

Exploring protein conformations by cluster-guided iterative multiple independent molecular dynamics simulations

Chengtao Ding(丁成涛), Guanglin Chen(陈光临), Qingguo Gong(龚庆国), and Zhiyong Zhang(张志勇)
Citation: Chin. Phys. B, 2026, 35 (**5**): 058702. DOI: 10.1088/1674-1056/ae0898
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Inhibitory effect of Relatlimab on LAG3-FGL1 interaction investigated by molecular dynamics simulation

Qing Xie(谢晴), Xue-Feng Liu(刘雪峰), Yu-Qing Wang(王雨晴), and Chen-Xiang Wang(王辰祥)
Chin. Phys. B, 2026, 35 (**4**): 048701. DOI: 10.1088/1674-1056/ae0308

Simulation on mechanochemical coupling of rotary biomotors F_1 and V_1

Liqiang Dai(戴立强), Yao-Gen Shu(舒咬根), and Zhong-Can Ouyang(欧阳钟灿)
Chin. Phys. B, 2026, 35 (**2**): 028702. DOI: 10.1088/1674-1056/ae2674

Mechanism of loop-2 in facilitating microtubule depolymerase activity of kinesin-8 motors

Xiao-Xuan Shi(史晓璇), Yao Wang(王瑶), Jie Wang(王杰), Yu-Ru Liu(刘玉如), and Ping Xie(谢平)
Chin. Phys. B, 2026, 35 (**2**): 028701. DOI: 10.1088/1674-1056/adf4a8

Reconfiguration of B-DNA structure induced by ethanol

Yue Huang(黄悦), Yipeng Chen(陈以鹏), Jing Li(李静), Rongri Tan(谈荣日), and Huanhuan Qiu(邱环环)
Chin. Phys. B, 2025, 34 (**8**): 088707. DOI: 10.1088/1674-1056/adbed9

Molecular simulation study on phase separation of immunoglobulin G

Lv-Meng Hu(胡吕梦), Yuan-Qiang Chen(陈远强), Hong-Ming Ding(丁泓铭), and Yu-Qiang Ma(马余强)
Chin. Phys. B, 2025, 34 (**8**): 088701. DOI: 10.1088/1674-1056/add50b

---

Exploring protein conformations by cluster-guided iterative multiple independent molecular dynamics simulations

Chengtao Ding(丁成涛)[1,2,†], Guanglin Chen(陈光临)[3,†], Qingguo Gong(龚庆国)[2], and Zhiyong Zhang(张志勇)[2,3,‡]

[1] *Department of Clinical Laboratory Center, The First Affiliated Hospital of Anhui University of Chinese Medicine, Hefei* 230031, *China*
[2] *Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei* 230026, *China*
[3] *Department of Physics and Anhui Center for Fundamental Science in Theoretical Physics, University of Science and Technology of China, Hefei* 230026, *China*

(Received 28 July 2025; revised manuscript received 11 September 2025; accepted manuscript online 18 September 2025)

Enhanced sampling methods in molecular dynamics (MD) simulations have been gaining popularity in the past decades because they can explore conformations of proteins more efficiently than conventional MD simulations. In this paper, we implement a protocol of enhanced sampling that combines iterative multiple independent MD simulations and cluster analysis. After a set of independent simulations, the combined trajectory is divided into clusters. The representative structures picked from the clusters are utilized to start the next cycle of MD simulations. By using different strategies to pick the representative structures, the enhanced sampling can be either targeted or non-targeted. Two multi-domain proteins, *Escherichia coli* adenylate kinase (AdK) and the three-domain (PHD-Bromo-PWWP) structure in the BS69 protein, were selected to test the method. The data indicate that conformations of the proteins can be efficiently explored, and the results show better agreement with the experimental data than those obtained through conventional MD simulations.

Keywords: enhanced sampling method, molecular dynamics simulation, cluster analysis, free energy calculation, multi-domain protein

PACS: 87.15.ap, 87.14.E–, 87.15.B–

DOI: 10.1088/1674-1056/ae0898 **CSTR:** 32038.14.CPB.ae0898

1. Introduction

Dynamic changes in proteins have aroused increasing interest in recent years, as they often play important roles in various biological functions.[1–3] To better understand the relationship between structure and function, x-ray crystallography, cryo-electron microscopy (cryo-EM), and nuclear magnetic resonance (NMR) have been widely used. However, it is still difficult to capture the conformational dynamics of a large protein in atomic detail. Although some alternative experimental techniques, such as small-angle x-ray scattering (SAXS),[4] fluorescence resonance energy transfer (FRET),[5] and chemical cross-linking with mass spectrometry (CXMS)[6] can be applied, molecular dynamics (MD) simulation is a very popular computational tool for understanding the conformational dynamics of proteins.[7–9]

With the advances in both hardware and software, MD simulations have accessed longer timescales for proteins at larger length scales.[10–13] However, MD simulations could be computationally expensive, which is an obstacle to future applications. For a relatively large protein, a simulation beyond microseconds (μs) will take up many computer resources, whereas a shorter simulation will have insufficient sampling. To overcome this obstacle, many enhanced sampling methods have been developed.[14–16] These kinds of methods usually require direct modifications of the complicated MD code.

In recent years, a new kind of sampling technique has been proposed,[17–22] which is based on iterative multiple independent MD (MIMD) simulations. These methods consist of many cycles. In each cycle, short MIMD simulations are conducted, and then from these trajectories, some conformations (called seeds) are selected to start the next cycle of MIMD. Unlike the previous techniques, this category of sampling methods does not need to change the MD code. Their efficiency would depend on how to select seed conformations for the next cycle, and different criteria have been tried.[23]

In this paper, we present a protocol that combines iterative MIMD and cluster analysis. In each cycle, trajectories generated by MIMD simulations are consolidated, and the conformations are divided into a certain number of clusters. Representative structures, each picked from one cluster, are used to start the simulations of the next cycle. There are various algorithms available for cluster analysis.[24] We can define different metrics to determine the clusters and use different criteria to pick the representative of each cluster, which would depend on the purpose of enhanced sampling.

