1. Results for docking of β-D-Galactose **to CBM32 using the complex X-ray structure as the receptor.** Population vs. Binding Energy plots for the Docking of β-D-Galactose to CBM32 using the CADM (Panel A) and the WSBDM (Panel B). Values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure (PDB ID 2v72). In lower panel, we show structures for the predicted CBM32 β-D-Galactose complexes, superimposed onto the reference structure. Panel C shows the structure corresponding to the highest population CADM prediction with an RMSD of 3.5Å. Panel D, shows the lowest energy WSBDM prediction with RMSD of 2.5Å and Panel E shows the second lowest energy WSBDM prediction with an RMSD of 1.1Å. Predictions are shown in

balls and stick representations, while the reference ligand position is shown as sticks. Atom labels marked with an * correspond to the predicted structures.

The results for the WSBDM using the four identified WSs to build the grid are shown in Fig. 1 right panel. The results are clearly better than those described above. Only two clusters are found, which are very close in energy (difference is only 0.2 kcal/mol, i.e. well below the method accuracy) and with similar populations. However, the closest to the reference structure, with only 1.1Å RMSD, and therefore a very good prediction as shown in Fig. 1E, ranks second in both Pop and ΔG_B. The highest Pop predicted structure instead shows the β-D-Galactose slightly displaced and rotated ca. 60 degrees, from the reference structure, as shown in Figure 1D. Similar results are obtained, when the receptor grid was built using a random selected snapshot from the CBM32 ligand free MD simulation (See SI Figure 1). Although for the WSBDM the correct structure still ranks second (and this time with even lower population), it performs significantly better than the CADM.

We now turn our attention to CBM40, presenting six well defined WS in its CBS region, as characterized in our previous work and shown in Table 1. The three sites with highest WFP, WS1, WS3 and WS6 are all three displaced by the ligand O8, C6 and the acid carboxylate respectively. Figure 2 shows the population vs. binding energy plots for Sialic Acid docking to CBM40 crystal structure using the CADM. The results show that the method ranks first the wrong complex structure (both in energy and population, with a RMSD of 4.7 Å compared to the reference), but correctly predicts the complex as the second best choice. The WSBDM, using all six sites to modify the grid, shows better results. The first energy ranking and highest population cluster is now the best predicted complex, with a RMSD of only 0.9 Å. The predicted structure

(shown in SI Fig. S2) shows the sugar ring correctly placed and oriented, with its main side chain only slightly shifted. It should be noted however, that its population is similar to clusters showing wrong predicted structures. As for CBM32 similar results are obtained using random selected snapshot from the ligand free protein MD simulation (data not shown).

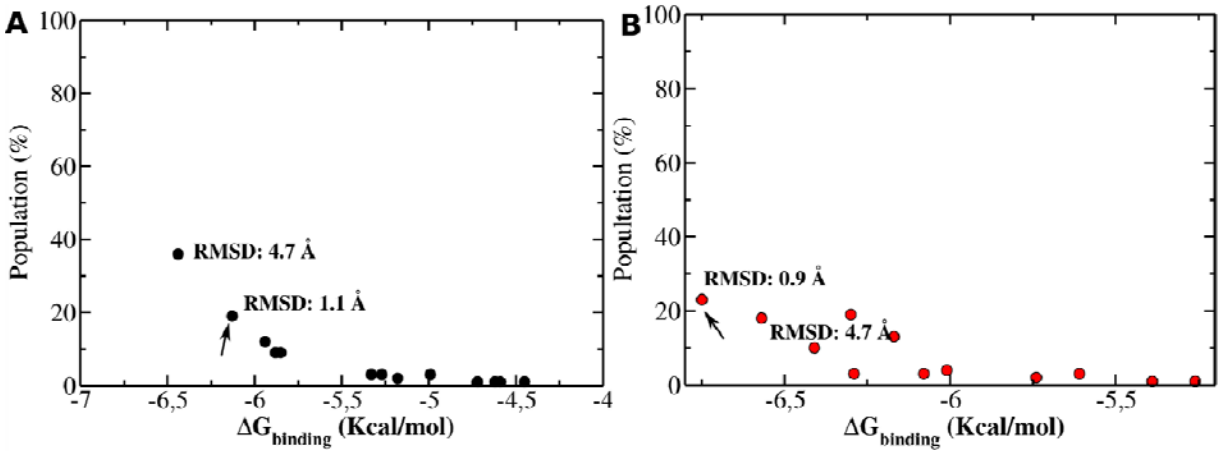


Figure 2. Results for docking of Sialic Acid **to CBM40 using the protein complex X-ray structure.** Population vs. Binding Energy plot for the Docking of Sialic Acid to CBM40 using the CADM (panel A) and using the WSBDM (panel B). Values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure PDB ID 2v73

In summary, the results for monosaccharide docking to the CBM modules show that the WSBDM significantly improves the docking predictions. However, the correct result is still not always clearly standing apart in terms of predicted binding energy and population, compared to wrong predictions (i.e. false positives).

2.2 Galectin-3

Gal-3 displays a CBS that usually hosts a disaccharide, which can be either lactose or N-Acetyl-Lactosamine. As analyzed in our previous work and shown in Table 1, seven clear WS can be identified in the CBS of Gal-3. Three WS are clearly found in the Galactose (Gal) binding site, three in the N-Acetylglucosamine (GlcNAc) binding site, and one more WS is between both. The two WS displaying the highest WFP are clearly replaced each by one hydroxyl group O3 of the GlcNAc, and O6 of Gal respectively. All other WS are also shown to be close to hydroxyl groups of the ligand except for WS7 which is closer to the CH_3 of the acetyl group. In the real case it is very difficult to know which monosaccharide binds where. Therefore, even for a monosaccharide docking the whole CBS and its associated WS could be used and analyzed. Keeping this in mind we will now compare the conventional and biased docking methods in their ability to correctly determine Gal-3-Disaccharide complex. We begin the analysis by individually docking each monosaccharide (Gal and GlcNAc) to the whole CBS of Gal-3.

2.2.1 Galactose docking to Gal-3. The results for docking of Gal to Gal-3, with the CADM and WSBDM, and using the crystal structure to build the corresponding receptor grid, are shown in Figure 3. The data from Figure 3A, shows that using the CADM a low energy high population cluster (46%) is clearly identified (marked as an arrow), having a RMSD 3.3 Å with respect to the reference complex structure. This is the closest prediction to the reference, and thus puts in evidence that the prediction is not very good. Using the WSBDM (Figure 3B), considering all seven WS found in the CBS, the results are much better, since the low energy high population cluster is clearly apart from other predictions, it has a high population (ca. 90%) and very low RMSD against the reference structure. A closer look at the predicted structures (Figure 3B and 3C), shows that for the CADM best prediction the Galactose ring is correctly positioned in the corresponding binding site (and not in that of GlcNAc), but it appears shifted and rotated about

180 degrees; while the WSBDM predicted complex shows a perfect match against the reference. In summary the results clearly show that WSBDM is able to correctly predict the Gal-3:Gal structure, while CADM is not.

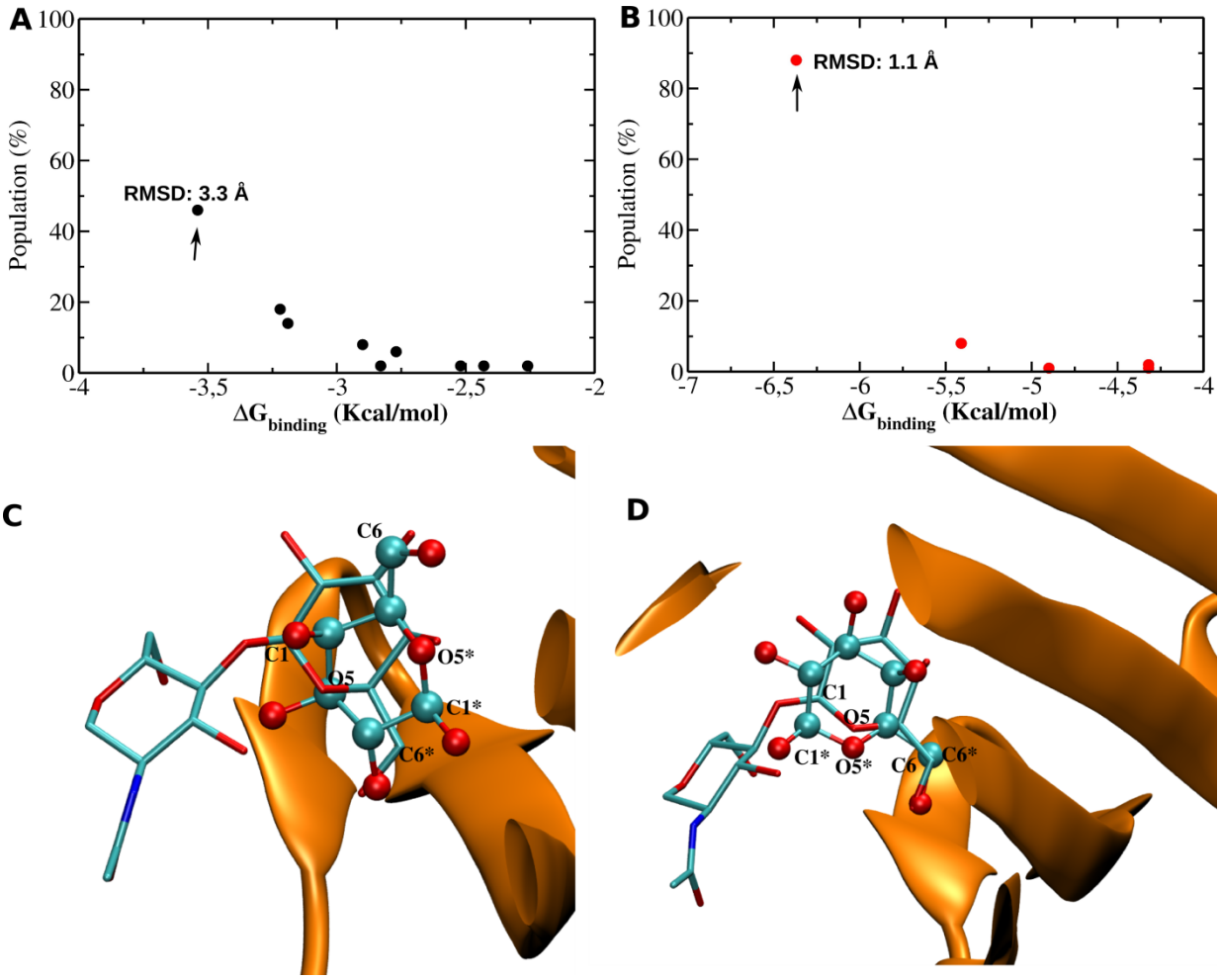


Figure 3. Results for docking of Galactose to Gal-3 using the protein complex X-ray structure as the receptor. Population vs. Energy plot for the Docking of Gal to Gal-3 CBS using the CADM (Panel A) and the WSBDM (Panel B). Values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure, PDB ID 1A3K. Lower panel shows the structures for the predicted Galactose:Gal-3 complexes, superimposed onto the reference structure. Panel C shows the structure

corresponding to the highest population CADM prediction with an RMSD of 3.3Å, Panel D, shows the lowest energy WSBDM prediction with RMSD of 1.1Å. Predicted structures are shown in balls and stick representations, while the reference ligand position is shown as sticks. Atom labels marked with an * correspond to the predicted structures.

