Bioinformatics, 31(12), 2015, 1959–1965
doi: 10.1093/bioinformatics/btv093
Advance Access Publication Date: 13 February 2015
Original Paper

OXFORD

Structural bioinformatics

Mobility-based prediction of hydration structures of protein surfaces

Norbert Jeszenői[1], István Horváth[2], Mónika Bálint[3], David van der Spoel[4] and Csaba Hetényi[5,*]

[1]Department of Genetics, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, [2]Chemistry Doctoral School, University of Szeged, Dugonics tér 13, 6720 Szeged, [3]Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary, [4]Uppsala Center for Computational Chemistry, Science for Life Laboratory, Department of Cell and Molecular Biology, University of Uppsala, Box 596, SE-75124 Uppsala, Sweden, and [5]MTA-ELTE Molecular Biophysics Research Group, Hungarian Academy of Sciences, Pázmány sétány 1/C, 1117 Budapest, Hungary

*To whom correspondence should be addressed.
Associate Editor: Anna Tramontano

Received on November 4, 2014; revised on January 9, 2015; accepted on February 10, 2015

Abstract

Motivation: Hydration largely determines solubility, aggregation of proteins and influences interactions between proteins and drug molecules. Despite the importance of hydration, structural determination of hydration structure of protein surfaces is still challenging from both experimental and theoretical viewpoints. The precision of experimental measurements is often affected by fluctuations and mobility of water molecules resulting in uncertain assignment of water positions.

Results: Our method can utilize mobility as an information source for the prediction of hydration structure. The necessary information can be produced by molecular dynamics simulations accounting for all atomic interactions including water–water contacts. The predictions were validated and tested by comparison to more than 1500 crystallographic water positions in 20 hydrated protein molecules including enzymes of biomedical importance such as cyclin-dependent kinase 2. The agreement with experimental water positions was larger than 80% on average. The predictions can be particularly useful in situations where no or limited experimental knowledge is available on hydration structures of molecular surfaces.

Availability and implementation: The method is implemented in a standalone C program MobyWat released under the GNU General Public License, freely accessible with full documentation at http://www.mobywat.com.

Contact: csabahete@yahoo.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Water molecules located on protein surfaces play fundamental structural and functional roles in biology. For example, hydrogen bonds formed by waters stabilize protein structure (Nisius and Grzesiek, 2012) and affect folding (Cheung *et al.*, 2002; Levy and Onuchic, 2006). Surface water molecules are mediators of the assembly of β-amyloid protofilaments of Alzheimer's disease (Thirumalai *et al.*, 2011) and there is evidence that structurally conserved waters are parts of electron transfer networks (Antonyuk *et al.*, 2013) such as respiratory chain (de la Lande *et al.*, 2010). Structures of many G-protein-coupled receptors are also stabilized by hydration (Angel *et al.*, 2009). A recent study (Xu and Leitner, 2014) suggests that structural water molecules are also involved in thermal

conductance of proteins, in photochemistry, as well as playing a fundamental role in charge transfer, allostery and energy flow (Fang *et al.*, 2009).

Water molecules are often considered essential parts of the protein structure (Petsko and Ringe, 2009) and the first hydration shell is a key determinant of the solubility and aggregation of solute molecules (Israelachvili and Wennerström, 1996). Protein–protein and protein–ligand interactions are influenced by surface-bound water molecules, and therefore, knowledge of their location is of great importance during structure-based drug design (Baron *et al.*, 2012; García-Sosa, 2013). Tightly bound water molecules can affect the chemical diversity of designed ligands (García-Sosa and Mancera, 2006) leading to simple rules for the use of water molecules in drug design (García-Sosa *et al.*, 2005) and also in interpretation of ligand-based pharmacophore models (Lloyd *et al.*, 2004). Inclusion of explicit water molecules in drug design (Mancera, 2007) have been thoroughly studied and was found to be of central importance in ligand–protein docking (Roberts and Mancera, 2008; Thilagavathi and Mancera, 2010).

Although hydration structure is important, it has hitherto proven to be very difficult to determine at the atomic level by experimental means largely due to mobility and complexity of interactions of water molecules located on a protein surface. The residence of a water molecule on the surface and its exchange with bulk are affected not primarily by the strength of protein–water interactions, but it is 'rather a topography that prevents the water molecule from exchanging by a cooperative mechanism' (Halle, 2004a). Importantly, such a cooperative mechanism of exchange also governs several water–water interactions that can often be detected between surface water molecules (Finney, 1977). It is problematic to handle (and to predict) the residence of water molecules in the hydration layer of a protein using merely thermodynamic or kinetic approaches (Halle, 2004a).

Crystallography is the prime experimental method for detection of water positions, via electron density maps and used as the *de facto* standard (Savage and Wlodawer, 1986). However, there are still numerous limitations of this method coming from low resolution of large structural assemblies (Finney, 1977), assignment problems (Afonine *et al.*, 2013; Badger *et al.*, 1997), and artifacts due to cryogenic temperatures used (Halle, 2004b).

