Article

Characterization of Protein Flexibility Using Small-Angle X-Ray Scattering and Amplified Collective Motion Simulations

Bin Wen,[1] Junhui Peng,[1] Xiaobing Zuo,[2] Qingguo Gong,[1] and Zhiyong Zhang[1,*]
[1]Hefei National Laboratory for Physical Science at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, People's Republic of China; and [2]Advanced Photon Source, Argonne National Laboratory, Chicago, Illinois

ABSTRACT Large-scale flexibility within a multidomain protein often plays an important role in its biological function. Despite its inherent low resolution, small-angle x-ray scattering (SAXS) is well suited to investigate protein flexibility and determine, with the help of computational modeling, what kinds of protein conformations would coexist in solution. In this article, we develop a tool that combines SAXS data with a previously developed sampling technique called amplified collective motions (ACM) to elucidate structures of highly dynamic multidomain proteins in solution. We demonstrate the use of this tool in two proteins, bacteriophage T4 lysozyme and tandem WW domains of the formin-binding protein 21. The ACM simulations can sample the conformational space of proteins much more extensively than standard molecular dynamics (MD) simulations. Therefore, conformations generated by ACM are significantly better at reproducing the SAXS data than are those from MD simulations.

INTRODUCTION

Flexibility within a protein is often critical for its function. For example, a multidomain protein consists of two or more domains connected by flexible linkers (1,2) that determine the extent of interdomain motions and, further, lead to large-scale functionally relevant conformational transitions. Structure determination of the multidomain protein containing flexible linkers is experimentally difficult. It could be rather challenging to use x-ray crystallography to solve a structure with multiple conformations, since this methodology is more applicable to a well-folded protein with a single dominant state. Although solution nuclear magnetic resonance (NMR) is limited to proteins of moderate molecular weight, electron microscopy (EM) usually works best for large-size biomolecular complexes. Small-angle x-ray scattering (SAXS) has been identified in recent years as a promising technique for structure elucidation of proteins (3–6). Although SAXS resolution is inherently low, since a complex 3D structure is reduced to a 1D scattering profile that is orientationally averaged, it can still provide valuable information regarding, for example, the size and shape of the protein. In principle, SAXS has no size limits, which has been successfully demonstrated in various systems from individual proteins to complexes.

SAXS is particularly useful in characterizing the flexibility of a protein in solution. However, traditional analysis methods, such as DAMMIN (7) or GASBOR (8), which use SAXS data to build a single molecular envelope, cannot provide a picture of the highly dynamic protein. Therefore, in recent years, several studies have explored the possibility of combining experimental SAXS data with computational simulations to interpret protein dynamics in solution (9–16). Many of these methods share a similar strategy, namely, using computer simulations to generate a large pool of protein conformations and then selecting the ensemble of structures that best reproduce the SAXS data.

Among the computational techniques, molecular dynamics (MD) simulation, which has been very successfully used in the study of protein dynamics (17–19), is gaining in popularity. However, the computational cost of MD is generally expensive. For a multidomain protein, because an MD simulation at a timescale of microseconds is time-consuming, a timescale of nanoseconds is usually used. On the other hand, under physiological conditions, the protein could be trapped in its locally stable states in the MD simulation while conformational transitions between the different states are rarely sampled due to the frustrating nature of the protein energy landscape (20). Thus, the inefficient sampling of protein conformations in the MD simulation due to the aforementioned issues may fail to interpret the experimental SAXS data properly. To overcome this problem, various methods have been utilized, such as rigid-body modeling (9–11,14), coarse-grained (CG) simulations (13–15), and enhanced sampling techniques (11,14,21).

We previously developed a sampling method called amplified collective motions (ACM) that utilizes a few collective modes obtained from an elastic network model (ENM) to guide the atomic MD simulation (22). The ENM (23,24) is a residue-based CG model that can efficiently calculate collective modes that describe functionally

Submitted February 24, 2014, and accepted for publication July 1, 2014.

*Correspondence: zzyzhang@ustc.edu.cn

Bin Wen and Junhui Peng contributed equally to this work.

Editor: James Cole.

0006-3495/14/08/0956/9 $2.00

http://dx.doi.org/10.1016/j.bpj.2014.07.005

relevant domain motions in proteins (25–28). In ACM, the collective motions obtained from the ENM are accelerated by coupling them to a high-temperature bath. With this strategy, the protein would be able to escape from the traps and explore different conformational states on the energy landscape in a relatively short simulation time. Applications to different proteins support the ability of ACM simulations to sample the conformational space much more extensively than can standard MD simulations (22,29–31). In this article, we combine the simulation results of ACM and SAXS data to reveal various conformational states of multidomain proteins in solution. The results show that the ACM sampling does a much better job than MD at reproducing the SAXS data.

In the next section, we introduce computational details of ACM and control MD simulations, SAXS data acquisitions, and the SAXS-based ensemble optimization method (EOM). In the Results and Discussion section, the protocol is applied to two multidomain proteins, bacteriophage T4 lysozyme (T4L) and tandem WW domains of the formin-binding protein 21 (FBP21-WWs). The ACM method is then compared with some other simulation techniques in combination with SAXS data. The final section is devoted to concluding remarks.

THEORY AND METHODS

Conformational sampling using MD and ACM simulations

T4L

T4L is a two-domain protein with 164 amino acid residues. The N-terminal (residues 13–65) and C-terminal (residues 75–162) domains are connected by an α-helix. The active site between the two domains is responsible for oligosaccharide binding. The many available experimental structures of T4L and its variants indicate the presence of a hinge-bending domain motion that opens or closes the active site (32).

MD simulation. An open conformation of T4L with a resolution of 2.7 Å was chosen from Protein Data Bank (PDB) entry 178L (32). The structure, containing four mutations (C54T, C97A, D127C, and R154C), was changed back to the wild-type form, and the simulation was then set up using the GROMACS-4.5.5 package (33) and the CHARMM27 force field (34). The protein was placed in a cubic box, with a minimum distance of 1.3 nm between the solute and the box boundary. The box was then filled with TIP3P water molecules (35). The energy of the system (protein and waters) was minimized by the steepest-descent method, until the maximum force was <1000 kJ mol^{-1} nm^{-1}. Eight Cl^- ions were added by replacing the same number of waters with the most favorable electrostatic potential to compensate the net positive charges on the protein. The final system (protein, waters, and ions) was minimized again using the steepest descent followed by the conjugate-gradient method, until the maximum force was <100 kJ mol^{-1} nm^{-1}. The simulation was conducted by using the leap-frog algorithm (36) with a time step of 2 fs. The initial atomic velocities were generated according to a Maxwell distribution at 300 K. An equilibration simulation with positional restraints (using a force constant of 1000 kJ mol^{-1} nm^{-2}) was carried out for 100 ps and followed by a production run of 20 ns. The simulation was performed under the constant NPT condition. Each of the three groups (protein, solvent, and ions) was coupled to a thermostat at 300 K using the velocity-rescaling algorithm (37) with a relaxation time of 0.1 ps. The pressure was coupled to 1 bar with a relaxation time of 0.5 ps and a compressibility of 4.5×10^{-5} bar^{-1}. All the bonds in the protein were constrained using the P-LINCS algorithm (38). Twin range cutoff distances for the van der Waals interactions were set to be 0.9 and 1.4 nm, respectively, and the neighbor list was updated every 20 fs. The long-range electrostatic interactions were calculated by the PME algorithm (39), with an interpolation order of 4 and a tolerance of 10^{-5}.

