Ligand Binding and Release Investigated by Contact-Guided Iterative Multiple Independent Molecular Dynamics Simulations

Xin-fan Hua1, Xin-zheng Du1, Zhi-yong Zhang*

National Science Center for Physical Sciences at the Microscale, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei 230026, China

(Dated: Received on October 12, 2020; Accepted on November 4, 2020)

Binding and releasing ligands are critical for the biological functions of many proteins, so it is important to determine these highly dynamic processes. Although there are experimental techniques to determine the structure of a protein-ligand complex, it only provides a static picture of the system. With the rapid increase of computing power and improved algorithms, molecular dynamics (MD) simulations have diverse of superiority in probing the binding and release process. However, it remains a great challenge to overcome the time and length scales when the system becomes large. This work presents an enhanced sampling tool for ligand binding and release, which is based on iterative multiple independent MD simulations guided by contacts formed between the ligand and the protein. From the simulation results on adenylate kinase, we observe the process of ligand binding and release while the conventional MD simulations at the same time scale cannot.

Key words: Molecular dynamics simulation, Enhanced sampling, Binding and releasing process, Adenylate kinase

I. INTRODUCTION

There is an inextricable link between biological functions and binding/release process of biomolecules. For example, many experiments have shown that dephosphorylation of adenosine triphosphate (ATP) is accompanied by its binding and release with adenylate kinase (AdK). Keeping the balance of ATP and adenosine diphosphate (ADP) by changing the conformation of AdK is an important cellular energy metabolism reaction.

Over the last years, there are traditional methods that could determine protein structures. For instance, X-ray crystallography could gain many different states and detailed structural information [1–14]. These experimental methods, however, can only provide static structure pictures that have little dynamic information about the ligand binding and release process. Although some work shows that a molecular description of a high-energy enzyme state may be captured by traditional methods like nuclear magnetic resonance (NMR) spectroscopy [15], elucidating the dynamic information of ligand binding and release is still a great challenge.

Molecular dynamic (MD) simulation is a powerful tool to investigate molecular motions [16–18] which could gain insight into the detailed information of important functional biomolecules such as AdK. A lot of work has proven the flexibility of AdK conformation [19–21]. The possibly minimum energy pathway between the open and close states has been captured by a coarse-grained plastic network [22], which predicts the order of domain motions. Some researches focus on determining whether the binding of the ligand is necessary for the transition [23], which suggests that the open state has lower free energy than the closed one. Ping et al. have run long time-scale MD simulations for two systems. One is for the closed AdK in complex with an...
inhibitor Ap5A, and the other is the free open structure. They proposed a possible catalytic cycle of the enzyme [24].

Although considerable MD work about AdK has been done, conventional simulations can only generally reach the timescale around microseconds that is hard to capture the rare events on the millisecond time scales or even longer. To overcome this limitation, one approach is to develop more powerful computer architecture to overcome the gap of time-scale. The Shaw group has achieved significant advance in this area who is famous for “Anton” that is a chip specifically designed for molecular simulations [25, 26]. The other way is to develop more advanced algorithms, such as enhanced sampling techniques seeking the “rare events” in MD. For instance, Gaussian accelerated MD [27] is a commonly used methodology, which applies bias potential to smoothen the system potential energy surface in order to accelerate sampling. Steered MD drive the coordinates changing by adding external force to the system [28]. Other than this, adaptive sampling [29] is an algorithm inspired by exploration-exploitation dilemma. In the case of an unknown ligand binding site, adaptive sampling is predominantly explored, which finds a small area of interest (ligand binding site) by exploring large and undetermined spaces (protein-ligand interactions).

Our work on the enhanced sampling is based on iterative multiple independent molecular dynamics (MIMD) simulations. We have incorporated new criteria that mainly consist of the feature based on contacts to guide the iterative MIMD simulations. We would like to develop a tool to study the process of ligand binding and release, which has been tested extensively on AdK.

