JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Diverse Fragment Clustering and Water Exclusion Identify Protein Hot Spots

John L. Kulp, III,[†,⊥] John L. Kulp, Jr.,[‡] David L. Pompliano,[‡] and Frank Guarnieri*[,‡,∥,§]

[†]Chemistry Division, Naval Research Laboratory, Washington, D.C. 20375-5342, United States
[‡]BioLeap, Inc., 238 West Delaware Avenue, Pennington, New Jersey 08534, United States
[∥]Department of Physiology and Biophysics, Virginia Commonwealth University, Richmond, Virginia 23298, United States
[§]Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02218, United States

Supporting Information

ABSTRACT: Simulated annealing of chemical potential located the highest affinity positions of eight organic probes and water on eight static structures of hen egg white lysozyme (HEWL) in various conformational states. In all HELW conformations, a diverse set of organic probes clustered in the known binding site (hot spot). Fragment clusters at other locations were excluded by tightly-bound waters so that only the hot-spot cluster remained in each case. The location of the hot spot was correctly predicted irrespective of the protein conformation and without accounting for protein flexibility during the simulations. Any one of the static structures could have been used to locate the hot spot. A site on a protein where a diversity of organic probes is calculated to cluster, but where water specifically does not bind, identifies a potential small-molecule binding site or protein−protein interaction hot spot.

Protein−protein, protein−DNA, and protein−ligand interactions are often governed by focused regions of significant binding affinity, commonly referred to as "hot spots". Locating and characterizing these structural subsets of the interaction surface would inform efforts to inhibit such associations, a key therapeutic aim for many diseases. While protein flexibility appears to be an important consideration for problems such as quantitative assessment of ligand binding, it is not clear whether an expensive exploration of a broad range of protein conformations is necessary for robust hot-spot identification. Using a grand canonical Monte Carlo method,[1] we calculated where small organic fragments clustered on the surface of hen egg white lysozyme (HEWL), and found that the location of the highest affinity cluster was at the sugar substrate binding site of HEWL. In contrast to a recent report by Carlson,[2] our computational method for finding hot spots was insensitive to protein flexibility.

Hot spots correspond to sites on the protein where multiple diverse small organic molecules cluster, as established by the experimental studies of protein structures in various organic solvents by the Ringe[3] and Fesik[4] laboratories. In these studies, a solitary fragment was usually found to bind to multiple sites on the protein. Ringe et al.[5] found that acetonitrile, for example, binds to nine different sites on elastase. The overlap of fragment binding sites from different experiments also produces multiple fragment clusters. The challenge is to distinguish which of these clusters corresponds to hot spots. The experimental methods are expensive, in both time and equipment. Computational approaches for predicting hot spots are maturing.[2,6,7] The success criteria for such methods are robustness across protein families, minimum production of false positives, and the ease and celerity of implementation.

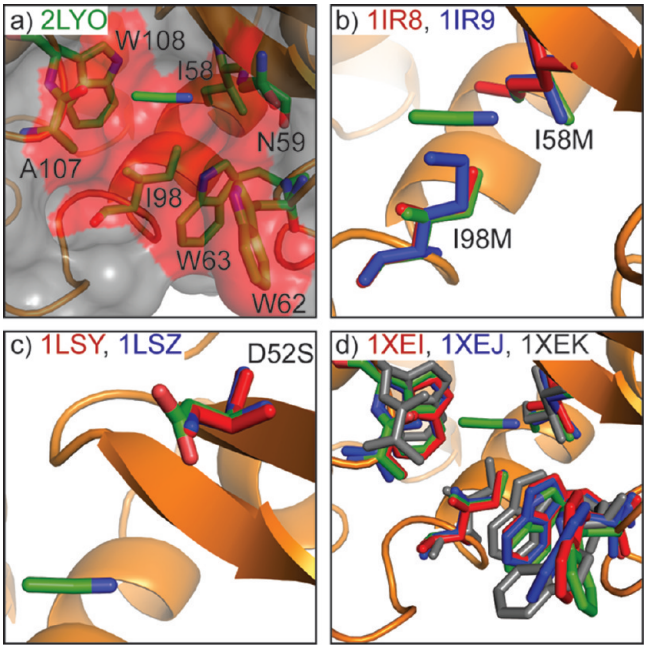

Figure 1. Structural overlays represent conformational changes in the binding pocket of HEWL. (a) The binding site of HEWL highlighting the hot spot (red) and key amino acid residues involved in binding substrate.[8] (b−d) Each structure aligned to 2LYO (green) with emphasis on the conformational changes within the hot spot; (b) 1IR8 (red, I58M mutation) and 1IR9 (blue, I98M mutation), (c) 1LSY (red, D52S mutation) and 1LSZ (blue, D52S mutation with sugar ligand bound), (d) 1XEI (red, 17.6% hydrated), 1XEJ (blue, 16.9% hydrated), and 1XEK (gray, 9.4% hydrated). The surface, ribbon, and the acetonitrile found in the HEWL structure were rendered with 2LYO structure.

