



# Quantitatively Characterizing the Ligand Binding Mechanisms of Choline Binding Protein Using Markov State Model Analysis

Shuo Gu<sup>1</sup>, Daniel-Adriano Silva<sup>1,2</sup>, Luming Meng<sup>1</sup>, Alexander Yue<sup>1</sup>, Xuhui Huang<sup>1,3,4\*</sup>

**1** Department of Chemistry, Institute for Advance Study and School of Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, **2** Department of Biochemistry, University of Washington, Seattle, Washington, United States of America, **3** Division of Biomedical Engineering, Institute for Advance Study and School of Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, **4** Center of Systems Biology and Human Health, Institute for Advance Study and School of Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

## Abstract

Protein-ligand recognition plays key roles in many biological processes. One of the most fascinating questions about protein-ligand recognition is to understand its underlying mechanism, which often results from a combination of induced fit and conformational selection. In this study, we have developed a three-pronged approach of Markov State Models, Molecular Dynamics simulations, and flux analysis to determine the contribution of each model. Using this approach, we have quantified the recognition mechanism of the choline binding protein (ChoX) to be ~90% conformational selection dominant under experimental conditions. This is achieved by recovering all the necessary parameters for the flux analysis in combination with available experimental data. Our results also suggest that ChoX has several metastable conformational states, of which an apo-closed state is dominant, consistent with previous experimental findings. Our methodology holds great potential to be widely applied to understand recognition mechanisms underlining many fundamental biological processes.

**Citation:** Gu S, Silva D-A, Meng L, Yue A, Huang X (2014) Quantitatively Characterizing the Ligand Binding Mechanisms of Choline Binding Protein Using Markov State Model Analysis. *PLoS Comput Biol* 10(8): e1003767. doi:10.1371/journal.pcbi.1003767

**Editor:** Thomas Weigl, Max Planck Institute of Colloids and Interfaces, Germany

**Received:** November 29, 2013; **Accepted:** June 22, 2014; **Published:** August 7, 2014

**Copyright:** © 2014 Gu et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was funded by Hong Kong Research Grants Council ECS 609813, M-HKUST601/13, AoE/M-09/12, and T13-607/12R. National Science Foundation of China: 21273188 and National Basic Research Program of China (973 Program 2013CB834703). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* Email: xuhuihuang@ust.hk

## Introduction

Protein-ligand recognition plays a key role in many aspects of biological processes, such as enzyme catalysis, substrate translocation, and drug therapy [1]. Current studies indicate two prevailing models to address the recognition process: the induced fit model (where ligand binding induces conformational changes of the protein) and the conformational selection model (where the ligand selects a pre-existing conformation of the protein to bind), both of which describe extreme situations (SI Fig. S1) [2–10]. Recent studies, however, suggest that many realistic systems show characteristics of both mechanisms [11–17].

A better understanding of the role of the two models may lead to an increased utilization of protein engineering techniques – for example we may fine tune their respective contributions to allow the creation of new properties [18]. In particular, augmenting the relative contribution of the induced fit mechanism might increase the binding specificity of a protein receptor. Direct applications of such protein engineering can also lead to better chemical sensors [19].

Hammes *et al.* have developed an analytical model based on flux analysis to determine the contribution of conformational selection and induced fit mechanism [20]. However, difficulties arise in obtaining the thermodynamic and kinetic parameters necessary for the flux analysis from experiments. For example, it

is difficult for experiments to directly examine which conformational ligands choose to bind in the conformational selection model, or observe protein conformational changes upon the ligand binding in the induced fit model. Recent progress of NMR techniques such as paramagnetic relaxation enhancement and residual dipolar coupling enable the detection of metastable conformations of the apo protein in solution, and further provide dynamic information for transitions between these conformations [21–24]. However, it is still difficult to apply these techniques to monitor the dynamics of the ligand binding process. In any case, as long as one obtains necessary kinetic and thermodynamic parameters, the flux through each pathway can be quantified, allowing one to assign a percentage to the involvement of induced fit or conformational selection for a particular recognition process.

Quantifying the flux is difficult by direct Molecular Dynamics (MD) simulations as well. The current timescale of MD simulations, mostly on the order of tens to hundreds of nanoseconds, is far too short to witness many biological events which occur on the order of milliseconds to seconds [25]. Only if a specific protein-ligand recognition process occurs very quickly can direct MD simulations be efficient, as our previous study on the binding mechanism of L-Lysine-, L-Arginine-, L-Ornithine-binding protein (LAOBP) demonstrated [26]. For that particular system, we used a total of 13  $\mu$ s MD simulations and Markov State

## Author Summary

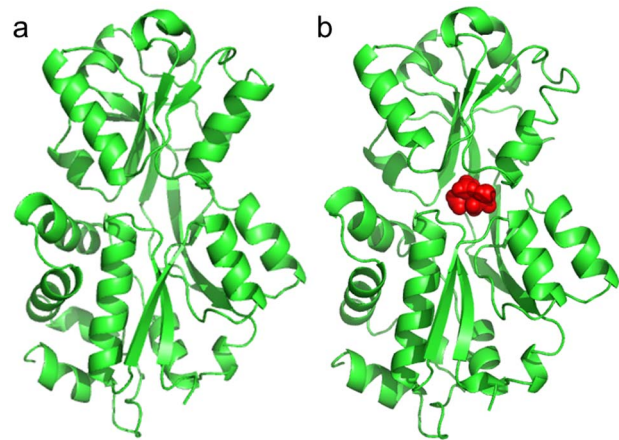
Molecular recognition plays important roles in numerous biological processes including gene regulation, cell signaling and enzymatic activity. It has been suggested that molecular recognition employs a variety of mechanisms, ranging from induced fit to conformational selection. In many realistic systems, conformational selection and induced fit are not mutually exclusive. An analytical flux analysis has been developed to determine the contribution of each model, but it is extremely challenging to obtain the necessary kinetic parameters for this flux analysis through experimental techniques. In this work, we have developed an approach integrating Markov State Models, molecular dynamics simulations, and flux analysis to tackle this problem. Using this approach, we have quantified the recognition mechanism of the choline binding protein to be ~90% conformational selection dominant in the experimental conditions. Our methodology provides a way to quantify the molecular recognition mechanisms that are extremely difficult to be directly accessed by experiments. This opens up numerous possibilities for *in silico* design to fine tune the recognition event either to increase the degree of conformational selection or induced fit, so that new properties could be created to accommodate the needs of protein engineering, drug development and beyond.

Models (MSMs) to examine the binding events between arginine and LAOBP that occurs at a couple of microseconds.

In this work, we propose a novel approach to quantify the flux following conformational selection and induced fit model for a particular molecular recognition process. By combining the techniques of MSMs for apo protein dynamics, direct MD simulations of the protein-ligand binding and flux analysis, we offer a systematic method for finding the necessary kinetic parameters to quantitatively measure the portion of each flux through two binding pathways. Our application of such an approach in choline binding protein (ChoX) [27], a periplasmic binding protein (PBP) [28] from the ATP-binding cassette transporter ChoVWX, demonstrates that our method can explore the free energy landscape and successfully quantify the independent contribution of two concurrent binding mechanisms – induced fit and conformational selection – in a complex realistic protein-ligand system.

As the periplasmic component of the choline import system, ChoX binds choline or acetylcholine before transferring it to the transmembrane domain of ChoVWX. Currently, X-ray crystallography has characterized several structures of this protein, all of which are in the closed or semi-closed conformations, whether there is a ligand bound or not (Fig. 1) [27,29]. In comparison to other PBPs, this unique property of ChoX – that it can apparently stay at its closed state without the help of the ligand – made the researchers raise the hypothesis that the binding mechanism of ChoX and its ligand follows the conformational selection model [30].

