Towards the Structural Characterization of Intrinsically Disordered Proteins by SAXS and MD Simulation

Tomotaka Oroguchi, Mitsunori Ikeguchi, and Mamoru Sato*
1Graduate School of Nanobioscience, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

E-mail: msato@tsurumi.yokohama-cu.ac.jp

Abstract. Dynamical structures of intrinsically disordered proteins (IDPs) and multidomain proteins that include large ID regions between the domains are unable to be determined by such conventional methods as X-ray crystallography and electron microscopy. Small-angle X-ray scattering (SAXS) is suitable to determine low-resolution structures of proteins and protein complexes in solution, but the structural data on protein dynamics are averaged over the structural ensemble in protein solution. To overcome this problem, we have developed a novel method, named MD-SAXS, of the combined use of SAXS and molecular dynamics (MD) simulation to analyze protein dynamics in solution of multimeric protein complexes and multidomain proteins toward the structural characterization of IDPs. Here we show validity of the method through the structural characterization of restriction Endonuclease EcoO109I.

1. Introduction
IDPs are characterized by lack of stable tertiary structure when the protein exists as an isolated polypeptide chain under physiological conditions. They adopt well-defined structures upon binding to their targets (coupled folding and binding). This is a novel target recognition mechanism unlike the “rock and key” and induced-fit models. IDPs are mostly found in eukaryotes and generally interact with several target molecules through the coupled folding and binding, thus playing important roles as protein hubs in intracellular networks [1]. However, structural characterization of IDPs by such conventional methods as X-ray crystallography and electron microscopy is difficult due to the extremely high flexibility of IDPs in solution, thus expecting development of a novel method to determine the IDP structure in solution at an atomic resolution. In view of this consideration, we have developed a novel method, named MD-SAXS of the combined use of SAXS and molecular dynamics (MD) simulation to analyze protein dynamics in solution of multimeric protein complexes and multidomain proteins.

Restriction endonucleases (REases) provide anti-viral protection for bacteria by degrading the foreign DNA of invading bacteriophages [2]. These enzymes recognize specific nucleotide sequences and cleave both strands of DNA. EcoO109I is a type II REase isolated from Escherichia coli and recognizes 7 base pairs of double-stranded DNA (ds-DNA). The crystal structures of both DNA-free and DNA-bound forms of EcoO109I have been determined (Figure 1), and SAXS analysis of the DNA-free form have revealed that EcoO109I consists of dimerization and catalytic domains (Figure 1D) and the homodimer (subunits A and B in Figure 1A) is the functional unit of the enzyme in solution.
A comparison between the crystal structures of the DNA-free and DNA-bound forms of EcoO109I shows that the enzyme undergoes large structural changes upon DNA binding. These structural changes can be explained mainly by two movements. The first is the movement of the two catalytic domains in subunits A and B closer together, allowing interaction with DNA (Figure 1A-C). The second is the twisting motion between the two subunits in the counter-clockwise direction, making the two catalytic domains fit together on the major groove of DNA from both sides (Figure 1D-F).

In this study, we carried out a 150-ns MD simulation for the DNA-free form of the enzyme with the explicit water molecules in order to elucidate the relationship between the intrinsic dynamics of EcoO109I in solution and its structural changes occurring upon DNA binding [3]. From the trajectory of the MD simulation, the SAXS profile, taking scattering from explicit hydration water molecules into account, was calculated and compared with the experimentally observed SAXS profile. The time-resolved analysis of the radius of gyration estimated from the Guinier plot demonstrates that the structure of EcoO109I is considerably flexible. Analyses of atomic fluctuations within the subunit and in the dimer identified two rigid cores that undergo rigid-body motions relative to each other. The motions of the two rigid cores were characterized using principal component analysis, and were compared with the DNA-induced structural changes observed in the crystallographic study [2]. Motions in the catalytic domains were also examined using principal component analysis.

![Image of crystal structures](image)

**Figure 1.** The crystal structures of (A) DNA-free and (B) DNA-bound forms colored according to subunits A (green) and B (purple). Panels (D) and (E) show a side view of the DNA-free and DNA-bound forms, respectively. Panels (C) and (F) show front and side views of the superimposition of these two forms, respectively.

