Supporting information

for

ProBiS H2O MD Approach for Identification of Conserved Water Sites in Protein Structures for Drug Design‡

Marko Jukič [a,b] , Janez Konc [a,c], Dušanka Janežič [b,*] and Urban Bren [a,b,*]

[a] University of Maribor, Faculty of Chemistry and Chemical Engineering, Laboratory of Physical Chemistry and Chemical Thermodynamics, Smetanova ulica 17, SI-2000 Maribor, Slovenia.

[b] University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljaška 8, SI-6000 Koper, Slovenia.

[c] National Institute of Chemistry, Hajdrihova 19, SI-1000, Ljubljana, Slovenia.

KEYWORDS: ProBiS, conserved water molecules, molecular dynamics, water clustering, in silico drug design, water site identification.

‡This paper is dedicated to the memory of Professor Maurizio Botta, an excellent researcher and an exceptional colleague with whom we had long-standing fruitful collaboration.

WATER ANALYSIS SOFTWARE REVIEW

Approaches can be roughly divided into four groups as is elaborated below, namely calculations with explicit water models (molecular dynamics - MD, RISM), calculations with implicit water models (probe-based), experimental methods (ProBiS H2O) and descriptor-based methods.

The first approach is MD using explicit water models with subsequent trajectory analysis.[1] An example is WaterMap (WScore; description of ligand and protein desolvation) from Schrödinger.[2] WaterMap uses MD with explicit solvent and restrained protein to model the solvation effects. WaterMap then clusters and characterises each identified hydration site energetically using inhomogeneous fluid solvation theory. Using this approach, it was described that the specific high-energy hydration sites could be used to identify potentially interesting binding sites and druggability of targets.[3] Another statistical mechanics-based approach to identify and energetically evaluate discrete hydration sites has been reported by Cui *et al*.[4] SPAM software utilizes explicit solvent MD simulations to capture discrete hydration sites and provides energetic evaluation from the distribution of interaction energies between water and macromolecular environments. In order to alleviate sampling problems in explicit solvent MD calculations, an alternative bordering on the implicit solvent representations was reported by Sindhikara *et al*.[5] Sindhikara *et al*. also published an analysis of biomolecular solvation by three-dimensional reference interaction site model (3D-RISM) theory.[6] 3D-RISM calculation on a static solute is followed by post-processing to identify individual solvation sites (*Placevent* algorithm) with final geometric and thermodynamic evaluation.[7] In order to tackle the sampling problems, Monte Carlo based methods have also been applied in identification of conserved water locations, e.g. Grand Canonical Monte Carlo[8] and Just Add Water moleculeS or JAWS.[9]

The second approach can be traced to Dowser software that performs cavity detection on protein structures, hydrates all the pockets and selects preferable water locations on the basis of energy cut-offs.[10] A more recent method is represented by SZMAP/GAMEPLAN by OpenEye Scientific Software Inc.[11] Herein analysis of sites near the protein or ligand surface is

performed using semi-continuum solvation theory where a single explicit water probe is translated through space and interactions calculated to find 3D grid maps that try to address how water displacement in specific region influences the binding affinity. The software ultimately addresses the understanding of water structure in the immediate environment of the examined ligand. There is also WaterFLAP from Molecular Discovery Inc. based on Fingerprints for Ligands And Proteins – FLAP algorithm.[12] The FLAP algorithm analyses a macromolecular space by using GRID molecular interaction fields obtained from spatial translation of a series of chemical probes. 3D-field data is then condensed to pharmacophoric points that are further enumerated and processed. Waters are then assessed as to their structural, displaceable, or bulk character. A protocol of docking a water molecule into a site of interest, filtering and clustering was reported as WaterDock and data used to produce the improved DOWSER++ software.[13,14]