MSMs are a powerful approach to automatically identify metastable states from short MD simulations and calculate the equilibrium thermodynamic and kinetic properties [31–49]. It divides the protein conformational space into a number of non-overlapping metastable states such that the transitions within each state are fast but transitions across different states are slow. Time is coarse grained (with the smallest unit of  $\tau$ , termed as the lag time) to ensure the model is Markovian so that the probability of transitioning from state  $i$  to  $j$  only depends on  $i$  but not any



**Figure 1. A cartoon representation of the choline-binding protein ChoX in (a) apo-closed (PDB ID: 2RF1) and (b) holo-closed (PDB ID: 2REG) states.** The ligand choline is shown in red spheres.

doi:10.1371/journal.pcbi.1003767.g001

previously visited states. With the help of MSMs, one can extract long time dynamics from short simulations and directly obtain many useful parameters of thermodynamics and kinetics [50–60], which can be further utilized in the flux analysis. For example, Noé *et al.* have performed the first flux analysis based on the transition path theory and MSMs to investigate the major pathways for the folding of the WW domain [55]. More recently, Noé and coworkers have also used MSMs to study the mechanisms of protein-ligand association where protein does not undergo substantial conformational changes [61].

The flux analysis proposed by Hammes *et al.* [20] is an useful tool to calculate and compare flux in both induced fit and conformational selection pathways, allowing one to analyze the contribution of these two limiting mechanisms in a complicated binding event. To conduct the flux-based approach, many kinetic parameters are required – namely, the binding constants and transition rates between different metastable conformational states of the system.

The combination of MSMs and MD simulation allows us to obtain such parameters, which are difficult to be directly measured from experimental assays [20]. In our study, the three-pronged approach of MSMs, MD simulation, and flux analysis was successfully applied in ChoX binding event to quantify both limiting recognition mechanisms.

## Results/Discussion

### Construction and validation of MSMs for apo ChoX

We used MD simulations to investigate the free energy landscape of apo ChoX. In particular, we have generated initial twenty 100-ns simulations, ten from apo-closed (PDB ID: 2RF1) and ten from apo-semiclosed (PDB ID: 2REJ) crystal structures [29]; and one hundred 50-ns additional simulations, seeded from random conformations of the previous twenty trajectories. In total, we collected 7  $\mu$ s of apo simulations and constructed a MSM using the MSMBuilder package [33] (see Methods for the details of model construction). The implied timescale plots flatten at  $\sim 15$  ns, indicating that the model is Markovian with this or longer lag time (SI Fig. S2) [33]. We thus selected 20 ns as the lag time to construct our MSM. Since the macrostate-MSM underestimates the kinetics, we computed all the quantitative properties reported

in this work such as equilibrium state populations and other kinetic properties based on the 500-state microstate-MSM. To better visualize the conformational dynamics of ChoX, we have also lumped the 500 microstates into 5 metastable states, denoted as S1 to S5 with descending populations (see the Methods section for details).

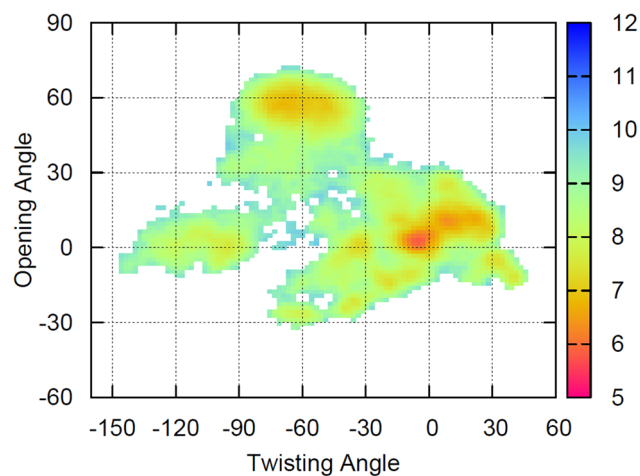
### Free energy landscape of apo ChoX

The projection of the free energy landscape of apo ChoX on the domain-domain opening and twisting angle is plotted in Fig. 2. It is clear that the most populated region is near (0°, 0°), corresponding to an apo-closed crystal structure. Our simulations demonstrate that such a conformation lies in the most dominant metastable state S1 with a population of ~47%. This result is consistent with previous X-ray crystallography research, which noted that ChoX could exist in the closed conformation without the help of a ligand [29]. In addition to this metastable state, we found other states with different degrees of opening or twisting. Such metastable conformations were not discovered by X-ray crystallography, possibly due to the very compact crystal environment and strong contacts between unit cells present in the available apo structures of ChoX (SI Fig. S3) [29].

We also studied kinetics for the transitions between these metastable states. The mean first passage times (MFPTs) were calculated for each pair of states (SI Table S1), and these timescales are on the order of microseconds.

### Structural and dynamic features of metastable states of apo ChoX

The structural features of the five metastable states are displayed in Fig. 3. The most populated state S1 is a closed conformation very similar to those discovered crystal structures. S2 is an open-and-twisted conformation with an essential hydrogen bond between N229 and G232 at the back of the hinge region. S3 is a closed conformation with a small opening at the side of the domain-domain interface, which is large enough to allow the diffusion of the ligand to the binding site (Fig. 3b). S4 is twisted to a very large degree. S5 is another closed structure similar to S1



**Figure 2. Projections of the free energy landscape onto the opening and twisting angles for the apo ChoX.** The free energy profiles were obtained by averaging over contributions from five different metastable states of MSMs weighted by their equilibrium populations. The unit of the color bar is kT. See SI Fig. S9 and Methods for the definition of opening and twisting angles. doi:10.1371/journal.pcbi.1003767.g002

with a different orientation of the helix containing residues 262–275. Further investigations show that hydrogen bonds may play an important role to stabilize these metastable conformations. One example from the metastable state S2 involves N229 and G232. When N229 was mutated to alanine or G232 was mutated to bulkier tyrosine to diminish the hydrogen bonds, fast transitions (within 50-ns) were observed from S2 to S1 (SI Fig. S4) compared to the wild type (~2.07  $\mu$ s), demonstrating the critical contribution of the hydrogen bond (N229-G232) to the metastability of S2.

### Ligands can selectively bind to the closed ChoX conformation to reach the bound states

We performed ten 50-ns simulations for each of the five metastable states by introducing ligands to the system. To increase the chance of observing a binding event within the length scale of MD simulations, we have added 20 ligands to the system (at a concentration of ~0.069M). In each simulation, 20 ligands were randomly placed in the simulation box away from the binding site with the minimum distance of 17 Å and the protein conformations were randomly chosen from each metastable state.

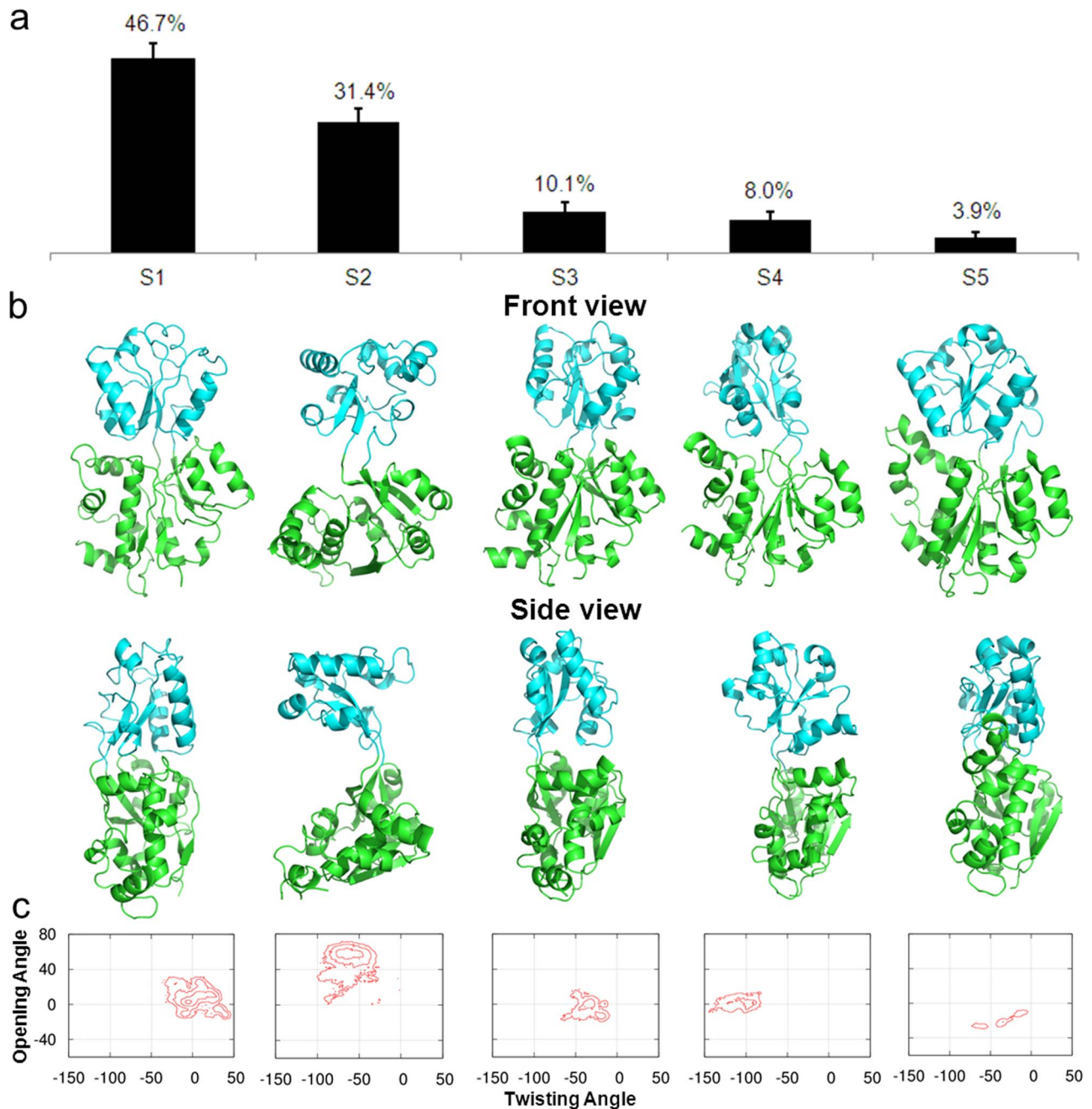
For S1 with closed protein conformations (S1+L), we discovered two out of ten simulations where the ligand recognized the target and bound to it. In order to enhance the sampling, we have performed additional twenty 50-ns MD simulations and three of them were identified with binding events.

