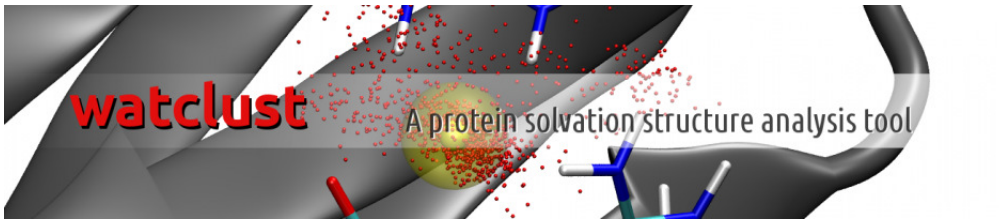

Lysozyme active site solvation structure: analysing molecular interactions

Basic Tutorial

Structural Bioinformatics Group
Author: Juan Pablo Arcon

19-FEB-2015

1 Introduction

The purpose of this tutorial is the determination of the solvation structure of the hen egg white lysozyme active site using the WATCLUST program. For this sake we just need to obtain a pdb file of the crystallized protein and perform a molecular dynamics (MD) simulation of the apo protein in explicit water. Then, we will be able to use the water clustering program WATCLUST to obtain the so called *water sites*, i.e. confined space regions, adjacent to the protein surface, where the probability of finding a water molecule is higher than in the bulk solvent.

The appropriate length of the MD simulation depends on each protein and users should carefully check convergence of the obtained parameters in each case. Since our aim is to present the WATCLUST program, we are not going to analyze the performance of the explicit water MD simulation in the present tutorial. You can choose any program for performing the MD simulation, as long as it is supported by the VMD platform (`http://www.ks.uiuc.edu/Research/vmd/`). If you do not want to run these simulations, you may use the files previously obtained by our group, which are available for download in `http://sbg.qb.fcen.uba.ar/wt/resources_tutorial_1.tar.gz`.

2 Obtaining the reference structure

The first step requires obtaining a pdb file with the crystalline structure of lysozyme. We downloaded the pdb file identified as 1LZB from the RCSB Protein Data Bank (`http://www.rcsb.org/`). This pdb will be our reference structure as well as our starting structure to perform the molecular dynamics simulation.

*** You can obtain this file from our resources (`http://sbg.qb.fcen.uba.ar/wt/resources_tutorial_1.tar.gz`) under the name ref.pdb. This file is just the 1LZB.pdb file after adding hydrogen atoms to the protein with the tLeap module of the AMBER molecular dynamics package (`http://ambermd.org/`) and removing the crystallized water molecules, while keeping the ligand. Since we are going to dismiss the hydrogen atoms to perform the alignments against this reference, it is the same to use either pdb files.***

3 Molecular dynamics simulation

We used the AMBER molecular dynamics package to prepare the starting structure and perform the MD simulation. The protocol consisted in removing the NAG ligand from the 1LZB pdb, adding the hydrogen atoms and solvating the protein with a TIP3P water box (initial structure); relaxing the structure to avoid close contacts produced after the addition of hydrogen atoms and water molecules (minimization); equilibrating the structure in order to reach the working temperature of 300K and the water density at the desired pressure of 1 atm (equilibration); and, finally, executing the production run of 20 ns at constant pressure.

*** You can go on with the tutorial using our set of snapshots (see below). If you want to replicate our simulation, the input files for AMBER are labeled as min.in for the minimization, from eq1.in to eq5.in for the equilibration steps and md.in for the production run. The parameter file is lys.parm7 and the initial structure coordinates are in lys_0.rst7.***

4 Water clustering analysis

Once the simulation is concluded, we are ready to use the WATCLUST program. The resulting MD trajectory is the source of the data that will be used in WATCLUST to analyze the solvent structure. As already mentioned, the same analysis can be done with a trajectory obtained with any MD program. We selected 2,000 snapshots from our whole simulation and group them into a binpos file (lys_md.binpos), after making the roto-translational alignment of the system for all the trajectory. You can use any file format supported by VMD to load the snapshots of your simulation.

