Understanding Molecular Recognition by Kinetic Network Models Constructed from Molecular Dynamics Simulations

Xuhui Huang[1] and Gianni De Fabritiis[2]

[1]*Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

[2]*Computational Biochemistry and Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain*

(Dated: August 14, 2012)

Abstract

Molecular recognition, the process by which biological macromolecules selectively bind, plays an important role in many biological processes. Molecular simulations hold great potential to reveal the chemical details of molecular recognition and to complement experiments. However, it is challenging to reconstruct the binding process for two-body systems like protein-ligand complexes because the system's dynamics occurs on significantly different timescales due to several physical processes involved, such as diffusion, local interactions and conformational changes. In this chapter, we review some recent progress on applying Markov state models (MSMs) to two-body systems. Emphasis is placed on the value of projecting dynamics onto collective reaction coordinates and treating the ligand dynamics with different resolution models depending on the proximity of the protein and ligand. We also discuss some future directions on constructing MSMs to investigate molecular recognition processes.

I. INTRODUCTION

Molecular recognition is the process by which macromolecules selectively interact. Virtually all biological phenomena depend in some way on specific molecular recognition, and thus an understanding of the process is of central importance in the study of biology. One critically important factor is that proteins exist as a statistical ensemble of conformers, which are transitory excited-states (having higher free energy) in the protein in normal solvated conditions; however, these excited states can become preferred upon binding, by shifting the equilibrium distribution towards them. For example, a thermally-accessible conformer that is 2 kBT higher in free energy would exist in just 13% of the molecules in solution (according to Boltzmann probability), yet upon binding could become the most favored state.

There are two popular models aiming to explain the mechanisms of molecular recognition based on a dual dynamic mechanism: "induced-fit" (see reaction R4-R3 in Figure 1) and "conformational selection" (see reaction R1-R2 in Figure 1). In the induced fit model introduced by Koshland [1], the apo protein only exists in the unbound form and the interactions with the ligand induce the protein to reach the bound state. In the conformational selection model [2–8], the protein's intrinsic dynamics may lead it to sample not only the unbound state but also the minor bound state. The ligand may then selectively bind to the pre-existing bound conformation and further increase its population. These two models are not mutually exclusive and both mechanisms may play a role as binding and folding are both search processes over a rugged free energy surface. For example, by binding to protein A, protein B may be stabilized in an excited conformation B* which can facilitate binding to other proteins or ligands determining a cellular signaling cascade.

Many molecular recognition processes involve significant conformational changes of one or both binding partners. For example, Periplasmic Binding Proteins (PBPs) can undergo a large-scale hinge bending motion between two domains from an open to a closed state upon substrate binding [9–12]. In these systems, the interplay between protein structure and dynamics upon substrate binding may ultimately determine the binding mechanisms.

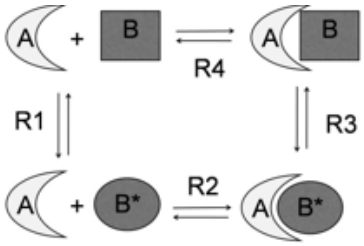

FIG. 1. Conformational selection (R1-R2) v.s. induced fit (R3-R4). A schematic diagram for the two popular binding mechanisms are displayed.

Computer modeling has been shown to be a valuable approach to complement experimental techniques to reveal the chemical details of molecular recognition mechanisms. Markov state models (MSMs) are kinetic network models that hold great potential for understanding the mechanisms of molecular recognition events from computer simulations.

Although MSMs have been successfully applied to study conformational dynamics of one-body systems such as a single protein or RNA[13–17], constructing MSMs to investigate the protein-ligand binding process is challenging because the ligand dynamics normally occurs on two significantly different timescales due to its interactions with the protein. In particular, a ligand's dynamics tend to be very slow when interacting with a protein, but ligands typically diffuse very quickly in solution. Therefore, the standard methods for constructing MSMs through a uniform clustering at a single resolution are often insufficient for properly describing ligand binding. In this chapter, we will review some recent progress on constructing MSMs for two-body systems associated with large conformational changes where the ligand dynamics occurs at a mixture of different resolutions.

II. METHODOLOGY

A. Projected dynamics MSMs

The use of reaction coordinates to project the high dimensional space of a molecular systems into a small dimensional space has been used for many years, especially in the setting of biased dynamics[18]. These biased dynamics schemes offer the advantage of speeding up the global dynamics provided that the reaction coordinates is a good one (no other degree of freedom is slower). A good reaction coordinate also provides a way to compute realistic energetic maps of the phenomena.

