J|A|C|S COMMUNICATIONS

Published on Web 12/15/2010

Full Protein Flexibility Is Essential for Proper Hot-Spot Mapping

Katrina W. Lexa and Heather A. Carlson*

Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1065, United States

Received September 2, 2010; E-mail: carlsonh@umich.edu

Abstract: A traditional technique for structure-based drug design (SBDD) is mapping of protein surfaces with probe molecules to identify "hot spots" where key functional groups can best complement the receptor. Common methods, such as minimization of probes or calculation of grids, use a fixed protein structure in the gas phase, ignoring both protein flexibility and proper competition between the probes and water. As a result, the potential surface is quite rugged, and many spurious local minima are identified. In this work, we compared rigid and fully flexible proteins in mixed-solvent molecular dynamics, which allows for flexibility and full solvent effects. We were surprised to find that the large number of local minima are still found when a protein's conformational sampling is restricted; the dynamic averaging of probes and competition with water do not smooth the potential surface as one might expect. Only when a protein is allowed to be fully flexible in the simulation are the proper minima located and the spurious ones eliminated. Our results indicate that inclusion of full protein flexibility is critical to accurate hot-spot mapping for SBDD.

Protein flexibility is an important component of protein−ligand binding, but it is often neglected in structure-based drug design (SBDD). Many traditional techniques for SBDD rely on solvent mapping performed through grids or probe minimization. Most computational solvent-mapping techniques[1−4] do not account for the impact of protein flexibility on ligand binding, which prevents accurate mapping of hot spots. Also, they typically do not allow for active competition between the solvent probes and water and thus ignore proper solvation effects. In this communication, we demonstrate that the conformational diversity inherent to proteins strongly affects the outcome of hot-spot mapping.

An experimental method that explores protein surfaces using water and organic solvent as probes is the multiple-solvent crystal structure (MSCS)[5] technique. Potential protein unfolding is typically prevented through cross-linking. The results of this procedure obtained with various solvents can be superimposed to design custom ligands by linking fragments. We have developed a protocol for using mixed-solvent molecular dynamics (MixMD) to map hot spots in a manner similar to MSCS. Our multiple protein structure (MPS) method[6−8] for creating binding-site pharmacophore models based on conformational ensembles has demonstrated success in mapping protein systems for drug design.[9,10] MixMD expands the MPS concept to simultaneously allow protein flexibility and competition between the probes and water.

Several similar efforts have incorporated MSCS concepts into computational methods, but each has notable limitations. FTMap[11] is modeled after MSCS, but while it can be used with ensembles like MPS,[12] neither ligand nor on-the-fly protein flexibility is used during probe mapping. A recent study from Seco et al.[13] utilized MD with mixed water and isopropyl alcohol to detect binding sites and predict potential druggability. However, the method was unable to reproduce many known binding sites. SILCS is a mapping method that incorporates a ternary solvent system (benzene, propane, and water) with MD to map sites.[14] However, the method was validated on only one protein for which MSCS data to make a proper evaluation was not available. Therefore, these methods are in their infancy and require significant development to provide a robust tool for SBDD. This study presents initial findings based on our MixMD protocol that should have a significant impact for others developing similar techniques.

Hen egg-white lysozyme (HEWL) is a canonical model system that allows for appropriate testing and validation of MixMD to identify hot spots. An MSCS of HEWL was produced using acetonitrile (CCN) as the organic solvent.[15] The high-quality electron density available for this structure allows an accurate assessment of MixMD data to be made. Below, we demonstrate how occupancy grids for both the probe and water can be directly compared to electron density.

MixMD simulations. The starting structure of HEWL in CCN and water (PDB entry 2LYO)[15] was obtained from the Protein Data Bank.[16] We performed all-atom MD simulations of the HEWL protein in the presence of multiple solvents using standard procedures for *sander* in Amber 10[17] at 300 K [see the Supporting Information (SI) for detailed methods]. Pre-equilibrated solvent boxes with an even distribution of 50% (w/w) CCN and water were used. The simulation setup was completed in tLeAP using the ff99SB force field,[18] TIP3P water,[19] neutralizing ions, a 10 Å van der Waals cutoff, and CCN parameters from Grabuleda et al.[20] A time step of 2 fs was implemented. The temperature was controlled through an Anderson thermostat,[21] and SHAKE was applied. Three different protocols for protein flexibility were evaluated for proper sampling and convergence: all-atom-restrained, backbone-restrained, and fully flexible HEWL. Five independent simulations with 10 ns of production time each were performed for every system, initiated from the same solvent configuration. Though it might have enhanced the sampling to have a different starting location for the solvent in each simulation, this would have made it more difficult for us to properly evaluate convergence in the simulations.

