Published in final edited form as:

J Chem Theory Comput. 2016 September 13; 12(9): 4578–4592. doi:10.1021/acs.jctc.6b00411.

Dissecting the Influence of Protein Flexibility on the Location and Thermodynamic Profile of Explicit Water Molecules in Protein-Ligand Binding

Ying Yang and Markus A. Lill*

Department of Medicinal Chemistry and Molecular Pharmacology, College of Pharmacy, Purdue University, 575 Stadium Mall Drive, West Lafayette, Indiana 47906, United States

Abstract

Explicit water molecules in the binding site of proteins play a crucial role for protein-ligand association. Recent advances in computer-aided drug discovery methodology allow for an accurate prediction of the localized position and thermodynamic profile of water molecules (i.e. hydration sites) in the binding site. The underlying calculations are based on MD simulations of explicit water molecules in a restrained protein structure. However, the ligand-binding process is typically associated with protein conformational change that influences the position and thermodynamic properties of the hydration site. In this manuscript, we present the developments of two methods to incorporate the influence of protein conformational change on hydration sites either by following the conformational transition step-by-step (method I) or to match the hydration sites of the two transition end states using local coordinate systems (method II). Using these methods, we highlight the difference in the estimated protein desolvation free energy with and without inclusion of protein flexibility. To the best of our knowledge, this is the first study that explicitly studies the influence of protein conformational change on the position and thermodynamic profiles of water molecules, and provides methodology to incorporate protein flexibility into the estimation of the desolvation free energy.

Graphical abstract

Keywords

protein flexibility; hydration site; WATsite; molecular dynamics; desolvation free energy

*To whom correspondence may be addressed. mlill@purdue.edu Phone: (765) 496-9375 Fax: 765 494-1414.

INTRODUCTION

Water molecules in the binding site of proteins are important for molecular recognition as they can either bridge interactions between protein and ligand, or can be displaced by the bound ligand.^{1–5} Both mechanisms contribute enthalpically and entropically to the binding free energy.^{2, 4} The explicit consideration of water molecules has gained increasing attention in drug-design projects over the last decade. Several computational tools, such as WaterMap⁶, 3D-RISM⁷ and WATsite⁸, can predict the localized position and thermodynamic profile of water molecules (i.e. hydration site) in the active site using either the ligand-bound (holo) or ligand-free (apo) protein conformation. This information can be utilized in multiple ways in computer-aided drug discovery. For example, efforts are documented to replace the implicit solvent molecular mechanics description of the protein desolvation free energy in binding affinity predictions with explicit binding site water molecules that are displaced upon ligand binding.^{9–12} Also, hydration site information has been used to select important pharmacophore elements thereby reducing the size of the pharmacophore model.¹³ Furthermore, the locations and thermodynamic profile of hydration sites can guide hit-to-lead selection and lead optimization.¹⁴

Typically a single starting structure, either apo or holo form, of the target protein is used for subsequent molecular dynamics (MD) simulations and hydration site prediction in WaterMap or WATsite.^{9, 13, 15} Those studies demonstrated the feasibility of these hydration-site approaches to improve binding free energy predictions. The previously mentioned studies implicitly assumed that structural differences between apo and holo structure are small and do not influence hydration site prediction. As our recent study¹⁶, however, indicated hydration site prediction is largely affected by small conformational variations in the initial protein structure used for the computational analysis. For most protein systems the rigid receptor assumption is invalid and we have to assume that large variations in hydration site predictions exist for protein systems with significant conformational change between apo and holo structure. To the best of our knowledge, all previous hydration-site related studies neglect the conformational transition of the protein upon ligand binding. Thus, the use of hydration site information from a single protein conformation is most likely incorrect for ligand binding (cf. Figure 1).

For example, there may be hydration sites predicted in the apo protein conformation, which are not present in the holo form (water α in Figure 1a). These hydration sites are not directly displaced by the bound ligand; however, they contribute to the protein desolvation free energy since the ligand-induced conformational change causes their disappearance. On the other hand, hydration sites may be predicted in the holo form of the protein and displaced upon ligand binding (water β in Figure 1a) although it is actually not present in the ligand-free apo state of the protein. Thus, adding water β 's desolvation energy to the free energy of binding upon replacement by the bound ligand is actually incorrect. Also, the protein's desolvation free energy is the free energy difference between the water being released to the bulk solvent due to ligand binding versus the water bound in the ligand-free apo state. When the thermodynamic profile of a hydration site in apo and holo state differs, using the energies from the holo state is again incorrect for the estimation of the free energy of binding (water γ , δ in Figure 1b). Therefore, we aim to analyze and incorporate the

influence of protein conformational change during ligand binding on hydration site location and thermodynamic profile.

In the current study, we developed two methods to dissect the changes in location and free energy of hydration sites upon protein conformational change. Method I involves the simulation of the conformational change by Multiply-Targeted Molecular Dynamics (MTMD), and provides a detailed transition of each hydration site throughout the trajectory. In method II, the locations of hydration sites are specified by local coordinate systems defined by nearby protein residues. Using these hydration-site specific coordinate systems (HSSCS), hydration sites in the apo protein structure are directly associated with those in the holo form without the necessity to follow the conformational transition path. In addition, we compare the results from both methods. Whereas method I provides a detailed explanation of appearance and disappearance of hydration sites during protein conformational change, method II is computationally more efficient, identifies a larger number of hydration-site pairs compared to method I and can be applied to a large number of ligands that bind to a diverse ensemble of protein conformational states.

MATERIALS AND METHODS

Protein Systems and Preparation

The ligand-free and ligand-bound form of yeast guanylate kinase (GUA, PDBID: 1EX6, 1EX7)^{17, 18} and the SH2 domain of (Pp60) Src kinase system (SH2, PDBID: 1BKL, 1O42)¹⁹ were used to test our methods. The GUA system undergoes large conformational change upon ligand binding, and the heavy-atom Root Mean Square Deviation in the binding site (RMSD_{BS}) between the apo and holo form is 4.53 Å. On the other hand, with an RMSD_{BS} of 1.62 Å, the SH2 system represents a system with smaller conformational change upon ligand binding. The binding site was defined to contain all residues which have at least one atom within 6 Å of any ligand atom in the X-ray structure of the holo form.

The apo protein conformations were prepared as input for Multiply-Targeted Molecular Dynamics (MTMD) simulations, keeping the crystallographic waters. The program Reduce²⁰ was used to adjust the side-chain conformations of ASN, GLN, and HIS, and tautomers and protonation states of the HIS residues. The protein was then solvated in an octahedron of water molecules using the SPC water model²¹ with a minimum distance of 15 Å between any protein atom and the faces of the octahedron. Chlorine and sodium ions were added to neutralize the simulation systems. The holo protein structures with the ligand removed from the active site were only used as the end-point reference structure for MTMD.

Multiply-Targeted Molecular Dynamics (MTMD) Simulations of Ligand-Induced Protein Conformational Change

All simulations were performed with Amber14²² using an NPT ensemble with the Amber ff99SB-ILDN force field²³, periodic boundary conditions, and a time-step for integrating the equations of motion of 2 fs. The electrostatic interactions were calculated using the Particle Mesh Ewald method^{24, 25}. The Lennard-Jones interactions were truncated at a distance of 10 Å. The covalent bonds including a hydrogen atom were constrained using the SHAKE

algorithm²⁶. Temperature coupling was performed using a Langevin thermostat²⁷ with collision frequency of 2 ps^{-1} at 300 K, and isotropic position scaling with a pressure relaxation time of 1 ps for pressure coupling at 1 bar.

