BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 20 2014, pages 2978–2980
doi:10.1093/bioinformatics/btu424

Structural bioinformatics

Advance Access publication July 1, 2014

PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering

Hitesh Patel[1], Björn A. Grüning[2], Stefan Günther[2,*] and Irmgard Merfort[1]

[1]Pharmaceutical Biology and Biotechnology, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Stefan-Meier-Str. 19, D-79104 Freiburg, Germany and [2]Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Hermann-Herder-Str. 9, D-79104 Freiburg, Germany

Associate Editor: Janet Kelso

ABSTRACT

Summary: Conserved water molecules play a crucial role in protein structure, stabilization of secondary structure, protein activity, flexibility and ligand binding. Clustering of water molecules in superimposed protein structures, obtained by X-ray crystallography at high resolution, is an established method to identify consensus water molecules in all known protein structures of the same family. PyWATER is an easy-to-use PyMOL plug-in and identifies conserved water molecules in the protein structure of interest. PyWATER can be installed via the user interface of PyMOL. No programming or command-line knowledge is required for its use.

Availability and Implementation: PyWATER and a tutorial are available at https://github.com/hiteshpatel379/PyWATER. PyMOL is available at http://www.pymol.org/ or http://sourceforge.net/projects/pymol/.

Contact: stefan.guenther@pharmazie.uni-freiburg.de

Received on February 16, 2014; revised on April 21, 2014; accepted on June 25, 2014

1 INTRODUCTION

Water molecules play a crucial role in protein structures, their stabilization, flexibility and activity. One approach to finding conserved water molecules is to calculate energy terms of the molecules in protein structure models. That can be achieved by energy minimization, molecular dynamics simulations or quantum mechanics. GRID determines energetically favorable binding sites for water molecules (Goodford, 1985); Rank uses the HINT force field and classifies water molecules based on the number of hydrogen bonds and their binding strength (Kellogg and Abraham, 2000; Kellogg and Chen, 2004). These methods are precise and can be applied to a single available structure but they are computationally expensive and only applicable for expert users. If many crystal structures are available for a protein, conserved water molecules can be also determined via cluster analysis and electron density of existing crystal structures. The method WaterScore uses a regression analysis to establish a statistical correlation between the structural properties of water molecules of a free protein crystal structure compared with the ligand complexed form (García-Sosa *et al.*, 2003). Bös and Pleiss predicted six conserved water molecules located at the Ω-loop in class A β-lactamases with the WatCH software (Sanschagrin and Kuhn, 1998) and showed that these conserved water molecules reduce the flexibility of the Ω-loop (Bös and Pleiss, 2008). A similar approach has been used to identify and analyze the consensus water sites in thrombin and trypsin. These highly conserved water molecules generally have more protein atom neighbors, a more hydrophilic environment and hydrogen bonds to the proteins, making them less mobile (Sanschagrin and Kuhn, 1998). Major histocompatibility complex (MHC) class I proteins have three highly conserved water molecules, which are believed to be involved in stabilizing the twisted β-turn, modulating peptide recognition and determining the position of N-terminal segment of the α2 helix (Ogata and Wodak, 2002). Cluster analysis of water molecules in alanine racemase reveals the consensus water sites in the active site and at the interface of two monomers, where they play a structural role in maintaining and stabilizing the alanine racemase dimer (Mustata and Briggs, 2004). By a similar method, Loris *et al.* identified conserved waters in a large family of microbial ribonucleases RNase T1 and in legume lectin crystal structures (Loris *et al.*, 1994, 1999). Using deuterium exchange mass spectroscopy and molecular dynamics simulations, Teze *et al.* predicted conserved water molecules channels to be probably involved in the function of family 1 glycosidases (Teze *et al.*, 2013).