2.2.2 N-Acetylglucosamine Docking to Gal-3.

The same calculations as described above for Gal were now performed but using GlcNAc as the ligand. The results (shown in SI Figure S3) for the CADM are not as good as those shown in the case of Gal. There is not a clear low energy high population cluster. The highest population cluster (22%) is close to the lowest energy one and it has an RMSD to the reference structure of ca. 5 Å. While the lowest RMSD cluster, that closest to the correct structure, is not easily recognizable. The results with the WSBDM are slightly better but still not satisfactory. Although in this case two structures stand out, they show RMSDs of 3.8 Å and 5Å with respect to the reference structure. So both methods fail to correctly dock GlcNAc inside Gal-3 CBS. The reason for this failure possibly originates in the fact that GlcNAc tends to be docked closer or even inside the galactose binding site, as evidenced by the RMSD of ca 2.5-2.6 Å observed for the best clusters when the galactose position in the x-ray complex structure is used as the reference. This observation prompted us to test whether the docking of GlcNAc into Gal-3 could be improved, when Gal is already placed in its binding site. Thus we performed the corresponding docking simulation with both methods, and the additional restraint imposed by the GlcNAc-Gal glycosidic bond. The corresponding results, shown in SI figure 7, show that the CADM still fails to place the GlcNAc correctly. The WSBDM performs better, correctly identifying in this case the complex structure (RMSD of 1.0 Å) as the highest population cluster, ranking second in energy (less than 0.5 kcal/mol difference to the best binding cluster).

In summary, although galactose can be reasonable well docked in the Gal MBS and the results slightly improve with the WSBDM including all WS, this is not the case for GlcNAc.We now turn our attention to the results obtained using as ligand the whole N-Acetyl-Lactosamine disaccharide.

2.2.3 Disaccharide (N-Acetyl-Lactosamine) docking to Gal-3

The results for Docking of the disaccharide N-Acetyl-Lactosamine to Gal-3, with the CADM and WSBDM using all seven WS, and using either the crystal structure or a random selected snapshot from the free protein MD simulation to build the corresponding receptor grid, are shown in Figure 4. The results clearly show that the CADM (Panels A, C and E) is unable to dock the disaccharide, even if a re-docking is performed (i.e. if the grid is built using the protein structure from the complex crystal). No clear cluster stands out, and the highest population or lowest energy clusters show an RMSD of more than 5Å compared to the reference system. The superimposed structure (Figure 4E) shows that the CADM predicted complex is shifted placing the N-Acetylglucosamine over the galactose binding site. On the other hand, the WSBDM is clearly able to correctly fit the ligand in place in both cases (Panels B and D). Either with the crystal structure, or a random selected snapshot taken from the free protein MD simulation, a complex structure stands out in the population vs. energy plot, showing a ligand heavy atom RMSD against the reference structure of less than 1Å. The corresponding predicted structure superimposed on the reference, shown in Figure 4F, is striking because of its perfect fit. The WSBDM places both sugar rings correctly, and even the N-Acetyl side chain is correctly oriented.

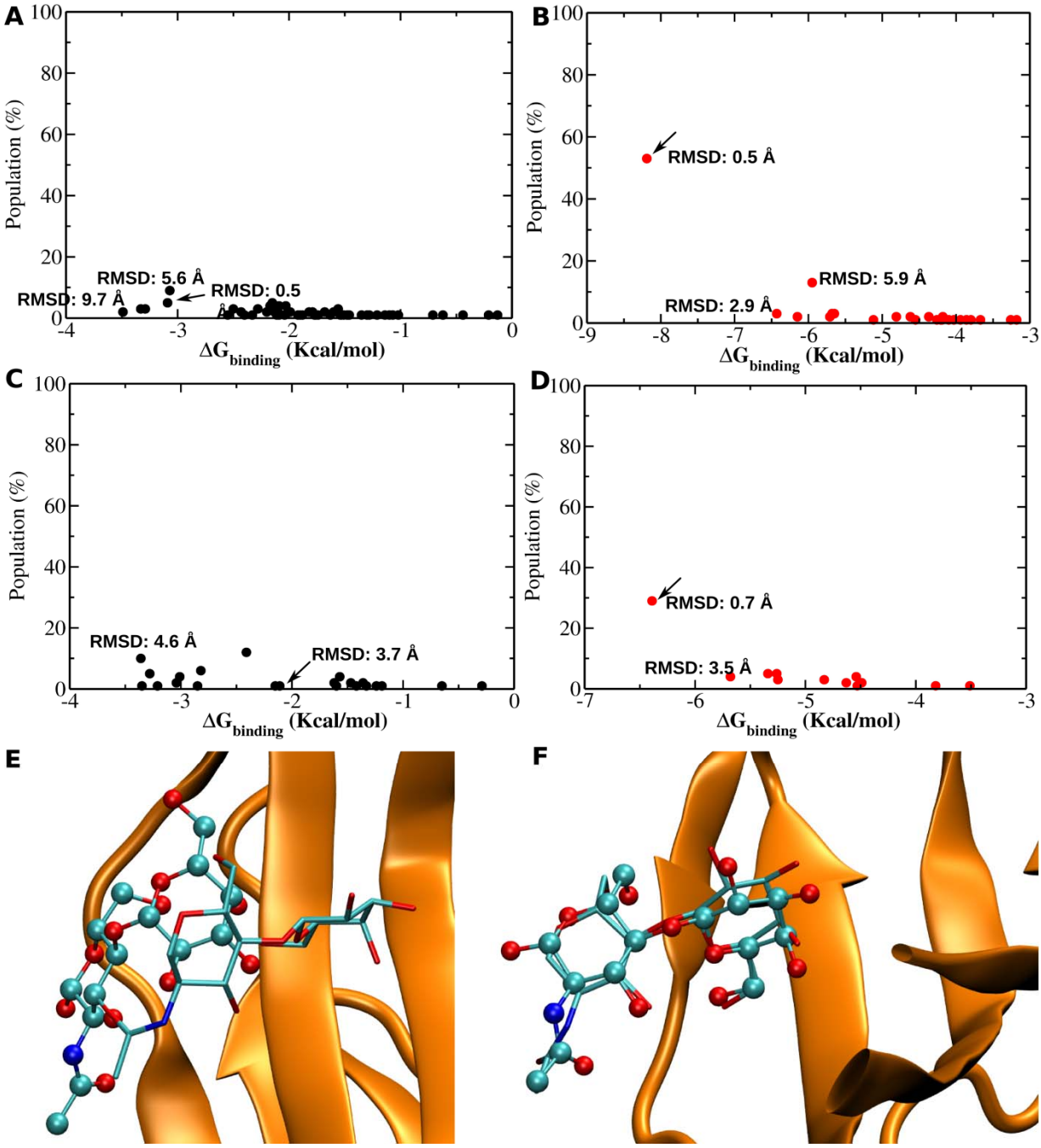


Figure 4. Results for docking of N-Acetyl-Lactosamine to Gal-3. Population vs. Binding Energy plots for the docking of N-Acetyl-Lactosamine to Gal-3 using the CADM to either the complex X-ray structure (Panel A) or a random selected snapshot from the ligand free MD simulation (Panel C); and using the WSBDM to either the complex X-ray structure (Panel B) or a random selected snapshot from the ligand free MD simulation (Panel D). Values next to the

dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure (PDB ID 1A3K). Lower panel shows the structures for the predicted N-Acetyl-Lactosamine-Gal-3 complexes, superimposed onto the reference structure. Panel E shows the structure corresponding to the first ranking CADM prediction with an RMSD of 4.6Å, Panel F, shows the lowest energy WSBDM prediction with RMSD of 0.5Å. Predictions are shown in balls and stick representations, while the reference ligand position is shown as sticks.

The excellent performance of the WSBDM using all seven WS shown above, prompted us to analyze the impact on the predicting capability of the method in relation to the number of WS chosen to be included to build the modified grid. Based on our previous work, where it is shown that the replaced water sites are usually those with highest WFP and smaller R_{90},[35] a clear rationale emerges for selecting the WS to be used in the WSBDM. Figure 5 shows the results for docking N-Acetyl-Lactosamine to Gal-3 CBS, including in the WSBDM either, one (corresponding to either WS4 or WS2 the two highest WFP regions), two (WS2 and WS4) or three WS, corresponding to WS1, WS2 and WS4.

