Grand Canonical Monte Carlo Simulation of Ligand−Protein Binding

Matthew Clark,* Frank Guarnieri, Igor Shkurko, and Jeff Wiseman

Locus Pharmaceuticals Four Valley Square, 512 Township Line Road, Blue Bell, Pennsylvania 19422

Received June 27, 2005

A new application of the grand canonical thermodynamics ensemble to compute ligand−protein binding is described. The described method is sufficiently rapid that it is practical to compute ligand−protein binding free energies for a large number of poses over the entire protein surface, thus identifying multiple putative ligand binding sites. In addition, the method computes binding free energies for a large number of poses. The method is demonstrated by the simulation of two protein−ligand systems, thermolysin and T4 lysozyme, for which there is extensive thermodynamic and crystallographic data for the binding of small, rigid ligands. These low-molecular-weight ligands correspond to the molecular fragments used in computational fragment-based drug design. The simulations correctly identified the experimental binding poses and rank ordered the affinities of ligands in each of these systems.

INTRODUCTION

The accurate prediction of protein−ligand binding energy is a key goal for enabling efficient drug design and optimization. A large body of work has been devoted to computing ligand−protein binding energies in the context of computational drug design, with the result that binding enthalpies are readily predicted by molecular mechanics computations. The multiple-copy-simultaneous-search (MCSS) method, for example, can efficiently search for binding sites across a protein surface and locate minimum enthalpy poses.[1−3] To compute the potency of ligand−protein binding, however, it is necessary to know the free energy of binding, and experimental ligand−protein binding thermodynamics routinely show that enthalpies are uncorrelated with free energies, for example, for ligands of T4 lysozyme[4,5] and cyctochrome *c* peroxidase.[6] Thus, computed enthalpies cannot be expected to reliably predict relative ligand binding affinities, and the prediction of potency remains a key goal in computational drug design.

Force-field-based approaches have demonstrated the ability to predict binding free energies within 1 or 2 kcal/mol using standard force fields coupled with free energy techniques.[7] The main approaches for computing free energies of ligand−protein binding have been through thermodynamic integration and free energy perturbation via molecular dynamics and Monte Carlo simulations. These topics have been extensively discussed in reviews and recent research.[8−15] Such free energy calculations typically study a single molecule or a series of molecules in a predesignated binding locale and pose. While these methods have generated high quality results, they are too time-consuming to be practical for locating all possible binding sites or quickly exploring a variety of poses in a large number of sites.

The second basic approach predicts free energies using ensemble average data from Monte Carlo simulations in parametrized equations. The "Linear Interaction Energy" method of Åqvist[16] has been used by several groups to produce very good results for large numbers of HIV inhibitors[17] and thrombin.[18] However, these methods require calibration using a set of known ligand−protein binding energies, as well as prior knowledge of the binding site.

In this report, we describe an alternative approach that computes binding free energies for a large number of poses over the protein surface by a method that is both thermodynamically rigorous and sufficiently rapid for practical application to drug design. It does not require a calibration set or the preidentification of a binding site. The method utilizes Monte Carlo sampling to construct grand canonical ensembles based on ligand−protein binding free energies[19] and is a derivation of a method used by Guarnieri and Mezei to explore water binding to DNA.[20,21] These workers demonstrated the ability to accurately compute the positions and relative energies of water bound to DNA, showing that the water is more tightly bound to the minor groove than the major groove. The current report describes the basic methodology for the computation of gas-phase binding affinities and demonstrates the ability of this approach to identify high-affinity poses for small, rigid ligands on the basis of a search of the entire protein surface. In practice, these small ligands represent molecular fragments that would subsequently be used for assembling full molecules in a ligand design process.

Two model systems were chosen to demonstrate the method because there exists extensive crystallographic and thermodynamic data for small, rigid ligands. The first system is the binding of hydrophobic ligands in an artificial cavity of T4 lysozyme,[4,5] which is used to illustrate the identification of binding sites and the rank ordering of ligand affinities. Because the shape of the T4 lysozyme cavity is strongly influenced by the ligand, the fit of the ligand in the cavity is very tight, and this test case represents a particularly stringent test of Monte Carlo methods for locating binding sites.

The second test system is taken from multiple-solvent-crystal-structure (MSCS) studies that map the binding of hydrophilic ligands to the surface of thermolysin.[22,23] This experiment parallels the computational system closely be-

* Corresponding author e-mail: mclark@locuspharma.com.

10.1021/ci050268f CCC: $33.50
Published on Web 11/24/2005

cause both methods, in effect, titrate the concentration of the ligand to identify multiple binding pockets. Because the binding pocket in this case is hydrophilic, the system also provides the opportunity to demonstrate the correction of simulated ligand binding affinities for tightly bound water.

METHODS

Grand Canonical Simulation. The basis of the method is to create a series of grand canonical ensembles of ligands interacting with a protein in a large simulation box using Monte Carlo sampling. The ligand poses are sampled throughout the box and over the entire protein surface. The free energy used to generate the ensemble is annealed, as contrasted with an annealing of temperature, in a series of steps. The output of the calculation is an ensemble of ligand poses at each free energy level in the annealing schedule. Annealing is generally run at descending free energy levels, with each new level starting from the last ensemble generated from the previous one. The ligands and protein are treated as rigid models, and conformationally flexible ligands are analyzed in multiple discrete conformations. Several hundred different ligands are subjected to this process, each running on a node of a large cluster computer.

The derivation of the thermodynamics of the grand canonical ensemble is available in several textbooks.[24] While the approach has been widely used in surface science,[25−27] its use in chemistry has been limited.[28] In the grand canonical ensemble, the system energy is commonly specified as the excess chemical potential relative to the chemical potential of a reference state, $\mu - \mu_{ideal}$. Instead of directly using the excess chemical potential to set the system energy, we use the unitless property B defined by the excess chemical potential and the number of ligands in the system, N, as shown in eq 1.[29]

$$B = (\mu - \mu_{id})/kT + \ln(\langle N \rangle) \quad (1)$$

This definition of B is convenient because B is then related to the concentration of molecules in the system, as shown below. In our simulations, it is specifically B that is annealed rather than the chemical potential. One of the features of the grand canonical ensemble is that the number of molecules in the system is varied by attempts to insert and delete molecules until the system density equilibrates to the selected chemical potential. The probability of accepting an insertion of a molecule into the system, and thus transitioning from state i to j, in terms of B is given in eq 2.

$$\frac{p_{ij}}{p_{ji}} = \frac{V \exp[B - (E_j - E_i)/kT]}{N_j} \quad (2)$$

Equation 2 can be factored to emphasize the relation of B to the system concentration, providing eq 3.

$$\frac{p_{ij}}{p_{ji}} = \frac{V}{N_j} \exp(B) \exp(-\Delta E/kT) \quad (3)$$

For an ideal gas, ΔE is always zero. At $B = 0$, eq 3 will, therefore, result in a concentration of one molecule per system volume V. The reference state for the free energy at $B = 0$ is, therefore, set by multiplying V/N in eq 3 by a reference concentration to make the unitless value B represent the energy with respect to this reference state.

In the grand canonical ensemble, ligands can translate and rotate as well as enter or leave the system. These moves are attempted randomly, and the probability of accepting them is given by eqs 4−6. In these equations, the term ΔE refers to the change in potential energy for the move being attempted, while B and T represent the dimensionless chemical potential and temperature of the system. Equation 4 gives the probability of accepting the insertion of a new particle into the system. In this case, if the energy change is negative, creating a favorable interaction, the value of the exponent is positive and the insertion is likely to be accepted. Equation 5 gives the probability of a chosen particle leaving the system. If the ligand being removed has a favorable interaction, ΔE is positive, and the exponent is negative, making the removal unlikely. For insertion, the number of particles changes from N to $N + 1$, and for deletion, the particles change from $N + 1$ to N. Equation 6 is the probability of accepting the translation or rotation of a molecule without changing the number N. Conformational changes can be included in the simulation using the criteria of eq 6; the amount of sampling required to compute converged free energies, however, increases exponentially with the degrees of freedom, and conformational sampling is not addressed in this study.

$$\alpha_{insert} = \min[1, \exp(-\Delta E/kT + B)\, V/(N + 1)] \quad (4)$$

$$\alpha_{delete} = \min[1, \exp(-\Delta E/kT - B)\, N/V] \quad (5)$$

$$\alpha_{move} = \min[1, \exp(-\Delta E/kT)] \quad (6)$$

A new configuration is generated using one or more of these methods; a change in energy ΔE from the previous to the new state is computed; then, a random number is generated with a uniform distribution between 0 and 1, and finally, if the probability computed above is greater than the random number, the new configuration is accepted.