The third approach the representative of which is the ProBiS H2O method and PyMOL plugin uses experimental structures to assemble water location data. Water location data can thus be collected from accessible and curated databases such as RCSB PDB, wwPDB or PDBj.[15-17] Deposited RCSB PDB data was also used in development of DRoP.[18] DRoP collects a set of similar protein structures on the basis of user selection, and then a scoring/filtering step enumerates water-protein contacts with subsequent global alignment of all structures with Ce-align algorithm. Water data can also be found at The Cambridge Crystallographic Data Centre (CCDC) with their IsoStar[19] and SuperStar databases of small-molecule crystal structures and intermolecular interactions respectively.[20] Namely, the IsoStar database of small-molecule crystal structures was used to develop AcquaAlta approach.[21] Data on geometry of water interactions towards generic functional groups was collected to perform *ab initio* interaction energy calculations to construct an empiric hydration propensity scale. PyWATER is another PyMOL plugin for identification of PDB structures with similar sequences.[22, 23] Plugin follows the standard protocol where water filtering to eliminate waters with high normalized B-factors is performed before a sequence-independent structure-based superimposition. Final hierarchical clustering with calculated degree of conservation between structures affords discrete conserved water location data.

As the fourth approach, descriptor based approaches have also been developed. Better than identifying conservation trends, descriptor based approaches evaluate specific waters obtained from experimental data. WaterScore was developed to establish a statistical correlation between structural properties of water molecules (B-factor, the solvent-contact surface area, total hydrogen bond energy and the number of protein atomic contacts) in the *apo*- and *holo*- protein complexes. It was found, that on the basis of the used descriptors, bound and displaceable waters could be discriminated.[24] Similarly, Consolv program employed a hybrid k-nearest neighbours/genetic algorithm classifier to predict conserved water molecules by examining their environment (B-factor, the number of hydrogen bonds between the water molecule and protein, density and hydrophilicity of neighbouring protein atoms) of water molecules in experimentally obtained crystal structures.[25] Recently, a modern consensus approach has also been reported by Wolber *et al*.[26] PyRod uses MD data to identify different waters on the basis of their environment and calculates molecular interaction fields with the ultimate goal to enhance current pharmacophore-based approaches.

MD experiments

To generate the data for conserved water site identification, molecular dynamics (MD) simulation was employed and trajectories exported as pdb snapshots. Published experimental crystal complexes were used as starting complexes (RCSB PDB Database). Using a software package from Schrödinger (Small molecule discovery suite - SMD, 2018-3), bond assignation and correction of missing hydrogens was performed first. Overlapping atoms were adjusted and missing residues modelled followed by a hydrogen bond optimization using PROPKA (pH=7.4, from Schrödinger SMD) and capping performed (N-Ac; $CONMe_2$). All MD experiments were conducted in Desmond (v 5.50; OPLS3e force field; nonbonding interactions cutoff radius = 9 Å) software package in replicates and where amenable repeated using Yasara Structure & WHAT IF software. Protein charge was neutralized by the addition of Na^+ or Cl^- ions and explicit solvation performed using SPC, SPC/E, TIP3P and TIP4P water models within orthorhombic or cubic systems with periodic boundaries set 10 Å from macromolecule extremes. Next, a relaxation protocol consisting of two stages of minimization followed by three stages of MD equilibration with gradually diminishing atomic restraints was ran. First two stages used NVT ensemble simulation with Brownian dynamics at 10 K and a 12ps NVT ensemble simulation at 10 K with solute non-hydrogen atoms restrained. Next three stages were 12 ps NPT ensemble simulations at 10 K and 300 K with non-hydrogen solute atoms restrained followed by a 24 ps NPT ensemble simulation at 300 K without restraints. Finally, NPT (periodic boundary conditions in number of atoms, pressure and temperature) ensemble production run at 300 K and 1 atmosphere was initiated. Simulation time with the time step of snapshots was adjusted depending on the experiment. Energy parameters of the systems as well as root-mean-square deviation (RMSD) values for protein backbone and ligands were monitored. Data for conserved water site study were generated by snapshot recording (pdb format) throughout production run in 10 ps intervals (1000 snapshots per 10 ns).

SOFTWARE USAGE

The user firstly, adds the plugin to their PyMol installation of choice and upon its first usage sets up the database by an automat-ic download button that contacts the RCSB using restful services for pre-calculated sequence clusters as well as our online reposito-ry where ProBiS binary resides (“setup db” button; Figure S1). Besides a standard analysis scenario using PDB database for water analysis4, users can now simply input a custom structure (or MD starting trajectory snapshot) in pdb format and select a series of MD trajectory snapshots in pdb format to be applied for con-served water site study (“MD traj” button).