II. METHODS

A. Iterative MIMD simulations

Our work on the enhanced sampling is based on multiple sets of independent MD simulations (MIMD) that is used to accelerate transition without external perturbations [30–38], in order to find poor events in simulations by an iterative process. In iterative MIMD, one runs a short simulation firstly to obtain an initial trajectory. Then, making a selection of conformations (called seeds) based on predefined metrics. After that, running a number of independent simulations for resampling starts from the selected seeds. The cycles are repeated until convergence. One of the most important issues in iterative MIMD is how to improve sampling efficiency by construct metrics for selecting seeds [39]. The criteria that have been commonly chosen are root mean square deviation (RMSD), bond, angle, radius of gyration, or domain distances, which are mainly used in domain motions and induced fit simulations. The outlier flooding methods tries to find the outlier in the sparse distributions on the conformation space [40]. The complementary coordinates (Co–Co) MD [41] and the two-ended Data-Driven Accelerated (teDA2) method [42, 43] are based on principal components analysis to construct the criterion. Frontier expansion sampling finds the seeds by integrating the Gaussian mixture model and the convex hull algorithm [44].

B. Contact-guided iterative MIMD

We calculate a contact if a heavy atom in one residue is within 6.0 Å of any heavy atom in a given ligand. At each cycle, we calculate contacts between the protein residues and the ligands, and then pick those conformations based on those contacts to start the next cycle of simulations. It should be noted that different residue contacts have different contribution to ligand binding and release process. In order to get better accuracy, we can make use of the information of the native contacts in the protein-ligand complex. If we have a target structure, we can define weight from this structure to distinguish the importance of interactions between the residues and the ligand. When the target structure is not available, we may use any experimental information that indicates some of the contacts. Here is the equation of the weight:

\[ w_i = \frac{C_i}{\sum_{i=1}^{N} C_i}, \quad i = 1, \ldots, N \]  

(1)

where \( C_i \) presents contact number of the residue \( i \). Two criteria were then constructed to study the ligand binding and release.

1. The centroid distance between a residue and the ligand (FIG. 1(a))

The blue and the red balls present the atoms in the residue and the ligand, respectively. The centroid distance is denoted as \( D \), and its weight is defined by the contact number between the residue and the ligand (Eq.(1)). The picked conformation with the maximal or the minimal value would depend on the initial structure. If it is an open and unbounded structure, we want to drive the ligand to approach the protein as close as possible, therefore in each cycle we choose these conformations with the minimal \( \sqrt{\sum_{i=1}^{N} w_i D_i^2} \) to start the next run. If the initial is a closed and bounded complex, we want to release the ligand, so in each cycle these conformations with the maximal \( \sqrt{\sum_{i=1}^{N} w_i D_i^2} \) are used to
FIG. 1 The criteria used in iterative MIMD to study ligand binding and release. (a) The centroid distance between a residue and the ligand. The blue and the red ball represent the residue and the ligand, respectively. The centroid distance is denoted as $D$, and its weight is defined by the contact number between the residue and the ligand. (b) The root-mean square deviation of the centroid distance between a residue and the ligand. The centroid distance in the target structure, which can be an unbounded or bounded conformation, is denoted as $D_0$. For a conformation with the centroid distance $D$, $d = D - D_0$. Its weight is also defined by the contact number between the residue and the ligand.

2. The root-mean square deviation of the centroid distance between a residue and the ligand (FIG. 1(b))

The centroid distance in the target structure, which can be an unbounded or bounded conformation, is denoted as $D_0$. In each cycle, for a conformation with the centroid distance $D$, the RMSD of the centroid distances $\sqrt{\sum_{i=1}^{N} w_i d_i^2}$, where $d = D - D_0$, are computed.

We then take those conformations with minimal RMSD that represent the most similar ones to the target structure to start the next cycle. We refer to this method for short as rms-dist.

The flow chart of our method using the aforementioned criteria is shown in FIG. 2.