The key outcomes of a grand canonical simulation are the states of the system, collected as snapshots characterized by N; the total number of ligands in the system; and E, the average energy per ligand molecule. After equilibration, the resulting ensemble of ligands corresponds to the free energy represented by the given B value.

At high energy levels, the system fills with ligands to form a condensed phase about the protein. This is a result of the attractive interactions among the ligand molecules. The ligand−ligand interactions limit the number of ligands in the system, which would otherwise increase exponentially, but also result in artifacts due to the stabilization of ligand clusters at intermediate energy levels. To increase the efficiency of inserting molecules into the condensed phase, the cavity bias algorithm of Mezei was implemented.[30] This algorithm biases attempts at inserting molecules to favor empty cavities and disfavor attempts in areas filled by molecules or proteins; the inverse of the bias is applied to the acceptance criteria to maintain detailed balance. The efficiency of moves is increased with a force-bias algorithm, which biases moves to the direction of the force gradient.[31,32] These improvements in sampling have demonstrated the

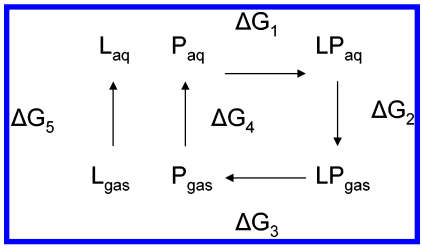

Figure 1. Thermodynamic cycle for ligand−protein binding. ΔG_1 is the experimentally measured value.

ability to accurately sample Lennard-Jones fluids near their triple point and have also proven effective with water.[32,33]

The grand canonical sampling, data collection, and analysis software was developed internally. The intermolecular potentials used for organic compounds were the nonbonded terms of the AMBER[34] molecular mechanics force field for protein residues, with CHELPG-computed[35] charges at the 6-31G* level for ligands. Water was represented by the TIP3P model.[36] All internal coordinates were fixed during the simulations. The nonbonded potentials, Lennard-Jones and electrostatic terms, were applied with a residue-based nonbonded cutoff of 20 Å for protein−ligand and ligand−ligand interactions. Periodic boundary conditions using the minimum image convention were implemented to avoid boundary effects at the edge of the box.

Figure 1 illustrates the thermodynamic cycle for the free energy of ligand−protein binding [LP]. ΔG_1 is the binding of the ligand to the protein in the solution phase. The free energy of binding computed by the grand canonical Monte Carlo algorithm described herein is ΔG_3, the free energy of binding in a vacuum. ΔG_4 and ΔG_2 are the solvation energies of the protein itself and the protein−ligand complex, respectively. These are assumed to be approximately equal in the current treatment, and the total contribution is, thus, zero. For a series of relatively similar molecules, this contribution may not be zero but relatively constant, allowing the computation of the relative binding of the series. ΔG_5 is the solvation energy of the ligand. This value can be significant, ranging from −20 to 3 kcal/mol across 500 common small organic molecules. In this study, ΔG_5 was computed using the GB/SA solvation model and was subtracted from the binding energy for the ligand to produce the solvation-corrected binding energy.[37]

LIGAND−PROTEIN STUDIES

T4-Lysozyme. A series of cocrystal structures and thermodynamic binding data are available for T4 lysozyme.[4,5] Nine cocrystal structures of T4 lysozyme with small ligands and the apo form, PDB codes 181L, 182L, 183L, 184L, 185L, 186L, 187L, 188L, 189L, and 1NHB, were prepared by removing the ligand structure, adding hydrogen atoms, and reorienting it to fit in the smallest rectangular box aligned along the Cartesian axes. The ligands listed in Table 1 were prepared by energy minimization with MacroModel,[38] computing charges by the CHELPG method using Gaussian 98[39] and then computing the solvation free energy using the GB/SA method implemented in MacroModel. Since the simulation process samples a rigid ligand against a rigid protein, conformationally flexible ligands were prepared in several conformations. Iso- and *n*-butyl benzene were prepared in both the crystallographic and minimum energy conformations. Propyl benzene was prepared in two minimum energy conformations in addition to the crystallographic conformation for *n*-butyl benzene with the terminal carbon removed, shown in Figure 2. Ethyl-substituted toluenes were modeled in two conformations shown in Figure 3. Single conformations were used for the remaining ligands.

The simulations were carried out in an orthorhombic box of 50.4 × 70.8 × 54.4 Å dimensions, providing a gap of 7.5 Å from the protein to the nearest side of the box.

The grand canonical simulations for T4 lysozyme were carried out at 298 K with 4 million steps of convergence at each *B* level, followed by 1 million steps for data collection. During the data collection, snapshots of the system configuration were taken every 10 000 steps for a total of 100 snapshots. The simulation was started at $B = 15$ and was reduced stepwise by 1 *B* until no ligands were left in the system. The starting value is chosen to produce a condensed phase to tightly pack the ligand about the protein. Simulated annealing was carried out 10 times for each ligand to ensure sampling of relevant poses in the tight pocket.

Thermolysin. The MSCS approach provides experimental information closely approximating simulated annealing in the grand canonical method, and the experimental results reported using this method can be used to evaluate the simulation results.[40] In this method, small organic molecules are soaked into crystal structures to evaluate their binding

Table 1. Experimental and Computed Binding Free Energies for T4-Lysozyme Ligands Computed for the 186L Crystal Structure

ligand	PDB	Δ*G* kcal/mol (exp)	*B* (solvation corrected)	runs locating binding site	RMS deviation (Å)
benzene	181L	−5.19	−9.81	10	0.69
benzofuran	182L	−5.46	−18.1	9	0.69
ethyl benzene	1NHB	−5.76	−20	6	2.96
indene	183L	−5.13	−14.8	7	0.48
indole	185L	−4.89	−12.8	10	1.86
isobutyl benzene	184L	−6.51	*a*	0	*a*
m-ethyl toluene		−5.12	−15.9	3	*b*
m-xylene		−4.75	−15.6	8	*b*
N-butyl benzene	186L	−6.70	*a*	0	*a*
o-ethyl toluene		−4.56	−12.6	1	*b*
o-xylene	188L	−4.60	−16.3	9	1.65
p-ethyl toluene		−5.42	*a*	0	*b*
p-xylene	187L	−4.67	−10.6	6	1.07
propyl benzene		−6.55	−22.3	6	*b*

[a] No poses observed in simulation; see text. [b] No crystal structure available.

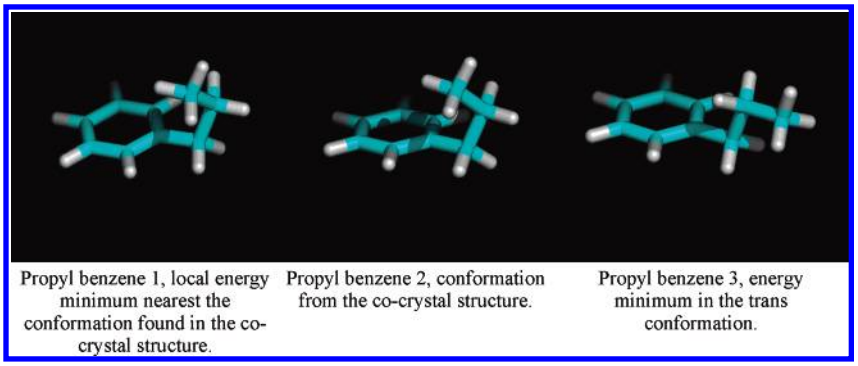

Figure 2. Conformers of *n*-propyl benzene.