Now, to perform the solvent structure analysis, follow the instructions below.

4.1. Open the vmd program.

4.2. Load the reference pdb:

- Go to "File" > "New Molecule".
- Choose Load files for: "New Molecule" from the drop down list.
- Click the "Browse" button, select the *ref.pdb* file and click OK.
- Choose Determine file type: "PDB" from the drop down list.

- Click the "Load" button.

4.3. Load the trajectory file:

- Go to "File" > "New Molecule".
- Choose Load files for: "New Molecule" from the drop down list.
- Click the "Browse" button, select the *lys.parm7* file and click OK.
- Choose Determine file type: "AMBER7 Parm" from the drop down list.
- Click the "Load" button.
- Choose Load files for: "lys.parm7" from the drop down list.
- Click the "Browse" button, select the *lys_md.binpos* file and click OK.
- Choose Determine file type: "Scripps binpos" from the drop down list.
- Click the Load button.

You must now have two active molecules loaded in your VMD Main window (Figure 1): the reference pdb and the trajectory of the MD simulation.

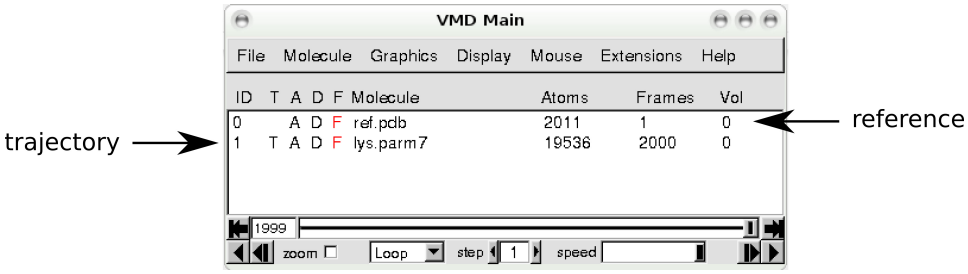

Figure 1: VMD main window.

As you may see on the VMD OpenGL Display, the trajectory is not aligned with the reference photo. Let us draw the proteins as "New Cartoon" and color them as "Secondary Structure" for better visualization (Figure 2).

We want to align the trajectory to the reference photo and then calculate the *water sites*. To do this, we will use the WATCLUST program, whose installation adds a new module available from the VMD Main window.

4.4. Open the WATCLUST module (Figure 3):

- Go to "Extensions" > "Analysis" > "WATCLUST".

In the "Selections section" we need to indicate where the trajectory is loaded

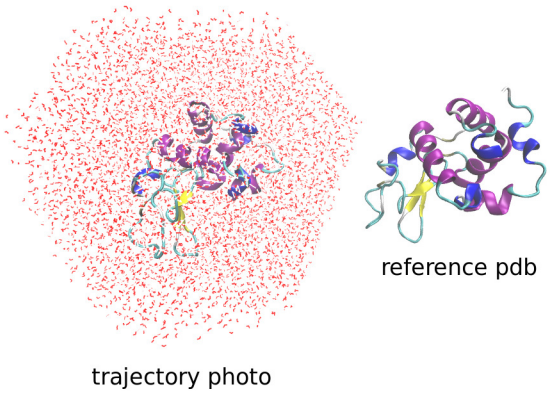

Figure 2: VMD OpenGL Display showing the solvated trajectory and the reference protein structure.

