JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Robust Identification of Binding Hot Spots Using Continuum Electrostatics: Application to Hen Egg-White Lysozyme

David H. Hall,[†,∥] Laurie E. Grove,[‡,∥] Christine Yueh,[†,∥] Chi Ho Ngan,[†,∥] Dima Kozakov,[*,†] and Sandor Vajda[*,†,§]

Departments of [†]Biomedical Engineering and [§]Chemistry, Boston University, Boston, Massachusetts 02215, United States

[‡]Department of Sciences, Wentworth Institute of Technology, Boston, Massachusetts 02115, United States

ABSTRACT: Binding hot spots, protein regions with high binding affinity, can be identified by using X-ray crystallography or NMR spectroscopy to screen libraries of small organic molecules that tend to cluster at such hot spots. FTMap, a direct computational analogue of the experimental screening approaches, uses 16 different probe molecules for global sampling of the surface of a target protein on a dense grid and evaluates the energy of interaction using an empirical energy function that includes a continuum electrostatic term. Energy evaluation is based on the fast Fourier transform correlation approach, which allows for the sampling of billions of probe positions. The grid sampling is followed by off-grid minimization that uses a more detailed energy expression with a continuum electrostatics term. FTMap identifies the hot spots as consensus clusters formed by overlapping clusters of several probes. The hot spots are ranked on the basis of the number of probe clusters, which predicts their binding propensity. We applied FTMap to nine structures of hen egg-white lysozyme (HEWL), whose hot spots have been extensively studied by both experimental and computational methods. FTMap found the primary hot spot in site C of all nine structures, in spite of conformational differences. In addition, secondary hot spots in sites B and D that are known to be important for the binding of polysaccharide substrates were found. The predicted probe−protein interactions agree well with those seen in the complexes of HEWL with various ligands and also agree with an NMR-based study of HEWL in aqueous solutions of eight organic solvents. We argue that FTMap provides more complete information on the HEWL binding site than previous computational methods and yields fewer false-positive binding locations than the X-ray structures of HEWL from crystals soaked in organic solvents.

The analysis of ligand binding sites of proteins is often the starting point for function identification and drug discovery. The sites generally include smaller regions called hot spots that are major contributors to the binding free energy and hence are crucial to the binding of any ligand at that particular site.[1] In drug design applications, such hot spots can be identified by screening for the binding of fragment-sized organic molecules.[2] Since the binding of these small-molecule probes is very weak, it is usually detected by NMR spectroscopy[3,4] or X-ray crystallography.[2,5−8] Individual probe molecules can bind at a number of locations, but clusters of different probes occur only at hot spots.[2,4] Although the origin of this weakly specific binding is not fully understood, the phenomenon itself has been well-established. For example, using their structure−activity relationships by the NMR method, Fesik et al. observed that for a diverse set of targets, nearly 90% of fragment-sized ligands bind exclusively to protein sites that are known also to bind druglike small molecules.[4] Similar conclusions have been made using the multiple-solvent crystal structures (MSCS) method, which is based on determining the structure of a protein by X-ray crystallography in aqueous solutions of several organic solvents and superimposing the structures to identify clusters of overlapping probe molecules.[2,5−8]

The hot spots of the model protein hen egg-white lysozyme (HEWL) have been extensively studied by both experimental[3,9−13] and computational[14,15] methods. HEWL recognizes its substrate, polymeric carbohydrates from bacterial cell walls, in an active site that can accommodate up to six saccharide units, such as 2-*N*-acetylglycosamine (NAG) or *N*-acetyl-D-muramic acid, in subsites A, B, C, D, E, and F. The X-ray structures of HEWL with various polysaccharides show that the most important sites, in the order of occupancy, are C, B, D, and A,[16−19] with C being the highest-affinity site.[20] X-ray structures have also been determined with a number of small organic molecules, including ethanol,[9] bromoethanol,[10] dimethyl sulfoxide (DMSO),[11] urea,[12] and acetonitrile (CCN).[13] CCN binds only at site C,[13] but all of the other compounds are found at a number of locations. However, overlapping the structures shows a cluster only at site C, which is also the highest-occupancy site in all structures, whereas the other binding sites are frequently located at crystal contacts.[11] These types of false-positive sites were eliminated by Liepinsh and Otting, who used ^{1}H NMR spectroscopy to measure the nuclear Overhauser effect (NOE) between eight different small organic compounds and H atoms of HEWL.[3] They found all eight molecules in site C, interacting with residues Asn59, Trp63, Ile98, Ala107, and Trp108. This site is the same identified by Wang et al. as binding a single CCN molecule (PDB entry 2LYO).[13] The NMR-based mapping also found that residues Trp62, Val109, and Ala110 participate in the binding of probe molecules. These residues lie within the B and D sites, respectively, in agreement with the X-ray data, which show Trp62, Ile98 (also part of site C), Asp101, and Asn103 in site B and Asn46, Asp52, Val109, and Ala110 in site D.[16]