Two multi-domain proteins were investigated to demonstrate the method. The first is *Escherichia coli* adenylate kinase (AdK),[25,26] which is a model system for studying con-

[†] These authors contributed equally to this work.
[‡] Corresponding author. E-mail: zzyzhang@ustc.edu.cn
© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

formational transitions since both the open and closed structures are known. We searched its transition pathways and compared the results with other sampling methods. The second system is the three-domain (PHD-Bromo-PWWP) structure in the BS69 protein.[27] Since there is no experimental structure available, we employed this approach to illustrate how well it can explore the protein conformations, and our results show better agreement with the experimental data than those obtained by the conventional MD simulation.

2. Materials and methods

2.1. Procedure of the cluster-guided iterative MIMD

Our protocol (denoted as cluster-MIMD) is to run a series of independent MD simulations combined with a clustering algorithm. The flow chart is as follows (Fig. 1).

1. Obtain an initial structure of the protein and set up the system.

2. Perform a preliminary MD simulation, and a conformation is saved every 1 ps.

3. Divide the conformations in the trajectory into M clusters. From each cluster, a representative conformation is picked.

4. Starting from each of the M representative conformations, we reassign atomic velocities and carry out an independent short simulation. A combined trajectory is constructed that contains $M \times 100$ conformations.

5. Repeat steps 3 and 4 for N cycles until satisfactory sampling results are obtained.

In this work, we set $M = 12$ and each independent short simulation was carried out for 100 ps. The selection of these parameters refers to another MIMD method called PaCS-MD.[17,23,28]

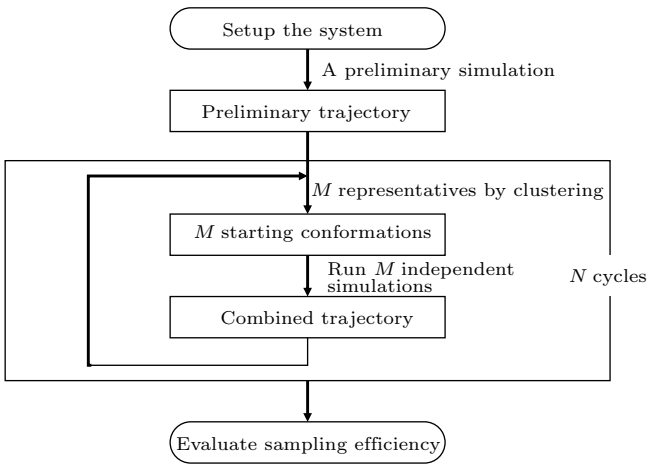


Fig. 1. Flowchart of the cluster-guided MIMD simulation.

2.2. Clustering methods

Various clustering algorithms can be used. In this work, we implemented a standard K-means algorithm[29] to divide the conformations into a certain number of clusters. Different metrics, such as root mean square deviation (RMSD), angles, and center-of-mass (COM) distances, could be chosen to measure "distances" between conformations and determine the clusters. One may try different ways to pick the representative in each cluster for starting the next cycle of MIMD simulations. The first is to pick the conformation that is closest to the cluster center (denoted as CEN), and the second is to pick the conformation that is farthest from the initial structure (denoted as F2I). If a target structure of the protein is available, a third strategy can be used where the conformation closest to the target structure (denoted as C2T) is selected.

2.3. Computational details of AdK

AdK is a monomeric phosphotransferase that can regulate the ATP concentration in cells. The protein contains three domains: a central domain (CORE), an AMP-binding domain (AMPbd), and an ATP-binding domain (LID).[25,26] AdK in *Escherichia coli* has 214 amino acid residues, in which the CORE contains residues 1–29, 60–121, and 160–214, the AMPbd includes residues 30–59, and the LID consists of residues 122–159. AdK has two major conformational states. The free protein is an open conformation,[26] with the AMPbd and the LID both far away from the CORE (Fig. 2(a)). When binding with substrates, the AMPbd and the LID will move toward the CORE and form a closed conformation (Fig. 2(b)).[25,26] AdK is known as a representative example with a large-scale conformational transition.[30–33] In this paper, we used both the conventional MD simulation (cMD) and the proposed cluster-MIMD to demonstrate the transition pathways of AdK.

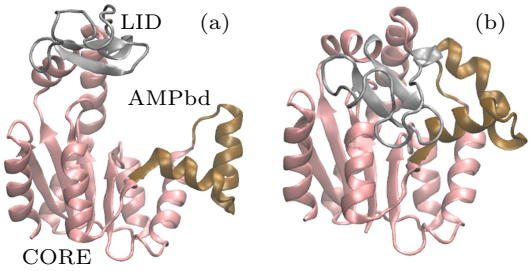


Fig. 2. Crystal structures of AdK. (a) The open state (PDB entry: 4AKE). (b) The closed state (PDB entry: 1AKE) with the inhibitor removed. The CORE, the AMPbd, and the LID are colored in pink, ochre, and silver, respectively.

cMD Starting from both the open and the closed structures of free AdK, MD simulations were carried out using the GROMACS-2018 package[34] and the force field CHARMM36m.[35]

