

# Cluster Analysis of Hydration Waters Around the Active Sites of Bacterial Alanine Racemase Using a 2-ns MD Simulation

Hung-Chung Huang,<sup>1</sup> Daniel Jupiter,<sup>1</sup> Meikang Qiu,<sup>2</sup> James M. Briggs,<sup>3</sup> Vincent VanBuren<sup>1</sup>

<sup>1</sup> Department of Systems Biology and Translational Medicine, College of Medicine, TX A&M Health Science Center, Temple, TX 76504

<sup>2</sup> Department of Electrical Engineering, University of New Orleans, New Orleans, LA 70148

<sup>3</sup> Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001

Received 11 September 2007; revised 6 November 2007; accepted 7 November 2007

Published online 19 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.20893

## ABSTRACT:

Structural data produced by a 2-ns molecular dynamics (MD) simulation on *Geobacillus alanine racemase* (AlaR; PDB: 1SFT) was used to study hydration around the two AlaR active sites. AlaR is a crucial enzyme for bacterial cell wall biosynthesis. It has been shown previously that the potency of an inhibitor can be increased by incorporating a functional group or atom that displaces hydration sites close to the substrate binding pocket of its target enzyme. The complete linkage algorithm was used for cluster analysis of the active site water positions from 126 solvent configurations sampled at regular intervals from the 2-ns MD simulation. Crystal waters in the 1SFT X-ray structure occupy most of the tightly bound water sites that were discovered. We show here that tightly bound water sites can be identified by cluster analysis of MD-generated coordinates starting with data supplied by a single X-ray structure, and we predict a highly conserved hydration site close to the carboxyl oxygen of L-Ala substrate. This approach holds

promise for accelerating the drug design process. We also discuss an analysis of the well-known notion of residence time and introduce a new measure called retention time.

© 2007 Wiley Periodicals, Inc. *Biopolymers* 89: 210–219, 2008.

**Keywords:** alanine racemase; molecular dynamics; cluster analysis; complete linkage; water hydration sites

This article was originally published online as an accepted preprint. The “Published Online” date corresponds to the preprint version. You can request a copy of the preprint by emailing the *Biopolymers* editorial office at [biopolymers@wiley.com](mailto:biopolymers@wiley.com)

## INTRODUCTION

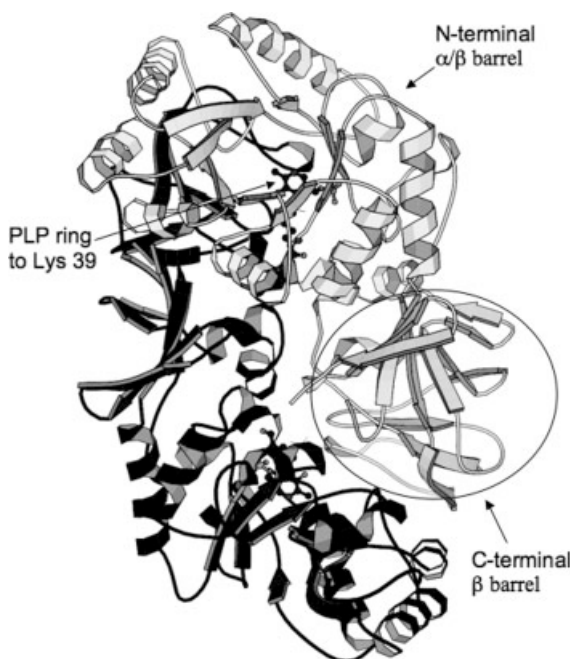
Alanine racemase (AlaR), a pyridoxal 5'-phosphate-dependent (PLP-dependent) enzyme, is a bacterial enzyme that contributes to the formation of peptidoglycan precursors via racemization of L-Alanine (L-Ala) to D-Alanine (D-Ala). This is the first step in the biosynthesis of the peptidoglycan. Because of its crucial role in bacterial cell wall biosynthesis and its nearly unique expression in bacteria, AlaR is considered to be an appropriate target for the design of inhibitors that will act as antibiotics.

The AlaR enzyme from *Geobacillus stearothermophilus* is a homodimer, with subunits 1 and 2 interacting with each other noncovalently, head to tail (Figure 1). The AlaR monomer consists of an N-terminal  $\alpha/\beta$  barrel domain and a C-terminal  $\beta$  barrel domain. A coenzyme PLP ring is covalently attached to the catalytic residue Lysine 39 (Lys39) in each subunit. The active site is formed by the catalytic residues from the N-terminal domain of one monomer and from the

Correspondence to: H.-C. Huan; e-mail: [hc.jhuang@tamu.edu](mailto:hc.jhuang@tamu.edu) or V. VanBuren; e-mail: [vanburen@tamu.edu](mailto:vanburen@tamu.edu)  
Contract grant sponsor: National Institutes of Health  
Contract grant number: AI46340  
This article contains supplementary material available via the Internet at <http://www.interscience.wiley.com/jpages/0006-3525/suppmat>.



© 2007 Wiley Periodicals, Inc.



**FIGURE 1** AlaR (*G. Stearotherophilus*) is a homodimer. The two monomers are shown as dark and light ribbons; Lysine 39 with PLP covalently attached is shown in both active sites as a ball-and-stick model. The C-terminal domain of the light ribbon AlaR monomer is circled for clarity of the domain designation. The figure was rendered with MolScript.<sup>35</sup>

C-terminal domain of the other subunit. There are two active sites (A and B) in each homodimer.

We performed a 2-ns molecular dynamics (MD) simulation on the AlaR X-ray structure (1SFT)<sup>1</sup> from *G. stearotherophilus* solvated in a periodic water box with the A-active site empty and the B-active site filled with an *L*-Ala substrate. Structural frames created in the course of an MD simulation may be explored and studied in order to investigate the relationship between the structure and the function of this enzyme. Such data can be used to assist in the design of new drug leads for inhibition of this important bacterial enzyme. In particular, hydration site studies around the enzymatic active sites may be especially helpful in modifying the current AlaR inhibitors to enhance their specificity and potency.

