WATsite: A Hydration Site Prediction Program with PyMOL Plugin

User Guide

Ying Yang, Bingjie Hu, Markus Lill
Department of Medicinal Chemistry and Molecular Pharmacology
College of Pharmacy
Purdue University
575 Stadium Mall Drive
West Lafayette, IN 47907
Email: mlill@purdue.edu

Table of Contents

Chapter 1

1 WATsite

In this chapter we will describe the installation process of WATsite and running WATsite from the command line.

1.1 WATsite version 2.0

Changes and updates that have been made to WATsite:

- A new clustering algorithm DBSCAN has been added, and user can specify which clustering method to use.
- Four water models (SPC, SPC/E, TIP4P, TIP4P-EW) are now implemented, and user can specify which water model to use.
- Changes made to adapt newer versions of GROMACS (4.5.5, 4.6.1, and 5.0.4) and AmberTools (10, 12, and 14) (e.g., GROMACS-5.0 and later versions no longer support *genbox*, but use *gmx solvate* instead).
- A new routine aligns the whole protein then shifts the hydration sites based on the protein alignment. The hydration sites can also be aligned using the binding site residues (within 5 Å of the residues defined in the file MyBindingSite.pdb).
- The entropy value of the bulk solvent is now the average over three 100 ns bulk solvent simulations.
- WATsite can handle now large systems (e.g., with more than 99999 atoms).
- Memory allocation error in hydroentropy (if user defines a large binding site using MyBindsingSite.pdb) was fixed.
- If user changes the simulation length, the input to read the simulation trajectory and to predict hydration sites will be changed now automatically.

1.2 Citing WATsite

When using WATsite please cite the following references:

1) Hu, B.; Lill, M.A., Protein Pharmacophore Selection using Hydration-site Analysis. *J Chem Inf Model* **2012**, 52, 1046-60.
2) Hu, B.; Lill, M. A., Watsite: Hydration Site Prediction Program with Pymol Interface. *J Comput Chem* **2014**, 35, 1255-60.
3) Yang, Y.; Hu, B.; Lill, M.A., Analysis of Factors Influencing Hydration Site Prediction based on Molecular Dynamics Simulations. *J Chem Inf Model* **2014**, 54, 2987-95.
4) Yang, Y.; Lill, M.A., Dissecting the Influence of Protein Flexibility on the Location and Thermodynamic Profile of Explicit Water Molecules in Protein-Ligand Binding. *J. Chem. Theory Comput.* **2016**, 12, 4578-92.

1.3 Installation of WATsite program

The package WATsite.tar.gz needs to be downloaded and extracted using

tar –zxf WATsite.tar.gz

The location of WATsite will be noted as

$WATSITEHOME

in the following.

1.3.1 Prerequisites

To utilize WATsite, the following programs have to be installed:

1.3.1.1 GROMACS package

WATsite performs the molecular dynamics simulations using Gromacs (tested on version 4.5.5, 4.6.1 and 5.0.4) (http://www.gromacs.org). Install Gromacs using the default location /usr/local/gromacs. The gromacs directory will be noted as

$GROMACS_HOME

in the following.

1.3.1.2 AmberTools

WATsite needs tleap, antechamber and parmchk to prepare the ligands and proteins. We have tested WATsite with AmberTools 10, AmberTools 12, and AmberTools 14 (http://ambermd.org).

First, install AmberTools using the default location /usr/local/amber14. The AmberTools directory will be noted as

$AMBERHOME

in the following.

1.3.1.3 Reduce

WATsite uses Reduce [1] to optimize the hydrogen bond network in proteins. Reduce can be downloaded from http://kinemage.biochem.duke.edu/software/reduce.php. Install Reduce to the default location /usr/local/reduce. The directory that contains the executable "reduce" will be noted as

$reduce_exe_dir

in the following.

1.3.1.4 PyMOL

Pymol will be used to define the protein binding site as well as the shifting of hydration sites. Install Pymol to the default location /usr/local/pymol. The PyMOL executable pymol will be noted as

$PYMOL_EXE

in the following.

1.3.2 Compiling WATsite

Compile the source code using the *makefile*, go to $WATSITEHOME and type:

make –f makefile

1.4 Running WATsite

Note: For running WATsite using the PyMOL plugin, you can directly move to Chapter 2.