For the open AdK, the periodic boundary condition was used with a dodecahedron box type, and the minimum distance between the solute and the box boundary was 1.2 nm. TIP3P water molecules[36] were filled in the box. The en-

ergy of the protein-water system was minimized by the steepest descent method until the maximum force was less than 1000 $kJ \cdot mol^{-1} \cdot nm^{-1}$. To neutralize the negative charges on the protein, four Na^+ ions were added by replacing the same number of water molecules with the most favorable electrostatic potential. The final system consisted of 45642 atoms in total. The energy was minimized again using the steepest descent followed by the conjugate gradient method until the maximum force was less than 200 $kJ \cdot mol^{-1} \cdot nm^{-1}$. Subsequently, a 100 ps equilibration simulation was performed, with the initial atomic velocities randomly assigned at 300 K. Positional restraints were added to the heavy atoms, with a force constant of 1000 $kJ \cdot mol^{-1} \cdot nm^{-2}$. Finally, a free MD with 240 ns was conducted. The production run was carried out using a time step of 2 fs under constant NPT conditions. Protein, solvent and ions were coupled separately to a temperature bath of 300 K using the velocity rescaling algorithm,[37] with a coupling time of 0.1 ps. The pressure was kept at 1 bar with a coupling time of 0.5 ps and a compressibility of 4.5×10^{-5} bar^{-1}. All the covalent bonds in the protein were constrained using the P-LINCS algorithm.[38] The van der Waals interactions were treated with twin-range cutoff distances (0.9 nm and 1.4 nm, respectively), and the pair list was updated every 20 fs. The PME algorithm[39] was used to calculate the long-range electrostatic interactions with a tolerance of 10^{-5} and an interpolation order of 4.

The setup procedure of the closed AdK was nearly the same as that of the open AdK, except that the minimum distance between the solute and the box boundary was 1.65 nm. The final system contained 45603 atoms in total, which is roughly the same as the open AdK system. In this way, the box size should be sufficiently large for simulating domain motions of the protein.

Cluster-MIMD The simulations also started from both the open and the closed structures of AdK. All the three strategies for picking the representative conformations were tried. Therefore, there were six cluster-MIMD simulations in total. For each cluster-MIMD, the preliminary MD simulation was 100 ps, and every cycle consisted of 12 independent short (100 ps) MD simulations. The parameters of these MD simulations were the same as those of cMD. The recorded 1200 snapshots per cycle were divided into 12 clusters using metrics of two inter-domain distances. One is the COM distance between the LID and the CORE, and the other is the COM distance between the AMPbd and the CORE. The two distances are good indicators of domain motions in AdK. Two hundred cycles were run, so the total simulation time of each cluster-MIMD was 240 ns.

Visualization of sampling Based on the inter-domain distances, we defined a two-dimensional (2D) plane. For conformations from either the cMD or the cluster-MIMD simulations, the distances were calculated. Thus, all conformations can be projected onto the plane to show the sampling efficiency of the different simulations.

Free energy calculation The sampling results obtained by cluster-MIMD are biased, and therefore the free energy landscape (FEL) needs to be calculated to reflect the true distribution of the sampling. The FEL formed by the specified reaction coordinates was calculated using umbrella sampling. The conformational space sampled by cluster-MIMD is evenly divided into windows according to the reaction coordinates. In each window, the structure nearest to the center is selected as the reference structure of the umbrella sampling. To make sure the conformational space is efficiently sampled, the umbrella potential

$$V_i^{\text{umbrella}} = \sum_j \frac{1}{2} k_{ij} \left(q_j - q_{ij0}\right)^2 \tag{1}$$

is applied, where k_{ij} is the force constant for the jth reaction coordinate in the ith window. q_j and q_{ij0} are the jth reaction coordinate of instantaneous and reference structure in the ith window. The unbiased probability density of conformations ρ_0 can be calculated by combining the results of windows using the WHAM equation.[40] Finally, the free energy F is calculated as

$$F = -k_{\text{B}} T \ln \rho_0, \tag{2}$$

where k_{B} is the Boltzmann constant and T is the temperature.

2.4. Computational details of PHD+Bromo+PWWP in BS69

The BS69 protein is a specific repressor of adenovirus E1A transcription activation, which contains 562 residues, including a PHD domain, a Bromo domain, a PWWP domain, and a MYND domain.[41] The crystal structure of the Bromo+PWWP two-domain in BS69 (residues 154–371, denoted by B2D) has been solved,[27] but the structure of the PHD+Bromo+PWWP three-domain (residues 98–371, denoted by B3D) is unavailable because the linker (residues 148–157) between PHD and B2D is very flexible. To investigate the solution structure of B3D, we conducted SAXS experiments (described in supporting information). A structural model was built by the following procedure. First, a homology model of the PHD domain was obtained using MODELLER,[42] and then the relative orientation between PHD and B2D was determined by fitting the SAXS data using the BUNCH program[43] in the ATSAS 2.8 package.[44] Finally, the atomic coordinates of the linker were generated by MODELLER. Starting from this structural model of B3D, we conducted both cMD and cluster-MIMD simulations.

cMD and cluster-MIMD The simulation details of B3D were nearly the same as those of AdK. The a99SB-disp force

field[45] was used. The minimum distance between the solute and the box boundary was 1.6 nm, in order to simulate large scale domain motions.

A preliminary MD simulation of 60 ns was run to relax the structural model of B3D. Then a follow-up MD simulation of 60 ns was conducted as a control.

A cluster analysis was done on the preliminary 60-ns MD trajectory, and then a follow-up cluster-MIMD was run for 50 cycles, so the total simulation time was 60 ns as well. The conformations were clustered using a metric of two inter-domain distances. One is the COM distance between PHD and Bromo, and the other is the COM distance between PHD and PWWP. Due to the absence of knowledge about the target conformation, only CEN and F2I were used to pick the representative structures. During the 50 cycles of cluster-MIMD, a mixed strategy was used by running one cycle of F2I followed by two cycles of CEN, and so on alternatively.