The prepared apo protein structure was first energy minimized for 5000 steps using the steepest descent algorithm. Throughout the equilibration process the temperature of the system was progressively increased from 0 K to 300 K within 50 ps, and then held constant for 500 ps. During the time-course of the MTMD simulation, an additional energy term based on the mass-weighted RMSD was added as a restraint term. The original apo protein conformation was used as the reference structure at the start of the MTMD simulation, and the holo conformation as the end-point reference structure.

A 4 ns MD simulation was performed with restraints on the apo conformation alone, followed by a 20 ns MTMD simulation where the RMSD restraints to the apo conformation were gradually decreased while increasing the corresponding RMSD restraints to the holo conformation. Finally, the MTMD simulation concluded with another 4 ns MD simulation with restraints only to the holo conformation. Since many different protein conformations may have the same RMSD and we are especially interested in the hydration sites in the protein binding site, both the binding site RMSD restraints and all heavy atoms RMSD restraints were applied in the MTMD simulation (Figure 2). Thus, four energy terms (both binding site and all heavy atoms RMSD for apo and holo protein) were added to the restraint term in the energy function. As shown in Figure 2, the MTMD simulation is initially restrained to the apo conformation for 4 ns, i.e. low RMSD to the apo conformation and high RMSD with respect to the holo form (heavy atom RMSD to binding site residues and all residues are 4.5 Å and 3.8 Å for the GUA system, respectively). During the next 20 ns the RMSD restraints to the holo protein conformation are gradually increased (RMSD to apo decreased) forcing the protein system slowly to the holo state, i.e. reduced RMSD values with respect to holo structure (and increasing RMSD to apo form). In the last 4 ns period of the MTMD simulation the system remains restrained to the protein holo conformation. Coordinates were saved every picosecond, generating a protein conformational change trajectory with 28,000 frames.

Extracting Trajectories of Overlapping MTMD Windows

The MTMD trajectory was split into multiple overlapping 4 ns windows to predict the hydration sites. A 3.5 ns overlap to the previous window was chosen to ensure a gradual transition of hydration site locations. As shown in Figure 2, the first set of hydration sites was predicted using the first 4 ns of the MTMD trajectory (window #1), then the second set of hydration sites was predicted using the window between 0.5 ns and 4.5 ns (window #2), and the last set of hydration sites was predicted using the last 4 ns of the MTMD trajectory (window #49).

Theory of Hydration Site Identification and Desolvation Free Energy Prediction

The localized positions of water molecules (i.e. hydration sites) in the active site were identified using the program WATsite based on any MD trajectory. The details of the underlying method have been described elsewhere^{8, 13}. Briefly, a 3D grid was placed over

the user-defined binding site using a grid spacing of 0.25 Å. In each snapshot, the positions of the oxygen atoms of all water molecules in the binding site were determined and a Gaussian distribution function centered on the oxygen atom was used to distribute the occupancy of the water molecule onto the 3D grid. The occupancy distribution was then averaged over the production run, and grid points were filtered out if their occupancy is lower than twice the corresponding value in bulk solvent. Then a density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm²⁸ was used to identify the locations of the hydration sites. The DBSCAN algorithm is efficient and can be applied to arbitrary shapes of clusters. A hydration site is defined if the cluster contains a minimum of 100 grid points. This cutoff value was defined based on the analysis of several X-ray structures (PDBID: 3T8G, 3T74, 3T87, 3T8H, 3T8C, 3T8D, 4H57, 4D9W) for which we were able to reproduce 90% of the water locations in the crystal structures using this criterion. Increasing the minimum number of grid points in a cluster as criterion, would result in too many crystallographic water molecules not being reproduced by the predicted hydration sites. On the other hand, if the minimum number of grid points in a cluster is reduced, we may generate hydration sites with relatively low occupancy increasing the noise in predicting favorable water locations.

The desolvation free energy of each hydration site was determined by separately analyzing the enthalpy and entropy contributions of the water molecules inside a hydration site:

$$\Delta G_{hs} = \Delta H_{hs} - T \Delta S_{hs} \quad (1)$$

H_{hs} and S_{hs} are the enthalpic and entropic change of transferring a water molecule from the bulk solvent into the hydration site of the protein binding site. Details on the calculations of both terms can be found in our previous publication¹³. Briefly, the enthalpic change was estimated by the change of the average interaction energies:

$$\Delta H_{hs} \approx \Delta E_{hs} = E_{hs} - E_{bulk} \quad (2)$$

where E_{hs} is the average sum of van der Waals and electrostatic interactions between each water molecule inside a given hydration site with the protein and all the other water molecules, and E_{bulk} is the interaction energy of a water molecule with all other water molecules in the bulk solvent phase. The entropic change was computed by²⁹

$$\Delta S_{hs} = S_{hs} - S_{bulk} \quad (3)$$

$$S = -R \int p_{ext}(q) \ln p_{ext}(q) dq \quad (4)$$

where R is the gas constant, and $p_{ext}(q)$ is the external mode probability density function (PDF) of the water molecules' translational and rotational motions during the molecular dynamics simulation. TS_{bulk} is the entropy of a pseudo-hydration site in the bulk solvent. To

estimate TS_{bulk} , a 100 ns simulation of a water box with 13734 explicit SPC²¹ water molecules was performed following the same procedures and parameters as the solvated protein simulation described above. Five pseudo-hydration sites are randomly chosen, and the average TS_{bulk} has been -3.8 kcal/mol (standard error: 0.035 kcal/mol).

Development of Two Methods for the Hydration Site Analysis Involving Protein Flexibility

Scheme 1 shows the overall procedure of the two methods we have developed for the analysis of hydration site locations and thermodynamic profiles incorporating protein flexibility. Hydration sites predicted using the ligand unbound (apo) protein structure are called apoHS, and using the ligand bound (holo) protein conformation, holoHS.

Method I consists of three steps: **a.** MTMD simulation with explicit water molecules was used to simulate the conformational change of a protein from its apo structure to the holo form. **b.** The MTMD trajectory was then split into overlapping windows of 4 ns length with neighboring windows separated by 0.5 ns (Figure 2). Our hydration site prediction program WATsite⁸ was used to predict the hydration sites of overlapping MTMD windows. **c.** The changes in hydration sites locations along the MTMD frames and their associated thermodynamic profiles were then analyzed. The potential disappearance, appearance and changes in thermodynamic profiles of hydration site waters during the conformational change were determined.

In detail, using the 49 sets of hydration sites predicted from the overlapping MTMD windows, we aimed to connect the hydration sites in the holo protein conformation to the corresponding hydration sites in the apo form defining continuous paths between the starting and ending locations. The transition of hydration sites was dissected by identifying the hydration-site pairs between neighboring MTMD windows. The closest two hydration sites between the current and the previous MTMD windows were considered as a hydration-site pair if the distance between them is smaller than 1 Å. The 1 Å threshold for defining similar hydration site locations was chosen as the hydration sites are defined as spheres with radius of 1 Å^{6, 13}. If no hydration-site pair was identified between the neighboring windows n and $n-1$, the hydration site comparison was performed between the current window n and window $n-2$. This protocol was continued (comparison n with $n-3$, then n with $n-4$, finally n with $n-5$) until hydration sites were paired or no pairing was identified up to fifth previous window $n-5$. As the RMSD of the protein conformational change from apo to holo for GUA system is about 4.5 Å, using 49 MTMD windows means that the protein changes about 0.5 Å in RMSD throughout five MTMD windows. As our previous study indicated¹⁶, hydration sites are incomparable for protein binding site residues with RMSD larger than 0.5 Å. Thus, we limit the comparison of the positions of hydration sites to the previous five MTMD windows. Once all paired hydration sites are identified, the transition of hydration sites is tracked from the protein apo form to its holo form, and in opposite direction, incorporating the conformational change of the protein upon ligand binding. If one hydration site is not paired to any hydration site in the previous five windows, it is considered as a new hydration site water that appears from bulk solvent throughout the protein's conformational change (β in Figure 1). Hydration sites that are not paired with any hydration site in the subsequent five windows are considered as sites that disappear during conformational change to the bulk

solvent, i.e. the ligand-induced conformational change of the protein displaced the hydration site water into bulk solvent (α in Figure 1).