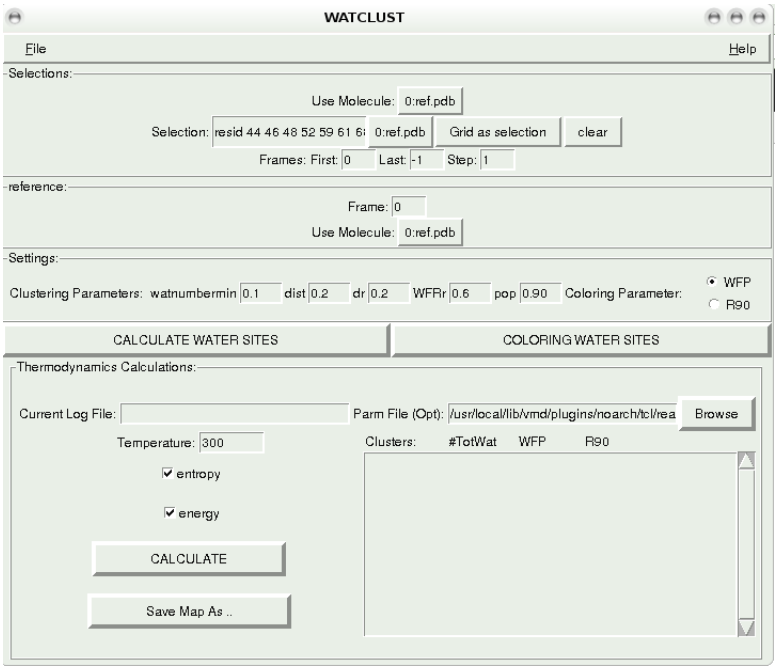

Figure 3: WATCLUST plugin window. Use the "Selections" block to define the active site, the "reference" block to choose the reference frame/pdb and the "Settings" block to adjust the parameters for the clustering algorithm.

(Figure 4).

4.5. Choose *lys.parm7* in the "Use molecule" box.

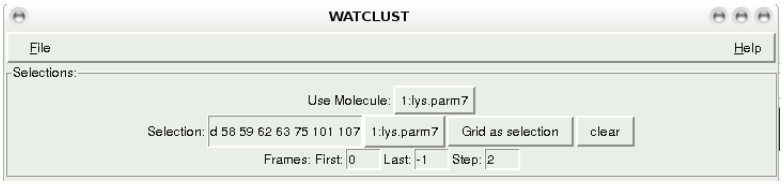

Figure 4: Selections block from the WATCLUST plugin window.

Since we want to explore the solvation of the active site, we need to select the atoms/residues corresponding to this site of interest. The WS will then be determined adjacent to these residues. In this case, we can define the active site accurately because we start from a cocrystal structure. For this sake, we look where the NAG ligand is located in the original pdb and get the residues in direct contact with it. You can make your own decision based on your experience of the site structure and its mobility during the simulation. We were quite restrictive in our selection:

4.6. Type "resid 58 59 62 63 75 101 107" in the "Selection" box. Once again, select *lys.parm7* in the box right next to the selection. For the moment, do not pay attention to the other two boxes.

For the present case, we will take into account 1,000 frames evenly distributed throughout the whole time scale of the MD simulation. Since our binpos file contains 2,000 fotos we shall choose a step of 2 frames.

4.7. Set "First: 0, Last: -1, Step: 2".

The reference photo to which the whole trajectory is going to be aligned is selected in the "reference" section (Figure 5).

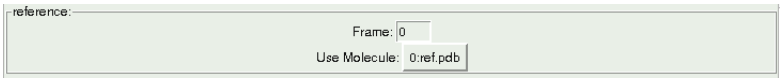

Figure 5: Reference block from the WATCLUST plugin window.

4.8. Choose "ref.pdb" in the Use Molecule box. Leave "Frame: 0" because ref.pdb has only one photo.

Now we are going to choose the parameters for the WS determination algorithm and the post processing data (Figure 6). For details check the User Guide at `http://watclust.wordpress.com/using-watclust/`.

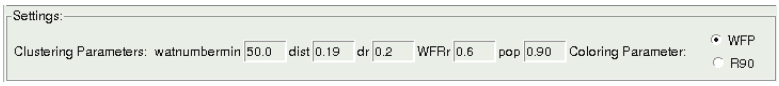

Figure 6: Settings block from the WATCLUST plugin window.