Ensemble optimization method (EOM) To fit the SAXS data, this method selects a small ensemble of conformations of the protein from a large structural pool by minimizing the following residual:

$$\chi^2 = \frac{1}{K-1}\sum_{i=1}^{K}\left[\frac{\mu I(q_i) - I_{\exp(q_i)}}{\sigma(q_i)}\right]^2, \tag{3}$$

where K is the number of data points in the SAXS profile, and $\sigma(q)$ are experimental errors. The momentum transfer q equals $4\pi\sin\theta/\lambda$, in which 2θ is the scattering angle and λ is the wavelength. For every conformation in the ensemble, its theoretical scattering profile is computed. $I(q)$ is the average of them, and μ is a scaling factor.

In this paper, we used an advanced version of EOM 2.0[46] to select ensembles that are consistent with the SAXS data from the structural pool, which was taken from either the cMD or the cluster-MIMD trajectories. Each selected ensemble consists of 10 representative conformations. The ensemble selection process was performed using the built-in genetic algorithm, which ran for 50 generations. At the end of the genetic algorithm, the selected representative conformations and the corresponding χ^2 values reached stable values.

3. Results and discussion

3.1. Cluster-MIMD of AdK

Sampling efficiency We conducted both open-to-close and close-to-open simulations of AdK. All the trajectories were projected onto the plane of inter-domain distances (Fig. 3). The open state of AdK is located at (30.9, 22.3) Å, whereas the closed state is located at (20.8, 18.2) Å. During the 240 ns cMD simulation starting from the open structure, the CORE-LID distances are from 24.3 Å to 36.1 Å, and the CORE-AMPbd distances are from 19.4 Å to 25.0 Å (Fig. 3(a)). That is, the simulation only samples the region around the open state but does not reach the closed state.

Cluster-MIMD simulations were carried out using the three strategies for picking representative structures. When choosing these CEN conformations for the next cycle of MIMD, the domain motion between the CORE and the LID becomes slightly larger from 22.2 Å to 35.7 Å (Fig. 3(b)), whereas that between the CORE and the AMPbd is nearly the same as that in cMD (Fig. 3(a)). When picking the F2I conformation in each cluster and using them for the next cycle, the domain motion of the LID is increased significantly (Fig. 3(c)), with the CORE-LID distances from 19.6 Å to 38.0 Å. However, the CORE-AMPbd distances are from 20.2 Å to 24.9 Å, indicating that the domain motion of the AMPbd does not increase. If we set the closed AdK structure as the target and choose these C2T conformations for the next cycle, the CORE-LID distances are from 20.7 Å to 33.8 Å and the CORE-AMPbd distances are from 19.0 Å to 24.0 Å (Fig. 3(d)). Compared to the other two strategies, the cluster-MIMD using C2T better approaches to the closed structure. It should be noted that the crystal structure of the closed AdK is in complex with an inhibitor,[26] thus the free AdK could not become that close in the above simulations. MD simulations at the time scale of microsecond, conducted by Li *et al.*,[32] also support this notion.

We then conducted simulations starting from the closed state of AdK. During the 240 ns cMD simulation, the protein quickly escapes from the closed state and transits to the open state. The CORE-LID distances are mainly from 22.4 Å to 34.2 Å, and the CORE-AMPbd distances are mainly from 18.6 Å to 23.2 Å (Fig. 3(e)). These results of cMD (Figs. 3(a) and 3(e)) suggest that the free AdK prefers an open structure, which is consistent with previous studies.[47–49] All the three cluster-MIMD simulations starting from the closed AdK (Figs. 3(f)–3(h)) have better sampling than the cMD simulation (Fig. 3(e)). They all reach the open state of the protein, and among them, the simulation using F2I can most extensively sample the domain motions in the protein (Fig. 3(g)).

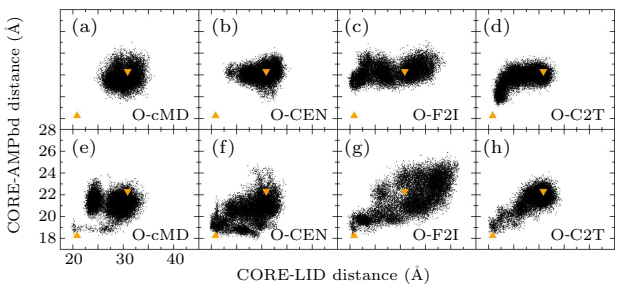


Fig. 3. Sampling efficiency of cluster-MIMD measured by projecting protein conformations onto a 2D plane of inter-domain distances. The projections of cMD starting from (a) the open structure (down-triangle in orange) and (e) the closed structure (up-triangle in orange) shown as controls. (b) Starting from the open structure using CEN. (c) Starting from the open structure using F2I. (d) Starting from the open structure using C2T. (f) Starting from the closed structure using CEN. (g) Starting from the closed structure using F2I. (h) Starting from the closed structure using C2T.

We performed another set of cluster-MIMD simulations (Fig. S1 in the supporting information), and the results are consistent with those shown in Fig. 3.