### 2. Methods

#### 2-1: MD simulations

All simulations were performed at 298 K using MARBLE [6]. The initial structure for the simulation of EcoO109I was generated from the crystal structure (protein data bank (PDB) accession code: 1WTD [2]). The protein and crystal water molecules were immersed in a water box, followed by energy minimization of 5000 steps of the whole system. The total number of atoms was 108,249. Two series of product runs were performed under the NPT condition with $T = 298.15$ K and $P = 1.0$ atm: (i) a 150-ns run without any restraints; and (ii) a 20-ns run with restraints on the positions of all the protein atoms to the crystal coordinates. The more details about MD procedures are described in the literature [3].

#### 2-2: Calculation of the SAXS profile

The experimental data of the SAXS profile were taken from the previous study [2]. The calculation procedure is as follows: (i) MD simulations of the pure solvent system $V$ are performed in addition to simulations of the protein solution (protein molecules in solution) system $U$ and (ii) the MD-SAXS excess intensity $I$ is obtained by a subtraction of the pure solvent scattering intensity $I^0$ from the protein solution intensity $I^p$. This procedure is similar to that used in typical SAXS experiments: the two sets of data $I^0$ and $I^p$ are collected and the difference between them is used for further data analysis. $I^0$ and $I^p$ are calculated by using only atoms included in region “v”, which corresponds to...
the area inside a sphere centered at the center of mass of the protein molecule (Figure 2A). As proven in the literature [3], the scattering from region “bo” corresponding to the area outside the sphere and the cross term of scattering between region v and region bo are equivalent between the protein solution and the pure solvent when the radius \( R \) of region o is large enough to include not only the entire solvation layer around the protein but also some of the bulk region. Therefore, those scattering terms become zero by subtraction. Then, the objective MD-SAXS excess intensity \( I(Q) \) is expressed as

\[
I(Q) = I^U(Q) - I^V(Q)
\]

(1)

where

\[
I^\eta(Q) = \langle \tilde{I}(Q) \rangle_\eta
\]

(\( \eta = U, V \)).

\( Q \) is the scattering vector, which is related to the scattering angle \( 2\theta \) and the wavelength of the incident radiation \( \lambda \) by the equation \( Q = |Q| = 4\pi \sin \theta / \lambda \). \( \langle \rangle_U \) and \( \langle \rangle_V \) represent the configurational averages over the ensembles of the protein solution and the pure solvent, respectively. \( \tilde{I}(Q) \) in these brackets is the instantaneous scattering intensity, which is calculated for each snapshot in the trajectories of the protein solution and pure solvent systems using by the following equation:

\[
\tilde{I}(Q) = \left\{ \int_v d^3r' \int_v d^3r (\tilde{\rho}(r') - \rho_0)(\tilde{\rho}(r) - \rho_0)e^{-iQ(r-r')} \right\}_\Omega
\]

(3)

where \( \tilde{\rho}(r) \) is the electron density at position \( r \), and \( \rho_0 \) is the average electron density of the bulk in the protein solution. In this study, the time-averaged density of the solvent in the region bo was used as \( \rho_0 \). \( \langle \rangle_\Omega \) denotes the orientational average. For the fast computation, equation (3) is calculated using the multipole expansion:

\[
\exp(-iQ \cdot r) = 4\pi \sum_{l=0}^\infty \sum_{m=-l}^l i^l j_l(Qr)Y_{lm}^*(\Omega_r)Y_{lm}(\Omega_q),
\]