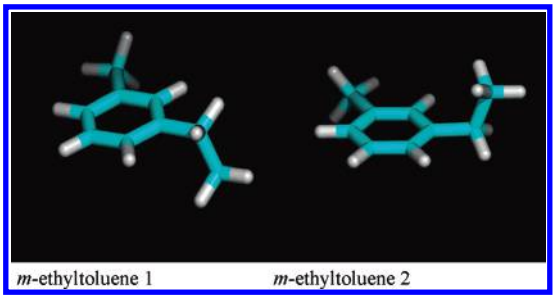

Figure 3. Conformers of *m*-ethyl toluene.

locations. The observed binding locations are presumed to be low-energy binding sites for the moiety under study. The evaluation of the overlap of these low-energy ligands has been used to identify binding sites.[3] The crystal structures resulting from the MSCS process carried out using a variety of concentrations of 2-propanol and acetone binding to thermolysin have been reported by English et al.[22,23] The thermolysin structure 1TLX was prepared and simulated in a box measuring 87.6 × 70.4 × 71.2 Å. Two conformations of the side chain of Tyr157 were present, and structures representing both were prepared. Acetone and 2-propanol were prepared as described above. 2-Propanol was prepared with three rotamers of the hydroxyl group separated by 60°. A single rotamer was used for acetone. The thermolysin protein was simulated at 298 K with 4 M steps of convergence for each *B* level. During the data collection phase, 2 million steps were performed and snapshots of the system configuration were taken every 20 000 steps for a total of 100 snapshots. The descending *B* schedule started at $B = 2$ and was lowered in steps of 1.0 *B* until the average population was less than 0.1.

RESULTS AND DISCUSSION

Simulated Annealing. One of the novel aspects of the grand canonical method is the use of the condensed phase to ensure that the entire protein surface is sampled by at least one pose of the ligand during the annealing process. The protein is immersed in a bath of ligand molecules, which is forced into the protein cavities under pressure. Annealing the *B* parameter to lower values lowers the free energy of the system, allowing weakly binding fragments to evaporate from the system until, ultimately, the tightest interaction is broken and no ligands remain in the system.

The effect of annealing *B* on the number of molecules in the simulation box is shown in Figure 4 for benzene surrounding the T4 lysozyme. At $B = 3$, a transition takes place, and the benzene molecules not interacting with the

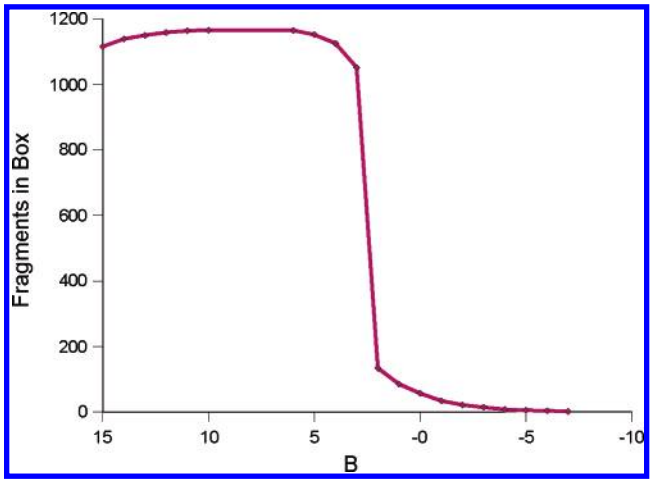

Figure 4. Number of ligands in the simulation as a function of *B* for benzene in T4 lysozyme. A condensed phase forms at low *B* values and undergoes a sharp transition as *B* is increased.

protein leave the system. As *B* is lowered further, the number of ligands in the system drops exponentially. When the number of ligands, averaged over the previous 10 000 convergence steps, is less than 0.1 ligands, the simulation is terminated.

The effect of annealing *B* on the distribution of ligand binding poses is shown in Figure 5 for propyl benzene interacting with T4 lysozyme. In Figure 5a, propyl benzene at −6.3 *B* exists as a condensed phase in the simulation box. At the phase transition point, shown in panel b, the bulk disappears and only ligands binding to the protein remain. Panels c−e show the "titration" of ligands against the protein until −22.3 *B*, where the only remaining ligands are those in the crystallographically observed binding site. This result demonstrates the relation of the titration curve exemplified by Figure 4 to the potency and specificity of the ligand−protein interactions, exemplified by Figure 5.

Since we are annealing *B*, which is related to the concentration of ligands in the system, and computing the occupancy of interaction sites on the protein surface, the simulation results can be treated as if they were a titration curve in solution to determine the free energy of interaction. Measuring the *B* value at which the pose transitions from full occupancy to nearly zero is analogous to measuring the binding constant for that pose from a titration curve. For example, for the pose in the T4 binding site, in Figure 5f, which shows the number of snapshots containing an example of the binding pose at each *B* level, the ligands are first sampled in the binding site at −6.3 *B*. Because of the tight nature of the binding pocket, the random insertions did not

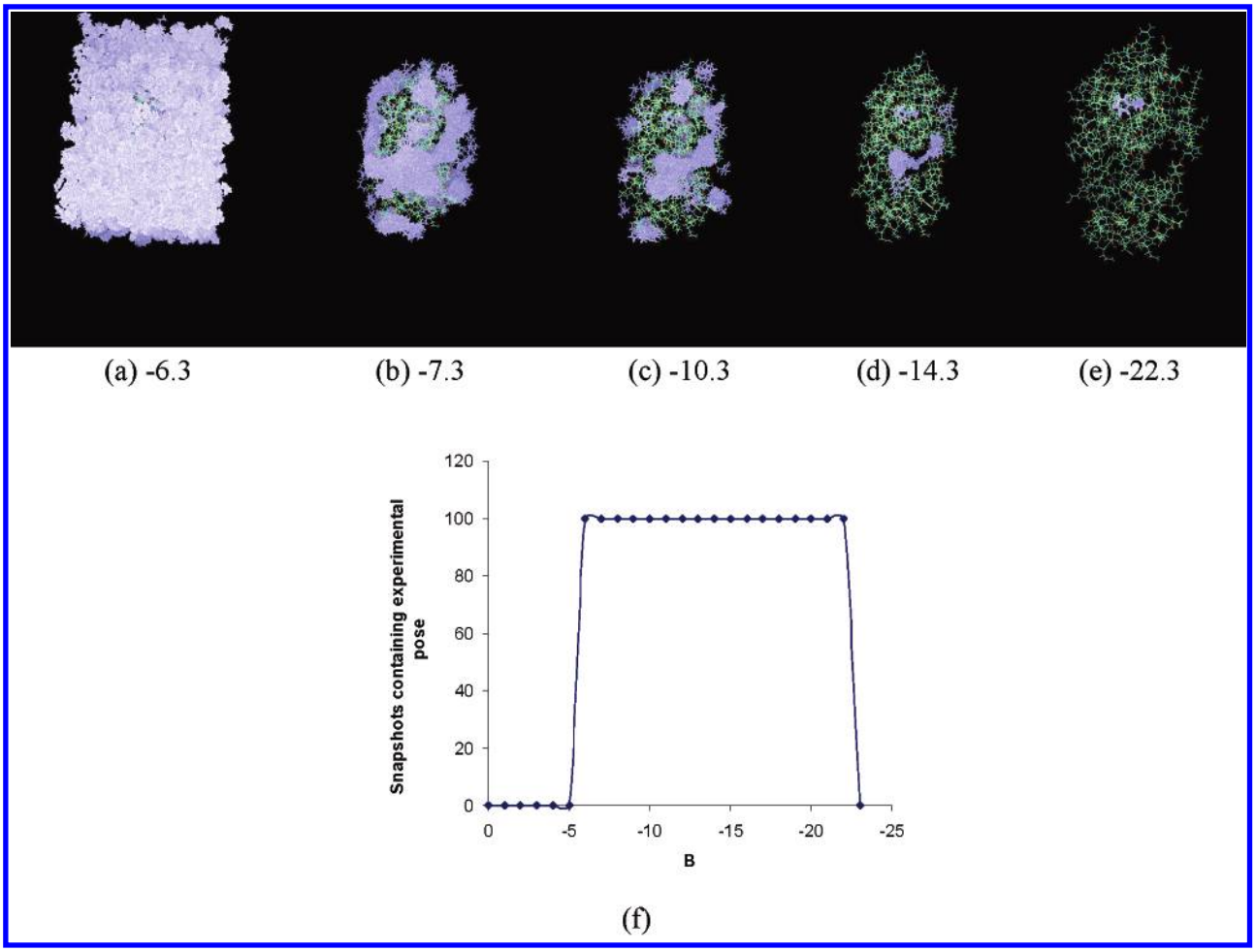

Figure 5. Population of the simulation box at four energy levels (*B*), with the population of propyl benzene in the binding site at each level. These snapshots illustrate the transition from the condensed phase to protein−bound ligand molecules.

sample the binding site for several *B* levels. Because the binding pose represents a strong interaction, when it is successfully inserted, it persists in each of 100 subsequent snapshots and then abruptly disappears at −23.3 *B*. The value of −22.3 *B*, therefore, is assigned for the free energy of binding in this pocket. In-house software analyzes the simulation snapshots to track the occupancy of each pose and assign the appropriate binding free energy.

