Carbohydrate-Binding Proteins: Dissecting Ligand Structures through Solvent Environment Occupancy

Diego F. Gauto,[†] Santiago Di Lella,[†,§] Carlos M. A. Guardia,[†] Darío A. Estrin,[†] and Marcelo A. Martí*[,†,‡]

Departamento de Química Inorgánica, Analítica, y Química Física, INQUIMAE-CONICET, Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón II, C1428EHA Ciudad de Buenos Aires, Argentina, and Instituto Superior de Investigaciones Biológicas (INSIBIO−CONICET), Facultad de Bioquímica, Química y Farmacia, Universidad Nacional de Tucumán, Chacabuco 462, T4000INI San Miguel de Tucumán, Tucumán, Argentina

Received: February 9, 2009; Revised Manuscript Received: April 30, 2009

Formation of protein ligand complexes is a fundamental phenomenon in biochemistry. During the process, significant solvent reorganization is produced along the contact surface and many water molecules strongly bound to the protein's ligand binding site must be displaced. Both the thermodynamics and kinetics of this process are complex and a clear understanding at the microscopic level has been not achieved so far. Special attention has been paid to the structure of water molecules on carbohydrate recognition sites of various proteins, and many studies support the idea that displacement of these water molecules should have a crucial effect on the binding free energy.

Molecular dynamics (MD) simulations in explicit water solvent is a very promising approach for this type of studies. Using MD simulations combined with statistical mechanics analysis, thermodynamic properties of these water molecules can be computed and analyzed in a comparative view. Using this idea, we developed a set of analysis tools to link solvation with ligand binding in a key carbohydrate binding protein, human galectin-1 (hGal-1). Specifically, we defined water sites (WS) in terms of the thermodynamic properties of water molecules strongly bound to protein surfaces. In the present work, we selected a group of proteins whose ligand bound complexes have been already structurally characterized in order to extend the analysis of the role of the surface associated water molecules in the ligand binding and recognition process. The selected proteins are concanavalin-A (Con-A), galectin-3 (Gal-3), cyclophilin-A (Cyp-A), and two modules CBM40 and CBM32 of the multimodular bacterial sialidase.

Our results show that the probability of finding water molecules inside the WS, $p(\upsilon)$, with respect to the bulk density is directly correlated to the likeliness of finding an hydroxyl group of the ligand in the protein−ligand complex. This information can be used to analyze in detail the solvation structure of the carbohydrate recognition domain (CRD) and its relation to the possible protein ligand complexes and suggests addition of OH-containing functional groups to displace water from high $p(\upsilon)$ WS to enhance drugs, specially glycomimetic-drugs, protein affinity, and/or specificity.

Introduction

Formation of protein ligand complexes is a fundamental process in biochemistry. In their natural environment, proteins are surrounded by water molecules, interacting with them and modifying thereby the solvent structure.[1,2] Therefore, on protein surfaces, water molecules are not placed randomly but tend to occupy specific positions and orientations which are dependent on their interactions with the protein surface and its particular physicochemical properties.[3]

During the ligand binding process, significant solvent reorganization is produced along the contact surface. In some cases, water molecules strongly bound to the protein's ligand binding region must be displaced to allow proper contact between the structures, while some are retained bridging protein−ligand interactions.[4−8] Both the thermodynamics and kinetics of this solvent reorganization process are complex and a clear understanding at the microscopic level has not been achieved so far.[9]

Enormous attention has been paid to the structure and dynamics of the water molecules on the carbohydrate recognition sites of various proteins. The corresponding contact surfaces are highly hydrophilic, especially when compared to the hydrophobic ones typically found for protein−protein and protein drug interactions. In this context, the effects of displacement of the well-ordered waters in Concanavalin A with different trimannosides complexes were evaluated,[10] showing that rearrangement of these water molecules contributes favorably to the binding affinity. In another study,[11] it was shown that for a range of proteins the amount of heat released due to the binding of saccharide ligands was significantly reduced when D_2O was used as the solvent. A more general view was provided by Dunitz[12] who showed that the release of a highly ordered water whose entropic contribution to the free energy change is up to 2 kcal/mol at 300 K. This estimation is based on the comparison of standard heats of hydration between a number

* To whom correspondence should be addressed. Phone: +54 11 45763378, ext. 124. Fax: +54 11 45763341. E-mail: marcelo@qi.fcen.uba.ar.
† Departamento de Química Inorgánica, Analítica, y Química Física, INQUIMAE-CONICET, Universidad de Buenos Aires.
‡ Departamento de Química Biológica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires.
§ Universidad Nacional de Tucumán.

10.1021/jp901196n CCC: $40.75
Published on Web 06/01/2009

of anhydrous and hydrated inorganic salts. Studies of potent cyclic nonapeptide HIV protease inhibitors provided further support for this entropic effect,[13] and other works confirmed this idea, by means of molecular dynamics simulations, where the authors demonstrated that ligands designed to displace ordered water molecules exhibited enhanced affinities.[14,15] All these studies show the relevance of the water molecules strongly bound in the protein ligand binding site and support the idea that displacement of these water molecules should have a crucial effect on the binding free energy. Therefore, we expect that the study of these water thermodynamic properties would yield useful information about the carbohydrate binding mechanism and thermodynamics.