A number of computational methods have been proposed for prediction of hydration structure on protein surfaces. Such methods generally require the 'dry' protein structure as an input and provide predictions for hydration structure using a variety of algorithms. A large group of the methods uses fast and simplified approaches disregarding exchange (mobility) between surface and bulk water molecules and dynamics of the hydration structure. They assume a static picture of hydration shells and focus on finding appropriate binding sites of water molecules on the protein surface using scoring schemes, energy calculations (Schymkowitz *et al.*, 2005), prior knowledge (Pitt and Goodfellow, 1993), H-bonding information (Vedani and Huhta, 1991) or artificial neural networks (Ehrlich *et al.*, 1998). Several studies (Makarov *et al.*, 1998; Truchon *et al.*, 2014; Virtanen *et al.*, 2010) have dealt with construction and use of density distribution functions of hydration shells for different atom types occurring in proteins. Limitations of generalized, density-based approaches were discussed in detail (Henchman and McCammon, 2002). These methods ignore dynamics and cooperativity governing hydration.

With advancement of computational infrastructure and force fields, the efficiency and chemical accuracy of atomic level Monte-Carlo and molecular dynamics (MD) simulations has increased enormously in the past decades (Michel *et al.*, 2009; Pettitt and Karplus, 1987) enabling their applications in cutting edge drug design projects (Dror *et al.*, 2012). It has become a routine task to generate MD trajectories with explicit water molecules for virtually any protein of interest. Atomistic simulations of MD hold a conceptual advantage over the static or density-based (trained) methods as the mobility, a key determinant of hydration structure is described directly at atomic level. Whereas such benefits of atomic MD calculations have been extensively used in analyses (Schoenborn *et al.*, 1995), there are not many MD-based methods for prediction of the hydration structure (Abel *et al.*, 2008; Cui *et al.*, 2013; Henchman and McCammon, 2002). These approaches focus on all individual positions of hydrating water molecules and apply various evaluation schemes such as the definition of time averaged positions (Henchman and McCammon, 2002) for calculation of the hydration structure. In this study, we introduce a mobility-based atomic-level method for prediction of hydration structure of molecular surfaces using only 'dry' protein structures as input. Our method was tested on 20 proteins, and the corresponding computational procedures are provided in a program MobyWat, which can be used in conjunction with any MD software that can produce all-atom MD trajectories.

2 Algorithm

2.1 Prediction

Logging molecular movements of all water molecules during a time period provides mobility information required by the prediction process used here. Such a log-book (a trajectory) is preferably generated by MD calculations with an explicit water model. Generation of molecular trajectories was performed by the GROMACS (Hess *et al.*, 2008; Pronk *et al.*, 2013) MD package in this study. During additional post-MD and preparatory steps a standard protocol was followed (Supplementary Methods S1.2).

Mobility information of the trajectory is transformed into the hydration structure of the protein surface during the prediction process outlined in Figure 1. All predictions can be performed with the program MobyWat designed and written in C implementing the prediction protocols of this study. Detailed descriptions of the algorithms can be found in Supplementary Algorithm S2.1 and also in the User's Manual of the program.

Briefly, during the prediction procedure, MobyWat performs clustering of water molecules in candidate pools filtered from the corresponding MD frames. Besides the usual spatial position-based (POS) clustering, an identity (ID)-based algorithm was also introduced with ranking variants named all-inclusive (IDa) and elitist (IDe, Supplementary Algorithm S2.1.5). The procedure ends up in prediction lists including the coordinates and mobility values of water molecules in Protein Databank (PDB) format. A merged (MER) prediction list can be also produced combining the results of the above IDa, IDe and POS predictions.

2.2 Validation

The identification of matches between experimental and predicted water positions is used for validating algorithms of MobyWat. From the matches, a success rate (SR_X) value is calculated for a prediction list (X = IDa, IDe, POS or MER, Eq. 1). The higher the SR_X value, the more successful a prediction is in comparison with crystallographic water positions. For comparison and estimation of the effect of clustering, per frame SR_n values

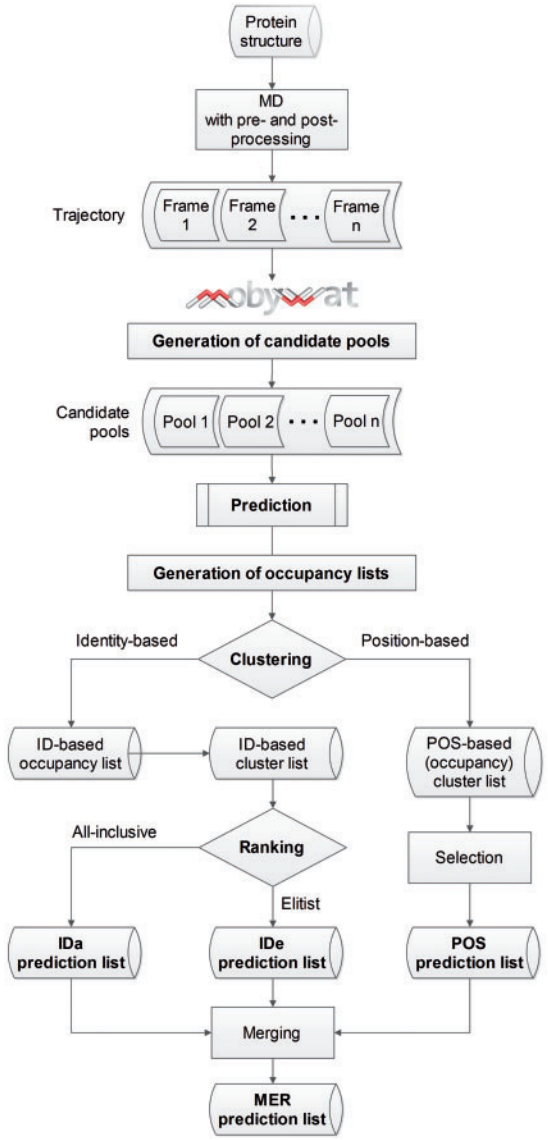

Fig. 1. The prediction process

are also calculated for each candidate pool using the analysis mode of MobyWat ($X = n$, Eq. 1).