T4 Lysozyme Binding Energies. The system comprising T4 lysozyme and the associated ligands reported by Morton and Matthews was chosen to demonstrate the computation of ligand−protein binding free energies because it is one of the few systems that provide a well-defined binding location, cocrystal structures, and experimental binding free energies for ligands of an appropriately small size.[4,5] Because the dynamic range of binding is only 2.1 kcal/mol, this data is illustrative but does not by itself fully validate the ability to predict binding energies. Methods for assembling small molecular fragments into full-sized ligands in order to access more extensive validation data are outside the scope of this paper. Because the binding pocket is small, however, this system provides a stringent test of the Monte Carlo method for locating binding poses. T4 lysozyme also has the advantage that the ligands are small and rigid, thus reducing issues associated with conformational flexibility and an entropy loss of rotatable bonds upon protein binding. Simulations were carried out to compute the binding free energy of each ligand to the protein structure with the largest cavity, 186L, in addition to its native protein structure. It was found that the 186L protein structure is the best single structure to successfully sample the largest number of ligands in the observed binding site during the Monte Carlo process. The binding site was not preidentified before the simulation was run; the free energy calculation was carried out over the entire protein.

Table 1 summarizes the ligands studied, their measured binding free energies, and the *B* values computed for binding. The *B* values are corrected for the solvation energies of the ligands using the GB/SA solvation energy as discussed in the Methods section. The simulated annealing was performed 10 times to assess sampling efficiency in the tight binding pocket. Table 1 shows the number of runs out of 10 in which the ligand was found in the binding site and the average root-mean-square (RMS) fit of the lowest energy poses to the observed pose. Figure 6 plots the predicted binding in *B* units versus the observed free energy of binding for the 11 ligands that were sampled successfully. These values were the lowest *B* values observed from the 10 annealing runs for each ligand. The agreement between the computed *B* and observed free energy of binding provides an r^2 value of 0.57, and a standard error of 0.4 kcal/mol. For this system, we conclude that the grand canonical Monte Carlo method both locates ligands in the observed binding pocket and rank orders their binding free energies with reasonable accuracy.

The binding free energy of benzene to T4 was studied in detail by Hermans and Wang using slow-growth thermody-

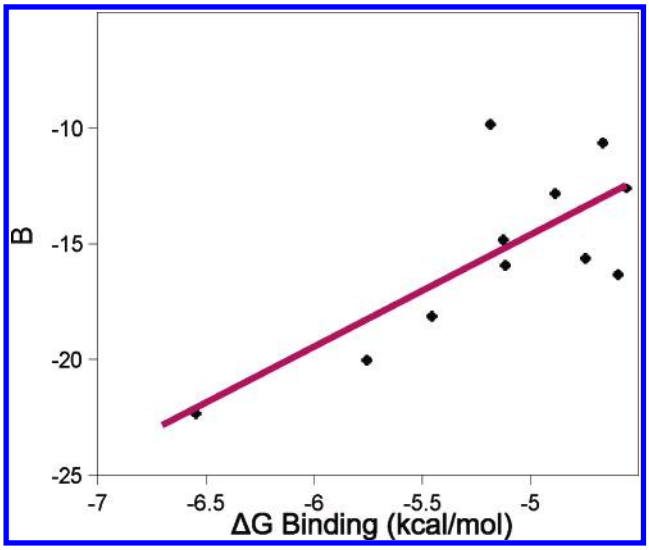

Figure 6. Observed binding of T4 lysozyme ligands vs computed *B* values. The r^2 is 0.57, with a standard error of 0.4 kcal/mol.

namic integration.[41] While that work did not compute free energies of other T4 ligands, it reported a binding free energy of −7 to −9 kcal/mol, which is comparable to our value of −9.6 kcal/mol.[42] In addition, the reported mean binding energy of −15.2 kcal/mol compares well to the energy of −16.0 kcal/mol found in this work. This energy difference, which accounts for most of the observed difference in the free energies, may be due to the use of different force fields and differences in the computation of charges between the two methods. The computed $T\Delta S$ values are very close between the two methods, with Hermans and Wang reporting 6.7 kcal/mol as compared to the grand canonical computation of 6.4 kcal/mol. The T4−benzene system was also studied using the double-decoupling method.[43] In that study, $T\Delta S$ values of 7.3−7.9 kcal/mol were computed, but the binding free energy was not reported. The results from the grand canonical simulation are in general agreement with these two other free energy computation methods, even though those methods allow the protein motion during the dynamics simulations and in this work the protein structure was held fixed.

Sampling Efficiency. Of the 14 ligands studied, all but three were found in the binding site during at least 1 of the 10 Monte Carlo sampling runs. The ligands *n*-butyl benzene, *p*-ethyl toluene, and isobutyl benzene were not found. This represents a "false negative" rate of 21%; that is, these ligands would be predicted to not bind to the protein as a result of their absence from the simulation. This result can be explained by the fact that Monte Carlo simulation is inherently limited in finding low-entropy poses, that is, ligands in very tight pockets, and may not sample the exact six-dimensional orientation and translation required to insert the molecule into the very confined protein binding site. This "limiting entropy" for sampling arises because the chance of sampling the correct rotation/translation coordinates in a relatively large simulation box can be very small. Even for propyl benzene, illustrated in Figure 5f, no poses are inserted into the binding site in the first six *B* levels. However, when it is inserted, it is accepted and is unlikely to be removed until a lower *B* level.

The ability to sample an observed pose is dependent on the probability of inserting a molecule into that pose, or inserting it nearby and moving the molecule to the observed pose by a series of translations and rotations. In the T4 lysozyme system, the tightness of the pocket prevents inserting nearby, and the correct orientation and position must be attempted by random selection to fit into the tight T4 pocket. The probability of sampling *n*-butyl benzene in the correct pose was measured by attempting to insert at the crystallographic Cartesian position, but with varying rotation angles for the insertion. Out of 5.4×10^6 attempts at random orientations, only 189 were accepted; the probability of sampling the available rotational space is, thus, 3.5×10^{-5}. Since the volume of the binding site in Cartesian space is about 2.9×10^{-7} of the volume of the entire simulation box, the overall chance of attempting an insertion at the correct location and rotation is about 1×10^{-11}. The cavity bias used along with the simulation of the condensed phase increases the likelihood of inserting into vacant sites by an order of magnitude to about 1×10^{-9}, and indeed, a single pose of *n*-butyl benzene was sampled in the binding site after an additional 4×10^9 sampling steps, which is qualitatively consistent with the above statistical predictions. The binding of *n*-butyl benzene to T4 lysozyme represents a worst case where there is essentially no freedom of movement for the bound ligand.

Morton and Matthews suggest that the two factors that correlate to the observed binding are the packing of the proteins in the crystal and the energy of reorganization of the protein in response to the ligand. Our results are consistent with their observation that the ligand−protein interaction energy is more important for binding than the energy of rearranging the protein structure. Thus, in this study, binding affinities were computed for a fixed protein—the cognate structure for the largest ligand—and predicted affinities correlated well with observed affinities. Morton and Matthews inferred that the shifts of 1.5−2.5 Å observed for helix F in the cocrystal structures of the various ligands represent a low cost in free energy. However, the effect of protein structural change may be responsible for the larger error for benzene, the smallest ligand. The largest F-helix move is between the benzene and *n*-butyl benzene structures. The 186L protein structure used in this study may not provide the close van der Waals contacts of the native structure, resulting in an underprediction of the binding energy of benzene.