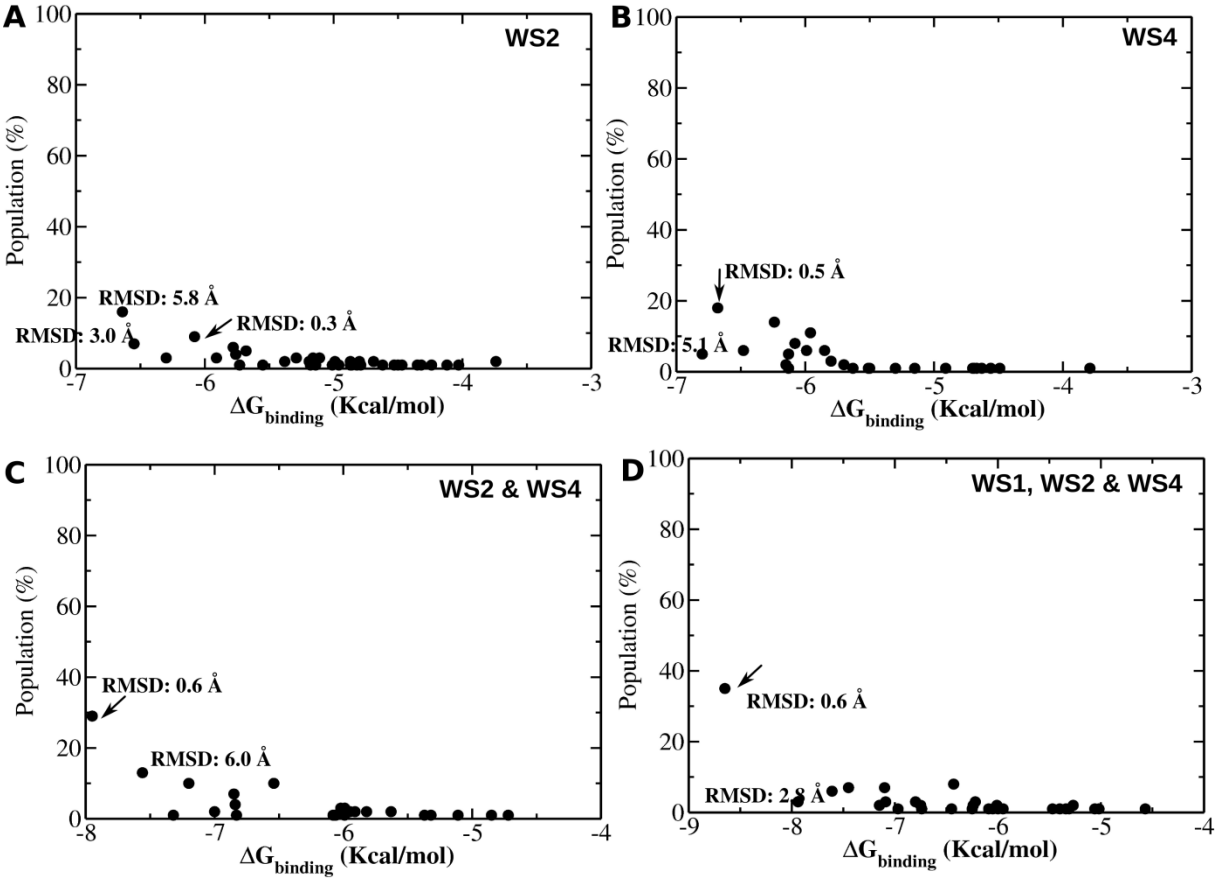


Figure 5. **Results for docking of N-Acetyl-Lactosamine to Gal-3 using different number of WSs to build the receptor grid.** Population vs. Energy plot for the Dockingd of N-Acetyl-Lactosamine to Gal-3 CBS using the WSBDM. In panel A, the WSBDM grid was built using just WS2. In Panel B, the WSBDM grid was built using just WS4. In panel C, the WSBDM grid was built using WS2 and WS4. In panel D, the WSBDM grid was built using WS1, WS2 and WS4. Values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the complex reference structure PDB ID 1A3K.

The results show that when only one site is used the results are highly dependent on which WS is used. For example using only WS4 the highest population cluster (ranking second in terms of energy) is able to correctly predict the complex structure (the RMSD with respect to the reference structure is only 0.5Å). On the other hand when only WS2 is used the highest

population and lowest energy cluster does not correctly predict the complex structure. The correct structure appears as the second highest population and with higher energy than several other structures. Using both WS the results improve, with a clear complex standing out (30% population and the lowest energy) and an RMSD with respect to the reference structure of only 0.6 Å. Finally, using three WS the results are slightly, but not significantly better. Therefore, it seems that at least one WS for each monosaccharide is needed, in order to allow significant improvement of the prediction. In summary for the present case the WSBDM significantly improves the prediction quality, and docking of the disaccharide seems to be a better strategy than docking of each monosaccharide independently.

2.3 Trimannoside Docking to Concanavaline-A

Concanavaline-A (ConA) is a thoroughly studied lectin that binds a trimannoside ligand.[67,44]-As described in our previous work,[35] and shown in Table 1, eleven WS with high WFP can be identified in the ligand binding site. Three WS, namely WS7, WS8 and WS9 are replaced by the first mannose O5, O6 and O3, which also establish a total four strong HB with the protein. The second mannose O2 clearly displaces WS1, and O4 possibly displaces WS11. Finally, the third mannose O3 must displace WS5, while O4 must displace WS6. Based on the previous results for disaccharide docking to Gal-3 we decided to dock directly the trimannoside. The results for CADM and WSBDM for docking of the 3,6-di-O-(α-D-Mannopyranosyl)-α-D-Mannopyranoside (trimannoside) into ConA are shown in Figure 6.

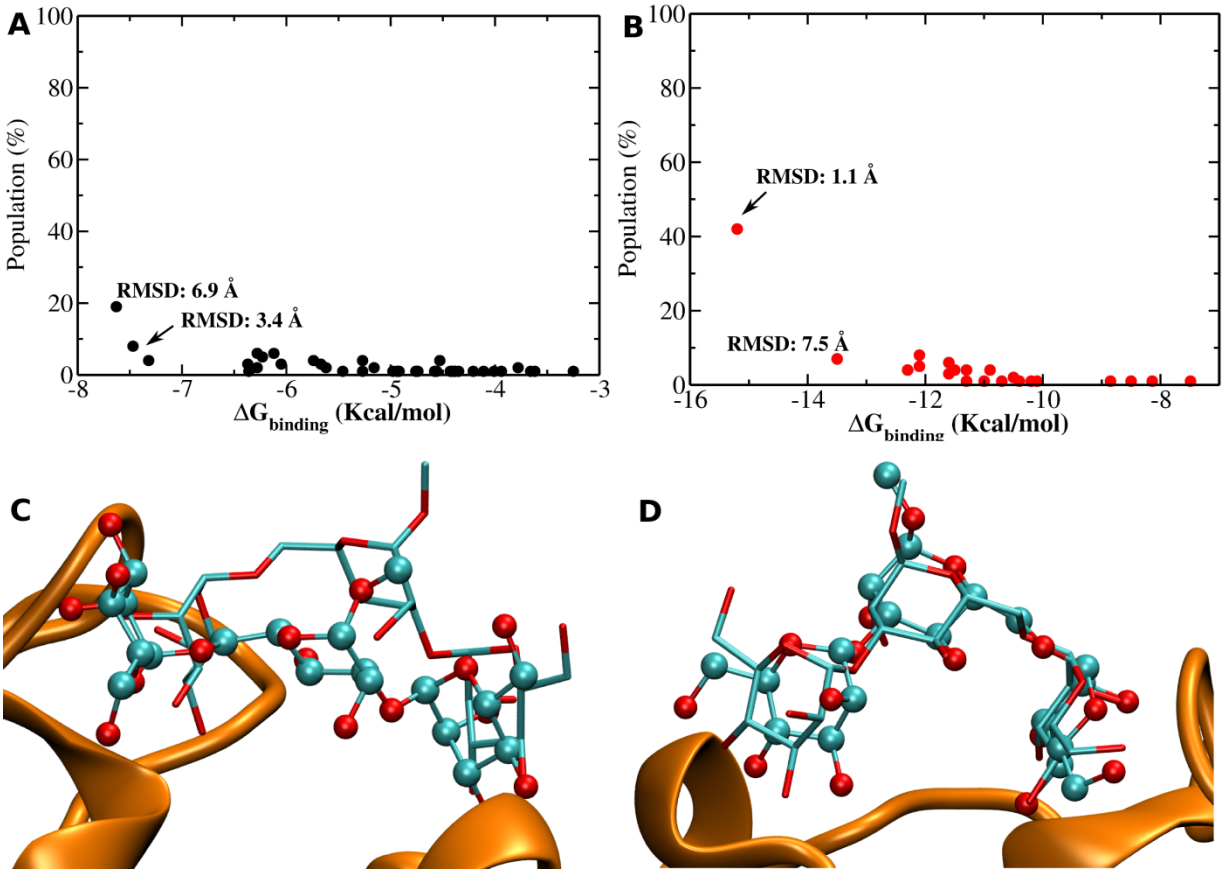


Figure 6 Results for docking of the Trimannoside to ConA**.** Population vs. Binding Energy plots for the docking of the trimannoside to ConA using the CADM (Panel A) and the WSBDM (Panel B). Values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure (PDB ID 1ONA). Structures for the predicted trimannoside-ConA complexes, superimposed onto the reference structure. Panel C shows the structure corresponding to best (in terms of RMSD) CADM prediction with an RMSD of 3.4Å, Panel D, shows best ranking WSBDM prediction with RMSD of 1.1Å. Predictions are shown in balls and stick representations, while the reference ligand position is shown as sticks.

The results show that with the CADM the highest population-lowest energy cluster does not yield the correct result. The RMSD to the reference complex is 6.9 Å. Only the second population ranked cluster, which also displays low binding energy values, predicts a close to correct structure, with an RMSD of ca. 3.4 Å. This structure, shown in Fig 6C, shows the trimannoside considerable shifted and closer to the protein, although it is correctly oriented (each monosaccharide is closest to its binding site). Again, the results are significantly improved with the WSBDM, the highest scoring lowest energy cluster is clearly separated from other results and displays an RMSD of only 1.1 Å with respect to the reference structure. As shown in Figure 6D the structure is very similar to the crystallographic one, with two mannoses perfectly in place and the third only slightly shifted.

As for Gal-3, we performed the same calculations using a random selected snapshot from the MD simulation, instead of the crystallographic structure to build the grid. The results shown in SI figure S4, similarly show that the CADM is unable to correctly predict the complex structure. The results for the WSBDM are better, but not as good as those obtained with the crystal structure. Instead of a clear best prediction, three cases stand out, two with lowest energy having fairly low RMSDs of roughly 2Å. Visual inspection of the predicted structures show that the trimannoside is nonetheless very well positioned, and probably the slightly higher RMSD (compared to the results obtained with the crystallographic structure) has partial contributions from the protein motions during the MD, that do not allow a perfect alignment between crystal structure and the MD selected snapshot, since they display a backbone RMSD of 2.1 Å.

Again as for Gal-3, we used Concanavilne A to analyze the relation between the number of WS that are used to build the WSBDM grid and the accuracy of the results obtained. It is important to choose the WS in a straightforward way that is not biased by our knowledge on the structure

of the complex. The first choice, as previously mentioned, should fall in those WS having the highest WFP and lowest R_{90}. However, for large CBS that bind tri and tetrasaccharides, selecting WS that cover the whole CBS seems also a good choice. The data from our previous work and Table 1 shows that the five WS with high WFP (ca 10 times that of the bulk or more) are located in two groups, each at one extreme of the CBS. The first group harbors WS7, WS8 and WS9, while the other WS3 and WS6. Thus we performed trimmanoside docking calculations using the WSBDM, building the grid either with one WS (selecting the best WS of each group, i.e. WS8 or WS3), with two WS (combining both of them), with three WS (selecting the best WS of each group, WS3, WS8 and WS1) and also with the five highest WS. The results are shown in Figure 7.