Water plays a crucial role in the structure, stability, dynamics, and function of biological macromolecules.<sup>2–6</sup> Protein dynamics are tightly connected to the dynamics of surrounding and internal water molecules.<sup>6</sup> Although bulk solvent is important for protein folding via the hydrophobic effect,<sup>2,3</sup> hydration waters have different dynamic behavior when compared to the waters in bulk.<sup>4</sup> Interior waters have *residence times* in the range of  $10^{-8}$  to  $10^{-2}$  s, as described in the experimental study by Otting et al.<sup>7</sup> In contrast, hydration of the protein surface in solution is by water molecules

with residence times in the subpicosecond to subnanosecond range, according to Otting et al.<sup>7</sup> and our theoretical study (data not shown). Our analysis identified waters in AlaR that display interior-like properties. Most of these are in the interior of the protein in identical locations to those in the crystal structure. Calculated residence times in our study were sometimes greater than the length of the simulation. Some of the buried waters inside the protein reside in a site for the entire length of the simulation (2 ns, see Results and Discussion), which yields a calculated residence time that is infinite. Waters that remain localized in a cluster for the entire simulation thus cannot be assessed for either an accurate residence time or for an accurate *retention time* (an alternative metric described in the Methods), but we can assess these measures as greater than 2 ns (the length of the simulation). The range of  $10^{-8}$  to  $10^{-2}$  s for residence times of interior waters is further supported by the work of Garcia and Hummer; at  $T = 300$  K they identified two waters that do not escape and 17 water molecules which reside in the protein interior for time periods longer than 500 ps in their 1.5-ns period of MD simulation on cytochrome *c*.<sup>8</sup>

It has been shown that specific, conserved, and tightly protein-bound water molecules are important for substrate and ligand recognition in numerous protein structures and that knowledge of their role in protein function can assist in the drug design process. For example, the base specific binding of the Trp repressor and operator DNA is achieved by a number of water mediated hydrogen bonds.<sup>9</sup> Water also allows plasticity in molecular recognition, so that class I human major histocompatibility complex (MHC I) can bind peptides of widely different side-chain chemistry, via water mediated contacts to bridge gaps between receptor and ligand.<sup>10</sup> A water molecule bound to the absolutely conserved Tyr146 allows thymidylate synthase to discriminate between substrate and product nucleotides.<sup>11</sup> Finally, a higher affinity cyclic urea inhibitor for HIV-1 protease was obtained by incorporating a carbonyl oxygen to displace a conserved water molecule in this enzyme.<sup>12,13</sup>

The concept of replacing and mimicking tightly bound water molecules in conserved water sites (as determined by coordinates derived from X-ray crystallography or other methods) for drug design purposes has become widespread.<sup>12–16</sup> Although the problem of identifying these important water sites from the available structural information of MD simulation has not been fully explored, some researchers have utilized MD simulation to calculate the residence times of the waters at the protein-solvent interface. These calculations have succeeded in rediscovering most of the hydration sites derived from NMR or X-ray experiments, as well as discovering new sites.<sup>17–19</sup>

The residence time analysis methodology developed by Brunne et al.,<sup>17</sup> originally for the study of water sites around the BPTI molecule, is based on analyzing those water molecules which can be located within a given shell around any particular atom within the protein of interest. A similar analysis by Lounnas and Pettitt employs a grid with a mesh size of 1 Å, covering the protein of interest, in their case the myoglobin simulation system.<sup>18,19</sup> Thus, the above two methods need reference points in order to study water hydration; protein atoms and grid points, respectively.

In this study, we use a simple strategy that can be applied to any area of the system of interest without using a reference point. Inspired by Sanschagrin and Kuhn's work<sup>20</sup> analyzing conserved X-ray derived water sites, we used the cluster analysis technique to study and compare water hydration sites between the empty A-active site and the L-Ala bound B-active site. This is done using structural information from the AlaR MD simulation, and allowed us to identify the hydration sites conserved in both active sites, as well as those sites that are displaced by substrates. A list of cluster sites and associated waters (for which relative 3D coordinates are available on request) is provided to assist with rational drug design targeting of alanine racemase (Supplementary Tables I and II). Although the cluster analysis technique has been applied to the analysis of water positions in protein crystal structures,<sup>20,21</sup> this is the first time that the cluster analysis with complete linkage algorithm has been utilized to analyze the animation waters in MD trajectories

## METHODS

Lounnas and Pettitt discovered a connected cluster of hydration around myoglobin using a 170-ps MD simulation of ~11K atoms in the simulation box.<sup>18</sup> As argued by Brunne et al., although one would have to simulate for a long time in order to sample all possible different conformations for a small and flexible peptide, a 1.4-ns simulation is adequate to obtain a qualitative picture of water of hydration for a stable structure like BPTI (~10K atoms in the simulation system).<sup>17</sup> Given these arguments, and that we have a much larger and more stable system (due to ~68K atoms in the much bigger simulation box), we performed a 2-ns MD simulation of AlaR with simulation conditions and parameters similar to those described in Mustata et al.<sup>22</sup> A notable difference is that in our experiment, we have an empty A-active site and the substrate L-alanine in the B-active site; whereas Mustata et al. simulated D-alanine in one active site and the weak inhibitor propionate in the other. The substrate L-Ala in the B-active site of our system was modeled by making use of the (R)-1-aminoethylphosphonic acid (L-Ala-P) in the AlaR complex structure (pdb code: 1BD0),<sup>23</sup> which was superimposed on the reference structure for the MD simulation (pdb code: 1SFT).<sup>1</sup>