C. Simulation system

The structure of the AdK has been well studied by traditional experiment methods [1–14], which has become a benchmark to investigate conformational transition in protein [45]. Taken the AdK in E. Coli as an example, there are structurally three domains [2–4, 7, 13]. A large center domain called the CORE contains the residues from 1–29, 60–121 and 160–214, an AMP-binding domain called the AMPbd contains the residues from 30–59, and an ATP-binding domain that looks like a lid contains the residues from 122–159 (called the LID). AdK is an important monophosphate transferase that plays a key role in controlling the level of ATP by catalyzing the reaction $\text{ADP-Mg}^{2+} + \text{ADP} = \text{ATP-Mg}^{2+} + \text{AMP}$. Two systems were built, which were described in detail in another work [46]. Briefly, one system is AMP and ATP-Mg$^{2+}$ placed near the open AdK ($O_{\text{AMP+ATP-Mg}}$), and the other is ADP and ADP-Mg$^{2+}$ binding to the closed AdK ($C_{\text{ADP+ADP-Mg}}$).

D. Conventional MD simulations

All the MD simulations were performed using the GROMACS-4.5.5 package [47] and the CHARMM27 force filed [48]. The parameters were described in another work [46].

E. Contact-guided iterative MIMD of AdK

From our previous work [46], nine contacts between the residues and the ligands were defined in $C_{\text{AMP+ATP-Mg}}$. In the CORE, ATP-Mg$^{2+}$ is in contact with K13 (78), R119 (98), and K200 (16) while AMP is in contact with R88 (109). In the AMPbd, AMP is in contact with R36 (27) and K57 (24). In the LID, ATP-Mg$^{2+}$ is in contact with R123 (94) while AMP is in contact with R156 (25) and D158 (17). These contacts were used to calculate metrics in the dist or rms-dist criterion.

In the cycle 0, a 100 ps MD simulation was run, and the generated conformations were ranked by the aforementioned metrics. The criterion was calculated using the $C_a$ atoms, and the first $n=12$ structures with the most suitable values were picked. It has been said that,
in sampling methods using iterative MIMD simulations, 10–20 seed conformations would be suitable for starting the next cycle of MIMD [38]. We did a preliminary test as well. In one cycle of the releasing process, the conformations were sorted in descending order of the metrics (FIG. S1(a) in the supplementary materials) and those conformations with the largest metrics should be picked. In one cycle of the binding process, the conformations were sorted in ascending order of the metrics (FIG. S1(b)) and those conformations with the smallest metrics should be picked. The results indicate that, in both the releasing and the binding process of AdK, 12 is a fairly reasonable number of seeds to be picked in each cycle. For other systems, we may perform the same test to determine an appropriate number of $n$.

Starting from these structures, $n$ independent MD simulations were run. Each simulation was 100 ps and a configuration was saved every 1 ps, so there were totally $N=1200$ structures in every cycle. We repeated the procedure for 200 cycles, or until the criterion value was smaller than a cutoff value. In practice, we set a very small cutoff value, so the simulations were always run for the maximal number of cycles.

III. RESULTS AND DISCUSSION

A. Ligand binding/release in AdK using different criteria

The process of ligand binding and release is highly related to the interactions between the key residues and the ligands. Thus, we monitor the changes of these contacts during simulations, in order to analyze the details.

1. Ligand release simulations

The closed bounded conformation of AdK with ADP and ADP·Mg$^{2+}$ is the initial structure (FIG. 3(a)), and the target conformation was set as the open state of the AdK (FIG. 3(d)). Two independent simulations were conducted using the two different criteria (FIG. 1). Each simulation has 200 cycles. FIG. 3 (b) and (c) show the conformation in the final cycle using the dist and the rms-dist criteria, respectively. We also plotted the contact distances as the function of the cycle number (FIG. 4).