FEL of AdK The conformational space of AdK sampled by cluster-MIMD was divided into 102 windows with bin size of 1.0 Å along each reaction coordinate. In each window, umbrella sampling was performed with $k_{i1} = 1000.0\ \mathrm{kJ\cdot mol^{-1}\cdot nm^{-2}}$ and $k_{i2} = 500.0\ \mathrm{kJ\cdot mol^{-1}\cdot nm^{-2}}$ (Eq. (1)), and the simulation time was 1 ns. Under this umbrella potential and simulation length, the sampling range slightly exceeded the preset window size. From the sampling results of cluster-MIMD, conformations near the center of each window were selected as the initial configurations for umbrella sampling, adequate overlap between neighboring windows could be ensured, thereby guaranteeing convergence of the umbrella sampling. The umbrella potential was applied using the pull code of GROMACS. The FEL was calculated using WHAM developed by the Grossfield lab.[50]

Figure 4 shows the FEL and some representative conformations of AdK are shown in Fig. 5. The FEL can be divided into closed and open states. The conformational space of the closed state (a and b in Fig. 4) is limited, with both the LID and the AMPbd compact (Figs. 5(a) and 5(b)). These two closed conformations are somewhat more open than the crystal structure 1AKE (up-triangle in Fig. 4), with a root mean square deviation (RMSD) of 2.5 Å and 3.3 Å, respectively. That is to say, the free protein can hardly reach the closed structure in complex with the inhibitor. The conformational space of the open state (e, f, and g in Fig. 4) is significantly larger. Energy barriers between them are relatively low, and thus conformational changes can occur frequently. The conformations e and f (Figs. 5(e) and 5(f)) have the lowest free energy values, and therefore, the free protein prefers open structures. They are similar to the experimental structure 4AKE (down-triangle in Fig. 4), with an RMSD of 1.3 Å and 2.1 Å, respectively. Some conformations that are even more open than 4AKE are explored, such as the conformation g (Fig. 5(g)).

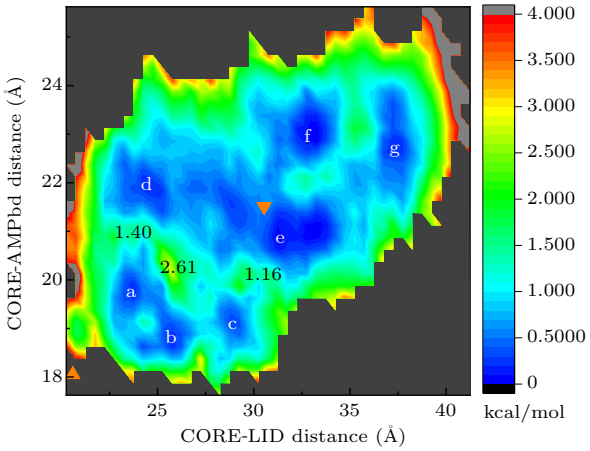


Fig. 4. The FEL of AdK. The orange down-triangle and up-triangle indicate the open and closed experimental structures of AdK. Some representative states are labeled a–g, and the corresponding structures are shown in Fig. 5. Free energy values (in kcal/mol) of some barriers are shown.

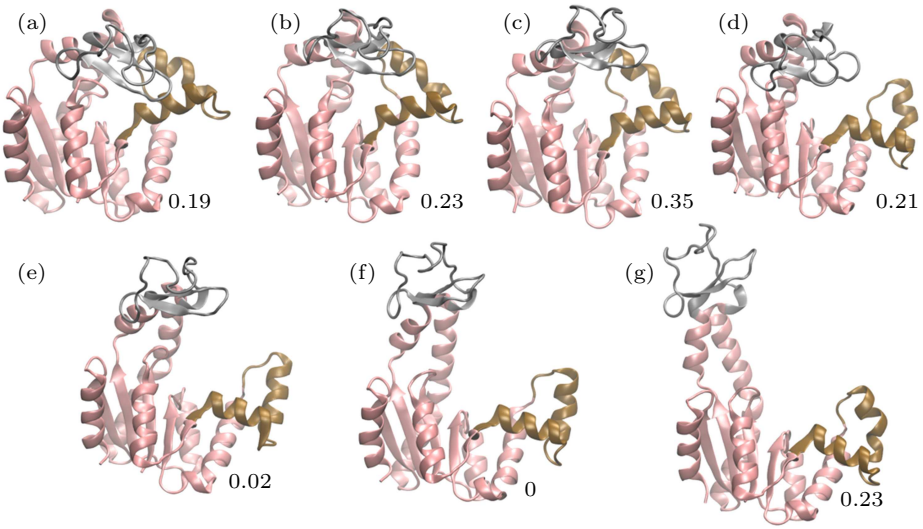


Fig. 5. Representative states of AdK in conformational transition. The value at the bottom-right of each structure is the free energy (in kcal/mol) of this state. The locations of these states are labeled in Fig. 4.

In addition to the open and closed structures, two intermediate states were detected. The conformation c (Fig. 4) represents a state in that the AMPbd essentially stays closed while the LID is becoming open (Fig. 5(c)). In the other intermediate (conformation d in Fig. 4), the AMPbd is significantly open but the LID is closed (Fig. 5(d)). There are two transition pathways from the closed to the open state. In one pathway, the LID opens first (b to c) and then the AMPbd (c to e). The energy barrier from b to c is low and that from c to e is about 0.81 kcal/mol. In the other pathway, the AMPbd opens first (a to d) and then the LID (d to e). The energy barrier from a to d is about 1.21 kcal/mol but that from d to e is low. That is to say, the former pathway should be more favorable than the latter. There are also two transition pathways from the open to the closed state. In one pathway, the AMPbd closes first (e to c) and then the LID (c to b). The energy barrier from e to c is

about 1.14 kcal/mol and that from c to b is low. In the other pathway, the LID closes first (e to d) and then the AMPbd (d to a). The energy barrier from e to d is low and that from d to a is about 1.19 kcal/mol. The results indicate that the open-to-close transition is harder than the close-to-open transition, and the opening/closing of the AMPbd may be the rate-limiting step.