Method II is based on the observation that a water molecule in the binding site is often stabilized by direct contacts with a single or small number of residues, especially due to specific hydrogen bond interactions. Thus, in this method hydration sites in apo and holo form of the protein are associated based on their similarity in interactions with nearby residues using a local coordinate system, named hydration-site specific coordinate systems (HSSCS). This allows to associate the hydration sites in the holo protein conformation to the corresponding hydration sites in the apo form without explicit simulation of the protein conformational change process. In detail, the eight closest binding site residues are identified for each hydration site (Figure 3A). Since three points are required to create a coordinate system, any three out of the eight residues are used for the definition of HSSCS, giving a total of 56 combinations. The centroids of the residues (denoted as a, b, c in Figure 3B and C) define a plane R. The x-axis is defined by vector $\vec{ab}$ and the y-axis is on the plane R and perpendicular to vector $\vec{ab}$. The z-axis is the perpendicular vector to both the x- and y-axis in a right handed coordinate system. The Cartesian coordinates of each hydration site are projected onto the HSSCS, and the residues defining the HSSCS along with the local coordinates of the hydration site are stored. It should be noted that the sequence of three centroids does influence the local coordinate system. For example, the HSSCS with x-axis defined by vector $\vec{ab}$ or $\vec{bc}$ are different. Thus, there are six permutations for each combination of three residues; thus, a hydration site can be represented by 336 (6×56) HSSCS. Centroids of residues instead of potential hydrogen bond donor/acceptor atoms are chosen to define the coordinate system for the following reasons: First, potential hydrogen bond donor/acceptor atoms may only be good reference points for hydration site that are stabilized predominantly via hydrogen bonding. However, for hydration sites in a hydrophobic cavity, it is unclear which atoms to pick as reference points for defining the coordinate system, as the cavity is formed by a multitude of atoms. Second, there are cases where a hydration site is close to one polar and two nonpolar residues in the apo form. However, in the holo form, this polar residue is no longer part of the protein binding site, while the other two nonpolar residues are still stabilizing the hydration site. Third, for cases where a hydration site is stabilized by hydrogen bonding with a single polar residue, using three atoms (e.g., C α , hydrogen bond donor/acceptor atoms, or main chain N/O) on the same polar residue as reference points may cause numerical instabilities in defining the hydration site position in this particular coordinate system. The reason for this observation is that these atoms are sometimes closer to each other compared to the potentially hydrogen-bonded hydration site. Consequently, the coordinate axes may change significantly with small variance in the atom positions and the same hydration site may artificially have very different coordinates in the two Cartesian coordinate systems. Fourth, the positions of the atoms are thermally fluctuating much larger than the residue centroids. Thus, differences in the hydration site locations using atom-based coordinate systems may be representing the thermal fluctuation rather than conformational transitions between two stable substates associated with apo and holo form of the protein. As Cartesian coordinate system can be very sensitive to the chosen points when they approach co-linearity, we checked the scalar

product of the angle theta formed by vector $\vec{ab}$ and $\vec{ac}$, where a, b and c are the coordinates of the three centroids. For the holo protein conformation of the GUA system, none of the coordinate systems out of 13440 has a scalar product larger than 0.9 ($\cos 25^\circ = 0.906$). For the apo protein conformation of the GUA system, only 3.6% of all coordinate systems (354 out of 9744) have a scalar product larger than 0.9. We need to mention that a voting scheme is used here over all coordinate systems. Such a small percentage of sensitive coordinate system is not believed to significantly influence the overall outcome.

To identify the associated hydration sites in the apo and holo protein conformation, all hydration sites from the apo form are compared to those in the holo form. For any hydration site comparison, distances in HSSCS coordinates are calculated for which the three residues and the permutation are the same. Any two hydration sites with a distance smaller than 2 Å were identified and stored as possible pair. We noticed that two close hydration sites between nearby windows in Cartesian coordinates with distance 0.2 Å have larger distance variations (from 0.1 Å to 0.9 Å) in HSSCS coordinates. Thus, we used 2 Å as the distance criteria for pairing hydration sites using HSSCS coordinates. Using these local coordinate systems, pairs are identified even if the RMSD of the three residues defining the HSSCS between apo and holo form is larger than 10 Å in the Cartesian coordinate system. Different pairs may be identified depending on the three residues defining the HSSCS (Figure 3A). Furthermore, hydration-site pairs that are associated with HSSCS defined by close residues are typically of more importance than using HSSCS of more distant residues because the interaction strength is likely to be larger in the former case, especially when the hydration site forms hydrogen bonds with those residues. Thus, to identify the most likely hydration-site pair between apo and holo form, a hydration-site similarity score was defined that depends on the number of HSSCS in which the pair i-j is matched (Figure 3A) and the distance ($d_{(i,j)-0}$) between the hydration site and the origin of the HSSCS averaged over the apo, d_{i0} , and holo conformation, d_{j0} (i.e., centroid **a** in Figure 3C and D).

$$\text{Similarity score } S_{i,j} = \sum_k^{\text{matched permutations}} \begin{cases} 1 & d_{(i,j)-0} < d_{\text{cutoff}} \\ e^{-0.1(d_{(i,j)-0} - d_{\text{cutoff}})} & d_{(i,j)-0} \geq d_{\text{cutoff}} \end{cases} \quad (5)$$

A distance cutoff (d_{cutoff}) of 4 Å was chosen. The pair with highest similarity score were selected to be the associated hydration sites between the apo and the holo form.

The highest similarity score might be different from apo-to-holo and holo-to-apo direction. For one hydration-site pair, 672 (336×2) is the highest possible similarity score if they are paired in all HSSCS and all distances to the origin of HSSCS, $d_{(i,j)-0}$, are smaller than the distance cutoff (d_{cutoff}). The similarity score was normalized to a range of 0% to 100% by dividing the raw value by the highest possible similarity score (672), yielding the percent similarity score $s_{i,j}$. Finally, all hydration-site pairs with a percent similarity score larger than 1% were used and discussed in the result section. We also extended method II, allowing one hydration site to be paired with multiple sites, and titled this modification method IIb.

Estimation of Buried Apolar Surface Area of Protein-Ligand Complexes

We estimated the buried apolar Solvent Accessible Surface Area (SASA) using the `get_area` command in PyMol³⁰. The buried apolar SASA was computed by summing up the apolar SASA of the protein and the ligand, and then subtracting the apolar SASA of the complex.

Estimation of Desolvation Free Energy of the Protein upon Ligand Binding

In previous studies^{6, 13}, hydration sites are predicted using the holo protein structure alone, and thus the protein desolvation free energy is estimated using the hydration sites in the holo form. Any hydration site to any of the ligands' heavy atoms with distance (d_{hs-lig}) smaller than a distance cutoff (e.g., $d_{lig_cutoff} = 2.24 \text{ \AA}$) are considered as been displaced upon ligand binding. The distance cutoff is adapted from Abel et. al. since it is very unlikely that the contact distance between a water-oxygen atom and a ligand carbon atom is less than $0.8 * (1.4 \text{ \AA} + 1.4 \text{ \AA}) = 2.24 \text{ \AA}$ (assuming the radii of carbon and oxygen atoms are approximately $1.4 \text{ \AA}$).⁶ The contribution of each displaced hydration site is based on the closeness to the ligand:

$$\Delta G_{desolv_rigid} = \sum_{holoHS, lig} \Delta G_{holoHS} \left(1 - \frac{d_{hs-lig}}{d_{lig_cutoff}} \right) \quad (6)$$

The desolvation free energy of the protein binding site (G_{desolv_rigid}) can then be estimated by summing up the free energies of the hydration sites in the holo form, with respect to bulk solvent, that are displaced upon ligand binding.