$$\mathrm{SR}_X = 100\frac{\text{Number of matches in } X}{\text{Number of water molecules in the reference pool}}\%, \quad \text{where } X = \begin{cases} \text{IDa/IDe/POS/MER (prediction list in validation)}, \\ n \text{ (denotes the nth candidate pool in analysis)}. \end{cases} \tag{1}$$

Further details on validation including selection and calibration of tolerance values are described in Supplementary Algorithm S2.2, and Figure S1. Twenty reference protein systems used for validation and external tests are listed in Tables S1 and S8.

3 Results and Discussion

3.1 Sampling versus predictions

MobyWat predictions are based on atomic mobility data of all water molecules obtained from MD simulations. In this study, mobility of a predicted water molecule is defined by its occupancy value (Supplementary Eq. S2). Occupancy can be counted using a collection (sample) of hydrated protein structures. Such a sample can be collected as a series of hydrated experimental structures of the same protein (Carugo, 1999; Patel *et al.*, 2014), or generated by computational methods. Sample collection from experimental structures is not an option for this purpose as the number of hydrated structures is limited to available entries available in the PDB. In addition, if there are hydrated PDB structures available, then comparative analysis can be performed by other tools (García-Sosa *et al.*, 2003; Patel *et al.*, 2014) which proved to be useful for selection of consensus or conserved water molecules.

However, in most of the cases, only a single structure of the same protein is available. Thus, computational generation of hydration states of a protein is presently the only tractable approach to produce an appropriate sample even if only a 'dry' protein surface is available lacking experimentally determined positions of water molecules. Among computational techniques atomic level MD simulation with an explicit water model is the obvious choice of sampling method. The user needs to supply only a 'dry' protein structure and a series of hydrated protein structures are resulted as an MD trajectory. MD-generated, raw hydration structures are sometimes used even as references in comparison with other methods (Ross *et al.*, 2012). However, important parameters such as the minimal length of an MD simulation necessary for a predictive sampling have not been determined. To address this question, 1-μs-long MD simulations were performed for the protein systems of the validation set producing a sample of 1000 frames spaced at 1 ns. SR_n values were calculated for each pool according to Eq. 1 and plotted in Figure 2A for Alzheimer's amyloid precursor protein (system 2FMA). Descriptive statistics of SR_n values are provided for all validation systems in Supplementary Table S4. The descriptive statistics show a good performance of raw MD sampling with mean SR_n values ranging between 44.6 and 72.7. The SR_n values fluctuate randomly during the 1 μs time-scale of the trajectory (Fig. 2A). This finding can be explained by the short residence time of water molecules in the hydration shell of protein surface (Halle, 2004a). During 1 μs water molecules can change their positions many times, and occurrence of frames with large SR_n values (with a lot of matching water positions) is unpredictable and non-deterministic.

In summary, MD provides an appropriate sampling with good SR_n values. However, the performance of a 'prediction' based on a single frame (randomly) picked from a trajectory is non-deterministic. Thus, a valid prediction cannot be guaranteed if using only one frame. Processing several frames of a trajectory may be a better way to maximize SR and arrive at valid predictions. Accordingly, validation, calibration and measurement of the performance of prediction algorithms are described in the forthcoming sections.

3.2 Validation, performance and robustness

The prediction parameters dmax, ctol and ptol (Supplementary Table S3) were calibrated for all four types of prediction algorithms implemented in MobyWat. The calibration process is documented in Supplementary Results S3.2. Optimal sampling conditions were also determined, as the final step of the validation process. Using calibrated values of parameters, MobyWat predictions were performed for all proteins by processing 1000 coordinate frames from 1-μs-long trajectories. The results are shown for system 2FMA (Fig. 2A), and for all systems of the Validation set (Supplementary Table S4). The SR values yielded by the predictions were significantly higher than the mean SR_n from raw MD, and in many cases they were close to the maximal SR_n values. Thus, all four algorithms

