Multiple Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore and Conformation Space Networks

Alex Dickson*,†,‡ and Samuel D. Lotz†

†*Department of Biochemistry & Molecular Biology, Michigan State University, East Lansing, MI*

‡*Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI*

Received August 3, 2016; E-mail: alexrd@msu.edu

Abstract: We report simulations of full ligand exit pathways for the trypsin-benzamidine system using unbiased dynamics, generated using the sampling technique WExplore. The averaged exit flux yields a ligand exit rate of 180 μs, which is within an order of magnitude of the experimental value. We obtain broad sampling of ligand exit pathways, including three distinct exit channels, two of which are formed through large, rare motions of the loop regions in trypsin. We use our broad set of ligand bound poses to investigate general properties of ligand binding, and observe a trade-off between the direct stabilizing effect of ligand-protein interactions, and the indirect destabilizing effect that the ligand induces in the protein-protein interactions. Significantly, the crystallographic binding poses are distinguished not only because their ligands induce large stabilizing effects, but also because they induce relatively low indirect destabilizations.

The pathways traveled by ligands as they bind to their molecular receptors are important to drug design. Although the binding thermodynamics is purely determined by the endpoints of these pathways, analysis of the entire paths can reveal binding transition states that govern the kinetics of the binding process. Under-appreciated until recently, long residence times have been shown in a handful of systems to be more predictive of in vivo efficacy than the thermodynamics alone.[1,2] Conversely, fast binding and release could also be preferable in some applications, including enzyme inhibition,[3] and for systems where fast clearance of the drug is essential. Robust methods that can predict structure-kinetics relationships would thus be of tremendous value to drug design efforts. Unfortunately, structural details of ligand-binding transition states are difficult to capture experimentally, and ligand binding and release typically occur on timescales that are inaccessible to conventional molecular simulation.

Recently, a handful of cutting-edge applications of molecular dynamics, using either specialized hardware,[4,5] large parallel sampling efforts synthesized with Markov state models,[6–8] or customized enhanced sampling algorithms,[9–12] have been applied to study full ligand binding or unbinding pathways. These have revealed an intricate interplay between the conformations of the ligand and receptor, and are beginning to reveal how biological molecules are controlled by exogenous factors, which is important both for our understanding of biology, and for our ability to design drugs that elicit a desired biomolecular response. Despite some progress, the principles that govern the general relationship between ligand binding and protein stability or protein activity remain elusive. General biophysical properties of protein-ligand interactions are needed to elucidate and predict phenomena such as allosteric signaling networks,[13] and ligand-induced stability changes.[14] This necessitates a general knowledge of how ligand binding is coupled with conformational change in the binding site.

The binding of the ligand benzamidine to trypsin has in recent years served as the system of choice to demonstrate emerging enhanced sampling approaches to study ligand binding.[6,8,9,11,12,15,16] Long simulations of ligand binding synthesized with Markov state models obtained binding rates that showed good agreement with experiment,[6,8,15] but the unbinding rates were consistently over-predicted, owing to the steep free energy barrier of ligand unbinding. Particularly, Plattner and Noé used hundreds of microseconds of simulation to show a dynamic picture of trypsin with two main binding channels and multiple long-lived trypsin conformations.[8] Approaches using metadynamics with path-based order parameters have also obtained unbinding rates,[11] but these were significantly under-predicted, although again the binding rates showed excellent agreement. Teo et al[12] used the Adaptive Multilevel Splitting method to obtain excellent agreement with the experimental rate with modest computational cost, but did not observe some of the long time-scale conformational transitions seen by previous investigations.