(4)

where \( j_l \) and \( Y_{lm} \) denotes a spherical Bessel function and a spherical harmonic function, respectively. \( Y_{lm}^* \) is a complex conjugate of \( Y_{lm} \). \( \Omega_r \) is the solid angle in real space, \( \Omega_q \) is the solid angle in reciprocal space, \( r = (r, \Omega_r) \). Then, equation (3) is rewritten as:

\[
\tilde{I}(Q) = 4\pi \sum_{l=0}^\infty \sum_{m=-l}^l i^l f_l(Q)Y_{lm}^*(\Omega_k)Y_{lm}(\Omega_q) \left| j_l(Qs) \right|^2 - 2\rho_0 \sum_{k} f_k(Q) \int_S (s \cdot n) \left( j_k(Qs') \right)^2
\]

\[
+ \rho_0^2 \int_S^2 dS' \int_S dS \left[ \left( \frac{\mathbf{n} \cdot \mathbf{n}}{Q^2} \right)^2 \left( \frac{\mathbf{s}' \cdot \mathbf{n}}{Q^2} \right)^2 \right]
\]

\[
+ \rho_0^2 \int_S^2 dS' \int_S dS \left[ \left( \frac{\mathbf{n} \cdot \mathbf{n}}{Q^2} \right)^2 \left( \frac{\mathbf{s}' \cdot \mathbf{n}}{Q^2} \right)^2 \right]
\]

(5)

The first term of the right hand side of equation (5) denotes the scattering from all explicit atoms within the region v, the second term denotes the cross term of the scattering between these atoms and the bulk solvent in the region bo, and the third term originates from the scattering from the infinite bulk. \( N \) is the number of atoms within the region v. \( r_k \) and \( \Omega_k \) is the position and the solid angle in real space of the \( k \)-th atom, respectively, and \( s_k \) and \( s' \) denotes \( r - r_k \) and \( r - r' \), respectively. \( f_k(Q) \) denotes the atomic form factor of the \( k \)-th atom. S represents the closed surface of the region v, and \( n \) is the unit normal vector outward from the infinitesimal surface element dS of the surface S. The more detailed derivations and definitions of the equations (3) and (5) are described in the literatures [3-5].
3. Results
During the 150-ns MD simulation of the DNA-free form of EcoO109I, the whole structure fluctuated largely. The root-mean-square displacement of the C\(\alpha\) atoms (C\(\alpha\)-RMSD) from the crystal structure of the DNA-free form reached \(~5.0\) Å within the first 20 ns of the simulation. However, the C\(\alpha\)-RMSD of individual rigid cores, i.e., the elements of rigid-body motions, was \(~1.0\) Å during the simulation, indicating that the rigid cores were stable during the simulation, and that there were certain rigid-body motions between them.

3-1: Comparison of the MD Simulations with the Experiment in SAXS data
To test the validity of the structural ensemble sampled by the MD simulation, we calculated the SAXS profile from all the snapshots in the trajectory, taking scattering from the explicit hydration water molecules into account and compared it with the experimental profile. As shown in Figure 3A, the simulation-derived profile was in agreement with the experimentally observed SAXS profile.

In the range of \(R_gQ < 1.3\), the scattering intensity \(I(Q)\) is approximated to (Guinier approximation),

\[
I(Q) \approx I(0) \exp (-Q^2 R_g^2 / 3),
\]

where \(I(0)\) and \(R_g\) are the scattering intensity at \(Q = 0\) (zero-angle) and the radius of gyration, respectively, which are estimated from \(\ln[I(Q)]\) versus \(Q^2\) plot (Guinier plot) [7]. Figure 3B compares the Guinier plots of the simulation-derived and experimentally observed SAXS profiles, showing that the \(R_g\) (referred to as ‘protein-water \(R_g\)’ or simply ‘\(R_g\)’) estimated from the SAXS profile calculated from the structural ensemble sampled by the MD simulation (\(R_g = 28.2\) Å) was in agreement with that obtained from the experimentally observed SAXS profile (\(R_g = 28.1 \pm 0.3\) Å).

We then calculated time-evolution of the protein-water \(R_g\) to investigate the fluctuation of the overall structure in the 150-ns simulation. \(R_g\) (the cyan curve in Figure 3C) exhibited large fluctuation with a period of about 50-100 ns. There was an approximately 1.8-Å difference in \(R_g\) between the most expanded and closed conformations; however, \(R_g\) averaged over all time-windows (the blue line in Figure 3C) is in agreement with \(R_g\) obtained from the experimental profile (the pink horizontal line).