Molecular dynamics (MD) simulations in explicit water solvent appear as a very promising approach for the study of water−protein interactions. Through this methodology, it is feasible to provide a detailed description of the structural and dynamical properties of the protein surface bound individual water molecules. Moreover, using MD simulations combined with statistical mechanics analysis, thermodynamic properties of these water molecules can be computed and analyzed in a comparative view.[16] For example, the standard free energy of tying up a water molecule in the binding site of protein complexes has been studied by using the double-decoupling method, a methodology that allows the calculation of the energies involved in transferring a specific water molecule from the protein surface to the bulk water.[17] Also, using the inhomogeneous fluid approach, Lazaridis et al.[18−20] were able to compute the energetic and entropic contributions to the water−protein interaction free energies for a selected group of water molecules tightly bound at protein surfaces. Using a similar idea in a previous work from our group we computed the thermodynamic properties of water molecules located at several specific sites, called "water sites" (WS), on the surface of human galectin-1 carbohydrate recognition domain (CRD), and showed their relevance for carbohydrate recognition.[21]

Water sites are defined as confined space regions close to the protein surface showing a high probability for finding water molecules inside them (water finding probability, WFP). The positions of the WS are defined by the coordinates of the maximum probability point using as a reference a surface residue of the protein which is able to interact favorably with the water. Our work showed that those WS with higher WFP tend to occupy the same position of hydroxil groups of the carbohydrate ligand in the protein−ligand complex. Therefore, by studying the properties of the WS on protein surfaces, carbohydrate binding modes may be predicted and better and rational design of glycomimetic drugs may be envisaged.[22−25]

A similar and promising line of research is described in a recent work[26] in which a map of water molecules occupancy was constructed in the active site of Factor Xa based on explicit solvent MD dynamics study. For each identified hydration site, the authors computed its thermodynamic properties using the inhomogeneous solvation theory. These data were then used to construct a model to compute the free energy differences for selected pairs of Factor Xa ligands. The results correlated exceptionally well with the experimental binding energy data, further demonstrating the fundamental role played by water molecule displacement in ligand affinity.

In the present work, we selected a group of proteins whose ligand bound complexes have been already structurally characterized in order to analyze the role of the surface-associated water molecules in the ligand binding and recognition process. The selected proteins are concanavalin-A (Con-A), galectin-3 (Gal-3), cyclophilin-A (Cyp-A), and two modules CBM40 and CBM32 of the multimodular bacterial sialidase. The choice of these proteins was motivated by their physiopathological relevance and the fact that in all cases water molecules tightly bound are expected to play a key role in determining protein function.

Galectins are a family of animal lectins defined by common consensus sequences and structures and possess specificities for β-galactosil and N-acetyllactosamine (LacNAc) residues in oligosaccharides. The family includes fourteen proteins that have been implicated in diverse biological roles such as the regulation of inflammation, modulation of cell adhesion, growth, and death. On the basis of their structures, they can be classified in three groups: (i) the monomeric or homodimeric type galectins (Gal-1, -2, -5, -7, -10, -11), which exist in an equilbrium between a monomeric and homodimeric state; (ii) the chimera type galectins, which include only Gal-3 and consist of a galectin type domain connected to a terminal domain; (iii) the tandem repeat type galectins (Gal-4, -6, -8, and -9) consisting of two CRD in a single subunit connected by a small linker peptide.[27−31] The particular structure of Gal-3, together with the data showing that significant differences are observed between the previously studied Gal-1 and Gal-3 binding to internal LacNAc (and lactose) residues of oligosaccharides, makes Gal-3 a good choice for the present study.[32,33]

Regarding Con-A, it is a lectin that binds trimmanoside ligands and other carbohydrates and was already studied using MD simulations with two different bound ligands.[16] The difference between the two trimmanoside ligands resides in the addition of a hydroxyl group in trimmanoside-2 that displaces a water molecule bridging the protein-trimmanoside-1 interaction. In particular, the thermodynamics properties of this strongly bound water molecule were studied and related to the differential binding affinity of both ligands. In yet another study by the same authors, the thermodynamic properties of several water molecules found in the binding interface of Cyp-A bound to either the undecapeptide Cyclosporin A (CsA) or (5-hydroxynorvaline)-2-cyclosporin.[18] CsA is an immunosuppressive drug that when bound to Cyp-A inhibits calcineurin activity. The availability of previous water thermodynamic data for these proteins and their completely different binding characteristics, compared with the galectins, prompted their use in this study.

Finally, we selected two recently structurally characterized carbohydrate binding domains that to our knowledge have not been the subject of detailed solvation studies or studied using MD techniques. Both domains belong to a large multimodular sialidase from *Clostridium perfringens*. Clostridial sialidases, classified as minor toxins produced by the bacteria, are believed to cause, among other problems, life-threatening hemolysis during blood transfusions of infected individuals[34] The NanJ gene encodes a large sialidase whose N-terminal two modules belong to the 32 and 40 families of carbohydrate binding modules, respectively (CBM32 and CBM40). Both modules, recently structurally characterized,[34] display a β-sandwich fold with a simple carbohydrate binding site on one side of the sandwich. CBM32 has a preference for terminal galacto configured sugars and its structure was obtained with lactose and calcium ion bound. CBM40 binds preferentially sialic acid as in the resolved structure. Interestingly, recent studies have indicated that targeting the binding activity of CBMs in pathogenesis-related glycoside hydrolases with multivalent ligands/substrates may be a viable approach to developing carbohydrate-based therapeutics that inhibit these toxins.[35]

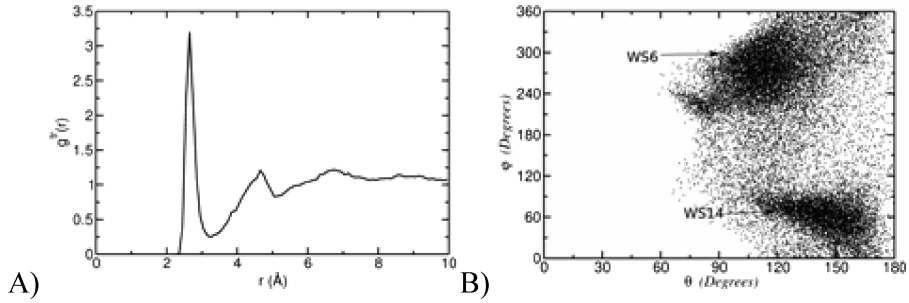

Figure 1. Translational radial distribution function $g(r)$ (A) and angular bidimensional plot distribution of WFP (B) for water molecules around concanavalin Asp16 carboxyl.