Briefly, the simulation consisted of a system including the AlaR protein (coordinates based on 1SFT), the L-Ala substrate, 196 crystal

waters (labeled with the "X" prefix for its initiation from the "X-ray" resolved water site although these waters may leave the crystal water position during the simulation), 10 sodium ions (to neutralize the molecular system for the particle mesh Ewald method<sup>24</sup>), and 18,512 solvent TIP3P<sup>25</sup> water molecules (in two segments SLV1 and SLV2 with 9256 waters each due to the 9999 limit of CHARMM<sup>26</sup> on each segment; waters in these two segments were labeled with "S" and "V" prefix respectively in our analysis), for a total of 68,227 atoms, in a rectangular box of dimensions 108 × 80 × 80 Å<sup>3</sup>. The CHARMM22 force field<sup>27</sup> was used to represent all protein atoms explicitly. The SHAKE<sup>28</sup> algorithm was used to constrain the bonds involving hydrogen atoms and the Verlet<sup>29</sup> algorithm was used to solve the equations of motion. The system was prepared and energy minimized with the CHARMM program.<sup>26</sup> The heating, equilibration, and the subsequent production phases were carried out with the NAMD program<sup>30</sup> with explicit solvent and periodic boundary conditions in the NPT ensemble. A 2-ns production run was performed and the MD trajectory snapshots were saved every 0.4 ps resulting in 5000 frames.

To select essentially independent frames, we examined 126 structural frames including the frame at time 0 (with 16-ps intervals between consecutive frames) for our cluster analysis of the simulation waters. The number of frames examined (i.e., 126) is at least 10 times the number of PDB X-ray structures examined by Sanschagrin and Kuhn,<sup>20</sup> and the average of the consecutive pairwise root mean square deviation (RMSD) values on the protein atoms of those 126 frames is around 1.06 Å, which indicates a substantial difference among the examined MD frames. We also chose to use complete linkage clustering, as it produces compact, globular clusters and allows specification of a maximum diameter for clusters.<sup>20</sup> In complete linkage clustering, the distance between two water clusters, *i* and *j*, is defined to be equal to the greatest distance between a water molecule in cluster *i* and a water molecule in cluster *j*. The "OC" cluster analysis program was utilized to implement complete linkage cluster analysis.<sup>31</sup> The complete clustering process that we used can be summarized as follows:

1. Calculate the distances between all initial clusters. In this study, initial clusters are made up of individual waters of interest (i.e., those around the active sites). This step creates a pairwise distance matrix for the waters of interest.
2. Join the two closest clusters and recalculate the intercluster distances.
3. Repeat step 2 until all waters are in clusters of diameter less than the cutoff diameter, and any new clustering will create clusters of diameter greater than the cutoff. Our choice of cutoff is discussed below.

The active site residues are formed by "Lys39, Tyr43, Arg136, His166, Asn203, and Tyr354" in the N-terminal domain and "Tyr265, Cys311, and Met312" in the C-terminal domain of the monomer subunits. The A-active site is formed by the residues from the N-domain of monomer 1 and the C-domain of monomer 2; the B-active site is formed by the residues from the N-domain of monomer 2 and the C-domain of monomer 1.

Before performing cluster analysis of the water sites, the saved MD coordinates were superimposed via the backbones of the active site residues at time frame 0 using Kabsch's algorithm,<sup>32,33</sup> via a transformation matrix calculated by the VMD program.<sup>34</sup> This was

done so that the water coordinates could all be viewed in the same reference frame. A superimposition was performed for each active site, in order to accurately derive transformed water coordinates around each of the two active sites; that is, each frame was superimposed separately onto each of the active sites. The 3D figures in this article were rendered with the MolScript<sup>35</sup> (Figure 1) or Raster3D programs<sup>36</sup> (Figures 2 and 3).

According to Kuhn and coworkers,<sup>37</sup> water molecules within 3.6 Å of protein surface atoms are considered to be first shell waters. We used 4 Å to account for higher structural flexibilities, as our work involves the superimposition of over one hundred frames and thus admits much variability, whereas the work of Sanschagrín and Kuhn involved at most 10 homologous PDB structures.<sup>20</sup> After the superimposition step, all waters in the first shell around both the active site residues and the L-Ala substrate were identified. Pairwise distances between these water molecules were then computed.

A maximum cluster diameter of 2.4 Å was chosen by Sanschagrín and Kuhn<sup>20</sup> as the cutoff for the complete linkage clustering. This diameter was chosen given the assumption that water molecules have an approximate effective diameter of 3.2 Å, so a cluster diameter of 2.4 Å allows a 50% overlap of water radii within a cluster (i.e., 1.6 Å nonoverlapping regions plus 0.8 Å overlapping regions, from center to center). Because of the high variability found in large MD structural data sets, and in accordance with visual inspection showing that some confirmed water sites are represented by clusters with diameters closer to 4 Å, we allowed a maximum cluster diameter of 4 Å.

Deviations from parameter values used by Sanschagrín and Kuhn<sup>20</sup> were made to calibrate our analysis according to our preliminary results. A maximum cluster diameter of 4 Å was chosen as the cutoff for the complete linkage clustering because some well-defined water clusters were discovered near that constraint; for example, a hydration site with longer than 2-ns calculated residence time was formed around MD/B-active site by clustered waters with maximum distance of 3.7 Å among them (Table I). Further, we used more than one hundred structural frames to determine the hydration sites (compared to around 10 X-ray structures utilized by Sanschagrín and Kuhn<sup>20</sup>), and it is reasonable to use a larger cluster diameter constraint to reflect the greater flexibility of water dynamics in simulated data, which is more attenuated in X-ray structures. Clusters for which a given water molecule appeared in more than 45% of the examined MD frames were marked as conserved hydration sites. This threshold was used to include the mildly conserved S5033 (47.6%) in A-site as contrasted to the highly conserved X7 (100%) in B-site where both water sites are in equivalent position (Table I). We also refer to the set of maximally dense conserved clusters determined in this manner as “microclusters” to emphasize that many of the water molecules within a cluster physically overlap.<sup>20</sup> The centroids (i.e., the average coordinates of the waters in the cluster) of the identified microclusters are used to represent the conserved and stable hydration sites around the active sites of AlaR. The centroids are given in the coordinate system determined by the superimposition step.