However, when a protein conformational change is associated with ligand binding, the hydration site free energy profiles in the apo protein state need to be considered as reference for the ligand-free state, i.e. the thermodynamic properties of the apo hydration sites, that correspond to the holo hydration sites that are replaced by the ligand, have to be used to calculate the protein desolvation free energy. To fully take the influence of protein conformational change on protein desolvation into account, three types of hydration site transitions are considered for the estimation of protein desolvation free energy:

Type i. Hydration sites that are displaced upon ligand binding in the protein holo form and match to a corresponding hydration site in the apo form (γ and δ in Figure 1). The free energy of the paired hydration sites in the apo form (G_{γ/δ_apo}) is used.

Type ii. Hydration sites that exist in the protein apo form but disappear in the protein holo form (α in Figure 1). These hydration sites are contributing to the free energy of protein desolvation although they are not directly displaced by the ligand. The reason for this treatment is that ligand binding induces conformational change of the protein which is responsible for the desolvation effect. Ligand binding therefore indirectly replaces those water molecules. This water replacement into bulk solvent, however, may cost or gain free energy (G_{α}) to ligand binding and therefore has to be considered for computing the desolvation free energy.

Type iii. If the hydration site does not have a corresponding site in the apo form, i.e. it only appears from bulk solvent during protein conformational change, its replacement

does not contribute to the free energy of ligand binding (β in Figure 1), and G_β is therefore not considered in the estimation of the free energy of protein desolvation. If a type (iii) hydration site is not displaced by the ligand (ϵ in Figure 1), it unfavorably ($-G_\epsilon$) contributes to the free energy of protein desolvation.

Thus, the overall desolvation free energy is computed by

$$\Delta G_{\text{desolv- flexible}} = \sum_{\text{type- } i} \left[\Delta G_{\gamma/\delta-\text{ apo}} \left(1 - \frac{d_{hs-\text{ lig}}}{d_{\text{lig- cutoff}}} \right) \right] + \sum_{\text{type- } ii} \Delta G_\alpha - \sum_{\text{type- } iii} \Delta G_\epsilon \quad (7)$$

In the modified method IIb, multiple pairings are allowed for one hydration site, the percent similarity score (s) of each pair is used to scale the contribution from different corresponding hydration-site pairs in the apoHS. For type i hydration sites ($G_{\gamma/\delta}$), for example, when multiple hydration sites in apoHS (e.g. l, m, n) are paired with a hydration site (e.g. k) in the holoHS, the contribution from displaced hydration site k to the protein desolvation free energy is estimated by

$$\Delta G_k = \Delta G_l \left(\frac{s_{k,l}}{s_{k,l} + s_{k,m} + s_{k,n}} \right) + \Delta G_m \left(\frac{s_{k,m}}{s_{k,l} + s_{k,m} + s_{k,n}} \right) + \Delta G_n \left(\frac{s_{k,n}}{s_{k,l} + s_{k,m} + s_{k,n}} \right) \quad (8)$$

The desolvation free energy from type i hydration sites are calculated using equation 8, and the contributions from type ii and iii hydration sites are calculated as in equation (7).

RESULTS AND DISCUSSION

Current methods predict hydration sites and their thermodynamic profile based on an initial protein structure, typically a holo form of the protein. Our previous study¹⁶ indicated that even for proteins with relatively small conformational change ($0.5 \text{ \AA} < \text{RMSD}_{\text{BS}} < 1.0 \text{ \AA}$) between apo and holo conformations large variations in location and thermodynamic profile of the hydration sites is typically observed. Here, we propose two methods to identify hydration sites in the protein holo form (holoHS) that originated from corresponding hydration sites in the apo form (apoHS) but changed their location due to the conformational transition between both protein states. In method I, we simulate a protein conformational transition pathway, and predict hydration site changes observed throughout the transition. The hydration sites in the apo form are connected to those in the holo form via locally adjacent, intermediate hydration-site pairs. In method II, we assign hydration sites to the surrounding protein residues, and match the hydration site in the apo form to corresponding sites in the holo form based on their conserved interactions with those nearby residues.

Protein Conformational Change from the Apo to the Holo Form Simulated by MTMD

Protein systems that undergo significant conformational change upon ligand binding are commonly seen in drug discovery, and many of them are assumed to occur via the induced-

fit mechanism. When ligands induce large conformational change, it is important to understand what influence this conformational change has on binding site water molecules. Here, we simulated the protein conformational change by applying RMSD restraints on the heavy atoms of the protein slowly shifting the system from apo to holo conformation.

Figure 4 shows an overlay of the midpoint MD frame from each 4 ns MTMD window representing the conformational transition pathway for the GUA and SH2 systems. The GUA system undergoes a large conformational change of the loop region that interacts with the phosphate group of the ligand (Figure 4A). Movie S1 shows the formation of the ligand binding pocket during the conformational change simulated by MTMD. The RMSD_{BS} between the apo and holo form of the GUA protein system is 4.53 Å, thus the hydration sites predicted using these two protein conformations are not overlapping using the Cartesian coordinates of the superimposed protein structures (Figure S1A). With an RMSD_{BS} of 1.62 Å, the conformational change of the SH2 system is relatively small compared to the GUA system, but the hydration sites predicted using the apo and the holo forms are still spatially distinct (Figure S2A). As the SH2 system undergoes smaller conformational change compared to GUA, the MTMD simulation length was reduced to 20 ns including 12 ns with continuously changing RMSD restraints to apo and the holo form (separate RMSD restraints for binding site and whole protein were used similarly as in the GUA simulations). In the first and the last 4 ns period of the MTMD simulation the system remains restrained to the protein apo and holo conformations, respectively.

Method I: Identify Hydration Site Transition Pathways through MTMD windows

For each 4 ns MTMD window, one set of hydration sites is predicted. However, a simple overlay of all predicted hydration sites from all windows is unable to directly identify the transition pathways for each site during the protein conformational change (Figure S1B & S2B). To identify the pathways of all hydration sites throughout the apo to holo transition pathway, we determined all hydration-site pairs between nearby MTMD windows. As described in the Materials and Methods section, a hydration-site pair is defined if the closest two hydration sites from two adjacent windows have a distance smaller than 1 Å.

Based on the spatial relocation, appearance and disappearance of hydration sites during the protein conformational change, three types of hydration site transitions are identified: (i) Hydration sites that move continuously along with active site residues during conformational change or remain largely unchanged in their location, (ii) hydration sites observed in the apo structure but that disappear during the conformational change to the holo form (i.e. transition into bulk solvent), and (iii) hydration sites that are not observed in the apo structure but appear during the conformational change to the holo form (i.e. transition from the bulk solvent to the binding site). For the calculation of desolvation free energies, only water molecules that are located to hydration sites in the apo or holo form are relevant. Hydration sites that only appear in intermediate windows do not contribute to the computation of protein desolvation free energy, and will therefore not be discussed here.