Two recent studies focused on the computational identification of the main hot spot in site C. Lexa and Carlson performed molecular dynamics (MD) simulations on HEWL in a 50% CCN/50% water mixture.[14] Since CCN density was found in site C only when allowing for protein flexibility, they concluded that full flexibility is essential for proper hot-spot mapping. The problems with this conclusion are that binding of a single compound is not sufficient to identify a hot spot and that Lexa and Carlson used the CCN-bound structure of HEWL, in which the large and generally open binding site is somewhat contracted around the small ligand. This narrowing of the binding site

Received: August 21, 2011
Published: November 17, 2011

Table 1. Rankings of Hot Spots Identified by FTMap and Their Respective Numbers of Probe Clusters

			secondary sites[a]			
PDB entry	structural notes	site C[a]	site B	site D	total probe cluster count	distance to CCN (Å)[b]
2LYO	holo, CCN bound	1 (34), 6 (4)	3 (13)	2 (16)	67	1.1
2LYM	apo	1 (22), 3 (13), 6 (6)	2 (15), 7 (5)	4 (12)	73	1.4
1IR8	Ile58 → Met58	1 (33), 7 (2)	3 (11), 5 (10)	2 (14)	67	1.6
1IR9	Ile98 → Met98	1 (25), 6 (6)	2 (18)	4 (11)	60	2.5
1LSY	Asp52 → Ser52	1 (34), 6 (5)	2 (13)	3 (12)	64	1.3
1LSZ	Asp52 → Ser52, holo	1 (32), 8 (3)	2 (11), 4 (10)	3 (11), 5 (6)	73	1.2
1XEI	17.6% dehydrated	1 (34)	4 (11)	2 (14)	59	1.1
1XEJ	16.9% dehydrated	1 (35), 6 (6)		4 (10)	51	0.7
1XEK	9.4% dehydrated	1 (18), 5 (7), 9 (4)		3 (12)	41	1.3

[a] The number shown in parentheses after each ranking number is the total probe cluster count for that particular CS. [b] The distance is that between the center carbons of the CCN molecules from the 2LYO structure and the lowest-energy CCN from mapping results.

may have also caused the failure to identify site C when the simulations were run assuming a rigid protein. To address these concerns, Guarnieri et al. performed grand canonical Monte Carlo simulations to map eight HEWL structures using eight different organic solvents as probes.[15] Individual simulations were run for each probe molecule without allowing for protein flexibility, and the bound probe positions were clustered to identify the hot spots as in the MSCS experiments. In this approach, the simulations were started in neat solvent and ended at very low concentrations, essentially in vacuum, without considering solvation effects. To eliminate false binding of probe molecules at water binding sites, simulations were also run using water as the probe molecule, and clusters of organic solvent molecules within 1 Å of a water cluster were discarded. In spite of this indirect approach, the simulations showed excellent results for each of the eight HEWL structures, with a single cluster of probe molecules found within site C, indicating that the main hot spot could be detected using a static protein.

In this work, we applied the protein mapping method FTMap[21] to the HEWL structures considered by Guarnieri et al.[15] as well as the HEWL apo structure. FTMap has three major advantages over the other two computational methods. First, it directly samples the potential probe binding sites on a dense grid around the target protein using empirical free energy functions that model the competition with water by adding a continuum electrostatic term. By accounting for solvation while sampling, the method avoids the need for a separate calculation of water binding positions. Second, the sampling is extremely efficient because of the use of fast Fourier transforms in energy evaluations, which allows for the sampling of millions of probe positions on the protein surface. Because of the resulting efficiency, we have made FTMap available as a public server, in contrast to the other two methods, which require lengthy runs and substantial computational resources. Third, since we can map a protein using many different probe molecules, FTMap reliably identifies all hot spots of a target protein while eliminating false positives.[21,22] In this manner, both primary and secondary hot spots are identified, which is crucial for accurate protein mapping since, as shown for HEWL, the secondary hot spots in sites B and D are known to be important for ligand binding.

