

Identification of key residues in protein functional movements by using molecular dynamics simulations combined with a perturbation-response scanning method

Jun-Bao Ma(马君宝), Wei-Bu Wang(王韦卜), and Ji-Guo Su(苏计国)
Citation: Chin. Phys. B, 2021, 30 (**10**): 108701. DOI: 10.1088/1674-1056/abf12f
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

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

Modeling hydrogen exchange of proteins by a multiscale method

Wentao Zhu(祝文涛), Wenfei Li(李文飞), and Wei Wang(王炜)
Chin. Phys. B, 2021, 30 (**7**): 078701. DOI: 10.1088/1674-1056/abe377

Multi-scale molecular dynamics simulations and applications on mechanosensitive proteins of integrins

Shouqin Lü(吕守芹), Qihan Ding(丁奇寒), Mingkun Zhang(张明焜), and Mian Long(龙勉)
Chin. Phys. B, 2021, 30 (**3**): 038701. DOI: 10.1088/1674-1056/abc540

Folding nucleus and unfolding dynamics of protein 2GB1

Xuefeng Wei(韦学锋) and Yanting Wang(王延颋)
Chin. Phys. B, 2021, 30 (**2**): 028703. DOI: 10.1088/1674-1056/abbbfa

Effect of C_{60} nanoparticles on elasticity of small unilamellar vesicles composed of DPPC bilayers

Tanlin Wei(魏坦琳), Lei Zhang(张蕾), Yong Zhang(张勇)
Chin. Phys. B, 2020, 29 (**4**): 048702. DOI: 10.1088/1674-1056/ab7d9f

Structural and thermal stabilities of Au@Ag core-shell nanoparticles and their arrays: A molecular dynamics simulation

Hai-Hong Jia(贾海洪), De-Liang Bao(包德亮), Yu-Yang Zhang(张余洋), Shi-Xuan Du(杜世萱)
Chin. Phys. B, 2020, 29 (**4**): 048701. DOI: 10.1088/1674-1056/ab7da9

Identification of key residues in protein functional movements by using molecular dynamics simulations combined with a perturbation-response scanning method

Jun-Bao Ma(马君宝), Wei-Bu Wang(王韦卜), and Ji-Guo Su(苏计国)†

Key Laboratory for Microstructural Material Physics of Hebei Province, School of Science, Yanshan University, Qinhuangdao 066004, *China*

(Received 6 January 2021; revised manuscript received 25 February 2021; accepted manuscript online 24 March 2021)

The realization of protein functional movement is usually accompanied by specific conformational changes, and there exist some key residues that mediate and control the functional motions of proteins in the allosteric process. In the present work, the perturbation-response scanning method developed by our group was combined with the molecular dynamics (MD) simulation to identify the key residues controlling the functional movement of proteins. In our method, a physical quantity that is directly related to protein specific function was introduced, and then based on the MD simulation trajectories, the perturbation-response scanning method was used to identify the key residues for functional motions, in which the residues that highly correlated with the fluctuation of the function-related quantity were identified as the key residues controlling the specific functional motions of the protein. Two protein systems, i.e., the heat shock protein 70 and glutamine binding protein, were selected as case studies to validate the effectiveness of our method. Our calculated results are in good agreement with the experimental results. The location of the key residues in the two proteins are similar, indicating the similar mechanisms behind the performance of their biological functions.

Keywords: protein functional movements, molecular dynamics simulations, perturbation-response scanning method

PACS: 87.15.ad, 87.14.E–, 87.15.hp **DOI:** 10.1088/1674-1056/abf12f

1. Introduction

In the field of structural biology, it is generally accepted that the sequence and structure of proteins determine the functions they can achieve, and the performance of protein functions usually involves specific conformational movements. Therefore, it is of great significance to study the functional movements encoded in protein structures.[1–5] In the process of protein conformational motions, there may exist some functionally important sites together constituting the key residue network, which mediates and controls the functional motions in proteins. Finding these key residues in the tertiary structure of proteins is of great help in understanding the molecular mechanism behind the related physiological processes. These key residues could also be used as the candidate target sites for future drug design as well as for the exploring of enzyme activity.[1,6–8] Due to the less labor cost and lower time consuming compared with the experimental approaches, the development of effective computational methods to identify the key residues for functional motions attracts great interest of researchers.

Molecular dynamics (MD) simulation has long been proven to be an effective method to investigate the conformational dynamics in proteins.[1,9–16] In MD simulations, given the initial state of the protein, the classical mechanics is numerically calculated to get the motion trajectory of the protein at the atomic level for a certain time. It is convenient for us to observe the protein conformational motions in detail. Based on the MD simulation trajectories, the analysis of the dynamic cross-correlations between residues is a frequently used method to reveal the functional correlated motions and identify the related key sites in proteins. McCammon employed the dynamic cross-correlation analysis to investigate the coupling between the local and collective motions, which controls the conformational transformations associated with the protein biological functions.[17] Kamberaj and van der Vaart applied MD simulation combined with the inter-residue cross-correlation analysis to detect the key interactions responsible for the DNA-induced unfolding of the inhibitory helix 1 in Ets-1 molecule.[18] Kasahara *et al*. proposed a multi-modal dynamic cross correlation analysis method, in which the multi-modal dynamic properties of protein atoms were captured, to explore the correlation network and key sites in the Est1-DNA complex.[19] However, in protein collective motions, the residues in well-packed domains usually move as a whole, and thus all the residues within a domain exhibit strong positive dynamic cross-correlations and the residues between different domains often exhibit negative cross-correlations. It is difficult to pick out the functionally key sites from these numerous highly correlated residues. The conventional cross-correlation analysis method only considers the residue fluctuations deviated from their equilibrium positions, but the energy communications within proteins are directly related to the corre-