The pathways of the ligand binding to S1 can be mainly characterized as conformational selection, and these binding simulations achieved similar conformation compared to the X-ray bound structure with a RMSD as small as 1.6 Å of protein C $\alpha$  atoms which are within 8 Å to the binding site (Fig. 4a). In addition, we examined the distances of the ligand choline to four essential residues at the binding site after the ligand binds to S1, and the values are similar to those from crystal structure. These results indicate that MD simulations have the capability to predict the bound state *in silico*.

### The ligand binding can also induce the conformational changes of the metastable state S3 to reach the bound state

In addition to these ligands that can directly bind to the closed conformation S1, we also found in other simulations that the ligand can interact with the conformation from the state S3 and, at the order of tens of nanoseconds induce the conformational change to the bound conformation S1L. Recall S3 was a closed and twisted state with a side-opening cavity ready for a ligand to insert. We also demonstrated that, from a distance analysis and an overlay of S3L with the holo crystal structure, the ligand stays close to Y119 and W205 (Fig. 5b). One trajectory was discovered with a transition from S3L to S1L, which suggests the possibility that the ligand can bind ChoX through an induced fit mechanism. Since only a single transition event was observed among ten simulations of S3+L, we have performed additional twenty 50-ns MD simulations of S3L complex to enhance the sampling. Among these twenty simulations, two of them displayed the transitions from S3L to S1L (SI Fig. S5).

For the remaining metastable states S2, S4 and S5, no ligands were found to bind to the correct binding site to form the complex: S2L, S4L or S5L (SI Fig. S6). However, in order to examine whether or not these protein-ligand complexes (S2L, S4L or S5L) if exist can induce the transitions to the bound state S1L, we have modeled these complexes using the AutoDock Vina [62], and initiate ten 50-ns MD simulations from each of these docked conformations. As shown in SI Fig. S7, none of these simulations contained any transitions to the bound state (S1L). These results



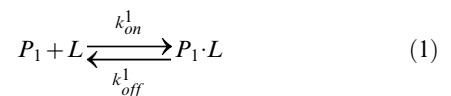
**Figure 3. (a) The equilibrium populations of five metastable states obtained from the MSM for apo ChoX. (b) Representative conformations from these five states. The two domains of the protein are colored in cyan and green respectively, and two viewpoints, front and side views, are shown. (c) Projections of free energy landscape on the protein opening and twisting angle for each metastable conformational state. The interval between two adjacent contour levels is 1 kT. doi:10.1371/journal.pcbi.1003767.g003**

indicate that the direct transitions from S2L, S3L and S5L to the bound state S1L are unlikely to occur.

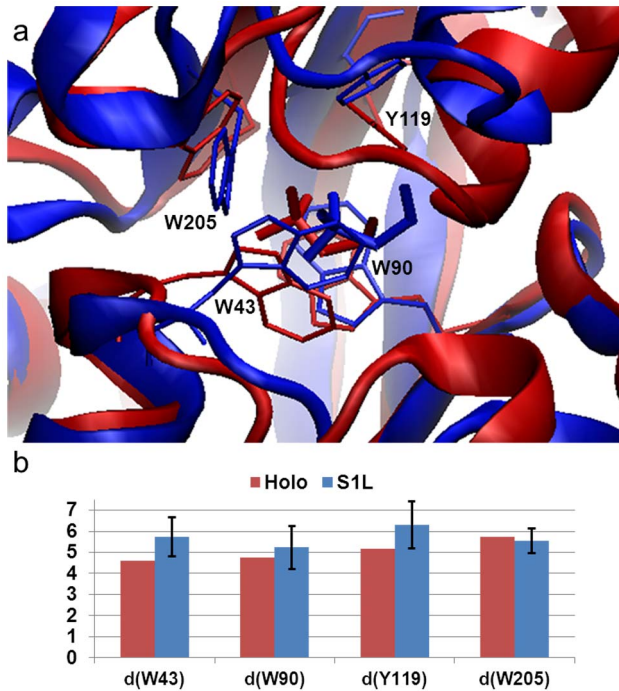
### Quantifying the binding flux

From the MD simulations discussed above, we can obtain a rough picture of the hybrid mechanism of conformational selection and induced fit for ChoX. However, the great challenge is to quantify the percentage of each mechanism in complicated scenarios. In order to achieve this, we have applied the flux analysis theory [20], where the flux going through each pathway is

utilized to quantify binding mechanisms. At first, the conformational selection pathway can be described as (SI Fig. S1):



where  $P_1$  represents the closed conformation S1 of ChoX, and other



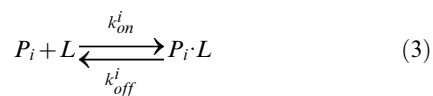
**Figure 4. Choline can selectively bind to the closed protein conformation S1 to reach the holo state.** (a) The X-ray bound conformation (red, PDB ID: 2REG) is superimposed with a conformation selected from MD simulations in the presence of ligands and with smallest RMSD to the bound state (blue). Four critical residues in the binding sites that are in direct contact with the ligand are highlighted in stick representation. (b) Distances (Å) between the center of mass of ligand and four centers of mass of critical residues in the binding site obtained from the holo structure (red bars) and five MD trajectories in which only those conformations after the ligand binding have been included in the analysis (blue bars). doi:10.1371/journal.pcbi.1003767.g004

metastable conformations  $P_i$  ( $i = 2-5$ ) can interconvert with  $P_1$ :

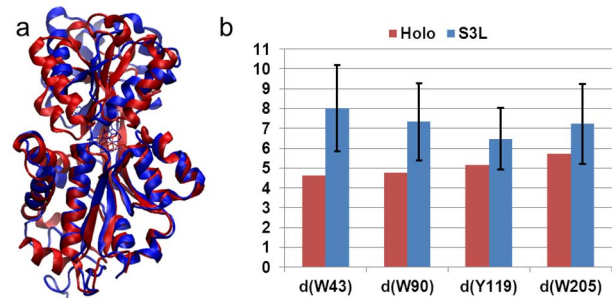
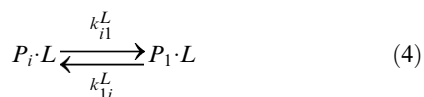


In this work, we consider the conformational selection mechanism in a general context where the ligand selects to bind a certain metastable protein conformation including the ground state ( $P_1$  in this case).