In each cycle, the distances between the residues and the ligands are presented by the average value of the 1200 conformations. For each contact, the black dotted line represents the distance between the residue and the ligand in the target conformation (FIG. 4), the red line represents the distance change in the simulation using the dist criterion, and the blue line represents the distance change in the simulation using the rms-dist criterion. In the dist simulation, the distance between R123 in the LID and ADP·Mg$^{2+}$ is increased from 6.7 Å to 12.4 Å that is 1.0 Å larger than the target distance. The distance between R156 and ADP is increased from 9.7 Å to 15.6 Å that is only 0.6 Å smaller than the target, and the distance between D158 and ADP is increased from 8.7 Å to 12.2 Å that is 0.9 Å smaller than the target. On the contrary, these contact distances do not change much in the dist simulation. In the rms-dist simulation, the distance between R36 in the AMPbd and ADP is increased from 7.6 Å to 12.1 Å that is 1.9 Å smaller than the target, and the distance between K13, R119, K200 in the CORE and ADP·Mg$^{2+}$ do not change much, whereas those in the dist simulation are increased significantly. The results indicate that, in the dist simulation without the target, only three contacts between the CORE (K13, R119, and K200) and ADP·Mg$^{2+}$ are broken while the other contacts keep stable, so the ligands cannot release off the pockets. However, in the rms-dist simulation using the O$_{ADP+ADP·Mg}$ as the target, all the contacts in the LID and the AMPbd and one contact in the CORE (R88) are broken, which means that the ligands are released off the pockets during the simulation. The three con-
FIG. 4 The centroid distances between the residues and the ligands as the function of the cycle number in the release process. In each panel, the black dotted line represents the distance between the residue and the ligand in the target conformation, the red line represents the distance change in the simulation using the dist criterion, and the blue line represents the distance change in the simulation using the rms-dist criterion.

Contacts between the CORE and ADP·Mg$^{2+}$ are relatively stable because they exist in the target structure.

All results of these contacts support that the rms-dist criterion is better in simulating the release process than the dist criterion. We have found that there are obvious sequential changes in the residues from different domains. The three contacts in the LID (R123, R156, and D158) have a rapid increase from the cycle 75 to 100, and then reach equilibrium in the remaining cycles. As a comparison, the two residues in the AMPbd have a substantial increase from the cycle 100, and tend to be in equilibrium in the following 20–30 cycles. This result may support that the releasing process of the LID is earlier than that of the AMPbd while the former undergoes a larger movement during the simulation, which may indicate that the releasing process of the AMPbd is the rate-limiting step. As for the CORE, the distance between R88 and ADP·Mg$^{2+}$ has two rapid increases that are from the cycle 30 to 50 and cycle 100 to 120, respectively, which could correspond to the increase of contact distance in the LID and the AMPbd. It should be noted that, R88 is located between the two domains, which may indicate that this part of the CORE accompanies the movement of the LID and the AMPbd simultaneously.

2. Ligand binding simulations

The open conformation of AdK with the unbounded AMP and ATP·Mg$^{2+}$ was the initial structure (FIG. 5(a)), and the target conformation was set to the closed state (FIG. 5(d)). FIG. 5 (b) and (c) show the conformation in the final cycle using the dist and the rms-dist criterion, respectively. We also plotted the contact distances as the function of the cycle number (FIG. 6).

In each cycle, the distances between the residues and the ligands are presented by the average values of the 1200 conformations. For each contact, the black dotted line presents the distance between the residue and the ligand in the target conformation, the red line and the blue line present the distance change in the simulation using the dist and the rms-dist criterion, respectively. In the rms-dist simulation, the distance between R123...
B. Comparison between the target and non-target criterion

By using two different criteria in the iterative MIMD simulations, we want to not only investigate the binding and releasing process but also evaluate the performance of the criteria. We divide our criteria into two categories, one uses the target conformation and the other does not. First of all, we discuss the distance changes between the residues and the ligands in the 200 cycles of the release process. We can observe that the distances in the rms-dist simulations can be closer to the target values in almost all the residue-ligand pairs than those in the dist simulations, except for the residue K200. In fact, when we compare the distance between K200 and ADP-Mg$^{2+}$, it is just a little closer to the target value in the dist simulation than that in the rms-dist simulation. Then we talk about the binding process, the distances between R123, R156, R158, K57, R88, R119 and its corresponding ligands can reach the target values in the rms-dist simulations. We can find the similar result in the binding and release process because the criterion is constructed by the target conformation. Despite this, the criterion without the target conformation still has a fairly good performance in the process of binding and releasing.