Our hypothesis is that successful computational hot spot identification requires (i) distinguishing higher- versus lower-affinity sites using a free-energy based ranking of fragment

Received: April 28, 2011
Published: June 17, 2011

dx.doi.org/10.1021/ja203929x | J. Am. Chem. Soc. 2011, 133, 10740–10743

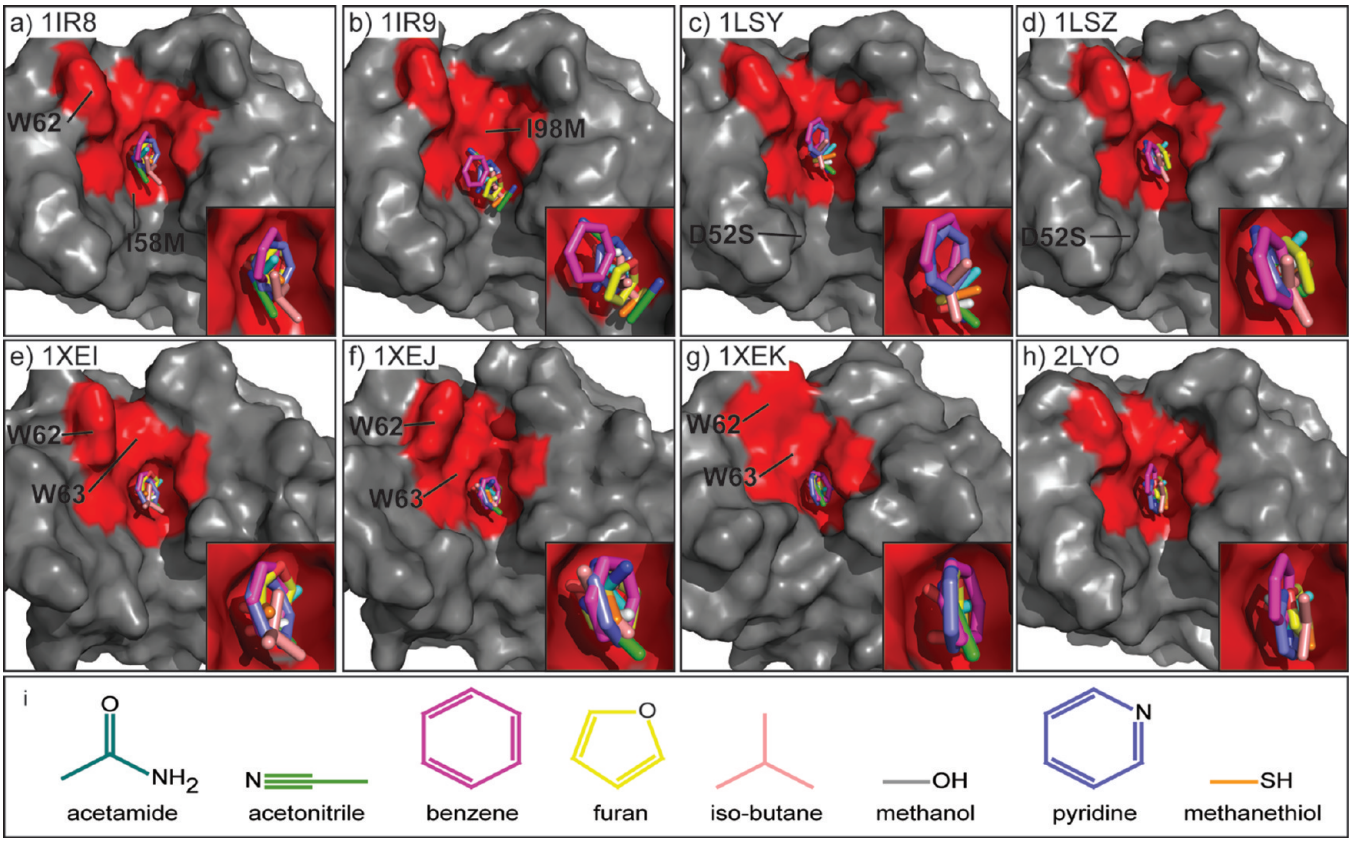

Figure 2. Fragments cluster at a single site in the hot spot (highlighted red), the site where most of the key binding interactions between the protein and sugar substrate occur. Panels show clusters of fragments for each structure: (a) 1IR8, (b) 1IR9, (c) 1LSY, (d) 1LSZ (e) 1XEI, (f) 1XEJ, (g) 1XEK, and (h) 2YLO. (i) Fragment structures are uniquely colored at each carbon atom.