Prediction of Hot Spots. The positions of the solvent from the *sander* trajectories were converted into occupancy grids using *ptraj*. In this way, we were able to directly compare our solvent "density" results to electron density data obtained in the crystallography study. This allowed for an equivalent comparison of solvent positions during the simulation with the solvent occupancy from crystal studies, which is a more even assessment than simply using the solvent coordinates given. (In Figures 1 and 2, crystallographic coordinates for CCN and water have often been used in place of electron density to avoid confusion arising from overlaying many grids.) Technically, the data most equivalent to crystallographic density would be an occupancy grid based on all atoms of the simulation (protein, water, CCN, and counterions), but we have made the simplification of examining only solvent-occupancy grids.

10.1021/ja1079332

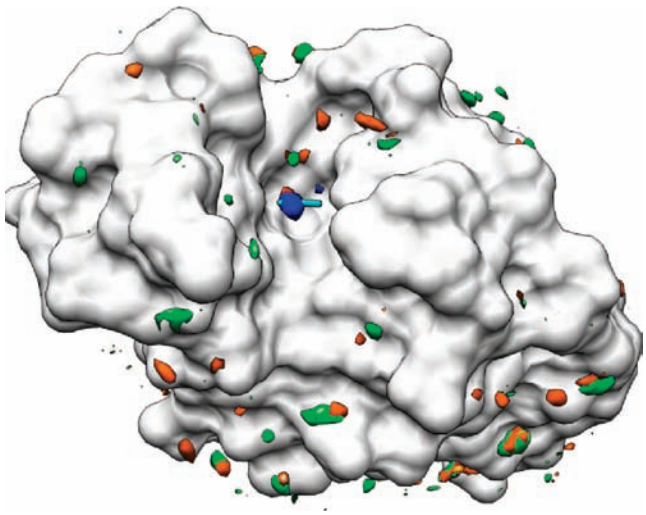

Figure 1. Results from unrestrained vs restrained protein simulations using CCN and water to solvate HEWL (white). The single hot spot identified by MSCS is shown in stick; CCN (cyan). The probe density from the fully restrained simulation is shown in orange, the backbone-restrained density in green, and fully flexible density in blue. Many incorrect local minima in green and orange can be seen, but the correct position alone dominates the blue data from the simulation of the fully flexible protein.

In our initial simulation using mobile solvent and a fixed protein, we aimed to establish a minimum sampling time required for the solvent to reproduce the MSCS results. We assumed that the mapping would identify the position for CCN and that longer sampling times would be required as more flexibility was allowed for the protein. Instead, we were surprised to find that our simulation of the rigid protein converged to multiple, trivial minima (Figure 1). Though the CCN hot spot in the crystal structure was mapped with weak occupancy, it was equal to or less than many incorrect sites. When we added side-chain flexibility (backbone still fixed), a variety of incorrect sites were again located, but the correct location was not! Only the inclusion of full protein flexibility afforded the correct location for the CCN hot spot and eliminated the trivial minima.

It appears that the numerous local minima obtained when gas-phase minimizations of probe molecules are performed are not an artifact of the vacuum; instead, they are an artifact of using a rigid protein conformation. A rugged landscape is observed, even in the presence of mobile solvent and side chains. The abundant local minima cannot be distinguished from the binding site, and probe mapping cannot successfully differentiate between irrelevant and druggable hot spots. With full receptor flexibility included, MixMD appropriately reproduces the one hot-spot binding site seen in the crystallographic data for CCN. The agreement between the simulation data and experimental electron density validates MixMD as an accurate mapping tool (Figure 2).

In addition to the CCN hot spot, MixMD reproduced the locations of low-B-factor (<33 Å) water. The only locations that were not reproduced were on surfaces of the protein that were involved in crystallographic contacts (Figure 2B). A few locations were seen where significant water occupancy in the interior of the protein did not correlate with water coordinates in the crystal structure, but those locations were in excellent agreement with unfulfilled density in the crystal structure (Figure 2A). The location of positive density on the $F_o - F_c$ map may in fact correspond to water positions. While not all of the unfulfilled density corresponds to solvent molecules, the locations identified by MixMD water maps may indicate positions where water should have been placed.