The FTMap server (http://ftmap.bu.edu/) currently uses 16 small molecules as probes (ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and *N*,*N*-dimethylformamide). FTMap performs four steps as follows. (1) The rotational/translational space of each probe is systematically sampled on a grid around the fixed protein, consisting of 0.8 Å translations and 500 rotations at each location. The energy function includes a stepwise approximation of the van der Waals energy with attractive and repulsive contributions and an electrostatics/solvation term based on the Poisson−Boltzmann continuum model with dielectric constants (ε) of 4 and 80 for the protein and solvent, respectively. The 2000 best poses for each probe are retained for further processing. (2) The 2000 complexes are refined by off-grid energy minimization during which the protein atoms are held fixed while the atoms of the probe molecules are free to move. The energy function includes the bonded and van der Waals terms of the CHARMM potential[23] and an electrostatics/solvation term based on the analytic continuum electrostatic (ACE) model,[24] as implemented in CHARMM. (3) The minimized probe conformations are grouped into clusters using a simple greedy algorithm and a 4 Å root-mean-square deviation clustering radius. Clusters with less than 10 members are excluded from consideration. The retained clusters are ranked on the basis of their Boltzmann-averaged energies. The six clusters with the lowest average energies are retained for each probe. (4) To determine the hot spots, FTMap finds consensus sites (CSs), which are regions on the protein where clusters of different probes overlap.[21] Therefore, the probe clusters are clustered again using the distance between the cluster centers of mass as the distance measure and 4 Å as the clustering radius. The CSs are ranked on the basis of their numbers of clusters, with duplicate clusters of the same type also considered in the count. The largest CS defines the most important hot spot, with smaller CSs identifying secondary hot spots that generally also contribute to ligand binding. It was shown for a large variety of proteins that the CSs determined by this algorithm agree very well with the hot spots identified by X-ray crystallographic or NMR techniques.[21,22,25−27]

The nine different HEWL structures mapped are summarized in Table 1. In short, the structure derived from PDB entry 2LYO corresponds to the experiment of Wang et al.,[13] with the singly bound CCN molecule residing in site C; the corresponding apo structure is also listed (PDB entry 2LYM).[13] All of these structures except the apo structure were also studied by Guarnieri et al.[15] Prior to mapping, all water molecules and heteroatoms were removed from the structures.

For all nine structures, FTMap predicted that the highest-ranked CS, CS1, occupies site C (Table 1). In particular, FTMap successfully identified site C as the top-ranked CS for the unbound HEWL structure (Figure 1A). Furthermore, in all cases, CS1 was found to have a very high probe cluster count, which is a good indicator of druggability.[4,27] Besides correctly locating site C, FTMap was also able to predict additional hot spots in sites B and D (Figure 1B and Table 1), as identified by the NMR mapping studies.[3]

In all cases, FTMap correctly identified the main hot spot (site C) as CS1. However, inspection of Table 1 does reveal some differences in both the total probe cluster count and the secondary site