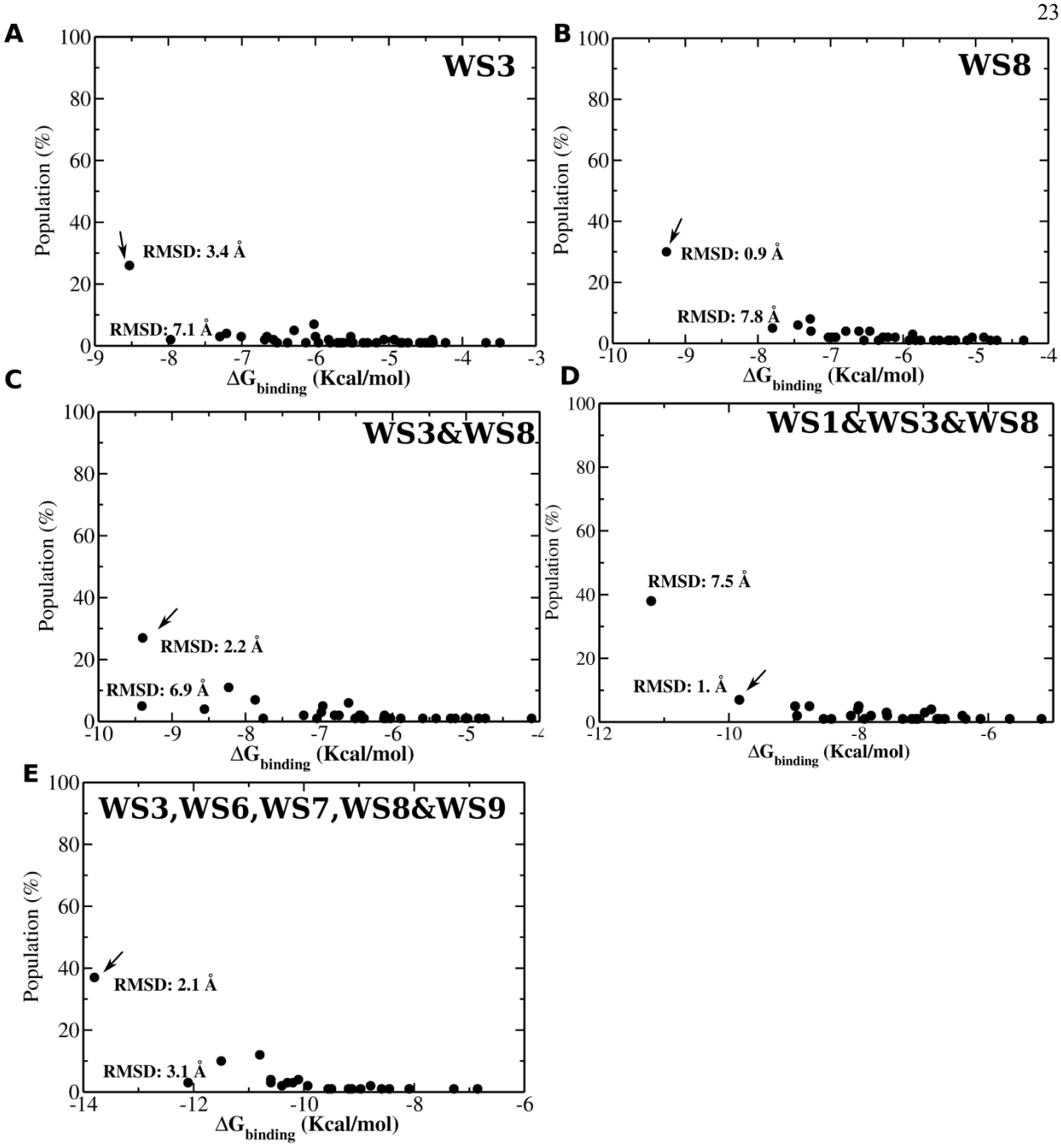


Figure 7 Results for docking of the Trimannoside to ConA using different number of WS. Population vs. Binding Energy plots for the docking of the trimannoside to ConA using the WSBDM. Panel A the grid was built using solely WS3. In panel B the grid was built using only WS8. In panel C the grid was built using WS3 and WS8 and in panel D the grid was built using WS1, WS3 and WS8. Finally, in panel E the grid was built using WS3, WS6, WS7, WS8, and

WS9. Values next to the dots represent the ligand heavy atom RMSD between the predicted complex structure and the reference complex structure (PDB ID 1ONA).

The results show that when only one WS is used the method still could be able to correctly predict the complex structure, but that the results are strongly dependent on which WS is used. If only WS8 is used the best ranking complex displays a very low RMSD of 0.9Å against the reference, but if only WS3 is used, the final results are not satisfactory. The results with two and three WS are clearly better than those obtained with CADM, but when using two WS the best complex has higher RMSD against the reference, compared to the case where only WS8 was used to build the grid. With three WS, although the best binding energy result is wrong, the second ranking cluster corresponds to the correct complex. Finally, when using the five highest scoring WS the results are very similar (even better in terms of RMSD) to those obtained with all the WS, with the best ranking prediction clearly standing out and a very low RMSD against the reference of 2.1Å. Altogether these results suggest when only few WS are used, the predictions vary a lot. This is not unexpected since not all WS are replaced by OH groups from the ligand. Thus it seems to be a better choice to use many WS and at least one for each monosaccharide.

2.4 Maltotriose docking to SPD

As another test case of trisaccharide binding we selected the SPD protein. This protein has a C-type lectin carbohydrate binding domain whose complex structure with its ligand maltotriose has been structurally characterized,[68] but where no previous information on the solvent structure in relation to the ligand is available. MD simulations in explicit water of the uncomplexed protein allowed identification of five WS in the CBS (Table 1), two of them showing very high WFP. The results for docking of maltotriose to the SPD CBS using the CADM and WSBDM (Shown as SI Figure S5) show that the CADM lowest energy structure has an RMSD of 4.1Å against the

reference structure, with only one of the three monosaccharides (number 3) correctly placed. Very close in energy and population to this structure, is a second predicted complex which is closer to the correct structure (RMSD of 1.8Å), which places all three monomers correctly. The WSBDM on the other hand clearly ranks the closest to the reference structure (with an RMSD of 1.8Å) first, and with significant lower energy and higher population than other predictions. Visual inspection of the predicted complex structure in relation to the reference clearly show that all three monosaccharides are correctly positioned and oriented in their respective binding sites, with the third displaying a perfect match and the first and second, slightly shifted. Clearly the WSBDM allows the correct prediction of the complex structure even when no detailed knowledge and analysis of the WS is performed.

2.5 Tetrasaccharide docking to CD44

As the final test case we selected CD44, a cell surface receptor that binds hyaluronic acid. The crystal structure of the corresponding complex shows the carbohydrate recognition domain of CD44 bound to an octasaccharide. However, only four monosaccharides are in contact with the protein, and where thus used for the docking calculations, resulting in the following two repeating disaccharide subunit as ligand: β-D-Glucuronic acid and β-D-N-Acetylglucosamine, linked via alternating β-1,4 and β-1,3 glycosidic bonds. Moreover, since both the complex structure and ligand free (apo) CBS structures are available, we tested the variability of the results using these two and several random selected snapshots from the apo protein MD simulation to build the grids. To perform the WSBDM calculations, as for the previous case, we firstly determined the solvent structure adjacent to the ligand binding site. The MD simulation of the apo protein in explicit solvent shows that the CBS harbors five WS (as shown in Table 1), two with very high WFP and two with medium WFP (ca. 5 times that of the bulk). The results

for docking of the hyaluronan tetrasaccharide on CD44:HA_8 complex crystal structure, CD44 apo protein crystal structure, and five random selected snapshot of CD44 apo protein MD simulation, using the CADM and the WSBDM (using the five identified WS) are shown in Figure 8 and SI Figure 6.

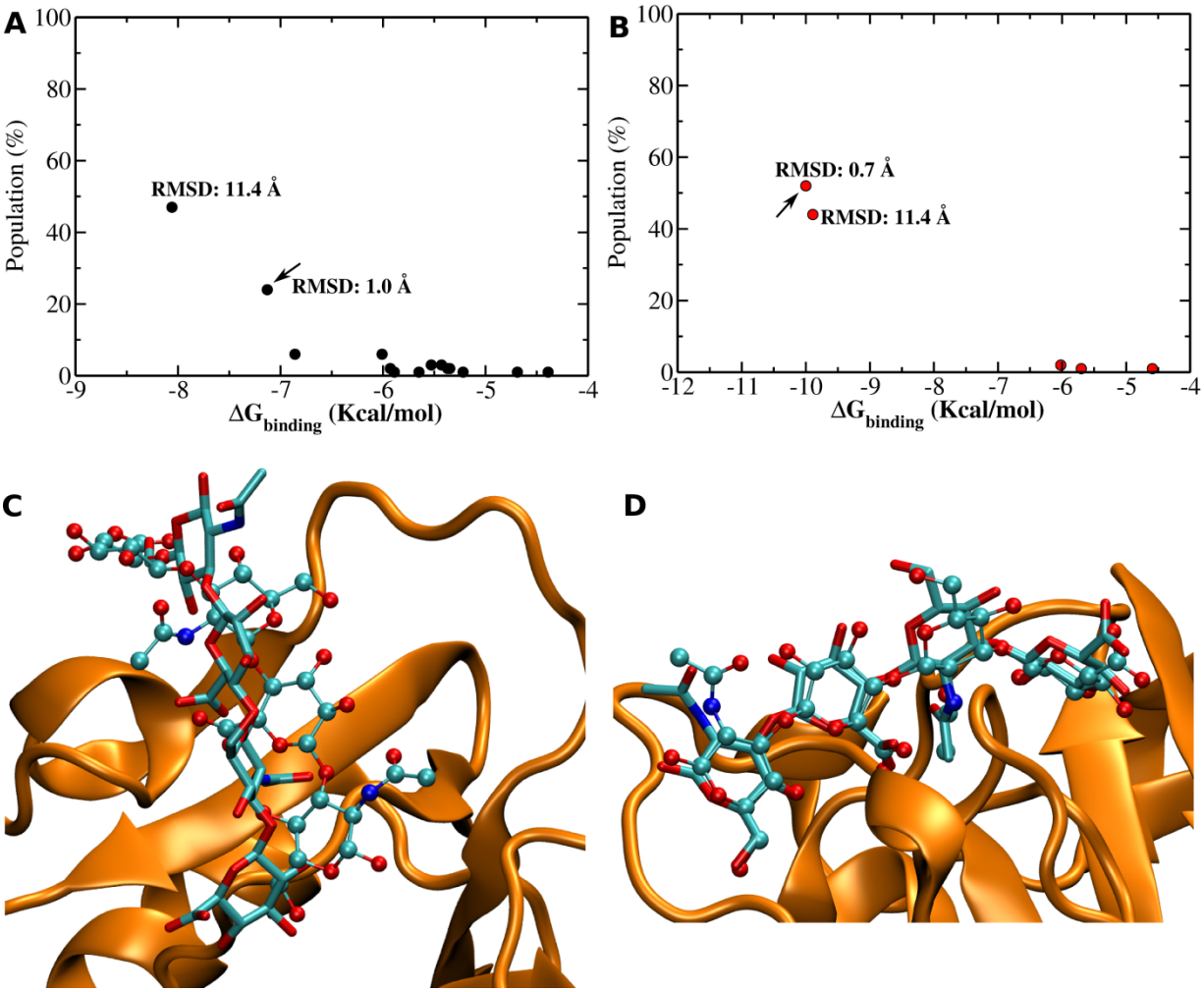


Figure 8. Results for docking of hyaluronan tetrasaccharide to CD44. Population vs. Energy plot for the Docking of the hyaluronan tetrasaccharide to CD44 using the CADM (panel A) and the WSBDM (panel B) Values in parenthesis represent the ligand heavy atom RMSD between the predicted complex structure and the complex reference structure PDB ID 2JCP. The lower panel shows the structures for the predicted hyaluronan tetrasacharide complexes, superimposed

onto the reference structure. Panel C shows the structure corresponding to the bnest ranking CADM prediction with an RMSD of 11.4 Å, Panel D, shows the best ranking WSBDM prediction with RMSD of 0.7 Å. Predictions are shown in balls and stick representations, while the reference ligand position is shown as sticks.