We will consider that a water site is formed when it hosts a water molecule during at least 5% of the whole simulation time (i.e. there is a water molecule inside the site in 5% of the trajectory snapshots). Thus, 50 out of our 1,000 snapshots should have a water molecule inside the water site. Then, to define this parameter,

4.9. Set "watnumbermin" = 50.

The clustering algorithm starts with the first water molecule from the first snapshot and finds all water molecules that are closer than a threshold distance, the "dist" adjustable parameter, in all subsequent snapshots. This procedure is performed iteratively (see the Supplementary information from the WATCLUST paper for details on this issue). The "dist" parameter should be small enough to avoid an excessively disperse WS, but large enough to avoid the finding of many WS which are too close to each other. Here, we will use a reasonable distance threshold of 0.19Å for exploring near water molecules.

4.10. Set "dist" = 0.19.

The last three adjustable parameters from the "Settings" block (*dr*, *WFRr* and *pop*) are used by the statistical calculations for each determined WS. The probability graphics *g(r)* and *WFP(r)* will be constructed with an increment of 0.2Å ($dr = 0.2$), the water finding probability (WFP) will be calculated at a distance of 0.6Å from the center of each water site ($WFRr = 0.6$) and the dispersion of the site will consider 90% of the water molecules in the water site ($pop = 0.9$). This last parameter corresponds to the so called R_{90} radius.

4.11. Set thus "dr" = 0.2; "WFRr" = 0.6; "pop" = 0.9.

Now, we are ready to perform the water clustering calculation.

4.12. Click on "CALCULATE WATER SITES".

Once the calculation has finished (it should take around 30 minutes on an Intel Core i5-2400 CPU @ 3.10 GHz), you should see the trajectory aligned to the reference photo in the VMD OpenGL Display according to the selection made in the "Selections" block (Figure 7).

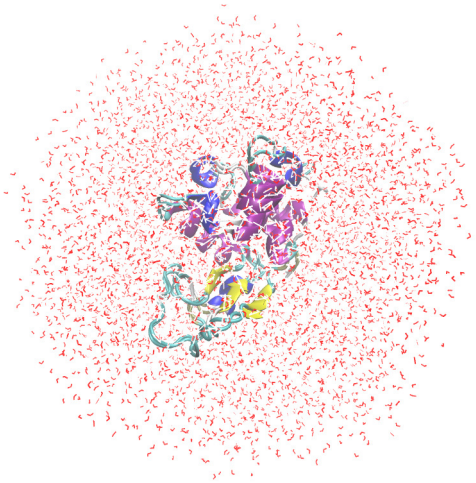

Figure 7: Trajectory aligned to the reference structure according to the selection made.

You should also have a list of the identified water sites in the "Clusters" box (Figure 8).

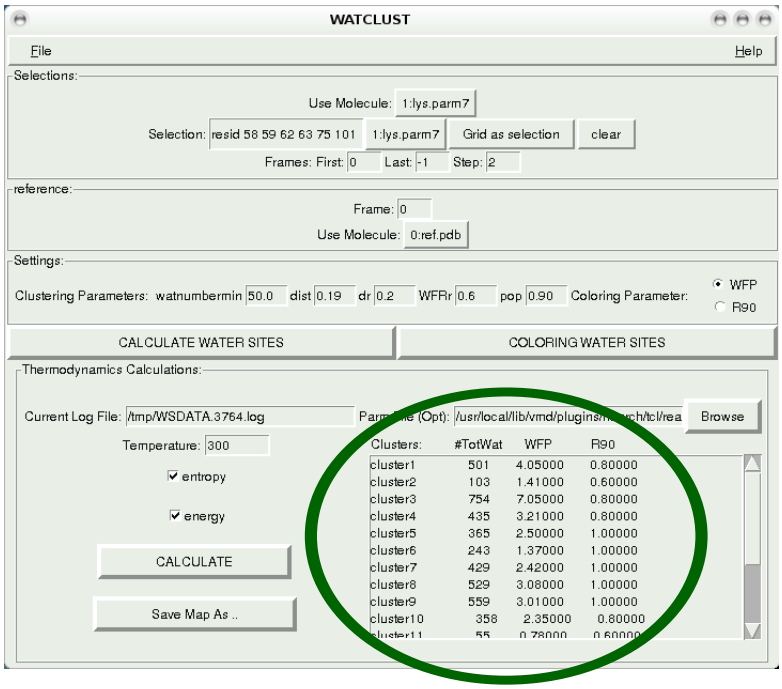

Figure 8: Water sites obtained are listed in the "Clusters" box (green circle) as *clusters* of water molecules. They are identified with a consecutive number. The box also shows the total number of water molecules in the WS (#TotWat), its WFP and the R_{90} value.