†Corresponding author. E-mail: jiguosu@ysu.edu.cn

© 2021 Chinese Physical Society and IOP Publishing Ltd

lations between fluctuations of the distance between pairwise residues. Based on the theory of physics, the inter-residue potential energy is a function of the distance between them, and thus the fluctuations in the distance between pairwise residues are associated with the energy changes. Erman *et al*. have pointed out that although the residues in the same domain have high positive correlations, the energy or signal transmission in protein allosteric movements is directly related to the changes in the distance of residue pairs.[20] Therefore, the conventional cross-correlation analysis method is not very effective in identifying the key residues that energetically control the functional motions of proteins, and new analysis methods should be developed, in which the fluctuations in the distance of residue pairs should be taken into account.

Besides MD simulations, the normal mode analysis (NMA) of the elastic network model (ENM) is another simple and effective theoretical method in investigating the intrinsic dynamics encoded in protein systems. ENM has been extensively applied in revealing protein large-scale functional motions,[21] flexible docking,[22] flexible fitting of cryo-electron microscopy maps,[23] enhanced sampling in MD simulations.[24] Based on ENM, several perturbation-response scanning methods have been proposed to identify the key residues for functional motions in proteins. Atilgan *et al*. predicted the key residues by observing the conformational changes of the whole protein when an external force is applied to a certain region in the system.[25] Zheng *et al.* identified the key residues by analyzing the fluctuation amplitude of residues, the motion pattern of the system and the changes of the corresponding frequency when the force constant of the inter-residue interactions is perturbed.[26–28] Ming *et al.* determined the functional sites of the protein by looking for the perturbation sites that can significantly change the distribution of the protein conformational ensemble.[29,30] Our group developed an ENM-based perturbation-response scanning method to identify the key residues controlling protein specific functions.[1,31–34] In our method, a physical quantity directly related to the specific function of proteins was introduced, and the residues whose perturbations induce significant changes in the mean-square fluctuations of the predefined function-related physical quantity were identified as the functionally key sites. The essence of our method is to evaluate the cross-correlation between the fluctuation of the function-related quantity with the fluctuation of the pairwise residue distance.

In the present work, our perturbation-response scanning method based on the function-related quantity was introduced into MD simulations to identify the key sites in protein functional motions. In this approach, the motion trajectories of the protein system were firstly simulated for a certain period of time by using all-atomic MD simulations. Based on the MD simulation trajectories, a number of conformational snapshots were extracted and the average structure of these sampled conformations were calculated. Then, a physical quantity that is directly related to protein specific function was introduced to quantitatively describe different functional states of the system. The construction of function-related quantity is quite flexible as discussed in our previous paper.[34] In this study, the centroid distance between the two domains situating at opposite sides of the ligand-binding pocket in the studied proteins was selected as the function-related quantity. According to our perturbation-response scanning method, all the residues in the protein system were then perturbed one by one, and the cross-correlation between the fluctuation of the predefined functional quantity with the fluctuation in the distance of the perturbation site involved residue pairs was calculated. The fluctuation of the functional quantity as well as the fluctuation of the pairwise residue distance can be computed based on the conformational snapshots and the average structure of MD simulations. The residues that highly correlated with the fluctuation of the function-related quantity were identified as the key residues controlling the specific functional motions of the protein. The presented method detected the fluctuations in the distance between pairwise residues, which essentially reflect the inter-residue energy exchanges in the protein system, instead of the fluctuations of the individual residues. Compared with the conventional cross-correlation analysis method, our method can better reveal the key residues energetically controlling protein functional motions. In this work, two proteins, i.e., the heat shock protein 70 (Hsp 70) and glutamine binding protein (GlnBP), were studied as examples to verify the effectiveness of this method. Both of these two proteins undergo obvious open-closed movements when performing their biological functions. The proposed method was used to identify the key residues controlling the functionally conformational transformations.

2. Materials and methods

2.1. The molecular dynamics simulation

The coordinate files of the studied proteins, i.e., Hsp 70 and GlnBP, downloaded from protein data bank (PDB) with accession codes 3c7n and 1ggg, respectively, were assigned as the initial structures for MD simulations. The MD simulations of the protein systems were carried out by using GROMACS 5.1.4 software with CHARMM27 force field. The studied proteins were solvated with SPC water model in the center of the cubic boxes, and the shortest distance between the proteins and the edge of the boxes were set to be 1.6 nm. The counter-ions were then added into the water boxes to make the system to be neutralized. The protein system was firstly minimized by the steepest descent method to remove the bad contacts between

atoms. After that, 500 ps isothermal-isovolumetric (NTV) and 5 ns isothermal-isobaric (NTP) MD simulations with position constraints were performed. Finally, the position constraints were released, and 100 ns MD equilibrium simulations were carried out to obtain the trajectories. All simulations were performed at 1 atm and 300 K. The step size of 2 fs was used in the simulations.