We will demonstrate the usage of WATsite using the example 1ela[2] (elastase enzyme) under

$WATSITEHOME/example/

1.4.1 Setup the RunScript

Go to the directory $WATSITEHOME/bin and open file *RunScript* in the editor. Change the environmental variables in the *RunScript* according to your system installation:

```
#!/bin/tcsh
setenv WATSITEHOME /home/yang570/WATsite2.0
setenv AMBERHOME /usr/local/amber14
setenv GROMACS_HOME /usr/local/gromacs
setenv PYMOL_EXE /usr/local/pymol/pymol
setenv reduce_exe_dir /usr/local/reduce
setenv withligand 0
setenv ligcharge 0
setenv watermodel 'spc'
setenv ncpus 8

$WATSITEHOME/bin/WATsite.sh
```

Specify the correct location for $WATSITEHOME, $AMBERHOME, $GROMACS_HOME, $reduce_exe_dir and $PYMOL_EXE. The $withligand variable defines whether the WATsite simulation is with bound ligand in the binding site or not. $withligand = 0 means simulation without a ligand, whereas $withligand = 1 means simulation with a ligand. When $withligand is set to 1, the net charge of the ligand $ligcharge also needs to be specified.

Currently four water models have been implemented in WATsite. User can choose from 'spc', 'spce', 'tip4p', 'tip4pew', and make changes to the variable $watermodel accordingly.

If user wants to use multiple nodes using MPI, change the $ncpus variable to the number of CPUs that will be used in the WATsite simulation. If $ncpus = 1, then the standard mdrun will be performed instead of mdrun_mpi.

1.4.2 Change the length of the GROMACS MD simulation

To increase the number of snapshots used for performing the hydration site analysis the length of the MD simulation can be extended. The mdp files controlling the length of the GROMACS simulation are located under $WATSITEHOME/mdp_files/posre. The detailed explanation of mdp options can be found on the GROMCACS website: http://manual.gromacs.org/online/mdp_opt.html

Briefly, to change the number of time steps of the MD simulation, the value of nsteps on the 15th line in *md.mdp* file has to be changed. The default value of nsteps is 2,000,000 that performs a 4 ns simulation and generates a trajectory that consists of 4,000 snapshots. A 4 ns simulation was demonstrated to be appropriate for converged hydration site analysis in a previous publication [3]. User also needs to change the total number of snapshots in $WATSITEHOME/dat/HydroEntropy_mod.bcf. The number of frames in the trajectory file has to be changed accodingly in section "$Input_MDsnapshots:" (line 7).

1.4.3 Change the water model for MD simulation

If user changes the water model for MD simulation, the parameter water model (on line 8) of $WATSITEHOME/dat/HydroEntropy_mod.bcf needs to be set correctly. If user set $watermodel = 'spc' or 'spce' in the RunScript, the parameter needs to be 0; whereas, parameter needs to be 1 if $watermodel is set to be 'tip4p' or 'tip4pew'

1.4.4 Change the clustering method of hydration site identification

There are two options in WATsite to identify hydration sites using different clustering algorithms: Quality threshold (QT) clustering algorithm and density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm [4]. QT clustering method is superior in identifying hydration sites with a large occupancy value during the simulation, whereas DBSCAN clustering method is advantageous in identifying more hydration sites with lower occupancy throughout the simulation.

To specify the clustering algorithm used for hydration site identification please specify the chosen option in section "$Cluster_method:" (line 23) in the file *HydroEntropy_mod.bcf* under $WATSITEHOME/dat. Using the default value of "1" hydration sites are identified using the DBSCAN clustering algorithm. For using QT clustering, please change this value to "2".

1.4.5 Setup protein system

The protein structure of the example system in pdb format and the ligand structure in both pdb and mol2 format are located in the directory $WATSITEHOME/example/1ela. Hydration sites can be predicted in the absence or presence of the ligand in the protein binding site. In both cases, the ligand structure in pdb format (MyBindingSite.pdb) is required to define the active site while the mol2 format of the ligand structure is only required when predicting hydration sites with the ligand present in the protein binding site.

1.4.6 Hydration site prediction without ligand

Make a sub-directory apo_hs in the folder $WATSITEHOME/example/1ela.

mkdir apo_hs

Copy 1ela_protein.pdb into apo_hs as MyProtein.pdb:

cp 1ela_protein.pdb apo_hs/MyProtein.pdb

We will use the co-crystallized ligand 1ela_ligand.pdb to define the protein binding site. Copy 1ela_ligand.pdb into apo_hs as MyBindingSite.pdb:

cp 1ela_ligand.pdb apo_hs/MyBindingSite.pdb

At this point, the RunScript should be edited based on the user's system installation, and the variable $withligand is set to 0 since hydration sites will be predicted without the ligand (see section 1.4.1 for details).