We also calculated time evolution of another \(R_g\) (referred to as ‘protein \(R_g\)’), which was obtained only from the protein structure (the black curve in Figure 3C). The time course of the protein \(R_g\) clearly shows almost the same time dependency as that of the protein-water \(R_g\); this indicates that the fluctuation of the protein-water \(R_g\) reflects the fluctuation of the protein moiety, and a 2-Å difference between the protein-water \(R_g\) and the protein \(R_g\) is attributable to the hydration shell on the surface of the protein molecule.

In order to analyze the effects of structural changes on the protein-water \(R_g\) during the 150-ns MD simulation, we performed another 20-ns MD simulation in which positions of the protein atoms were restrained to the original crystal positions and only the solvent atoms were sampled (referred to as the ‘restraint-MD simulation’). As shown in Figure 3C (the yellow-green curve), we calculated the protein-water \(R_g\) for the time window of 500 ps shifting from \(t = 0-20\) ns for the restraint-MD
simulation. In this simulation, the protein-water $R_g$ was ~27.6 Å, and the average of the protein-water $R_g$ (the green horizontal line in Figure 3C) was still smaller than the experimental value. The difference in the protein-water $R_g$ between the restraint-MD simulation and the experimental value indicates that EcoO109I adopts a more expanded conformation in solution, as observed in the 150-ns MD simulation, than in crystalline state. All these results show the validity of the structural ensemble sampled by the 150-ns MD simulation. In the following section, we investigate the detailed dynamics of EcoO109I during the 150-ns MD simulation.

Figure 3. Comparison of the experimental and simulation-derived SAXS profiles of the DNA-free form of EcoO109I in solution. The SAXS profiles from the experiments (pink) and the 150-ns MD simulation (blue) are shown in Logarithmic plot (A) and Guinier plot (B). Panel (C) shows time evolution of the protein-water $R_g$ for the 150-ns MD trajectory (cyan) and the 20-ns restraint-MD trajectory (yellow-green). The pink line shows the protein-water $R_g$ estimated from the experiment. The protein $R_g$ without water molecules was also calculated for the 150-ns trajectory (black).

3-2: Intrinsic Dynamics of DNA-free Form of EcoO109I

For further analyses of structural fluctuations of EcoO109I in the simulation, the principal component analysis was applied for the 150-ns trajectory. The first two principal modes accounted for the dominant portion of the total fluctuations of this protein (Figure 4A).

The eigenvector of the first principal mode represents the open-close motion (the red arrows in Fig. 4 D). The time course of projections onto the first principal mode is shown in Figure 4B, in which positive and negative projection values correspond to open and closed conformations, respectively. Because the $R_g$ of the open conformations were larger than those of the closed conformations, there was a significant correlation in the time courses between the projections onto the first principal mode and the time-dependent $R_g$ shown in Figure 3C (the correlation coefficient = 0.74). Moreover, we found that the open-close motion represented by the first principal mode changed the width of the intervening space between the two catalytic domains, $d_{bc}$ (Figure 4B-C). In Figure 6E, the superposition of the structures with the largest and smallest projections onto the first principal mode is shown. The structure with the largest projection value (green in Figure 4E) has a $d_{bc}$ of 20.9 Å, which is slightly larger than the diameter of DNA (20 Å). In contrast, the structure with the smallest projection value (pink in Figure 4E) has a $d_{bc}$ of 5.2 Å, which is close to the crystal structure in the DNA-bound form (2.7 Å). Because the $d_{bc}$ of the crystal structure in the DNA-free form is 6.4 Å (smaller than t DNA), the open-close motion represented by the first principal mode is necessary for the enzyme to access DNA.
Figure 4. Principal component analysis for motions in the 150-ns MD trajectory. Panel (A) shows individual contributions of the first 10 principal components to the total fluctuations. (B) Snapshots in the trajectory were projected onto the first principal mode. The DNA-free (red open circle) and DNA-bound (blue open circle) crystal structures are also projected onto the first principal mode. (C) The distance between the two catalytic domains, $d_{bc}$, in the 150-ns simulation is shown. (D) The amplitude and direction of the first principal mode are represented by red arrows on the average structure. (E) The superposition of the structures with the largest projection (green) and smallest projection (pink) onto the first principal mode.