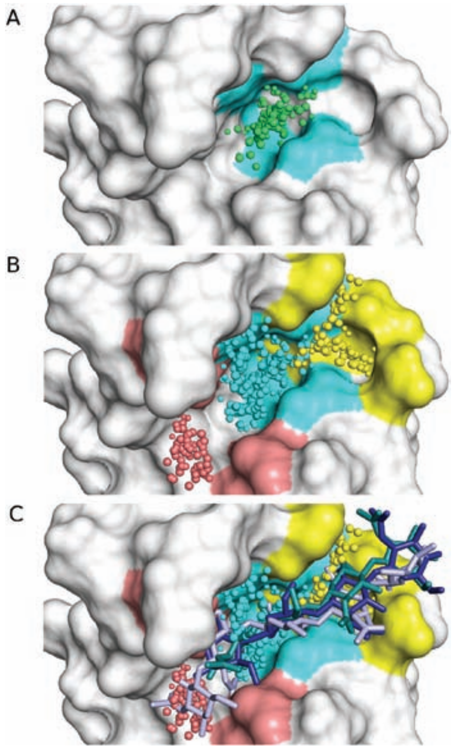

Figure 1. Centers of probe clusters found by mapping the HEWL apo structure (PDB entry 2LYM) using FTMap. (A) Probe clusters (green spheres) in CS1. The residues of site C are shown in cyan. (B) Probe clusters in all large CSs. The clusters (represented as spheres) are colored according to their locations as follows (see Table 1): CS1, CS3, and CS6 in Site C, cyan; CS2 and CS7 in site B, yellow; CS4 in site D, salmon. The patches on the protein surface indicate the three sites. (C) Three representative polysaccharide ligands (sticks) superimposed within the binding site, with probe clusters shown as spheres. The ligands come from PDB entries 1LZC, 1SF6, and 1SFB.

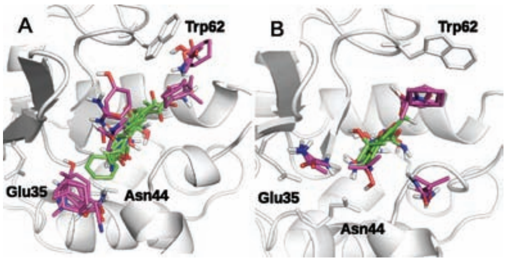

Figure 2. Effect of conformational changes on the mapping results. Centers of probe clusters (shown as sticks) for (A) the apo structure 2LYM and (B) the structure 1XEK with the smallest binding cavity. The clusters of CS1 are shown in green and those of the other CSs in magenta.

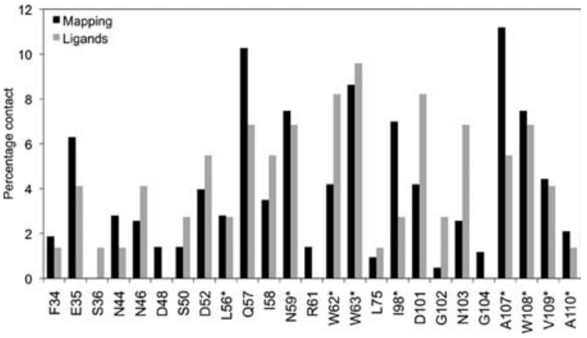

Figure 3. Distributions of nonbonded interactions between HEWL residues and the probes in the CSs shown in Table 1 from the mapping of the apo structure 2LYM (black) and between the HEWL residues and eight representative ligands from PDB files 1LZB, 1LZC, 1AT6, 1HEW, 1SF6, 1SF7, 1SFB, and 1SFG (gray). The residues detected via NOE ^{1}H NMR experiments are marked with an asterisk.[3]

assignments, particularly for structures 1XEI, 1XEJ, and 1XEK. This series of structures has a collapse of the binding site cavity from largest (1XEI) to smallest (1XEK), and that size difference is reflected by the trend in the total probe cluster count (Table 1), which also decreases from 59 to 41. Furthermore, no CSs were found for site B for structures 1XEJ and 1XEK. This result can be explained by the rotation of Trp62 by ~90° around C_β, thereby blocking off the top of the cavity (Figure 2). Despite this conformational change, the primary hot spot was still detected without accounting for flexibility in the computational mapping.

The CSs predicted by FTMap were also in agreement with known ligand binding sites. Figure 1C illustrates the overlap between the CSs and polysaccharide substrates (shown as sticks) from three X-ray structures. To quantify further the similarity between the mapping results and known ligands, the distribution of nonbonded interactions between the protein residues and probe molecules was calculated. Figure 3 shows the frequency of contacts between the probe molecules and each residue in the ligand binding site of HEWL. As shown, Ala107, Gln57, Trp63, Asn59, Trp108, and Ile98 participate in the highest number of nonbonded interactions with the probe molecules. These residues are the same as those identified using NMR spectroscopy, which are indicated by an asterisk in Figure 3.[3] For comparison, Figure 3 also shows interaction frequencies observed in eight HEWL structures cocrystallized with polysaccharide ligands ranging from $(NAG)_3$ to $(NAG)_6$. Three of these substrates are shown in Figure 1C. Figure 3 confirms that the residues with the highest percentage of nonbonded interactions are in good agreement with those predicted by the mapping. For both the mapping results and the ligands, the residues with the highest percentage of nonbonded interactions are in site C, with those in site B also exhibiting a high frequency of contacts. Indeed, the ligands interact with a number of residues outside of site C, all of which are identified by the CSs predicted by FTMap. Thus, FTMap identified residues in both the primary and secondary hot spots that are important for ligand binding, in contrast to the other two computational methods, which restricted consideration to site C.[14,15]