A different approach is to use MSM to analyze a set of unbiased trajectories using a low dimensional space to build the Markov model. In this case, the reaction coordinate does not have to be perfect as the dynamics is only projected into this space but the kinetics are well recovered provided that the runs are long enough. This is the approach for instance of Ref. [19], in which the binding pathway of a small molecule is constructed by using a simple reaction coordinate, the three-dimensional position of one of its atoms.

B. Automated methods for constructing MSMs for one-body systems

In many studies, MSMs are constructed by grouping conformations into a number of metastable states and then counting the transitions between these states without projecting the dynamics onto certain reaction coordinates. Automated methods based on a splitting-and-lumping scheme have been developed to construct MSMs for one-body systems[20, 21]. Since these methods have been discussed in detail in other chapters, we briefly review the general procedure here: first, a geometric clustering is applied to divide the MD conformations into a large number of small clusters. This assumes that conformations within the same cluster are kinetically similar because of their structural similarity. Next, clusters that can interconvert quickly are grouped together into the same metastable state to construct an MSM model. Finally, we can calculate thermodynamic

and kinetic properties of interest if the model is Markovian.

C. Constructing MSMs for two-body systems

The above splitting-and-lumping algorithm for one-body systems is often unideal for two-body systems because the dynamics in these systems occur at a mixture of different timescales due to the interactions between the binding partners (see Fig. 5). For example, the ligand diffuses freely in the solvent in the un-bound state. While in the bound state, the ligand forms stable interactions with the protein and its dynamics is slow and strongly correlated with the protein conformation. A kinetically-relevant, uniform clustering at a single resolution as in the splitting-and-lumping algorithm is often difficult to achieve for these two-body systems. If the resolution of the clustering is too low, one cannot split enough in the region where the ligand binds to the protein. On the other hand, if the clustering resolution is too high, there may not be a sufficient number of conformations in each cluster in the unbound region (e.g. many clusters in the unbound region end up containing a single conformation).

In order to address this issue, Silva *et al.*[22] have performed independent clustering at two different resolutions: a high-resolution clustering (or larger number of clusters) on conformations where the ligand binds to the protein and a low-resolution clustering (or smaller number of clusters) on conformations where the ligand diffuses in solution. Kinetic lumping was then used within each region to generate a set of metastable states. Finally, the two sets of metastable states were combined into a single MSM. In this algorithm, a hard distance cut-off (5 $\bar{A}$ between the ligand and protein) is set to separate the fast and slow motion regions for the ligand. This algorithm was shown to be useful for dealing with protein-ligand binding systems, but it may introduce errors on the boundary between the two regions due to the hard distance separation.

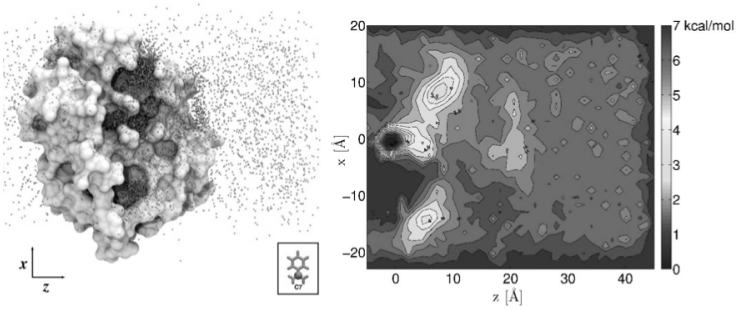

FIG. 2. The spatial positions visited by diffusion of Benzamidine show clearly few metastable states, but only a MSM analysis of the trajectories can recover the free energy profile in three-dimension (here projected in two dimensions for clarity). These figures are adapted from [19].

III. EXAMPLE TRYPSIN-BENZAMIDINE BINDING

In this section, we use the molecular recognition process of trypsin-benzamine as an exemplary case of rigid binding. In Ref [19], a kinetic model for the binding process of serine protease beta-trypsin inhibitor benzamidine was obtained from extensive high-throughput all-atom molecular dynamics (MD) simulations using the ACEMD[23] software on the GPUGRID distributed computing network[24].