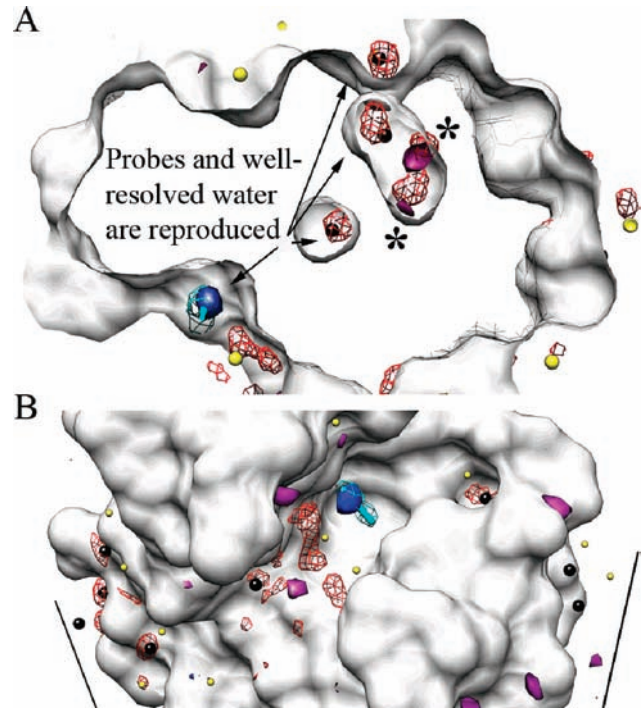

Figure 2. MixMD data from the fully flexible simulation agrees well with the densities from MSCS experiments using CCN and water. All of the snapshots have been superimposed on the crystal structure of HEWL (white surface). The MixMD density for water is shown in red mesh, and crystallographic waters are colored black and yellow for B-factors below and above 33 Å, respectively. The MSCS coordinates for CCN are shown in stick form (cyan with its electron density in mesh, $2F_o - F_c$ shown at 1.5σ); the highest occupancy for CCN in the fully flexible simulation matches perfectly and is shown as the solid blue surface. Unsatisfied electron density in the crystal structure (positive $F_o - F_c$ shown at 3σ) is shown in solid purple. (A) Water maps within the protein highlight interior waters conserved in the crystal structure and reproduced in our simulation. The large asterisks (*) denote two highly occupied regions of the interior water map that correspond to unfulfilled density in the crystal structure. (B) Maps of the protein surface show good correspondence between the crystallographic and MixMD densities for both CCN and well-resolved waters. For the 16 crystallographic waters with B-factors below 33 Å (black spheres), five occur at symmetry-packing interfaces. MixMD missed four of these five, which was expected because the contacts were not present in our simulation. The other 11 best-resolved waters are well-reproduced.

Convergence of Sampling. Though the 10 ns sampling time used in the simulations is relatively short by current standards, it is important to stress that long trajectories are inappropriate in a mixed solvent. Modest time scales (long enough to allow solvent equilibration and convergence but short enough to avoid possible unfolding of the protein) are needed. Furthermore, an accurate MD technique built on short time scales makes this method more accessible for practical applications in a pharmaceutical setting.

We calculated the maximal occupancy location of each probe type during each individual simulation using the *ptraj* grid utility. These positions for CCN over the last 2 ns of production were compared among independent simulations of the same initial system (see Figure S1 in the SI). Excellent convergence across the five independent MixMD simulations of the fully flexible HEWL was seen. However, the individual simulations of the rigid and backbone-fixed simulations did not agree on a common location for the CCN hot spot, reflecting a propensity for solvent molecules to become trapped within local minima along the protein surface. For the fully flexible simulation, these points were all within <1 Å, which is within the limits of accuracy when a 0.5 Å grid is used. Not only did the locations agree with one another, they were in excellent

agreement with the position for CCN in the crystal structure. In contrast, there was no agreement between the five independent MixMD simulations of the rigid and backbone-fixed HEWL (see Figure S1). Those simulations also failed to identify the correct location for the CCN hot spot, except for one trajectory of the rigid HEWL.

To further evaluate the sampling, we calculated the ratio of the number of solvent probes to water molecules at the edges of the box. Far from the protein, there should be no bias between the solvents, and the ratio of their occupancies should approach N_p/N_w, the ratio of the number of CCN probes to the number of waters in the simulation.[22] All of the systems demonstrated good convergence according to this metric, with the fully flexible system being the least biased (see Tables S1 and S2 in the SI). The fact that CCN and water exchange freely at the box edge indicates that the mixed solvent system inherently samples evenly, but the pronounced differences at the protein surface indicate that solvent molecules become trapped and poorly sample the rugged potential surface of a rigid or semirigid protein.

Initial Preparation of the Mixed-solvent Environment. The above results were obtained with a pre-equilibrated 50% (w/w) solution, but we also examined other choices for the mixed-solvent environment. Two protocols for initiating the mixed-solvent box were compared. The first used the pre-equilibrated 50% (w/w) mixed solution, providing an even distribution of the two solvents (data shown above). The second method aimed to reproduce the MSCS experiment, where the CCN has to displace water from the surface of the protein. The waters were placed in a shell around the protein, and the CCN molecules were placed outside the water shell, resulting in a layered solvent environment.