Next, we aligned the equivalent amino acid residues of the A- and B- active sites, again using Kabsch’s algorithm as described earlier for superimposition of different frames. Using these derived coordinates, we superimposed and compared microclusters separately identified in the two active sites. The conserved water sites (i.e., microclusters appearing in equivalent positions in both sites)

and displaced clusters (i.e., microclusters appearing in the A-site, but displaced in the B-active site by the substrate) were determined.

We define retention time as “the total length of time a specific water stays in the hydration site during the simulation”. Retention time during the 2-ns period was determined for each microcluster as follows. First, we identify the water which appears for the longest time in a given cluster. Then we multiply the percentage of studied frames in which this water appears, by 2000-ps. We also follow the method described in Brunne et al.<sup>17</sup> and Rocchi et al.<sup>38</sup> to determine residence time, an alternative measure of water localization. (We note here that, as our simulation was run for 2 ns, each of the 125 time steps is 2 ns/125 = 16 ps long. Thus, the intervals range from 0 ps to 2 ns, increasing in increments of 16 ps). According to Rocchi et al.,<sup>38</sup> the resulting function of the time step  $P(t)$ , as described by the equations below, is a decaying exponential.

$$P(t) = \sum_{j=1}^{N_w} \frac{1}{N - m + 1} \sum_{t_0=0}^{\Delta t_s(N-m)} P_j(t_0, t_0 + t) \quad (1)$$

$$P(t) = P(0)e^{-t/\tau} \quad (2)$$

where  $N$  is the total number (not including  $t = 0$ ) of frames selected for analysis along the MD trajectory (125),  $N_w$  is the number of water molecules,  $m$  is an index of the time steps (0, 1, ...,  $N$ ),  $t$  is the length of the time interval, where  $t = m\Delta t_s$  and  $\Delta t_s = 16$  ps in our analyses.  $P_j(t_0, t_0 + t)$  is a binary function equal to one if water molecule  $j$  remains in the referred cluster site over the entire interval from time  $t_0$  to time  $t_0 + t$ , and zero otherwise.  $\tau$  is the “typical relaxation time” which corresponds to the average time water molecules reside in the cluster; we estimated  $\tau$  by fitting the survival function [Eq. (1)] with the single relaxation time function provided by Eq. (2), as described in more detail below.

The majority of previous efforts on residence time studies have conducted shorter MD simulations on the order of several hundred picoseconds<sup>17,19,38–47</sup> although longer MD simulations have also been performed.<sup>8,48–51</sup> In the 2-ns period of our MD simulation, there is cessation of the decaying exponential behavior at some time interval for each of the 15 microclusters identified in this study. We visually determined this time interval for each microcluster by considering the function  $P(t)$  from 0 to that time interval and using this data to estimate  $\tau$ , the residence time.

## RESULTS AND DISCUSSION

### Tightly Bound Water Sites Determined by Complete Linkage Clustering

Some advantages of the complete linkage clustering algorithm applied to the MD simulated waters include:

1. The hydration site analyses can be focused on the area of interest (e.g., enzyme active site or ligand binding pocket in receptor) in the simulated biological system and there is no need to analyze all the waters in the

Table I Microclusters with Conserved Waters Around Active Sites of AlaR

Most-Conserved Water ID <sup>a</sup>	Cluster ID	Max. Distance <sup>b</sup>	Water Mol. <sup>c</sup>	No. of Frames <sup>d</sup>	Percentage <sup>e</sup>	Retention Time <sup>f</sup>	Residence Time <sup>g</sup>	Equivalent Cluster Site <sup>h,i</sup>	Equivalent Water Position in A Site of 1SFT?	Equivalent Water Position in B Site of 1SFT?
MD/A site										
X18	5654	4.0	5	63	50.0	1000	470	—	+	+
X21	5621	3.5	2	95	75.4	1508	249	—	+	+
S5033	5599	3.2	13	60	47.6	952	219	+(3994)	+	+
X17	5597	3.2	2	126	100.0	2000	Infinite	+(4026)	+	+
V7273	5570	2.9	2	68	54.0	1079	39	—	+	+
X63	5564	2.9	2	85	67.5	1349	1598	—	—	—
X15	5504	2.5	2	97	77.0	1540	1608	—	+	+
X33	5374	2.0	1	125	99.2	1984	1141	+(3923)	+	+
MD/B site										
X158	4026	3.7	1	122	96.8	1937	>2000	+(5597)	+	+
X6	4009	3.4	1	80	63.5	1270	220	—	—	+
X7	3994	3.2	1	126	100.0	2000	Infinite	+(5599)	+	+
S7793	3986	3.1	3	88	69.8	1397	251	—	+	+
X182	3979	2.9	1	63	50.0	1000	94	—	+	+
S4356	3969	2.8	3	70	55.6	1111	323	—	+	—
X139	3923	2.5	2	114	90.5	1810	256	+(5374)	+	+

<sup>a</sup> The ID of the water which appears longest in the cluster. X represents crystal water, and S and V represent water from the SLV1 (Bulky Solvent portion 1) or SLV2 (Bulky Solvent portion 2) segments, respectively (Note: waters are in two segments due to CHARMM's limit). The letter representing the water class is followed by a unique water ID.

<sup>b</sup> Maximum pairwise distance between waters in the cluster (in Å).