The analysis of 495 trajectories of free diffusion of benzamidine around trypsin each of 100 ns of length lead to 187 trajectories (37%) which successfully recovered the bound pose in the binding pocket with an RMSD compared to the crystal structure of less than 2 $\bar{A}$. Several clusters of benzamidine on the surface of trypsin can be observed in Figure 2, which indicates a rather more complex pathway of binding than expected instead of a of simple pathway directly from the bulk. Some trajectories reach the bound crystallographic pose just after 10-15 ns of simulation while some reach the binding pocket only after 90 ns, but the majority of the trajectories do not enter the binding site within 100 ns, as should be expected in such a short time frame. Nevertheless, these simulations provide enough data to carry out a detailed quantitative analysis of the binding pathway.

An aggregate of 50 microseconds of trajectory data have been used to construct a

Markov State Model (MSM) of the binding process of benzamidine to trypsin. The MSM was built using the three-dimensional reaction coordinate defined by the coordinates of the C7-atom of benzamidine (Figure 2). A projection in two dimensions of the energetic profile is also shown highlighting the secondary and the main binding sites (Figure 2). This surface is recovered directly by solving the stationary distribution of the MSM. Using a formula derived in[19], it is possible to compute directly from the energetic map the standard free energy of binding of the ligand of approximately 5.2 kcal/mol compared to an experimental one of 6.2 kcal/mol. A kinetic model can also be built to measure on and off rates which compare well with experiments [19].

An analysis of the slowest eigenvectors of the MSM also allows the reconstruction of the binding pathway. Considering the slowest modes, we see transitions from site S0 to S1 (Figure 3a) and collectively from sites S0/S1 to sites S3/S4 (Figure 3b), corresponding to the diffusion of the ligand from bulk to the first structural contact with the protein. At a slower timescale, there are transitions between sites S2 and S3 (Figure 3c). Site S2 is a secondary binding pocket but not directly involved in the binding pathway. Finally, the rate-limiting step of the process is the transition to the bound site S4 (Figure 3d) and preferentially coming from S3 interestingly rolling on the surface of the protein.

The case of Trypsin-Benzamidine represents a best case scenario where both the ligand and protein are relatively inflexible. While the methodology is not limited to this case, more flexible ligands would require substantially more time to bind. Conformational changes in the protein could also forbid binding all-together until certain loops open. All these factors imply that while the current methodology is very promising, more work is necessary in order to efficiently resolve complex molecular recognition processes.

IV. EXAMPLE LAO PROTEIN BINDING

In this section, we use the Lysine-, Arginine-, Ornithine-binding (LAO) protein as an example to demonstrate the power of MSMs for studying protein-ligand binding mechanisms. The LAO protein is one of the Periplasmic Binding Proteins (PBPs), which

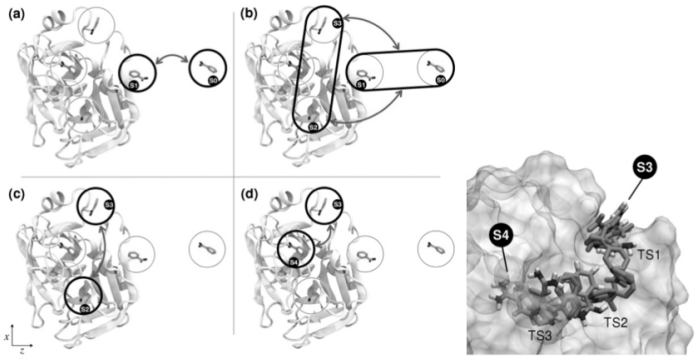

FIG. 3. Main binding modes for benzamidine on trypsin. (a) the encounter of the ligand with the protein, (b) binding to two secondary binding sites, (c) exchange between the secondary binding sites, (d) final pathway of binding into the catalytic site, which show a curious rolling of the ligand on the surface of the protein as the most probable path. These figures are adapted from [19].

is an attractive class of systems for studying the mechanisms of molecular recognition events[25, 26]. With more than 100 crystal structures available, different PBPs can bind to a large variety of substrates including amino acids, sugars, small peptides, etc. However, all PBPs share similar tertiary structures containing two globular domains connected by a hinge region with the binding site at the domain-domain interface. They can undergo a large-scale hinge bending motion from an open to a closed state upon ligand binding (see Figure 4). These features make PBPs a good model system to investigate the coupling between ligand binding and protein conformational changes.