Here we use our own technique, WExplore,[17] to investigate a broad set of ligand release pathways in the trypsin-benzamidine system. This and related methods have been used to study protein unfolding, hydration changes near a fluorophore,[18] long time-scale conformational transitions in a RNA helix-helix junction[19] and to generate the ensemble of unbinding pathways of small ligands from the protein FKBP.[20] Like MSM approaches, it uses trajectories that are run with the unbiased Hamiltonian and are suitable for a network-based conformation analysis,[21–23] but it is based on a weighted ensemble of trajectories, and obtains unbinding rates by a different mechanism that does not rely on a Markovian assumption of transitions between regions. A set of trajectories are run in parallel, each with a statistical weight, and these are actively managed every 20 ps using cloning and merging steps that maximize the heterogeneity of the trajectory set. As in the original weighted ensemble algorithm,[24] during cloning the weights are split, and during merging the weights are added. Observables are then computed using weighted averages. One such observable is the flux of trajectories that cross into the unbound state (defined here as all states where the minimum protein-ligand distance is greater than 10 Å). In the nonequilibrium ensemble where trajectories are initiated in the binding site and are terminated in the unbound state, the flux of trajectories into the unbound state per unit time is equal to the unbinding rate.

Figure 1 shows the predicted mean first passage time (MFPT = $1/k_{\text{off}}$) as a function of simulation time for five independent WExplore runs. Each run uses 48 trajectories total that are cloned and merged repeatedly throughout the simulation, and the total sampling time for each run averages 820 ns. A total of 4.1 μs of simulation time is used to generate the average curve (thick orange line) that obtains a final prediction of 180 μs, using the last 10% of the data. Significantly, the individual results differ over eight orders of magnitude, owing to large differences in the weight of the trajectories that break out of the binding pocket. Large, downward jumps in residence time occur (e.g. Run 3, Run 5) when a new exit point is recorded that has a significantly higher weight than the others recorded so far. As such, we expect that extensions of runs 2 and 4 forward would eventually converge toward the mean, although we have found that multiple shorter runs are more efficient than single long ones, as the weight distributions within a run are much more highly correlated than those between the runs. Despite this variability, the averaged trajectory flux gives a MFPT is within an order of magnitude of the experimental value of 1700 μs (Table S1).

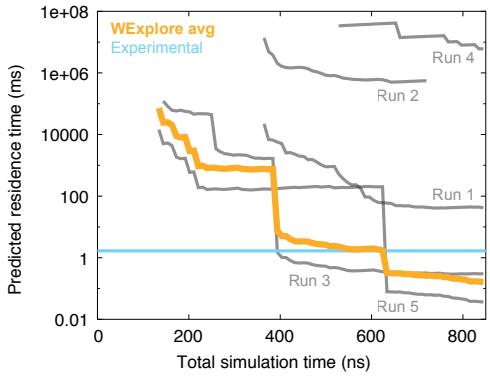

Figure 1. **Mean-first passage time of ligand unbinding.** Predicted residence times are shown for all five WExplore runs (grey). The residence time computed using the average probability flux across all WExplore simulations is shown as a thick orange line, and shows reasonable agreement with the experimentally determined residence time,[25] shown as a horizontal blue line.

As these simulations are conducted using the unbiased Hamiltonian, we can use conformation space networks to synthesize our findings.[21–23] Figure 2A shows the complete network of states visited by all five simulations, created by clustering using a set of 50 ligand-protein distances (see "Clustering" in SI). Node sizes show the state probabilities; the biggest nodes in the top right are the bound states closest to the crystal structure used to initialize the simulations (PDBID 3PTB). Nodes are colored here by solvent accessible surface area (SASA), which reveals a large number of states that are kinetically far from the crystal state, but are still completely buried inside the protein. We find three transition paths that connect the bound and unbound basins (Figure 2B). Path 1 is the direct exit pathway that has been found by all previous investigations. In Path 2, the loop shown in blue (residues 209-218) undergoes a conformational change and creates an alternative pathway for benzamidine release. This path was previously observed by Plattner and Noé,[8] and significant loop motions in this region were also observed using metadynamics.[11] Path 3 involves a similar conformational change in another loop shown in orange (residues 179-190) creating a large set of bound, buried states that have not been previously observed.