The two previous studies correctly placed CCN probes within the CCN binding site (site C).[14,15] FTMap also predicted that the top CS, CS1, contains at least one cluster of CCN molecules for each of the nine structures, with the mapping results from the 2LYO, 1IR8, 1LSY, 1XEI, and 1XEJ structures containing two CCN clusters. To quantify the position of CCN within each cluster in relation to that found in the bound structure (PDB entry 2LYO), the distance between the middle carbons of the lowest-energy CCN probe in the highest ranking FTMap-predicted cluster and the CCN molecule from the 2LYO structure was measured for each structure (Table 1). In seven of the nine cases, this distance is 1.5 Å or less, which indicates very good agreement considering that we mapped ligand-free HEWL structures. For the mutant 1IR8 and 1IR9 structures, the antiparallel β-sheet located near Ile58 somewhat protrudes into the binding cavity, thereby slightly altering the position of the CCN. In each case, the probe representative located closest to that from the 2LYO structure has the lowest energy. As shown in Figure 4, the CCN

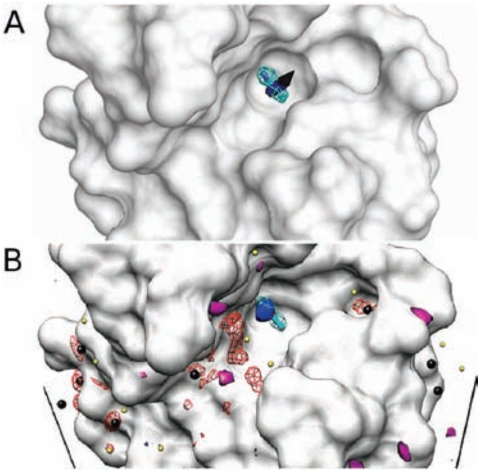

Figure 4. Mapping results for the 2LYM structure using FTMap (A) are shown in comparison to those obtained using MD simulations on the 2LYO structure (B).[14] The experimentally determined density of CCN is shown in cyan, and the density corresponding to the computational mapping results is shown as cobalt blue.

position predicted by FTMap for the apo structure of HEWL is nearly identical to that found in the MD simulations by Lexa and Carlson for the holo structure[14] as well as to the one in the X-ray structure (the CCN molecule and density from the X-ray crystallographic data are shown in cyan).[13] However, as already discussed, we emphasize that the bound position of a single probe compound is not necessarily a hot spot and that hot spot identification requires simulations with at least six to eight different probes.[2,21] Since MD requires substantial computational resources, it would have been very difficult to carry out such a large number of simulations. In addition, while both MD and FTMap accurately predicted the CCN binding site, to obtain results in agreement with the experimental data, the MD simulations had to allow for protein flexibility.[14] Without full flexibility, multiple CCN binding sites were located, with only weak occupancy in site C, suggesting that the simulations did not provide adequate sampling.

FTMap was able to identify the primary hot spot of HEWL as the top-ranked consensus site in nine different structures of HEWL. This hot spot coincides with site C of HEWL, which is known to be a key site for ligand binding.[13,3] Furthermore, each top-ranked CS included at least one cluster of CCN molecules in close proximity to the experimentally determined position of CCN in the 2LYO structure. The results confirm that when a diverse set of probes are used in computational mapping, the primary hot spot is identified and spurious minima are avoided. Although considering protein flexibility may become important for detailed characterization of the binding site,[27] the hot spots showed remarkable robustness to conformational changes and were consistently found in a variety of HEWL structures. An additional advantage of FTMap is that, in contrast to previous computational methods,[14,15] it also detected secondary hot spots in sites B and D that are known to be important for the binding of polysaccharide substrates. The predicted probe−protein interactions agree well with those seen in complexes of HEWL with various polysaccharides and also agree with the results of an NMR-based study of the protein in aqueous solutions of eight different organic solvents.[3] We thus conclude that FTMap provides more complete information on the HEWL binding site than the two recently published computational methods that focused only on site C.[14,15] As discussed, soaking of HEWL crystals in an organic solvent such as ethanol[9] or DMSO[11] shows binding at site C but also at a number of locations, primarily in crystal contacts, that are clearly false positives. Thus, it appears that the computational approach is even more reliable than the X-ray based one. FTMap has been implemented as a server, which is freely available at http://ftmap.bu.edu/.