On the other hand, the induced fit pathway can be described by a two-step process (SI Fig. S1), where the ligand first binds to any metastable conformation  $P_i$  other than the closed conformation  $P_1$ :



And the binding will further induce the conformational change to reach the bound state:



**Figure 5. MD simulations initiated from the state S3 that exhibit spontaneous ligand binding.** (a) A representative conformation of S3L (blue) is overlaid with the X-ray bound state (red, PDB ID: 2REG). (b) The same as Fig. 4b except that the distances computed from 7 50-ns ligand binding MD simulations that the ligand binds state S3. doi:10.1371/journal.pcbi.1003767.g005

The flux flowing through each of the above two pathways can be derived from the flux analysis theory as the following (see Methods for the details of the derivation):

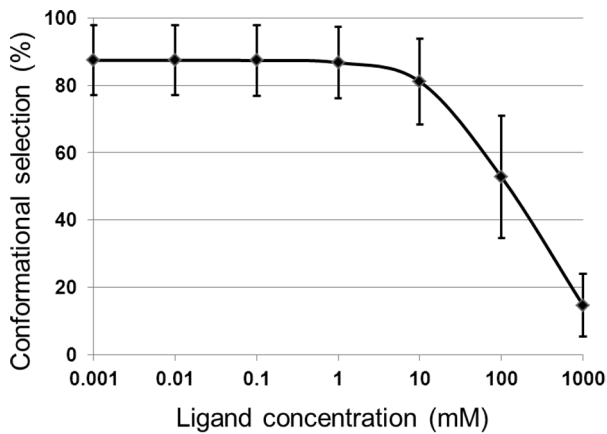
$$F_{CS} = \left( \frac{1}{\sum_{i=2}^5 k_{i1}[P_i]} + \frac{1}{k_{on}^1[P_1][L]_f} \right)^{-1} \quad (5)$$

$$F_{IF} = \sum_{i=2}^5 \left( \frac{1}{k_{on}^i[P_i][L]_f} + \frac{1}{k_{i1}^L[P_i \cdot L]} \right)^{-1} \quad (6)$$

where  $F_{CS}$  and  $F_{IF}$  represent the flux through conformational selection and induced fit pathway respectively.  $k_{on}^i$  and  $k_{off}^i$  are the kinetic rate constant for ligand binding/unbinding to state  $i$  respectively;  $k_{i1}$  is the rate constant for the transition from state  $i$  to state 1;  $k_{i1}^L$  is the rate constant for the transition from the complex  $S_i \cdot L$  to the bound state  $S_1 \cdot L$ ; and  $[L]_f$  is the free ligand concentration.

In this study, we have derived important parameters from MSMs and MD simulation that are missing in the flux analysis. Specifically,  $k_{i1}$  can be derived from the transition probability matrix. For MD simulations starting from state S2, S4 and S5, we didn't observe any successful binding events, therefore  $k_{on}^2$ ,  $k_{on}^4$ ,  $k_{on}^5$  are all set to be zero. For S3, there exist multiple binding and unbinding events in our ten 50-ns MD simulations. Therefore, we have obtained  $k_{on}^3$  and  $k_{off}^3$  by computing the fraction remaining in the unbound state S3 as a function of time followed by fitting to Eq. 19 (SI Fig. S8).  $k_{on}^1$  can be derived from MD simulations with ligands in a similar way to  $k_{on}^3$  (Eq. 20). At last, we need to derive the values for  $k_{i1}^L$  ( $i = 2-5$ ). Since the simulations of S2L, S4L and S5L didn't show any transitions from each state to S1L (SI Fig. S7), these values ( $k_{21}^L$ ,  $k_{41}^L$ ,  $k_{51}^L$ ) are all set to be zero. For  $k_{31}^L$ , we can obtain its value from twenty S3L MD simulations since the ligand binding further induced the conformational changes to the bound state in a fraction of these trajectories (Eq. 24).

We can then proceed to measure the percentage contribution of conformational selection and induced fit mechanism using equations (5) and (6). At the protein concentration fixed to 1  $\mu$ M, Fig. 6 shows the contribution of conformational selection to the binding pathway depending on the ligand concentration. Conformational selection is dominant for a wide range of the ligand concentration, accounting for around 90% of the binding



**Figure 6. The percentage of conformational selection mechanism as a function of ligand concentration.** Conformational selection is dominant for most of the ligand concentration range for ChoX. doi:10.1371/journal.pcbi.1003767.g006

event at the concentration of choline in the lab conditions ( $\mu\text{M}$  scale) [63]. Therefore, we conclude that the binding mechanism of choline to ChoX is dominated by conformational selection.

## Conclusions

In summary, we here propose a novel method that combines MSMs, MD simulations and flux analysis to quantify the binding mechanism of conformational selection and induced fit in complex binding events. In the case study of choline binding to ChoX, we were able to derive all the necessary parameters using MD simulations and MSMs. Based on these parameters, the percentage of each limiting binding mechanism could be quantitatively calculated as a function of ligand concentration. It would be difficult, using common experiments, to obtain these necessary parameters to elucidate these mechanisms. Finally, once the mechanism is quantified, one can further apply other techniques (e.g. *in silico* design) to the biological system to fine tune the binding event either to increase the degree of conformational selection or induced fit, so new properties of macromolecules could be explored and created to accommodate the needs of protein engineering and beyond.

## Methods

### Definition of the opening and twisting angle

We project the free energy landscape of ChoX onto two dihedral angles: the opening and twisting angles. The opening angle is defined as the angle between the normal vectors of the two planes formed by the center of mass of the following groups of C $\alpha$  atoms (SI Fig. S9a):

Plane A: residues 31–114 & 234–316; 185–194; 159–166.

Plane B: residues 118–230; 185–194; 159–166.

The twisting angle planes are:

Plane C: residues 31–114 & 234–316; 185–194; 46–55.

Plane D: residues 118–230; 185–194; 46–55 [64].

From Principal Component Analysis (PCA) of the apo MD simulations, we found strong correlations between the opening angle and the first eigenvector ( $R^2 = 0.77$ ), as well as between the twisting angle and the second eigenvector ( $R^2 = 0.53$ ) (SI Fig. S9b). In this work, the degrees of apo-closed and holo-closed crystal structures were shifted to the ( $0^\circ$ ,  $0^\circ$ ) point in corresponding settings.

## MD simulations of the apo ChoX protein

The GROMACS 4.5.4 [65] software and Amber99sb force field [66] were used for all the MD simulations. The procedure is as follows: the protein (and ligand, when present) was solvated in a dodecahedron box with 14, 450 SPC water molecules [67] and enough counter ions to neutralize the system. The system was first minimized with a steepest descent algorithm, followed by a 200 ps MD simulation with position restraints for the heavy atoms. All the simulations were performed at NVT ensemble with 300K of temperature using V-rescale thermostat [68]. The cut-offs for both VDW and short-range electrostatic interactions were set to 10 Å and long-range electrostatic interactions were treated with the Particle-Mesh Ewald method [69]. The time-step was 2 fs and the neighbour list was updated every 10 steps. Water molecules were constrained by the SETTLE algorithm [70] and all protein bonds were constrained by the LINCS algorithm [71].

## Markov state model construction

Using the MSMBuilder package [33], we first applied the k-centre algorithm to cluster all the conformations into 500 microstates based on the C $\alpha$  atoms of residues in proximity to the binding site (i.e. within 8 Å of the ligand as defined by the holo crystal structure, PDB ID: 2REG). The generated microstates were small, with the average RMSD values to its central conformation in each state of about 1.9 Å. The transition probability matrix ( $\mathbf{T}$ ) was obtained by counting the number of transitions observed in the MD trajectories. We then examined the transition probability matrix, and removed two disconnected microstates from our model. The implied timescales ( $\tau_k$ ) obtained from the transition probability matrix  $\mathbf{T}$  indicates the aggregated timescales for transitions between groups of microstates.

$$\tau_k = -\frac{\tau}{\ln \mu_k(\tau)} \quad (7)$$

where  $\mu_k$  is an eigenvalue of the transition matrix with the lag time  $\tau$ .

We have examined the implied timescale plots for this 500-microstate model to select a lag time that ensures the model to be Markovian. As shown in SI Fig. S2, the implied timescale curves plateau at around 15 ns, indicating the model is Markovian with this or longer lag times. Thus we chose a lag time of 20 ns for our final MSM. In order to better visualize the conformational dynamics of apo ChoX, we have further lumped all the microstates into 5 metastable macrostates using the Robust Perron Cluster Analysis (PCCA+) algorithm [72].