Ligand Poses. The lowest-energy ensembles of poses obtained by the simulation closely match the crystallographically observed positions. Figure 7 shows the computed poses within the binding site superimposed on the corresponding cocrystal structures. The figure emphasizes the fact that the grand canonical method simulates the full ensemble of poses available at the simulation temperature, while crystallography provides only a single, averaged position. Therefore, the average RMS fits for the computed versus observed poses, which are provided in Table 1, represent the RMS deviation from the crystallographic position versus the full ensemble of predicted poses. Benzene, Figure 7a, is the smallest ligand and has an RMS fit of 0.69 Å, for example. The figure suggests that there is freedom to move about in the pocket and that the aromatic ring may freely rotate. This is advantageous in a drug design context because it provides information about the range of poses available at the simulation temperature, as compared to the minimum pose

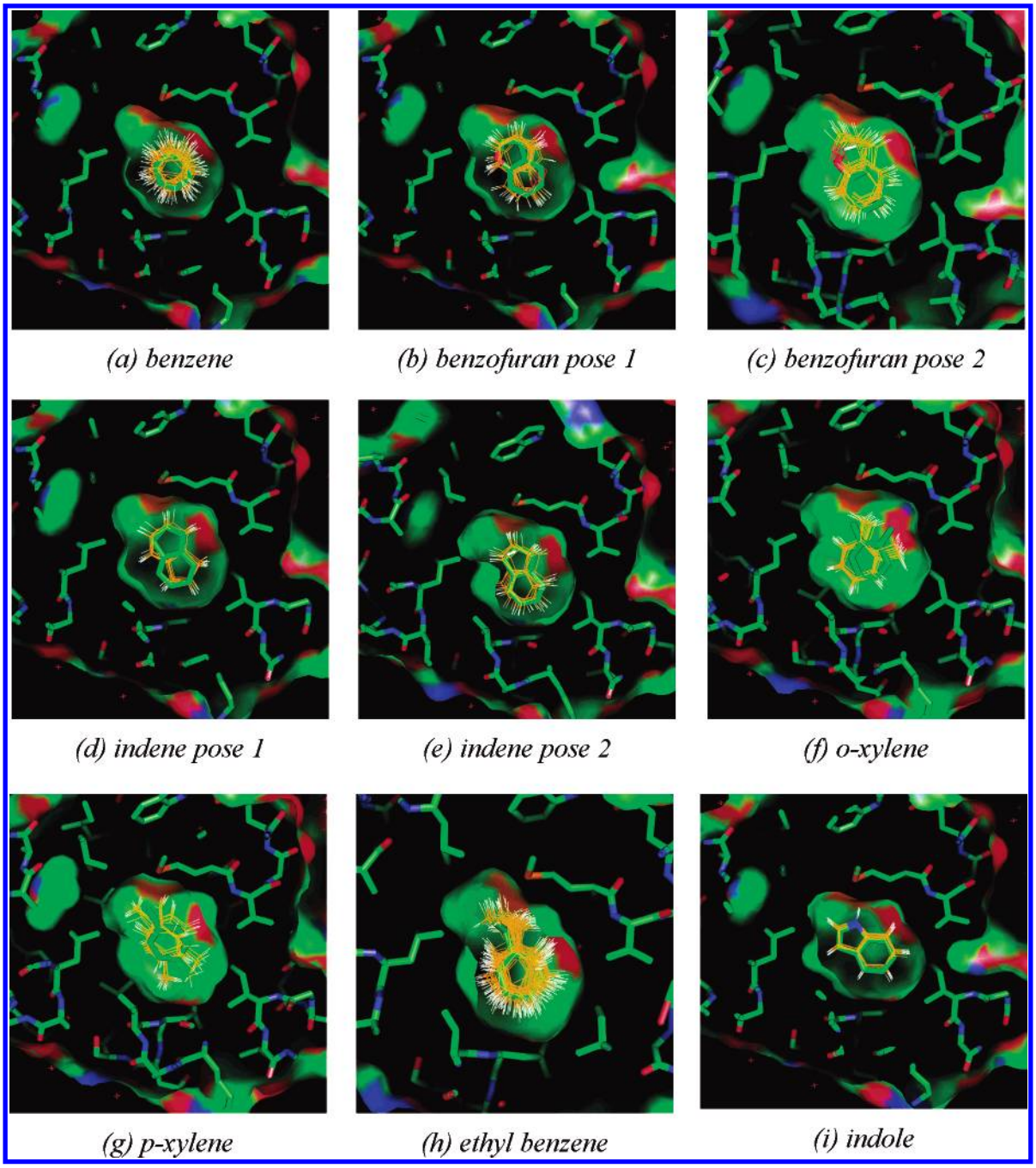

Figure 7. Ensembles of the ligands of T4 lysozyme overlaid on their cognate crystal structures.

computed from potential energy. A simple metric such as the RMS deviation between predicted and crystallographic poses may not be appropriate to evaluate the computed poses since the crystallographic experiment cannot evaluate the nature of the energy surface nearby the observed pose and thereby suggest the possible variance of the position.

In the crystallographic work, the authors noted that the exact orientations of benzofuran, indene, and indole are somewhat ambiguous.[4,5] Indole, Figure 7i, is found in the position selected in the original study. The pose of indene with the lowest computed energy (Figure 7d), however, has the positions of the five- and six-membered rings reversed from that assigned to the crystal structure so that this alternative pose may be preferred, particularly since this computed free energy correlates well with the experimental binding affinity. In our simulations, the pose originally assigned by crystallography is found at an energy level 5 *B* higher in energy than the minimum energy pose (Figure 7e).

Benzofuran is found in two low-energy poses (Figure 7b and c), and our results corroborate the assignment made by the original authors. Benzofuran cannot make a hydrogen bond with the protein in any orientation; this ligand was assigned an orientation with the oxygen facing away from the sulfur of Met 109 to avoid a repulsive interaction, and we find this pose at 3 *B* lower in energy than the alternative, inverted pose. While four poses may be possible in all, rotated 180° about the long and short axes, only the two most consistent with the crystallographic data were found in our 10 simulated annealing runs.

The results for benzofuran highlight an additional feature of Monte Carlo sampling as implemented herein. A single simulated annealing run often results in convergence to a

single pose in a given binding site, and interconversion with alternative poses may not occur. In this situation, multiple annealing runs may be necessary to observe alternative low-energy poses. This effect has been previously described for simulated annealing and stochastic conformational searching where a number of annealing runs or stochastic iterations are required to find the set of relevant conformations.[44,45] In the case of benzofuran, when the ligand inserts into the binding site in one favorable orientation, there is little opportunity for it to rotate 180° to another orientation, and successive simulations are required to identify the alternate orientations in the tight pocket. Improvements to the current sampling paradigm, therefore, would minimize this phenomenon in addition to minimizing the impact of the limiting entropy phenomenon.

The ensemble of poses observed for *p*-xylene is interesting in the context of the reported ambiguity of the crystal structure. The computed poses shown in Figure 7g are rotated by about +30 and −30° from the pose assigned by the crystallographers, resulting in an average pose that directly overlays the reported structure. The average RMS deviation across the full ensemble of computed poses is 1.07 Å.

Ethyl benzene (Figure 7h) has the highest RMS deviation, 2.96 Å, because the lowest energy ensemble allows some rotation about the center of the aromatic ring. This freedom sweeps the ethyl group, which increases the RMS due to its lever-arm effect, and is an excellent example of the thermodynamic basis of the current method. The method is not attempting to locate energy minima but rather the ensemble of poses available at the simulation temperature, and the ensembles produced are expected to surround the observed crystallographic pose and thereby create higher RMS differences than minimization-based or docking methods.

o-Xylene (Figure 7f) is twisted and offset from the crystallographic position. This pose has an RMS deviation of 1.65 Å and is the largest real deviation between the predicted and observed poses.

Comparison to MSCS Experiments. The T4 lysozyme binding site is an enclosed, largely hydrophobic binding site, and there is no evidence for significant solvation of the site, either from crystal structures or from our simulations. To demonstrate the grand canonical method for a more open, solvent-accessible binding site, a series of thermolysin crystal structures generated from MSCS experiments was studied. In parallel with grand canonical Monte Carlo sampling, the MSCS technique provides experimental information on the binding of small organic molecules over the entire protein surface.[40,46] Small ligands are soaked into crystal structures at several concentrations, and the positions and orientations are observed using X-ray crystallography. The ligands occupy more positions on the protein surface as the concentrations are increased, and the relative energies of each pose can be inferred from the concentration at which they first appear in the crystal structures. In the MSCS experiments, some low-energy binding sites are found at the interfaces between protein molecules in the crystals. Since our computation is based on a single isolated protein, poses observed in the crystal structure that interact with multiple proteins are not addressed here.