In the previous work by Li *et al.*,[32] they aimed to comprehensively understand AdK's conformational transition through long MD on a timescale of micro-seconds and bias-exchange metadynamics (BE-META) simulations. They quantitatively mapped the FEL of AdK on transition pathways and found some intermediate states. When starting from the open structure, the conformation is sometimes trapped in a state in which the LID closes and contacts the AMPbd while the AMPbd stays in the open or semi-open state. After closure of the LID, the AMPbd can start closing. There is another pathway observed in their MD simulations, that is, the AMPbd closes first and then the LID closes afterward. The simulations starting from the closed state have a conformational transition pathway in that the LID opens rapidly while the AMPbd can stay in the closed state for a while before opening. Cluster-MIMD can obtain similar results, but our simulation time is shorter than theirs.

Cluster-MIMD can further improve efficiency by adjusting the number of clusters. During the exploration of the conformational space, the sampling region starts from the initial conformations and gradually extends farther away. The cluster operations in each iteration ensure that the sampling regions of independent simulations in the next iteration are spatially separated, preventing redundant sampling of the same regions. However, this operation is ineffective during the first few iterations, since the sampling region remains close to the initial conformations regardless of the number of clusters used; thus, similar results can be obtained even with fewer clusters. As the sampling region moves farther from the initial positions, the space to be sampled increases. To adequately cover the conformational space and ensure convergence of umbrella sampling, it becomes necessary to increase the number of clusters. Therefore, an efficient strategy is to start the iterations with a smaller number of clusters and increase the cluster count as the sampling regions become sparse. Similarly, for simple systems with a small target sampling space, fewer clusters can be used, whereas for systems with a larger target space, the number of clusters should be increased according to the sparsity of the sampled regions.

3.2. Cluster-MIMD of B3D

In the 60 ns cMD simulation, the protein only samples a limited region in the 2D plane of the inter-domain distances (Fig. 6(a)), with the PHD-PWWP distances from 46.6 Å to 65.8 Å and the PHD-Bromo distances from 41.4 Å to 58.9 Å. However, the cluster-MIMD achieves significantly broader sampling (Fig. 6(b)) than cMD, with the PHD-PWWP distances from 44.4 Å to 70.5 Å and the PHD-Bromo distances from 28.2 Å to 58.8 Å.

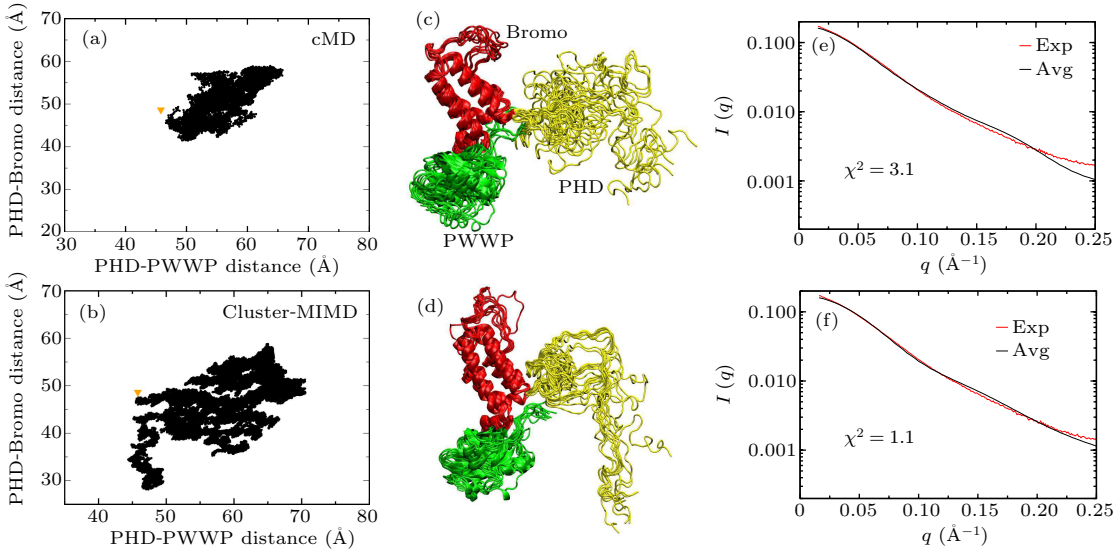


Fig. 6. Cluster-MIMD is in better agreement with the SAXS data of B3D than cMD. Projections of inter-domain distances in (a) cMD and (b) cluster-MIMD. The structural ensemble selected from (c) cMD and (d) cluster-MIMD using EOM. All conformers are superimposed on the Bromo domain (colored in red). The back-calculated SAXS profile of (e) the cMD ensemble and (f) the cluster-MIMD ensemble fitted to the experimental data (black).

EOM was performed on each MD trajectory to select an ensemble of 10 conformers to best fit the experimental SAXS data of B3D (Figs. 6(c) and 6(d)). Their simulated SAXS profiles are plotted (Figs. 6(e) and 6(f)) to compare with the experimental data. In the ensemble from cMD (Fig. 6(c)), the ten conformers are similar in that the PHD domain moves far from B2D. The simulated SAXS profile has a χ^2 of 3.1 to the experimental data (Fig. 6(e)). In the ensemble from cluster-MIMD (Fig. 6(d)), the PHD domain moves towards the Bromo domain, and the N-terminus of PHD is close to the PWWP do-

main. The simulated SAXS profile is in good agreement with the experimental data with a χ^2 of 1.1 (Fig. 6(f)). In summary, cluster-MIMD does a better job in reproducing the experimental data of B3D than cMD, in a relatively short simulation time.