binding, (ii) clustering of chemically diverse fragments, and (iii) excluding sites where tightly bound waters block fragment binding that otherwise might appear to achieve high affinity. Further, omitting one or more of these factors is the key source of the false positives. The degree of sensitivity to protein conformational variability is reflected in its impact on the above requirements.

An interesting exception to the requirement for a diversity of probes in identifying hot spots is the work by Wang and colleagues who crystallized HEWL in a solution of acetonitrile.[9] In a solvent mixture of 90% acetonitrile and 10% water, a single acetonitrile molecule was bound at the center of the sugar binding pocket, the location where interactions with the protein provide most of the binding energy for the natural sugar ligand. While not representative of a biologically relevant environment, this unambiguous structure is a good test case for an algorithm designed to predict the propensity of organic fragments to bind to high affinity protein sites. Initially using the Wang structure,[9] we additionally simulated seven other structures that range in conformational and mutational states.

Lexa and Carlson[2] reported that full protein flexibility of HEWL is required in order to reproduce acetonitrile binding results[9] when using molecular dynamics simulations and occupancy as an affinity metric.[2] We wondered whether the dependency of the method upon protein flexibility applies in general to other computational clustering methods.

To address this question, we asked if the small-molecule binding site(s) identified in a Monte Carlo-based clustering method would depend upon the protein conformer used in the simulation. Our assumption is that protein flexibility can be equated to a series of discrete protein conformations, strictly valid in the limit of a large number of conformations. From the Protein Data Bank[10] we collected eight structures of HEWL, which contained conformational and mutational variations in the binding site: (1) cocrystallized with acetonitrile (Figure 1a); (2 and 3), two different isoleucine to methionine mutations at the binding site (Figure 1b); (4 and 5) the catalytic D52S mutation with and without bound ligand (Figure 1c); and (6−8) three wild-type structures induced into significantly different binding site conformational states by dehydration (Figure 1d). Ohmure and co-workers confirmed that several isoleucine residues in the binding site of HEWL could tolerate methionine residue substitutions (Figure 1b).[11] Hadfield and colleagues demonstrated that mutating the catalytic aspartate residue to serine residue virtually eliminated the enzymatic activity of HEWL while maintaining ligand binding similar to the wild-type protein (Figure 1c).[12] Nagendra and co-workers obtained a set of HEWL crystal structures at different hydration levels (Figure 1d).[13] They demonstrated that the dramatic collapsing of the binding pocket as a function of dehydration mirrored the conformational changes observed when the protein was carrying out its enzymatic function. This shed light on the interplay between water binding and catalysis. Figure 1 demonstrates a range of conformational states adopted by these various HEWL structures overlaid on the structure[9] used in the Lexa and Carlson study.[2]

Each annealing of chemical potential (ACPS) consists of a sequence of grand canonical ensemble Monte Carlo simulations

dx.doi.org/10.1021/ja203929x |J. Am. Chem. Soc. 2011, 133, 10740–10743

and will determine where a probe molecule is likely to bind on the protein as a function of imposed chemical potential on the system. ACPS accurately pinpoints water binding positions, as reported by Guarnieri[1] in the original development of the method. Further, the method effectively predicts fragment binding patterns in proteins, as described recently.[14−16] The only inputs to the algorithm are the protein structure, a chemical fragment, and atomic force field parameters (Amber).[17] The simulations require 10−20 h of computing time per fragment, depending on the size of the protein. Each fragment simulation runs on a single CPU core; thus, multiple fragments and protein structures are run in parallel so that this cluster analysis is relatively inexpensive. ACPSs were run on each of the eight HEWL structures, using the eight probe molecules shown in Figure 2, and water. The annealing schedule used in these studies specifies a parameter B, $B = \mu_{ex}/kT + \ln\langle N\rangle$, where $\langle N\rangle$ is the average number of fragments in the system and μ_{ex} is the excess chemical potential associated with protein interactions. The value of B at which occupancy drops below 50% is used to calculate μ_{ex}—the excess free energy per molecule—for ranking sites by affinity. The value of B is decremented from +100 to negative values until no fragments remain in the system.