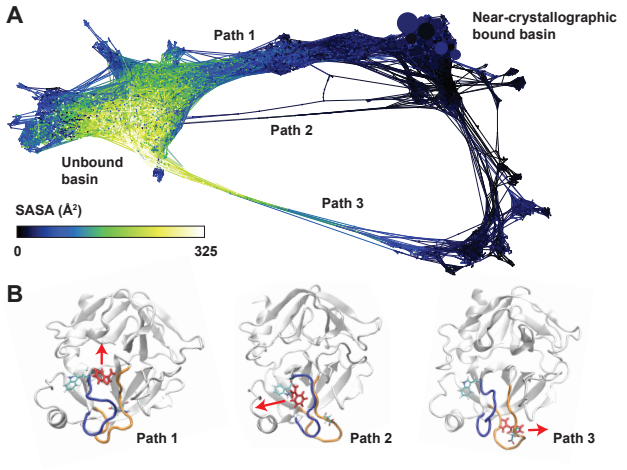

Figure 2. **Trypsin-benzamidine unbinding network shows three exit pathways.** (A) The conformation space network of the trypsin-benzamidine system is shown. The size of the nodes corresponds to the weight of the states, and the node color shows the SASA of a representative structure from that region. The bound and unbound basins are connected by three discrete transition paths, which are labeled. (B) Representative structures are shown that characterize the mechanism of the three transition paths. Benzamidine is shown in red, and the general direction of exit is shown with a red arrow for each pathway. Residues TRP208 and ASP186 are shown in licorice representation, and the loop regions 179-190 and 209-218 are shown in orange and blue, respectively.

We break up our network into communities using a fast stochastic modularity-based community detection algorithm[26] (Figure 3A). We obtain seven communities: two of each representing the bound (B,B*), and path 3 (P3,P3*) states, and one of each representing unbound (U), path 1 (P1) and path 2 (P2). To study these communities we first profile the entire set of ligand-protein hydrogen bonds (H-bond) in the network. For each H-bond that we observe in our simulations, Figure 3B shows the frequency with which it is observed in each of the seven communities. 276 unique acceptor-donor pairs are found with 8621 H-bonding instances total (see "Hydrogen Bond Profiling" in SI). B and B* distributions are dominated by a few high frequency pairs, while U has many low to moderate frequency pairs. The remaining unbinding pathway communities (P1, P2, P3, and P3*) have somewhat heterogeneous distributions but feature some high frequency interaction pairs that are mostly non-overlapping between pathways. This suggests that each pathway may be characterized uniquely by only a few specific interactions. The highest-weighted structures for the highest frequency pairs in each community are shown in "Specific Structure Analysis" in the SI.

Each of the three pathways is not observed by every WExplore simulation (Figure S3). Path 1 is observed in runs 2, 3 and 5, Path 2 is observed only in run 1 and Path 3 is observed only in run 4. Figure S4 shows the free energy of each state, which shows Path 1 to be by far the most probable, Path 2 to be the next-most probable, and Path 3 to be the least probable, consistent with Figure 1. However, this result underscores the ability of WExplore to discover alternative bound conformations, even those that are separated by large free energy barriers, requiring significant rearrangement of local protein structure.