In order to determine whether the simulated proteins arrive at the equilibrium states, the time-evolution of the root mean square deviation (RMSD) of the proteins was calculated. The RMSD calculation is expressed as follows:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{r}_i - \boldsymbol{R}_i)^2}, \tag{1}$$

where N is the number of atoms in the protein, $\boldsymbol{r}_i$ is the coordinate vector of the i-th atom in the instantaneous conformational frame, and $\boldsymbol{R}_i$ is the corresponding coordinate vector in the reference conformation.

When the simulation time is long enough, the ensemble average can be calculated based on the time average according to the ergodic hypothesis. For the simulated trajectories, all the frames were extracted after the system arriving at the equilibrium state with small RMSD fluctuations. These frames were superposed to obtain the average conformation by using the "trjconv" command in GROMACS. Then based on the instantaneous frames and the average conformation, the dynamic cross-correlation between the fluctuations of the i-th and j-th residues can be calculated according to the following formula:[18]

$$\mathrm{DCC}(i,j) = \frac{\langle(\boldsymbol{r}_i(t) - \bar{\boldsymbol{r}}_i)\cdot(\boldsymbol{r}_j(t) - \bar{\boldsymbol{r}}_j)\rangle_t}{\sqrt{\langle(\boldsymbol{r}_i(t) - \bar{\boldsymbol{r}}_i)^2\rangle_t \langle(\boldsymbol{r}_j(t) - \bar{\boldsymbol{r}}_j)^2\rangle_t}}, \tag{2}$$

where $\boldsymbol{r}_i(t)$ and $\boldsymbol{r}_j(t)$ are the coordinate vectors of the i-th and j-th atoms in the instantaneous frame, respectively; $\bar{\boldsymbol{r}}_i$ and $\bar{\boldsymbol{r}}_j$ are the coordinate vectors of the i-th and j-th atoms in the average conformation, respectively, and $\langle\cdot\rangle_t$ denotes the time average.

2.2. The perturbation-response scanning method combined with molecular dynamics simulations to identify the key residues for functional motions

In order to identify the key residues involved in the specific function of the protein, a physical quantity directly related to the specific function was firstly introduced. There are various choices to construct the function-related quantity, such as the geometric shape of the protein, the volume of ligand-binding pocket, the thermodynamic properties of the system, the inter-residue interaction distributions, or other quantities directly related to protein specific functions. In the present work, the studied proteins, i.e., Hsp 70 and GlnBP, perform their functions through the open-closed conformational transition of the ligand binding pocket. Therefore, a physical quantity that is able to quantitatively describe the opening and closing of the binding pocket was constructed. For these two studied proteins, the centroid distance between the two domains on the opposite sides of the binding pocket was chosen as the function-related quantity.

Using the function-related quantity, a perturbation-response scanning method has been proposed based on the ENM in our previous studies.[34] ENM simplifies the inter-residue interactions into harmonic potential, and thus the protein structure is represented by a network of springs. In our perturbation-response scanning method, each residue in the protein was perturbed, and the changes in the mean-square fluctuation (MSF) of the predefined function-related physical quantity were calculated. Supposing the function-related quantity is denoted as ξ, the change in the MSF of ξ in response to the residue perturbation can be computed by using the linear response theory and expressed as[34]

$$\begin{aligned}&\frac{\partial\langle(\Delta\xi)^2\rangle}{\partial\gamma_{ij}}\\ &= -\frac{1}{2k_{\mathrm{B}}T}\left[\langle(\Delta\xi)^2(R_{ij}-R_{ij}^0)\rangle - \langle(\Delta\xi)^2\rangle\langle(R_{ij}-R_{ij}^0)^2\rangle\right]\\ &= -\frac{1}{k_{\mathrm{B}}T}\left[\langle\Delta\xi\Delta R_{ij}\rangle\right]^2,\end{aligned} \tag{3}$$

where γ_{ij} is the force constant between the i-th and j-th residues, k_{B} is the Boltzmann constant, T is the absolute temperature, and R_{ij} and R_{ij}^0 are the instantaneous and equilibrium distance between the i-th and j-th residues, respectively. Equation (3) only accounts for the perturbation of the interaction between the i-th and j-th residue pair. If a residue in the protein was perturbed, the responses to all the pairwise interactions involving the perturbed residue were calculated one by one using Eq. (3) and then summed up to measure the effects of the residue perturbation.

In this study, our perturbation-response scanning method based on the function-related quantity was combined with MD simulations to identify the key sites in protein functional motions. The cross-correlation between the fluctuation of the functional quantity and the fluctuation of the pairwise residue distance, i.e., $\langle\Delta\xi\Delta R_{ij}\rangle$ in Eq. (3), was calculated with time-average approach based on the equilibrium MD simulation trajectories according to Eq. (2). The average conformation and the conformation at each moment used for calculation are obtained by MD simulation. Then based on these MD simulation frames and the average conformation, $\Delta\xi$, ΔR_{ij} and $\langle\Delta\xi\Delta R_{ij}\rangle$ were computed. $\Delta\xi = \xi - \xi'$, where ξ is the C_α centroid distance between domains in each instantaneous conformation, and ξ' is the C_α centroid distance between domains in the average conformation. $\Delta R_{ij} = R_{ij} - R_{ij}^0$, where R_{ij} is the distance between the C_α atom of the i-th and j-th interacting residues in each instantaneous conformation, and R_{ij}^0 is the C_α atom dis-

tance between the i-th and j-th interacting residues in the average conformation. In this work, whether there exist interactions between two residues is determined by a cutoff distance (8.0 Å is adopted). If the C_α distance between two residues in the average conformation is less than the cutoff value, they are considered to be interacted, and otherwise they are not interacted. We have utilized 7.0 Å and 9.0 Å for the same calculations, and it is found that the cutoff value has little effects on the calculation results (data no shown).