The 500-state microstate-MSM was used to compute all the quantitative properties we report in this work. To obtain the populations of metastable states ( $P_1, \dots, P_5$ ) from the 500-state microstate-MSM, we simply sum over the equilibrium populations of all the microstates that belong to a certain metastable state:  $P_i = \sum_{k \in i} p_k$ . To compute the MFPT between a pair of metastable state  $i$  and  $j$  from the microstate-MSM, we first set MFPTs of all the microstates that belong to the destination metastable state  $j$  to be zero:  $t_k, \text{ where } k \in j = 0$ . We then computed MFPTs starting from any of the microstate that belong to the metastable state  $i$ :  $t_l, \text{ where } l \in i$ . Finally, we obtained the MFPTs from  $i$  to  $j$  by performing a weighted average over all the microstates that belong to  $i$ :  $MFPT_{ij} = \sum_{l \in i} p_l t_l, \text{ where } l \in i$ .

### Mean first passage time calculation

The MFPT, determined by the following formula, calculates the average transition times between a pair of states [73].

$$MFPT_{if} = \sum_j P_{ij}(\tau + MFPT_{jf}) \quad (8)$$

where  $P_{ij}$  is the transition probability from state  $i$  to state  $j$ ,  $\tau$  is the lag time of the transition probability matrix  $T$ , and  $MFPT_{jf}$  is the mean first passage time of the state  $j$  to final state  $f$ . The boundary condition is  $MFPT_{ff}=0$ .

### Deriving force field parameters for choline

In order to simulate the process of choline binding to the ChoX protein, we need to obtain the force field parameters for choline. We followed the same procedure as we previously published [74] to derive both bonded and non-bonded force field parameters of choline. Specifically, we have obtained the stretching, bending and torsion parameters by fitting against the Quantum Mechanics (QM) calculations performed using the Density Functional Theory with B3LYP/6-31G\* in the Gaussian software [75]. Similar with previous studies [76], we have employed the Restrained Electro-Static Potential (RESP) method to derive the partial charges from the QM calculations with HF/6-31G\*. We have listed all the force field parameters in the format that is compatible with the GROMACS 4.5 software package in SI Text S1, S2.

### Deriving parameters for flux analysis

The two limiting ligand-binding mechanisms: conformational selection and induced fit can be described by Eq. (1)–(4). Following the flux analysis theory developed by Hammes *et al.* [20], we can then derive the fractional flux passing through each pathway. We note that this flux analysis can be considered as a special case of transition path theory and yields consistent path fluxes for serial and parallel pathways [55,77]. In particular, if one pathway is consisted of parallel reaction paths, its flux can be written as:

$$F = \sum F_i \quad (9)$$

If one pathway contains serial segments, the flux is:

$$F = \frac{1}{\sum \frac{1}{F_i}} \quad (10)$$

Now, let's consider the conformational selection pathway, which involves two steps. First, different protein metastable conformations ( $S_i$ ,  $i=2-5$ ) can all interconvert to the closed state ( $S_1$ ) at rates  $k_{i1}$  (Eq. 2). Therefore, the flux for this segment is:  $F_1 = \sum_{i=2}^5 k_{i1}[P_i]$ , where  $[P_i]$  is the concentration of the state  $S_i$ . Next, the ligand can selectively bind to the closed protein conformation  $S_1$ , to reach the bound state. For this part, the flux is:  $F_2 = k_{on}^1 [P_1][L]_f$  (Eq. 1). Therefore, the flux of conformational selection  $F_{CS}$  can be derived as:

$$F_{CS} = \left( \frac{1}{F_1} + \frac{1}{F_2} \right)^{-1} \quad (11)$$

The detailed expression of  $F_{CS}$  can be found in Eq. 5.

Following the similar procedure [20], we can also derive the flux for the induced fit pathway. In particular, there exist independent parallel pathways to reach the bound state, where the ligand can first bind to a certain metastable conformational state ( $S_i$ ,  $i=2-5$ ),

and further induced the conformational change to the bound state. The flux of each of these pathways ( $F_i$ ) can be written as:

$$F_i = \left( \frac{1}{k_{on}^i [P_i][L]_f} + \frac{1}{k_{i1}^L [P_i \cdot L]} \right)^{-1} \quad (12)$$

where  $k_{on}^i$  and  $k_{i1}^L$  denote to the rate for the ligand binding to state  $S_i$  ( $i=2-5$ ) and the transition rate from the complex  $S_i \cdot L$  to the bound state  $S_1 \cdot L$  respectively;  $[L]_f \approx [L]$  when  $[L]$  exceeds. When  $[L]$  and  $[P]$  are comparable,  $[L]_f$  is given by Eq. 13:

$$[L]_f = \frac{[L]}{1 + \sum_{i=1}^5 \frac{[P_i]}{K_d^i}} \quad (13)$$

The overall flux for the induced fit pathway ( $F_{IF}$ ) can then be written as:

$$F_{IF} = \sum_{i=2}^5 F_i \quad (14)$$

The fractional flux for conformational selection pathway is:

$$F_{CS}\% = \frac{F_{CS}}{F_{CS} + F_{IF}} \quad (15)$$

In order to obtain  $F_{CS}\%$ , we need to derive a series of parameters:  $k_{i1}$  ( $i=2-5$ ),  $k_{on}^i$  ( $i=1-5$ ),  $k_{i1}^L$  ( $i=2-5$ ).  $k_{i1}$  are the transition rate constants from state  $i$  to state 1, which can simply be derived from MFPT [58]:

$$k_{i1} = 1/MFPT_{i1} \quad (16)$$

where  $MFPT_{i1}$  is the mean first passage time of state  $i$  to final state 1. The uncertainties of this set of rates are obtained from bootstrapping the MD dataset (containing  $N=138$  trajectories) with replacement for  $N$  times.

For  $k_{on}^i$ , we use our MD simulations with ligand to derive their values. Based on Eq. (1) and (3), we can write the rate equation as:

$$\frac{d[P_i]}{dt} = -k_{on}^i [P_i][L] + k_{off}^i [P \cdot L] \quad (17)$$

In our simulations, the initial ligand concentration  $[L]_0$  is twenty times larger than the initial protein concentration  $[P]_0$ :  $[P]_0 \ll [L]_0$ , therefore,  $[L] \approx [L]_0$ . The forward reaction, which is only dependent on  $k_{on}$ , can then be written as:

$$\frac{d[P_i]}{dt} = -k_{on}^i [P_i][L]_0 + k_{off}^i ([P]_0 - [P_i]) \quad (18)$$

We can solve Eq. 18 to obtain  $[P_i]$ :

$$\frac{[P_i]}{[P_i]_0} = 1 - \frac{k_{on}^i [L]_0}{k_{on}^i [L]_0 + k_{off}^i} (1 - e^{-(k_{on}^i [L]_0 + k_{off}^i)t}) \quad (19)$$

We can then examine our ligand MD simulations initiated from different protein metastable conformations to obtain  $k_{on}^i$  and  $k_{off}^i$ . For MD simulations starting from state  $S_2$ ,  $S_4$  and  $S_5$ , we didn't observe any binding events (for distances between the center

of mass of ligand and center of mass of four critical residues, W43, W90, Y119, W205 [29], in the binding site to be less than 12 Å), therefore  $k_{on}^2$ ,  $k_{on}^4$ ,  $k_{on}^5$  all equal zero (however we estimated their upper limits in Table 1). For S3, there exist multiple binding and unbinding events in our ten 50-ns MD simulations. We have thus computed the fraction remaining in the unbound state ( $\frac{[P_i]}{[P_1]_0}$ ) as a function of time, and further fit Eq. 19 to it in order to obtain  $k_{on}^3 = 4.19 \pm 2.06 \times 10^8 \text{ M}^{-1} \text{ s}^{-1}$  and  $k_{off}^3 = 4.50 \pm 3.09 \times 10^7 \text{ s}^{-1}$  (see SI Fig. S8 for the parameter fitting). We can then obtain  $K_d^3 = 0.107\text{M}$ . Thus,  $[P_3 \cdot L]$  can be calculated by  $P_3 \times [L]_f / K_d^3$ . For S1, we have observed multiple binding events; however none of the unbinding events from our MD simulations, due to the high stability of the bound state. We can then simplify Eq. 18 by only considering the forward reaction and  $[P_1]/[P_1]_0$  can be written as:

$$\frac{[P_1]}{[P_1]_0} = e^{-k_{on}^1 L t} \quad (20)$$

We then obtain  $k_{on}^1 = 6.33 \pm 2.53 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$ .