When acetone was cocrystallized with thermolysin at concentrations of 50, 60, and 70%, the lowest-energy pose, observed at the 50% concentration, was observed in the S_1' pocket shown at the center of Figure 8.[22] This low-energy position was simulated in our calculations to have an affinity corresponding to −22.75 *B*. Since the grand canonical simulation produces ensembles of poses at each energy level, the *B* value for the crystallographic position was assigned by finding the simulated pose with the lowest energy and a center of mass closer than 2 Å from the observed center of mass. A less-bound site appears in the crystal structures at a 70% concentration, to the upper right in Figure 8, and poses nearby this region are observed in the simulation at −17.75 *B*. The latter poses are on the surface, not in a binding pocket, and the exact crystallographic pose was not observed in the simulation. However, the rank order of binding of the two sites is reproduced.

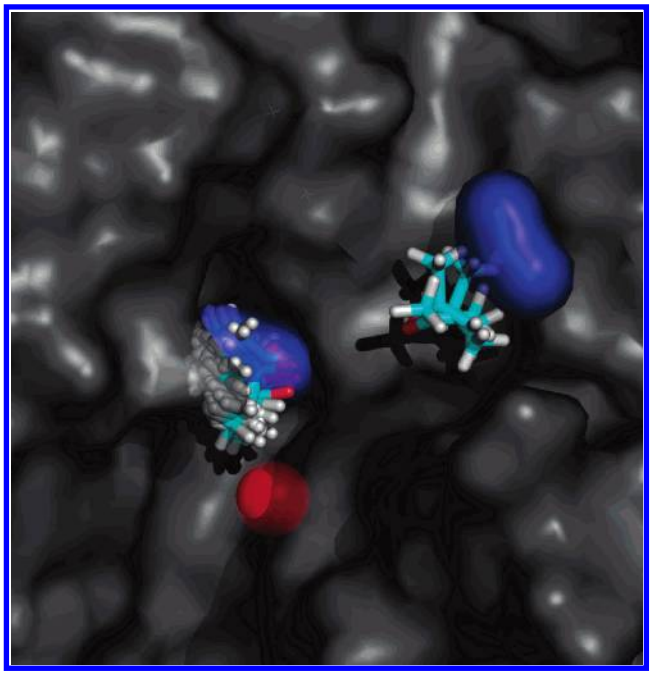

Figure 8. Computed acetone poses on thermolysin at position 1, −22.75 *B* (left), and position 3, −17.75 *B* (right). The crystallographically observed poses are shown as blue surfaces. The zinc atom is shown in red.

In addition, the crystal structure includes multiple crystallographic water molecules. When simulated, the most tightly bound water was observed at −35 *B*, interacting with Zn in the binding site shown in Figure 9. This observation is consistent with the absence of acetone in this site in the crystal structures; this tightly binding water may be difficult for ligands to displace. The binding energies and concentrations are summarized in Table 2.

In another MSCS study, 2-propanol was cocrystallized with thermolysin at a variety of concentrations from 2 to 100%.[23] Twelve poses of 2-propanol are observed in the various concentrations described as poses 1−12 in the original work. Poses 1, 5, 3, 8, and 9 are observed in the binding cleft at concentrations from 5 to 100% and were also found in the grand canonical simulation; the remaining poses occur at protein−protein interaction sites that arise from crystal packing and are not well-represented in simulations of an isolated protein. The binding energies for simulated poses are given in Table 3. The poses are shown in Figure 10. The minimum energy pose, 1 in Figure 10, is observed experimentally at a 5% concentration in the deep S_1' pocket. In contrast, the computationally lowest-energy

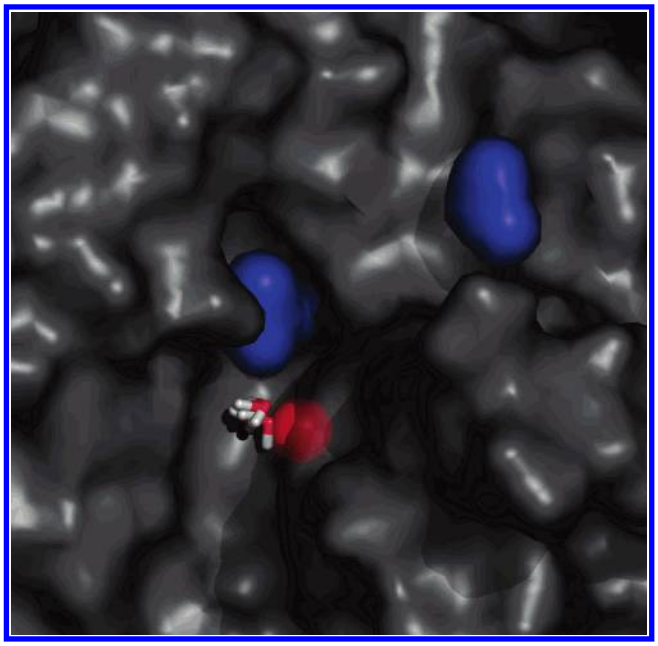

Figure 9. Tightly bound water near the zinc ion at −35 *B*. The crystallographically observed poses of acetone are shown as blue surfaces.

Table 2. Binding Energies of Acetone to Thermolysin in the Binding Pocket

position	% acetone	*B*, acetone
1	50%	−22.75
3	70%	−17.75

Table 3. Binding Energies of Isopropanol and Water to Thermolysin at the Five Sites in the Binding Pocket

position	% 2-propanol	*B*, 2-propanol
1	5	−18.5
3	80	−15.5
8	90	−4.5
5	90	−8.5
9	100	−7.5

pose of 2-propanol in the binding site is identified at −23.5 *B*, interacting with the zinc ion in a pose that is observed below pose 1 (Figure 11). However, as was described for simulations of acetone, water is also strongly bound at that site at −35 *B* (Figure 9), and the computed affinity of water is sufficient to prevent 2-propanol from binding in this location.

Pose 1, observed at a 5% concentration, is found in the simulation at −18.5 *B*, Figure 10.

The pose corresponding to 2-propanol position 3 appears at −15.5 *B*, shown in Figure 12. This position appears in the crystal structure at an 80% 2-propanol concentration. Even though the free energy difference among the sites is small, 2-propanol is correctly predicted to be more binding in sites 1 and 3 than in sites 5, 8, and 9.

One position observed in the crystal structure for both 2-propanol and acetone, identified as position 2 in the original work, is completely within the protein and was not observed in the simulation. This pose occupies a very small, completely included pocket that may be too small to access with Monte Carlo sampling. The issues for sampling this pocket are the same as those discussed for T4 ligands. Consistent with our simulations, this binding site was not found computationally in the original work, or in a subsequent MCSS study.[3]

CONCLUSIONS

The application of the grand canonical ensemble method for computing ligand binding free energies was demonstrated by simulating the binding of small ligands to T4 lysozyme and thermolysin. The computed free energy and entropy is comparable to more complex computations of free energy of the T4−benzene complex. The grand canonical method requires fewer choices of simulation parameters and constraints than most other free energy algorithms. The ultimate purpose of this method is the design of high-affinity ligands, which is done by predicting the affinities and preferred binding poses of small molecular fragments and then computationally assembling the fragments into higher molecular weight ligands. The systems simulated herein were selected because of the availability of both crystallographic and thermodynamic data for low molecular weight ligands. In effect, therefore, we are demonstrating a method for identifying low molecular building blocks as the first step in a full-fledged computational drug design methodology.

The method was shown to be effective in these test systems for identifying binding sites and ranking the free energies of multiple binding sites as well as multiple ligands within each binding site. As suggested by these results, one of the key advantages of the grand canonical Monte Carlo method is that it does not require prior identification of a binding mode before carrying out the computation. A second advantage is that the computation of free energies does not require calibration or tuning for different ligands or different proteins, making the process straightforward to execute and independent of training data. The only parameters set in the current simulations were the number of steps of equilibration and data collection, in addition to the choice of force field.

While the results of the simulations in the current test systems are consistent with the experimental results, there is only limited thermodynamic data available for the binding of low molecular weight ligands to protein, and the results presented are insufficient for a complete validation of the accuracy of the predictions. The primary purpose of this paper, therefore, is to describe and exemplify a new method. To access sufficient data to fully validate the method, it will be necessary to demonstrate the computational assembly of molecular fragments into full-sized ligands and then assess the accuracy versus extensive structure−activity data. This may also require additional terms to account for the entropy loss upon binding the fragments to each other. One of the results of our simulations is that the low-energy ligand-binding positions are identified as an ensemble of poses, each with a comparable free energy. This result is much different than the more conventional experience in identifying the lowest-enthalpy pose, which is generally highly localized.