From Eq. (3), it is found that our method evaluates the cross-correlation between the fluctuation of the function-related quantity with the fluctuations in the distance between pairwise residues, which essentially reflect the inter-residue energy exchanges in the protein system. Compared with the conventional cross-correlation analysis method, our method can better reveal the key residues energetically controlling protein functional motions. It should be pointed out that Eq. (3) is derived under the assumption that the inter-residue interactions are simplified as harmonic potentials without any frustration. Many theoretical and simulation studies have shown that the large-scale collective motions in proteins can be well described by the low frequency normal modes of this simple harmonic model.[35–38] In the present study, we only focused on the large-scale functional motions encoded in protein structure, and the identified key sites are those essential for the collective motions. It is reasonably speculated that the anharmonic interactions may have minor impacts on the calculation results.

3. Results and discussion

3.1. The functionally key sites in the nucleotide binding domain of heat shock protein 70

Studies have shown that heat shock protein (Hsp) can improve the heat resistance of cells, and when combined with other protein molecules, it carries out the functions such as preventing misfolding, repairing and degradation of amino acid chains.[39–41] Hsp 70 is composed of the substrate binding domain (SBD) and nucleotide binding domain (NBD). The hydrolysis of ATP at NBD causes the conformation changes from the open state to the closed state, and then the information is transmitted to the SBD to trigger the release of the substrate.[1,42–45] Hsp is ubiquitous in the biological world, and the study of its operating mechanism, especially the ATP regulation mechanism, is of great significance to our understanding of the functional movements of the protein.

The tertiary structure of the NBD of Hsp 70 was downloaded from the PDB with accession code 3c7n. Hsp 70 NBD consists of four subdomains, IA, IB, IIA and IIB, as shown in Fig. 1(a). It can be clearly seen that these four subdomains surround around the nucleotide binding pocket. The open-closed motions of Hsp 70 NBD play important roles in the binding and hydrolysis of the ligand ATP. During the performance of the biological function of Hsp 70 NBD, the four subdomains, especially the IB and IIB subdomains, undergo distinct open-closed conformational movements. In this study, the C_α centroid distance of the subdomains IB and IIB was defined as the physical quantity directly related to the function of Hsp 70 NBD. Then our functional-quantity-based perturbation-response scanning method was combined with MD simulations to predict the key residues controlling the open-closed motions of Hsp 70 NBD.

The 100 ns motion trajectory of Hsp 70 NBD was obtained by using MD simulations. It is found that the motion trajectory is in a relatively stable state after 20 ns by observing the RMSD curve, as shown in Fig. 1(c). Therefore, the last 80 ns of the simulation trajectory was used in our calculation. From the simulation trajectory, a total of 1600 conformational frames were extracted, and these frames were superposed to obtain the average conformation. Then, all the residues in the protein were perturbed one by one, and the change in the fluctuation of the MSF of the functional quantity, i.e., the C_α centroid distance of the subdomains IB and IIB, in response to the residue perturbation was calculated with our perturbation-response scanning method combined with MD simulations. Based on the frames extracted from MD simulation trajectory, as well as the average conformation, the cross-correlation between the fluctuation of the functional quantity and the fluctuation of the pairwise residue distance was calculated according to Eqs. (2) and (3). The calculation result are shown in Fig. 1(d).

From Fig. 1(d), it is found that there are 19 residues with the absolute value of $\partial\langle(\Delta\xi)^2\rangle/\partial\gamma_{ij}$ larger than 0.01, which are identified as the key residues for functional motions marked by the number of 1–19 in the figure. These key residues are Gly4, Pro5, Gly12, Thr13, Asn57, Ala60, Met61, Lys71, Val139, Glu192, Gly201, Gly202, Gly203, Asp214, Gly230, Lys257, Arg258, Arg261 and Ser381, which are numbered by 1–19, respectively, in Figs. 1(b) and 1(d). The positions of these predicted key residues in the conformation are shown in Fig. 1(b). According to the position of these predicted key residues in the conformation and their roles in the open-closed movement of the protein, theses key residues can be divided into four categories.

(1) Residues Asn57, Ala60, Met61, Lys257, Arg258 and Arg261 are located at the opening of the ligand binding pocket. These residues are considered to be the key residues controlling the relative movements between the IB and IIB subdomains, which is important for the binding of the ligand ATP as well as the release of the hydrolysis products. Woo *et al.* mutated the residue Arg261 and found that the mutation obviously contributes to the open-closed conformational transition, as well as the binding and dissociation of the ligand ATP.[46]

Ung *et al.* found that the alpha helix (residues 230–280) of IIB subdomain also bend around the axis at the residues 262–267.[47]

(2) Residues Gly12, Thr13, Lys71, Gly201, Gly202, Gly203 and Gly230 are located at the substrate binding site. These residues are considered to be directly related to the substrate binding and the associated signal transduction. These residues are located at the interface between the IA and IB subdomains as well as that between the IIA and IIB subdomains, which regulate the conformational movements between these subdomains. Mayer *et al.* found that the residues located in this region obviously participate in the binding of ATP and affect the signal transmission.[48] Ung *et al.* mutated residues such as Gly230 and found that the mutation significantly affect the chaperone activities of Hsp 70.[47]