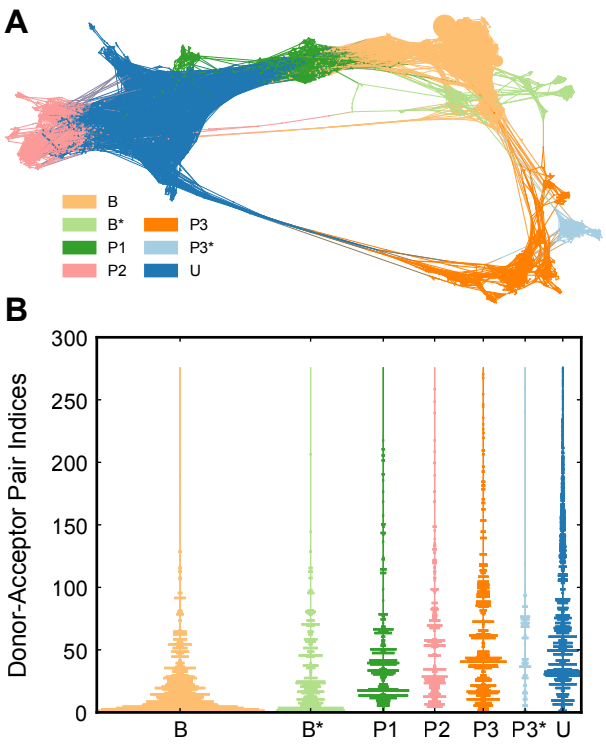

Figure 3. **Community Detection and Hydrogen Bonding Frequencies.** (A) Network plot showing communities of the network. The labels B and B* correspond to the two bound state communities and U corresponds to the unbound states. P3* is classified as a distinct component of the P3 pathway. (B) Violin bar plots of hydrogen bond frequencies. The vertical axis shows the donor-acceptor pairs sorted by their frequency in the whole network. Each violin shows the frequencies observed within each community.

The large set of bound but buried states generated here presents a unique opportunity to examine general properties of ligand-protein interactions across many heterogeneous ligand-protein conformations. Specifically, we examine the relationship between ligand-protein interactions and protein-protein interactions by examining the set of protein atoms that are close enough to directly interact with the ligand (< 4 Å); we call this set of atoms "D_4" (Figure 4C). This selection is unique for each of the 4000 nodes in the network, as the ligand takes on a wide range of conformations in different regions of the protein and the local protein structure also varies significantly. We examine the interaction energies of this selection with its surroundings, and compare it to the interaction energy of the same selection in a set of 10 apo structures. The apo structures chosen are the 10 highest probability states that have a minimum protein-ligand distance greater than 5 Å (Figure S5). These differences in interaction energies reveal the direct and indirect impacts of ligand binding on protein stability.

Throughout the network there is a tradeoff between the stabilizing effect of D_4-Ligand interactions and the destabilizing effect the ligand on D_4-Protein interactions, where "protein" is the entire set of protein atoms that does not include D_4. We also examine the change in D_4-D_4 interaction energies, as well as the change in D_4-Solvent interaction energies and find no significant trends (Figure S6). In the lower right portion of the unbound ensemble, both energies trend to zero, which is expected since the ligand forms fewer interactions with the protein (Figure 4A) and is less disruptive to protein-protein interactions (Figure 4B). Significantly, around the high-probability near-crystallographic bound states the destabilizing effect of the ligand is minimized, which implies that ΔE_{int} D_4-Protein could be a valuable quantity to examine during drug-design efforts (Figure 4E).

The simulations here are performed strictly in the nonequilibrium unbinding ensemble. Using a previous framework[27,28] we can define two basins, B and U, that define the bound and unbound ensembles. B can be defined as the set of conformers where the ligand is within a certain root mean squared distance (say, 3 Å) away from its crystallographic pose, and U is defined as the set of conformers where the minimum protein-ligand distance is farther than 10 Å. Our sampling is composed of two types of paths: $B \to B$ paths, and $B \to U$ paths. By the microscopic reversibility principle, the $B \to U$ ensemble and the $U \to B$ ensemble are identical under equilibrium conditions, however, our simulations will not completely agree with those of the nonequilibrium *binding* ensemble, as those would include $U \to U$ pathways, and neglect $B \to B$ pathways. In Figure S7 we identify nodes in our network that correspond to states previously observed by Buch et al[6] in simulations that mostly approximate the binding ensemble. Two of their states ("S2" and "S3") are found in this work, but the third is not, which implies that it is not represented in the $U \to B$ ensemble, instead lying mostly in the $U \to U$ ensemble.

Plattner and Noé identified two unbinding pathways for trypsin, in one of which the ligand exits through the 209-218 loop, as in our Path 2. This alternative binding pathway was shown to be preferred for alternative trypsin conformations, the highest probability of which was called the "red state". We obtained three representative structures of the red state, and calculate the RMSD to the red state residues 209-218 for each node in the network, averaged over the three conformations. We find some clusters show good local alignment to the red state loop structures, although the global alignments are poor (Figure S8). Interestingly, a large cluster of states showing good local alignment lies at the foot of Path 2 in our conformation space network.