Computational Methods

Setup of the Systems and Equilibration. Protein coordinates were retrieved from the Protein Data Bank, corresponding code 1ona for Con-A, 1a3k for Gal-3, 1mik for Cyp-A, and 2v73 and 2v72 for modules CBM40 and CBM32, repectively. For each protein, only one monomer was simulated. Hydrogen atoms were added with the LEaP module of the AMBER 9 program.[36] Standard protonation states were assigned to tritratable 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 immersed in a truncated octahedral box of TIP3P waters. Each system was first optimized using a conjugate gradient algorithm for 2000 steps, followed by 200 ps long constant volume MD thermalization 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 processes, the protein Cα atoms were restrained by a 1 kcal/mol harmonic potential. After equilibration, 20 ns long production MD simulations were performed for each protein. No atomic or geometrical restraints were applied during each of the production runs.

Simulation Parameters. All amino acid parameters correspond to the Amber f99SB force field,[37] while Glycam-04 was used for ligand parameters.[38] Pressure and temperature were kept constant with the Berendsen barostat and thermostats,[39] respectively. All simulations were performed with periodic boundary conditions using the particle mesh Ewald 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.

Water Site Identification. In order to identify the presence of water sites on the protein surface, the following methodology was performed. First, we obtained radial distribution functions $g(r)$ for water molecules around selected potential hydrogen bonding donor/acceptor atoms in the recognition domain of each protein. These functions allow the identification of the region corresponding to the first solvation shell of the chosen atoms. Second, both translational angular $g(\theta)$ and dihedral $g(\phi)$ distribution functions were constructed taking the data of only those water molecules occupying the first solvation shell with respect to the reference atom, as defined by the peak in the $g(r)$ functions. Visual analysis of the bidimensional plots allows clear identification of the WS, as the regions with high WFP.

As an example, Figure 1 shows both the radial distribution function for the carboxyl oxygen of Asp16 in concanavalin, and its bidimensional angular scatter plot. As clearly presented in the $g(r)$, the first peak corresponds to the first solvation shell. The bidimensional angular plot evidence the presence of two regions of high WFP, defining two WS. This methodology has been successfully applied previously in our group to identify the WS in Gal-1.[21]

Using the above-mentioned analysis for each potential WS, the maximum WFP point is determined. For all subsequent calculations, this point defines the center of the corresponding WS. A water molecule is defined as being inside the corresponding WS if its oxygen distance to the WS center is less than 0.6 Å, a value approximately corresponding to a volume of 1 Å^3 for the WS. The corresponding translation distribution function $g(r)$ and angular bidimensional plot distribution for all the sites studies in this work are shown in the Supporting Information (Figures S1 to S28). In order to eliminate ambiguities in WS definitions, angular bidimensional plot analysis was complemented by visual analysis of the three-dimensional grid clustering of the WFP performed by Visual Molecular Dynamics software[40] and analysis of the convergence of the computed parameters. Checking for convergence of the reported quantities (described above) was also performed using different segments of the simulation or different references to define a given WS ensuring reproducibility in the WS definition and properties.

Calculation of Structural Dynamic and Thermodynamic Properties for the WS. In order to compute the potential energy associated with the interaction of water molecules in the WS with the protein and the rest of the solvent, for each snapshot along the simulation van der Waals and electrostatic potential contributions were calculated between the water located inside the WS and either the protein (E_p) or the other solvent molecules E_w. Contributions were considered up to a distance of 8 Å to the WS. This cutoff has already shown to yield reasonably converged results.[21] For each WS, the mean interaction energies $\langle E_x \rangle$ were computed over the whole simulation. Total mean interaction energies $\langle E_t \rangle$ of a water molecule inside the WS were then computed. Using as a reference the interaction energy of a water molecule in bulk water ($\langle E_{wat-bulk} \rangle$), the differential mean interaction energy is computed $\langle \Delta E \rangle$ as

$$\langle \Delta E \rangle = \langle E_t \rangle - \langle E_{wat\text{-}bulk} \rangle$$

This difference in energy corresponds to the gain in potential energy of transferring a water molecule from the bulk solvent to the corresponding WS. As already reported, energetic values calculated for WS are reasonably converged in the time scale of our simulations, as shown in SF 29.[21]

For each WS, we also computed the probability $p(v)$ of finding water molecule inside it. They were then normalized with respect to that of the bulk water that corresponds to the water density at the corresponding temperature and pressure values. The values were computed using an arbitrary volume of 1 Å^3 for the WS.

TABLE 1: Thermodynamic and Structural Parameters for Water Sites around Gal-3 Carbohydrate Recognition Domain[a]

WS	reference atom	resid	$<E_p>$	$<E_w>$	$<E_t>$	$<\Delta E>$	WFP	R_{90}	R_{min}
1	NE2	Hie158	−10.1	−12.3	−22.4	−4.9	7.0	2.2	1.6
2	NH1	Arg162	−25.0	−3.5	−28.5	−11.4	9.1	1.5	0.4
3	NH1	Arg162	−11.0	−10.5	−21.5	−5.4	4.1	3.3	1.8
4	ND2	Asn174	−13.8	−10.4	−24.2	−6.4	18.5	1.5	0.3
5	OE2	Glu184	−19.0	−5.0	−24.0	−6.9	7.0	2.4	1.3
6	OE1	Glu184	−10.4	−11.6	−22.0	−4.7	2.3	2.9	0.8
7	NH2	Arg186	−7.0	−14.7	−21.7	−4.8	6.6	4.1	2.3

[a] Energies (E_p, E_w, E_t and ΔE) are in kcal/mol, R_{90} and R_{min} are in angstroms.