(3) Residues Glu192 and Asp214 are located at the hinge connecting the subdomains IA and IIA. During the open-closed movement of Hsp 70 NBD, these residues control the overall conformational transition of the subdomains and provide a pivot for the opening and closing of the pocket. Stetz *et al.* found that a few residues in this region are highly co-evolved, and these residues participate in the stimulation of ATPase activity.[49]

(4) Residues Gly4, Pro5, Val139 and Ser381 are located at the hinge region for the internal movements of the IA domain, which mediate the relative motions between different regions within the IA domain. This region is also directly connected to the SBD in Hsp 70, and therefore the residues in this region mediate the allosteric signal transduction between NBD and SBD.[50]

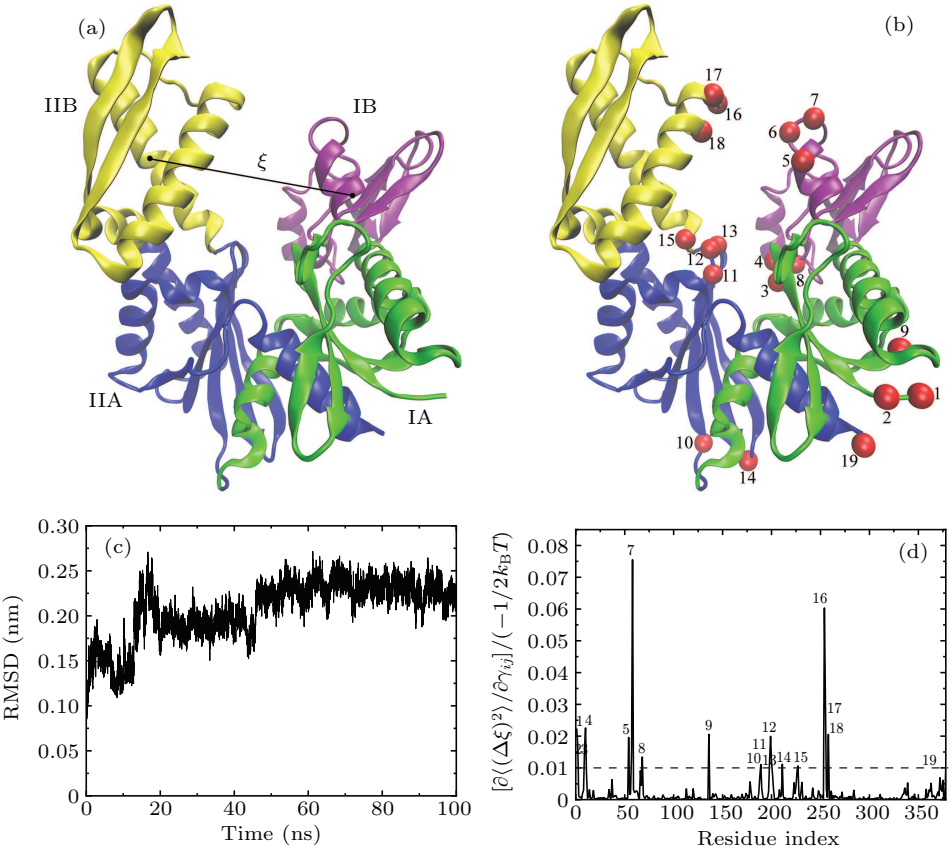


Fig. 1. (a) The tertiary structure of Hsp 70 NBD, which is composed of four subdomains, i.e., IA (green), IB (purple), IIA (blue) and IIB (yellow). The C_α centroid distance of the subdomains IB and IIB was defined as the physical quantity directly related to the function of Hsp 70 NBD, which is represented by ξ in the figure. (b) The location of the 19 key residues for functional motions in the structure of Hsp 70 NBD identified by our method. (c) Time-evolution of the RMSD for the C_α atoms of Hsp 70 NBD during the MD simulation of the system. (d) The values of $\partial\langle(\Delta\xi)^2\rangle/\partial\gamma_{ij}$ in response to each residue perturbation. The key residues for functional motions were identified as the sites whose $\partial\langle(\Delta\xi)^2\rangle/\partial\gamma_{ij}$ values are larger than 0.01. A total of 19 key sites were identified, which are marked by the numbers of 1–19 as shown in the figure. These residues are Gly4, Pro5, Gly12, Thr13, Asn57, Ala60, Met61, Lys71, Val139, Glu192, Gly201, Gly202, Gly203, Asp214, Gly230, Lys257, Arg258, Arg261 and Ser381, respectively.

3.2. The functionally key sites in glutamine binding protein

GlnBP is ubiquitous in organisms, which belongs to the E.coli permease system. It is a specific binding protein recognizing extracellular fluid substrates, and its main role is to transport glutamine (Gln) into the cell. Upon the binding of Gln, GlnBP closes and wraps Gln to form a complex, which can be recognized by the permease system and finally completes the task of Gln transporting.[51–53] GlnBP undergoes significant conformational changes during this process. In this study, the tertiary structure of GlnBP was downloaded from the PDB with accession code 1ggg. GlnBP consists of two

subdomains named as A and B, as shown in Fig. 2(a). During the binding of Gln to GlnBP, subdomains A and B undergo obvious open-closed movements. In the present work, the C_α centroid distance of subdomains A and B was defined as the physical quantity directly related to the function of GlnBP, and then the functional-quantity-based perturbation-response scanning method was combined with MD simulations to predict the key residues involved in the openclosed movements of GlnBP.