4.13. Click the "COLORING WATER SITES" button to visualize the sites obtained. Hide the trajectory representations (lys.parm7) for better visualization (Figure 9).

The different water sites found by the program are represented as VDW balls. As the "Coloring parameter" was set to WFP, the different colors of the water sites represent variations in their relative water finding probability: red for low WFP values, white for intermediate WFP and blue for high WFP. Notice that water sites were calculated only for the active site of the enzyme, as we wanted.

A new molecule named *watcent.pdb* has been added in the "VMD Main" menu after the clustering calculation. There is a single representation for each water site, as VDW balls, colored by WFP.

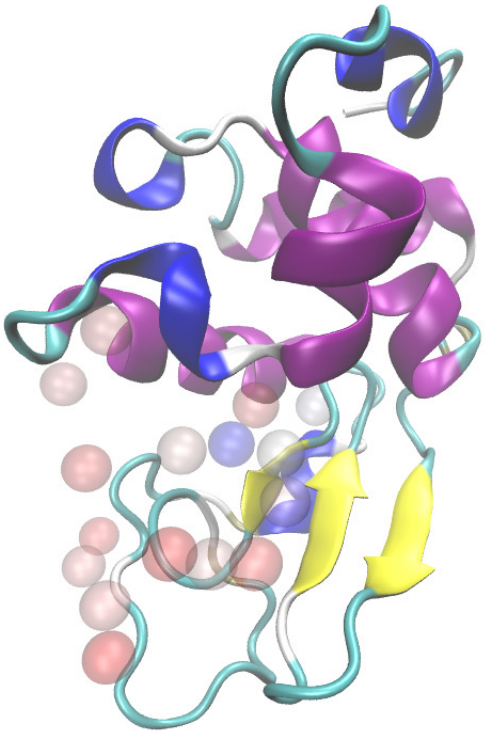

Figure 9: The reference protein structure with the obtained water sites represented as VDW spheres and colored according to their WFP.

There are plenty of capabilities at this point. We are going to select those water sites that have the larger water finding probability, identify the residues from the protein interacting with these sites and, finally, superimpose the original ligand to the sites in order to investigate if the protein-water interactions are similar to those of the protein-ligand complex.

For choosing the water sites with high WFP values we recognize from the representation those VDW balls with blue and white colors and select them.

4.14. Standing on the VMD OpenGL Display, press "1" on your keyboard and select the blue and white VDW balls with the mouse (Figure 10).

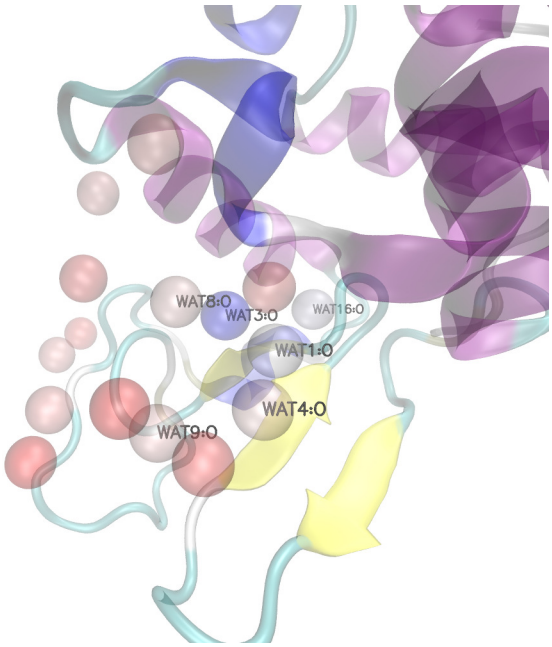

Figure 10: Selection of the water sites with higher WFP values from the OpenGL Display.