<sup>c</sup> Number of different water molecules visiting the cluster area.

<sup>d</sup> The number of frames in which the designated water is present in the cluster.

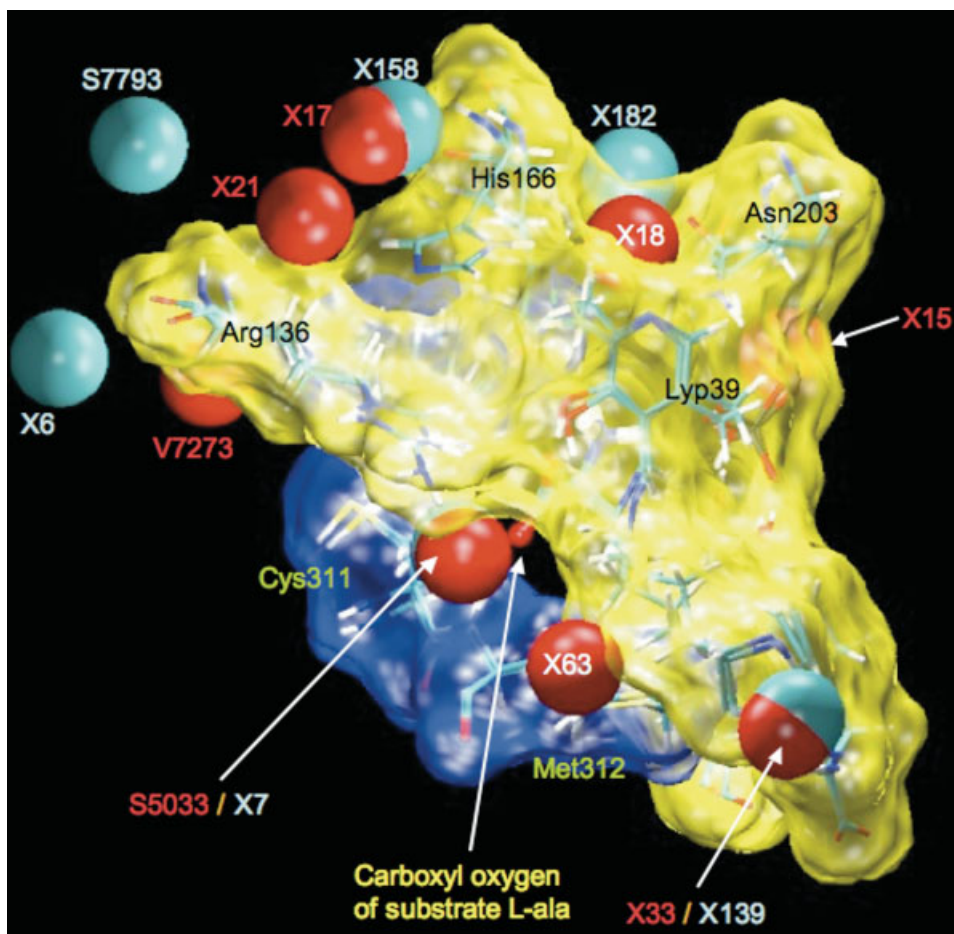
<sup>e</sup> The percentage (%) of studied frames in which the above designated water is seen in the cluster during the 2-ns simulation.

<sup>f</sup> Retention time (in ps) for the water which appears longest.

<sup>g</sup> Residence time (in ps) for the cluster site.

<sup>h</sup> The equivalent cluster site in the other active site.

<sup>i</sup> + for presence and — for absence.



**FIGURE 2** The visualization of conserved hydration sites after the active sites were superimposed. All of the amino acid residues surrounding the active site residues were removed for clarity. The hydration sites are denoted by the IDs of the most-conserved water visiting the site (see Table I). The hydration sites around A-active site are colored in red and the ones around the B-active site are cyan blue.

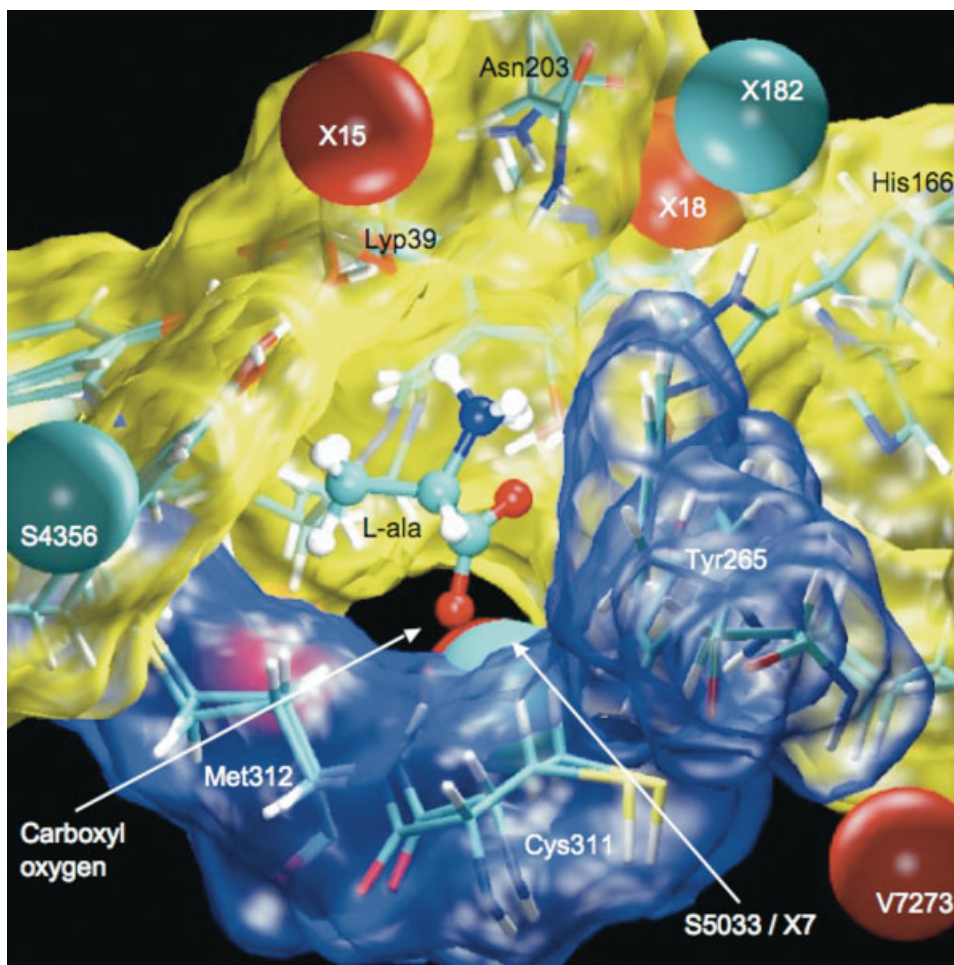
whole MD simulation system. This can cut down the computing cost tremendously.