The 100 ns equilibrium motion trajectory was sampled by using MD simulation from the initial structure of GlnBP. It is found that the protein system gradually arrives at the equilibrium state after 10ns by observing the RMSD curve as shown in Fig. 2(c). Thus, the last 90 ns of the simulation trajectory was used in our calculations. A total of 1800 conformational frames were obtained from the simulation trajectory, and the average conformation was obtained by superposing these conformational frames. The C_α centroid distance of the subdomains A and B was used as the physical quantity directly related to the functional movements. All the residues in the protein were perturbed one by one, and the change in the MSF of the functional quantity due to the perturbation was evaluated. Based on the conformational frames extracted from the MD simulation, the cross-correlation between the fluctuation of the functional quantity and the fluctuation of the distance of residue pairs was calculated. The calculation result is shown in Fig. 2(d).

It can be seen from Fig. 2(d) that there are 17 residues with the value of $\partial\langle(\Delta\xi)^2\rangle/\partial\gamma_{ij}$ larger than 0.015, which are marked by 1–17. These residues are identified as the key residues directly related to the functional movements of the protein. These key residues are Thr11, Ala12, Pro47, Met48, Asp49, Ala67, Val101, Val104, Lys105, Leu107, Asp108, Lys129, Asp152, Tyr217, Lys218, Thr223 and Glu224, respectively, as shown in Figs. 2(b) and 2(d). These 17 residues can be grouped into three clusters according to their location on the conformation of the protein as well as their roles in the functional movements of the protein.

(1) Residues Thr11, Ala12, Ala67, Tyr217, Lys218, Thr223, and Glu224 are located at the opening of the ligand binding pocket. These residues are considered to be the key residues related to the relative movements between the A and B subdomains, and these residues are responsible for the opening and closing of the ligand binding pocket. The experimental results of Sun *et al.* showed that the residues near Ala67 are responsible for the binding of the ligand, which is consistent with our prediction.[54] Several other experimental results also showed that the conformational transition between the open and closed states of the ligand-binding pocket is related to the breakage of the hydrogen bond and the changes of the distance between these key residues.[55–57]

(2) Residues Pro47, Met48 and Asp49 are situated at the hinge region responsible for the relative motion between different parts within the subdomain A. During the open-closed conformational transition of the protein, the tip region of the subdomain A performs more obvious movements compared with the remaining part. These key residues control the relative large motions for the tip region of the subdomain A.

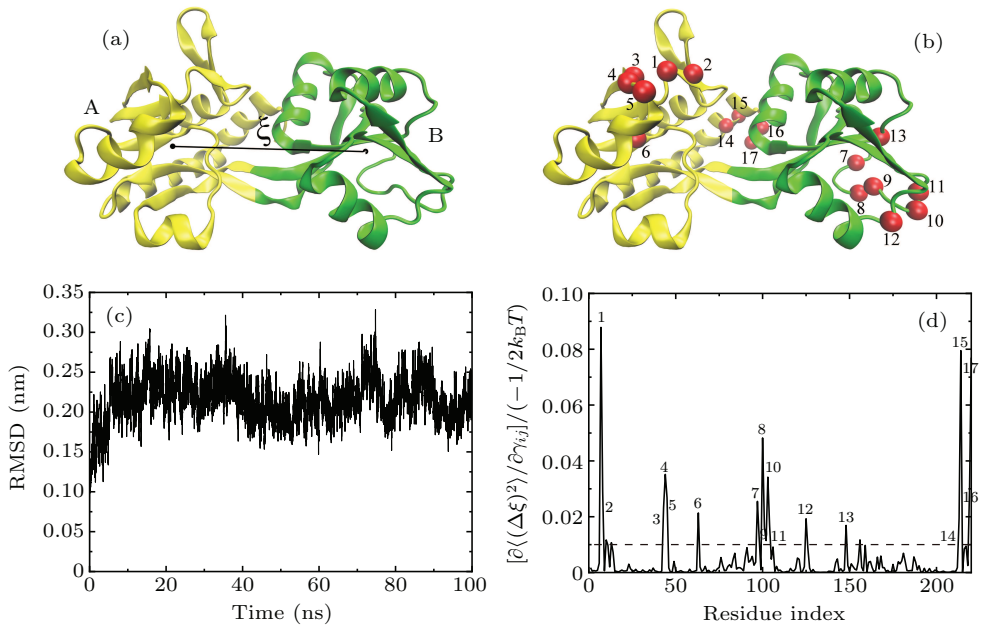


Fig. 2. (a) The tertiary structure of GlnBP, which is composed of two subdomains, i.e., A (yellow) and B (green). The C_α centroid distance of subdomains A and B is defined as the functional quantity directly related to the openclosed of GlnBP, which is expressed by ξ in the figure. (b) The positions of the 17 key residues related to the open-closed movements of GlnBP predicted by our method. (c) Time-evolution of the RMSD for the C_α atoms of GlnBP during the MD simulation of the system. (d) The value of $\partial\langle(\Delta\xi)^2\rangle/\partial\gamma_{ij}$ after disturbing each residue one by one. The sites whose value of $\partial\langle(\Delta\xi)^2\rangle/\partial\gamma_{ij}$ are larger than 0.015 were identified as functional key residues. A total of 17 key sites were identified, which are numbered 1–17 as shown in the figure. They are Thr11, Ala12, Pro47, Met48, Asp49, Ala67, Val101, Val104, Lys105, Leu107, Asp108, Lys129, Asp152, Tyr217, Lys218, Thr223 and Glu224, respectively.