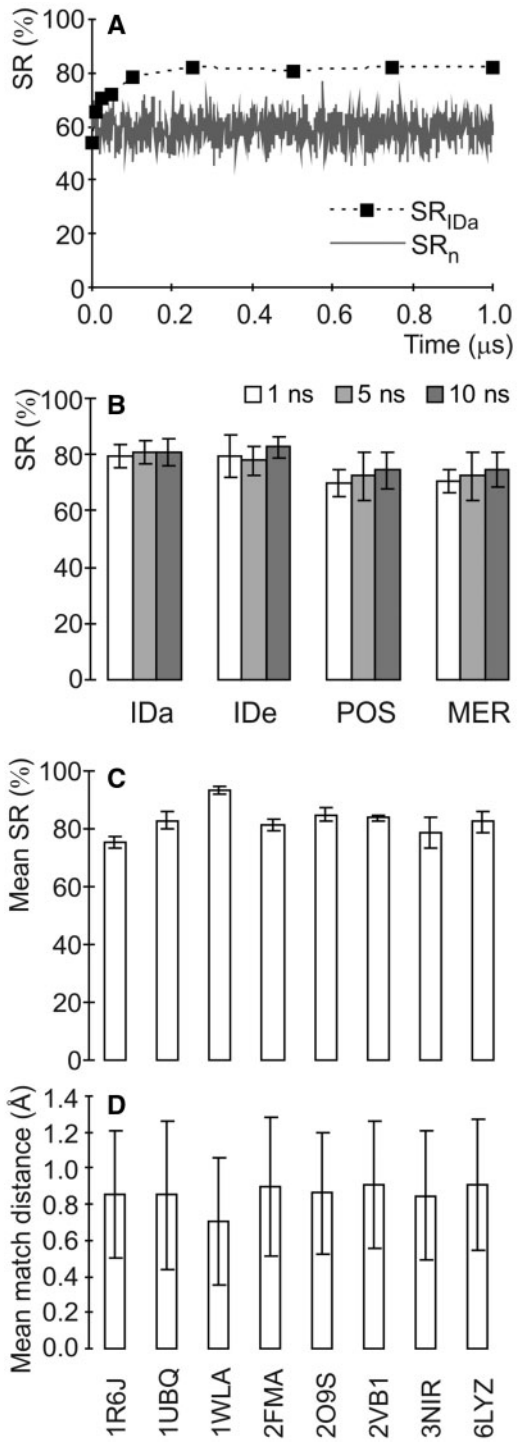

Fig. 2. (**A**) Success rates of Alzheimer's amyloid precursor protein (system 2FMA) calculated for the pools of the raw MD trajectory frames (SR_n) and resulting from IDa prediction of MobyWat (SR_{IDa}). MD trajectory of 1 μs with 1000 frames was used as a sample. (**B**) Effect of sampling time on the performance of prediction algorithms. Ten thousand frames were used for prediction with sampling times 1, 5 and 10 ns. Mean values are calculated from SRs obtained for the Validation set. Standard deviations are shown as error bars. (**C**) Reproducibility of MD sampling in terms of mean SR values calculated from three independent MD runs for each protein system. (**D**) Mean distances in matched pairs of predicted and reference water oxygen atoms plotted for all systems. Error bars denote standard deviations

Table 1. Success rates (%): statistics calculated for raw MD sampling and prediction results achieved by MobyWat

PDB ID[a]	Raw MD[b] (SR_n in Eq. 1)			MobyWat[b,c]			
	Min.	Mean	Max.	SR_{IDa}	SR_{IDe}	SR_{POS}	SR_{MER}
Validation set							
1R6J	41.4	52.4	64.1	71.8	76.2	64.6	65.8
2FMA	39.4	61.5	80.3	80.3	83.6	77.1	77.1
2O9S	46.2	62.2	77.9	87.5	85.6	78.9	78.9
2VB1	44.9	59.9	71.7	82.6	84.1	79.7	80.4
3NIR	33.9	59.0	80.4	80.4	83.9	71.4	71.4
Mean	41.2	59.0	74.9	80.5	82.7	74.3	74.7
SD^V	4.9	3.9	7.0	5.7	3.7	6.3	6.1
Test set 1							
1UBQ	28.6	53.9	82.4	85.7	80.0	68.6	74.3
1WLA	31.4	68.5	94.3	94.3	88.6	82.9	82.9
6LYZ	32.2	54.2	72.9	78.0	81.5	71.2	71.2
Mean	30.7	58.8	83.2	86.0	83.4	74.2	76.1
SD^E	1.9	8.3	10.7	8.2	4.6	7.6	6.0

[a]Mean and standard deviation (SD) values of success rates were calculated for systems of external test and validation separately. [b]Sampling conditions: 10 ns MD run time, 1.0001×10^4 frames. [c]Success rates of MobyWat predictions were calculated with default mtol = 1.5 Å, $b_{max} = 30.0$ Å^2, $d_{max} = 3.5$ Å, $p_{tol} = 2.5$ Å and c_{tol} according to Supplementary Table S5.

resulted in valid predictions. Whereas sampling of 1-μs-long trajectories provided good predictions, such simulations with explicit waters can be computationally demanding. Figure 2A shows that SR_{IDa} values exceeded the SR_n curve and reached a plateau relatively early, after 100–200 ns sampling time. This finding suggested that shortening the sampling time should be possible without a large drop in SR of the prediction. Increasing the sampling frequency (frame count) is also a logical step to achieve reliable predictions with shortened sampling time. Indeed, results in Table 1 reveal that 10-ns-long trajectories with increased frame count yielded mean SR values of >80% for the Validation set, similarly to the 1-μs-long runs (Supplementary Table S4).

Figure 2B shows that the good performance of ID-based prediction algorithms was preserved at 1, 5 and 10 ns sampling times averaged for all systems used in Validation set. In the cases of MER and POS, there is a 5% increase in average SR values if comparing trajectories of 1 and 10 ns length. In summary, the ID-based algorithms outperformed POS and MER predictions, and they provide good predictions even at 1 ns sampling time (Table 1, Fig. 2B).