The solid agreement with experimental rates, the broad sampling of pathways and poses, and the relative efficiency of our technique bode well for future applications of WExplore. Drug-like ligands can have residence times approaching minutes or hours, which will be prohibitive to straightforward molecular dynamics for the foreseeable future, but is well within the residence time that we predict for benzamidine dissociating via Path 3, which involves substantial rearrangements of the protein that occur on extremely long timescales. Further testing is needed on ligand dissociation events that occur on longer time scales, which could reveal important information about the optimization of kinetic properties for drugs under development. (Un)binding pathways can also reveal important molecular motions in the receptor that can be used to design new ligands that stabilize alternative receptor conformations. As an example, many states are identified here where the ligand is still deeply buried (SASA ≈ 0) that are kinetically far from the crystallographic starting structure. It is easy to imagine this approach being used to identify such states, which can serve as templates for the design of new ligands that bind via an induced-fit mechanism.

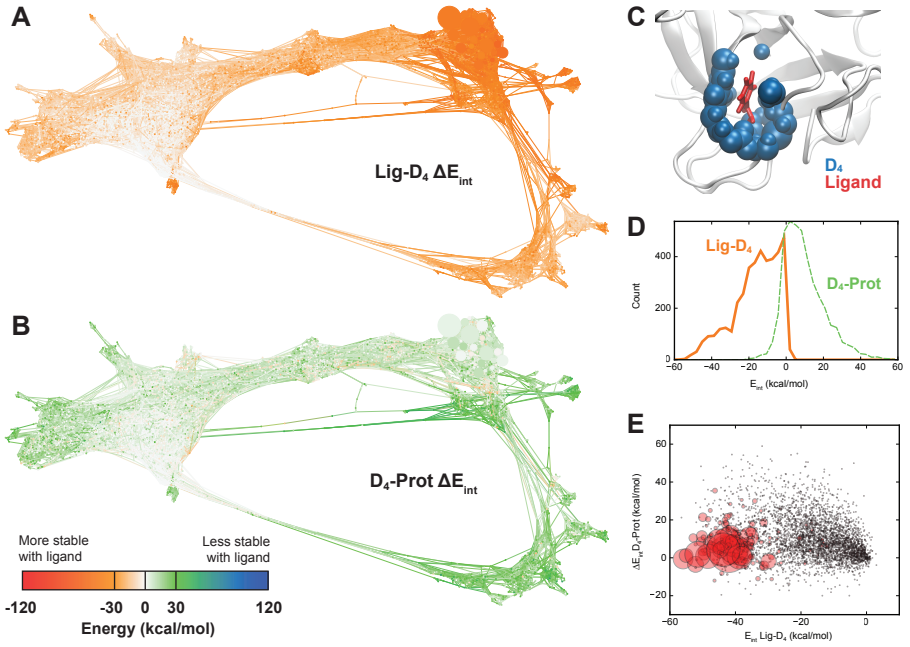

Figure 4. **Protein-ligand contacts disrupt protein-protein interactions.** (A) Conformation space network (CSN) of trypsin-benzamidine colored by the interaction energy between the ligand and a selection of protein atoms that are within 4 Å of the ligand (D_4). The color scale is shown in the bottom left: red and orange indicate stabilizing interactions, white indicates no interaction, and green and blue indicate destabilizing interactions. (B) A CSN colored by the difference in the D_4-protein interaction energy, comparing each node to an ensemble of apo structures, with no ligand. Green nodes indicate that the corresponding D_4 atoms have a lower energy in the apo structures. (C) A visualization of the ligand (in red licorice representation), and the D_4 selection for that pose. (D) Probability distributions for Ligand-D_4 interaction energies (solid orange) and D_4-Protein interaction energies (dashed green). (E) Scatter plot of Ligand-D_4 interaction energies versus D_4-Protein interaction energies. The size of the circles is proportional to the statistical weight of each state, which shows that the high-probability regions are distinguished by their low D_4-Protein destabilization energies as well as their favorable Ligand-D_4 interaction energies.