The major limitation to the method identified in these results is that the computations did not identify the known binding sites for all the ligands in the T4 lysozyme system. This result represents an effect that is inherent, in principle, in any method of searching for very small binding pockets across a protein surface, and the T4 lysozyme system is a stringent test. The results presented for T4 lysozyme, for example, show that the efficiency of locating binding sites

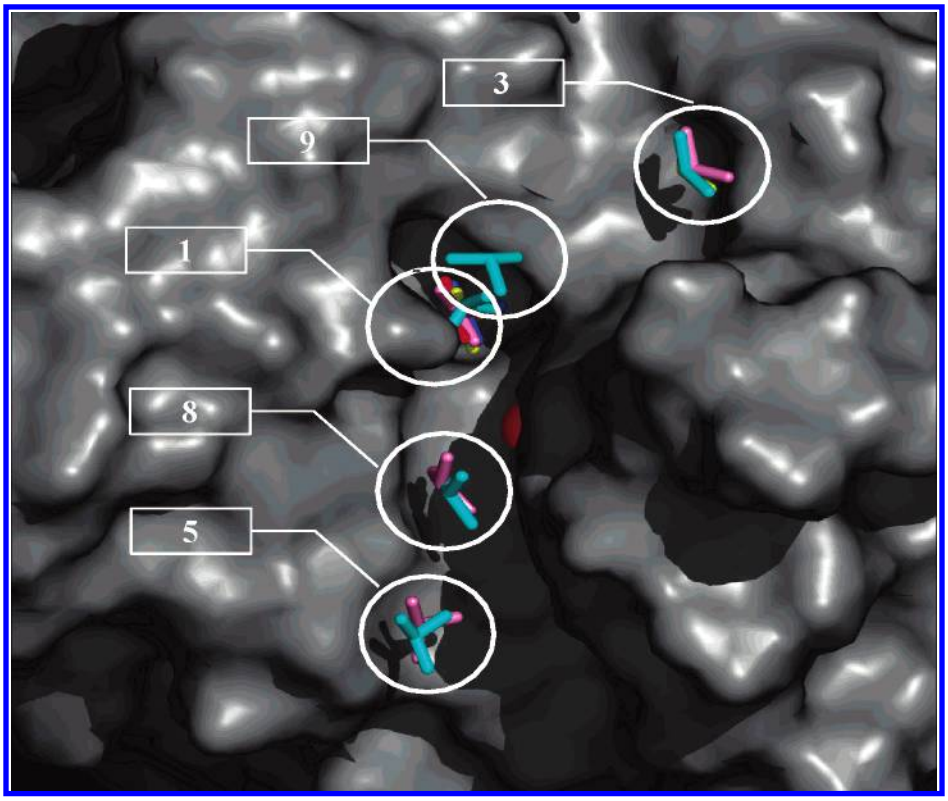

Figure 10. Crystallographically observed poses of 2-propanol on thermolysin, numbered after ref 23. Position 1 is observed at a 5% concentration, position 3 at 80%, and positions 5 and 8 at 90%. The Zn ion is shown in red.

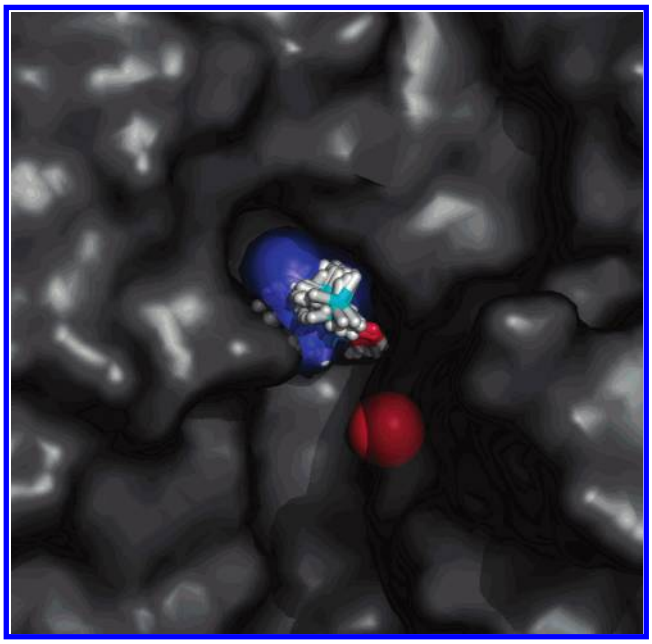

Figure 11. Computed 2-propanol poses near site 1 of thermolysin at −18.5 *B*. The crystallographically observed pose in this region is shown as a blue surface. The zinc atom is shown in red.

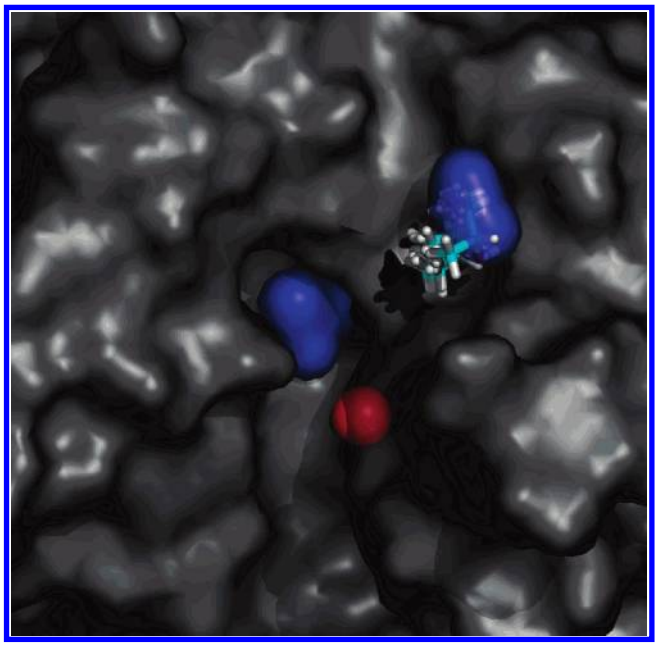

Figure 12. Computed 2-propanol poses near site 3 on thermolysin at −15.5 *B*. The crystallographically observed poses are shown as blue surfaces. The zinc atom is shown in red.

depends on the choice of crystal structure, and the native cocrystal structure for a particular ligand protein is not necessarily the optimum structure. Binding site entropy is a key element, therefore, for understanding the implications of protein flexibility for drug design. Because of the general importance of this concept, we have elected to defer a more in-depth discussion of the problem and its solution to a separate manuscript.

The simulation of 2-propanol and water binding to thermolysin identified the crystallographic water molecules that bind most tightly to this protein and correctly predicted the ability of water to block the binding of 2-propanol to the metal ion in the binding cleft. The simple thermodynamic solvation correction cycle used herein has proven to be sufficient for successful drug design if the exclusion of fragment binding by tightly bound water is included in the model.[47] As can be discerned from the predicted affinities for T4 lysozyme ligands, however, the free energies computed with the method as currently described are accurate only in a relative sense and not in an absolute sense. To

achieve the prediction of absolute binding free energies, it will be necessary to evaluate the efficacy of more extensive solvation corrections and to assess the accuracy of alternative force field models.

ACKNOWLEDGMENT

The authors acknowledge the participation of Charles Karney, Richard Bryan, and John Culp Jr. of Sarnoff, Corp.; Stephan Brunner, Frank Hollinger, David Mosenkis, Judith LaLonde, Subao Rong, Ted Fujimoto, Jen Ludington, Bill Chiang, Qiang Wang, George Talbot, Paolo Carnevali, Sia Meshkat, Keith Milligan, and Zenon Konteatis of Locus Pharmaceuticals; and Mihaly Mezei of Mt. Sinai Medical Center for their work in development of the grand canonical software and methodology.

REFERENCES AND NOTES

(1) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 29−34.

(2) Caflisch, A.; Miranker, A.; Karplus, M. Multiple Copy Simultaneous Search and Construction of Ligands in Binding Sites: Application to Inhibitors of HIV-1 Aspartic Proteinase. *J. Med. Chem.* **1993**, *36*, 2142−2167.

(3) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational Mapping Identifies the Binding Sites of Organic Solvents on Proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *98*, 4290−4295.