As the importance of conserved water molecules is generally accepted, a simple and rapid tool for their identification in protein structures is desirable before structural bioinformatics studies are performed. Up to now two tools are available: Sanschagrin and Kuhn presented a command line tool called WatCH (Sanschagrin and Kuhn, 1998) and C. A. Bottoms *et al.* described a semiautomated method to identify conserved solvent sites and studied conserved waters in six protein families (Bottoms *et al.*, 2006). No other comprehensive open-source tools with a graphical user interface (GUI) are available for such kind of analyses. Here, we present a tool, PyWATER, with an intuitive graphical user interface that can be seamlessly integrated into PyMOL (http://sourceforge.net/projects/pymol), including all important steps that are crucial for identifying the conserved water molecules.

2 IMPLEMENTATION

PyWATER is written in python and depends on the scipy and the numpy modules. At first, the high resolution 3D protein structures, determined by X-ray crystallography, which are

*To whom correspondence should be addressed.

similar to a query structure, are retrieved and superimposed onto the query structure to identify the consensus water molecules in most of the structures. The protein data bank (PDB) identifiers of these structures are obtained by an XML query using the RESTful web services of PDB (Bernstein *et al.*, 1977), which weekly clusters data of protein chains by BlastClust (Altschul *et al.*, 1990) with different sequence identity cutoffs. By default, PyWATER uses 95% sequence identity cutoff. The threshold can be further adjusted by the user. PDB structure files for all sequences in the same cluster are fetched from the PDB.

Then the quality of an overall crystal structure is assessed by its resolution. By default, only structures with $\leq$2.0 Å resolution are selected. Users can further modify the resolution cutoff. The tool also accepts a user-defined list of pdb chains to superimpose. In such case, it bypasses the retrieval of pdb clusters data and filtering by resolution cutoff. Further, the quality of water refinement in a crystal structure is assessed by mobility or normalized B-factor (Buerger, 1960). Based on either one of these two criteria, water molecules are filtered. Mobility reflects the crystallographic temperature factor (B-factor) and the occupancy. From each structure, water oxygen atoms with mobility values $\geq$2.0 are discarded (Sanschagrin and Kuhn, 1998).

$$Mobility_i = \frac{B_i/\langle B\rangle}{O_i/\langle O\rangle}$$

Normalized B-factor is calculated as (Carugo, 1999):

$$Normalized\ B-factor_i = \frac{B_i - \langle B\rangle}{\sigma(B)}$$

In the equations above, B is the temperature factor and O is the occupancy of all atoms of a protein crystal structure.

From each structure, water oxygen atoms having normalized B-factor $\geq$1.0 are also neglected. In case >50% of the water molecules are discarded by mobility or normalized B-factors filter, the whole PDB structure is not considered for further calculation and filtered out. The user can also deselect this filter step by choosing 'No Refinement' in the drop-down list.

The selected crystal structures are superimposed using the backbone as reference to bring all the structures in the same frame as the query protein structure. Then, hierarchical clustering is performed on the 3D Cartesian coordinates of all the water molecules from the superimposed structures. For this step euclidean distance metric to calculate distances between 3D Cartesian coordinates and by default, the complete linkage algorithm are used, aiming to form flat clusters by the inconsistency method with a cutoff of 2.4 Å. For this hierarchical clustering, the fclusterdata method of scipy.cluster.hierarchy python module is used. The inconsistency cutoff threshold of 2.4 Å ensures that at most, only one water molecule from each superimposed structure is present in the cluster. For each cluster, the degree of conservation is calculated. The degree of conservation of water molecule is calculated as the number of water molecules in a cluster divided by the total number of structures superimposed for water clustering. Clusters having a degree of conservation of >0.7 were considered as conserved. This default cutoff can also be changed by the user.

The query protein structure is saved in the PDB file format with only conserved water molecules. Optionally the user can save the superimposed structures used for the clustering analysis. As a result, a PyMOL session is presented showing conserved water molecules with all the hydrogen bonds formed by them. All conserved water molecules are colored according to their degree of conservation. A log file with all input parameters, program messages, warnings and errors is saved. An additional file is generated showing the degree of conservation of each cluster with related atom numbers of water oxygen atoms from the superimposed protein structures. These data are useful to statistically analyze the conserved water molecules in more detail, to manipulate the input parameters or to analyze side chains in close contact.