To further examine the robustness of the definition of the successful binding events, we have compared it with a more specific definition: the heavy atoms of the ligand form contact with atoms belonging to at least three critical residues (among W43, W90, Y119, W205 [29]) in the binding pocket. These two sets of rates only differ slightly (SI Table S2). Furthermore, we have compared the fraction of conformational selection computed based on two different definitions of ligand binding. As shown in SI Fig. S10, the results from two different definitions are also consistent.

In order to examine whether or not the protein undergoes any major conformational changes while collecting data for estimating the binding rates (e.g.  $k_{on}^1$ ,  $k_{on}^3$ , and  $k_{off}^3$ ), we have projected protein conformations in each binding MD simulation onto a pair of reaction coordinates (opening and twisting angles). As shown in SI Fig. S11, the protein remains in its initial metastable state during the whole course of all the binding MD simulations. These results confirm that our estimations of binding rates are clean.

In order to obtain  $K_d^1$ , which is the dissociation constant for the ligand directly binding to S<sub>1</sub>, we have constructed a thermodynamic cycle to compute its value (SI Fig. S12):

$$-\Delta G_d = \Delta G_p + \Delta G_1 \quad (21)$$

where the free energy  $\Delta G_1$  is directly related to  $K_d^1$ :  $\Delta G_1 = kT \ln K_d^1$ .

$\Delta G_d$  can be computed from the disassociation constant for the ligand binding  $K_d$  by  $\Delta G_d = -kT \ln K_d$ , where  $K_d$  has been obtained experimentally as 2.7  $\mu\text{M}$  [63].  $\Delta G_p$  is the free energy difference associated with the conformational transitions from different metastable states to the closed state S<sub>1</sub>:

$$\Delta G_p = \sum_{i=2}^5 \Delta G_{i1} \times f_i \quad (22)$$

where  $\Delta G_{i1}$  is the free energy difference between S<sub>i</sub> and S<sub>1</sub>, and  $f_i$  denotes to the equilibrium population of S<sub>i</sub>:

$$\Delta G_p = \sum_{i=2}^5 -kT \ln(f_i/f_1) \times f_i \quad (23)$$

From Eq. 21–23, we can compute the value of  $K_d^1$  as 4.53  $\mu\text{M}$ .

At last, we need to derive the values for  $k_{i1}^L$  ( $i=2-5$ ). As discussed before, there are no binding/unbinding events for S<sub>2</sub>, S<sub>4</sub> and S<sub>5</sub>, and these values ( $k_{21}^L$ ,  $k_{41}^L$ ,  $k_{51}^L$ ) are all set to be zero. For  $k_{31}^L$ , we can obtain the value from the MD simulations initiated from the S<sub>3</sub> state with ligand.

$$\frac{[P_3 \cdot L]}{[P_3 \cdot L]_0} = e^{-k_{31}^L t} \quad (24)$$

Eq. 24, helps us to calculate the forward transition rate  $k_{31}^L = 2.60 \pm 1.60 \times 10^6 \text{ s}^{-1}$ .

For the ligand binding/unbinding ( $k_{on}^1$ ,  $k_{on}^3$  and  $k_{off}^3$ ) and transition from S<sub>3</sub>L to S<sub>1</sub>L ( $k_{31}^L$ ). We have performed the bootstrapping analysis to obtain the error bars. In particular, we bootstrapped the MD dataset (containing N MD simulation trajectories with N = 30, 9 and 20 for  $k_{on}^1$ ,  $k_{on}^3/k_{off}^3$ , and  $k_{31}^L$  respectively) with replacement for N times.

For rates that are estimated to be zero with no observed transitions (e.g.  $k_{on}^2$ ,  $k_{21}^L$ ), we have estimated the upper limit of these rates by assuming one binding/transition event occurs

**Table 1.** Parameters for the flux analysis.

Parameter	Value	Parameter	Value
P <sub>1</sub>	0.467 ± 0.038 $\mu\text{M}$	$k_{on}^1$	6.33 ± 2.53 × 10 <sup>7</sup> M <sup>-1</sup> s <sup>-1</sup>
P <sub>2</sub>	0.314 ± 0.034 $\mu\text{M}$	$k_{on}^2$	< 3.06 × 10 <sup>7</sup> M <sup>-1</sup> s <sup>-1</sup>
P <sub>3</sub>	0.101 ± 0.022 $\mu\text{M}$	$k_{on}^3$	4.19 ± 2.06 × 10 <sup>8</sup> M <sup>-1</sup> s <sup>-1</sup>
P <sub>4</sub>	0.080 ± 0.019 $\mu\text{M}$	$k_{on}^4$	< 3.06 × 10 <sup>7</sup> M <sup>-1</sup> s <sup>-1</sup>
P <sub>5</sub>	0.039 ± 0.013 $\mu\text{M}$	$k_{on}^5$	< 3.06 × 10 <sup>7</sup> M <sup>-1</sup> s <sup>-1</sup>
k <sub>21</sub>	5.78 ± 2.29 × 10 <sup>5</sup> s <sup>-1</sup>	$k_{21}^L$	< 2.11 × 10 <sup>6</sup> s <sup>-1</sup>
k <sub>31</sub>	7.19 ± 5.40 × 10 <sup>5</sup> s <sup>-1</sup>	$k_{31}^L$	2.60 ± 1.60 × 10 <sup>6</sup> s <sup>-1</sup>
k <sub>41</sub>	1.81 ± 1.00 × 10 <sup>5</sup> s <sup>-1</sup>	$k_{41}^L$	< 2.11 × 10 <sup>6</sup> s <sup>-1</sup>
k <sub>51</sub>	2.97 ± 3.09 × 10 <sup>6</sup> s <sup>-1</sup>	$k_{51}^L$	< 2.11 × 10 <sup>6</sup> s <sup>-1</sup>
		$k_{off}^3$	4.50 ± 3.09 × 10 <sup>7</sup> s <sup>-1</sup>

doi:10.1371/journal.pcbi.1003767.t001



during our accumulated 500-ns simulation time. We also show that applying these upper limits of the rates to our flux analysis does not change qualitatively the conclusion of this paper: the fraction of the conformation selection slightly decreases from 93% to 82% at experimental concentrations ( $\sim 1 \mu\text{M}$ ) as shown in SI Fig. S13.

We have listed all the necessary rate constants with uncertainties/upper limits in Table 1.

## Supporting Information

**Figure S1** A schematic diagram representing the conformational selection and induced fit models of protein-ligand recognition. (TIF)

**Figure S2** Implied timescales of microstate and macrostate models. (a). Fifteen slowest implied timescales as a function of lag time computed from the 500-state microstate-MSM. (b). Implied timescales as a function of lag time computed from the 5-state macrostate-MSM. (TIF)

**Figure S3** Crystal contacts experienced by the ChoX in the (a) apo-closed (PDB ID: 2RF1) and (b) apo-semiclosed (PDB ID: 2REJ) X-ray structures. The central unit cell contains two ChoX molecules (in ribbon representations). The surrounding unit cells are shown in surface representations in orange. (TIF)

**Figure S4** MD simulations of two single mutants of apo ChoX, Asn229Ala and Gly232Tyr, that exhibit an accelerated conformational change from the open to the closed state. We have performed three 50-ns MD simulations for each mutant. One of these MD simulations for each mutant is projected onto the opening and twisting angles with a step size of 5-ns. The projections of the apo ChoX free energy landscape (the same as Fig. 3c, and each macrostate is assigned a different color) are also displayed in the same figure. (TIF)

**Figure S5** Protein conformational changes are displayed for three MD simulations where transitions from S3L to S1L occur. The projections of the free energy landscape onto the opening and twisting angles are shown for state S3 (blue) and S1 (red). Each arrow corresponds to a 10-ns segment of the MD simulation. The black cross represents the holo crystal structure. The middle and right panels correspond to the two additional MD simulations containing the transitions from S3L to S1L. (TIF)

**Figure S6** Superimposition of representative snapshots from macrostates S2 (a), S4 (b) and S5 (c) in blue with the X-ray structure of the ChoX bound state (red, PDB ID: 2REG). Each protein conformation is displayed in both front and side views. We did not observe any stable binding events in our MD simulations. A binding event is defined as when distances between the center of mass of the ligand and center of mass of four critical residues (W43, W90, Y119 and W205) in the binding site all to be less than 12 Å. Therefore we consider that the ligands do not bind to these metastable states (S2, S4 and S5). (TIF)

**Figure S7** The projections of protein conformational change on the opening and twisting angles during the course of MD simulations are displayed in black arrowed lines (two trajectories in one panel). The simulations initiated from S2L, S4L, and S5L are plotted in (a), (b), and (c) respectively. The projections of the apo protein free energy landscape are also displayed as the

background. Each arrow corresponds to a 10-ns segment of the MD simulation. The black cross corresponds to the holo crystal structure.