Molecular dynamics (MD) simulations have shown that the ligand dynamics in the LAO system indeed displays a mixture of different timescales. Silva *et. al*[22] performed a set of sixty-five 200-ns MD simulations of the ligand Arginine binding to the LAO protein. From these simulations, they calculated the ligand rotational autocorrelation functions for three conformational states: unbound state, encounter complex, and bound state. As shown in Figure. 5, the ligand can rotate quickly when it undergoes free diffusion in the solvent, but the ligand rotation is largely restrained when it binds to

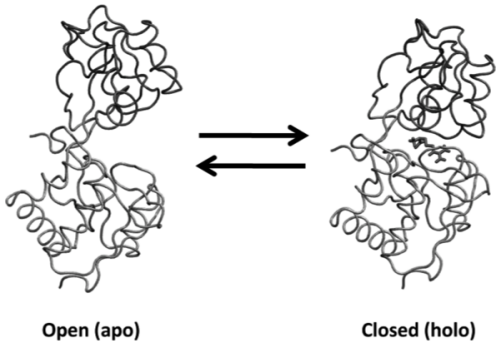

FIG. 4. The Lysine-, Arginine-, Ornithine-binding (LAO) Protein undergoes large domain displacement from the open (left, PDB id: 2LAO) to the closed (right, PDB id: 1LAF) state upon the binding of Arginine (sticks). This figure is reproduced from [22].

the protein. Therefore, when they later constructed MSMs from these MD simulations, they performed structural clustering at two different resolutions in the "splitting" stage of the splitting-and-lumping algorithm. In the low-resolution (or fewer clusters) region, the dynamics of the ligand is fast, so that only the center of mass motion of the ligand is considered. In the high-resolution (or more clusters) region, the dynamics of the ligand is constrained due to its strong interactions with the protein, so that motions of all ligand heavy atoms are considered. Finally, they performed kinetic lumping at each region to generate a set of metastable states, and combine them into a single 54-state MSM. This MSM is shown to reproduce the structure of the bound state, experimental binding free energy, and association rate with reasonable accuracy[22].

MSMs[22] suggest a two-step binding mechanism for the LAO protein with a number of intermediate states and parallel binding pathways (see the ten most probable binding pathways predicted by the MSM as shown in Figure. 6). In the first step, the ligand binds to the protein to form an encounter complex. In the encounter complex state, the protein is partially closed and only weakly interacts with the ligand. RMSD analysis shows that the structure of individual protein domains in the encounter complex is very similar to those in the unbound and bound X-ray structures (with RMSD mostly $< 2\bar{A}$). Therefore the conformational change from either unbound or bound state to the encounter complex

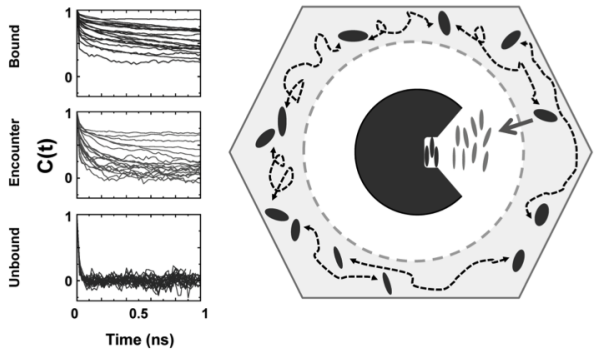

FIG. 5. A figure demonstrating the challenge for constructing kinetic network models for two-body systems where the ligand dynamics occur at a mixture of different timescales as shown by the rotational autocorrelation functions of the ligand in the LAO protein system. The decay of ligand rotational autocorrelation functions is much faster in the unbound state (inside hexagon in right panel and bottom of the left panel) than the encounter complex (inside dashed circle in right panel and middle of the left panel) and the bound state (in binding pocket in right panel and top of the left panel). On the right panel, a schematic figure illustrates the ligand positions in different states. This figure is reproduced from [22].

conformation may be achieved through domain rigid body rotations. All major pathways pass through the encounter complex state, which serves as a gatekeeper for binding. This process is dominated by conformational selection. In the second step, the protein-ligand interactions induce conformational changes to reach the bound state.