As a measure of how dispersed are water molecules localized during the simulations within the regions defined for WS, we computed for each WS the radius at which 90% of the time a water molecule can be found inside. Therefore, the computed R_{90} results in a measurement of the size of the WFP in volume units for each WS. The volume of the WS also provides a measurement of its dispersion and therefore of the associated translational entropy, as shown by Lazaridis et al.[16,20] The greater the volume (larger R_{90}) the more conformational freedom have the water molecules inside the WS and therefore they are expected to have bigger entropy.

Finally, to analyze the correlation between the WS position respect to the protein surface and the structure of the protein ligand complex, we calculated the so-called R_{min} values. For this sake, WS positions defined as the maximum WFP point with respect to the protein CRD surface during the MD trajectories of the simulated system were superimposed on the protein ligand complex crystallographic structure. R_{min} values were then computed as the distance of the WS position to the nearest heavy atom of the ligand in the superimposed structures. This procedure was performed with the help of the Visual Molecular Dynamics software of the Illinois University.[40]

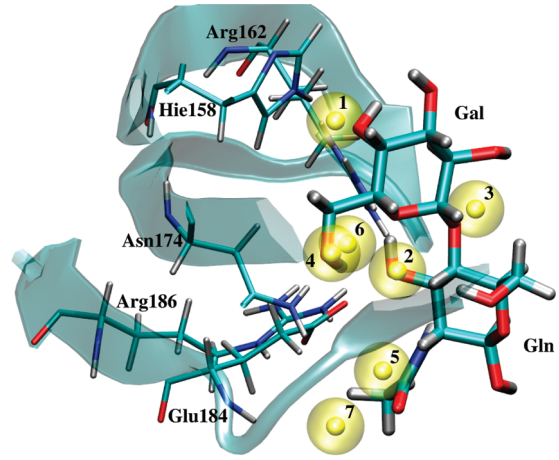

Figure 2. Structure of the Gal-3 LacNac complex superimposed on the WS positions relative to the protein structure. WS are shown as yellow spheres and numerated from 1 to 7.

Results and Discussion

The results are organized as follows: for each protein a brief description of the ligand binding site is given and the detected WS are structurally and thermodynamically characterized. Second, results are interpreted in terms of their respective ligand binding characteristics and subsequently discussed. Finally, all the data is analyzed in a comparative way.

Structural and Dynamical Characterization of WS in Gal-3 CRD. Gal-3 overall structure shows the typical β-sandwich structure. The CRD region in Galectin-3 is defined by residues 144 to 188. From these residues, we selected nine (Arg144, Asp148, His158, Glu165, Arg162, Asn174, Lys176, Glu184, and Arg186) as those capable of establishing strong HB with water molecules for the determination of the WS. Using these residues as references and the analysis protocol described in Computational Methods, we were able to find seven WS with significant water occupation probability in the CRD of Gal-3. For each WS, the thermodynamic properties were computed as described in the methods section. The results together with the best reference atom used to define the WS are shown in Table 1.

Most of the identified WS for Gal-3 have more than five times the probability of finding a water molecule than in the bulk solvent. Sites with highest WFP values correspond to WS 2 and 4, which also display a very low R_{90} of less than 1.5 Å. In a second group, we can find WS 1, 5, and 7 with WFP about 7 times compared with the bulk solvent.

In order to analyze the role of the WS in connection to ligand binding, we compare the position of the WS in the three reported X-ray Gal-3 complex structures corresponding to the Gal-3 LacNac complex (PDBid 1KJL) the Gal-3-lactose complex (PDBid 2NN8) and the Gal-3 complex of a *N*-acetyllactosamine derivative bearing a phenyl group at position 3′ (PDBid 1KJR). A quantitative measurement of this replacement can be expressed as the minimum distance between the center of the corresponding WS and any ligand heavy atom, R_{min}, as also reported in Table 1. Visual analysis of the complex shows that two WFP sites, WS2 and WS4, are perfectly replaced by oxygen atoms O3 of the glucose (or *N*-acetyl-Glc) and O6 of the galacatose saccharides, respectively. Moreover, water molecules in the WS show the same HB as the hydroxil groups of the ligand. Although not with a perfect match, WS1 is occupying a similar space and establishing the same HB as GalO4, while WS5 is replaced by carbonyl oxygen from the O-acetyl group, and WS7 is displaced by the Acetyl group of NacGlc itself. Finally, WS 3, which is close to the ether oxygen between both carbohydrates, appears to replace it during the ligand binding process (Figure 2). In summary the results for Gal-3 show that most of the found WS are displaced by the ligand, and the higher the WFP the higher the chance that a ligand oxygen occupies its place.

Structural and Dynamical Characterization of WS in Con-A Ligand Binding Site. The structure of ConA consists also of mostly β-sheets forming a β-sandwich. The ligand binding site lies on top-side of the small sandwich sheet. From the ligand binding region, we selected seven residues (Tyr12, Pro13, Asn14, Thr15, Asp16, Leu99, and Arg222) as those capable of establishing strong HB with water molecules for the determination of the WS. We identified 11 WS with high WFP in the ligand binding site. The computed thermodynamic parameters for the corresponding WS are shown in Table 2.