AUTHOR INFORMATION

Corresponding Author

midas@bu.edu; vajda@bu.edu

Author Contributions

[‖]These authors contributed equally.

ACKNOWLEDGMENT

This investigation was supported by Grant GM064700 from the National Institute of General Medical Sciences.

REFERENCES

(1) Clackson, T.; Wells, J. *Science* **1995**, *267*, 383.
(2) Mattos, C.; Ringe, D. *Nat. Biotechnol.* **1996**, *14*, 595.
(3) Liepinsh, E.; Otting, G. *Nat. Biotechnol.* **1997**, *15*, 264.
(4) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. *J. Med. Chem.* **2005**, *48*, 2518.
(5) Allen, K. N.; Bellamacina, C. R.; Ding, X.; Jeffery, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. *J. Phys. Chem.* **1996**, *100*, 2605.
(6) English, A. C.; Done, S. H.; Caves, L. S.; Groom, C. R.; Hubbard, R. E. *Proteins: Struct., Funct., Bioinf.* **1999**, *37*, 628.
(7) English, A. C.; Groom, C. R.; Hubbard, R. E. *Protein Eng.* **2001**, *14*, 47.
(8) Mattos, C.; Bellamacina, C. R.; Peisach, E.; Pereira, A.; Vitkup, D.; Petsko, G. A.; Ringe, D. *J. Mol. Biol.* **2006**, *357*, 1471.
(9) Lehmann, M. S.; Mason, S. A.; McIntyre, G. J. *Biochemistry* **1985**, *24*, 5862.
(10) Yonath, A.; Podjarny, A.; Honig, B.; Traub, W.; Sielecki, A.; Herzberg, O.; Moult, J. *Biophys. Struct. Mech.* **1977**, *4*, 27.
(11) Lehmann, M. S.; Stansfield, R. F. *Biochemistry* **1989**, *28*, 7028.
(12) Pike, A. C.; Acharya, K. R. *Protein Sci.* **1994**, *3*, 706.
(13) Wang, Z.; Zhu, G.; Huang, Q.; Qian, M.; Shao, M.; Jia, Y.; Tang, Y. *Biochim. Biophys. Acta* **1998**, *1384*, 335.
(14) Lexa, K. W.; Carlson, H. A. *J. Am. Chem. Soc.* **2011**, *133*, 200.
(15) Kulp, J. L., III; Kulp, J. L., Jr.; Pompliano, D. L.; Guarnieri, F. *J. Am. Chem. Soc.* **2011**, *133*, 10740.
(16) Strynadka, N. C. J.; James, M. N. G. *J. Mol. Biol.* **1991**, *220*, 401.
(17) Von Dreele, R. B. *Acta Crystallogr.* **2005**, *D61*, 22.
(18) Ose, T.; Kuroki, K.; Matsushima, M.; Maenaka, K.; Kumagai, I. *J. Biochem.* **2009**, *146*, 651.
(19) Song, H.; Inaka, K.; Maenaka, K.; Matsushima, M. *J. Mol. Biol.* **1994**, *244*, 522.
(20) Lumb, K. J.; Cheetham, J. C.; Dobson, C. M. *J. Mol. Biol.* **1994**, *235*, 1072.
(21) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. *Bioinformatics* **2009**, *25*, 621.
(22) Landon, M.; Lieberman, R.; Hoang, Q.; Ju, S.; Caaveiro, J.; Orwig, S.; Kozakov, D.; Brenke, R.; Chuang, G.-Y.; Beglov, D.; Vajda, S.; Petsko, G.; Ringe, D. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 491.
(23) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
(24) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578.
(25) Landon, M. R.; Lancia, D. R.; Yu, J.; Thiel, S. C.; Vajda, S. *J. Med. Chem.* **2007**, *50*, 1231.
(26) Chuang, G.-Y.; Kozakov, D.; Brenke, R.; Beglov, D.; Guarnieri, F.; Vajda, S. *Biophys. J.* **2009**, *97*, 2846.
(27) Kozakov, D.; Hall, D. R.; Chuang, G.-Y.; Cencic, R.; Brenke, R.; Grove, L. E.; Beglov, D.; Pelletier, J.; Whitty, A.; Vajda, S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 13528.