3.3. Comparison with other sampling methods based on MIMD

Our sampling protocol is based on iterative MIMD simulations. There is a group of methods belonging to this category, such as PaCS-MD,[17–19] teDA2,[21] CoCo,[20] and PEX,[22] which are either targeted[17,21] or non-targeted.[18–20,22]

The original t-PaCS-MD[17] method calculates RMSD to a target conformation in each cycle and picks the first 10–20 conformations with the minimal RMSD values to start the next cycle. The teDA2 method[21] uses the minimal RMSD as a way of selecting conformations and applies principal component analysis (PCA) to reduce the dimensionality. We have performed t-PaCS-MD on AdK for 200 cycles (Figs. 7(a) and 7(b)). When starting from the open structure, the LID closes faster than the AMPbd. The protein spends most of the simulation time in sampling an area near the target closed structure (Fig. 7(a)). When starting from the closed structure, both the LID and the AMPbd can completely open but the former is faster. The protein is around the target open structure for the most of the simulation time (Fig. 7(b)). The results indicate that t-PaCS-MD may only sample a 'narrow' transition pathway between the initial and the target structure. In our protocol, a C2T simulation uses the target information as well. A cluster analysis is performed on conformations first, and then in each cluster, we pick the conformation closest to the target. This strategy may avoid aggregation of the starting conformations for the next cycle. Therefore, the C2T simulations can sample a broader area and obtain more intermediate structures (Figs. 3(d) and 3(h)) than the t-PaCS-MD simulations (Figs. 7(a) and 7(b)).

The nt-PaCS-MD method[18,19] does not rely on any target structure. It uses parameters like RMSD and backbone dihedral angles to identify those highly deviated structures for the next cycle of MIMD, so large conformational transitions could happen. The CoCo-MD[20] combines iterative MIMD simulations and complementary coordinates (CoCo) based on PCA. In each cycle, a boundary box in the PCA space is identified, and rarely sampled regions within the box are found. Then, the next cycle of MIMD simulations would start from the conformations in those rarely sampled regions. The frontier expansion sampling (FEXS) method[22] selects 'seed structures' distributed at the frontier of an already sampled region to start a new cycle of MIMD simulations by using the Gaussian mixture model and the convex hull algorithm. The CEN and F2I strategies are similar to the above methods in that the system tries to explore its conformational space without any target.

We conducted nt-PaCS-MD on AdK using a criterion called moving RMSD (mRMSD).[19] Technical details of this method have been introduced in supporting information. Starting from either the open or the closed structure, the results of nt-PaCS-MD (Figs. 7(c) and 7(d)) are fairly similar to those of cluster-MIMD (Fig. 3). The mRMSD-based nt-PaCS-MD shows higher efficiency in sampling domain motions of AdK than CEN (Figs. 3(b) and 3(f)) because the former always selects conformations with the largest mRMSD to start the next cycle. However, F2I (Figs. 3(c) and 3(g)) can generally sample larger domain motions than nt-PaCS-MD.

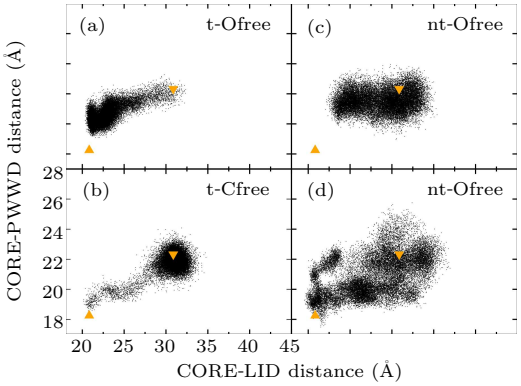


Fig. 7. Sampling efficiency of other methods measured by projecting protein conformations onto the 2D plane of inter-domain distances. (a) t-PaCS-MD starting from the open structure. (b) t-PaCS-MD starting from the closed structure. (c) nt-PaCS-MD starting from the open structure. (d) nt-PaCS-MD starting from the closed structure.

4. Conclusion and perspectives

In this work, we present a sampling protocol based on iterative MIMD guided by cluster analysis. The method is flexible and convenient to be implemented. First, there is no need to modify the simulation code, so any MD engine can be used straightforwardly. Different algorithms and metrics for clustering may be tried. In addition, the representative conformations can be selected with different strategies, depending on the requirements of the simulated system. If one wants to find transition paths between the initial and target conformations of a protein, the C2T strategy is used. Otherwise, the CEN or F2I strategy should be tried if one wants to explore the protein conformations without any target. These strategies have been found to be efficient for different cases. Our method can be combined with multiple independent umbrella sampling methods to calculate the free energy landscape.

Two multi-domain proteins, AdK and B3D, were selected to test the method. For AdK, with open and closed structures both available, the results demonstrate that cluster-MIMD is more efficient than cMD on the same time scale. The sampled intermediate states in the conformational transition pathways are consistent with previous work.[32] For B3D without a target, the protein conformations are extensively explored by using cluster-MIMD. The simulation can fit the experimental

SAXS data better than the cMD simulation. This protocol can be used for integrative structure modeling[51] by combing experimental data of proteins and their complexes.

Supporting information

Supporting information provides additional details of the nt-PaCS-MD method using moving RMSD and one figure illustrating the results of another set of cluster-MIMD on AdK.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (Grant No. 2021YFA1301504), Anhui University of Chinese Medicine 2024 Clinical Research Project (Grant No. 2024YFYLCZX26), the National Natural Science Foundation of China (Grant No. 91953101), and Chinese Academy of Sciences Strategic Priority Research Program (Grant No. XDB37040202). The Supercomputing Center of USTC provides computer resources for this project, and we are grateful to Mr. Yundong Zhang for his technical supports.