The results for the CADM show that there is a moderate variability in the quality of the results depending on which structure is used to perform the docking. Re-docking in the complex structure correctly allows identification of the correct complex (RMSD of only 1.0Å) with the lowest energy and highest population (ca. 40%), also docking to some of the MD snapshots allows prediction of the correct structure. Interestingly, for the apo structure the lowest energy complex shows the ligand upside down (see Figure 8C) and an RMSD of 11 Å, the correct structure coming second. Results for other MD snapshots gave results that were in the range as those presented, and are thus not explicitly shown. The WSBDM performs significantly better in all cases, yielding higher populations for the best ranked structure, and lower RMSDs values against the reference structure. Even for the docking of the apo structure the WSBDM ranks the correct structure first, although the upside down structure is second, and still displays high population. The best ranked structure of the WSBDM as shown in Figure 8D has an almost perfect match with the reference, for the whole tetrasaccharide.

In summary the results for ConA, SPD and CD44 show that the WSBDM is able to correctly predict and clearly identify the protein-carbohydrate complexes using as ligands three and even tetrasaccharides as ligands, a challenging task given the ligand size and flexibility. The CADM on the other hand is not always able to predict the correct complex, and best structure is usually mixed with false positives. Concerning the use of different structures, the results show that the

CADM presents more variability in the results (when compared to the biased method), and their quality strongly depends on the receptor structure used to build the grid. WSBDM results are more homogeneous (i.e less dependent on the selected receptor structure), especially concerning the best ranking complex.

3. Discussion

The aim of the present work was to analyze whether the information of the solvent structure adjacent to the ligand binding sites of carbohydrate binding proteins, as derived from explicit water Molecular Dynamics simulation and described by the identification and characterization of the Waters Sites, could be used to improve the performance of molecular docking methods for the prediction of protein-carbohydrate complex structures. For this sake, we compared the conventional docking method, with the presently developed and presented WS biased method, in their capacity to correctly predict the complex structures of five different known protein-carbohydrate structures, with ligands ranging from the mono to the tetrasaccharide. Overall we performed over thirty different docking calculations varying the size of the ligand, the receptor structure and the number of WS used to bias the scoring function, an overall summary and analysis of these results is presented below.

Overall analysis of the results. Table S1 shows a summary of the results obtained for both the CADM and WSBDM. For each docking calculation (characterized by the method, receptor, ligand structure and WS used) the following parameters are shown: i) the RMSD to the reference (i.e. crystal structure) of the best ranked (lowest energy) complex, together with its binding energy and population: and ii) The predicted complex with the lowest RMSD to the reference

structure, together with its ranking, binding energy, and population. To analyze the results we performed the following analysis shown in Figure 9A. First, we plotted the computed RMSD against the reference complex for the highest ranked (i.e. lowest energy) prediction, as predicted with the CADM (Black Columns) and WSBDM (Red column). Second, we plotted the RSMD for the prediction with the lowest RSMD among all predicted clusters, with CADM (Grey columns) and WSBDM (orange column).

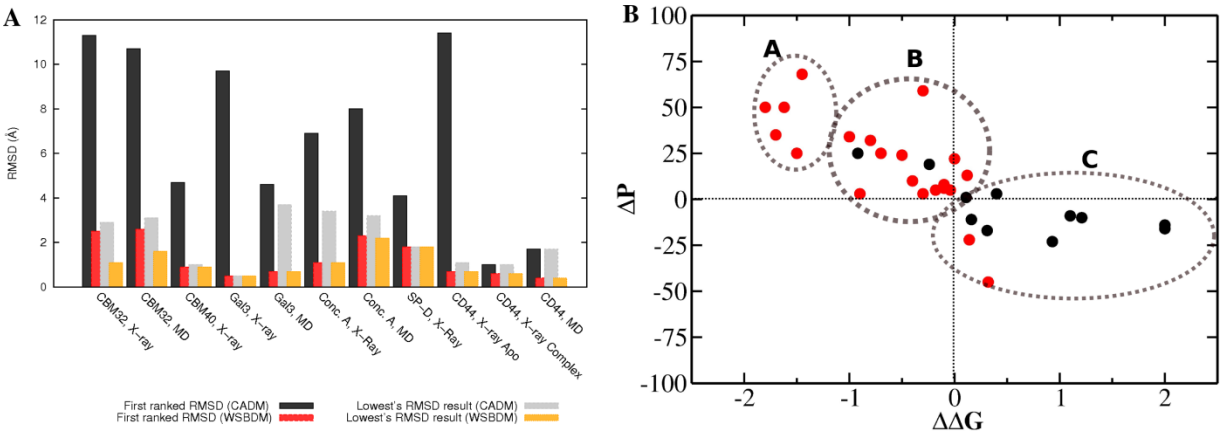


Figure 9. Panel A. For the docking calculations shown in Table S1, Black column shows the RMSD of the highest ranked (i.e. lowest energy) prediction, and grey column the lowest RMSD among all predicted clusters obtained with the CADM. Red column shows the RMSD of the highest ranked (i.e. lowest energy) prediction, and orange column the lowest RMSD among all predicted clusters obtained with the WSBDM. Panel B. $\Delta\Delta G_B$ vs. ΔPop plot for the docking calculations performed with the CADM (Black dots) and WSBDM (Red dots). For the definition of $\Delta\Delta G_B$ and ΔPop see text.

Results from figure 9A show that there is a clear and significant difference in the predictive power of both methods. While the CADM ranks first (Black columns), almost always predicted structures that are completely wrong (with RMSD above 4Å), the WSBDM first ranked

predicted complexes (red columns) are close (between 2-3Å) or even very close (less than 1 Å) to the reference structure. Moreover, for most cases in the WSBDM the complex that has the lowest RMSD value with respect to the crystallographic structure used as the reference (i.e. the best predicted complex) usually ranked first (compare red and orange columns). Interestingly, comparison between the highest ranked and best predicted structure for the CADM shows big differences, but the best predictions are in many cases close to those obtained with the WSBDM. Thus the main difficult of AutoDock4 seems to be not the capacity for predicting the correct protein-ligand complex structure, but correctly ranking different possibilities according to the predicted binding energy, a fact that was also observed in others works.[15,23,57] [13] This observation is consistent with the general result of the present work, which shows that by modifying the AutoDock4 scoring function that determines the binding free energy, significant improvement of the results can be achieved.

As a final analysis of the method predictive capacity in relation with its precision, we measured the difference 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. Thus a negative $\Delta\Delta G_B$ value, implies that best obtained 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 false positive. On the other hand, a positive $\Delta\Delta G_B$ means that the best complex has less binding energy than the first ranking complex, i.e the best prediction is wrong. Similarly, a positive ΔPop means that the best prediction has higher population compared to any other prediction, while a negative value in ΔPop means that best prediction has a smaller population compared to wrong predictions. The results located in the upper-left corner of the plot, correspond to those cases where the best

obtained complex is correctly ranked (has the lowest predicted ΔG_B and highest Pop), and since as previously shown the method usually is able to correctly dock the ligand inside the CBS (RMSD of less than 1 Å to the reference structure), they correspond to successful calculations, in the sense that they would have yielded a correct prediction of the corresponding protein-carbohydrate complex. The resulting $\Delta\Delta G_B$ vs. ΔPop plot is shown in Figure 9B.

A first glimpse on the plot undoubtedly shows that the WSBDM performs significantly better than the CADM, with almost all results falling in the upper left corner (groups A and B). Only two WSBDM calculations fall in the lower right corner (group C), corresponding to the discussed case of CBM32 using either the X-ray or an MD simulation structure to build the grid. More detailed analysis of the WSBDM results, allows identification of a first group of results (group A) where the correct results are predicted to have at least 1 kcal/mol $\Delta\Delta G_B$ and over 25% ΔPop, thus clearly standing out against wrong predictions. These results correspond to all re-docking calculations, i.e. using the receptor structure to build the grid taken from complex X-ray structure (except the monosaccharide binding modules). Group B harbors most of the WSBDM results obtained when a structure taken from the MD simulation is used to build the grid, a result which is not unexpected. For this cases the difference in binding energy and population are smaller. The only results obtained with the CADM that fall in this group are those for CD44.

Effect of the ligand size.

When comparing altogether the results in relation to the ligand size, correctly predicting monosaccharide binding seems to be more difficult than larger ligands. This is evidenced in

CBM32 where even the WSBDM fails to rank the correct structure first, and in Gal-3 where docking of the disaccharide is clearly a better option than docking each monosaccharide separately. Also important, the results for docking of the tri and tetrasaccharides are fairly accurate, with the predicted complexes having low RMSD against the reference and with the right complex always ranking first. Thus docking of large sugars seems to be a better idea than docking of individual monosaccharides. It should be noted however, this might not be always the case, especially as sugars become increasingly larger. The AutoDock4 genetic algorithm that explores possible ligand configurations starts to fail when an increasing number of torsional degrees of freedom are included. Considering that each additional monosaccharide adds at least two more torsional degrees of freedoms, for tetrasacharides or even larger ligands docking separately smaller parts of the sugar (di or tri-saccharides) would be a better idea than docking the whole polysaccharide.

Effect of the number and characteristic of the considered WS. Considering the number of WS the results show that the use of all identified WS seems to be a better choice than using only a few. And although it is possible to correctly predict the complex structure with only one WS, the results vary depending on both the system and the WS choice. However, in order to determine how many WS should be determined, characterized and used to bias the docking method for a particular case, a simple rule of thumb could be to use one or two WS per ligand monosaccharide subunit. However, it is important to select them distributed along the whole CBS. Finally, it should be noted that although selecting those WS having higher WFP seems a reasonable choice, the inclusion of other WS with lower WFP does not significantly affect the predictions. This is not unexpected since the modified functions already scales the bias according to the WFP.

How should we use the WSBDM? Although the results presented in the present work, are focused on the performance of the modified docking protocol, using the solvent structure information (i.e the definition and characteristics of the WS) derived from our previous work, [35,36] except for the SPD and CD44 cases. We will briefly describe how to apply the WSBDM to a particular problem starting from the separate structures of the receptor and ligand. The protocol has 4 steps. i) determining the WS, ii) selecting the WS to built the receptor grid, iii) perform the WSBDM iv) analyze the data. Scripts and programs to perform these tasks are freely available under request.

i) First, the receptor protein structure should be subjected to explicit water MD simulations during 20 to 40ns. From this simulation, WS adjacent to the proteins CBS should be determined and characterized using previously described protocol.[35,36]

ii) WS should be analyzed and those having significant WFP (usually higher than 5 times that of bulk solvent) should be selected. The selected WS should be well distributed along the whole CBS. The number of WS should be in the range of 1 to 2 per monosaccharide of the ligand. Grids should be built onto different receptor structures taken from the simulation, and using if possible different number of WS.

iii) For each grid, 100-200 individual docking runs should be performed and clustered, as described in methods.

iv) To analyze the data, binding energy vs. population plots should be built looking for clusters clearly standing out, thus having significantly lower energy and higher population than other results ($\Delta\Delta G_B$ > 1 kcal/mol and ΔPop > 25%). If no cluster stands out, a different snapshot or number of WS should be tried to build the grid. A trustable complex, should appear best (or

highly) ranked using several snapshots, and once found, it should remain top scoring with an increasing number WS used to build the modified grid.