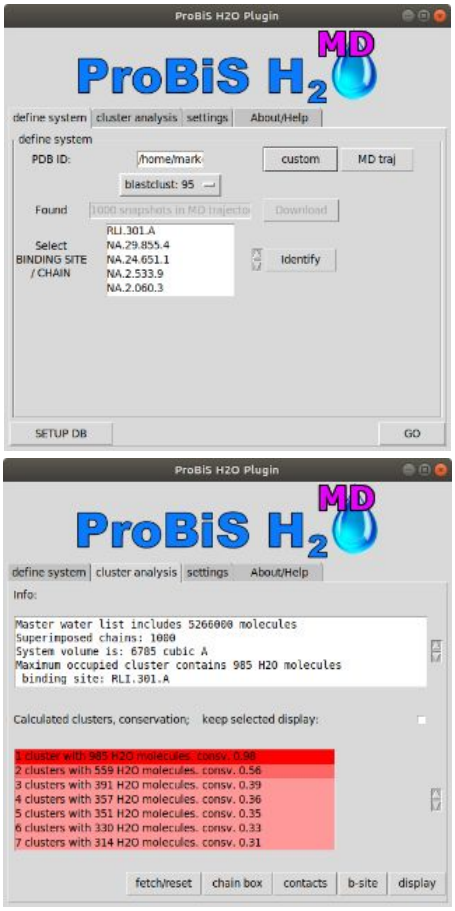

Figure S 1: ProBiS H2O MD usage: Top: Input window where PDB ID input field resides – herein a PDB ID structure can be automatically downloaded or a custom structure specified by the use of the neighbouring "custom" button. "MD traj" button helps the user specify a selection of trajectory snapshots to be used for water site analysis (alternatively a complete user-selected set of structures can be used; "Found" field displays the number of snapshots/structures applied in the calculation). Analysis is commenced by the "identify" button that directs the attention to a general system chain or a specific binding site of interest (they are detected automatically based on the query structure in PDB ID field) and calculation started by the "GO" button.; Bottom: Output window displays a list of identified clusters or water sites of interest, ranked by their conservation with respect to input structures. Each entry can be displayed ("display" button), neighbouring residues emphasized ("b-site" button), distances towards neighbouring residues measured ("contacts" button) and whole studied system highlighted ("chain box" button). "Fetch-reset" button serves to reset all visualisation settings to a starting query structure (PDB ID or complete user selected custom system). ProBiS H2O MD logo, reproduced with permission from logo author Marko Jukič.

After the selection of a query chain or binding site, a robust and local-superimposition step follows to generate water location data transposed to the query coordinate system. Thereafter, water molecules are collected, followed by the Three Dimensional Density Based Spatial Clustering of Applications with Noise (3D–DBSCAN) clustering algorithm, after which the user is presented with a color–coded list of identified water clusters sorted by their conservation (e.g. *number of water molecules in identified cluster / superimposed protein chains*).

REFERENCES

(1) Nguyen, T. T.; Viet, M. H.; Li, M. S. Effects of water models on binding affinity: evidence from all-atom simulation of binding of tamiflu to A/H5N1 neuraminidase. *The Scientific World Journal*, **2014**, 1-14.

(2) Small-Molecule Drug Discovery Suite 2019-3, Schrödinger, LLC, New York, NY, 2019. Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY.

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

(4) Cui, G.; Swails, J.M.; Manas, E.S. SPAM: a simple approach for profiling bound water molecules. *J. Chem. Theory. Comput*., **2013**, *9*, 5539–5549.

(5) Michel, J.; Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aid. Mol. Des*., **2010**, *24*, 639-658.

(6) Sindhikara, D. J.; Hirata, F. Analysis of biomolecular solvation sites by 3D-RISM theory. *J. Phys. Chem. B*, **2013**, *117*, 6718-6723.