Copy RunScript into apo_hs, and start job by typing:

```
cd apo_hs
nohup ./RunScript
```

1.4.7 Hydration site prediction with ligand

Make a sub-directory holo_hs under 1ela:

```
mkdir holo_hs
```

Copy 1ela_protein.pdb into holo_hs as MyProtein.pdb and 1ela_ligand.pdb into holo_hs as MyBindingSite.pdb:

```
cp 1ela_protein.pdb holo_hs/MyProtein.pdb
cp 1ela_ligand.pdb holo_hs/MyBindingSite.pdb
```

To predict the hydration sites with the presence of the binding site ligand, copy 1ela_ligand.mol2 into holo_hs as MyLigand.mol2:

```
cp 1ela_ligand.mol2 holo_hs/MyLigand.mol2
```

For demonstration purposes, we used the co-crystallized ligand for the simulation. However, the ligand can be any molecule that is placed in the protein binding site.

Finally, copy RunScript to holo_hs, change the variable $withligand to 1, and change the variable $ligcharge to 1 as the net charge of this ligand is positive 1. Start job by typing:

```
cd holo_hs
nohup ./RunScript
```

1.5 Analyze WATsite Results

Once the WATsite simulations are finished, the output files *HydrationSites.mol2* and *cluster.egy* will be located in the *apo_hs* and *holo_hs* simulation folders. As we mentioned, the new routine is able to align the whole protein or the selected binding site residues, and then shift the hydration sites based on the alignment. *HydrationSites.mol2* and *HydrationSites_byBS.mol2* store the information about the location of the hydration sites predicted by WATsite using alignment of the whole protein and the binding site residues respectively. The file *cluster.egy* contains the predicted entropy and enthalpy values for each hydration site. Please refer to section 2.3 for how to analyze and visualize these results files using the WATsite PyMOL plugin.

Chapter 2

2 WATsite PyMOL Plugin

In this chapter we will describe the process of setting up and using the WATsite PyMOL plugin for hydration site analysis.

2.1 Download and Installation

Note: WATsite and its prerequisites should be installed following instructions in section 1.2.

The WATsite PyMOL plugin provides an interface to prepare WATsite simulations, analyze and visualize WATsite results through the GUI PyMOL. The current version of our plugin is designed for Linux OS.

Download PyMOL from http://pymol.org and install it on your computer (cf. section 1.3.1.4). The location of PyMOL will be noted as

$PyMOL

in the following.

The file *WATsite_Settings.txt* located in $WATSITEHOME needs to be copied to your home directory.

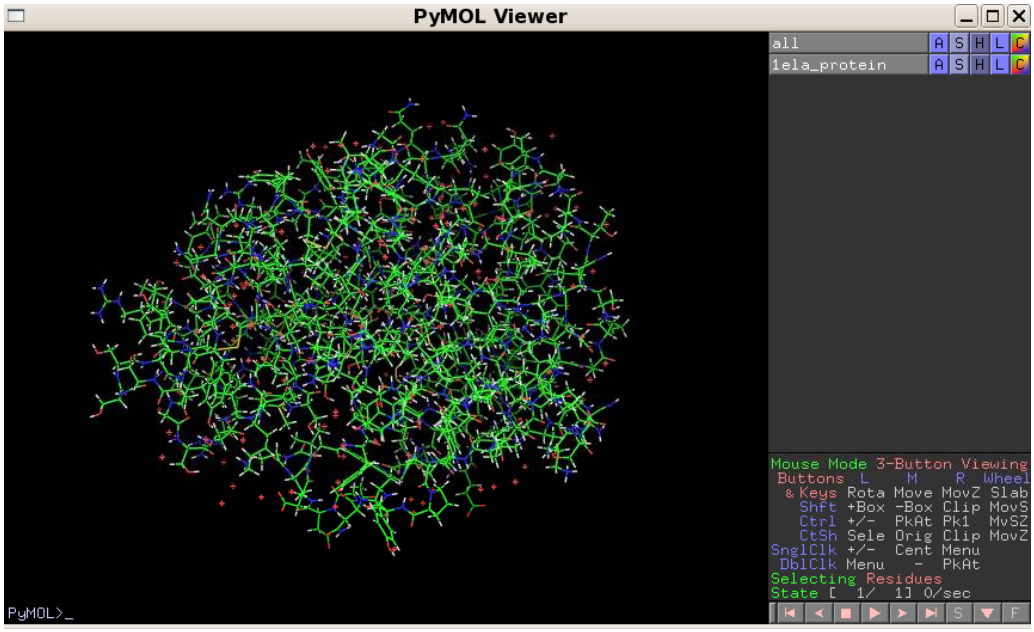
The plugin (*WATsite.py*) located under $WATSITEHOME needs to be placed in $PyMOL/modules/pmg_tk/startup/ if you installed the binary version of PyMOL or in $PyMOL/lib64/python/pmg_tk/startup if you compiled PyMOL from source on a 64bit version hardware. $PyMOL is the top-directory containing your local PyMOL installation.