To evaluate system-independence of our method, a test of the predictions was performed. Systems of Test set 1 (1UBQ, 1WLA and 6LYZ) have relatively moderate resolution and a low number of assigned water positions per protein surface area (Supplementary Table S1). The same set had been used earlier in a study (Virtanen *et al.*, 2010) applying a solvent density-based approach. Detailed comparison of our results using the standards of the earlier study (Supplementary Results S3.3) indicates that overall performance of MobyWat is good if compared with solvent density-based results. For comparability with the above validation results performance of MobyWat on Test set 1 was also evaluated using the standards of this study and the results are listed separately in Table 1. All four algorithms provide valid predictions with SR significantly higher than average values of SR_n. Moreover, the mean SR values of Test set 1 are comparable to or slightly higher than mean SR values obtained for Validation set (Table 1) indicating system-independence of the method.

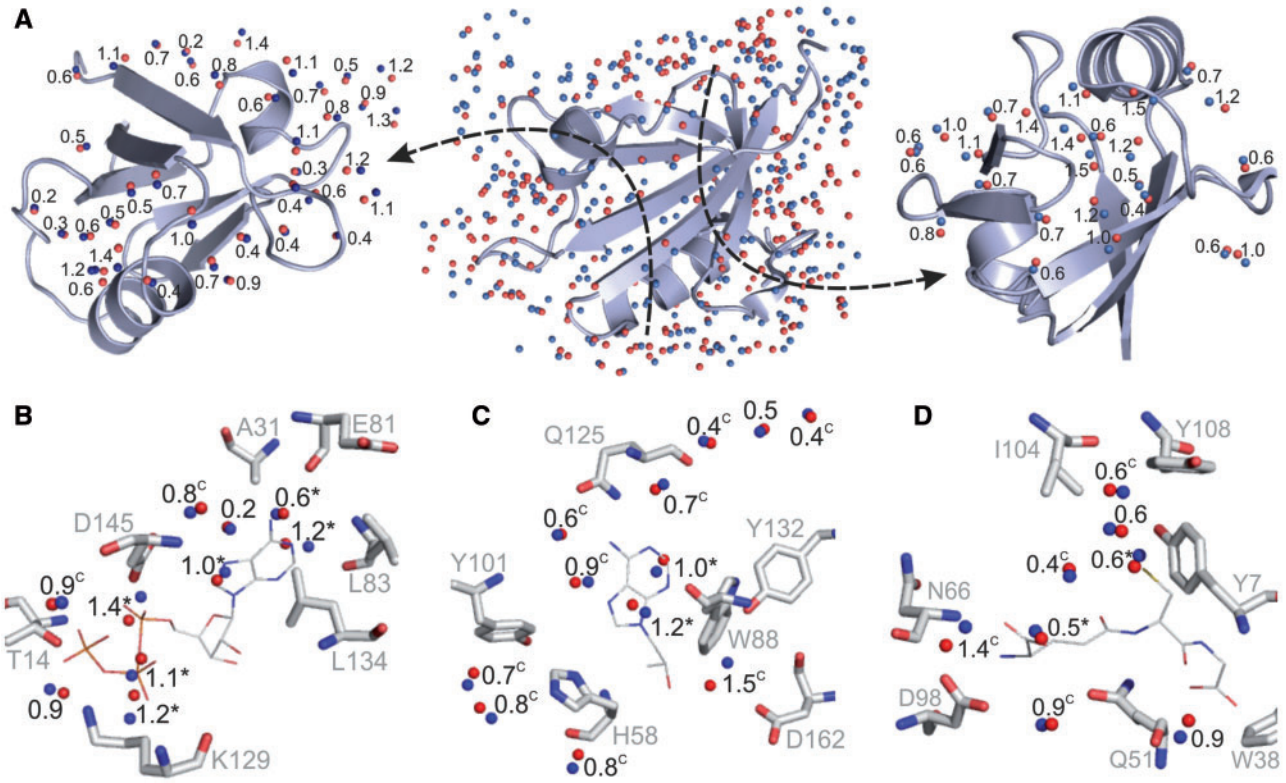

Fig. 3. (**A**) Prediction results for system 1R6J. (**B–D**) Featured binding sites of apo enzymes cyclin-dependent kinase 2 (system 1HCL, B), thymidine kinase (system 1E2H, C) and glutathione S-transferase (system 16GS, D). Ligands were inserted from superimposed ligand-bound enzyme structures (PDB codes 1HCK, 1E2I and 5GSS) for comparison with water positions. Match distances between crystallographic (red spheres) and predicted (blue spheres) water oxygen atoms are given in Å. Conserved and replaceable water molecules are marked with C and asterisk at the distance values, respectively

Reproducibility is also a key issue of robustness. As MobyWat operations are reproducible by their algorithmic definition, reproducibility tests can be performed for the MD sampling process. MD trajectories are inherently chaotic in practical applications due to hardware-dependent rounding of floating point calculations, the use of dynamic load balancing in parallel execution and so on. Therefore, it is common to repeat MD calculations with different starting atomic velocity values to test the convergence of trajectories. Practically, this can be done by selecting different seed numbers of the velocity generator routine. During the tests, three MD trajectories of all systems were produced using three different sets of initial velocities. For these trajectories, predictions were made using the top performer algorithms of Table 1.