TABLE 2: Thermodynamic and Structural Parameters for Water Sites around Con-A Ligand Binding Region[a]

WS	reference atom	resid	<E_p>	<E_w>	<E_t>	<ΔE>	WFP	R_{90}	R_{min}
1	OH	Tyr12	−5.1	−15.1	–20.1	−2.7	7.7	2.9	1.52
2	OH	Tyr12	−5.9	−13.6	–19.5	−2.1	8.8	2.8	1.2
3	O	Pro13	−7.4	−13.9	–21.3	−2.3	12.8	1.9	1.42
4	ND2	Asn14	−14.6	−8.7	–23.3	−6.3	1.1	1.4	1.49
5	N	Thr15	−12.8	−8.4	–21.2	−3.7	7.2	3.9	0.9
6	OD1	Asp16	−9.5	−11.3	–20.8	−3.5	9.3	2.2	1.47
7	N	Leu99	−5.5	−15.9	–21.4	−4.0	9.8	3.1	1.07
8	N	Leu99	−10.7	−14.1	–24.8	−7.4	20.3	1.6	0.73
9	N	Arg222	−6.3	−15.1	–21.4	−4.0	16.9	2.1	0.58
10	ND2	Asn14	−16.9	−7.1	–24.0	−6.6	3.4	2.5	2.78
11	ND2	Asn14	−17.0	−5.6	22.6	−5.2	3.1	1.8	1.91

[a] Energies (E_p, E_w, E_t and ΔE) are in kcal/mol, R_{90} and R_{min} are in angstroms.

TABLE 3: Thermodynamic and Structural Parameters for Water Sites around Cyp-A Binding Region[a]

WS	reference atom	resid	<E_p>	<E_w>	<E_t>	<ΔE>	WFP	R_{90}	R_{min}
1	NE1	Trp121	−5.1	−15.9	−20.8	−3.4	11.0	2.7	0.42
2	NH2	Arg55	−13.5	−11	−24.3	−6.9	11.5	1.8	0.9
3	NH1	Arg55	−13.5	−10.6	−24.6	−7.2	10.6	1.9	1.25
4	NE2	Gln63	−6.4	−15.6	−21.9	−4.5	3.3	2.7	1.04
5	OE1	Gln111	−7.3	−12.4	−19.7	−2.3	6.0	2.4	1.80
6	O	Asn102	−11.0	−9.1	−20.1	−2.7	10.1	2.4	1.20

[a] Energies (E_p, E_w, E_t and ΔE) are in kcal/mol, R_{90} and R_{min} are in angstroms.

TABLE 4: Thermodynamic and Structural Parameters for Water Sites around CBM32 Binding Region[a]

WS	reference atom	resid	<E_p>	<E_w>	<E_t>	<ΔE>	WFP	R_{90}	R_{min}
1	ND2	Asn73	−9.0	−13.5	−22.5	−5.1	18.9	1.6	1.49
2	ND2	Asn73	−4.6	−16.4	−21.0	−3.6	2.4	4.4	2.48
3	NH1	Arg68	−13.6	−10.1	−23.7	−6.3	6.2	2.1	0.90
4	NH1	Arg68	−14.9	−9.1	−23.9	−6.5	11.5	1.7	1.24

[a] Energies (E_p, E_w, E_t and ΔE) are in kcal/mol, R_{90} and R_{min} are in angstroms.

As shown in Table 2, except WS4, WS10, and WS11, all identified WS have over five times the bulk WFP. On the basis of the crystal structure of the trimmanoside bound ConA structure (PDBid 1ONA), we can analyze which WS are replaced by the trimmanoside ligand.

The first Mannose displays several OH groups establishing HB with the protein O4 is HB to Asp208 carboxyl, O3 to Arg222NH, the O6 with Leu99NH and Tyr100NH and O5 is also close to Leu99NH. From these, O6 is perfectly replaced by WS8, O5 by WS7 establishing the same HB with the protein backbone and O3 results perfectly replaced by WS9. All three WS show a high WFP.

Mannose 2 O2 HB to Tyr12 is replaced by WS1, also HB to Tyr12 and mannose 2 O4 HB to Asn14 is replaced by WS10 and WS11. Finally, Manoside3 O3-hydroxil must displace WS2, 3, and 5. From these, WS2 and WS3 are mainly HB to Pro13 carbonyl group while WS5 is HB to Thr15 NH. WS3 and WS5 are very close together, almost adjacent to each other. Mannoside 3 O4 must displace WS6 HB to Asp16 carboxyl group. Also interestingly, WS4, which has the smaller WFP, is possibly displaced by the carbon skeleton of the first mannoside as shown by the low R_{min} to C5.

Structural and Dynamical Characterization of WS in Cyp-A Ligand Binding Site. As mentioned in the Introduction, Cyp-A is the target of the undecapeptide CsA, which binds through several HB and hydrophobic interactions. In the CsA binding site, we detected several WS with high WFP. The computed thermodynamic parameters for the corresponding WS are shown in Table 3.

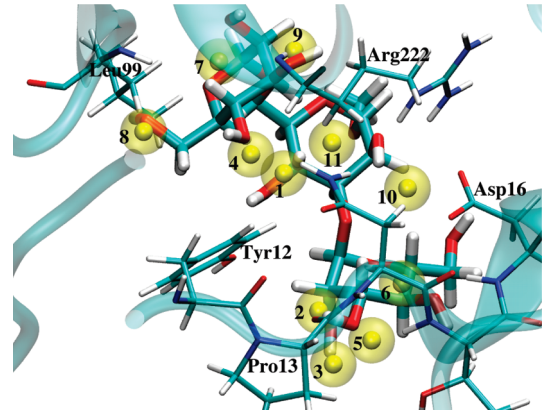

Figure 3. Structure of the Con-A trimmanoside complex superimposed on the WS positions relative to the protein structure. WS are shown as yellow spheres and numerated from 1 to 11.