(4) Morton, A.; Baase, W. A.; Matthews, B. W. Energetic Origins of Specificity of Ligand Binding in an Interior Nonpolar Cavity of T4 Lysozyme. *Biochemistry* **1995**, *34*, 8564−8575.

(5) Morton, A.; Matthews, B. A. Specificity of Ligand Binding in a Buried Nonpolar Cavity of T4 Lysozyme: Linkage of Dynamics and Structural Plasticity. *Biochemistry* **1995**, *34*, 8576−9888.

(6) Musah, R. A.; Jensen, G. M.; Bunte, S. W.; Rosenfeld, R. J.; Goodin, D. B. Artificial Protein Cavities as Specific Ligand-binding Templates: Characterization of an Engineered Heterocyclic Cation-binding Site that Preserves the Evolved Specificity of the Parent Protein. *J. Mol. Biol*. **2002**, *315*, 845−857.

(7) Rao, B. G.; Kim, E. E.; Murcko, M. A. Calculation of solvation and binding free energy differences between VX-478 and its analogs by free energy perturbation and amsol methods. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 23−30.

(8) Reddy, M. R.; Erion, M. D. *Free Energy Calculations in Rational Drug Design*; Kluwer Academic: New York, 2001.

(9) Wang, J.; Morin, P.; Wang, W.; Kollman P. A. Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc*. **2001**, *123*, 5221−5230. Tembe, B. L.; McCammon, J. A. Ligand−Receptor Interactions. *Comput. Chem.* **1984**, *8*, 281−283.

(10) Warshel, A.; Sussman, F.; King, G. Free energy of charges in solvated proteins: Microscopic calculations using a reversible charging process. *Biochemistry* **1986**, *25*, 8368.

(11) Beveridge, D. L.; DiCapua, F. M. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Chem.* **1989**, *18*, 431.

(12) Jorgensen, W. L. Free Energy Calculations, A Breakthrough for Modeling Organic Chemistry in Solution. *Acc. Chem. Res*. **1989**, *22*, 184−189.

(13) Kollman, P. A. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.

(14) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(15) Jorgensen, W. L. *BOSS*, version 4.2; Yale University: New Haven CT, 2000.

(16) Åqvist, J.; Marelius, J. The Linear Interaction Energy Method for Predicting Ligand Binding Free Energies. *Comb. Chem. High Throughput Screening* **2001**, *4*, 613−626.

(17) Rizzo, R. C.; Udier-Blagovic, M.; Wang, D.-P.; Watkins, E. K.; Kroeger Smith, M. B.; Smith, R. H., Jr.; Tirado-Rives, J.; Jorgensen, W. L. Estimation of Binding Affinities for HEPT and Nevirapine Analogues with HIV-1 Reverse Transcriptase via Monte Carlo Simulations. *J. Med. Chem*. **2002**, *45*, 2970−2987.

(18) Pierce, A. C.; Jorgensen W. L. Estimation of Binding Affinities for Selective Thrombin Inhibitors via Monte Carlo Simulations. *J. Med. Chem.* **2001**, *44*, 1043−1050.

(19) Guarnieri, F. Computational Protein Probing to Identify Binding Sites. U.S. Patent 6,735,530, May 11, 2004.

(20) Guarnieri, F.; Mezei, M. Simulated Annealing of Chemical Potential: A General Procedure for Locating Bound Waters. Application to the Study of the Differential Hydration Propensities of the Major and Minor Grooves of DNA. *J. Am. Chem. Soc*. **1996**, *118*, 8493−8494.

(21) Resat, H.; Mezei, M. Grand Canonical Monte Carlo Simulation of Water Positions in Crystal Hydrates. *J. Am. Chem. Soc.* **1994**, *116*, 7451−7452.

(22) English, A. C.; Groom, C. R.; Hubbard, R. E. Experimental and computational mapping of the binding site of a crystalline protein. *Protein Eng.* **2001**, *14*, 47−95.

(23) English, A. C.; Done, S. H.; Caves, L. S. D.; Groom, C. R.; Hubbard, R. E. Locating interaction sites on proteins: The crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 628−640.

(24) Friedman, H. L. A Course in Statistical Thermodynamics; Prentice Hall: Englewood Cliffs, NJ, 1985.

(25) Nagumo, R.; Takaba, H.; Nakao, S.-I. Prediction of Ideal Permeability of Hydrocarbons through an MFI-Type Zeolite Membrane by a Combined Method Using Molecular Simulation Techniques and Permeation Theory. *J. Phys. Chem. B* **2003**, *107*, 14422−14428.

(26) He, Y.; Seaton, N. A. Experimental and Computer Simulation Studies of the Adsorption of Ethane, Carbon Dioxide, and Their Binary Mixtures in MCM-41. *Langmuir* **2003**, *19*, 10132−10138.

(27) Meredith, J. C.; Johnston, K. P. Density Dependence of Homopolymer Adsorption and Colloidal Interaction Forces in a Supercritical Solvent: Monte Carlo Simulation. *Langmuir* **1999**, *15*, 8037−8044.

(28) Jayaram, B.; Beveridge, D. L. Grand canonical Monte Carlo simulations on aqueous solutions of sodium chloride and sodium DNA: excess chemical potentials and sources of nonideality in electrolyte and polyelectrolyte solutions. *J. Phys. Chem.* **1991**, *95*, 2506−2516.

(29) Adams, D. Grand Canonical Ensemble Monte Carlo for a Lennard Jones Fluid. *J. Mol. Phys.* **1975**, *29*, 307.

(30) Mezei, M. Grand-Canonical Ensemble Monte Carlo Simulation of Dense Fluids: Lennard-Jones, Soft Spheres and Water. *Mol. Phys.* **1987**, *61*, 565−582. Mezei, M. Monte Carlo Method for the Computer Simulation of Fluids. *Mol. Phys.* **1980**, *40*, 901.

(31) Rao, M.; Pangali, C. S.; Berne, B. On the Force-Bias Monte Carlo Simulation of Water: Methodology and Optimization and Comparison with Molecular Dynamics. *J. Mol. Phys.* **1979**, *37*, 79.

(32) Mezei, M. Distance-Scaled Force Biased Monte Carlo Simulation for Solutions Containing a Strongly Interacting Solute. *Mol. Simul.* **1991**, *5*, 405−408.

(33) Mezei, M. Direct Calculation of the Excess Free Energy of the Dense Lennard-Jones Fluid with Nonlinear Thermodynamic Integration. *Mol. Simul.* **1989**, *2*, 201−207.

(34) Cornell, W. D.; Cieplak, P. I.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D.; Spellmeyer, D. C.; Fox, T.; Caldwell J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(35) Breneman, C.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **1990**, *11*, 361−373.

(36) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*. **1983**, *79*, 926.

(37) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.

(38) *Macromodel 8.6*; Schrodinger Inc.: New York, 2004.

(39) *Gaussian 98*; Gaussian Inc.: Wallingford, CT, 1998.

(40) Allen, K. A.; Bellamacina, C. R.; Ding, X.; Jeffrey, C. J.; Mattos, C.; Petsko, G. A.; Ringe, D. An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J. Phys. Chem.* **1996**, *100*, 2605−2611.

(41) Hermans, J.; Wang. L. Inclusion of loss of translational and rotational freedom in theoretical estimates of free energies of binding. Application to a complex of benzene and mutant T4 lysozyme. *J. Am. Chem. Soc*. **1997**, *119*, 2707−2714.

(42) $\Delta G = kT(B - 6.4)$. The factor of 6.4 changes the reference state from 1 molecule per 10^6 Å^3 used in the simulation to a 1 M concentration, which is 1 molecule per 1660 Å^3.
(43) Boresch, S.; Tettinger, F.; Letigeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.
(44) Saunders, M. Stochastic search for the conformations of bicyclic hydrocarbons. *J. Comput. Chem.* **1989**, *10*, 203−208.
(45) Saunders, M.; Houk, K. N.; Wu, Y.-D.; Still, W. C.; Chang, G.; Guida, W. C. Conformations of cycloheptadecane. A comparison of methods for conformational searching. *J. Am. Chem. Soc.* **1990**, *112*, 1419−1427.
(46) Mattos, C.; Ringe, D. Locating and characterizing binding sites on proteins. *Nat. Biotechnol*. **1996**, *14*, 595−599.
(47) Moore, W. R., Jr. Maximizing discovery efficiency with a computationally driven fragment approach. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 355−364.

CI050268F