The corresponding three SR values were averaged for all systems and plotted in Figure 2C. Their standard deviations are found to be small compared with mean values for all systems, and MD sampling is therefore shown to be reproducible in terms of SR. Improvements in the quality of force fields, in particular the introduction of polarization, may improve the reproducibility of water prediction further (Lopes *et al.*, 2013).

During validations and tests, MobyWat automatically calculated SR values using a match tolerance of 1.5 Å which is the upper limit for the detection of matches between predicted and reference water molecule pairs (Section 2.2). To further quantify the precision of matches, statistics of distances of all matched pairs of the top performer algorithms were calculated (Fig. 2D). It can be seen that mean match distances are below 1 Å for all systems. Matching water positions of one of the systems is shown in Figure 3A, and three other systems are depicted in Supplementary Figure S4.

3.3 Featured test examples

Test set 2 containing 12 proteins was assembled to further check the performance of MobyWat predictions. Using prediction algorithm IDa, a mean SR of 87% was achieved for this set. The members of Test set 2 and the resulted SR values are listed in Supplementary Table S8. Below the prediction results obtained for three enzymatic systems of Test set 2 are discussed focusing on their active sites.

Cyclin-dependent kinase 2 (Cdk2) is a key enzyme in cell cycle control and a promising drug target in oncology (Akli *et al.*, 2011) that also affects senescence (Chenette, 2010). A change of the hydration structure of the active site of Cdk2 due to ligand binding has been reported with obvious implications for drug design (Schulze-Gahmen *et al.*, 1996). A good agreement was obtained between predicted (blue spheres, Fig. 3B) and experimental reference (red spheres) water positions verifying that MobyWat accurately predicted the hydration structure of the active site of apo Cdk2 (Fig. 3B). Notably, experimental water positions were used in comparisons of Figures. 3B–D without any restrictions on their B-factors. Insertion of the ligand (ATP, thin lines in Fig. 3B) from the superimposed ATP-bound Cdk2 structure reveals that six waters (marked with asterisks in Fig. 3) are displaced by the ligand during binding. Release of such water molecules has a favorable contribution to binding entropy of the ligand, and therefore, their identification is important for thermodynamics-driven engineering of new ligands. The results were not affected by the chemical nature of ligand binding as waters replaced by both the charged phosphate moieties and the non-charged adenine ring were found correctly. This finding is in agreement with our general results showing that prediction quality is independent on the type of interacting amino acids

(Supplementary Results S3.7). The second example (Fig. 3C) features the nucleoside binding pocket of thymidine kinase from Herpes simplex type 1. This enzyme has been involved in enzyme-prodrug gene therapy of cancer (Vogt *et al.* 2000). Besides two replaceable water molecules, MobyWat precisely predicted several conserved water positions (marked with C in Fig. 3) existing in both the apo and the ligand-bound enzyme structures. Similar to the cases of replaceable water molecules, locating conserved water sites precisely is also important during the design of new ligands. A complete chain of waters leading to the active site was also predicted correctly (top-right corner of Fig. 3C). The third binding pocket in Figure 3D belongs to glutathione S-transferase, an important detoxifying enzyme (Wu and Dong, 2012). Binding chemistry of glutathione, the peptidic ligand of this enzyme is remarkably different from the previous two ligands with heteroaromatic cores (Fig. 3B and C). However, the quality of MobyWat prediction of the surrounding water positions is similarly good as it was in the other two examples.

MobyWat produces a prediction list including water positions in increasing order of mobility scores (Supplementary Algorithm S2.1.5) where experimentally verified (positive) predictions are mostly located at the top of the prediction list. It was found (Supplementary Results S3.4) that 88% of positive predictions for whole protein surfaces are located in the top 50% of the prediction list. As active sites are the most important spots on enzymes, it was also checked how mobility scores work for these specific segments of the surface. 20 of 24 (85%) of the correctly predicted water positions shown in Figures. 3B–D are located in the top 15% of the prediction lists. Thus, in the cases of active sites investigated, the mobility scores short-list the positive candidates very efficiently at the top of the prediction list. This indicates that water molecules in the active sites of enzymes are predicted with higher fidelity than other water molecules residing on the surface. This result can in part be explained by the presence of conserved water molecules surrounding the ligands, most of which are located at the top 5% of prediction lists. Notably, half of replaceable water molecules occupying active sub-sites in the apo structures were also ranked at top 10%.

4 Conclusions

MD has become an indispensable tool of prediction of structure of proteins and protein–ligand complexes (Shan *et al.*, 2011; Söderhjelm *et al.*, 2012). However, there are only a few MD-based methods for the prediction of hydration structure using explicit simulation of water contacts. Here, we presented MobyWat, a freely available program validated and tested on more than 1500 experimental water positions in 20 different protein surfaces. The prediction process of MobyWat aims at finding the least mobile (most occupied) points of the hydration structure. It was shown that MD simulation is an appropriate sampling technique for such predictions. MobyWat performs predictions using mobility information cumulated in MD trajectories. Two predictive approaches were implemented and tested. The first approach uses only spatial information (coordinates) for a candidate water position. This can be done for example by averaging trajectory frames and producing solvent densities (Virtanen *et al.*, 2010) or by clustering water molecules along the trajectory and counting frequencies of their occurrence in candidate positions. In this study, a second approach was introduced based on identification records of water molecules rather than spatial positions. On average, the identity-based predictions provided higher success rate values than positional and merged algorithms. This is probably a consequence of the position-independent philosophy of the identity-based algorithms.