The "resid" numbers of the selected sites are: 1, 3, 4, 8, 9 and 16. The only site with an intense blue color is the one labeled with resid 3, suggesting its WFP is quite larger than the others. Checking the "Clusters" box we confirm that these are exactly the sites with the higher WFP values and that the site number 3 has the highest one (WFP = 7.1).

4.15. Lets keep only the selected sites in our visualization by selecting them in the Clusters box (Figure 11). Notice that all the sites are buried inside the active site.

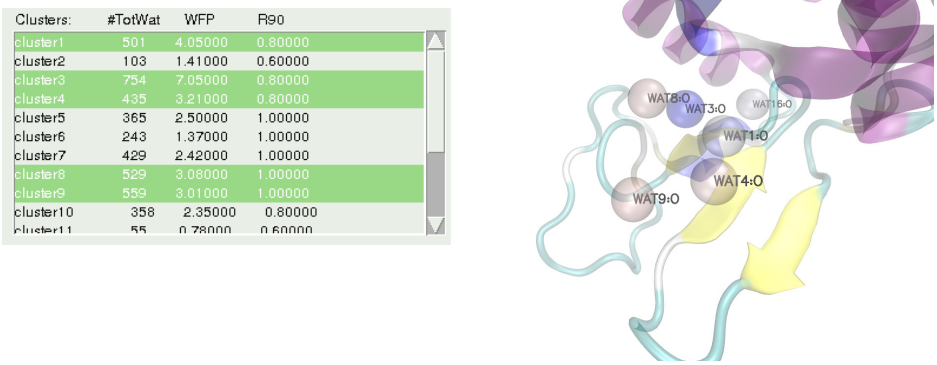

Figure 11: Selection of the water sites with higher WFP values from the "Clusters" box.

To investigate the interactions we have to analyze site by site. Lets start with water site number 3, the one with the highest WFP. Select it as before. If you select the site from the VMD OpenGL display (by pressing 1) you will obtain its coordinates from the terminal window where VMD is running.

Info) molecule id: 3
Info) trajectory frame: 0
Info) name: O
Info) type: O
Info) index: 2
Info) residue: 2
Info) resname: WAT
Info) resid: 3
Info) chain: X
Info) segname:
Info) x: 27.312000

Info) y: 29.113001
Info) z: 35.762001

We will use these coordinates to find which residues of the protein are close to the site.

4.16. In the "Graphical Representation" menu choose ref.pdb from the selected molecule. Create a new representation and type the following sentence in the "selected atoms" box:

not resname NAG and same residue as (((x-27.3)∧2 + (y-29.1)∧2 + (z-35.8)∧2) < 4∧2)

We are selecting every residue from the protein that has at least one atom inside the sphere of radius 4Å centered in the center of water site number 3 (the expression inside the parenthesis is the equation for a sphere of radius 4 and centered at x=27.3, y=29.1, z=35.8). We exclude the cocrystallized ligand (NAG) for now because we want to analyze the interaction between the water site and the protein.