ACM simulation. The ACM method was implemented in the GROMACS 4.5.5 package. Accelerated sampling of the structure was started after the equilibration simulation. Many parameters were the same as for the standard MD simulation, except that collective motions described by the ENM (23) were amplified by coupling them to a high-temperature bath. From an all-atom structure of the protein in the simulation, an ENM was built with CG sites located at the center of mass (COM) of residues. The potential energy function of the ENM takes the harmonic form

$$V = \sum_{i,j>i} \frac{1}{2} k_{ij} \Delta r_{ij}^2, \tag{1}$$

where Δr_{ij} is the fluctuation of the bond connecting residues i and j, and k_{ij} is the spring constant. For any two residues i and j, with their COM distance r_{ij}, the spring constant between them was

$$k_{ij} = \begin{cases} 1.0c & r_{ij} \le 0.7 \text{ nm} \\ 10^{-2}c & 0.7 < r_{ij} \le 1.1 \text{ nm} \\ 5 \times 10^{-4}c & 1.1 < r_{ij} \le 1.4 \text{ nm} \\ 0 & r_{ij} > 1.4 \text{ nm} \end{cases}, \tag{2}$$

where c could be any nonzero value, and the four-range spring constants described the interactions in the protein from strong to weak. The short cutoff distance, 0.7 nm, defined the first coordination shell, and the long cutoff distance, 1.4 nm, was chosen to avoid unrealistic large-amplitude fluctuations in some residues along particular directions (23). A middle cutoff value of 1.1 nm was set between the short and long cutoff distances. A Hessian matrix of the second derivatives of the overall potential (Eq. 1) was constructed and then diagonalized to yield a matrix of eigenvectors and corresponding eigenvalues. Each eigenvector with a nonzero eigenvalue is called a normal mode, and the corresponding eigenvalue is proportional to the squared frequency of the motion along the mode. Note that the value of c is not important here, because it only affects the eigenvalues, not the eigenvectors (collective modes). Usually only a few ENM modes with the lowest frequencies are dominant in collective motions of the protein. For T4L, we took the three slowest modes to define an essential subspace. At each time step, the velocity of each atom was divided into two parts, the part projected onto the essential subspace and the remainder. By modifying the weak coupling method (40), the component of velocity in the essential subspace was coupled to a high temperature of 800 K, whereas the remaining velocity was coupled normally to 300 K, and thus the updated velocity was the combination of these two components. During the ACM simulation, collective modes were updated on the fly by doing ENM calculations every 100 time steps according to the new generated protein conformation. The simulation time was 20 ns in total.

FBP21-WWs

As a structural component of the mammalian spliceosomal A/B complex, FBP21 plays an important role in pre-mRNA splicing (41). The protein consists of a matrin-type zinc finger and two group-III WW domains. Huang et al. have solved the NMR structure of the tandem WW domains (42), which contains 75 amino acid residues. The two domains, denoted as WW1 (residues 6–32) and WW2 (residues 47–73), respectively, are connected by a highly flexible linker. The above structure information and ^{15}N relaxation data both suggest a very mobile interdomain movement, which may enable cooperative binding of these domains with different ligands.

MD simulation. Model 1 of the NMR ensemble (PDB entry 2JXW) was selected as the initial structure. Besides the 75 residues, the protein sample for SAXS measurement has a Met at the N-terminus and an eight-residue His tag (LEHHHHHH) at the C terminus. We added these additional residues to the NMR structure by MODELER (43), and the system with 84 residues in total was used to start an MD simulation. The set-up procedures and parameters were the same as those in the MD simulation of T4L except as follows. A rhombic dodecahedron water box was used, and the minimum distance between the protein and the box boundary was 1.4 nm. Ninety-nine Na^+ and 91 Cl^- ions were added, not only to compensate for the net negative charges on the protein but also to mimic the salt concentration (300 mM) of the SAXS sample. The energy of the final system was minimized using the steepest descent and then the conjugate-gradient method, until the maximum force was <180 kJ mol^{-1} nm^{-1}. Initial atomic velocities for the equilibration simulation were generated according to a Maxwell distribution at 310 K, and the subsequent production run was 20 ns under the constant NPT condition. The four groups (protein, solvent, Na^+ ions, and Cl^- ions) were coupled separately to a reference temperature of 310 K.

ACM simulation. Parameters for the ACM simulation of FBP21-WWs were largely the same as those for T4L, except as follows. The velocities along the three slowest ENM modes were coupled to 500 K, whereas the rest of the velocities were coupled to 310 K. Note that to accelerate the collective motions of FBP21-WWs, we used a lower temperature than that used for T4L, because FBP21-WWs is more mobile than T4L, with easier transit between different conformational states. Those collective modes were updated on the fly every 50 time steps. The ACM simulation was 20 ns long.

SAXS data

Simulated SAXS profile of T4L

From various experimental structures of wild-type T4L and its mutants, we selected 38 structures that may represent possible conformations of the protein in solution (44). Each mutant was changed back to the wild-type form, and its theoretical SAXS curve was computed by the CRYSOL program (45). Thus, a multiconformational SAXS profile of T4L was obtained by taking the average,

$$I(q) = \frac{1}{N}\sum_{n=1}^{N} I_n(q). \qquad (3)$$

Here, $N = 38$ is the number of experimental structures, $I_n(q)$ is the theoretical SAXS profile of a single structure, n, and $q = 4\pi \sin\theta/\lambda$ is the momentum transfer, where 2θ is the scattering angle and λ is the wavelength.

Experimental SAXS data of FBP21-WWs

The SAXS experiment of FBP21-WWs was performed at the beamline 12ID-B of the Advanced Photon Sources at Argonne National Laboratory, with a wavelength of 1.033 Å. Data were acquired from three concentrations (1.0, 3.0, and 5.0 mg/mL) and analyzed by the ATSAS package (46,47). After subtracting buffer scattering, the data curves from different concentrations were scaled and merged using PRIMUS (48). GNOM (49) was employed for calculating the pair distance distribution function (PDDF). The radius of gyration (R_g) of the protein was estimated by Guinier plot.

SAXS fitting

The EOM (9) was selected to identify a small ensemble of representative conformations from a large pool of protein structures, such as an MD or ACM trajectory, to best fit the experimental SAXS data. The search procedure is achieved by minimizing the residual between the experimental and calculated SAXS curves:

$$\chi = \left\{\frac{1}{K-1}\sum_{m=1}^{K}\left[\frac{\mu I(q_m) - I_{exp}(q_m)}{\sigma(q_m)}\right]^2\right\}^{1/2}, \qquad (4)$$

where K is the number of data points in $I_{exp}(q)$, and $\sigma(q)$ are standard deviations of the experimental data. $I(q)$ is the average of the SAXS profiles (Eq. 3) of these conformations in the small ensemble, and μ is a scaling factor. In EOM, χ (Eq. 4) is minimized by using the genetic algorithm (50) to pick the optimal ensemble of structures.

RESULTS AND DISCUSSION

T4L

The simulated SAXS profile of T4L is shown in Fig. 1 (*black line*) to be somewhat different from the SAXS curve of either an open (Fig. 1, *red dashed line*) or a closed (Fig. 1, *green dashed line*) structure. To reproduce the simulated SAXS profile, one has to sample not only the open but also the closed conformations of T4L in simulations.

Starting from the open structure, T4L remains in its open state during the 20 ns MD simulation. Root mean-square deviations (RMSDs) of the C_α atoms of residues 1–162 are mostly <2.0 Å (Fig. 2 *a*, *black trace*). The relatively large RMSD values for the closed structure (Fig. 2 *a*, *red trace*) also indicate that T4L does not access the closed state in the MD simulation. Conversely, the protein transits between the open and closed states frequently during the 20 ns ACM simulation (Fig. 2 *b*). RMSDs of the respective N-terminal (residues 13–65) and C-terminal (residues 75–162) domains were also calculated. In the MD simulation, the RMSD of the N-terminal domain is ~0.6 ± 0.1 Å, and the values of the C-terminal domain are ~0.7 ± 0.1 Å. In

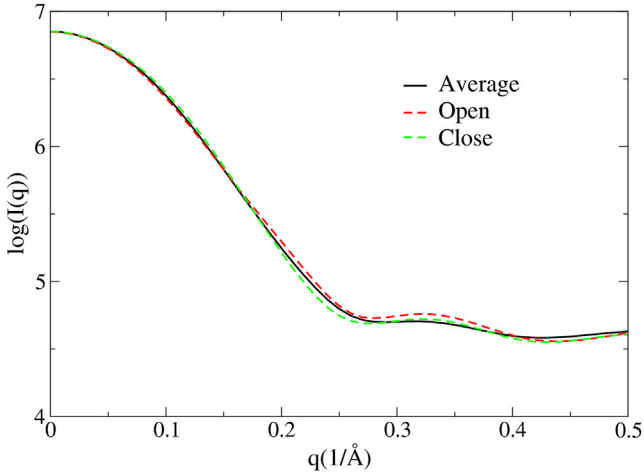

FIGURE 1 The simulated SAXS profile of T4L (*black line*) that is the average from the 38 experimental structures (Eq. 3). The theoretical SAXS curves of an open conformation (*red dashed line*) and a closed conformation (*green dashed line*) of the protein are shown for comparison. To see this figure in color, go online.