(TIF)

**Figure S8** Fraction of the unbound state as a function of time for the ligand binding to metastable state S3. Eq. 19 is fitted (solid line) to data obtained from MD simulations (points) to derive the kinetic parameters:  $k_{on}^3$  and  $k_{off}^3$ . (TIF)

**Figure S9** Definition of the twisting and opening dihedral angle. (a) The opening (left panel, side view) and twisting (right panel, top view) angles are defined as angles between pairs of planes. (b) The opening and twisting angles have a good correlation with the top two eigenvectors obtained from the Principal Component Analysis. The correlation coefficients  $R^2$  are 0.77 and 0.53 between the first eigenvector and the opening angle, and between second eigenvector and the twisting angle, respectively. The protein conformations from all the apo ChoX MD simulations are included in this analysis. (TIF)

**Figure S10** Fractions of conformational selection computed by using two sets of kinetic rates obtained by different definitions of successfully binding events. In the first definition (blue), the distances between the center of mass (c.o.m) of the ligand and the c.o.m of the side-chains of four critical residues in the binding pockets all have to be smaller than 12 Å. In the second definition (red), heavy atoms of the ligand form contact with atoms belonging to at least 3 critical residues in the binding pocket. (TIF)

**Figure S11** Protein conformational changes during the MD simulations in the presence of the ligand (a) S1+L. In particular, the ligand binding occurs in the first 5 panels. Only 9 out of 30 MD simulations of S1+L are displayed. (b) S3+L. The projections of the apo ChoX free energy landscape The projections of the free energy landscape onto the opening and twisting angles are shown for state S1 (Red), S2 (Green) and S3 (Cyan) as background. Each arrow corresponds to a 10-ns segment of the MD simulation. The black cross corresponds to the holo crystal structure. (TIF)

**Figure S12** The thermodynamic cycle used for  $K_d^1$  calculation. The ligand dissociation constant  $K_d$  is measured from experiments.  $\Delta G_d$  can also be obtained from a two-step process: protein conformational transition and ligand binding to state S1. Therefore, we can construct a thermodynamic cycle to obtain the value of  $\Delta G_I$ . (TIF)

**Figure S13** Fractions of conformational selection as a function of ligand concentration obtained from the flux analysis with original rates and with the upper limit of certain rates (see Table 1) are shown in blue and red respectively. (TIF)

**Table S1** Mean first passage times between pairs of macrostates. (PDF)

**Table S2** Rates computed by different definitions of successful ligand binding events. In the first definition, the distances between the center of mass (c.o.m) of the ligand and the c.o.m of the side-chains of four critical residues in the binding pockets all have to be smaller than 12 Å. In the second definition, heavy atoms of the

ligand form contact with atoms belonging to at least 3 critical residues in the binding pocket.

(PDF)

**Text S1** Force field parameters for choline (ffbonded\_choline.itp).

(PDF)

**Text S2** Force field parameters for choline (choline.rtp).

(PDF)

## References

- Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology* 5: 789–796.
- Koshland DE (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* 44: 98–104.
- Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R (2000) Folding and binding cascades: dynamic landscapes and population shifts. *Protein science : a publication of the Protein Society* 9: 10–19.
- Ma B, Kumar S, Tsai CJ, Nussinov R (1999) Folding funnels and binding mechanisms. *Protein engineering* 12: 713–720.
- Ma BY, Shatsky M, Wolfson HJ, Nussinov R (2002) Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Science* 11: 184–197.
- Arora K, Brooks CL (2007) Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proceedings of the National Academy of Sciences of the United States of America* 104: 18496–18501.
- Bahar I, Chennubhotla C, Tobi D (2007) Intrinsic dynamics of enzymes in the unbound state and, relation to allosteric regulation. *Curr Opin Struct Biol* 17: 633–640.
- Bui JM, McCammon JA (2006) Protein complex formation by acetylcholinesterase and the neurotoxin fasciculin-2 appears to involve an induced-fit mechanism. *Proceedings of the National Academy of Sciences of the United States of America* 103: 15451–15456.
- Tsai CJ, Kumar S, Ma BY, Nussinov R (1999) Folding funnels, binding funnels, and protein function. *Protein Science* 8: 1181–1190.
- Tsai CJ, Ma BY, Nussinov R (1999) Folding and binding cascades: Shifts in energy landscapes. *Proceedings of the National Academy of Sciences of the United States of America* 96: 9970–9972.
- Zhou HX (2010) From Induced Fit to Conformational Selection: A Continuum of Binding Mechanism Controlled by the Timescale of Conformational Transitions. *Biophys J* 98: L15–L17.
- Okazaki KI, Takada S (2008) Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. *Proceedings of the National Academy of Sciences of the United States of America* 105: 11182–11187.
- Wlodarski T, Zagrovic B (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences of the United States of America* 106: 19346–19351.
- Formanek MS, Ma L, Cui Q (2006) Reconciling the “old” and “new” views of protein allostery: A molecular simulation study of chemotaxis Y protein (CheY). *Proteins* 63: 846–867.
- Bakan A, Bahar I (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proceedings of the National Academy of Sciences of the United States of America* 106: 14349–14354.
- Bucher D, Grant BJ, McCammon JA (2011) Induced Fit or Conformational Selection? The Role of the Semi-closed State in the Maltose Binding Protein. *Biochemistry-U S A* 50: 10530–10539.
- Greives N, Zhou HX (2014) Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proceedings of the National Academy of Sciences of the United States of America* 10.1073/pnas.1407545111.
- Ross CL, Liang XW, Liu Q, Murray BE, Hook M, et al. (2012) Targeted Protein Engineering Provides Insights into Binding Mechanism and Affinities of Bacterial Collagen Adhesins. *J Biol Chem* 287: 34856–34865.
- Dwyer MA, Hellinga HW (2004) Periplasmic binding proteins: a versatile superfamily for protein engineering. *Curr Opin Struct Biol* 14: 495–504.
- Hammes GG, Chang YC, Oas TG (2009) Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America* 106: 13737–13741.
- Clore GM, Iwahara J (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chemical reviews* 109: 4108–4139.
- Prestegard JH, Bougault CM, Kishore AI (2004) Residual dipolar couplings in structure determination of biomolecules. *Chemical reviews* 104: 3519–3540.
- Ortega G, Castano D, Diercks T, Millet O (2012) Carbohydrate Affinity for the Glucose-Galactose Binding Protein Is Regulated by Allosteric Domain Motions. *J Am Chem Soc* 134: 19869–19876.
- Tang C, Schwieters CD, Clore GM (2007) Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature* 449: 1078–U1012.
- Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE (2009) Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 19: 120–127.
- Silva DA, Bowman GR, Sosa-Peinado A, Huang XH (2011) A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. *Plos Comput Biol* 7: e1002054.
- Oswald C, Smits SHJ, Hoing M, Bremer E, Schmitt L (2009) Structural analysis of the choline-binding protein ChoX in a semi-closed and ligand-free conformation. *Biol Chem* 390: 1163–1170.
- Fukami-Kobayashi K, Tateno Y, Nishikawa K (1999) Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. *J Mol Biol* 286: 279–290.
- Oswald C, Smits SH, Hoing M, Sohn-Bosser L, Dupont L, et al. (2008) Crystal structures of the choline/acetylcholine substrate-binding protein ChoX from *Sinorhizobium meliloti* in the liganded and unliganded-closed states. *J Biol Chem* 283: 32848–32859.
- Bermejo GA, Strub MP, Ho C, Tjandra N (2010) Ligand-Free Open-Closed Transitions of Periplasmic Binding Proteins: The Case of Glutamine-Binding Protein. *Biochemistry-U S A* 49: 1893–1902.
- Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 126: 155101.
- Noe F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* 18: 154–162.
- Bowman GR, Huang XH, Pande VS (2009) Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* 49: 197–201.
- Buchete NV, Hummer G (2008) Coarse master equations for peptide folding dynamics. *J Phys Chem B* 112: 6057–6069.
- Zheng W, Andreu M, Gallicchio E, Levy RM (2007) Simulating replica exchange simulations of protein folding with a kinetic network model. *Proc Natl Acad Sci U S A* 104: 15340–15345.
- Yao Y, Cui RZ, Bowman GR, Silva DA, Sun J, et al. (2013) Hierarchical Nystrom methods for constructing Markov state models for conformational dynamics. *J Chem Phys* 138: 124101.
- Huang X, Bowman GR, Bacallado S, Pande VS (2009) Rapid equilibrium sampling initiated from nonequilibrium data. *Proc Natl Acad Sci U S A* 106: 19765–19769.
- Huang X, Yao Y, Bowman GR, Sun J, Guibas LJ, et al. (2010) Constructing multi-resolution Markov State Models (MSMs) to elucidate RNA hairpin folding mechanisms. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*: 228–239.
- Bowman GR, Meng L, Huang X (2013) Quantitative comparison of alternative methods for coarse-graining biological networks. *J Chem Phys* 139: 121905.
- Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J Am Chem Soc* 132: 1526–1528.
- Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, Noe F (2013) Identification of slow molecular order parameters for Markov model construction. *J Chem Phys* 139: 015102.
- Noe F, Horenko I, Schutte C, Smith JC (2007) Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys* 126: 155102.
- Schutte C, Fischer A, Huisinga W, Deuffhard P (1999) A direct approach to conformational dynamics based on hybrid Monte Carlo. *J Comput Phys* 151: 146–168.
- Swope WC, Pitera JW, Suits F (2004) Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J Phys Chem B* 108: 6571–6581.