The detailed results for all hydrations site transitions are tabulated in Table S1 (GUA) and Table S2 (SH2). For the GUA system, our method predicted 29 hydration sites in the apo structure and 40 in the holo structure (Fig S1). 16 hydration sites belong to type (i), and 13

to type (ii), and 24 to type (iii) hydration sites. For the SH2 system (Fig S2), 51 hydration sites were predicted in the apo conformation, and 46 in the holo form. 25 hydration sites were classified as type (i), 26 as type (ii), and 21 as type (iii) hydration sites. Thus, only approximately half of the hydration sites are conserved throughout the conformational transition in the binding site, highlighting the significant dynamics in hydration site appearance and disappearance during the apo to holo transition.

Figure 5 shows examples of hydration site transitions in the GUA system. The thermodynamic properties (ΔG , $-\Delta T \Delta S$, and ΔH) (in kcal/mol) and occupancy of each hydration site is shown on the right. Even for type (i) hydration sites (e.g. Figure 5) that are conserved in the binding site (no appearance or disappearance), the location of the hydration site with respect to the whole protein may alter significantly. The water molecule in Figure 5A, for example, interacts with ASP-100 in the apo protein structure via a hydrogen bond. Throughout the transition to the holo form, GLU-69 approaches the hydration site, and a second hydrogen bond is formed to GLU-69 in addition to the hydrogen bond to ASP-100. Due to the second hydrogen bond with a formally charged residue the enthalpy of solvation becomes more negative, associated with a loss in conformational freedom, i.e. increase in $-\Delta T \Delta S$ (Figure 5A, right). Together with the increasing stability of the hydration site, the occupancy of this hydration site throughout a 4 ns MTMD window increases from about 60% in the apo conformation to 100% in the holo conformation. Although this hydration site is predicted to be conserved between apo and holo form, its predicted free energy of desolvation in the apo form is 2.5 kcal/mol less favorable than in the holo form. Thus, when this hydration site is displaced by a binding ligand, its free energy predicted in the holo form should not be used, as the desolvation free energy is defined with respect to the apo form of the protein.

An example of a type (ii) hydration site that disappears during the conformational change of the protein is shown in Figure 5B. The hydration site in this example forms hydrogen-bond interactions with the carboxyl group of GLN-102 in the apo protein conformation. The carboxyl group flips during the conformational change to the holo form, and the hydration site gradually loses interaction with the group and disappears after the 17 ns MTMD window into bulk solvent. As this hydration site is not predicted in the holo form of the protein, it will not be taken into account in the estimation of the free energy of binding of a ligand, if only the holo protein conformation would be considered. The hydration site, however, actually transits into the bulk solvent during the conformational change induced by the binding ligand. Therefore, its desolvation free energy needs to be considered as if being replaced by the binding ligand, i.e. direct replacement of a water molecule in the holo structure or due to induced conformational change equally contributes to the desolvation free energy of the protein upon ligand binding.

Type (iii) hydration sites appear during the conformation change to the holo structure (Figure 5C). The hydration site is not observed in the protein apo structure as the binding site residue TYR-50 on the flexible loop region is pointing to the bulk solvent. For clarity, we only show the surface of the holo conformation with the bound ligand in the active site. As the protein changes its conformation to the holo form, TYR-50 moves towards the active site, and a hydration site appears forming hydrogen bonds with TYR-50. If this hydration

site is displaced by a ligand and released into the bulk solvent, it will not contribute to the free energy of ligand binding since it originates from bulk solvent in the apo protein conformation and ends in the bulk solvent again upon ligand binding. If it is, however, not replaced by a bound ligand, it will contribute to the free energy of binding, as the interaction energy and entropy of the water molecule in the binding site (in the holo state) may differ from the corresponding free energy contributions in bulk solvent (in the apo state).

Supplementary movies show window-by-window density changes of the above examples: type (i) hydration site (Figure 5A) in Movie S2, type (ii) hydration site (Figure 5B) in Movie S3, and type (iii) hydration site (Figure 5C) in Movie S4. Examples for all three types of hydration sites for the SH2 system are displayed in Figure S3.

Method II: Associate Hydration Sites in Apo and Holo Forms using HSSCS

The protein desolvation free energy obtained from the explicit hydration site information has been successfully used to replace the implicit desolvation term in MMGBSA or in the scoring of docking poses^{9, 31–33}. As demonstrated in the previous section, for protein systems with significant conformational change upon ligand binding, changes in hydration site location and thermodynamic profile need to be considered for a more accurate calculation of protein desolvation free energies. Using MTMD simulations in method I, provides a continuous path between apo and holo hydration sites allowing to define clear associations between hydration sites in the two protein states. The method, however, is too computationally expensive to apply to a large ligand library, if different protein conformations are induced or stabilized by those bound ligands. To accelerate the process of predicting the desolvation free energies of the protein upon ligand binding, a second method is proposed that requires only the hydration site identification and profiling for the end points (apo and holo form). To associate the hydration sites in the holo protein conformation to the corresponding hydration sites in the apo form, the definition of hydration site specific coordinate systems (HSSCS) for each hydration site is introduced. Using the HSSCS, locations and thermodynamic profiles of hydration sites between the apo and the holo form are compared and associated based on their similarity in interactions with nearby residues. The highest similarity score for each hydration site is used to identify a hydration-site pair.

Comparison Between Method I and Method II

A comparison of both methods for identifying associated hydration sites is shown in Table 1 for the GUA system. There are 16 type (i) hydration-site pairs identified using method I, and 18 pairs identified using method II. The majority of hydration sites identified by method I (10 from 16) is reproduced by method II (highlighted in light purple). Hydration-site pairs that are predicted differently by the two methods are highlighted in yellow (method I) and green (method II). The free energy differences of predicted hydration site desolvation between the apo and holo protein conformations ranges from 0.1 to 3.5 kcal/mol. The percent similarity score is normalized to a range 0% to 100% and shown in Table 1.

Method II identifies hydration-site pairs for three additional holoHS (#9, #34, and #36) between apo and holo structure that are missed by method I. Analysis of those additional paired hydration sites highlights an issue encountered in method I. For example, a hydration-

site pair (#28(apoHS)–#9(holoHS)) predicted only by method II is shown in Figure 6. Hydration site #28 in the apo form and #9 in the holo form are stabilized by a hydrogen bond to SER-34. The hydration sites are paired by method II using, for example, the local HSSCS defined by the three nearby residues SER-34, TYR-50, and TYR-78 (Figure 6A). However, in method I, the hydration site disappears at window 12, and does not reappear before window 17 (Figure 6B). The reason for this temporary disappearance is the decreasing water density that does not meet the criteria of defining a hydration site (Figure 6C and 6D). Just before and after the hydration site's disappearance and re-appearance the occupancy of this hydration site is only about 20%. The window-by-window density change of this hydration site is also shown in Movie S5. It should be noted that WATsite hydration site identification is based on the water density on grid points and the clustering method. For areas in the binding site with less pronounced water density peaks, e.g. more mobile water molecules, the definition of the hydration sites is sensitive to the clustering algorithm. The sensitivity to the water density, clustering algorithm, and the number of nearby MTMD windows of hydration-site pair identification is the main reason why hydration site paths become discontinuous throughout the MTMD trajectory.