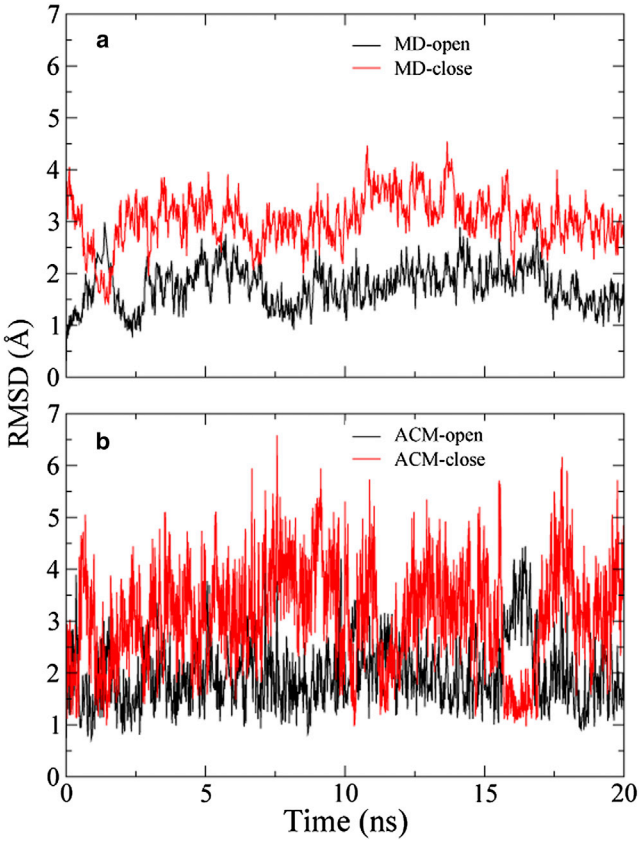

FIGURE 2 RMSD in the MD simulation (*a*) and the ACM simulation (*b*) of T4L. The values are calculated from C_α atoms of residues 1–162. In each panel, the RMSD curve for the open structure is colored black, and that for the closed structure red. To see this figure in color, go online.

the ACM simulation, the RMSD values of both domains are ~0.7 Å. That is to say, each domain in the ACM simulation is as stable as that in the MD simulation, which indicates that ACM does not break the internal structures of the domains. These RMSD results strongly affirm that the ACM method not only allows for extensive sampling of collective domain motions, but also preserves the local structures of the protein. In the ACM simulation, only a very few degrees of freedom (the first three slowest ENM modes) are coupled to the high temperature, whereas most of the degrees of freedom are coupled to room temperature, which distinguishes ACM from a high-temperature MD simulation. In the latter, the internal structure of each domain would be destroyed.

From experimental structures of T4L and its mutants in the PDB, 38 structures were selected to constitute a protein ensemble (32,44,51). Principal component analysis (PCA) was performed on the ensemble (52) using the C_α atoms of residues 1–162 to yield PCA modes describing collective motions of T4L. The results indicate that there are two PCA modes that contribute ~90% of the total fluctuation in the protein. One mode describes an open-closed domain motion, and the other represents a twist motion between the domains. The 38 experimental structures were projected onto the plane spanned by the above two PCA modes, which clearly form two distinct clusters along the open-closed mode (Fig. 3, *blue*). The cluster on the right contains closed structures and that on the left consists of open structures. The trajectories of the MD and ACM simulations were also projected onto the plane to compare their efficiency of sampling of the domain motions. The MD simulation starting from the open structure of T4L only samples a limited region on the left side of the plane (Fig. 3, *black*), which partially covers the cluster of open structures but not the cluster of closed structures. That is to say, the protein is trapped in the open state and conformational transitions do not occur during the 20 ns MD simulation. The ACM simulation (Fig. 3, *red*), which can already cover the two clusters of T4L structures, explores significantly larger areas on the plane than does MD. We estimated potential energies of the conformations in the respective MD and ACM trajectories by replacing explicit water molecules with an implicit generalized Born surface area solvent model (53). The energy differences between the MD and ACM simulations are marginal (Fig. S1 in the Supporting Material), which suggests that the protein conformations sampled by ACM have fairly low energies compared to those from MD under room temperature conditions. Thus, the ACM simulation is unlike a standard MD simulation under high temperature in that the latter would mainly sample the conformational space with high energies.

A pool of 2000 protein conformations was constructed from the respective MD and ACM trajectories of T4L. The theoretical SAXS profiles of all the structures were precomputed by CRYSOL (45) and were used to select a

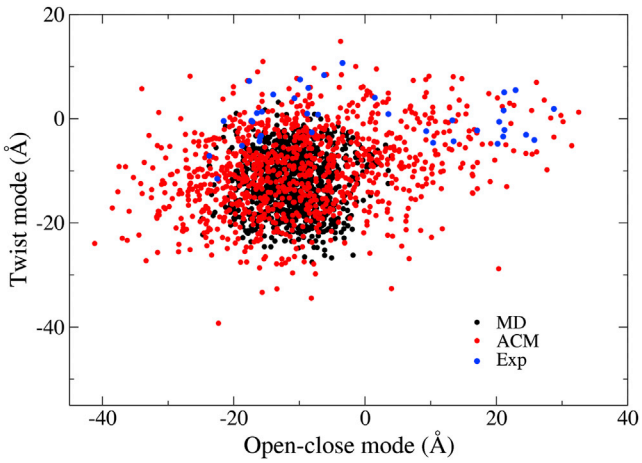

FIGURE 3 Projections of the T4L structures onto the 2D essential subspace defined by the open-closed and twist modes. PCA was performed on the ensemble of 38 experimental structures of the T4L, and the first two eigenvectors with the largest eigenvalues defined the essential subspace. Each point on the plane represents a conformation. The 38 experimental structures of T4L are colored blue. The projections of MD (*black*) and ACM (*red*) indicate their sampling efficiency. To see this figure in color, go online.

small number (up to 20) of conformations to fit the simulated SAXS curve of T4L (Eq. 3) by EOM (9). Fifty independent EOM calculations were run on the respective MD and ACM pools. The χ values (Eq. 4) plotted in Fig. 4 *a* clearly indicate that the small ensembles selected from ACM always have smaller χ than those from MD. The minimal χ determined by EOM for the MD pool is 0.179, and the corresponding ensemble contains all open conformations of T4L (Fig. 4 *b*). The EOM applied to the ACM pool obtains a minimal χ of 0.007, and the corresponding ensemble includes both open and closed conformations (Fig. 4 *c*). Since the simulated SAXS profile of T4L is the average from 38 experimental structures (Eq. 3) that consist of both open and closed conformations (Fig. 3, *blue*), the ACM simulation, which samples diverse conformations, is superior to the MD simulation at reproducing the SAXS profile.

FBP21-WWs

Fig. 5 shows the experimental SAXS curve of FBP21-WWs (Fig. 5 *a*), and the corresponding PDDF computed by GNOM (49). The shape of the PDDF (Fig. 5 *b*) suggests that the protein may be able to take an extended structure in solution, which is possible, since the linker between the two WW domains is very mobile (42). The R_g of the protein, estimated from a Guinier plot, is ~19.0 Å.

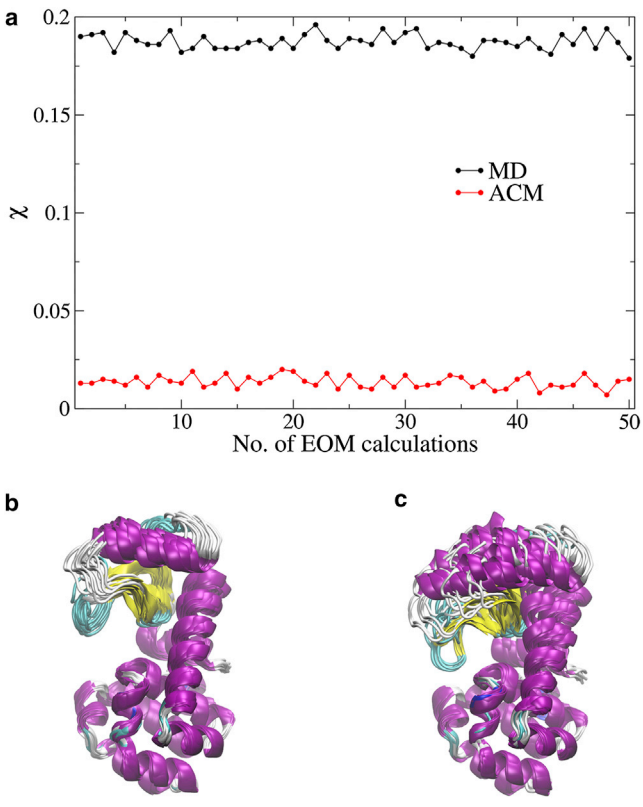

FIGURE 4 EOM analysis of T4L. (*a*) χ values of 50 independent EOM calculations for the respective MD (*black*) and ACM (*red*) trajectories. (*b* and *c*) Structure ensemble with the minimal χ = 0.179 from MD (*b*) and structure ensemble with the minimal χ = 0.007 from ACM (*c*). The structures are superimposed by the C-terminal domain (residues 75–162). All the structures, including those in the Supporting Material, were created by VMD (59). To see this figure in color, go online.