(3) Residues Val101, Val104, Lys105, Leu107, Asp108, Lys129 and Asp152 are located in the hinge region inside subdomain B. This subdomain mainly contains 3 alpha helices and 5 beta sheets. These residues mainly mediate the relative movements between different regions within B subdomain.

3.3. Summary of the location of the predicted key residues for the two studied proteins

Hsp 70 NBD and GlnBP studied in this work undergo the similar overall conformational motions between the subdomains in these systems when performing their biological functions. The functional motions for these two protein systems are mainly manifested as the open-closed movements of the domains around the ligand binding pocket, and the key residues predicted by our method are located at the similar regions in the structure of the proteins. The locations of these predicted key residues are summarized as follows:

(1) The opening of the pocket. When these two proteins achieve their biological functions, the subdomains around the ligand binding pocket exhibit obvious relative movements. Residues at this region control the open-closed movements of the pocket, which is very important for the binding of ligands. (2) The hinge region connecting different subdomains. These residues regulate the relative overall movements of the subdomains and serve as the pivot for the relative motions of the subdomains. (3) The substrate binding site. These key residues are related to substrate binding and signal transmission. These residues control the binding and release of the ligands, as well as the energy or signal generation and transmission from the ligand binding site to the remote regions of the proteins. (4) The hinge region within the subdomains. These residues are responsible for the relative motions between different regions within the subdomain. It should be noted that in this study, 0.010 and 0.015 were used as the threshold values to determine the key residues for Hsp 70 and GlnBP. In fact, there is no a uniform quantitative threshold to exactly determine which residue is essential and which is not. In our method, the larger the calculated value, the more important the residue for the protein specific function. Although the number of the predicted key residues varied for different threshold values, the regions where the key residues are located are basically unchanged (data not shown).

4. Conclusion

In this study, our perturbation-response scanning method was combined with MD simulations to identify the key residues responsible for the conformational movements of proteins. In our method, a physical quantity which is directly related to the function of the protein was introduced. By perturbing each residue in the protein, the cross-correlation between the fluctuation of the pre-defined function-related quantity and the fluctuation in the distance of residue pairs was calculated, and the key residues were identified as those highly correlated with the fluctuation of the function-related quantity. Two protein systems, i.e., Hsp 70 NBD and GlnBP, were investigated by using our method to identify the key residues related to the openclosed movements of the protein pockets.

For the two studied proteins, the key residues predicted by our method are located at the similar regions, which can be grouped into the following locations: (1) The opening of the pocket. (2) The hinge region connecting different subdomains in the protein structure. (3) The substrate binding site. (4) The hinge region responsible for the relative motion between different parts within the subdomains of the proteins. From the view of physics, the functional motions in proteins generally involve large-scale inter-domain movements. The residues within a domain usually move together with positive correlations, whereas the residues between different domains generally exhibit negative motion correlations. Thus, the hinge region between different domains usually undergoes negative correlated movements along with the functional motions of proteins, which cause large fluctuations of the inter-residues distance. The key residues identified by our method are consistent with the biological observations.

Our method essentially reflect the inter-residue energy exchanges in the protein system. Compared with the conventional cross-correlation analysis method, our method can better reveal the key residues energetically controlling protein functional motions. In this study, two protein systems, i.e., Hsp 70 and GlnBP, were used as case studies to test the validity of the proposed method. It needs further investigation to assess the effectiveness of our method by using more protein systems with proper statistical analysis method, which will be addressed in the following studies. In the present work, the C_α centroid distance between domains was used as the physical quantity directly related to protein specific function. It should be noted that the choice of the function-related quantity is quite flexible in our method, and other functional quantities will be tested in the future studies.

References

[1] Su J G, Xu X J, Li C H, Chen W Z and Wang C X 2011 *J. Chem. Phys.* **135** 174101
[2] Monod J, Changeux J P and Jacob F 1963 *J. Mol. Biol.* **6** 306
[3] Perkins J R, Diboun I, Dessailly B H and Lees J G and Orengo C 2010 *Structure.* **18** 1233
[4] Tsai C J, Del Sol A and Nussinov R 2009 *Mol. Biosyst.* **5** 207
[5] Wei X and Wang Y 2021 *Chin. Phys. B* **30** 028703
[6] Teilum K, Olsen J G and Kragelund B B 2009 *Cell. Mol. Life Sci.* **66** 2231
[7] Doshi U, Holliday M J, Eisenmesser E Z and Hamelberg D 2016 *Proc. Natl. Acad. Sci. USA* **113** 4735
[8] Zhou H, Dong Z and Tao P 2018 *J. Comput. Chem.* **39** 1481
[9] Ota N and Agard D A 2005 *J. Mol. Biol.* **351** 345
[10] Sharp K and Skinner J J 2006 *Proteins* **65** 347
[11] Kong Y and Karplus M 2009 *Proteins* **74** 145