As for the other proteins, a comparison of WS positions with the space occupied by the CsA ligand upon binding was performed. The WS with the highest WFP are WS 1, 2, 3, and 6. All these WS are displaced by the CsA ligand. WS1 is perfectly replaced, as shown by the low R_{min}, by the carbonyl of CsA residue 9 establishing a tight HB with Trp121NH. WS2 and WS3 are close together (almost overlapping) and both displaced by carbonyl of CsA residue 10 interacting with Arg55. WS6 is displaced by the hydrophobic side chain of CsA residue 1. WS5 also has relative high WFP and is probably displaced

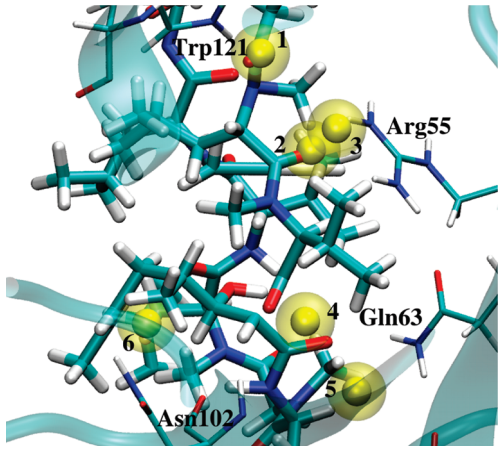

Figure 4. Structure of the Cyp-A CsA complex superimposed on the WS positions relative to the protein structure. WS are shown as yellow spheres.

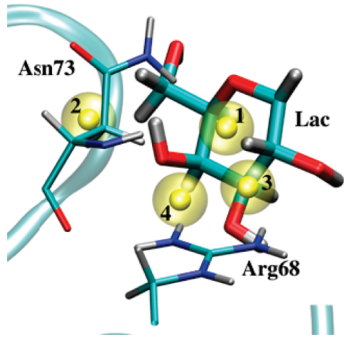

Figure 5. Structure of the CBM32-Lac complex superimposed on the WS positions relative to the protein structure. WS are shown as yellow spheres, numerated from 1 to 4.

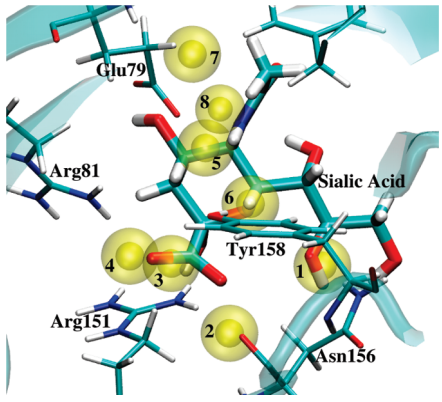

Figure 6. Structure of the CBM40 bound to sialic acid superimposed on the WS positions relative to the protein structure. WS are shown as yellow spheres, numerated from 1 to 8.

but not exactly by carbonyl of CsA residue 1. Finally WS4 must be also displaced by carbonyl of CsA residue 1.

Structural and Dynamical Characterization of WS in CBM32. CBM32 shows a β-sandwich fold with a simple carbohydrate binding site on one side of the sandwich and has a preference for galactosil terminated sugars. The main residues in the ligand recognition region are Arg68, His37, Asn73, and Trp40. Visual analysis of the trajectory allows detection of four WS, whose thermodynamic parameters are shown in Table 4.

The results show that WS1 and 4 display very high WFP, WS3 displays moderate WFP, and WS2 displays a WFP only slightly higher than the bulk solvent. Comparison with the ligand bound structure shows that WS1 is on the side of the ligand ring, in van der Waals contact (less than 2.5 Å) with all carbon atoms, clearly showing that it will be displaced by ligand. WS4 is surely replaced by LacO4, and WS3 by LacO3 as shown by the low R_{min}. Finally, WS2 is not too far from LAcC6, as shown in Figure 5.

Structural and Dynamical Characterization of WS in CBM40. As already mentioned, the CBM40 is a carbohydrate binding module, showing a β-sandwich fold with a simple binding site that recognizes preferentially sialic acid. The main residues responsible for the ligand recognition are Tyr70, Glu79, Arg81, Tyr150, Arg158, and Asn156. Analysis of the MD trajectory allows identification and characterization of 8 WS with high WFP, whose properties are shown in Table 5.

Table 5 shows that the WS1, 3 and WS6 have the highest WFP, more than 15 times that of the solvent. Comparison with the sialic acid bound structure shows that WS1 defined by the interaction with Asn156 is perfectly replaced by O8 from the acid, WS6 defined by Tyr158 is perfectly replaced by C6 of the sialic acid and WS3 defined by Arg151, is perfectly replaced by the acid group of the ligand. In a second group WS5 defined by Glu79 is replaced by Nitrogen atom of the N-acetyl moiety, and WS4 is also probably displaced by the ligand carboxylate oxygen atom, as shown by the R_{min} of less than 1.5 Å. Still further away from the ligand carboxylate but still at van der Waals contact from the oxygen appears last WS2. WS7 is also close to the ligand but too far to be displaced. Finally, WS8 must be displaced by the ligand's acetyl residue.

General Aspects of Structural and Thermodynamic Properties of the WS. Since the natural environment of proteins is in most cases aqueous, the binding process is often accompanied by significant solvent reorganization especially on the protein ligand contact surface. As already mentioned, many studies have shown the presence of water molecules strongly bound to protein reactive surfaces and support the idea that displacement of these water molecules from the ligand binding site should have a positive effect on the ligand binding free energy. On the basis of this idea, we hypothesized that the presence of tightly bound waters at proteins surfaces in the free protein may yield relevant information on the possible proteins ligand binding properties. To determine the presence of tightly bound waters, we defined the so-called WS, which correspond to regions of space at the protein surface with higher WFP than the bulk solution as determined by the water density. Our previous work on hGal-1 showed that those four WS with higher WFP tend to be occupied with oxhydril groups from the carbohydrate ligand in the protein−ligand complex.

To extend our analysis to other proteins, in this work we analyzed the thermodynamic and structural properties of additional 36 WS on five different proteins. From these WSs, 20 have R_{min} values of less than 1.5 Å to an oxygen of the ligand and 5 more an R_{min} of less than 2.0 Å, and other 5 WS are close to the ligand carbon atoms. This clearly shows that the identified WS are good predictors of the ligand binding site. Moreover, in many cases the HB between the water molecules in the WS are similar to those found between the ligand oxygen and the protein. Having characterized more than 40 WS allows statistic analysis of their thermodynamic and structural parameters.