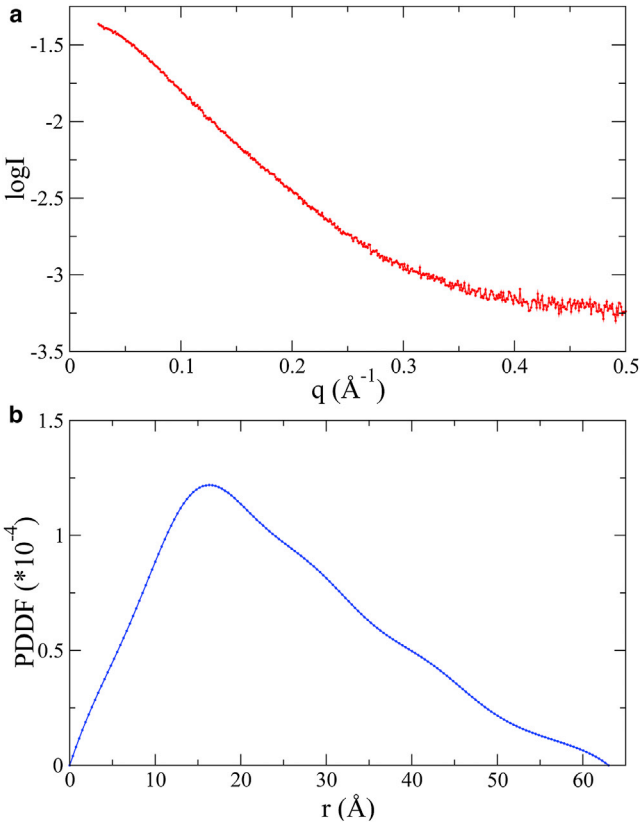

FIGURE 5 SAXS data of FBP21-WWs. (*a*) Plots of experimental SAXS curve, with data points up to q = 0.5 Å^{-1}. (*b*) PDDF calculated by GNOM (49). To see this figure in color, go online.

From the respective MD and ACM trajectories of FBP21-WWs, pools containing 2000 conformations were built. After precomputing the theoretical SAXS profiles of all the structures, 50 cycles of EOM were run to select from the MD and ACM pools small ensembles that best reproduce the experimental SAXS data. As in the case of T4L, the ensembles selected from the ACM pool of FBP21-WWs give a much better fit to the SAXS data than those from the MD pool, based on their χ values (Eq. 4) (Fig. 6 *a*). The starting model of FBP21-WWs is compact, and the two WW domains essentially stay close to each other during the 20 ns MD simulation, although their relative orientations change. Therefore, all the ensembles selected from the MD pool consist of compact structures (Fig. 6 *b*), and the minimal χ is 0.592. In the 20 ns ACM simulation, although the internal structure of each WW domain is well preserved, the distance between the two changes widely, as do the domain orientations. The ensembles selected from the ACM pool contain not only compact but also extended

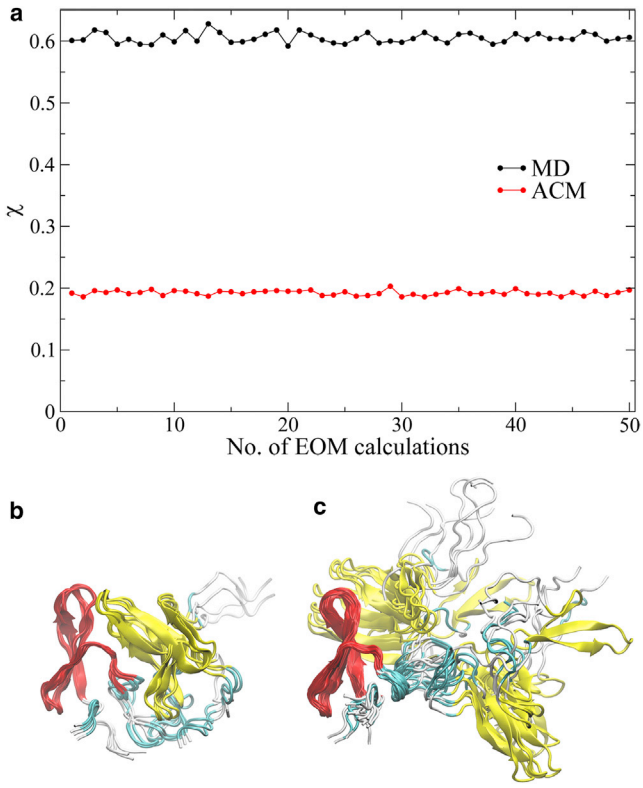

FIGURE 6 EOM analysis of FBP21-WWs. (*a*) χ values of 50 independent EOM calculations of the MD (*black*) and ACM (*red*) trajectory. (*b* and *c*) Structure ensemble with the minimal $\chi = 0.592$ from MD (*b*) and structure ensemble with the minimal $\chi = 0.186$ from ACM (*c*). The structures are superimposed by the WW1 domain (residues 6-32 (*red*)), to show the relative orientation of the WW2 domain (residues 47-73 (*yellow*)). To see this figure in color, go online.

structures (Fig. 6 *c*), and the minimal χ is 0.186, significantly smaller than that from the MD pool (Fig. 6 *b*). The results indicate that FBP21-WWs may transit between the compact and extended conformations in solution. The average R_g of those conformations in the ensemble from the ACM pool (Fig. 6 *c*) is around 19 Å, which is consistent with the Guinier analysis.

Convergence of ACM in fitting the SAXS data

It is clear that the ACM method can significantly enhance conformational sampling and does a better job of reproducing the SAXS data compared to normal MD. One may ask whether or not different ACM simulations of the same protein can offer similar results of SAXS fitting. We have performed multiple ACM simulations of T4L and FBP21-WWs that 1), start from different conformations; 2), accelerate different numbers of collective modes; 3), choose different high temperatures for ACM coupling, and 4), end with different simulation times. The results indicate a reasonable convergence of the SAXS-fitting calculations, that is, different ACM simulations can yield similar structure ensembles via the EOM. More details can be found in the Supporting Material.

Comparison with other sampling methods in combination with SAXS

To our knowledge, there are several methods that integrate SAXS data with computational modeling to characterize dynamic multidomain proteins in solution (9,11,13,14). In the EOM (9), if there is no outside trajectory of a multidomain protein, the program Pre_bunch will produce a large pool of conformations by rigid-body modeling. Individual domains are treated as rigid bodies, which are connected by self-avoiding linkers. We used Pre_bunch (54) to generate 10,000 conformations of FBP21-WWs and the EOM to pick from these a small ensemble to fit the SAXS data. Compared to the ensembles from the ACM trajectories (Figs. 6 *c* and S7), the ensemble from the structure pool generated by Pre_bunch consists of significantly more diverse conformations (Fig. S8). In Pre_bunch, only a simple interaction is considered, to avoid steric clashes in the generated models, so the two WW domains may take various orientations. However, the ACM simulations of FBP21-WWs are all-atom simulations with a refined molecular force field, so the conformations should be physically more reasonable than those from Pre_bunch, and clearly some clusters of structures exist in the ensembles (Figs. 6 *c* and S7). Therefore, although the ensembles from ACM and those from Pre_bunch have nearly the same χ values in fitting the SAXS data, the former are likely more realistic than the latter. Since the SAXS data are inherently low-resolution, the SAXS fitting from a large structure pool is susceptible to over-fitting. The ACM simulations may avoid this issue to some extent, because they can produce realistic conformations of proteins.

In the minimal ensemble search (11), rigid-body MD simulations (called BILBOMD) are used to generate a wide range of protein conformations for SAXS analysis. Additional strategies, such as reduced nonbonded interactions, large time-step size, and high-temperature coupling to domain linkers, are implemented to enhance sampling efficiency. The basis-set-supported SAXS (BSS-SAXS) reconstruction (13) developed by Yang et al. samples a large number of conformations using MD simulations based on a one-site-per-residue CG model. Hummer and co-workers have developed a method called ensemble refinement of SAXS (EROS) (14), in which the residue-level CG model is also used and the domains are represented as rigid bodies in replica-exchange Monte Carlo simulations. In the ACM simulation, the internal structure of each domain would be naturally preserved, since only the collective motions between domains are accelerated by high-temperature coupling; this obviates the need for rigid-body approximation. This may be one of the advantages of ACM, because it is not always intuitive to predetermine which parts should

be treated as rigid bodies in some proteins. The applications of ACM in this article are actually all-atom simulations including explicit solvent, which may make it possible to sample physically more reasonable conformations but with more computational cost than the aforementioned methods.