(7) Sindhikara, D. J.; Yoshida, N.; Hirata, F. Placevent: An algorithm for prediction of explicit solvent atom distribution—Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem*., **2012**, *33*, 1536-1543.

(8) Ross, G.A.; Bodnarchuk, M.S.; Essex, J.W. Water sites, networks, and free energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc*., **2015**, *137*, 14930–14943.

(9) Bodnarchuk, M. S. Water, water, everywhere… It's time to stop and think. Drug Discov. Today, **2016**, *21*, 1139-1146.

(10) Zhang, L.; Hermans, J. Hydrophilicity of cavities in proteins. *Proteins*, **1996**; *24,* 433-438.

(11) Bayden, A.S.; Moustakas, D.T.; Joseph-McCarthy, D.; Lamb M.L. Evaluating free energies of binding and conservation of crystallographic waters using SZMAP *J. Chem. Inf. Model*, **2015**, *55,* 1552–1565.

(12) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.*, **2007**, *47*, 279-294.

(13) Sridhar, A.; Ross, G. A.; Biggin, P. C. Waterdock 2.0: Water placement prediction for Holo-structures with a pymol plugin. *PloS one*, **2017**, *12*, 1-17.

(14) Morozenko, A.; Stuchebrukhov, A. A. Dowser++, a new method of hydrating protein structures. *Proteins*, **2016**, *84*, 1347-1357.

(15) Rose, P.W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A.R.; Christie, C.H.; Di Costanzo, L.; Duarte, J.M.; Dutta, S.; Feng, Z.; Green, R.K.; Goodsell, D.S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A.S.; Shao, C.; Tao, Y.; Valasatava, Y.; Voigt, M.; Westbrook, J.D.; Woo, J.; Yang, H.; Young, J.Y.; Zardecki, C.; Berman, H.M.; Burley, S.K. The RCSB protein data bank: integrative view of protein, gene and 3D structural information *Nucleic Acids Res.*, **2017**, *45*, D271-D281.

(16) Berman, H.M.; Henrick, K.; Nakamura, H. Announcing the worldwide Protein Data Bank *Nat. Struct. Biol*., **2003**, *10*, 980.

(17) Kinjo, A. R.; Bekker, G. J.; Suzuki, H.; Tsuchiya, Y.; Kawabata, T.; Ikegawa, Y.; Nakamura, H. Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res.*, **2016**, *45*, D282-D28.

(18) Kearney, B. M.; Johnson, C. W.; Roberts, D. M.; Swartz, P.; Mattos, C. DRoP: a water analysis program identifies Ras-GTP-specific pathway of communication between

membrane-interacting regions and the active site. *J. Mol. Biol.,* **2014**, *426*, 611-629.
(19) Cole, J. C.; Giangreco, I.; Groom, C. R. Using more than 801 296 small-molecule crystal structures to aid in protein structure refinement and analysis. *Acta Cryst. D*, **2017,** *73*, 234-239.
(20) Radoux, C. J.; Olsson, T. S.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Identifying interactions that determine fragment binding at protein hotspots. *J. Med. Chem*., **2016**, *59*, 4314-4325.
(21) Rossato, G.; Ernst, B.; Vedani, A.; Smiesko, M. AcquaAlta: a directional approach to the solvation of ligand–protein complexes. *J. Chem. Inf. Model.*, **2011**, *51*, 1867-1881.
(22) Patel, H.; Grüning, B. A.; Günther, S.; Merfort, I. PyWATER: A PyMOL plugin to find conserved water molecules in proteins by clustering. *Bioinformatics*, **2014**, *30*, 2978-2980.
(23) DeLano, W. L. Pymol: An open-source molecular graphics tool. CCP4 *Newsletter On Protein Crystallography*, **2002**, *40*, 82-92.
(24) García-Sosa, A. T.; Mancera, R. L.; Dean, P. M. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model.*, **2003**, *9*, 172-182.
(25) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm11Edited by B. Honig. *J. Mol. Biol*., **1997**, *265*, 445-464.
(26) Schaller D.; Pach, S.; Wolber, G. PyRod: Tracing Water Molecules in Molecular Dynamics Simulations. *J. Chem. Inf. Model*., **2019**, *59*, 2818-2829.