The average and deviation of the mean thermodynamic and structural values are shown in Table 6.

Analysis from Table 6 shows, for example, that all the identified WSs display a gain in energy $<\Delta E>$ when going from the bulk solvent to the protein surface. Also interesting is the fact that total interaction energy $<E_t>$ has smaller deviation than its two components, as reflected in the sd and min/max values.

TABLE 5: Thermodynamic and Structural Parameters for Water Sites around CBM40 Binding Region

WS	reference atom	resid	$\langle E_p \rangle$	$\langle E_w \rangle$	$\langle E_t \rangle$	$\langle \Delta E \rangle$	WFP	R_{90}	R_{min}
1	ND2	Asn 156	−6.7	−13.6	−20.3	−2.9	17.8	1.8	0.76
2	NH1	Arg151	−7.0	−12.2	−19.1	−1.7	8.5	3.1	1.70
3	NH2	Arg151	−11.5	−11.6	−23.0	−5.6	13.6	1.7	0.72
4	NH2	Arg81	−11.4	−12.7	−24.0	−6.6	7.5	2.1	1.38
5	OE1	Glu79	−12.4	−7.6	−20.0	−2.6	6.5	3.5	1.31
6	OH	Tyr158	−7.3	−14.3	−21.6	−4.2	18.8	1.7	0.77
7	ND1	His92	−9.1	−11.1	−20.2	−2.8	4.7	2.8	2.41
8	OH	Tyr158	−11.4	−7.2	−18.6	−1.0	1.7	1.6	2.27

[a] Energies (E_p, E_w, E_t and ΔE) are in kcal/mol, R_{90} and R_{min} are in angstroms.

TABLE 6: Average, Deviation, Minimum, and Maximum Values for the Thermodynamic and Structural Parameters Computed for All Described Water Sites[a]

	$\langle E_p \rangle$	$\langle E_w \rangle$	$\langle E_t \rangle$	$\langle \Delta E \rangle$	WFP	R_{90}	R_{min}
mean	−10.64	−11.11	−21.54	−4.15	8.77	2.39	1.32
sd	5.03	3.64	2.59	2.64	5.06	0.78	0.62
min	−25.00	−16.4	−28.5	−11.4	2.3	1.4	0.30
max	−3.54	−2.01	−14.31	−1.6	20.3	4.4	2.78

[a] Energies (E_p, E_w, E_t and ΔE) are in kcal/mol, R_{90} and R_{min} are in Å.

This is reasonable considering that a water molecule is limited in the amount of interactions it may establish. Consequently, establishing more interactions with the protein results in less water−water interactions. This is quantitatively measured by the correlation coefficient between $\langle E_p \rangle$ and $\langle E_w \rangle$ of −0.77. This limiting fact results in a limit for the energy gain in surface association, which in this case seems to be around 11.4 kcal/mol.

The data presented here also allow determining which is the parameter that better predicts that position of a ligand in the protein−ligand complex at least from a comparative viewpoint. Assuming that R_{min} is a good estimate of how well the WS reflects the ligand occupancy in the complex, we can correlate it with the other values. Interestingly, the correlation coefficient for any of the energetic parameters, $\langle E_p \rangle$, $\langle E_t \rangle$, or $\langle \Delta E \rangle$ with R_{min} have values between 0.2 and 0.35, indicating a poor correlation. On the other hand, correlation coefficient for the probability P1 and R_{90} shows values of 0.55 and 0.612, respectively. Although the correlation is far from perfect, it shows, in accordance with our previous results, that the probability that takes into account entropic contributions (and not the interaction energy) is a better predictor of which water molecules are likely to be displaced by the ligand.

Conclusions

Using MD simulations in explicit solvent, we have defined highly solvent occupied regions of space called hydration or water sites associated to the protein hydrophilic surfaces. We have analyzed the thermodynamic and structural properties of more than 40 WS on six different proteins, namely Gal-1 (in our previous work), Gal-3, Concanavaline A, Galectin 3, Ciclophylin A, and two modules CMB40 and CBM32 of the multimodular bacterial sialidase. Our results show that the probability of finding water molecules inside the WS, $p(v)$, with respect to the bulk density is directly correlated to the likeliness of finding a heavy atom (and mostly oxygen atom) of the ligand in the protein−ligand complex. This information can be used to analyze in detail the solvation structure of the CRD and its connection to the possible protein ligand complexes, and suggests addition of OH-containing functional groups to displace water from high $p(v)$ WS to enhance drugs, specially glycomimetic-drugs, protein affinity and/or specificity.

Acknowledgment. This work was supported by Grants PICT 2007-01650 and UBA08-X625 to M.A.M. and Grants PICT 2004-25667, CONICET 2518, and UBA X074 to D.A.E. S.D.L. is grateful to CONICET for a Doctoral Fellowship. M.A.M. and D.A.E. are staff members of CONICET. Computer power was provided by the CECAR at FCEN-UBA and by Open Science Grid supported by NSF and the U.S. Department of Energy's Office for Science.