In addition to ACM, there are other MD-based enhanced sampling methods, such as replica-exchange MD (55) and accelerated MD (aMD) (56), which also can be used for SAXS fitting. However, REMD of a protein like FBP21-WWs (with explicit solvent) would be computationally quite expensive, since many replicas must be run under a series of temperatures. aMD improves the sampling by using a boost potential to reduce energy barriers between different states of the protein. A standard MD simulation must be run first to determine a proper value of the boost potential. Our ACM method has only a minor additional computational cost compared to MD. Instead of altering the potential energy surface, ACM accelerates the collective motions and lets the protein cross the energy barriers more easily than does conventional MD. For conformational sampling of a multidomain protein, ACM is expected to be more efficient than aMD, since these collective modes, directly related to the domain motions, are excited in ACM, whereas the sampling in aMD would not focus along particular reaction coordinates. In this sense, aMD may work better than ACM for intrinsically disordered proteins, because there may be no collective motions in such proteins.

CONCLUSION

SAXS is an efficient and important complement to other techniques for structure elucidation, especially in the case of highly dynamic multidomain proteins. High-resolution techniques (x-ray crystallography and solution NMR) are able to solve the structures of individual domains. However, it would be difficult to crystallize a flexible multidomain protein, such as FBP21-WWs. Also, it is generally not easy to obtain NMR restraints between the domains connected by flexible linkers. A protein like FBP21-WWs is too small to be investigated by electron microscopy. Data can be collected faster by SAXS than by other techniques, and they provide useful information, such as the size and domain orientations of the multidomain protein.

Due to the low-resolution nature of SAXS, it should be combined with computational simulations to extract structure information about the multidomain protein. From a starting structure, a large number of protein conformations are generated by simulations, and an ensemble of structures is then selected from the pool to best reproduce the experimental SAXS data. In the case of simulations, a key issue is to sample the conformational space of the protein adequately, but this is a nontrivial problem. The study described in this article contributes a useful tool that combines the ACM sampling method and the SAXS data. Results of the two multidomain proteins, T4L and FBP21-WWs, support the idea that ACM simulations are significantly better than control MD simulations at reproducing the SAXS data and interpreting protein flexibility in solution. In the study of FBP21-WWs, it was found that the compact and extended conformations can coexist in solution, although this was not detected by NMR studies (42).

It should be noted that the ACM sampling is a nonequilibrium simulation and does not generate a proper Boltzmann ensemble. Therefore, the protein conformations produced by ACM need to be reweighted to recover the canonical distribution. This issue has been addressed elsewhere in the literature (57,58), where the idea of accelerating collective motions is combined with other sampling methods that can retain the correct ensemble. This strategy may help us to tackle the reweighting problem in our ongoing improvement of the ACM method. Alternatively, we can simply use the current version of ACM to efficiently generate possible conformations of the protein and then rely on appropriate SAXS fitting to recover the correct relative population between different states. The key issue of how to prevent overfitting can be tackled by determining a small number of clusters from the large structure pool. The weights of these clusters, which usually represent possible conformational states of the protein, are then optimized by best fitting the SAXS data using some advanced approaches, such as the Bayesian-based Monte Carlo algorithm (13) and the maximum-entropy method (14).

Generally, there is a trade-off between the sampling efficiency and the accuracy of the generated conformations. For a very large multidomain protein or complex, the all-atom ACM simulation with explicit solvent would be rather time-consuming. The protein may not be able to achieve an adequate sampling within a simulation time of nanoseconds. In this case, ACM can be combined with some simplified models, such as implicit solvent (53) and CG protein models (13–15), to achieve further acceleration of the conformational sampling. This would be one focus for future research.

SUPPORTING MATERIAL

Eight figures and Supporting Results and Discussion are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00720-6.

This work is supported by the National Key Basic Research Program of China (grants 2013CB910203 and 2011CB911104), the National Natural Science Foundation of China (grant 31270760), the Strategic Priority Research Program of the Chinese Academy of Sciences (grant XDB08030102), the Specialized Research Fund for the Doctoral Program of Higher Education (grant 20113402120013), Anhui Natural Science Foundation (grant 1208085MC38), and the Fundamental Research Funds for the Central Universities (WK2070000020).

REFERENCES

1. Ekman, D., A. K. Björklund, …, A. Elofsson. 2005. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.* 348:231–243.

2. Levitt, M. 2009. Nature of the protein universe. *Proc. Natl. Acad. Sci. USA.* 106:11079–11084.
3. Lipfert, J., and S. Doniach. 2007. Small-angle x-ray scattering from RNA, proteins, and protein complexes. *Annu. Rev. Biophys. Biomol. Struct.* 36:307–327.
4. Jacques, D. A., and J. Trewhella. 2010. Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls. *Protein Sci.* 19:642–657.
5. Mertens, H. D. T., and D. I. Svergun. 2010. Structural characterization of proteins and complexes using small-angle x-ray solution scattering. *J. Struct. Biol.* 172:128–141.
6. Rambo, R. P., and J. A. Tainer. 2010. Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle x-ray scattering. *Curr. Opin. Struct. Biol.* 20:128–137.
7. Svergun, D. I. 1999. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* 76:2879–2886.
8. Svergun, D. I., M. V. Petoukhov, and M. H. J. Koch. 2001. Determination of domain structure of proteins from x-ray solution scattering. *Biophys. J.* 80:2946–2953.
9. Bernadó, P., E. Mylonas, …, D. I. Svergun. 2007. Structural characterization of flexible proteins using small-angle x-ray scattering. *J. Am. Chem. Soc.* 129:5656–5664.
10. Förster, F., B. Webb, …, A. Sali. 2008. Integration of small-angle x-ray scattering data into structural modeling of proteins and their assemblies. *J. Mol. Biol.* 382:1089–1106.
11. Pelikan, M., G. L. Hura, and M. Hammel. 2009. Structure and flexibility within proteins as identified through small angle x-ray scattering. *Gen. Physiol. Biophys.* 28:174–189.
12. Bernadó, P., and M. Blackledge. 2010. Structural biology: Proteins in dynamic equilibrium. *Nature.* 468:1046–1048.
13. Yang, S., L. Blachowicz, …, B. Roux. 2010. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. USA.* 107:15757–15762.
14. Różycki, B., Y. C. Kim, and G. Hummer. 2011. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure.* 19:109–116.
15. Daily, M. D., L. Makowski, …, Q. Cui. 2012. Large-scale motions in the adenylate kinase solution ensemble: coarse-grained simulations and comparison with solution x-ray scattering. *Chem. Phys.* 396:84–91.
16. Hammel, M. 2012. Validation of macromolecular flexibility in solution by small-angle x-ray scattering (SAXS). *Eur. Biophys. J.* 41:789–799.
17. Karplus, M., and J. A. McCammon. 2002. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9:646–652.
18. Adcock, S. A., and J. A. McCammon. 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* 106:1589–1615.
19. Dror, R. O., R. M. Dirks, …, D. E. Shaw. 2012. Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* 41:429–452.
20. Onuchic, J. N., Z. Luthey-Schulten, and P. G. Wolynes. 1997. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* 48:545–600.
21. Ravikumar, K. M., W. Huang, and S. Yang. 2012. Coarse-grained simulations of protein-protein association: an energy landscape perspective. *Biophys. J.* 103:837–845.
22. Zhang, Z., Y. Shi, and H. Liu. 2003. Molecular dynamics simulations of peptides and proteins with amplified collective motions. *Biophys. J.* 84:3583–3593.
23. Atilgan, A. R., S. R. Durell, …, I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–515.
24. Bahar, I., and A. J. Rader. 2005. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.* 15:586–592.
25. Kitao, A., and N. Go. 1999. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* 9:164–169.
26. Berendsen, H. J. C., and S. Hayward. 2000. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* 10:165–169.
27. Ma, J. 2005. Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure.* 13:373–380.
28. Bahar, I., T. R. Lezon, …, E. Eyal. 2010. Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* 39:23–42.
29. He, J., Z. Zhang, …, H. Liu. 2003. Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions. *J. Chem. Phys.* 119:4005–4017.
30. Wriggers, W., Z. Zhang, …, D. C. Sorensen. 2006. Simulating nanoscale functional motions of biomolecules. *Mol. Simul.* 32:803–815.
31. Zhang, Z., P. C. Boyle, …, W. Wriggers. 2006. Entropic folding pathway of human epidermal growth factor explored by disulfide scrambling and amplified collective motion simulations. *Biochemistry.* 45:15269–15278.
32. Zhang, X. J., J. A. Wozniak, and B. W. Matthews. 1995. Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. *J. Mol. Biol.* 250:527–552.
33. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.
34. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.
35. Jorgensen, W. L., J. Chandrasekhar, …, M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.
36. Hockney, R. W., S. P. Goel, and J. W. Eastwood. 1974. Quiet high-resolution computer models of a plasma. *J. Comput. Phys.* 14:148–158.
37. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.
38. Hess, B. 2008. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* 4:116–122.
39. Essmann, U., L. Perera, …, L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.
40. Berendsen, H. J. C., J. P. M. Postma, …, J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.
41. Bedford, M. T., R. Reed, and P. Leder. 1998. WW domain-mediated interactions reveal a spliceosome-associated protein that binds a third class of proline-rich motif: the proline glycine and methionine-rich motif. *Proc. Natl. Acad. Sci. USA.* 95:10602–10607.
42. Huang, X., M. Beullens, …, Y. Shi. 2009. Structure and function of the two tandem WW domains of the pre-mRNA splicing factor FBP21 (formin-binding protein 21). *J. Biol. Chem.* 284:25375–25387.
43. Eswar, N., D. Eramian, …, A. Sali. 2008. Protein structure modeling with MODELLER. *Methods Mol. Biol.* 426:145–159.
44. de Groot, B. L., S. Hayward, …, H. J. C. Berendsen. 1998. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins.* 31:116–127.
45. Svergun, D., C. Barberato, and M. H. J. Koch. 1995. CRYSOL: a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28:768–773.
46. Konarev, P. V., M. V. Petoukhov, …, D. I. Svergun. 2006. ATSAS 2.1, a program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* 39:277–286.
47. Petoukhov, M. V., D. Franke, …, D. I. Svergun. 2012. New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* 45:342–350.