The result of the simulation is that each structure is saturated with fragment molecules at a range of μ_{ex} values (Table S1, SI). For each fragment type, we selected the most favorable (negative) μ_{ex} that results in fragment occupancy at multiple sites on the protein. Next, we retain fragments in sites where clusters form; a cluster is defined as a set of fragment types that are located within 2.5 Å of each other—the maximum allowed distance from the center of mass of one fragment to another in a cluster. Subsequently, the μ_{ex} for waters is selected below the value where the phase transition occurs—the bulk volume of the simulation box is voided, and only surface-bound waters remain (tightly bound waters). Fragments are eliminated that have a heavy atom within 1 Å of the oxygen of a bound water molecule. The μ_{ex} of water is lowered until at least some clusters survive elimination. Finally, there will often be more than one cluster site remaining. To reduce these sites, additional fragment types are added to increase the chemical diversity (Table S2, SI). Clusters are recalculated and eliminated by water exclusion until only one site remains. This procedure is agnostic to the final outcome and is systematically applied to minimize the number of clusters. Sites with the highest chemical diversity and affinity, not excluded by tightly bound waters, are consistently identified by this method.

Except for isobutane in the 1XEK structure (Table S2, SI), which was not found within the 2.5 Å cutoff and was excluded by a nearby water, all fragments that were run in the simulations were found clustered in the hot spot, despite conformational and chemical diversity variations within the binding site of the various protein structures used in the simulations.

The results of the analysis of the fragment probe data from the simulations is presented in Figure 2. Note that there are significant differences in the geometry among the HEWL structural conformers used in these simulations. For example, the I58M substitution in the binding site (Figure 2a) has a deeper pocket than does the I98M substitution (Figure 2b). The D52S mutation (Figure 2d), which includes a bound ligand, also has a deeper pocket in the binding site relative to the D52S apo structure (Figure 2c). Interestingly, the most hydrated wild-type structure (Figure 2e) has a tryptophan residue at the top left position of the binding pocket that is similarly orientated to the structures shown in Figures 2a−d, but at a lower level of hydration. As the level of hydration decreases, a quite dramatic conformational change occurs, resulting in the stacking of two tryptophan residues (Figure 2f). At the lowest level of hydration (Figure 2g), the tryptophan residue W62 clearly seen in Figure 2a−f disappears, having rotated completely into the protein, and the binding site dramatically collapses, eliminating the binding of the branched isobutane moiety. In spite of all these structural changes to the binding pocket, the fragment clustering robustly identified the hot spot.