## Acknowledgments

SG acknowledges Dr. Lu Zhang, Qin Qiao and Hanlun Jiang for helpful discussions. DAS acknowledges Dr. Gabriel Rocklin for helpful discussions. Computing resources were provided by the National Supercomputing Center of China in Shenzhen.

## Author Contributions

Conceived and designed the experiments: SG DAS XH. Performed the experiments: SG. Analyzed the data: SG LM XH. Contributed reagents/materials/analysis tools: SG DAS LM. Wrote the paper: SG DAS AY XH.

45. Prinz JH, Wu H, Sarich M, Keller B, Senne M, et al. (2011) Markov models of molecular kinetics: Generation and validation. *J Chem Phys* 134: 174105.
46. Zwanzig R (1983) From Classical Dynamics to Continuous-Time Random-Walks. *J Stat Phys* 30: 255–262.
47. Vitalis A, Caflisch A (2012) Efficient Construction of Mesostate Networks from Molecular Dynamics Trajectories. *J Chem Theory Comput* 8: 1108–1120.
48. Keller B, Hunenberger P, van Gunsteren WF (2011) An Analysis of the Validity of Markov State Models for Emulating the Dynamics of Classical Molecular Systems and Ensembles. *J Chem Theory Comput* 7: 1032–1044.
49. Pan AC, Roux B (2008) Building Markov state models along pathways to determine free energies and rates of transitions. *J Chem Phys* 129: 064107.
50. Morcos F, Chatterjee S, McClendon CL, Brenner PR, López-Rendón R, et al. (2010) Modeling conformational ensembles of slow functional motions in Pin1-WW. *PLoS computational biology* 6: e1001015.
51. Da LT, Avila FP, Wang D, Huang XH (2013) A Two-State Model for the Dynamics of the Pyrophosphate Ion Release in Bacterial RNA Polymerase. *Plos Comput Biol* 9: e1003020.
52. Da LT, Wang D, Huang XH (2012) Dynamics of Pyrophosphate Ion Release and Its Coupled Trigger Loop Motion from Closed to Open State in RNA Polymerase II. *J Am Chem Soc* 134: 2399–2406.
53. Qiao Q, Bowman GR, Huang X (2013) Dynamics of an Intrinsically Disordered Protein Reveal Metastable Conformations That Potentially Seed Aggregation. *Journal of the American Chemical Society* 135: 16092–16101.
54. Bowman GR, Voelz VA, Pande VS (2011) Taming the complexity of protein folding. *Curr Opin Struct Biol* 21: 4–11.
55. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proceedings of the National Academy of Sciences of the United States of America* 106: 19011–19016.
56. Zhuang W, Cui RZ, Silva DA, Huang XH (2011) Simulating the T-Jump-Triggered Unfolding Dynamics of trpzip2 Peptide and Its Time-Resolved IR and Two-Dimensional IR Signals Using the Markov State Model Approach. *J Phys Chem B* 115: 5415–5424.
57. Song J, Gao F, Cui RZ, Shuang F, Liang WZ, et al. (2012) Investigating the Structural Origin of Trpzip2 Temperature Dependent Unfolding Fluorescence Line Shape Based on a Markov State Model Simulation. *J Phys Chem B* 116: 12669–12676.
58. Buch I, Giorgino T, De Fabritiis G (2011) 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* 108: 10184–10189.
59. Chodera JD, Noe F (2014) Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol* 25C: 135–144.
60. Silva DA, Weiss DR, Pardo Avila F, Da LT, Levitt M, et al. (2014) Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proceedings of the National Academy of Sciences of the United States of America* 111: 7665–7670.
61. Held M, Metzner P, Prinz JH, Noe F (2011) Mechanisms of Protein-Ligand Association and Its Modulation by Protein Mutations. *Biophys J* 100: 701–710.
62. Trott O, Olson AJ (2010) Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J Comput Chem* 31: 455–461.
63. Dupont L, Garcia I, Poggi MC, Alloing G, Mandon K, et al. (2004) The Sinorhizobium meliloti ABC transporter cho is highly specific for choline and expressed in bacteroids from Medicago sativa nodules. *J Bacteriol* 186: 5988–5996.
64. Silva DA, Dominguez-Ramirez L, Rojo-Dominguez A, Sosa-Peinado A (2011) Conformational dynamics of L-lysine, L-arginine, L-ornithine binding protein reveals ligand-dependent plasticity. *Proteins* 79: 2097–2108.
65. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4: 435–447.
66. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins* 65: 712–725.
67. Berendsen HJC, Grigera JR, Straatsma TP (1987) The Missing Term in Effective Pair Potentials. *J Phys Chem-Us* 91: 6269–6271.
68. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126: 014101.
69. Cerutti DS, Duke RE, Darden TA, Lybrand TP (2009) Staggered Mesh Ewald: An extension of the Smooth Particle-Mesh Ewald method adding great versatility. *J Chem Theory Comput* 5: 2322.
70. Miyamoto S, Kollman PA (1992) Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J Comput Chem* 13: 952–962.
71. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18: 1463–1472.
72. Weber M, Kube S (2005) Robust Perron Cluster Analysis for various applications in computational life science. *Lect Notes Comput Sc* 3695: 57–66.
73. Singhal N, Snow CD, Pande VS (2004) Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J Chem Phys* 121: 415–425.
74. Zhang L, Silva DA, Yan YJ, Huang XH (2012) Force field development for cofactors in the photosystem II. *J Comput Chem* 33: 1969–1980.
75. M. J. Frisch GW T, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, et al (2004) Gaussian 03, Revision E.01.
76. Fuller JC, Jackson RM, Edwards TA, Wilson AJ, Shirts MR (2012) Modeling of Arylamide Helix Mimetics in the p53 Peptide Binding Site of hDM2 Suggests Parallel and Anti-Parallel Conformations Are Both Stable. *Plos One* 7: e43253.
77. E W, Vanden-Eijnden E (2006) Towards a theory of transition paths. *J Stat Phys* 123: 503–523.