Glossary

Abbreviations

MD	molecular dynamics
CRD	carbohydrate recognition domain
WS	water site
WPF	water finding probability
HB	hydrogen bond
LacNAc	*N*-acetyllactosamine
Lac	lactose
Con-A	concanavalin-A
Gal	galectin
Cyp-A	cyclophilin-A

Supporting Information Available: Radial distribution functions $g(r)$ and angular bidimensional plot distributions for all the sites studied in this work and interaction energy convergence plot for one example. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Teeter, M. M. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 577.
(2) Schoenborn, B. P.; Garcia, A.; Knott, R. *Prog. Biophys. Mol. Biol.* **1995**, *64*, 105.
(3) Henchman, R. H.; McCammon, J. A. *Protein Sci.* **2002**, *11*, 2080.
(4) Tame, J. R. H.; Murshudov, G. N.; Dodson, E. J.; Neil, T. K.; Wilkinson, A. J. *Science* **1994**, *264*, 1578.
(5) Sleigh, S. H.; Tame, J. R. H.; Dodson, E. J.; Wilkinson, A. J. *Biochemistry* **1997**, *36*, 9747.
(6) Nasimith, J. H.; Field, R. A. *J. Biol. Chem.* **1996**, *271*, 972.
(7) Loris, R.; Maes, D.; Poortmans, F.; Wyns, L.; Bouckaert, J. *J. Biol. Chem.* **1996**, *271*, 30614.
(8) Weisner, S.; Kurian, E.; Prendergast, F. G.; Halle, B. *J. Mol. Biol.* **1999**, *286*, 233.
(9) Asensio, J. L.; Siebert, H.-C.; von der Lieth, C.-W.; Laynez, J.; Bruix, M.; Soedjanaamadja, U. M.; Beintema, J. J.; Cañada, F. J.; Gabius, H.; Jiménez-Barbero, J. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 218.
(10) Clarke, C.; Woods, R. J.; Gluska, J.; Cooper, A.; Nutley, M. A.; Boons, G. *J. Am. Chem. Soc.* **2001**, *123*, 12238.
(11) Chervenak, M. C.; Toone, E. J. *Biochemistry* **1995**, *34*, 5685.
(12) Dunitz, J. D. *Science* **1994**, *264*, 670.
(13) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.;

Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* **1994**, *263*, 380.

(14) Watson, K. A.; Mitchell, E. P.; Johnson, L. N.; Son, J. C.; Bichard, C. J. F.; Orchard, M. G.; Fleet, G. W. J.; Oikonomakos, N.; Leonidas, D. D.; Kontou, M.; Papageorgioui, A. *Biochemistry* **1994**, *33*, 5745.

(15) Conelly, P. R.; Aldape, R. A.; Wilson, K. P. *Proc. Natl. Acad. Sci. U.S.A* **1994**, *91*, 1964.

(16) Li, Z.; Lazaridis, T. *J. Phys. Chem. B* **2005**, *109*, 662.

(17) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 7683.

(18) Li, Z.; Lazaridis, T. *J. Phys. Chem. B* **2006**, *110*, 1464.

(19) Rabinovich, G. A.; Liu, F. T.; Hirashima, M.; Anderson, A. *Scand. J. Immunol.* **2007**, *66*, 143.

(20) Li, Z.; Lazaridis, T. *J. Am. Chem. Soc.* **2003**, *125*, 6636.

(21) Di Lella, S.; Marti, M. A.; Alvarez, R. M. S.; Estrin, D. A.; Díaz Ricci, J. C. *J. Phys. Chem. B* **2007**, *111*, 7360.

(22) Mayo, K. H. *Drugs* **2008**, *11*, 1.

(23) Sörme, P.; Arnoux, P.; Kahl-Knutsson, B.; Leffler, H.; Rini, J. M.; Nilsson, U. J. *J. Am. Chem. Soc.* **2005**, *127*, 1737.

(24) Öberg, C. T.; Leffler, H.; Nilsson, U. J. *J. Med. Chem.* **2008**, *51*, 2297.

(25) Tejler, J.; Skogman, F.; Leffler, H.; Nilsson, U. J. *Carbohydr. Res.* **2007**, *342*, 1869.

(26) Abel, R.; Young, T.; Farid, R.; Berne, B. J.; Friesner, R. A. *J. Am. Chem. Soc.* **2008**, *130*, 2817.

(27) Cooper, D. N. W.; Barondes, S. H. *Glycobiology* **1999**, *9*, 979.

(28) Leffler, H. *Trends Glycosci. Glycotechnol.* **1997**, *9*, 9.

(29) Lopez-Lucendo, M. F.; Solis, D.; Andre, S.; Hirabayashi, J.; Kasai, K.; Kaltner, H.; Gabius, H. J.; Romero, A. *J. Mol. Biol.* **2004**, *343*, 957.

(30) Bianchet, M.; Ahmed, H.; Vasta, G.; Amzel, L. M. *Proteins* **2000**, *40*, 378.

(31) Houzelstein, D.; Gonçalves, I. R.; Fadden, A. J.; Sidhu, S. S.; Cooper, D. N. W.; Drickamer, K.; Leffler, H.; Poirier, F. *Mol. Biol. Evol.* **2004**, *21*, 1177.

(32) Seetharaman, J.; Kanigsberg, A.; Slaaby, R.; Leffler, H.; Barondes, S. H.; Rini, J. M. *J. Biol. Chem.* **1998**, *273*, 13047.

(33) Brewer, C. F. *Glycoconj. J.* **2005**, *19*, 459.

(34) Boraston, A. B.; Ficko-Blean, E.; Healey, M. *Biochemistry* **2007**, *46*, 11352.

(35) Thobhani, S.; Ember, B.; Siriwardena, A.; Boons, G. *J. Am. Chem. Soc.* **2003**, *125*, 7154.

(36) Case, D. A.; Darden, T. A.; Cheatham, T. E. III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Gohlke, H.; Yang, L.; Tan, C.; Hornak, V.; Cui, G.; Mathews, D. H.; Steinbrecher, T.; Seetin, M. G.; Sagui, C.; Babin, V.; Ross, C. W.; Kollman, P. A. AMBER 10 University of California: San Francisco, CA, 2006.

(37) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.

(38) Case, D. A.; Cheatham, T. E.; Darden, T. A.; Gohlke, H.; Luo, R.; Merz, K. M. J.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.

(39) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(40) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.

JP901196N