2. By using coordinates relative to the objects of interest (active site residues, in this case), and by the superimposition (both in building clusters, and comparing active sites A and B) we are able to obtain a much clearer picture of hydration than is obtainable using absolute schemes such as grids, or fixed reference points.
3. The complete linkage algorithm defines rounded clusters. This is an important consideration when examining stable hydration sites, as water molecules have roughly a rounded shape.

The microclusters discovered in this study, along with all the waters which visit them, are shown in Supplementary Tables I and II (for active sites A and B, respectively).

These water lists illuminate how water molecules move in and out of the spaces defined by each cluster in the course of the MD simulation for the 126 time frames examined. These data show that it is sometimes the case that more than one water molecule visits the area defined by each microcluster.

Conserved microclusters for active sites A and B are listed in Table I, respectively, along with all of those waters in each microcluster which appear in the cluster in more than 45% of the examined MD frames. In each microcluster there is only one such water, although it is possible that two waters visit at the same time on the same site with a maximum cluster diameter of 4 Å. The ID of this particular water molecule is used to label the water site defined by the microcluster, giving a sense of which water molecule most often appears in the spot specified by the microcluster (Figures 2 and 3).



**FIGURE 3** Some conserved waters in the A-active site (X15, X18, V7273, and S5033) compared with those in the B-active site (X7, X182, and S4356), near the L-Ala substrate, after superimposition of the sites. The molecule in the center with Ball-and-Stick display is the substrate L-Ala in the B-active site. All of the amino acid residues surrounding the active site residues were removed for clarity. The upper, yellow segment is from the N-terminal domain, and the lower, blue segment is from the C-terminal domain. This figure is viewed as from the back side of the display in Figure 2. The figure was created with VMD<sup>34</sup> and rendered with Raster3D.<sup>36</sup>

### Comparison Between Tightly Bound Water Sites at Active Sites A and B

Superimposition of the amino acid residues of the A-active site on those of the B-active site reveals those microclusters common to both sites (Figure 2). In Table I, the major water microclusters around the active site (determined by MD simulation and cluster analysis) are listed and we identify equivalent sites in the other active site and in the 1SFT X-ray structure. There are a total of three tightly bound water sites common to both active sites in the MD simulation: X17/X158, X33/X139, and S5033/X7, where the first cluster of each pair is found in the A-active site, and the second in the B-active site. These three sites also have equivalent water

positions in both chains of the 1SFT X-ray structure (Table I). Figure 3 shows the S5033/X7 hydration site in detail; this site is in close proximity to the carboxyl group of the substrate L-Ala and is especially important, as that position is known to be crucial to the catalytic function of AlaR.<sup>23,52</sup> The site denoted by S5033 in the A-active site has a short residence time of  $\sim 219$  ps and the site denoted by X7 in the B-active site has a calculated infinite residence time; this indicates that this hydration site became highly stable and conserved (like a part of the protein structure) due to the presence of L-Ala substrate in B-active site. We propose that a potent drug can be designed by integrating a functional group into either the substrate analog or a weak inhibitor of

AlaR, that will displace this highly conserved water site, thus bind tightly to the enzyme and disrupt its function.

### Existence of Highly Conserved Crystal Water Sites

Microclusters identified by our analysis are uniquely associated with the ID of the water with the longest retention time in the microcluster (Table I). Waters are labeled with an “X” if their initial position in the 2-ns MD simulation was defined by a known crystal water position (see Methods). Of the 15 identified microclusters, 11 were visited most often by waters which were initiated from crystal water positions, and 10 of these 11 sites have equivalent sites occupied by crystal waters in the 1SFT structure (Table I). The one exception is the position occupied by X63 in the A-active site; there is no crystal water in that position in either of the active sites of the X-ray structure, i.e., X63 was initiated from a crystal water position, but it moved to and stayed in a noncrystal water position for 67.5% of the time during 2-ns simulation (Table I).

In the three microclusters represented by X17/X158, X33/X139, and S5033/X7, a crystal water was localized to the equivalent position in both active sites of the X-ray structure. These three positions are highly conserved and tightly-bound sites. They appear in the X-ray structure as the crystal waters A516/B766, A532/B733, and A506/B765, respectively. Waters labeled with an “A” here are in the first chain of the 1SFT PDB structure (monomer subunit 1), while those labeled with a “B” is in the second PDB chain (monomer subunit 2). As seen from Table I, those waters that appear longest in this set of three microclusters, with the exception of S5033, appear in their respective clusters for more than 90% of the 2-ns period. This result suggests that our methodology is robust, as our method is able to detect some of the crystal waters common to active sites A and B, and that those waters are very tightly bound. Additionally, it can detect water clusters not identified in the X-ray structure. For example, cluster site 5564 in the A-active site is not associated with crystal waters in the X-ray structure.