Valid predictions do not require trajectories from long MD runs. The typical lifetime of a hydrogen bond is a few pico seconds only, virtually independent of the environment (van der Spoel *et al.*, 2006). Consequently, due to rapid exchange and equilibration of water positions relatively short simulations (e.g. 1–10 ns) with regular saving of coordinates suffice. Thus, with a moderate computational effort valid predictions can be achieved.

Limitations of mobility-based predictions were also investigated via an analysis of non-matched water positions of eight systems (Supplementary Results S3.7 and Appendix 2). The analysis identified location of waters above shallow protein sites and/or far from the surface to be a limiting factor in a few cases. Further work is on the way to overcome such limitations using a relative coordinate definition and testing combined MD sampling schemes.

MobyWat algorithms were coded in the portable C language. As the program has to perform calculations on numerous atoms in numerous frames (e.g. $10^4 \times 10^4$) special attention was paid to the efficient use of memory. MobyWat can be used in conjunction with any MD program as it reads frames from PDB files. However, for efficient use of memory and disk space MobyWat also reads and writes xdr-type portable binary trajectory files called xtc in GROMACS.

Mobility is often considered as a disturbing property hampering experimental determination of positions of water molecules on protein surfaces. In this study, it was shown that mobility can be utilized as an information source for prediction of hydration structure. If experimental determination of water structure is not available or incomplete, MobyWat can offer an alternative solution.

Acknowledgements

Part of the simulations were carried out on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the 'Abisko' supercomputer of the High Performance Computing Center North (HPC2N, Sweden, grant SNIC2013-26-6). We acknowledge PRACE for awarding us access to resources Monte Rosa based in Switzerland at CSCS Swiss National Supercomputing Centre, and NIIFI SC based in Hungary at NIIF National Information Infrastructure Development Institute.

Funding

The work was supported by the Hungarian Scientific Research Fund (OTKA K112807) and the MedinProt project of the Hungarian Academy of Sciences. We are thankful to the Gedeon Richter Pharmaceutical Plc. for a pre-doctoral scholarship (to N.J.).

Conflict of Interest: none declared.

References

Abel,R. *et al.* (2008) Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.*, **130**, 2817–2831.

Afonine,P.V. *et al.* (2013) Bulk-solvent and overall scaling revisited: faster calculations, improved results. *Acta Cryst.*, **D69**, 625–634.

Akli,S. *et al.* (2011) Cdk2 is required for breast cancer mediated by the low-molecular-weight isoform of Cyclin E. *Cancer Res.*, **71**, 3377–3386.

Angel,T.E. *et al.* (2009) Structural waters define a functional channel mediating activation of the GPCR, rhodopsin. *Proc. Natl Acad. Sci. USA*, **106**, 147367–14372.

Antonyuk,S.V. *et al.* (2013) Structures of protein–protein complexes involved in electron transfer. *Nature*, **496**, 123–126.

Badger,J. (1997) Modeling and refinement of water molecules and disordered solvent. *Methods Enzymol.*, **277**, 344–352.

Baron,R. *et al.* (2012) Hydrophobic association and volume-confined water molecules. In:Gohlke,H. (ed.) *Protein–Ligand Interactions*. Wiley-VCH Verlag GmbH & Co. KGaA, Wennheim.

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

Chenette,E.J. (2010) Senescence: a key role for CDK2. *Nat. Rev. Cancer.*, **10**, 84.

Cheung,M.S. *et al.* (2002) Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl Acad. Sci. USA*, **99**, 685–690.

Cui,G. *et al.* (2013) SPAM: a simple approach for profiling bound water molecules. *J. Chem. Theor. Comput.*, **9**, 5539–5549.

de la Lande,A. *et al.* (2010) Surface residues dynamically organize water bridges to enhance electron transfer between proteins. *Proc. Natl Acad. Sci. USA*, **107**, 11799–11804.

Dror,R.O. *et al.* (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.*, **41**, 429–452.

Ehrlich,L. *et al.* (1998) Prediction of protein hydration sites from sequence by modular neural networks. *Protein Eng. Des. Sel.*, **11**, 11–19.

Fang,C. *et al.* (2009) Mapping GFP structure evolution during proton transfer with femtosecond Raman spectroscopy. *Nature*, **462**, 200–204.

Finney,J.L. (1977) The organization and function of water in protein crystals. *Philos. Trans. R. Soc. Lond. B*, **278**, 3–32.

García-Sosa,A.T. (2013) Hydration properties of ligands and drugs in protein binding sites: tightly-bound, bridging water molecules and their effects and consequences on molecular design strategies. *J. Chem. Inf. Model.*, **53**, 1388–1405.

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. Mod.*, **9**, 172–182.

Garcia-Sosa,A.T., *et al.* (2005) Including tightly-bound water molecules in de novo drug design. Exemplification through the in silico generation of poly(ADP-ribose)polymerase ligands. *J. Chem. Inf. Model.*, **45**, 624–633.