References

[1] Yon J M, Perahia D and Ghelis C 1998 *Biochimie* **80** 33
[2] Berendsen H J and Hayward S 2000 *Curr. Opin. Struct. Biol.* **10** 165
[3] Kern D and Zuiderweg E R 2003 *Curr. Opin. Struct. Biol.* **13** 748
[4] Lipfert J and Doniach S 2007 *Annu. Rev. Biophys Biomol. Struct.* **36** 307
[5] Joo C, Balci H, Ishitsuka Y, Buranachai C and Ha T 2008 *Annu. Rev. Biochem.* **77** 51
[6] Sinz A 2003 *J. Mass Spectrom* **38** 1225
[7] Karplus M and McCammon J A 2002 *Nat. Struct. Biol.* **9** 646
[8] Adcock S A and McCammon J A 2006 *Chem. Rev.* **106** 1589
[9] Dror R O, Dirks R M, Grossman J P, Xu H and Shaw D E 2012 *Annu. Rev. Biophys.* **41** 429
[10] Anderson J A, Lorenz C D and Travesset A 2008 *J. Comput. Phys.* **227** 5342
[11] Friedrichs M S, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg A L, Ensign D L, Bruns C M and Pande V S 2009 *J. Comput. Chem.* **30** 864
[12] Harvey M J, Giupponi G and Fabritiis G D 2009 *J. Chem. Theory Comput.* **5** 1632
[13] Jung J, Naurse A, Kobayashi C and Sugita Y 2016 *J. Chem. Theory Comput.* **12** 4947
[14] Bernardi R C, Melo M C R and Schulten K 2015 *BBA-Gen. Subjects* **1850** 872
[15] Wang A H, Zhang Z C and Li G H 2019 *Chinese Journal of Chemical Physics* **32** 277
[16] Tang Y, Yang Z, Yao Y, Zhou Y, Tan Y, Wang Z, Pan T, Xiong R, Sun J and Wei G 2024 *Chin. Phys. B* **33** 030701
[17] Harada R and Kitao A 2013 *J. Chem. Phys.* **139** 035103
[18] Harada R and Kitao A 2015 *J. Chem. Theory Comput.* **11** 5493
[19] Harada R, Sladek V and Shigeta Y 2019 *J. Chem. Theory Comput.* **15** 5144
[20] Shkurti A, Styliari I D, Balasubramanian V, Bethune I, Pedebos C, Jha S and Laughton C A 2019 *J. Chem. Theory Comput.* **15** 2587
[21] Yuan Y, Zhu Q, Song R, Ma J and Dong H 2020 *J. Chem. Theory Comput.* **16** 4631
[22] Zhang J and Gong H 2020 *J. Chem. Theory Comput.* **16** 4813
[23] Harada R and Shigeta Y 2018 *Journal of Molecular Graphics & Modelling* **85** 153
[24] Xu R and Wunsch D 2005 *IEEE Transactions on Neural Networks* **16** 645
[25] Muller C W and Schulz G E 1992 *J. Mol. Biol.* **224** 159
[26] Muller C W, Schlauderer G J, Reinstein J and Schulz G E 1996 *Structure* **4** 147
[27] Wang J, Qin S, Li F, Li S, Zhang W, Peng J, Zhang Z, Gong Q, Wu J and Shi Y 2014 *Cell Research* **24** 890
[28] Harada R and Shigeta Y 2018 *Molecular Simulation* **44** 206
[29] Jain A K 2010 *Pattern Recognition Letters* **31** 651
[30] Arora K and Brooks C L I 2007 *Proc. Natl. Acad. Sci. USA* **104** 18496
[31] Ping J, Hao P, Li Y X and Wang J F 2013 *Biomed Research International* **2013** 628536
[32] Li D, Liu M S and Ji B 2015 *Biophysical Journal* **109** 647
[33] Zheng Y and Cui Q 2018 *J. Chem. Theory Comput.* **14** 1716
[34] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B and Lindahl E 2015 *SoftwareX* **1–2** 19
[35] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot B L, Grubmüller H and MacKerell A D 2017 *Nature Methods* **14** 71
[36] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 *J. Chem. Phys.* **79** 926
[37] Bussi G, Donadio D and Parrinello M 2007 *J. Chem. Phys.* **126** 014101
[38] Hess B 2008 *J. Chem. Theory Comput.* **4** 116
[39] Essmann U, Perera L, Berkowitz M L, Darden T, Lee H and Pedersen L G 1995 *J. Chem. Phys.* **103** 8577
[40] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 *J. Comput. Chem.* **13** 1011
[41] Hateboer G, Gennissen A, Ramos Y F, Kerkhoven R M, Sonntag-Buck V, Stunnenberg H G and Bernards R 1995 *EMBO J.* **14** 3159
[42] Eswar N, Eramian D, Webb B, Shen M Y and Sali A 2008 *Methods in molecular biology* (Clifton, N.J.) **426** 145
[43] Petoukhov M V and Svergun D I 2005 *Biophysical Journal* **89** 1237
[44] Franke D, Petoukhov M V, Konarev P V, Panjkovich A, Tuukkanen A, Mertens H D T, Kikhney A G, Hajizadeh N R, Franklin J M, Jeffries C M and Svergun D I 2017 *Journal of Applied Crystallography* **50** 1212
[45] Robustelli P, Piana S and Shaw D E 2018 *Proc. Natl. Acad. Sci. USA* **115** E4758
[46] Tria G, Mertens H D T, Kachala M and Svergun D I 2015 *Iucrj* **2** 207
[47] Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser E Z and Kern D 2004 *Nature Structural & Molecular Biology* **11** 945
[48] Maragakis P and Karplus M 2005 *Journal of Molecular Biology* **352** 807
[49] Henzler-Wildman K A, Thai V, Lei M, Ott M, Wolf-Watz M, Fenn T, Pozharski E, Wilson M A, Petsko G A, Karplus M, Hübner C G and Kern D 2007 *Nature* **450** 838
[50] Grossfield A http://membrane.urmc.rochester.edu/wordpress/?page_id=126
[51] Braitbard M, Schneidman-Duhovny D and Kalisman N 2019 *Annu. Rev. Biochem.* **88** 113