[12] Fornili A, Giabbai B, Garau G and Degano M 2010 *J. Am. Chem. Soc.* **132** 17570
[13] Chan C, Wen H, Lu L and Fan J 2015 *Chin. Phys. B* **25** 018707
[14] Shao D and Gao K 2018 *Chin. Phys. B* **27** 018701
[15] Zhang C and Zhou X 2020 *Chin. Phys. B* **29** 108706
[16] Zhang X, Guo Z, Yu P, Li Q, Zhou X and Chen H 2020 *Chin. Phys. B* **29** 078701
[17] McCammon J A 1984 *Rep. Prog. Phys.* **47** 1
[18] Kamberaj H and Van der Vaart A 2009 *Biophys. J.* **96** 1307
[19] Kasahara K, Fukuda I and Nakamura H 2014 *PLoS One.* **9** e112419
[20] Erman B 2013 *Proteins* **81** 1097
[21] Yang L W, Liu X, Jursa C J, Holliman M, Rader A J, Karimi H A and Bahar I 2005 *Bioinformatics* **21** 2978
[22] Rueda M, Bottegoni G and Abagyan R 2009 *J. Chem Inf. Model.* **49** 716
[23] Tama F, Miyashita O and Brooks III C L 2004 *J. Mol. Biol.* **337** 985
[24] Bernardi R C, Melo M C R and Schulten K 2015 *Biochim Biophys Acta Gen Subj.* **1850** 872
[25] Atilgan C, Gerek Z N, Ozkan S B and Atilgan A R 2010 *Biophys. J.* **99** 933
[26] Zheng W, Brooks B R, Doniach S and Thirumalai D 2005 *Structure* **13** 565
[27] Zheng W, Liao J C, Brooks B R and Doniach S 2007 *Proteins* **67** 886
[28] Zheng W and Tekpinar M 2009 *BMC Struct. Biol.* **9** 45
[29] Ming D and Wall M E 2005 *Phys. Rev. Lett.* **95** 198103
[30] Ming D and Wall M E 2005 *Proteins* **59** 697
[31] Su J G, Du H J, Hao R, Xu X J, Li C H, Chen W Z and Wang C X 2013 *J. Phys. Chem. B.* **117** 8689
[32] Su J G, Zhang X, Zhao S X, Li X Y, Hou Y X, Wu Y D, Zhu J Z and An H L 2015 *Int. J. Mol. Sci.* **16** 17933
[33] Su J G, Han X M, Zhang X, Hou Y X, Zhu J Z and Wu Y D 2016 *J. Biomol. Struct. Dyn.* **34** 560
[34] Zhang P F and Su J G 2019 *J. Chem. Phys.* **151** 045101
[35] Tirion M M 1996 *Phys. Rev. Lett.* **77** 1905
[36] Bahar I, Atilgan A R, Demirel M C and Erman B 1998 *Phys. Rev. Lett.* **80** 2733
[37] Lezon T R and Bahar I 2010 *PLoS Comput Biol* **6** e1000816
[38] Tang Q Y, Zhang Y Y, Wang J, Wang W and Chialvo D R 2017 *Phys. Rev. Lett.* **118** 088102
[39] Bettati S and Mozzarelli A 1997 *J. Biol. Chem.* **272** 32050
[40] Mayer M P and Bukau B 2005 *Cell. Mol. Life Sci.* **62** 670
[41] Rosenzweig R, Nillegoda N B, Mayer M P and Bukau B 2019 *Nat. Rev. Mol. Cell Biol.* **20** 665
[42] Boorstein W R, Ziegelhoffer T and Craig E A 1994 eJ. Mol. Evol. **38** 1
[43] Bukau B and Horwich A L 1998 *Cell.* **92** 351
[44] Bhattacharya A, Kurochkin A V, Yip G N, Zhang Y, Bertelsen E B and Zuiderweg E R 2009 *J. Mol. Biol.* **388** 475
[45] Kampinga H H and Craig E A 2010 *Nat. Rev. Mol. Cell Biol.* **11** 579
[46] Woo H J, Jiang J, Lafer E M and Sousa R 2009 *Biochemistry.* **48** 11470
[47] Ung P M U, Thompson A D, Chang L, Gestwicki J E and Carlson H A 2013 *PLoS Comput. Biol.* **9** e1003279
[48] Mayer M P and Bukau B 2005 *Cell. Mol. Life Sci.* **62** 670
[49] Stetz G and Verkhivker G M 2017 *PLoS Comput. Biol.* **13** e1005299
[50] Kumar D P, Vorvis C, Sarbeng E B, Ledesma V C C, Willis J E and Liu Q 2011 *J. Mol. Biol.* **411** 1099
[51] Feng Y, Zhang L, Wu S, Liu Z, Gao X, Zhang X, Liu M, Liu J, Huang X and Wang, W 2016 *Angew. Chem. Int. Edit.* **55** 13990
[52] Su J G, Jiao X, Sun T G, Li C H, Chen W Z and Wang C X 2007 *Biophys. J.* **92** 1326
[53] Pang A, Arinaminpathy Y, Sansom M S and Biggin P C 2003 *FEBS Lett.* **550** 168
[54] Sun Y J, Rose J, Wang B C and Hsiao C D 1998 *J. Mol. Biol.* **278** 219
[55] Loeffler H H and Kitao A 2009 *Biophys. J.* **97** 2541
[56] Hayward S and Kitao A 2015 *J. Chem. Theory Comput.* **11** 3895
[57] Lv D, Wang C, Li C, Tan J and Zhang X 2017 *Comput. Biol. Chem.* **67** 62