2.2 Prepare the System for Hydration Site Analysis under PyMOL

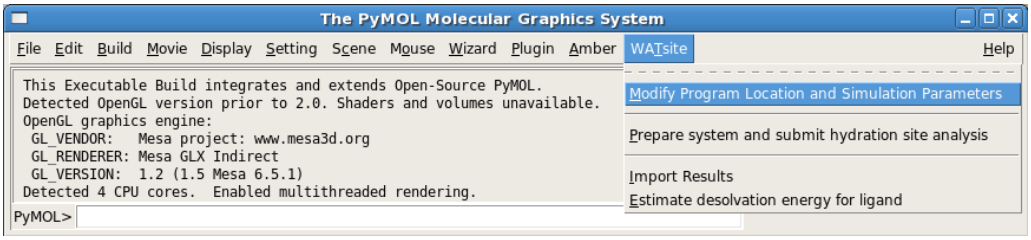
We will demonstrate the use of the WATsite PyMOL plugin using the protein system 1ela as example. Go the directory $ WATSITEHOME/example/1ela, open PyMOL loading 1ela_protein.pdb:

pymol 1ela_protein.pdb

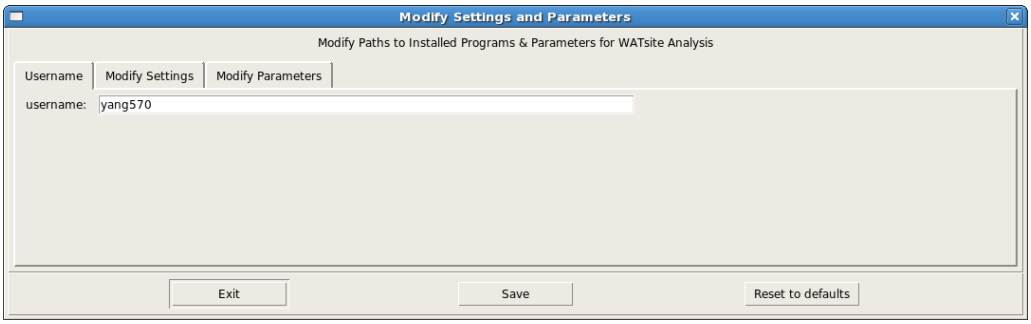
the following window will appear:



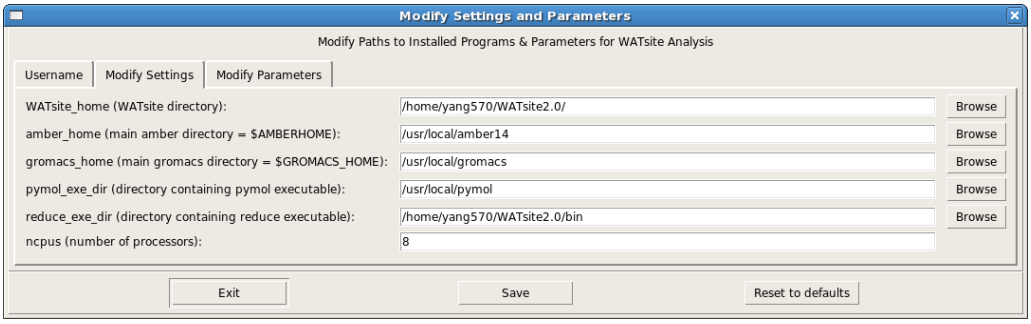
1. Click the "Modify Program Location and Simulation Parameters" submenu under WATsite menu,



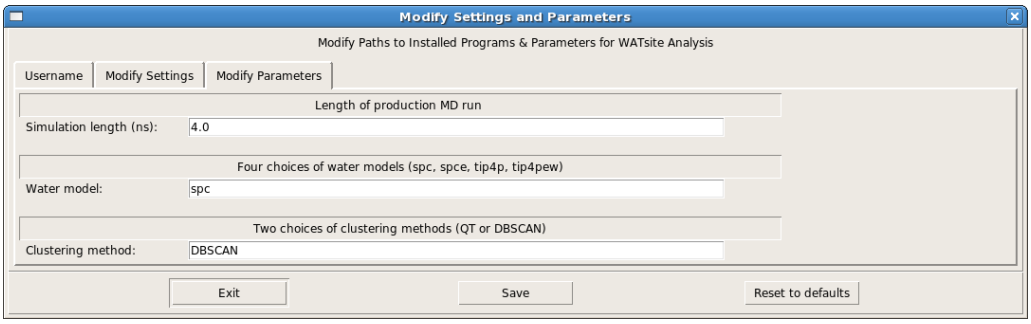
The following window will appear:



2. Click the "Modify Settings" tab, and specify the correct location for "WATsite_home" to $WATSITEHOME. Similarly, specify the correct location for Amber, Gromacs, PyMOL, and Reduce under the variables "amber_home", "gromacs_home", "pymol_exe_dir" and "reduce_exe_dir". Finally, specify the number of CPUs provided for running WATsite simulations.