48. Konarev, P. V., V. V. Volkov, …, D. I. Svergun. 2003. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* 36:1277–1282.

49. Semenyuk, A. V., and D. I. Svergun. 1991. GNOM: a program package for small-angle scattering data-processing. *J. Appl. Crystallogr.* 24:537–540.

50. Goldberg, D. E. 1989. Genetic Algorithms in Search. Kluwer Academic, Boston.

51. Mchaourab, H. S., K. J. Oh, …, W. L. Hubbell. 1997. Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. *Biochemistry.* 36:307–316.

52. Amadei, A., A. B. Linssen, and H. J. C. Berendsen. 1993. Essential dynamics of proteins. *Proteins.* 17:412–425.

53. Onufriev, A., D. Bashford, and D. A. Case. 2004. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins.* 55:383–394.

54. Petoukhov, M. V., and D. I. Svergun. 2005. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* 89:1237–1250.

55. Sugita, Y., and Y. Okamoto. 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314:141–151.

56. Hamelberg, D., J. Mongan, and J. A. McCammon. 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120:11919–11929.

57. Kubitzki, M. B., and B. L. de Groot. 2007. Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange. *Biophys. J.* 92:4262–4270.

58. Hu, Y., W. Hong, …, H. Liu. 2012. Temperature-accelerated sampling and amplified collective motion with adiabatic reweighting to obtain canonical distributions and ensemble averages. *J. Chem. Theory Comput.* 8:3777–3792.

59. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28.

SUPPORTING MATERIAL

Characterization of Protein Flexibility using Small-angle X-ray Scattering and Amplified Collective Motion Simulations

Bin Wen[1, #], Junhui Peng[1, #], Xiaobing Zuo[2], Qingguo Gong[1], and Zhiyong Zhang[1, *]

[1]Hefei National Laboratory for Physical Science at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China; [2]Advanced Photon Source, Argonne National Laboratory, Chicago, IL 60437

[#]Bin Wen and Junhui Peng contributed equally to this work.

*Corresponding author: Zhiyong Zhang, Tel: +86-551-63600854; Email: zzyzhang@ustc.edu.cn

Running title: Protein Flexibility in Solution

SUPPLEMENTARY RESULTS AND DISCUSSION

Convergence of ACM in fitting the SAXS data

T4L

We have investigated the issue of convergence by running multiple ACM simulations as follows.

Different starting structures. We have carried out an ACM simulation starting from a closed structure of T4L. The protein can transit between the closed and the open states back and forth within the 20 ns simulation time (Fig. S2a), like in the ACM simulation starting from the open structure (Fig. 3). EOM yields a structure ensemble containing both closed and open conformations of T4L, with the minimal χ of 0.008 (Fig. S2b).

Number of collective modes to be accelerated. This should be determined based on how many ENM modes are needed to describe the hinge-bending domain motions of T4L. We computed the overlap between the low-frequency ENM modes and the open-close/twist mode, respectively. The first three ENM modes have already shown a good convergence to significantly cover the collective domain motions of T4L (Fig. S3), with an overlap coefficient of 0.89 to the open-close mode and 0.81 to the twist mode (note that a coefficient of 1.0 means complete coverage). Therefore the ACM simulation using the three modes should be better than that using the two modes. On the other hand, there is a technical issue that prevents us from using very few (two or even one) collective modes. In this case, the temperature of one or two degrees of freedom would fluctuate wildly, which may distort the protein structure when the temperature is extremely high (see below the discussion of how to set the high temperature in ACM). For different proteins, the number of collective modes to be accelerated should be system dependent, but we suggest starting with three modes, and adding more if necessary. We have carried out an ACM simulation that coupled the first four collective modes at 800 K, which shows a larger sampling area in the essential subspace (Fig. S4a) than the three-mode ACM simulation does (Fig. 3). However, the EOM ensembles of the two simulations are rather similar (Fig. S4b and Fig. 4c).

High temperature for ACM coupling. We performed several ACM simulations, which couple the first three collective modes to different temperatures, respectively. If the temperature is not high enough, the protein cannot cross the energy barrier and reach the closed state, so the ensemble selected by EOM does not fit the SAXS data quite well (data not shown). Figure S5 shows the results of the ACM simulation at 1000 K, which samples a broader region in the essential subspace (Fig. S5a) than the ACM simulation at 800 K does (Fig. 3). The two ACM simulations yield very similar EOM ensembles that contain not only the closed but also the open conformations of T4L (Fig. S5b and Fig. 4c). Generally we have little information on the energy barriers of the protein, so it is not straightforward to determine an optimal temperature for ACM coupling. We usually try a relatively high temperature firstly in order to obtain efficient sampling, but it should be noted that a very high temperature may distort the local structures of the protein since there exists a leakage of

energy between the high-temperature degrees of freedoms to the room-temperature ones. In this case, the temperature should be decreased.

Different simulation times. We have extended the 20 ns ACM simulation of T4L to 40 ns. It is found that the 40 ns simulation covers a larger area in the essential subspace (Fig. S6a) than the 20 ns simulation does (Fig. 3), but their EOM ensembles are quite similar (Fig. S6b and Fig. 4c).

FBP21-WWs

We have finished a series of ACM simulations of FBP21-WWs, as those of T4L. Despite their differences, the EOM ensembles from various trajectories share some similar clusters of structures including both the compact and the extended conformations of the protein (Fig. 6c and Fig. S7). The results again indicate a fairly good convergence of ACM in fitting the SAXS data.

SUPPLEMENTARY FIGURES

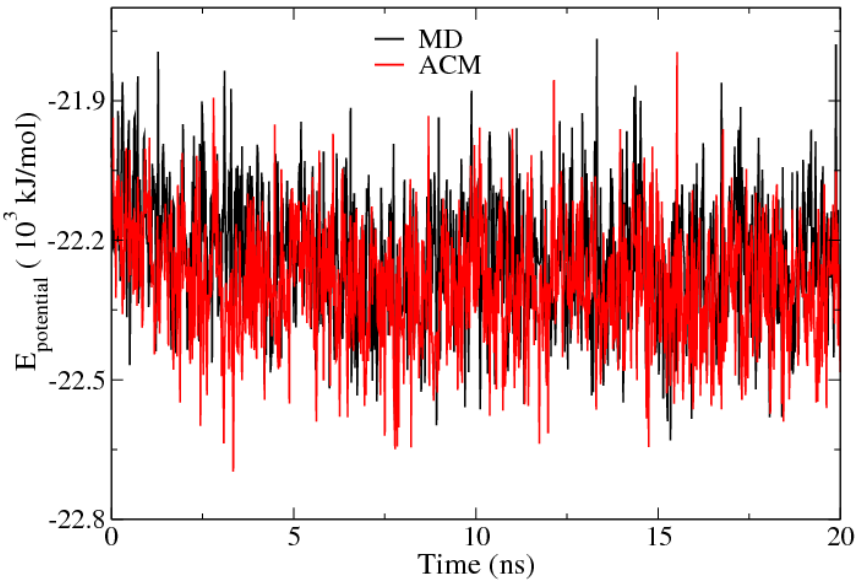

Figure S1. Potential energies of the MD (black solid line) and the ACM (red solid line) simulation, respectively. For each trajectory, the explicit water molecules in each frame were removed, and then the solvent contribution was estimated by using an implicit solvent model called the generalized Born surface area (GBSA) model. The calculations were done by using the "-rerun" option of the "mdrun" program in GROMACS-4.5.5 package. In the mdp file, the option "GBSA" was turned on.