V. DISCUSSION AND FUTURE PERSPECTIVE

One major advantage of MSMs is that they can dissect atomistic details of molecular recognition. For instance, Silva et al.[22] have observed roles for both conformational selection and induced fit in LAO binding, as well as an encounter complex intermediate state. RecentNMR studies by Tang *et. al*[27] have also suggested the duality of conformational selection and induced fit for the binding of PBPs. Using NMR with paramagnetic

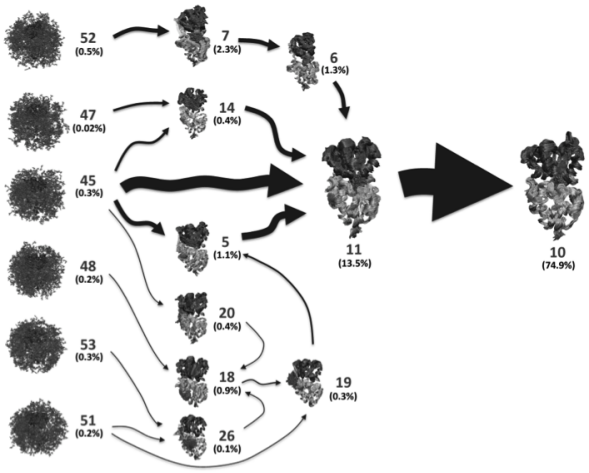

FIG. 6. Ten highest flux binding pathways from the unbound states (left) to the bound state (right) of the LAO protein are superimposed. The arrow sizes are proportional to the flux. State numbers and their equilibrium population calculated from a 54-state Markov State Model are also shown. This figure is reproduced from [22].

relaxation enhancement (PRE), they have identified a minor (5%) partially closed form in equilibrium with the major open form for another PDB, the maltose-binding protein[27]. Based on these observations, they proposed that this partially closed state may be available for the binding of the ligand through conformational selection and this binding could then facilitate the transition to the bound state via the induced fit mechanism. This model was proposed mainly based on experiments in the absence of the ligand. MSMs have the advantage that they can directly observe the interplay between protein conformational changes and ligand dynamics from simulations of ligand binding at atomic resolution. The other main advantage of MSMs is that they can help bridge the timescale gap between the experiments and atomistic MD simulations. For many two-body systems such as protein-ligand binding and protein-protein interactions, the association timescales (millisecond or longer) would be too long to be reached by straightforward atomistic MD simulations. MSMs built from many independent microsecond simulations, however, have already proven capable of capturing protein-folding events that occur at tens of milliseconds timescales[16]. They can thus likely be applied to study slow protein-ligand binding

events too. For the LAO protein discussed above, the timescale is fast enough to observe multiple binding and unbinding events within our sixty-five 200-ns simulations. Even for this case, it would still be challenging to extract a complete picture of the binding mechanism from a single long simulation, because one would need this single simulation to be at least tens of microseconds long so that many binding/unbinding transitions occur (the average transition time from the unbound to the bound state is 2 microseconds). While such a trajectory could be run, scaling to study even slower events (i.e. at millisecond timescales) would not be possible.

In the future, new algorithms are needed for better integrating different timescales of ligand dynamics when constructing the kinetic network models. Silva et al.[22] used a hard distance cut-off (5 $\bar{A}$ between the ligand and protein) to separate the slow and fast motion regions for the ligand, and then performed independent kinetic lumping for each before recombining the two sets of metastable states into a single MSM. As we discussed above, this algorithm may introduce errors on the boundary between the two regions due to this sharp distance cut-off. One potential way to avoid this problem is to directly integrate the geometric "splitting" and kinetic "lumping" steps during model construction. This may require the consideration of both the structural similarity and the kinetic connectivity when performing the clustering. Moreover, kinetic network models containing nodes at transition states could also greatly aid in understanding the mechanisms of molecular recognition events, even though these models are no longer Markovian. They are particularly useful for systems where sufficient sampling can already be achieved by straightforward MD simulations.

References

[1] Koshland, D. E., Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A 1958, 44 (2), 98-104.

[2] Kumar, S.; Ma, B.; Tsai, C. J.; Sinha, N.; Nussinov, R., Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci 2000, 9 (1), 10-9.

[3] Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R., Folding funnels and binding mechanisms. Protein Eng 1999, 12 (9), 713-20.

[4] Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R., Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. Protein Sci 2002, 11 (2), 184-97.

[5] Tsai, C. J.; Kumar, S.; Ma, B.; Nussinov, R., Folding funnels, binding funnels, and protein function. Protein Sci 1999, 8 (6), 1181-90.