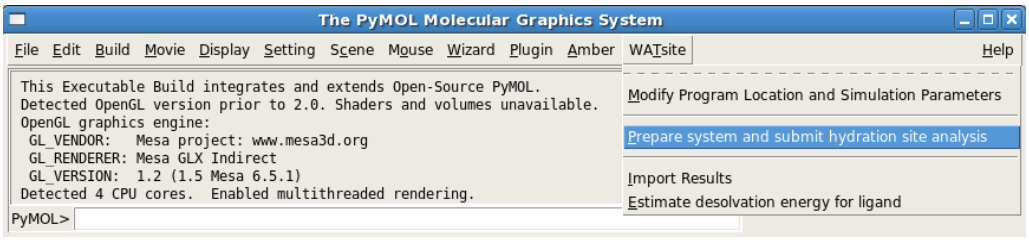


3. Click the "Modify Parameters" tab. The length of MD simulation can be changed (a minimum of 4 ns simulation length is recommended). User can choose from four water models: spc, spce, tip4p, tip4pew, and two clustering methods: QT or DBSCAN.

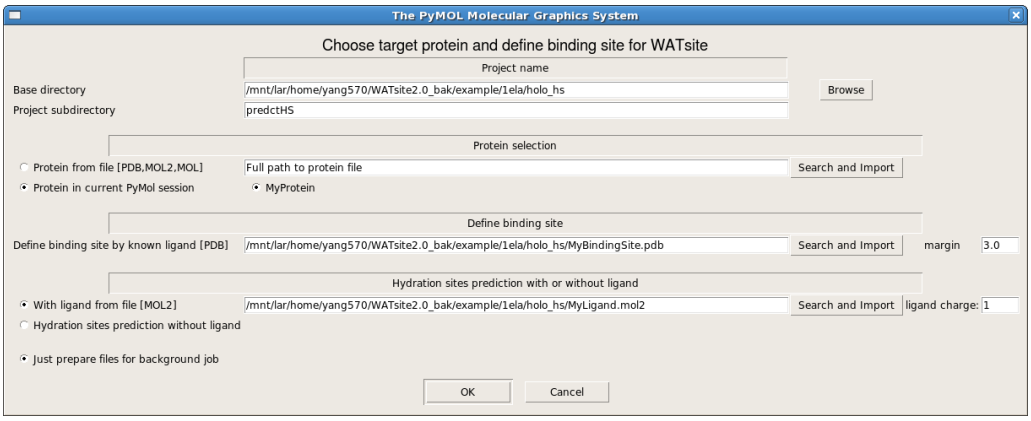


4. Click "Save" to save changes.

5. Clicking the "Prepare system and submit hydration site analysis" submenu under the WATsite menu,

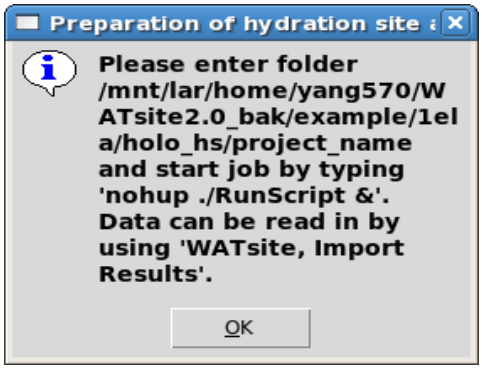


the following window will appear:



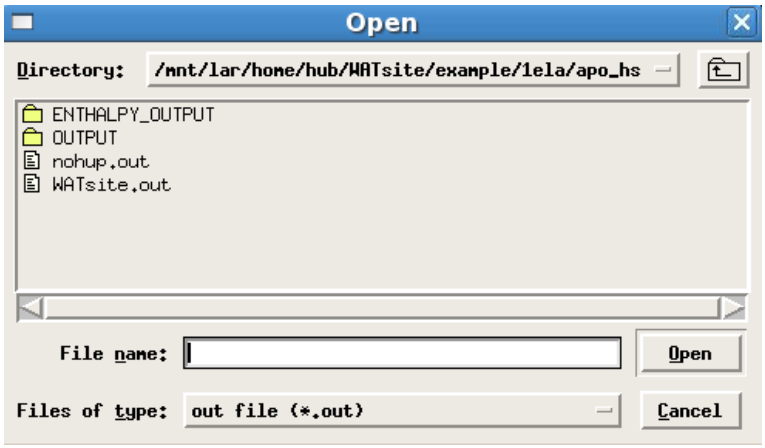
6. Make changes to the project name, select protein and binding site structure. The protein structure can be specified using a file or a structure already displayed in the current PyMOL session. To define the protein binding site, a ligand molecule positioned within the binding site needs to be defined, as well as a margin (in Å) for defining the binding pocket enclosing the ligand. The ligand provided in this step is only intended for defining the binding site. It is not included in the MD simulation and the subsequent hydration site identification process. Thus, the user can also construct a "pseudo-ligand" using the binding site residues to define the binding site. User can also select to predict hydration sites with or without the presence of ligand in the binding site. If the user intends to predict hydration sites at the interface between the protein and the ligand, the file location of the bound ligand and its net charge has to be specified within the PyMOL plugin. The specified ligand will then be included in the MD simulation and the following hydration site identification process. In the current version, no docking service for the user-specified ligand is provided. Thus, the provided ligand conformation needs to be a meaningful binding pose for the protein. If ligand is present, the net charge of the ligand has to be specified.

7. Press "OK" to prepare the files which yields the dialog, showing that files are prepared and the commands to run the program.

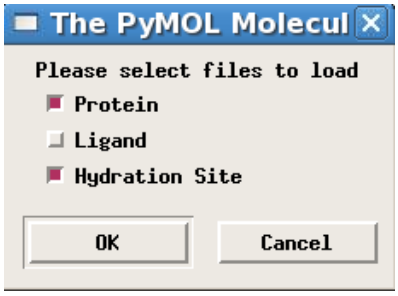


2.3 Analyze WATsite Results

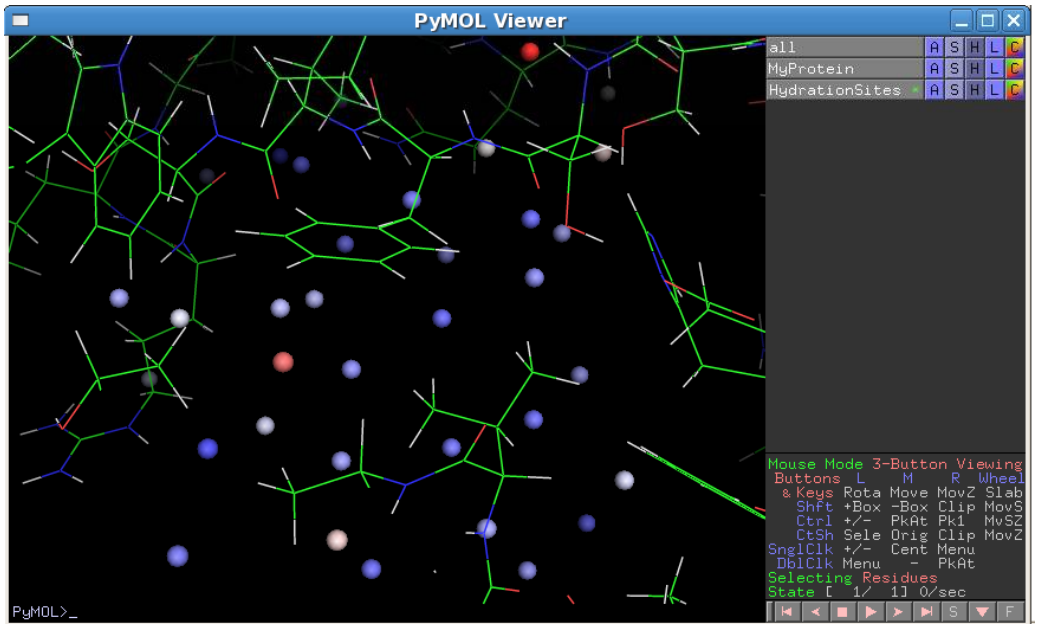
Go to directory $WATSITEHOME/example/1ela/apo_hs and open a new PyMOL session. Click the "Import Results" command under WATsite menu. The following window will appear:



Select the file *WATsite.out* and click "open". The following window will appear:

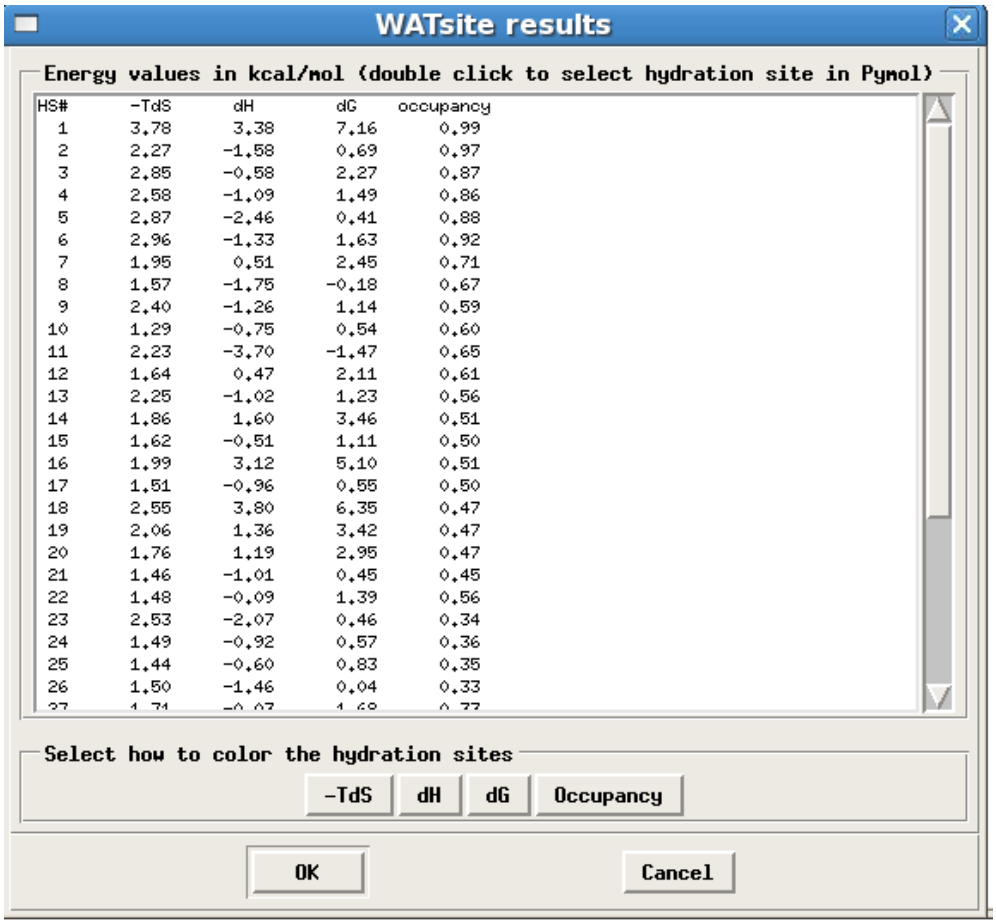


For hydration site prediction without the ligand, select only the "Protein" and the "Hydration Site" to load. The hydration sites will be loaded into PyMOL viewer and showed as non-bonded spheres:



By default, the hydration site is color coded based on their ΔG values, in a blue-to-red spectrum where blue indicates relatively low ΔG values and red indicates relatively high ΔG values.

The energy value of individual hydration site will also show up in the window below:



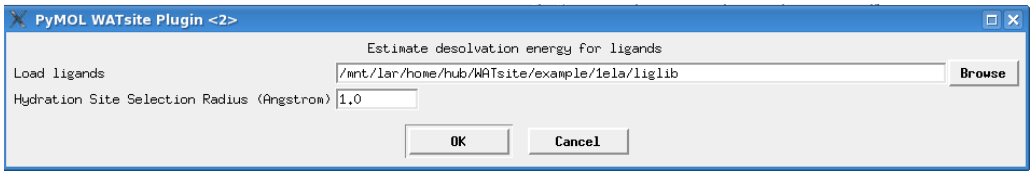
Double click the hydration site on the above table will select the hydration site in the PyMOL viewer. Additionally, you can select to color code the hydration sites by $-T\Delta S$, ΔH, ΔG or occupancy in the above window.