As a final remark, it is important to discuss the computational time required to use the WSBDM compared to the CADM. Once the grid is built, performing the docking calculations itself takes the same amount of time to utilize both methods. However, while building the grid with the conventional method requires only having a structure for the receptor, to build the WS biased grid, prior explicit water MD simulation of the receptor protein needs to be performed and analyzed. However, the computational time required to perform MD simulation is not extensive, requiring for the medium size proteins Gal-3, CBM30 or ConA, ca. 8 hs for each nanosecond on an 8 core cpu machine. Thus using 32 cores, where the amber code has been shown to scale linearly, it takes less than 1 week to perform over 50ns MD simulation.

In summary, analysis of the solvent structure adjacent to the binding sites of carbohydrate binding proteins, allows the identification and characterization of specific regions of space, called water sites, with significantly higher probability of finding a water molecule when compared to the bulk solvent. This information was used to modify the AutoDock4 scoring function, favoring those ligand conformations where the carbohydrate-OH groups match the position of the water site, resulting in the development of a Water Sites Biased Docking Method. The method is able to correctly predict the complex structures of several protein-carbohydrate complexes, with ligand ranging from the mono to tetrasaccharide. Altogether, the method performance shows that it significantly outperforms the non modified AutoDock4 in both 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. The resulting solvent structure biased docking protocol thus results in a powerful

tool for the design and optimization of glycomimetic drugs development, and for the basic understanding of protein carbohydrate complexes and their interactions. Moreover, the achieved improvement also underscores the relevance of the solvent structure for the protein carbohydrate recognition process.

4. Computational Methods

4.1 Set up of the systems and Molecular Dynamics parameters

Protein coordinates were retrieved from the Protein Data Bank, corresponding codes are: 1ONA for Con-A,[67] 1A3K for Gal-3,[69] 2JCP for CD-44,[53] 2GGU for SDP[68] and 2V73 and 2V72 for the modules CBM40 and CBM32,[64] respectively. For each system, only one monomer corresponding to the carbohydrate recognition domain harboring the CBS without the carbohydrate ligand was simulated, in order to determine the solvent structure adjacent to the CBS. No MD simulations of the protein-carbohydrate complexes were performed. Standard protonation states were assigned to titratable residues (Asp and Glu are negatively charged; Lys and Arg are positively charged). Histidine protonation was assigned favoring formation of hydrogen bond in the crystal structure. Each protein was then solvated by a truncated octahedral box of TIP3P waters, ensuring that the distance between the biomolecule surface and the box limit was at least 10Å. Each system was first optimized using a conjugate gradient algorithm for 2000 steps, followed by 200 ps. long constant volume MD equilibration during which the temperature of the system was slowly raised from 0 to 300 K. The heating was followed by a 200 ps. long constant temperature and constant pressure MD simulation to equilibrate the system density. During these temperature and density equilibration processes, the protein backbone atoms were restrained by 1 kcal/molÅ force constant using an harmonic potential centered at each atom starting position. No restraints were applied during the following production

simulations. For small mono and disaccharide binding proteins (CB32, CBM40 and Gal-3) 20ns long production MD simulations were performed, while 50ns long production MD simulations where performed for the systems harboring larger ligands (ConA, SPD and CD44). All simulations were performed with the amber package[70] of programs using the ff99SB force field[71] for all aminoacidic residues. No ligands were included in the MD simulations. Pressure and temperature were kept constant using the Berendesen barostat and thermostat[72] respectively, using the Amber default coupling parameters. All simulations were performed with periodic boundary conditions using the particle mesh Ewald (PME) summation method for long range electrostatic interactions. The SHAKE algorithm was applied to all hydrogen-containing bonds, allowing the use of a 2 fs. time step. These explicit water MD simulations were used to define and compute the water site properties (see below).

4.2 Water Sites definition, identification and characterization

Water sites (WSs) correspond to specific space regions, adjacent to the protein surface, where the probability of finding a water molecule is significantly higher than that observed in the bulk solvent. As shown in our previous works,[35,36,60] these regions can be readily identified by computing the probability of finding a water molecule inside the correspondingly defined region during an explicit solvent MD simulation. The region volume used to identify the WS is arbitrarily set to 1Å^3, and the WS center coordinates correspond to the average position of all the water oxygen atoms that visit the WS along the simulation. In other words, a water molecule is considered as occupying that WS, as soon as the distance between the position of its oxygen atom and the WS center is less than 0.6 Å. Once identified, for all putative WSs, we compute the following parameters: i) The Water Finding Probability (WFP), corresponding to the probability of finding a water molecule in the region defined by the WS (using the arbitrary volume value of

1Å^3) and normalized with respect to that of the bulk water which is considered to be the water density at the corresponding temperature and pressure values; thus the WFP is actually used as a cut-off value, to decide which putative WS are considered for further characterization. Only the WSs with a WFPs > 2 are retained. ii) R_{90}, corresponding to the radius the WS should have for a water molecule being found inside it 90% of the simulation time. This value is a measure of the WS dispersion, and is related with the mobility of the water molecules inside the WS. iii) R_{min}, computed as the distance between the WS position and the nearest heavy atom of the ligand in the superimposed structures of the free protein (where the WS have been identified) and the protein-ligand complex structure. Therefore, this parameter can be computed only in cases where the protein-ligand complex structure is previously known.[35]

4.3 Conventional AutoDock4 Docking Method (CADM) and Protocol. To perform conventional docking calculation we used the AutoDock4.2 program.[63] Briefly, the protocol employed for docking calculations is as follows: based solely on the protein receptor structure, the program firstly builds an energy grid for each ligand atom type, where the non-bonded protein-ligand interaction (including electrostatic and van der Waals contributions) are computed. Thus, during the docking calculation, the ligand binding energy estimations are calculated for each ligand position/conformation directly with the grid. Secondly, an initial set of ligand position/conformations are placed on the grid, and for each one the binding energy is computed. Bad conformations displaying poor interaction energy are eliminated, while best conformations are retained. New possible docking solutions are created from these best binding structures varying structural degrees of freedom. This Lamarckian type of genetic algorithm is continued until the best conformation or pose is obtained, corresponding to a putative ligand-protein complex. This procedure is called a docking run. Usually, for each protein-ligand pair hundreds

of runs are performed and the results are clustered according to their resulting ligand position/conformation, leading to a population parameter value for each putative complex structure (the population being the percentage of times an individual docking run results in a given binding mode for the used grid).[19,20,28,63] For the present calculation we kept all genetic algorithm parameters of the conformational search at their default values (150 for initial population size, 2.5 x 10^6 as the maximum number of energy evaluation, 2.7 x 10^4 as the maximum number of generations). For each protein-ligand pair we built the corresponding grids that represent the AutoDock4 scoring function using the ligand free protein structure, as provided either from the corresponding protein-ligand complex crystal structure (i.e. a re-docking calculation, or best case), or a random selected snapshot of the protein, taken from the explicit water MD simulation. The grid size and position were chosen so that they include the whole CBS. For this sake the grid center was placed in the geometric center of the CBS, computed as average coordinates (x, y and z) of all heavy atoms of all residues that compose the corresponding CBS. Residues that compose the CBS were defined as those residues with at least one heavy atom closer than 5Å from any heavy atom from the ligand in the corresponding protein-ligand complex reference structure. The grid size was then build extending 20 (for the mono and disaccharide binding proteins) and 25 Å (for the tri and tetra saccharide binding proteins) in each direction. The chosen grid spacing was 0.375 Å. For each structure 100 different docking runs were performed and the results were clustered according to the ligand-heavy atom RMSD using a cut-off of 2.0, as computed by the AutoDock4.2 program.

4.4 Water-Site-Biased Docking Method (WSBDM) and Protocol

In order to make use of the fact that carbohydrate -OH groups tend to occupy or replace the position of tightly bound waters, as characterized by the WSs on the protein surface, we

modified the AutoDock4 energy function, adding an additional energy term for each carbohydrate-ligand oxygen (OA type in AutoDock4's atom type nomenclature) to the original function, as described by equation 1 shown below.

$$\Delta G_O^M = \Delta G_O^{AD} - RT \sum_{i=1}^{N} \ln(WFP_i)\, e^{\frac{-\left((x-x_{WS,i})^2+(y-y_{WS,i})^2+(z-z_{WS,i})^2\right)^{\frac{1}{2}}}{R_{90i}}}$$

Equation (1)

Where, ΔG^M corresponds to the resulting modified scoring function, ΔG^{AD} corresponds to the original function, WFP*i* is the above defined water finding probability of the "ith" WS considered, X_{WS} ,Y_{WS} and Z_{WS} are the corresponding WS position coordinates, x, y and z each grid point coordinate, and R_{90} is the above defined volume for the corresponding WS. Therefore, each WS considered provides an interaction energy between the center of the WS position, and every OA atom (i.e. any carbohydrate oxygen), with a magnitude that is proportional to the Ln(WFP) and an amplitude that related to the WS size characterized by the R_{90}. The function is inspired in the fact that the likelihood that carbohydrate oxygen replaces the corresponding WS (measured by the R_{min} value) correlates with the WFP and the R_{90}, as shown in our previous work.[35] A comparative energy map of the resulting function can be shown in Figure 10 for Gal-3 CBS. The Figure clearly shows that conventional and biased energy grids are very similar at an isoenergetic value of -0.5 kcal/mol, although for lower energy values WS-biased grid shows the presence of energy wells in the places of the best WSs, which are not present in the original grid.