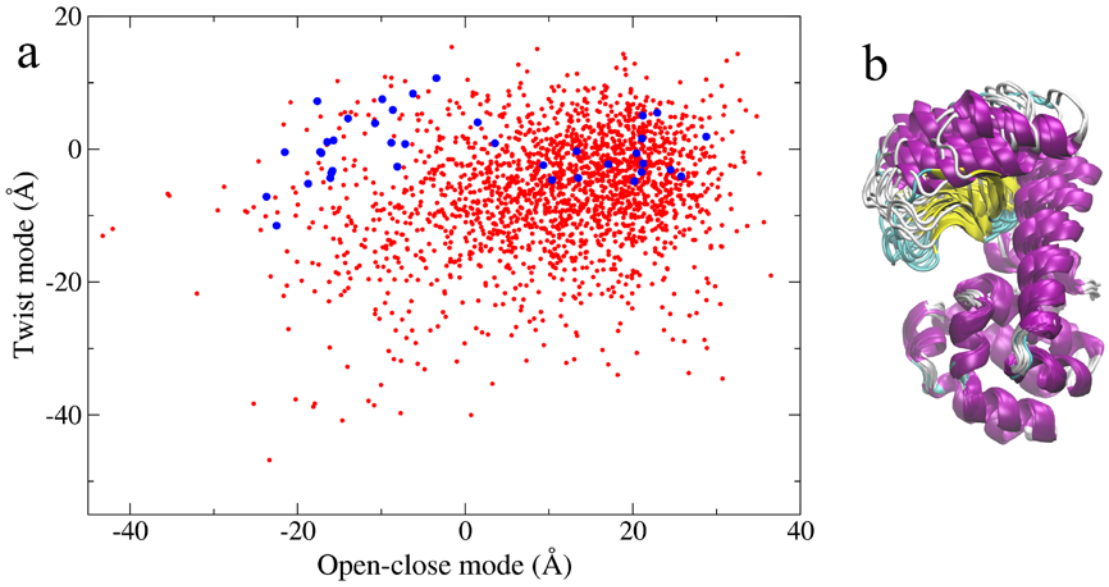

Figure S2. The ACM simulation of T4L starting from a closed structure. (a) Conformations in the trajectory are projected onto the 2D essential subspace (colored by red), and the 38 experimental structures of T4L are also show (colored by blue). (b) The structure ensemble selected by EOM with the minimal χ=0.008.

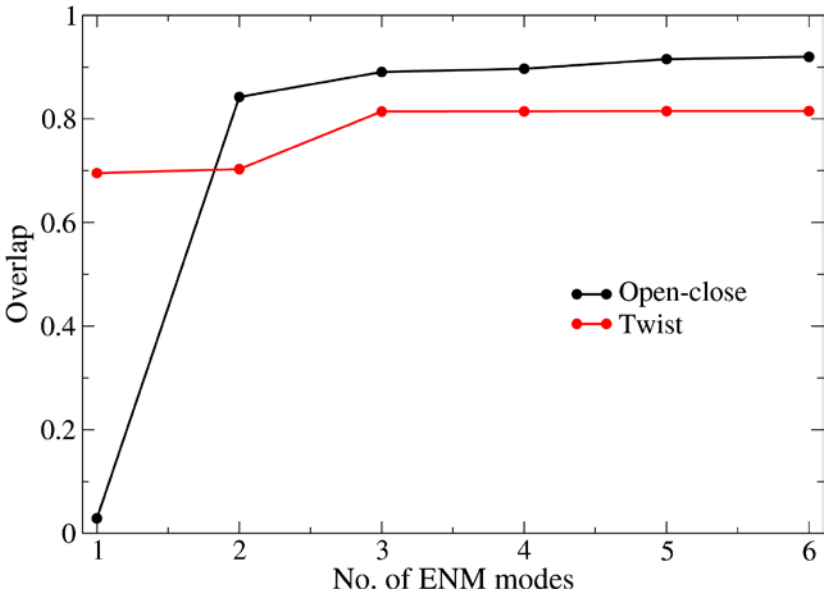

Figure S3. Overlap between the ENM modes and the open-close/twist mode of T4L, respectively. For the open/close or the twist mode, we projected it on the subspace formed by the slowest ENM modes (including from one to six modes, respectively), and obtained the overlap values.

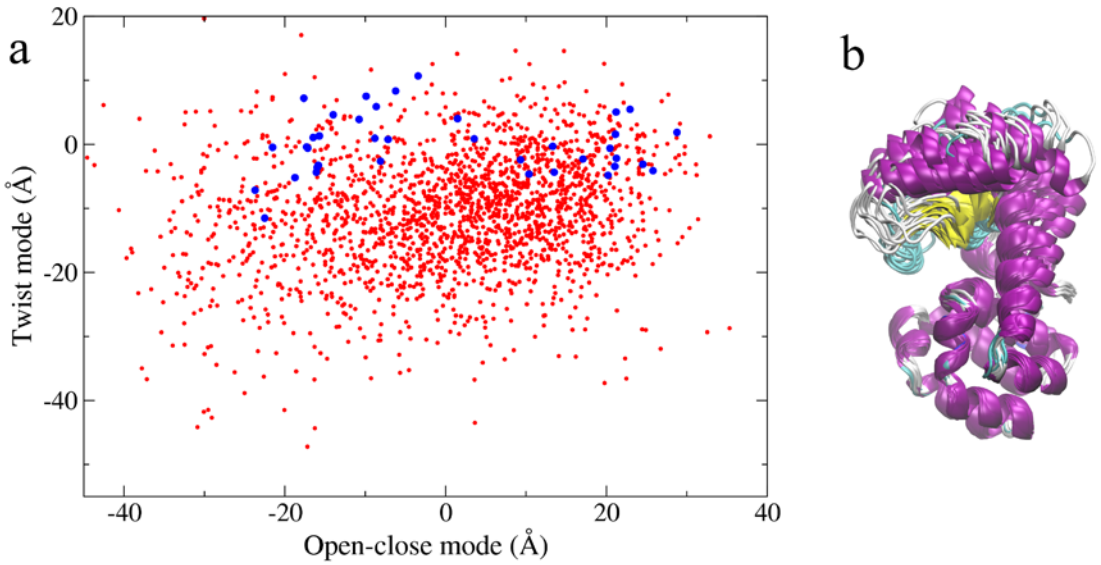

Figure S4. The ACM simulation of T4L that couples the first four ENM modes at 800 K. (a) Conformations in the trajectory are projected onto the 2D essential subspace (colored by red), and the 38 experimental structures of T4L are also show (colored by blue). (b) The structure ensemble selected by EOM, with the minimal χ=0.007.