Method IIb: multiple-to-multiple hydration-site pairings

We also observed that all the hydration-site pairs that differ in the two methods are located in proximity to charged residues such as GLU or ASP. Figure 7A & 7B show an example of hydration sites directly interacting with ASP-100 and GLU-69. There are nine hydration sites that are in close proximity with these two residues in the apo form while only five hydration sites are observed in the holo form. As shown in Figure 7C & 7E, less pronounced water density peaks are observed in the apoHS around GLU-69 and ASP-100 compared to the density of the comparable holoHS (Figure 7D & 7F). The reason for this discrepancy is the flexibility of the solvent-exposed carboxylate group of a GLU or ASP residue; a slight rotation can result in different water density especially at solvent-exposed locations. Analysis of the 4 ns MD trajectory of apo protein conformation (Movie S6) demonstrates the significant mobility of the water molecules in this region at the interface to bulk water, which results in the rather broad distribution of water density with weakly pronounced peaks. Due to the sensitivity of the clustering algorithm in those situations, five hydration sites are identified, although only 3–4 water molecules interact with residue GLU-69 at the same time. This also explains why we observe different hydration-site pairs. For example, both #2 and #12 interact with the residue ASP-100 in the apoHS (Figure 7A), whereas, #23 is observed to interact with the same oxygen atom in the holo form (Figure 7B). Hydration-site pair #12–#23 is identified in method I, while method II identified #2–#23 as a pair. Similarly, different hydration-site pairs are identified around GLU-69 by both methods (#1–#12 in method I; #4–#12 in method II). Both associations are meaningful because both #1 and #4 in the apoHS interact with GLU-69, and #12 interacts with the same residue in the holoHS.

The above observation, that multiple hydration site in the apoHS can be paired with one site in the holoHS, led to the conclusion that one-to-one hydration-site pairing is not ideal. Therefore, we adjusted method II by allowing one hydration site to be paired with multiple sites instead of only using the pairing with the highest similarity score. Table 2 shows all

hydration-site pairs with a similarity score larger than 1%. All hydration-site pairs identified by method I are now predicted by method II as well. HoloHS that are paired differently by method I and the original method II are circled in red (method I) and green (original method II; highest similarity score). It should be noted that the percentage similarity scores are significantly smaller than 100%. This is not surprising, as it is not possible that a hydration-site is within 4 Å of all eight residues used to define the HSSCS. Furthermore, it is unlikely that two hydration sites are paired in all 672 HSSCS, in particular if the protein changes conformation (cf. Figure 3A).

The results of comparing both methods for the SH2 system are shown in Table S3 & S4. All hydration-site pairs that are predicted by method I but not by original method II (Table S3) are predicted when allowing multiple pairings for one hydration site in method IIb. These multiple-to-multiple hydration-site pairings identified here make more sense from a physical perspective since several identified hydration sites may actually be the result of the same water molecule interacting with rotated side chain of the protein residue (Movie S6). Thus, the sensitivity of hydration site identification on water density and clustering algorithm may be reduced by method IIb.

The analysis also highlighted that method I fails to identify many hydration-site pairs that are identified by method IIb. In method I, we pair hydration sites from the apo to the holo structure with the assumption of a one-to-one correspondence. Furthermore, method I is neglecting the temporary disappearance of hydration site due to changes in water density during the conformational transition. In this context, it should also be noted that simulating the conformational change using MTMD may result in intermediate protein conformations that are unrealistic and thus unfavorable for stable water positions (Figure 6).

Prediction of Protein Desolvation Free Energy including Protein Flexibility

As no experimental data is available on water energies, a direct validation of the above prediction is not possible. Thus, we tried to correlate the predicted protein desolvation free energy with binding affinity even if the other components, such as the protein conformational energy and protein ligand interaction energy, are missing.

The desolvation free energies of the protein active site upon binding of several ligands to the SH2 domain of (Pp60) Src protein system. The apoHS and holoHS, and their pairings, were used to estimate the protein desolvation free energy. Ligands from different crystal structures (PDBID: 1O4A, 1O4B, 1O42, 1O43, 1O44, 1O45, 1O46, 1O47, 1O48, 1O49) were placed back into the active site of the protein holo conformation, and the desolvation free energies of different ligands were estimated using the standard rigid-protein method as well as flexible-protein method as presented in the current manuscript. In short, G_{desolv} using the rigid protein method is estimated by adding the desolvation free energies of the holoHS that are displaced by the bound ligand (within the cutoff distance of 2.24 Å to any heavy atom of the ligand). The flexible protein method utilized all hydration-site pairs information from method IIb, and the contribution from three types of hydration sites were computed using equations (7) and (8). The predicted G_{desolv} are then correlated with the experimentally measured binding affinities (pIC₅₀ values).

In addition to the protein desolvation free energy, we also computed the correlation between pIC_{50} versus buried apolar surface area which is frequently used as an estimate for dehydration costs (Figure 8A). A favorable correlation was observed, although the correlation coefficient is lower in absolute magnitude compared to using explicit hydration sites and protein flexibility in method II. A more negative G_{desolv} means a more favorable contribution to the free energy of binding, and therefore should be associated with a higher pIC_{50} value. Whereas G_{desolv} estimated from the flexible-protein method shows a meaningful correlation with the experimental affinity data, the predicted desolvation free energies using the rigid-protein method actually displays an inverse correlation to the experimental affinity (Figure 8). Clearly, the protein desolvation free energy is only one but important contribution to the free energy of binding, and future studies will aim to improve the predictive capacities of the flexible protein method by combining it with other energy terms such as the molecular mechanics direct protein-ligand interaction energy and terms from the generalized Born surface area (MM-GB/SA) method. In the current study, we just wanted to demonstrate the importance of including protein flexibility in hydration site prediction for the more accurate estimation of protein desolvation and its contribution to the free energy of binding.

CONCLUSION

In this manuscript, we presented the development of two methods to incorporate protein flexibility in the prediction of hydration site location and thermodynamic profile. Method I requires time-consuming MTMD simulations, but provides a detailed picture on the hydration site changes during the apo to holo transition. This method is useful for obtaining a detailed understanding of the changes in water density and thermodynamic properties of localized water molecules during the conformational change. It should, however, be noted that the identification of hydration-site transition pathways using method I is sensitive to the clustering algorithm and to changes in water density, so that temporarily low occupancy regions during the conformational transition may lead to the disappearance of hydration sites and therefore a loss in transition pathways. Method II is computationally more efficient compared to method I. It can be easily applied to a large library of compounds that bind to an ensemble of different holo structures of the same protein, and thus may be useful for the calculation of protein desolvation free energy in context of docking or post-processing methods such as MM-GB/SA.

Our study also highlights the large difference in the predicted desolvation free energy of the same hydration site between the apo and the holo protein conformation, which can be as large as 3.5 kcal/mol. Thus, using the hydration site information without inclusion of protein flexibility may lead to the wrong estimation of protein desolvation free energy. With the methods presented in this manuscript, we are able to incorporate protein flexibility into the estimation of the desolvation free energy. The implicit protein desolvation term used in the MM-GB/SA method can be replaced by the estimated desolvation free energy using hydration site replacement. Guimaraes et.al.³¹ have demonstrated an improved correlation between the MM-GB/SA results and experimental data when replacing the implicit desolvation term with explicit hydration site free energies. Thus, a potential new strategy which we will investigate in future studies is to first associate the hydration sites in the apo

and holo forms using our new flexible-protein method, estimate the desolvation free energy for each ligand in the protein system, and use this term in the context of MM-GB/SA replacing the implicit protein desolvation term.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Neha Rana and Amr Abdallah for critical reading of the manuscript. The authors gratefully acknowledge a grant from the NIH (GM092855) for partially supporting this research.