The highly conserved water X7 is present in 100% of the frames in the B-active site, and the mildly conserved water S5033 is present in 47.6% of the frames in the equivalent position in the A-active site. As both are very close to their respective active sites, it is highly likely that the difference in retention time is due to presence of L-Ala in the B-active site. The X7 water position is highly stabilized between substrate and protein, so a modified substrate analog (e.g., D-cycloserine) that contains an extra atom or functional group that would displace this water site is predicted to bind strongly in

the active site of AlaR and thus act as a potent inhibitor of enzymatic function.

We also note some changes in water positions due to the presence of substrate in the B-active site. For example, the positions represented by X15, X18, X21, and V7273 in the A-active site did not have waters with long retention times in the equivalent positions in the B-active site even though those positions were found in both chains of 1SFT X-ray structure (Table I). It seems that these water sites were disturbed by the presence of substrate in the active site. The structural environment of these tightly bound waters, and the conformational changes to the enzyme caused by the presence of the substrate, may be studied further to better understand the catalytic mechanism of AlaR. This knowledge may assist in the design process for inhibitor drugs targeting this enzyme.

### Retention Times and Residence Times

Residence time, as defined by Brunne et al.,<sup>17</sup> is often a useful way to analyze and understand the presence of waters in clusters. Below, we discuss some of the limitations of this method. In clusters which have only one water molecule, especially one which is tightly bound, the decaying exponential model is not a good fit to  $P(t)$ . For example, the cluster in the A-active site represented by X33, has only that single water molecule as a member of the cluster. It appears in 125 of the 126 frames that we examined. However, it does not appear in frame 91; and this single absence produces a curve that is a poor fit to a decaying exponential. This phenomenon occurs several times, in several of the clusters. The problem can be remedied by using one of the variants of residence time, where we count a molecule as present in a cluster for the entirety of the desired time step, even if it has actually been absent for some prespecified fraction of the time step.<sup>47</sup>

The logic and interpretation of retention time is straightforward, and this interpretation is sometimes markedly different from the interpretation of measures of residence time. For example, consider the microclusters represented by X21 and S5033 in the A-active site, and the microclusters represented by X6 and S7793 in the B-active site. For all four,  $P(t)$  is reasonably well fit by a decaying exponential, and the four resulting residence times are very similar; 249, 219, 220, and 251, respectively. Their retention times, however, are 1508, 952, 1270, and 1397, respectively. In this instance, retention time has a much greater variance than residence time. This difference together with the simple interpretation of retention time argues that it is useful to consider both residence time and retention time in the study of water clusters.

It may be argued that the free energy of buried water thermodynamics (although difficult to measure) is ultimately what determines which water molecule should be displaced in ligand design, or that kinetic properties like residence time may show some correlation with free energy measurement, but that stable waters with long residence times may just be trapped in the protein structure. Whether the water is stably located or forcibly trapped in the hydration site, both the free energy and measures of kinetics should be considered. Accessibility of binding sites is an important factor in ligand binding that is not reflected in the free energy of binding. Retention time is also not expected to be greatly affected by the trapping of water molecules in protein active sites, due to the escape of the water from these sites after the protein conformation is relaxed during the MD simulation. Therefore, the measure of retention time for specific waters offers a useful alternative evaluation of binding stability for waters of interest.

## CONCLUSION

Complete linkage clustering is appropriate for analyzing water sites on the surface of X-ray resolved biomolecular structures. It can also be utilized to analyze the large number of animated waters present during MD simulation. In our study, we discovered that certain waters visited the same area around the active sites of AlaR more than 45% of the simulation time. Most waters, however, moved through the sites randomly.

The conserved and tightly bound waters that we discovered occupy hydration sites inside the protein that are structurally or functionally important for the conformation and catalytic mechanisms of AlaR due to their proximity to the active site area. Of the conserved clusters, there are three pairs of clusters of particular interest: in each pair one cluster appears in the A-active site, while the other appears in the B-active site, in the equivalent position. One such pair of clusters, occupied by S5033 in the A-site and X7 in the B-site, is proposed to be especially important due to its close proximity to the functional regions of AlaR, around L-Ala and the PLP cofactor on Lys39. We propose that a potent inhibitory drug can be designed by linking a functional group in this position to the ligand for AlaR, and that a virtual screening process can be designed for this purpose in the future.

We also determined that hydration sites with similar residence times can be further characterized with another measure, the retention time of the most conserved water visiting the site. We predict that the best target hydration sites around the active sites for structure-based drug design are those that have a long residence time and include a water with a long retention time.

Researchers previously needed several homologous X-ray structures for a protein of interest in order to study hydration sites. We conclude from this study that the conserved and tightly bound water sites around protein active sites can be identified by the rich structural information produced by MD simulation performed using a single X-ray structure.

Authors would like to thank San Diego Supercomputer Center for the supercomputing time on MD simulation in this project. Salary support for H.C.H. was provided by start-up funds to V.V. from the Department of Systems Biology and Translational Medicine and the Dean of the College of Medicine, Texas A & M Health Science Center.