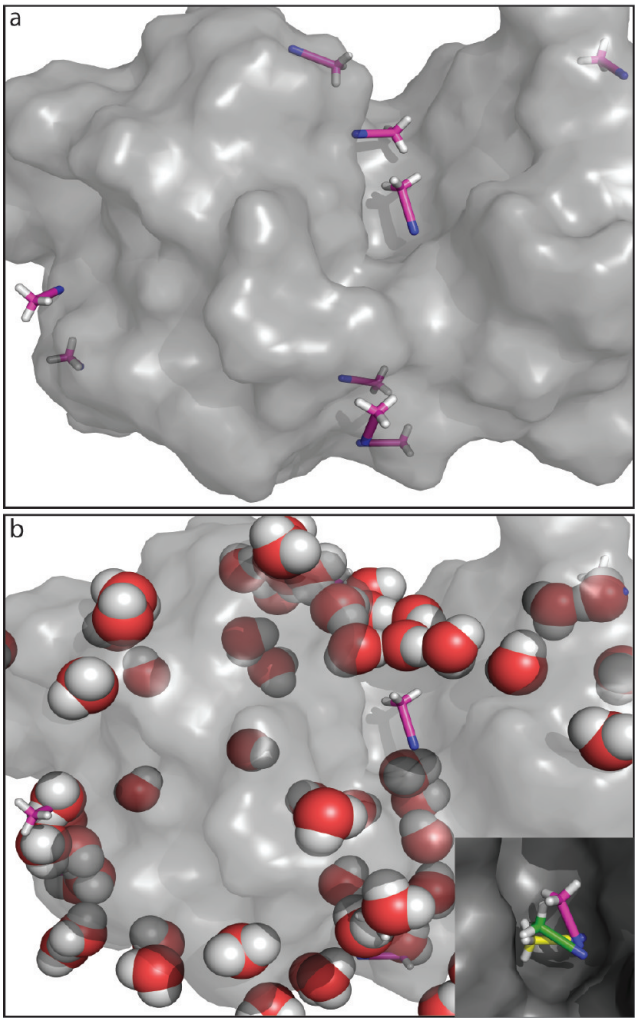

Figure 3. Water exclusion eliminated all other acetonitrile molecules except for the one in the hot spot of structure 2YLO. (a) Calculated positions of nine acetonitrile molecules binding at μ_{ex} of −22.53 kcal/mol. (b) Calculated positions of water molecules (μ_{ex} of −15.54 kcal/mol) overlapping all of the acetonitrile molecules except for the one in the binding site. (b) Inset displays acetonitrile molecule in the crystal structure (yellow carbons) and calculated acetonitrile molecules (magenta carbons, μ_{ex} −22.53 kcal/mol; green carbon, μ_{ex} −16.12 kcal/mol). Shaded molecules are on the back side of the protein.

Discriminating between false-positive clusters of small organic probes and true hotspots requires knowing whether water molecules are also tightly bound at the cluster site. If they are, then the cluster site is ruled out as a true hot spot. Figure 3a illustrates the calculated highest affinity binding sites of acetonitrile, and Figure 3b shows the calculated sites of tightly bound water superimposed on the acetonitrile map in the 2LYO structure.[9] The only calculated site for

acetonitrile which remains after excluding the sites that are also occupied by tightly bound waters is the site of acetonitrile binding found in the crystal structure (see inset).

Simulated annealing of chemical potential calculations were run on each static structure with eight small organic probes and water, for a total of 72 simulations. Consistent with our hypothesis, we found that, in all cases where different organic probes clustered at low free energies but where water was not present at higher free energies, the binding site was correctly located with no false positives. We have shown that a Monte Carlo technique using multiple fragment probes and a static protein produce the correct location of the hot spot independent of starting protein conformation or mutation state.

■ ASSOCIATED CONTENT

Supporting Information. Supplementary analysis, including details of simulated annealing of chemical potential, additional data, and aspects concerning fragment clustering and water exclusion. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
frankguarnieri@yahoo.com

Present Addresses
[⊥]BioLeap, Inc., 238 West Delaware Avenue, Pennington, New Jersey 08534, United States.

■ ACKNOWLEDGMENT

We thank Ian Cloudsdale and Rick Bryan for assistance and discussions. J.L.K. III acknowledges the Office of Naval Research for support.

■ REFERENCES

(1) Guarnieri, F.; Mezei, M. *J. Am. Chem. Soc.* **1996**, *118*, 8493.
(2) Lexa, K. W.; Carlson, H. A. *J. Am. Chem. Soc.* **2011**, *133*, 200.
(3) Mattos, C.; Ringe, D. *Nat. Biotechnol.* **1996**, *14*, 595.
(4) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. *Science* **1996**, *274*, 1531.
(5) Allen, K. N.; Bellamacina, C. R.; Ding, X. C.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. *J. Phys. Chem.* **1996**, *100*, 2605.
(6) Guvench, O.; MacKerell, A. D. *PLoS Comput. Biol.* **2009**, 5.
(7) Dennis, S.; Kortvelyesi, T.; Vajda, S. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4290.
(8) Jollès, P., Ed. In *Lysozymes: Model Enzymes in Biochemistry and Biology*; EXS, Vol. 75, Birkhäuser Verlag: Basel, Boston, 1996.
(9) Wang, Z.; Zhu, G.; Huang, Q.; Qian, M.; Shao, M.; Jia, Y.; Tang, Y. *Biochim. Biophys. Acta* **1998**, *1384*, 335.
(10) www.pdb.org; 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.
(11) Ohmura, T.; Ueda, T.; Hashimoto, Y.; Imoto, T. *Protein Eng.* **2001**, *14*, 421.
(12) Hadfield, A. T.; Harvey, D. J.; Archer, D. B.; Mackenzie, D. A.; Jeenes, D. J.; Radford, S. E.; Lowe, G.; Dobson, C. M.; Johnson, L. N. *J. Mol. Biol.* **1994**, *243*, 856.
(13) Nagendra, H. G.; Sukumar, N.; Vijayan, M. *Proteins* **1998**, *32*, 229.
(14) Clark, M.; Guarnieri, F.; Shkurko, I.; Wiseman, J. *J. Chem. Inf. Model* **2006**, *46*, 231.
(15) Moore, W. R. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 355.
(16) Konteatis, Z. D. *Expert Opin. Drug Discovery* **2010**, *5*, 1047.
(17) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179.