Garcia-Sosa,A.T., and Mancera,R.L. (2006) The effect of tightly-bound water molecules on scaffold diversity in the computer-aided de novo ligand design of CDK2 inhibitors. *J. Mol. Mod.* **12**, 422–431.

Halle,B. (2004a) Protein hydration dynamics in solution: a critical survey. *Philos. Trans. R. Soc. Lond. B*, **359**, 1207–1224.

Halle,B. (2004b) Biomolecular cryocrystallography: structural changes during flash-cooling. *Proc. Natl Acad. Sci. USA*, **101**, 4793–4798.

Henchman,R.H. and McCammon,J.A. (2002) Extracting hydration sites around proteins from explicit water simulations. *J. Comput. Chem.*, **23**, 861–869.

Hess,B. *et al.* (2008) GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, **4**, 435–447.

Israelachvili,J. and Wennerström,H. (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature*, **379**, 219–225.

Levy,Y. and Onuchic,J.N. (2006) Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 389–415.

Lloyd,D.G. *et al.* (2004) The effect of tightly bound water molecules on the structural interpretation of ligand-derived pharmacophore models. *J. Comput. Aided Mol. Des.*, **18**, 89–100.

Lopes,P.E.M. *et al.* (2013) Polarizable force field for peptides and proteins based on the classical drude oscillator. *J. Chem. Theor. Comput.*, **9**, 5430–5449.

Makarov,V.A. *et al.* (1998) Reconstructing the protein-water interface. *Biopolymers*, **45**, 469–478.

Mancera,R.L. (2007) Molecular modeling of hydration in drug design. *Curr. Opin. Drug Discov. Dev.*, **10**, 275–280.

Michel,J. *et al.* (2009) Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J. Am. Chem. Soc.*, **131**, 15403–15411.

Nisius,L. and Grzesiek,S. (2012) Key stabilizing elements of protein structure identified through pressure and temperature perturbation of its hydrogen bond network. *Nat. Chem.*, **4**, 711–717.

Patel,H. *et al.* (2014) PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics*, **30**, 2978–2980.

Petsko,G.A. and Ringe,D. (2009) *Protein Structure and Function.* Oxford University Press Inc., New York.

Pettitt,B.M. and Karplus,M. (1987) The structure of water surrounding a peptide: a theoretical approach. *Chem. Phys. Lett.*, **136**, 383–386.

Pitt,W.R. and Goodfellow,J.M. (1991) Modelling of solvent positions around polar groups in proteins. *Protein Eng.*, **4**, 531–537.

Pronk,S. *et al.* (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, **29**, 845–854.

Roberts,B.C. and Mancera,R.L. (2008) Ligand-protein docking with water molecules. *J. Chem. Inf. Model.*, **48**, 397–408.

Ross,G.A. *et al.* (2012) Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS ONE*, **7**, e32036.

Savage,H. and Wlodawer,A. (1986) Determination of water structure around biomolecules using x-ray and neutron diffraction methods. *Methods Enzymol.*, **127**, 162–183.

Schoenborn,B.P. *et al.* (1995) Hydration in protein crystallography. *Prog. Biophys. Mol. Biol.*, **64**, 105–119.

Schulze-Gahmen,U. *et al.* (1996) High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. *J. Med. Chem.* **39**, 4540–4546.

Schymkowitz,J.W.H. *et al.* (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl Acad. Sci. USA*, **102**, 10147–10152.

Shan,Y. *et al.* (2011) How does a drug molecule find its target binding site? *J. Am. Chem. Soc.*, **133**, 9181–9183.

Söderhjelm,P. *et al.* (2012) Locating binding poses in protein–ligand systems using reconnaissance metadynamics. *Proc. Natl Acad. Sci. USA*, **109**, 5170–5175.

Thilagavathi,R. and Mancera,R.L. (2010) Ligand–protein cross docking with water molecules. *J. Chem. Inf. Model.*, **50**, 415–421.

Thirumalai,D. *et al.* (2011) Role of water in protein aggregation and amyloid polymorphism. *Acc. Chem. Res.*, **45**, 83–92.

Truchon,J-F. *et al.* (2014) A cavity corrected 3D-RISM functional for accurate solvation free energies. *J. Chem. Theor. Comput.*, **10**, 934–941.

van der Spoel,D. *et al.* (2006) Thermodynamics of hydrogen bonding in hydrophilic and hydrophobic media. *J. Phys. Chem. B*, **110**, 4393–4398.

Vedani,A. and Huhta,D.W. (1991) An algorithm for the systematic solvation of proteins based on the directionality of hydrogen bonds. *J. Am. Chem. Soc.*, **113**, 5860–5862.

Virtanen,J.J. *et al.* (2010) Modeling the hydration layer around proteins: HyPred. *Biophys. J.*, **99**, 1611–1619.

Vogt,J. *et al.* (2000) Nucleoside binding site of herpes simplex type 1 thymidine kinase analyzed by X-ray crystallography. *Proteins*, **41**, 545–553.

Wu,B. and Dong,D. (2012) Human cytosolic glutathione transferases: structure, function, and drug discovery. *Trends Pharmacol. Sci.*, **33**, 656–668.

Xu,Y. and Leitner,M.D. (2014) Vibrational energy flow through the green fluorescent protein–water interface: communication maps and thermal boundary conductance. *J. Phys. Chem. B.*, **118**, 7818–7826.