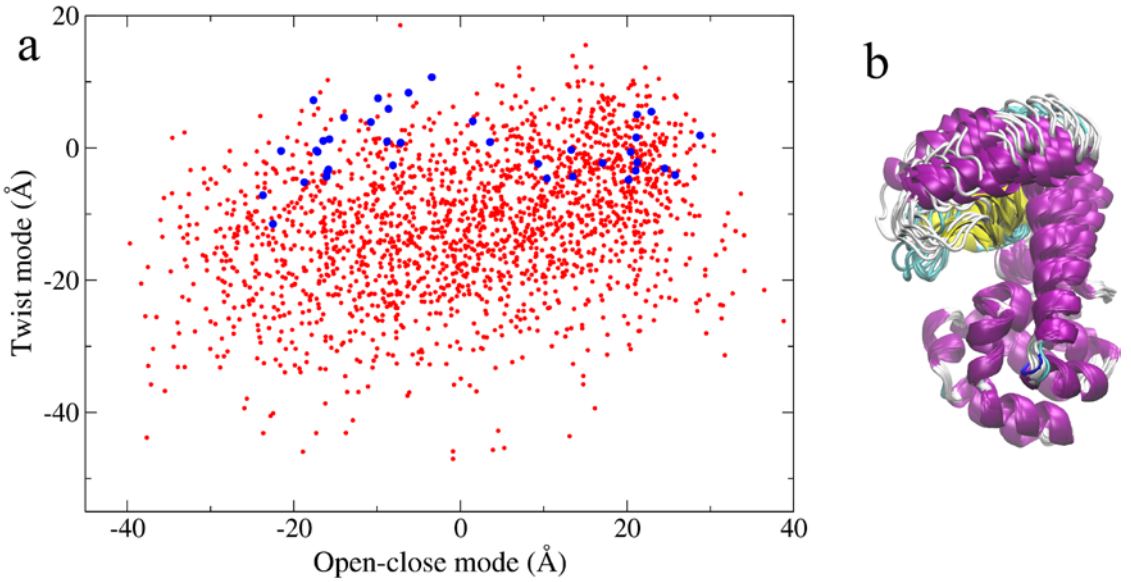

Figure S5. The ACM simulation of T4L that couples the first three ENM modes at 1000 K. (a) Conformations in the trajectory are projected onto the 2D essential subspace (colored by red), and the 38 experimental structures of T4L are also show (colored by blue). (b) The structure ensemble selected by EOM, with the minimal χ=0.008.

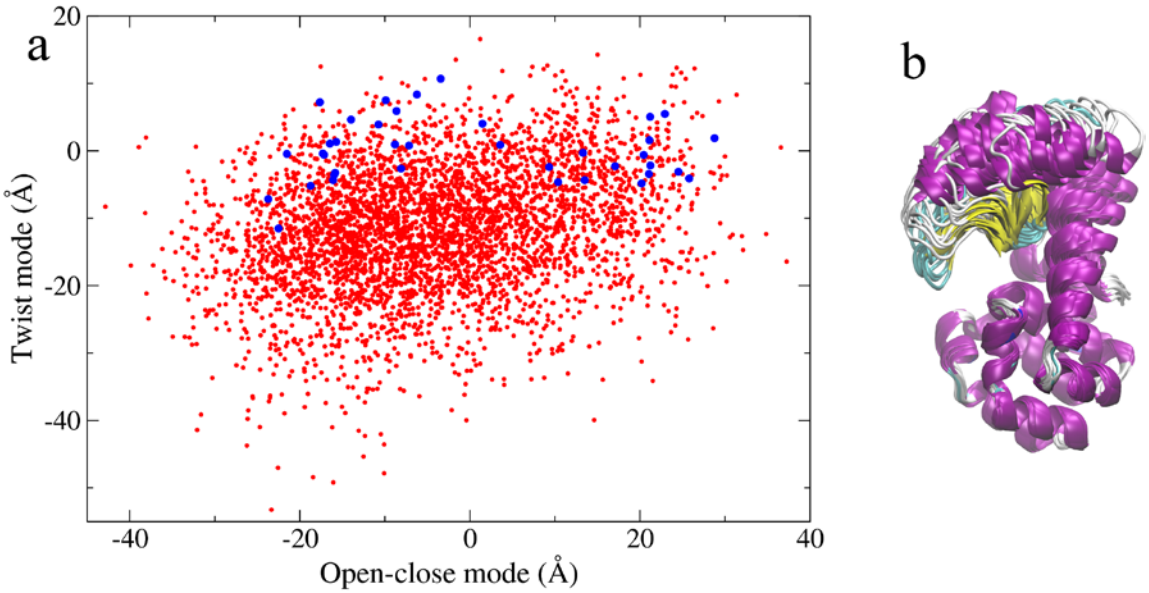

Figure S6. The ACM simulation of T4L with a simulation time of 40 ns. (a) Conformations in the trajectory are projected onto the 2D essential subspace, and (b) the structure ensemble selected by EOM, with the minimal χ=0.006.

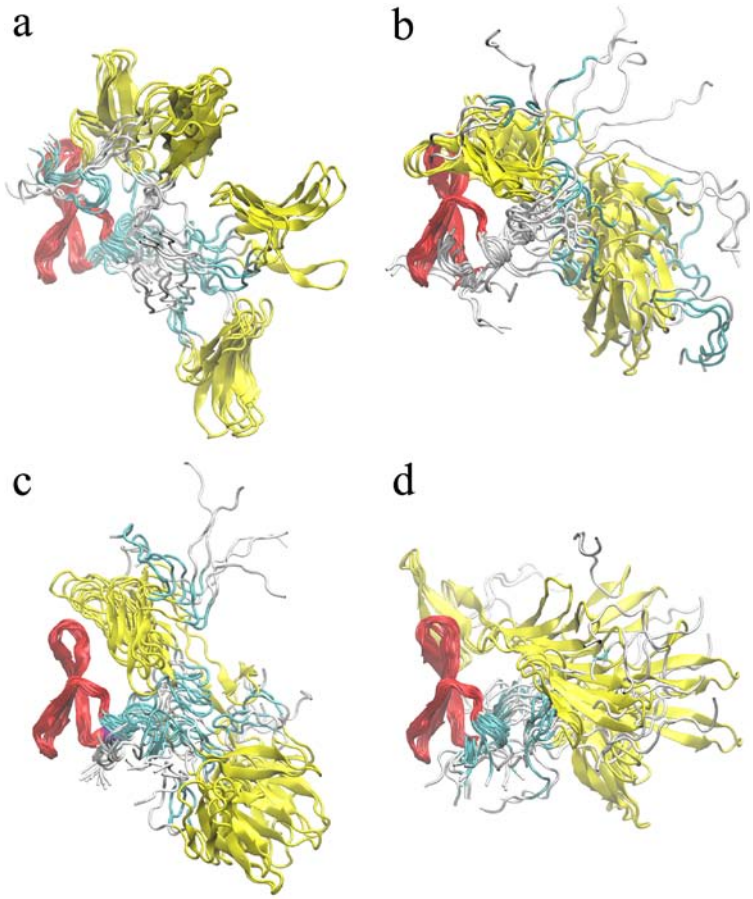

Figure S7. EOM ensembles from multiple ACM simulations of FBP21-WWs. (a) Structure ensemble with the minimal χ=0.165 from the ACM simulation starting from an extended structure of the protein. The simulation parameters were the same as those for Figure 6c. (b) Structure ensemble with the minimal χ=0.165 from the ACM simulation that coupled the first four ENM modes at 500K. (c) Structure ensemble with the minimal χ=0.164 from the ACM simulation that coupled the first three ENM modes at 400 K. (d) Structure ensemble with the minimal χ=0.170 from a 40 ns ACM simulation that is an extension of the original 20 ns simulation (Fig. 6c). The structures are superimposed by the WW1 domain (residues 6-32, colored by red), and the WW2 domain (residues 47-73) is colored by yellow.

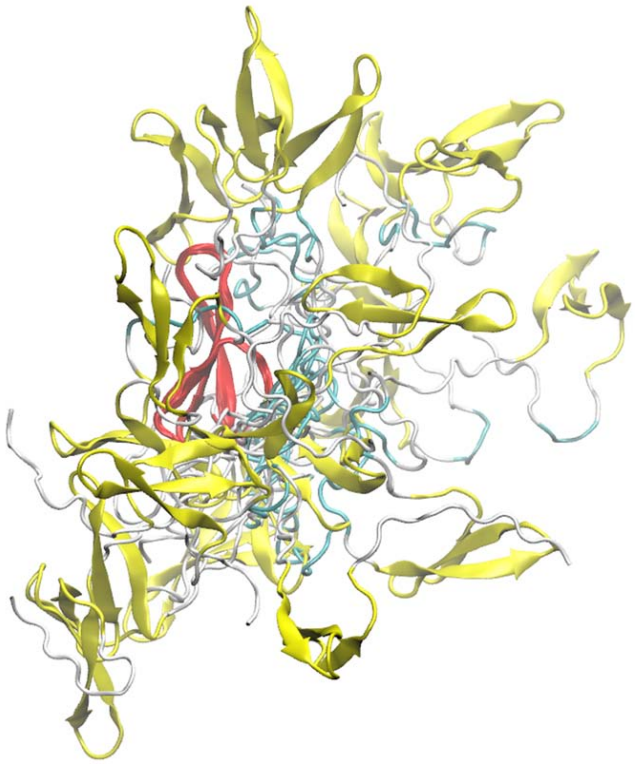

Figure S8. EOM ensemble of FBP21-WWs from the structure pool generated by Pre_bunch, with the minimal χ=0.164. The structures are superimposed by the WW1 domain (residues 6-32, colored by red), and the WW2 domain (residues 47-73) is colored by yellow.