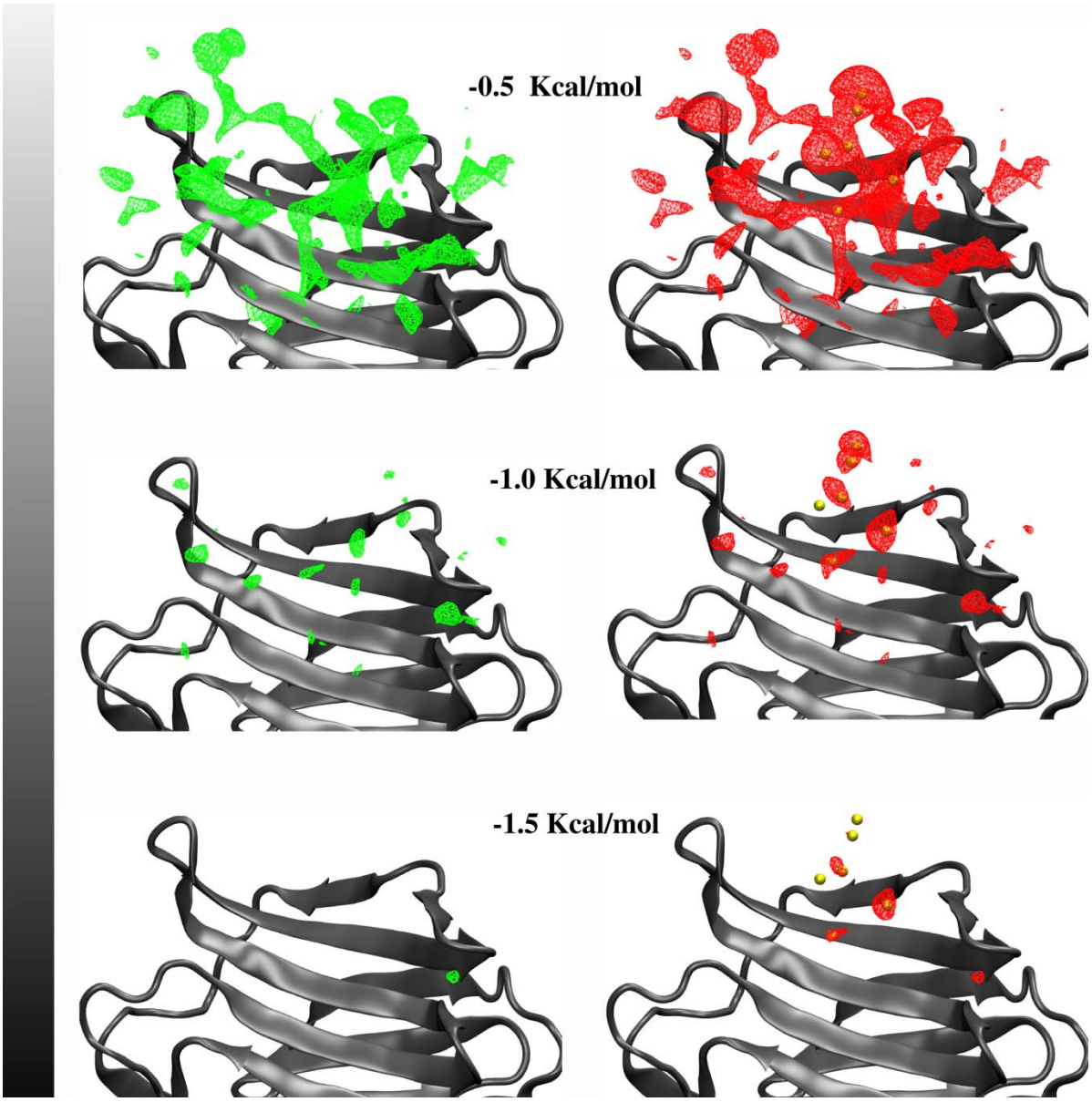


Figure 10. Conventional and WS biased interaction energy grids for the ligand Oxygen atoms in Gal-3 CBS. Ligand oxygen atom interaction energy grids drawn as isosurfaces with energy values of -0.5, -1.0 and -1.5 kcal/mol in top, middle and bottom panels respectively. Grids corresponding to the CADM and WSBDM are shown in left and right panels, respectively. The CBS corresponds to Gal-3. WS used to compute the biased grid are shown as small yellow spheres in the right panel.

The WSBDM is then employed in the same manner as the CADM but introducing the biased function computed with a given number of WS and their corresponding parameters when creating the grid. For strict comparison purposes, all other docking parameters, like grid size, position and number of docking runs were the same as those used in the CADM. Thus the computational time needed to perform a docking calculation with the WSBDM is exactly the same as that required by the CADM once the modified grid is built. All characterized WS used to build the modified grids are presented for each protein in Table 1. WFP, R_{90} and R_{min} values were computed as described previously.

WS numbering is arbitrary. WFP corresponds to the probability of finding a water molecule in the region defined by the WS and normalized with respect to that of the bulk water; R_{90}, corresponds to the radius the WS should have for a water molecule being found in its region 90% of the time. R_{min}, is computed as the distance between the WS position and the nearest heavy atom of the ligand in the corresponding protein-ligand complex structure. Data for CBM32, CBM40, Gal-3 and ConA have been already reported[35] while data for SPD and CD44 were computed in the present work.

Table 1. WS number and characteristics for each protein CBS**.** The table shows the protein name, Ws number and WFP in the three first columns respectively, while in the last two columns the R_{90} and R_{min} value are summarized for all WSs of the proteins**.**

Protein	WS	WFP	R90	Rmin
CBM32	1	18.9	1.6	1.5
	2	2.4	4.4	2.5
	3	6.2	2.1	0.9
	4	11.5	1.7	1.2
CBM40	1	17.8	1.8	0.8
	2	8.5	3.1	1.7
	3	13.6	1.7	0.7
	4	7.5	2.1	1.4
	5	6.5	3.5	1.3
	6	18.8	1.7	0.8
Galectin 3	1	7.0	2.2	1.6
	2	9.1	1.5	0.4
	3	4.1	3.3	1.8
	4	18.5	1.5	0.3

	5	7.0	2.4	1.3
	6	2.3	2.9	0.8
	7	6.6	4.1	2.3
Concanavalin A	1	7.7	2.9	1.5
	2	8.8	2.8	1.2
	3	12.8	1.9	1.4
	4	2.1	1.4	1.5
	5	7.2	3.9	0.9
	6	9.3	2.2	1.5
	7	9.8	3.1	1.1
	8	20.3	1.6	0.7
	9	16.9	2.1	0.6
	10	3.4	2.5	2.9
SPD	1	3.7	0.8	0.4
	2	8.0	0.9	0.7
	3	2.4	0.6	2.1
	4	22.0	0.7	0.3
	5	24.1	0.6	1.8

CD44	1	16.0	0.8	0.7
	2	5.7	0.8	1.0
	3	5.4	0.6	0.5
	4	12.7	1.0	0.8
	5	2.7	0.4	2.1

4.5 Data analysis

Results comparing both the CADM and WSBDM were analyzed in terms of their capability to correctly predict the protein carbohydrate complex. Two issues were considered: First, how close to the reference complex structure does the method dock the corresponding ligand, thus resulting in a measure of the method accuracy. Secondly, what is the method capability to distinguish the right complex from wrong predictions, a parameter that may be thought as the method precision. To determine the accuracy of the prediction, we computed the ligand heavy atoms RMSD of each predicted complex using the CADM or WSBDM, with respect to the position of the ligand in the corresponding complex crystal structure. For the cases where the receptor structure is taken from the crystal structure of the complex (i.e. a re-docking calculation) no prior structural alignment of the receptor structure to the reference is needed, while for those cases where the docking was performed on a MD snapshot, the predicted complex structure was first structurally aligned to the reference complex structure considering the protein CBS heavy atoms.

To analyze the method precision, it is important to remember that for each predicted complex AutoDock4[19,20,63] software yields two parameters, namely the predicted binding energy (ΔG_B) and the cluster population (Pop). The ΔG_B is defined as the free energy difference for the binding process, therefore the more negative (larger absolute) values, correspond to better binding conformations. %Pop is the percentage of individual docking runs that resulted in the same binding mode for a particular receptor structure (that defines the grid) ligand pair. The predicted complexes are ranked according to the predicted ΔG_B. However, several times, low population clusters appear close, and have even lower (i.e better) binding energies than higher population clusters which resemble more closely the real complex. Therefore, both parameters should be taken into account in order to have a reliable prediction. In other words, an optimal result should give a high population and low binding energy cluster, which significantly differs in both parameters from the others. As will be 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.

Supplementary Information

A summary of all the obtained results, and the Population vs. Energy plots for the data not shown in the manuscript can be found in the supplementary information. Scripts and programs to determine and compute Water Site properties and to modify the AutoDock4 grid are freely available under request. This material is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgements

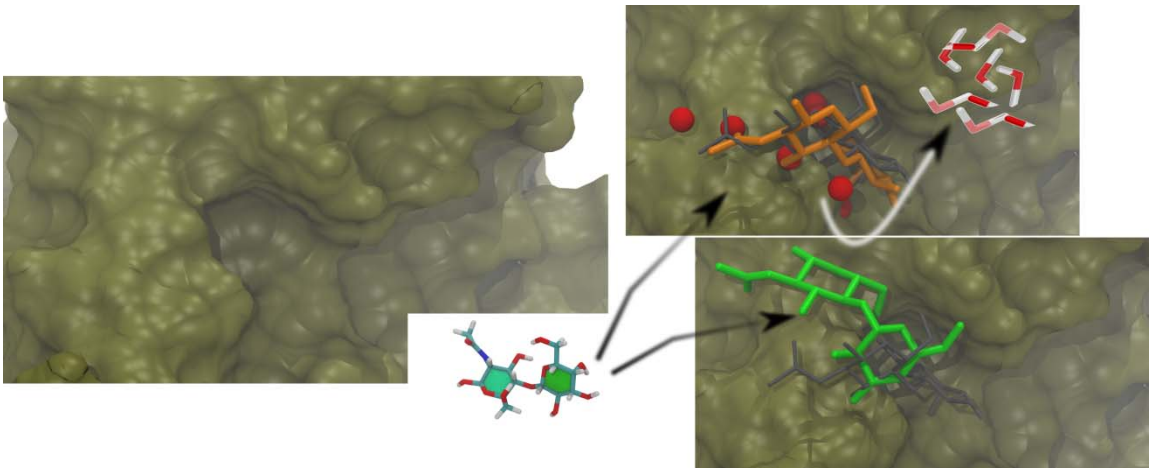
This work was supported by PICT-2010-416, UBACyT 2010-2012 and Bunge y Born to MAM. DFG, CM are fellowship from CONICET. SDL and MAM are staff members of CONICET.

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 Córdoba.

Abbreviations:

MD: Molecular Dynamics**, WSBDM:** Waters Site-Biased Docking Method, **CADM:** Conventional AutoDock4 Docking Method, **WS:** Water Site, **WSs:** Water Sites, **Gal-3:** Galectin-3, **CBM32:** Carbohydrate-Binding Module 32, **CBM40:** Carbohydrate-Binding Module 40, **ConA:** Concanavalin-A, **WFP:** Water Finding Probability, **RMSD:** Root Mean Square Deviation, **Pop:** Docking pose cluster population, **IFST:** Inhomogeneous Fluid Solvent Theory, **CBS:** Carbohydrate Binding Site. **Trimmanoside:** 3,6-Di-O-(α-D-Mannopyranosyl)-α-D-Mannopyranoside. **Sialic Acid:** α-D-N-acetylneuraminic acid. **Maltotriose:** α-D-Glucopyranosyl(l-4)-α-D-Glucopyranosyl(l-4)-α-D-Glucopyranose. **Galactose:** β-D-Galactose**. LAcNAc or N-Acetyl-Lactosamine:** β-D-Galactosyl-1,4-N-Acetyl-D-Glucosamine. **Hialuronan tetrasaccharide:** β-D-Glucuronyl(1-3)-2-Acetamido-β-D-Glucopyranosyl(1-4)-β-D-Glucuronyl(1-3)-N-Acetyl-β-D-Glucosamine

Table of Contents Art Work (TOC)



References

(1) Feinberg H, Taylor ME, Razi N, McBride R, Knirel YA, Graham SA, Drickamer K, Weis WI. 2011. Structural basis for langerin recognition of diverse pathogen and mammalian glycans through a single binding site. *J Mol Biol*. 405:1027-1039.