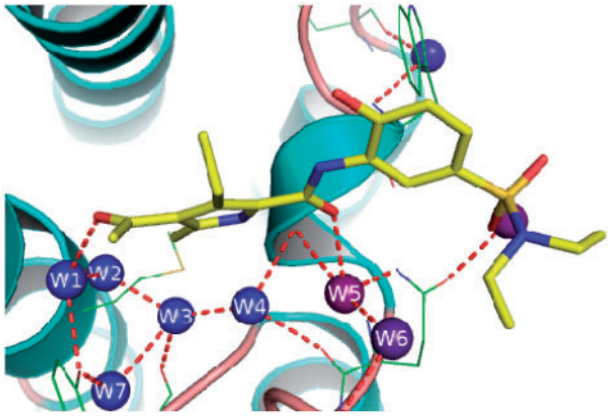

Fig. 1. For pdb entry 4LYW chain A, PyWATER found 31 conserved water molecules. W6 and W7 are predicted, in addition to known W1, W2, W3, W4 and W5, which are important in ligand binding site (Lucas *et al.*, 2013)

3 USAGE

PyWATER can be executed from the command line or used as PyMOL plug-in with a GUI. It can be installed in PyMOL, as a plug-in, by choosing 'Install' under 'Manage Plugins' under the 'Plugin' menu in PyMOL Tcl-Tk GUI and then selecting the 'pywater.py' file.

4 VALIDATION

PyWATER was validated on previously studied proteins such as thrombin, trypsin, BPTI (Sanschagrin and Kuhn, 1998) and bromodomain-containing protein 4 (Lucas *et al.*, 2013). Furthermore, protein families such as MHC class I proteins (Ogata and Wodak, 2002) and class A β-lactamases (Bös and Pleiss, 2008) were analyzed. PyWATER identified all water molecules as discussed by the authors. Figure 1 shows an example of a PyWATER run with PDB entry 4LYW chain A.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bernstein,F.C. *et al.* (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bös,F. and Pleiss,J. (2008) Conserved water molecules stabilize the Omega-loop in class A beta-lactamases. *Antimicrob. Agents Chemother.*, **52**, 1072–1079.

Bottoms,C.A. *et al.* (2006) Exploring structurally conserved solvent sites in protein families. *Proteins*, **421**, 404–421.

Buerger,M.J. (1960) *Crystal-structure Analysis*. Wiley, New York, NY.

Carugo,O. (1999) Correlation between occupancy and B factor of water molecules in protein crystal structures. *Protein Eng.*, **12**, 1021–1024.

García-Sosa,A.T. *et al.* (2003) 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.*, **9**, 172–182.

Goodford,P. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849–857.

Kellogg,G.E. and Abraham,D.J. (2000) Hydrophobicity: is LogP o/w more than the sum of its parts? *Eur. J. Med. Chem.*, **35**, 651–661.

Kellogg,G.E. and Chen,D.L. (2004) The importance of being exhaustive. Optimization of bridging structural water molecules and water networks in models of biological systems. *Chem. Biodivers.*, **1**, 98–105.

Loris,R. *et al.* (1999) Conserved water molecules in a large family of microbial ribonucleases. *Proteins*, **36**, 117–134.

Loris,R. *et al.* (1994) Conserved waters in legume lectin crystal structures. The importance of bound water for the sequence-structure relationship within the legume lectin family. *J. Biol. Chem.*, **269**, 26722–26733.

Lucas,X. *et al.* (2013) 4-Acyl pyrroles: mimicking acetylated lysines in histone code reading. *Angew. Chem. Int. Ed. Engl.*, **52**, 14055–14059.

Mustata,G. and Briggs,J.M. (2004) Cluster analysis of water molecules in alanine racemase and their putative structural role. *Protein Eng. Des. Sel.*, **17**, 223–234.

Ogata,K. and Wodak,S.J. (2002) Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. *Protein Eng.*, **15**, 697–705.

Sanschagrin,P.C. and Kuhn,L.A. (1998) Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Sci.*, **7**, 2054–2064.

Teze,D. *et al.* (2013) Conserved water molecules in family 1 glycosidases: a DXMS and molecular dynamics study. *Biochemistry*, **52**, 5900–5910.