ABBREVIATIONS

MD	molecular dynamics
MTMD	Multiply-Targeted Molecular Dynamics
HSSCS	hydrationsite specific coordinate systems
RMSD_{BS}	Root Mean Square Deviation in the binding site
GUA	yeast guanylate kinase
SH2	SH2 domain of (Pp60) Src kinase system
DBSCAN	density-based spatial clustering of applications with noise
apoHS	hydration sites predicted in the protein apo form
holoHS	hydration sites predicted in the protein holo form

SUPPORTING INFORMATION

Figure S1. Overlay of hydration sites identified in the GUA system. (A) Hydration sites predicted using the apo (Purple) and the holo (Red) protein conformation. (B) Hydration sites of all MTMD windows. (Purple: apo protein conformation, Red: holo protein conformation) (C) Surface representation of the apo protein conformation with the predicted hydration sites and the extracted ligand from the holo form. (D) Surface representation of the holo protein form with the predicted hydration sites and the bound ligand. The ligand binding pocket is observed in the holo form. **Figure S2.** Overlay of all hydration sites of the SH2 system. (A) Hydration sites predicted using the apo (purple) and the holo (red) protein conformation. (B) Hydration sites of all MTMD windows. (Purple: apo protein conformation; Red: holo protein conformation). **Figure S3.** Examples of hydration site transition pathways in the SH2 system. **Table S1.** Data for type (i), type (ii), and type (iii) hydration sites for the GUA system using method I. **Table S2.** Data for type (i), type (ii), and type (iii) hydration sites for the SH2 system using method I. **Table S3.** Comparison of matched hydration site results between the two developed methods for the SH2 system. **Table S4.** Heat map showing all hydration-site pairs identified using method IIb allowing for

multiple pairing with the same hydration site (SH2 system). A percent similarity score of each hydration-site pair is shown in the table and color coded from white (0%) to blue (100%). Hydration-site pairs that identified differently by method I are circled in red. Hydration-site pair with the highest similarity score is identified by original method II. Another apoHS with a higher similarity score identified by method II is circled in green. The other hydration site in holoHS with a higher similarity score paired with the same hydration site in apoHS is circled in cyan. **Movie S1.** Protein conformational change of GUA system simulated using Multiply-Targeted Molecular Dynamics (MTMD). The formation of ligand binding pocket is observed. **Movie S2.** Window-by-window density changes of type (i) hydration site corresponding to Figure 5A. **Movie S3.** Window-by-window density changes of type (ii) hydration site corresponding to Figure 5B. **Movie S4.** Window-by-window density changes of type (iii) hydration site corresponding to Figure 5C. **Movie S5.** Window-by-window density changes of the example in Figure 6. The hydration site disappears at window 12 due to the less pronounced water density, and reappear at window 17 as the water density increase. **Movie S6.** The 4 ns MD trajectory of apo protein conformation corresponding to Figure 7C. Broad distribution of water density with weakly pronounced peaks are observed in this example at the interface to bulk water.

This information is available free of charge via the Internet at <http://pubs.acs.org>

The authors declare no competing financial interest.

REFERENCES

1. Baron R, Setny P, McCammon JA. Water in Cavity-ligand Recognition. *J. Am. Chem. Soc.* 2010; 132:12091–12097. [PubMed: 20695475]
2. Breiten B, Lockett MR, Sherman W, Fujita S, Al-Sayah M, Lange H, Bowers CM, Heroux A, Krilov G, Whitesides GM. Water Networks Contribute to Enthalpy/Entropy Compensation in Protein-Ligand Binding. *J. Am. Chem. Soc.* 2013; 135:15579–15584. [PubMed: 24044696]
3. Hummer G. Molecular Binding: Under Water's Influence. *Nat. Chem.* 2010; 2:906–907. [PubMed: 20966940]
4. Ladbury JE. Just Add Water! The Effect of Water on the Specificity of Protein-Ligand Binding Sites and its Potential Application to Drug Design. *Chem. Biol.* 1996; 3:973–980. [PubMed: 9000013]
5. Snyder PW, Mecinovic J, Moustakas DT, Thomas SW 3rd, Harder M, Mack ET, Lockett MR, Heroux A, Sherman W, Whitesides GM. Mechanism of the Hydrophobic Effect in the Biomolecular Recognition of Arylsulfonamides by Carbonic Anhydrase. *Proc. Natl. Acad. Sci. U. S. A.* 2011; 108:17889–17894. [PubMed: 22011572]
6. Abel R, Young T, Farid R, Berne BJ, Friesner RA. Role of the Active-site Solvent in the Thermodynamics of Factor Xa Ligand Binding. *J. Am. Chem. Soc.* 2008; 130:2817–2831. [PubMed: 18266362]
7. Imai T, Hiraoka R, Kovalenko A, Hirata F. Locating Missing Water Molecules in Protein Cavities by the Three-dimensional Reference Interaction Site Model Theory of Molecular Solvation. *Proteins: Struct., Funct., Bioinf.* 2007; 66:804–813.
8. Hu B, Lill MA. Watsite: Hydration Site Prediction Program with Pymol Interface. *J. Comput. Chem.* 2014; 35:1255–1260. [PubMed: 24752524]
9. Abel R, Salam NK, Shelley J, Farid R, Friesner RA, Sherman W. Contribution of explicit solvent effects to the binding affinity of small-molecule inhibitors in blood coagulation factor serine proteases. *ChemMedChem.* 2011; 6:1049–1066. [PubMed: 21506273]

10. Bortolato A, Tehan BG, Bodnarchuk MS, Essex JW, Mason JS. Water Network Perturbation in Ligand Binding: Adenosine A(2A) Antagonists as a Case Study. *J. Chem. Inf. Model.* 2013; 53:1700–1713. [PubMed: 23725291]
11. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* 2015; 10:449–461. [PubMed: 25835573]
12. Guimaraes CRW. Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free-energy perturbation calculations. *Abstr. Pap. Am. Chem. Soc.* 2010; 240:547–559.
13. Hu B, Lill MA. Protein Pharmacophore Selection using Hydration-site Analysis. *J. Chem. Inf. Model.* 2012; 52:1046–1060. [PubMed: 22397751]
14. Horbert R, Pinchuk B, Johannes E, Schlosser J, Schmidt D, Cappel D, Totzke F, Schachtele C, Peifer C. Optimization of Potent DFG-in Inhibitors of Platelet Derived Growth Factor Receptor beta (PDGF-R beta) Guided by Water Thermodynamics. *J. Med. Chem.* 2015; 58:170–182. [PubMed: 25007344]
15. Li Z, Lazaridis T. Thermodynamics of Buried Water Clusters at a Protein-ligand Binding Interface. *J. Phys. Chem. B.* 2006; 110:1464–1475. [PubMed: 16471698]
16. Yang Y, Hu B, Lill MA. Analysis of Factors Influencing Hydration Site Prediction based on Molecular Dynamics Simulations. *J. Chem. Inf. Model.* 2014; 54:2987–2995. [PubMed: 25252619]
17. Blaszczyk J, Li Y, Yan HG, Ji XH. Crystal structure of unligated guanylate kinase from yeast reveals GMP-induced conformational changes. *J. Mol. Biol.* 2001; 307:247–257. [PubMed: 11243817]
18. Seeliger D, de Groot BL. Conformational Transitions upon Ligand Binding: Holo-Structure Prediction from Apo Conformations. *PLoS Comput. Biol.* 2010; 6:1–9.
19. Lange G, Lesuisse D, Deprez P, Schoot B, Loenze P, Benard D, Marquette JP, Broto P, Sarubbi E, Mandine E. Requirements for specific binding of low affinity inhibitor fragments to the SH2 domain of (pp60)Src are identical to those for high affinity binding of full length inhibitors. *J. Med. Chem.* 2003; 46:5184–5195. [PubMed: 14613321]
20. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *J. Mol. Biol.* 1999; 285:1735–1747. [PubMed: 9917408]
21. Berendsen H, Postma J, Van Gunsteren W, Hermans J. Interaction Models for Water in Relation to Protein Hydration. *Intermol. Forces.* 1981; 1:331–341.
22. Case, DA., Babin, V., Berryman, JT., Betz, RM., Cai, Q., Cerutti, DS., Cheatham, TE., Darden, TA., Duke, RE., Gohlke, H., Goetz, AW., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossváry, I., Kovalenko, A., Lee, TS., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, KM., Paesani, F., Roe, DR., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, CL., Smith, W., Swails, J., Walker, RC., Wang, J., Wolf, RM., Wu, X., Kollman, PA. *AMBER*. Vol. 14. San Francisco: University of California; 2014.
23. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* 2010; 78:1950–1958.
24. Darden T, York D, Pedersen L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* 1993; 98:10089–10092.
25. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* 1995; 103:8577–8593.
26. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* 1977; 23:327–341.
27. Loncharich RJ, Brooks BR, Pastor RW. Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers.* 1992; 32:523–535. [PubMed: 1515543]