IV. CONCLUSION

Even though we have obtained much knowledge on the process of ligand binding/release with the help of MD simulations, nowadays it is still hard to get enough information especially when studying a large system at a long time scale. Inspired by the work of Ye et al. [46], that emphasizes the importance of some key residue contact in the process of ligand binding/release, that is, the distances between ATP-Mg$^{2+}$ and the two residues increase in the first 20 rounds, then reduce to around initial level in the next 40 rounds, and finally maintain a stable value in the left simulation. These results are consistent with the work of Ye et al. [46], which indicates that the CORE should stay away first during the close process, in order to make room for the other domains to reach the right place. Two residues, R156 and D158 in the LID, also have a similar process. The distances reduce quickly at first 10 rounds, and reach equilibrium at about 20 rounds. The results are consistent with the hypothesis that the LID binds to the right place at the first 50 rounds. Three residues, R36, K57 and R123 in the AMPbd, also have the process of falling, keeping balance and then falling. In the dist simulation, R88 and K200 first descend, and then ascend when K13 has a similar but larger up-and-down process, which indicates that more room the CORE makes firstly, the easier binding process could happen.

FIG. 5 Conformations in the binding process. (a) The initial conformation, (b) the conformation in final cycle when using the dist criterion, (c) the conformation in final cycle when using the rms-dist criterion, and (d) the target conformation. In each structure, the CORE is colored in gray, the AMPbd is colored in red, the LID is colored in blue, ATP-Mg$^{2+}$ is colored in purple/pink and AMP is colored in green. The ligands and the contacted residues are shown in CPK mode.

in the LID and ATP-Mg$^{2+}$ is decreased from 11.8 Å to 6.9 Å that is 0.4 Å larger than the target distance, the distance between R156 and AMP is decreased from 17.3 Å to 10.6 Å that is 0.8 Å larger than the target, and the distance between D158 and AMP is decreased from 13.7 Å to 9.8 Å that is 0.8 Å larger than the target. In the dist simulation, the contact distance of R123-ATP-Mg$^{2+}$ reaches 7.7 Å, larger than that in the rms-dist simulation. However, the contact distances of R156-AMP/D158-AMP are 8.9 Å/6.9 Å, significantly smaller than those in the rms-dist simulation. As for the residues in the AMPbd, the distance between R36 and AMP is decreased from 12.0 Å to 9.5 Å that is 1.4 Å larger than the target distance, whereas the distance between K57 and AMP is decreased from 17.1 Å to 10.6 Å that is 1.0 Å larger than the target, in the rms-dist simulation. The two contact distances decrease to 8.8 Å and 10.5 Å in dist simulation. In the rms-dist simulation, the distance between R88 in the CORE and AMP is decreased from 7.8 Å to 7.6 Å that is 2.0 Å larger than the target, the distance between K13 and ATP-Mg$^{2+}$ is decreased from 6.0 Å to 5.8 Å that is 0.6 Å smaller than the target, the distance between K200 and ATP-Mg$^{2+}$ is increased from 13.8 Å to 14.3 Å that is 1.5 Å larger than the target, and the distance between R119 and ATP-Mg$^{2+}$ is decreased from 8.0 Å to 5.8 Å that is 0.5 Å larger than the target. It should be noted that K13 and K200 have similar process during the rms-dist simulation.
we designed new criteria by adding different weights on residue contacts. We have tested two criteria with or without the target.

We have validated the method using the AdK system. The results have shown that it can get an efficient sampling of the binding/release process. By investigating distance changes between residues and ligands during the simulations, we find some key residue contacts, such as R36 and K57, may indeed have great contributions to the binding and releasing processes, while some other contact may become the rate-limiting step during the transition, such as R119 and K200.

In many cases, a high resolution structure of the target protein-ligand complex may not be available. However, it may be relatively easy to obtain some information on interactions and contacts between some residues and the ligands. This limited number of experimental data may be incorporated in our method to guide the simulations of ligand binding and release.

**Supplementary materials:** Figure S1 shows how to pick a proper number of MIMD simulations in each cycle.

---

**V. ACKNOWLEDGMENTS**

This work was supported by the National Natural Science Foundation of China (No.91953101), the Strategic Priority Research Program of the Chinese Academy of Science (XDB37040202), the Hefei National Science Center Pilot Project Funds, and the New Concept Medical Research Fund of USTC. The authors thank the Supercomputing Center of USTC for providing the computer resources for this project.