Acknowledgement The authors thank Nuria Plattner and Frank Noé for sharing representative trypsin conformations from their research. We thank Pratyush Tiwary for a critical reading of the manuscript. We also acknowledge support from the High Performance Computing Center at Michigan State University.

Supporting Information Available: A Supplemental Information file (PDF) contains: a detailed description of Molecular Dyanamics methodology, WExplore sampling and the clustering analysis; structures for each community; a table summarizing previous trypsin-benzamidine results; statistics for the interactions in each community; and other Figures referred to in the main text. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. *Drug Discovery Today* **2013**, *18*, 667–673.
(2) Copeland, R. A. *Nature Reviews Drug Discovery* **2016**, *15*, 87–95.
(3) Yin, N.; Pei, J.; Lai, L. *Molecular bioSystems* **2013**, *9*, 1381–9.
(4) Shaw, D. E. et al. *Communications of the ACM* **2008**, *51*, 91–97.
(5) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. *Journal of the American Chemical Society* **2011**, *133*, 9181–9183.
(6) Buch, I.; Giorgino, T.; De Fabritiis, G. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 10184–10189.
(7) Chodera, J. D.; Noé, F. *Current Opinion in Structural Biology* **2014**, *25*, 135–144.
(8) Plattner, N.; Noé, F. *Nature Communications* **2015**, *6*, 7653.
(9) Limongelli, V.; Bonomi, M.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2013**, *110*, 6358–6363.
(10) Sun, H.; Tian, S.; Zhou, S.; Li, Y.; Li, D.; Xu, L.; Shen, M.; Pan, P.; Hou, T. *Scientific Reports* **2015**, *5*, 8457.
(11) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. *Proceedings of the National Academy of Sciences* **2015**, *112*, E386–E391.
(12) Teo, I.; Mayne, C. G.; Schulten, K.; Lelièvre, T. *Journal of Chemical Theory and Computation* **2016**, acs.jctc.6b00277.
(13) Lu, S.; Li, S.; Zhang, J. *Medicinal Research Reviews* **2014**, *34*, 1242–1285.
(14) Dai, R.; Geders, T. W.; Liu, F.; Park, S. W.; Schnappinger, D.; Aldrich, C. C.; Finzel, B. C. *Journal of Medicinal Chemistry* **2015**, *58*, 5208–5217.
(15) Doerr, S.; De Fabritiis, G. *Journal of Chemical Theory and Computation* **2014**, *10*, 2064–2069.
(16) Takahashi, R.; Gil, V. A.; Guallar, V. *Journal of Chemical Theory and Computation* **2014**, *10*, 282–288.
(17) Dickson, A.; Brooks III, C. L. *The Journal of Physical Chemistry B* **2014**, *118*, 3532–42.
(18) Laricheva, E. N.; Goh, G. B.; Dickson, A.; Brooks III, C. L. *Journal of the American Chemical Society* **2015**, *137*, 2892–2900.
(19) Dickson, A.; Mustoe, A.; Salmon, L.; Brooks III, C. *Nucleic Acids Research* **2014**, *42*, 12126–12137.
(20) Dickson, A.; Lotz, S. D. *The Journal of Physical Chemistry B* **2016**, acs.jpcb.6b04012.
(21) Rao, F.; Caflisch, A. *Journal of Molecular Biology* **2004**, *342*, 299–306.
(22) Huang, D.; Caflisch, A. *PLoS Computational Biology* **2011**, *7*, e1002002.
(23) Dickson, A.; Brooks III, C. L. *Journal of the American Chemical Society* **2013**, *135*, 4729–34.
(24) Huber, G. G. A.; Kim, S. *Biophysical Journal* **1996**, *70*, 97–110.
(25) Guillain, F.; Thusius, D. *Journal of American Chemical Society* **1970**, *92*, 5534–5536.
(26) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, *10008*, 6.
(27) Dickson, A.; Warmflash, A.; Dinner, A. R. *The Journal of Chemical Physics* **2009**, *131*, 154104.
(28) Vanden-Eijnden, E.; Venturoli, M. *The Journal of Chemical Physics* **2009**, *131*, 044120.

Graphical TOC Entry