## REFERENCES

- Shaw, J. P.; Petsko, G. A.; Ringe, D. *Biochemistry* 1997, 36, 1329–1342.
- Eisenberg, D.; McLachlan, A. D. *Nature* 1986, 319, 199–203.
- Kuntz, I. D., Jr.; Kauzmann, W. *Adv Protein Chem* 1974, 28, 239–345.
- Raschke, T. M. *Curr Opin Struct Biol* 2006, 16, 152–159.
- Levy, Y.; Onuchic, J. N. *Annu Rev Biophys Biomol Struct* 2006, 35, 389–415.
- Helms, V. *Chemphyschem* 2007, 8, 23–33.
- Otting, G.; Liepinsh, E.; Wuthrich, K. *Science* 1991, 254, 974–980.
- Garcia, A. E.; Hummer, G. *Proteins* 2000, 38, 261–272.
- Joachimiak, A.; Haran, T. E.; Sigler, P. B. *EMBO J* 1994, 13, 367–372.
- Wilson, I. A.; Fremont, D. H. *Semin Immunol* 1993, 5, 75–80.
- Fauman, E. B.; Rutenber, E. E.; Maley, G. F.; Maley, F.; Stroud, R. M. *Biochemistry* 1994, 33, 1502–1511.
- Wlodawer, A.; Miller, M.; Jaskolski, M.; Sathyanarayana, B. K.; Baldwin, E.; Weber, I. T.; Selk, L. M.; Clawson, L.; Schneider, J.; Kent, S. B. *Science* 1989, 245, 616–621.
- Lam, P. Y.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C. H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. *Science* 1994, 263, 380–384.
- Marrone, T. J.; Briggs, J. M.; McCammon, J. A. *Annu Rev Pharmacol Toxicol* 1997, 37, 71–90.
- Garcia-Sosa, A. T.; Firth-Clark, S.; Mancera, R. L. *J Chem Inform Model* 2005, 45, 624–633.
- Garcia-Sosa, A. T.; Mancera, R. L. *J Mol Model* 2006, 12, 422–431.
- Brunne, R. M.; Liepinsh, E.; Otting, G.; Wuthrich, K.; van Gunsteren, W. F. *J Mol Biol* 1993, 231, 1040–1048.
- Lounnas, V.; Pettitt, B. M. *Proteins* 1994, 18, 133–147.
- Lounnas, V.; Pettitt, B. M. *Proteins* 1994, 18, 148–160.
- Sanschagrin, P. C.; Kuhn, L. A. *Protein Sci* 1998, 7, 2054–2064.
- Mustata, G.; Briggs, J. M. *Protein Eng Des Sel* 2004, 17, 223–234.
- Mustata, G. I.; Soares, T. A.; Briggs, J. M. *Biopolymers* 2003, 70, 186–200.
- Stamper, G. F.; Morollo, A. A.; Ringe, D. *Biochemistry* 1998, 37, 10438–10445.

24. Darden, T. A.; York, D. M.; Pedersen, L. G. *J Chem Phys* 1993, 98, 10089–10092.
25. William, L. J.; Jayaraman, C.; Jeffry, D. M.; Roger, W. I.; Michael, L. K. *J Chem Phys* 1983, 79, 926–935.
26. Brooks, B.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J Comput Chem* 1983, 4, 187–217.
27. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J Phys Chem B* 1998, 102, 3586–3616.
28. Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J Comput Phys* 1977, 23, 327–341.
29. Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Oxford University Press: New York, 1987.
30. Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Kraetz, N.; Phillips, J. C.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J Comput Phys* 1999, 151, 283–312.
31. Barton, G. J. OC—A cluster analysis program; University of Dundee, Scotland, UK, 1993, 2002, [www.compbio.dundee.ac.uk/downloads/oc](http://www.compbio.dundee.ac.uk/downloads/oc).
32. Kabsch, W. *Acta Crystallogr Sect A* 1976, 32, 922–923.
33. Kabsch, W. *Acta Crystallogr Sect A* 1978, 34, 827–828.
34. Humphrey, W.; Dalke, A.; Schulten, K. *J Mol Graph* 1996, 14, 33–38.
35. Kraulis, P. J. *J Appl Crystallogr* 1991, 24, 946–950.
36. Merritt, E. A.; Murphy, M. E. *Acta Crystallogr Sect D* 1994, 50, 869–873.
37. Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. *J Mol Biol* 1997, 265, 445–464.
38. Rocchi, C.; Bizzarri, A. R.; Cannistraro, S. *Chem Phys* 1997, 214, 261–276.
39. Schoenborn, B. P.; Garcia, A.; Knott, R. *Prog Biophys Mol Biol* 1995, 64, 105–119.
40. Makarov, V. A.; Andrews, B. K.; Smith, P. E.; Pettitt, B. M. *Biophys J* 2000, 79, 2966–2974.
41. Bruge, F.; Parisi, E.; Fornili, S. L. *Chem Phys Lett* 1996, 250, 443–449.
42. Wei Gu, B. P. S. *Proteins: Struct Funct Genet* 1995, 22, 20–26.
43. Muegge, I.; Knapp, E. W. *J Phys Chem* 1995, 99, 1371–1374.
44. Phillips, G. N., Jr.; Pettitt, B. M. *Protein Sci* 1995, 4, 149–158.
45. Impey, R. W.; Madden, P. A.; McDonald, I. R. *J Phys Chem* 1983, 87, 5071–5083.
46. Garcia, A. E.; Stiller, L. *J Comput Chem* 1993, 14, 1396–1406.
47. Luise, A.; Falconi, M.; Desideri, A. *Proteins* 2000, 39, 56–67.
48. Likic, V. A.; Prendergast, F. G. *Proteins* 2001, 43, 65–72.
49. Sterpone, F.; Ceccarelli, M.; Marchi, M. *J Mol Biol* 2001, 311, 409–419.
50. Henchman, R. H.; Tai, K.; Shen, T.; McCammon, J. A. *Biophys J* 2002, 82, 2671–2682.
51. Damjanovic, A.; Garcia-Moreno, B.; Lattman, E. E.; Garcia, A. E. *Proteins* 2005, 60, 433–449.
52. Watanabe, A.; Yoshimura, T.; Mikami, B.; Hayashi, H.; Kagamiyama, H.; Esaki, N. *J Biol Chem* 2002, 277, 19166–19172.

*Reviewing Editor: J. McCammon*