[6] Tsai, C. J.; Ma, B.; Nussinov, R., Folding and binding cascades: shifts in energy landscapes. Proc Natl Acad Sci U S A 1999, 96 (18), 9970-2.

[7] Arora, K.; Brooks, C. L., 3rd, Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. Proc Natl Acad Sci U S A 2007, 104 (47), 18496-501.

[8] Bahar, I.; Chennubhotla, C.; Tobi, D., Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation. Curr Opin Struct Biol 2007, 17 (6), 633-40.

[9] Oh, B. H.; Ames, G. F.; Kim, S. H., Structural basis for multiple ligand specificity of the periplasmic lysine-, arginine-, ornithine-binding protein. J. Biol. Chem. 1994, 269 (42), 26323-26330.

[10] Ames, G. F., Bacterial periplasmic transport systems: structure, mechanism, and evolution. Annu Rev Biochem 1986, 55, 397-425.

[11] Pang, A.; Arinaminpathy, Y.; Sansom, M. S.; Biggin, P. C., Comparative molecular dynamics-similar folds and similar motions? Proteins 2005, 61 (4), 809-822.

[12] Stockner, T.; Vogel, H.; Tieleman, D., A salt-bridge motif involved in ligand binding and large-scale domain motions of the maltose-binding protein. Biophys. J. 2005, 89 (5), 3362-3371.

[13] Buchete, N. V.; Hummer, G., Coarse master equations for peptide folding dynamics. J Phys Chem B 2008, 112 (19), 6057-69.

[14] Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R., Constructing the equilib-

rium ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci U S A 2009, 106 (45), 19011-6.

[15] Huang, X.; Yao, Y.; Bowman, G. R.; Sun, J.; Guibas, L. J.; Carlsson, G.; Pande, V. S., Constructing multi-resolution markov state models (msms) to elucidate RNA hairpin folding mechanisms. Pac Symp Biocomput 2010, 228-39.

[16] Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S., Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). J Am Chem Soc 132 (5), 1526-8.

[17] Morcos, F.; Chatterjee, S.; McClendon, C. L.; Brenner, P. R.; Lopez-Rendon, R.; Zintsmaster, J.; Ercsey-Ravasz, M.; Sweet, C. R.; Jacobson, M. P.; Peng, J. W.; Izaguirre, J. A., Modeling conformational ensembles of slow functional motions in Pin1-WW. PLoS Comput Biol 2010, 6 (12), e1001015.

[18] Buch, I.; Sadiq, S. K.; De Fabritiis, G., Optimized Potential of Mean Force Calculations for Standard Binding Free Energies. Journal of Chemical Theory and Computation 2011, 7 (6), 1765-1772.

[19] Buch, I.; Giorgino, T.; De Fabritiis, G., Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America 2011, 108 (25), 10184-10189.

[20] Bowman, G. R.; Huang, X.; Pande, V. S., Using generalized ensemble simulations and Markov state models to identify conformational states. Methods 2009, 49 (2), 197-201.

[21] Huang, X.; Bowman, G. R.; Bacallado, S.; Pande, V. S., Rapid equilibrium sampling initiated from nonequilibrium data. Proc Natl Acad Sci U S A 2009, 106 (47), 19765-9.

[22] Oh, B. H.; Pandit, J.; Kang, C. H.; Nikaido, K.; Gokcen, S.; Ames, G. F.; Kim, S. H., Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand. J. Biol. Chem. 1993, 268 (15), 11348-11355.

[23] Harvey, M. J.; Giupponi, G.; De Fabritiis, G., ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. Journal of Chemical Theory and Computation 2009, 5 (6), 1632-1639.

[24] Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G., High-Throughput

All-Atom Molecular Dynamics Simulations Using Distributed Computing. Journal of Chemical Information and Modeling 2010, 50 (3), 397-403.

[25] Bucher, D.; Grant, B. J.; Markwick, P. R.; McCammon, J. A., Accessing a hidden conformation of the maltose binding protein using accelerated molecular dynamics. PLoS Comput Biol 2011, 7 (4), e1002034.

[26] Quiocho, F. A.; Ledvina, P. S., Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. Mol Microbiol 1996, 20 (1), 17-25.

[27] Tang, C.; Schwieters, C. D.; Clore, G. M., Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. Nature 2007, 449 (7165), 1078-82.