4.17. Choose the Drawing method as "Licorice". The representation should display something like Figure 12.

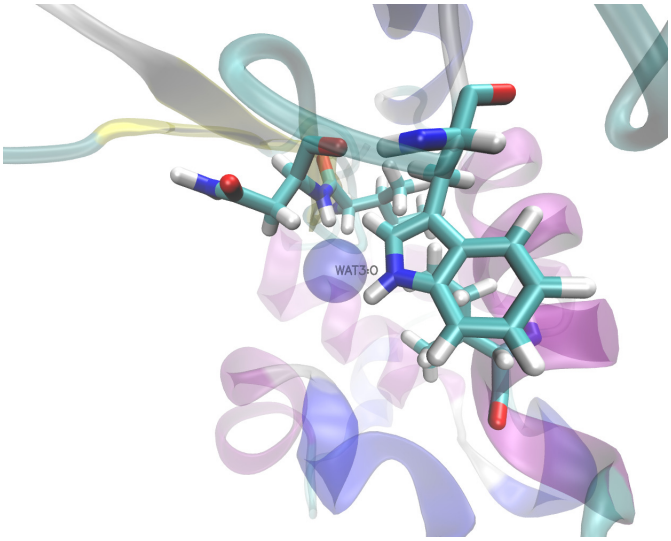

Figure 12: Protein residues surrounding WS number 3.

The residues from the protein in close contact with the water site number 3 are Ile58, Asn59, Trp63 and Ile98. Among all these, the interaction that stands out is the hydrogen bond between the site (the water molecules that form the site) and the backbone NH from Asn59 residue. Notice that two main features of the hydrogen bond interaction are satisfied: the N-H bond is directly pointing towards the site and a distance of 3.1Å between the N atom and the center of the water site is established (Figure 13).

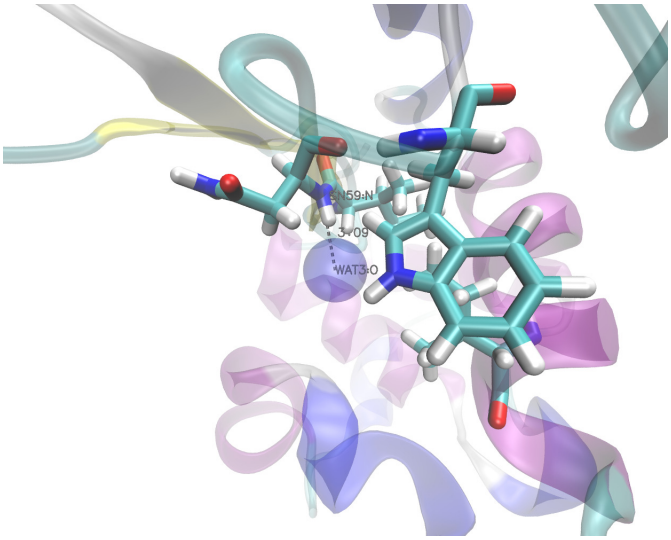

Figure 13: Hydrogen bond sampled by WS number 3.

One could imagine the indol NH from the Trp63 residue aiding in the formation of hydrogen bond interactions with the site. This is not evident looking at the reference snapshot and the mobility of the residue should be inspected through the dynamics simulation to check if this interaction is present. The other two residues do not seem to be determinant in the formation of the water site through intermolecular interactions.

This analysis should be repeated for each of the water sites of interest. Knowing the interacting residues in an active site is of great importance.

Finally, we could superimpose the original ligand (NAG) from the reference (1LZB pdb) to the water sites, in order to investigate which WS establish

the same interactions as the ligand with the protein.

4.18. Select all the sites from the "Clusters" box in the WATCLUST window and create a representation for the ligand: resname NAG as licorice in the ref.pdb representation (Figure 14).

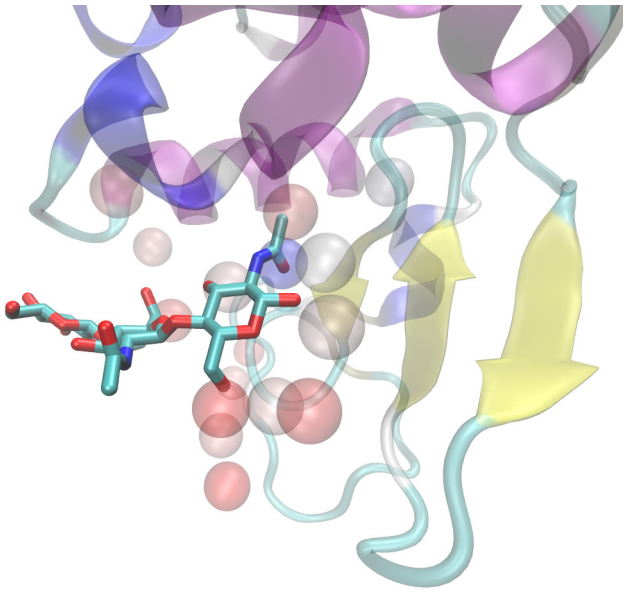

Figure 14: Reference complex structure with the ligand superimposed to the water sites.