(2) Drews J. 2000. Drug discovery: A historical perspective. *Science*. 287:1960-1964.

(3) Loging W, Rodriguez-Esteban R, Hill J, Freeman T, Miglietta J. 2012. Cheminformatic/bioinformatic analysis of large corporate databases: Application to drug repurposing. *Drug Discovery Today: Therapeutic Strategies*. 8:109-116.

(4) Powlesland AS, Quintero-Martinez A, Lim PG, Pipirou Z, Taylor ME, Drickamer K. 2010. Engineered carbohydrate-recognition domains for glycoproteomic analysis of cell surface glycosylation and ligands for glycan-binding receptors. *Methods Enzymol*. 480:165-179.

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

(6) Brooijmans N, Kuntz ID. 2003. Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct*. 32:335-373.

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

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

(9) 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 Inf Model*. 49:2564-2571.

(10) Amzel LM. 1998. Structure-based drug design. *Curr Opin Biotechnol*. 9:366-369.

(11) Shoichet BK, McGovern SL, Wei B, Irwin JJ. 2002. Lead discovery using molecular docking. *Curr Opin Chem Biol*. 6:439-446.

(12) Barril X, Javier Luque F. 2012. Molecular simulation methods in drug discovery: A prospective outlook. *J Comput-Aided Mol Des*. 26:81-86.

(13) Feliu E, Oliva B. 2010. How different from random are docking predictions when ranked by scoring functions? *Proteins*. 78:3376-3385.

(14) Seco J, Luque FJ, Barril X. 2009. Binding site detection and druggability index from first principles. *J Med Chem*. 52:2363-2371.

(15) Agostino M, Jene C, Boyle T, Ramsland PA, Yuriev E. 2009. Molecular docking of carbohydrate ligands to antibodies: Structural validation against crystal structures. *J Chem Inf Model*. 49:2749-2760.

(16) Kerzmann A, Fuhrmann J, Kohlbacher O, Neumann D. 2008. Balldock/slick: A new method for protein-carbohydrate docking. *J Chem Inf Model*. 48:1616-1625.

(17) Kerzmann A, Neumann D, Kohlbacher O. 2006. Slick--scoring and energy functions for protein-carbohydrate interactions. *J Chem Inf Model*. 46:1635-1642.

(18) Nurisso A, Kozmon S, Imberty A. 2008. Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin cel-iii. *Mol. Simul.* 34:469-479.

(19) Goodsell DS, Morris GM, Olson AJ. 1996. Automated docking of flexible ligands: Applications of autodock. *J Mol Recognit*. 9:1-5.

(20) Morris GM, Goodsell DS, Huey R, Olson AJ. 1996. Distributed automated docking of flexible ligands to proteins: Parallel applications of autodock 2.4. *J Comput Aided Mol Des*. 10:293-304.

(21) Li L, Chen R, Weng Z. 2003. Rdock: Refinement of rigid-body protein docking predictions. *Proteins*. 53:693-707.

(22) Mishra SK, Adam J, Wimmerova M, Koca J. 2012. In silico mutagenesis and docking study of ralstonia solanacearum rsl lectin: Performance of docking software to predict saccharide binding. *J Chem Inf Model*. 52:1250-1261.

(23) Agostino M, Yuriev E, Ramsland PA. 2011. A computational approach for exploring carbohydrate recognition by lectins in innate immunity. *Front Immunol*. 2:23.

(24) Trott O, Olson AJ. 2010. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31:455-461.

(25) Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N, Rizzo RC. 2006. Development and validation of a modular, extensible docking program: Dock 5. *J Comput Aided Mol Des*. 20:601-619.

(26) Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. 2004. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 47:1739-1749.

(27) 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.* 19:1639-1662.

(28) Huey R, Morris GM, Olson AJ, Goodsell DS. 2007. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* 28:1145-1152.

(29) Li Z, Lazaridis T. 2003. Thermodynamic contributions of the ordered water molecule in hiv-1 protease. *J Am Chem Soc*. 125:6636-6637.

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

(31) 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*. 130:2817-2831.

(32) Luccarelli J, Michel J, Tirado-Rives J, Jorgensen WL. 2010. Effects of water placement on predictions of binding affinities for p38α map kinase inhibitors. *J. Chem. Theory Comput.* 6:3850-3856.

(33) Michel J, Tirado-Rives J, Jorgensen WL. 2009. Prediction of the water content in protein binding sites. *The Journal of Physical Chemistry B*. 113:13337-13346.

(34) 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*. 131:15403-15411.

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

(36) Di Lella S, Marti MA, Alvarez RM, Estrin DA, Ricci JC. 2007. Characterization of the galectin-1 carbohydrate recognition domain in terms of solvent occupancy. *J. Phys. Chem. B*. 111:7360-7366.

(37) Young T, Abel R, Kim B, Berne BJ, Friesner RA. 2007. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc Natl Acad Sci U S A*. 104:808-813.

(38) de Beer SB, Vermeulen NP, Oostenbrink C. 2010. The role of water molecules in computational drug design. *Curr Top Med Chem*. 10:55-66.

(39) Varki A, Cummings, R., Esko, J., Freeze, H., Hart, G., and Marth, J. *Essentials of glycobiology*, 1999.

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

(41) Crocker PR, Paulson JC, Varki A. 2007. Siglecs and their roles in the immune system. *Nat Rev Immunol*. 7:255-266.

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

(43) Leffler H, Carlsson S, Hedlund M, Qian Y, Poirier F. 2004. Introduction to galectins. *Glycoconj J*. 19:433-440.

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

(45) Rabinovich GA. 2005. Galectin-1 as a potential cancer target. *Br J Cancer*. 92:1188-1192.

(46) Hirabayashi J. 2004. Lectin-based structural glycomics: Glycoproteomics and glycan profiling. *Glycoconj J*. 21:35-40.

(47) Di Lella S, Sundblad V, Cerliani JP, Guardia CM, Estrin DA, Vasta GR, Rabinovich GA. 2011. When galectins recognize glycans: From biochemistry to physiology and back again. *Biochemistry*. 50:7842-7857.

(48) Echeverria I, Amzel LM. 2011. Disaccharide binding to galectin-1: Free energy calculations and molecular recognition mechanism. *Biophys J*. 100:2283-2292.

(49) Taylor ME, Drickamer K. 2009. Structural insights into what glycan arrays tell us about how glycan-binding proteins interact with their ligands. *Glycobiology*. 19:1155-1162.

(50) Ernst B, Magnani JL. 2009. From carbohydrate leads to glycomimetic drugs. *Nat. Rev. Drug Discovery*. 8:661-677.

(51) Balzarini J. 2007. Targeting the glycans of glycoproteins: A novel paradigm for antiviral therapy. *Nat Rev Microbiol*. 5:583-597.

(52) Feinberg H, Mitchell DA, Drickamer K, Weis WI. 2001. Structural basis for selective recognition of oligosaccharides by dc-sign and dc-signr. *Science*. 294:2163-2166.

(53) Banerji S, Wright AJ, Noble M, Mahoney DJ, Campbell ID, Day AJ, Jackson DG. 2007. Structures of the cd44-hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. *Nat Struct Mol Biol*. 14:234-239.

(54) Feng L, Sun H, Zhang Y, Li DF, Wang DC. 2010. Structural insights into the recognition mechanism between an antitumor galectin aal and the thomsen-friedenreich antigen. *FASEB J*. 24:3861-3868.

(55) Frank M, Schloissnig S. 2010. Bioinformatics and molecular modeling in glycobiology. *Cell. Mol. Life Sci.* 67:2749-2772.

(56) von der Lieth CW, Freire AA, Blank D, Campbell MP, Ceroni A, Damerell DR, Dell A, Dwek RA, Ernst B, Fogh R, Frank M, Geyer H, Geyer R, Harrison MJ, Henrick K, Herget S, Hull WE, Ionides J, Joshi HJ, Kamerling JP, Leeflang BR, Lutteke T, Lundborg M, Maass K, Merry A, Ranzinger R, Rosen J, Royle L, Rudd PM, Schloissnig S, Stenutz R, Vranken WF, Widmalm G, Haslam SM. 2011. Eurocarbdb: An open-access platform for glycoinformatics. *Glycobiology*. 21:493-502.

(57) Agostino M, Sandrin MS, Thompson PE, Yuriev E, Ramsland PA. 2010. Identification of preferred carbohydrate binding modes in xenoreactive antibodies by combining conformational filters and binding site maps. *Glycobiology*. 20:724-735.

(58) Woods RJ, Tessier MB. 2010. Computational glycoscience: Characterizing the spatial and temporal properties of glycans and glycan-protein complexes. *Curr Opin Struct Biol*. 20:575-583.

(59) Saraboji K, Hakansson M, Genheden S, Diehl C, Qvist J, Weininger U, Nilsson UJ, Leffler H, Ryde U, Akke M, Logan DT. 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*. 51:296-306.

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

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

(62) Lazaridis T. 1998. Inhomogeneous fluid approach to solvation thermodynamics. 2. Applications to simple fluids. *J. Phys. Chem. B*. 102:3542-3550.

(63) 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.

(64) Boraston AB, Ficko-Blean E, Healey M. 2007. Carbohydrate recognition by a large sialidase toxin from clostridium perfringens†. *Biochemistry*. 46:11352-11360.

(65) von Schantz L, Hakansson 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*. 22:948-961.

(66) Ahmad N, Gabius HJ, Sabesan S, Oscarson S, Brewer CF. 2004. Thermodynamic binding studies of bivalent oligosaccharides to galectin-1, galectin-3, and the carbohydrate recognition domain of galectin-3. *Glycobiology*. 14:817-825.

(67) Loris R, Maes D, Poortmans F, Wyns L, Bouckaert J. 1996. A structure of the complex between concanavalin a and methyl-3,6-di-o-(alpha-d-mannopyranosyl)-alpha-d-mannopyranoside reveals two binding modes. *J Biol Chem*. 271:30614-30618.

(68) Crouch E, McDonald B, Smith K, Cafarella T, Seaton B, Head J. 2006. Contributions of phenylalanine 335 to ligand recognition by human surfactant protein d: Ring interactions with sp-d ligands. *J Biol Chem*. 281:18008-18014.

(69) Seetharaman J, Kanigsberg A, Slaaby R, Leffler H, Barondes SH, Rini JM. 1998. X-ray crystal structure of the human galectin-3 carbohydrate recognition domain at 2.1-a resolution. *J Biol Chem*. 273:13047-13052.

(70) Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ. 2005. The amber biomolecular simulation programs. *J. Comput. Chem.* 26:1668-1688.

(71) Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. 2006. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*. 65:712-725.

(72) Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*. 81:3684-3690.