2.4 Predict Desolvation Energy upon Ligand Binding

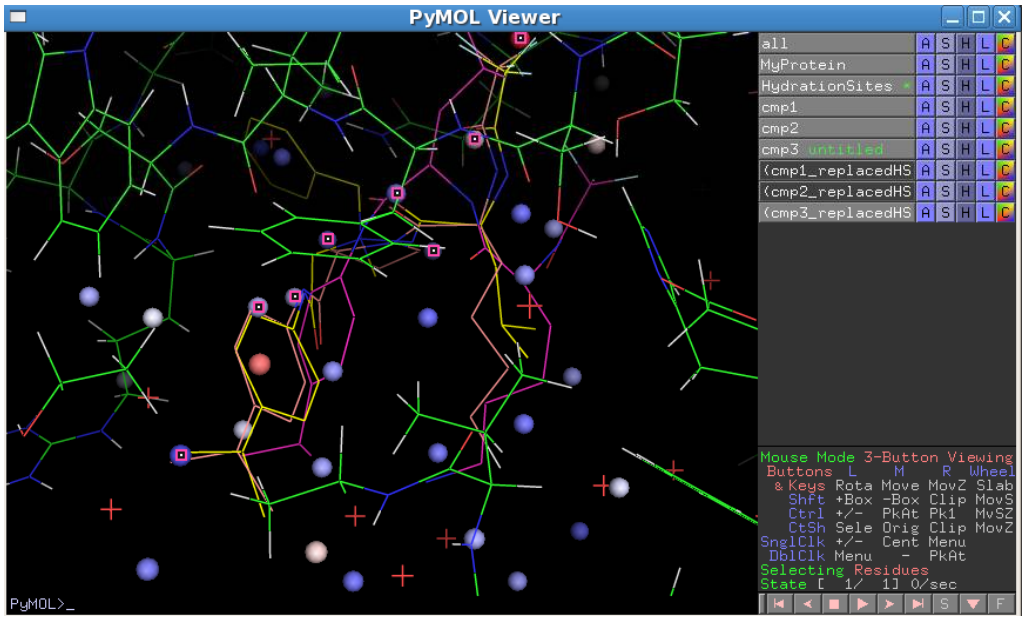
A major application of WATsite is to use the predicted hydration sites to estimate the desolvation free energies involved in replacing water molecules in the protein binding site upon ligand binding. For this purpose, a ligand library can be imported into PyMOL using the plugin and the desolvation free energy associated with replacing the binding site water molecules with each ligand is computed. For this tutorial a ligand library with three ligands is provided under \$WATSITEHOME/example/1ela/liglib. The ligands are known to bind with the enzyme elastase and have been pre-aligned into the 1ela binding site.

Import the WATsite results as described in the previous section (2.3) and press the OK button on the WATsite results dialog (see previous figure). Then select "Estimate

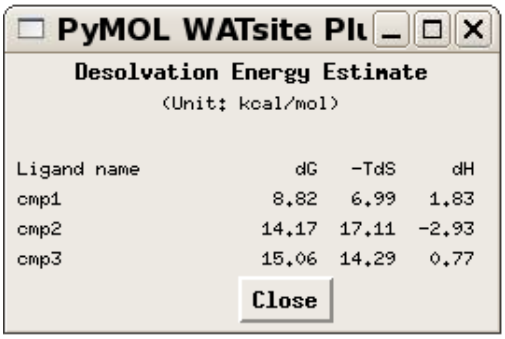
desolvation energy for ligand" under the WATsite menu. The following dialog will appear:



Provide the directory to the ligands as "$WATSITEHOME/example/1ela/liglib". Specify the distance cutoff within which the hydration sites overlapping with the ligand heavy atoms will be selected (default = 1Å). The ligands will be loaded into the PyMOL viewer:



The WATsite plugin will identify the hydration sites that are within the user-specified distance (1Å by default) to any of the ligand's heavy atoms and add up the free energies associated with the selected hydration sites. The hydration sites that were replaced by each individual ligand will also be shown as named selections in the PyMOL viewer (cmp1_replacedHS, cmp2_replacedHS, etc.). The estimated desolvation free energies will be displayed for all ligands in a separate window, for example:


PyMOL WATsite Plu
Desolvation Energy Estimate
(Unit: kcal/mol)
Ligand name dG -TdS dH
cmp1 8.82 6.99 1.83
cmp2 14.17 17.11 -2.93
cmp3 15.06 14.29 0.77
Close

3 References

1. Word, J., et al., *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1*. Journal of molecular biology, 1999. **285**(4): p. 1735-1747.
2. Mattos, C., et al., *Analogous inhibitors of elastase do not always bind analogously*. Nature Structural and Molecular Biology, 1994. **1**(1): p. 55-58.
3. Yang, Y., B. Hu, and M.A. Lill, *Analysis of Factors Influencing Hydration Site Prediction based on Molecular Dynamics Simulations*. J Chem Inf Model, 2014. **54**(10): p. 2987-95.
4. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei., et al., *A density-based algorithm for discovering clusters in large spatial databases with noise*. AAAI Press. 1996. pp. 226–231.
5. Hu, B. and M.A. Lill, *Protein Pharmacophore Selection using Hydration-site Analysis*. J Chem Inf Model, 2012. **52**(4): p. 1046-60.
6. Hu, B. and M.A. Lill, *Watsite: Hydration Site Prediction Program with Pymol Interface*. J Comput Chem, 2014. **35**(16): p. 1255-60.