At first glance, it is evident that some particular water sites occupy the same region as the polar groups from the ligand. There are five of this water sites (WS 1, 3, 8, 12 and 15) and three of them are among the water sites with higher WFP (WS 1, 3 and 8).

Lets now analyze which protein-ligand interactions are reproduced by the sites. The interactions from the original cocrystal (PDB ID 1LZB) are shown in Figure 15. There are six major protein-ligand interactions and all of them are hydrogen bonds. Four of these are well represented by water sites that interact with the same residue groups of the protein and occupy the same region of space as the polar interacting group of the ligand. Table 1 shows which water site mimic the protein-ligand interaction and the distance be-

tween the center of the water site and the corresponding ligand polar atom.

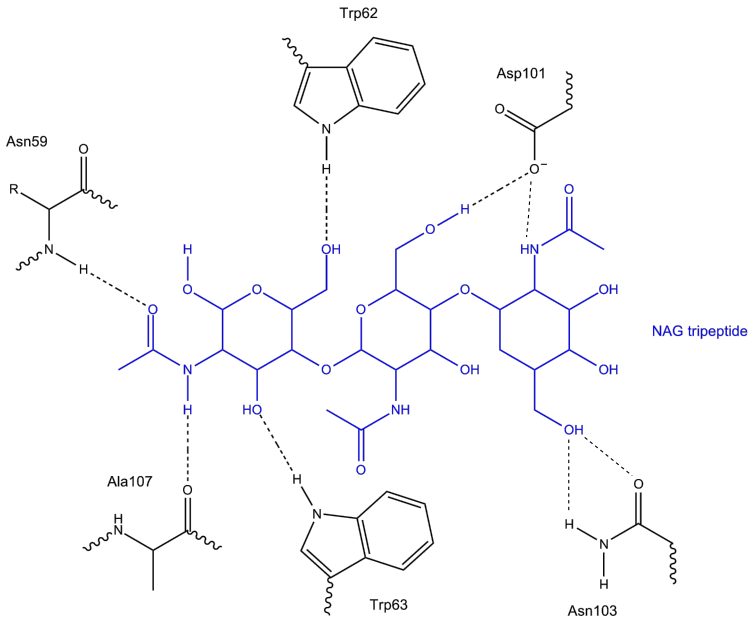

Figure 15: Scheme showing the protein-ligand interactions in the reference complex structure (PDB ID 1LZB).

Table 1: Protein-ligand interactions sampled by the different water sites

Protein residue	Ligand group	#water site	Distance (Å)
Asn59	C=O	3	0.2
Ala107	N-H	-	-
Trp63	-OH	8	1.1
Trp62	-OH	15	1.1
Asp101	-OH	12	2.7
Asn103	-OH	-	-

Note that there are not water sites representing the interactions with Ala107 nor Asn103. It happens that Ala107 is a mobile residue from a loop that is not able to fix water molecules in a certain region through hydrogen bonding along the MD simulation. As far as Asn103 is concerned, this residue is totally exposed to water and no water site is going to be established.

Finally, it is worth mentioning that there is one water site (number 14) located in a place where the ligand poses a partial apolar moiety: the methyl group of the terminal acetyl group. Always bear in mind that ligands can not optimize all of its groups to interact strongly with the protein.

Now that we understood the WATCLUST usage and how the WS are related to ligand polar groups, we may use the program to improve the available protein-ligand docking methods. This subject is treated in our next tutorial "Water sites improve docking prediction for hydrophilic drugs" (`http://sbg.qb.fcen.uba.ar/wt/tutorial_2.pdf`).