28. Ester, M., Kriegel, H-P., Sander, J., Xu, X. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press; 1996. p. 226-231.
29. Minh DDL, Bui JM, Chang CE, Jain T, Swanson MJ, McCammon JA. The Entropic Cost of Protein-Protein Association: A Case Study on Acetylcholinesterase Binding to Fasciculin-2. *Biophys. J.* 2005; 89:L25–L27. [PubMed: 16100267]
30. Version 1.8. Schrödinger, LLC; The PyMOL Molecular Graphics System.
31. Guimaraes CRW, Cardozo M. MM-GB/SA rescoring of docking poses in structure-based lead optimization. *J. Chem. Inf. Model.* 2008; 48:958–970. [PubMed: 18422307]
32. Hu BJ, Lill MA. PharmDock: a pharmacophore-based docking program. *J. Cheminf.* 2014; 6:1–14.
33. Kohlmann A, Zhu X, Dalgarno D. Application of MM-GB/SA and WaterMap to SRC Kinase Inhibitor Potency Prediction. *ACS Med. Chem. Lett.* 2012; 3:94–99. [PubMed: 24900440]

Figure 1.

Scheme for different scenarios of the influence of conformation change of the protein on explicit water molecules involved in protein-ligand binding. Free energy of each hydration site scales from blue (more favorable) to red (less favorable binding). (a) Positional change of hydration sites upon conformational change upon ligand binding: α leaves, β and ϵ enter the protein active site. Upon ligand binding, water β , although replaced by the ligand in the holo state, does not contribute to the desolvation energy as it is in the bulk solvent in the apo state. In contrary, water α , although not directly replaced in the holo state by the ligand,

contributes to the protein's desolvation free energy as it is replaced by the conformational change induced by ligand binding. Contrary, water ϵ contributes to the desolvation free energy. (b) Change of thermodynamic profiles of hydration sites during conformational change upon ligand binding: γ becomes less stable, and δ becomes more stable. The desolvation energy of those waters in the apo state (not holo state) needs to be taken into consideration for the free energy of binding calculation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 2.

Illustration of the 28 ns MTMD simulation of the protein conformational change from apo to holo structure of the GUA system by applying RMSD restraints to both the binding site residues and the whole protein structure. Using 0.5 ns intervals, 49 overlapping windows with a length of 4 ns were generated.

Figure 3.

Illustration of the Hydration Site-Specific Coordinate System (HSSCS) method. (A) All combinations of three out of eight closest residues are used to define a HSSCS. (B) The hydration sites associated with THR-34 and THR-35 are 4.7 Å apart between apo (cyan) and holo (magenta) form when measured in the Cartesian coordinate system of aligned apo and holo protein structure. (C)(D) Three red spheres (a, b and c) defining a plane R are used to create a HSSCS coordinate system. The x-axis is defined by spheres a and b, the y-axis is on the same plane and perpendicular to the x-axis, and the z-axis is perpendicular to the plane R. The two hydration sites are essentially co-localized in terms of the HSSCS.

Figure 4. Representative protein conformations of each 4 ns MTMD window for (A) GUA system, and (B) SH2 system, representing the conformational transition pathway simulated by MTMD simulation (purple: apo structure, red: holo structure).

Figure 5.

Examples of hydration site transition pathways for the GUA system. (A) type (i) hydration sites that remain in the binding site throughout the conformational transition; (B) type (ii) hydration sites that disappear throughout transition from apo to holo form; (C) type (iii) hydration sites that appear throughout the conformational change. The graphs on the right side show the thermodynamic properties (ΔG , $-T\Delta S$, and ΔH) (in kcal/mol) and occupancy of the hydration site throughout the transition (window 0 represents the apo protein structure, and window 48 the holo protein structure).

Figure 6.

Example of additional hydration site pair identified using method II which disappeared at intermediate windows using method I. (A) Hydration site #28 in the apo form (purple), and #9 in the holo form (red) are interacting with SER-34 via hydrogen bonding. (B) Transition trajectories of hydration site #28 in the apo form and #9 in the holo form using method I highlights the missing hydration site throughout the transition. (C) Hydration sites for window 11 and 16 are shown together with the water density at window 14 at which no hydration site was identified. (D) Overlay of thermodynamic properties and occupancy of hydration site #28 and #9.

Figure 7.

Example of different hydration-site pairs identified from method I and II. More hydration sites interact with GLU-69 and SP-100 in solvent-exposed apo form (A) than in the holo form(B). (C)–(F) Water density of each hydration site around GLU-69 or ASP-100. Less pronounced water density is observed in the protein apo form(C,E) than in the holo form(D,F).

Figure 8.

The correlation between buried apolar surface area and predicted protein desolvation free energies (G_{desolv}) versus the experimentally measured pIC_{50} values. (A) The buried apolar surface area is estimated using PyMol as described in the method section. (B) G_{desolv} using the rigid protein method is estimated with the hydration sites predicted using the protein holo form. (C) G_{desolv} from the flexible protein method is estimated using all identified hydration-site pairs between apoHS and holoHS using the equation (7) and (8).

Scheme 1.

Overview of the two methods for the analysis of hydration site locations and thermodynamic profiles during protein conformational change upon ligand binding. Three types of hydration sites can be identified based on their transition during conformational change.

Table 1

Comparison of matched hydration sites between the two developed methods (for the GUA system).

Method I				Method II				
Index of apoHS	ΔG (kcal/mol)	Index of holoHS	ΔG (kcal/mol)	Index of apoHS	ΔG (kcal/mol)	Index of holoHS	ΔG (kcal/mol)	Percent similarity score
8	2.2	2	2.0	8	2.2	2	2.0	29.95%
13	1.6	4	0.3	13	1.6	4	0.3	44.34%
25	0.5	7	0.7	25	0.5	7	0.7	3.61%
				28	2.1	9	2.9	28.23%
1	0.7	12	-0.7	4	2.4	12	-0.7	36.49%
16	0.2	13	1.5	18	1.2	13	1.5	27.24%
22	0.8	14	0.5	22	0.8	14	0.5	79.76%
6	1.1	15	2.4	6	1.1	15	2.4	6.04%
10	0.4	17	3.5	19	1.4	17	3.5	9.73%
21	2.4	18	2.4	9	0.0	18	2.4	3.81%
7	0.2	19	2.8	7	0.2	19	2.8	8.45%
20	-1.5	22	2.0	20	-1.5	22	2.0	47.86%
12	1.5	23	-1.0	2	-0.3	23	-1.0	25.87%
9	0.0	25	-0.7					
17	1.7	28	1.6	17	1.7	28	1.6	56.21%
5	0.3	30	1.9	5	0.3	30	1.9	13.06%
14	0.9	32	1.8	14	0.9	32	1.8	3.98%
				26	3.4	34	2.6	5.03%
				3	0.9	36	2.1	1.33%

Table 2