The densities of CCN obtained from the two solvent protocols showed good agreement (see Figure S2). Maximal occupancy positions were used to compare the coordinates of the experimental and simulation probes. For simulations of fully flexible HEWL, we found that the layered solvent produced a maximally occupied location 0.8 Å from the crystallographic C2 atom of CCN. The pre-equilibrated, evenly mixed solvent produced a maximally occupied location 0.9 Å from the crystallographic C2 atom of CCN. These maximally occupied locations were 0.5 Å away from each other. Again, this is within the limits of error of our grids for calculating the occupancy maps. It appears that either protocol may be appropriate for 50% (w/w) CCN and water, but the layered solvent showed a slight disagreement in the convergence of the five independent simulations (see Figure S3).

We also examined 90% and 10% (w/w) mixed solutions of water and CCN to determine whether maps are more accurate when more or fewer probes are present. Both the 90% and 10% mixtures identified the correct hot spot for CCN (see Figure S4). However, we found that the 50% mixtures gave better water maps and more complete sampling than either 90% or 10% mixtures of CCN and water (see Figure S5).

Conclusion. Our results demonstrate the need to include protein flexibility to achieve valid hot-spot mapping. MixMD simulations were successfully performed to determine the correct mapping procedure for locating truly relevant binding minima. MixMD was capable of locating hot spots for the CCN solvent probe, and it identified crystallographic waters with the lowest B-factors, crystal contact waters, and locations where water could have been modeled into the structure (unsatisfied density in the $F_o - F_c$ map). The information contained within individual MixMD trajectories can be combined into a consensus model retaining only the consistently important mapped sites. We have shown that only through the incorporation of protein flexibility and appropriate solvent competition can viable mapping results be obtained.

Acknowledgment. We thank Jeanne Stuckey and Thomas Goddard for their advice and assistance in interpreting electron density data. This work was supported by the National Institutes of Health (GM65372). K.W.L. thanks Rackham Graduate School, the Pharmacological Sciences Training Program (GM07767), and the American Foundation for Pharmaceutical Education for funding. Molecular graphics images were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco.

Supporting Information Available: Supplementary analysis, including detailed methods, additional data, and probe parameters, and complete ref 17. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Stultz, C. M.; Karplus, M. *Proteins* **1999**, *37*, 512–529.
(2) Dennis, S.; Kortvelyesi, T.; Vajda, S. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290–4295.
(3) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849–857.
(4) Guarnieri, F.; Mezei, M. *J. Am. Chem. Soc.* **1996**, *118*, 8493–8494.
(5) Mattos, C.; Ringe, D. *Nat. Biotechnol.* **1996**, *14*, 595–599.
(6) Meagher, K. L.; Carlson, H. A. *J. Am. Chem. Soc.* **2004**, *126*, 13276–13281.
(7) Bowman, A. L.; Lerner, M. G.; Carlson, H. A. *J. Am. Chem. Soc.* **2007**, *129*, 3634–3640.
(8) Meagher, K. L.; Lerner, M. G.; Carlson, H. A. *J. Med. Chem.* **2006**, *49*, 3478–3484.
(9) Damm, K. L.; Ung, P. M.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. *Biopolymers* **2008**, *89*, 643–652.
(10) Bowman, A. L.; Nikolovska-Coleska, Z.; Zhong, H.; Wang, S.; Carlson, H. A. *J. Am. Chem. Soc.* **2007**, *129*, 12809–12814.
(11) Brenke, R.; Kozakov, D.; Chuang, G. Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. *Bioinformatics* **2009**, *25*, 621–627.
(12) Landon, M. R.; Amaro, R. E.; Baron, R.; Ngan, C. H.; Ozonoff, D.; McCammon, J. A.; Vajda, S. *Chem. Biol. Drug Des.* **2008**, *71*, 106–16.
(13) Seco, J.; Luque, F. J.; Barril, X. *J. Med. Chem.* **2009**, *52*, 2363–2371.
(14) Guvench, O.; MacKerell, A. D., Jr. *PLoS Comput. Biol* **2009**, *5*, e1000435.
(15) Wang, Z.; Zhu, G.; Huang, Q.; Qian, M.; Shao, M.; Jia, Y.; Tang, Y. *Biochim. Biophys. Acta* **1998**, *1384*, 335–344.
(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(17) Case, D. A.; et al. *Amber 10*; University of California: San Francisco, 2008.
(18) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 712–25.
(19) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(20) Grabuleda, X.; Jaime, C.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 901–908.
(21) Andrea, T. A.; Swope, W. C.; Andersen, H. C. *J. Chem. Phys.* **1983**, *79*, 4576–4584.
(22) Aburi, M.; Smith, P. E. *J. Chem. Phys.* **2004**, *108*, 7382–7388.

JA1079332