The eigenvector of the second principal mode represents the twisting motion between the two subunits in the counter-clockwise direction (the red arrows in Figure 5B). The superposition of the structures with the largest and smallest projections is also shown in Figure 5C. As seen in the comparison of Figure 1F with Figure 5C, the direction of the second principal mode was significantly similar to the structural changes occurring upon DNA binding. Indeed, the correlation coefficient in the directions between the second eigenvector and the structural change in the DNA-free and DNA-bound crystal structures is 0.90 (Figure 5B). This indicates that structural changes upon DNA binding clearly correlate with the intrinsic dynamics of EcoO109I. In other words, the relative arrangement between the two subunits is specifically flexible in the counter-clockwise direction in the DNA-free state, and this flexibility effectively facilitates structural changes upon DNA-binding.

4. Conclusion

To elucidate the relationship between the intrinsic dynamics of EcoO109I and its function, we performed a 150-ns MD simulation of the DNA-free form of the enzyme and examined the resultant trajectory in detail. We calculated the SAXS profile from all the snapshots in the trajectory, taking scattering from the explicit hydration water molecules into account. The calculated SAXS profile is in agreement with the experimentally observed SAXS profile, and a difference between the protein-water $R_g$ and the protein $R_g$ is 2 Å and coincides with the hydration shell width on the surface of the protein molecule. These results support the validity of the structural ensemble sampled by the present simulation. In addition, a large fluctuation in $R_g$ was observed in the simulation, which indicates considerable flexibility in the overall structure of the enzyme.

From the principal component analysis applied to the trajectory, we have identified intrinsically functional motions in the DNA-free form of EcoO109I. The motion represented by the first principal mode is necessary for the intervening space between the catalytic domains to be accessed by DNA. The motion represented by the second principal mode has a significant correlation with the structural changes upon DNA binding, and is therefore necessary for EcoO109I to effectively achieve functionality. These two modes are dominant fluctuations in the DNA-free state.

Our present study using the MD-SAXS has revealed that the combination of SAXS and MD simulations is very powerful to identify the structural ensemble at an atomic resolution, and thus
elucidate the functional motions of the enzyme. In recent studies, we have extended our MD-SAXS approach to make it possible to apply it to the flexible proteins, the crystal structures of which are very difficult to be obtained. One example of such a protein is the e-subunit of F1-ATPase, which takes the extended form in the absence of ATP. We have revealed the structural ensemble of the ATP-free form of the e-subunit by our extended MD-SAXS approach (data not shown). Our approach can be also applied to determine multi-subunit protein complexes and multi-domain proteins in solution at an atomic resolution. We are now extending our method to analyze the extremely flexible proteins, IDP.

Figure 5. The second principal mode. (A) Snapshots in the trajectory were projected onto the second principal mode. (B) The amplitude and direction of the second principal mode are represented by red arrows on the average structure. The structural change between the DNA-free and bound forms is represented by cyan arrows. (C) The superposition of the structures with the largest and smallest projections onto the second principal mode.

Acknowledgements
This study was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), grants from the Japan Science and Technology Agency, and Research and Development of the Next-Generation Integrated Simulation of Living Matter, a part of the Development and Use of the Next-Generation Supercomputer Project of MEXT. All computations were carried out at Tsurumi campus of the International Graduate School of Arts and Sciences, Yokohama City University, Yokohama, Japan.

References
[1] Uversky V N and Dunker A K 2010 Biochimica et Biophysica Acta 1804 1231
[2] Hashimoto H, Shimizu T, Imasaki T, Kato M, Shichijo N, Kita K and Sato M 2005 J. Biol. Chem. 280 5605
[3] Oroguchi T, Hashimoto H, Shimizu T, Sato M and Ikeguchi M 2009 Biophys. J. 96 2808
[4] Seki Y, Tomizawa T, Khechinashvili N N and Soda K 2002 Biophys. Chem. 95 235
[5] Soda K, Miki Y, Nishizawa T and Seki Y 1997 Biophys. Chem. 65 45
[6] Ikeguchi M 2004. J. Comput. Chem. 25 529
[7] Guinier A 1939 Ann. Phys. 12 161