JOURNAL OF
SYNCHROTRON
RADIATION

ISSN 1600-5775

Classification of *ab initio* models of proteins restored from small-angle X-ray scattering

Mao Oide,[a,b] Yuki Sekiguchi,[a,b] Asahi Fukuda,[a,b] Koji Okajima,[a,b] Tomotaka Oroguchi[a,b] and Masayoshi Nakasako[a,b]*

[a]Department of Physics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan, and [b]RIKEN SPring-8 Center, 1-1-1 Kouto, Sayo-cho, Sayo-gun, Hyogo 679-5148, Japan. *Correspondence e-mail: nakasako@phys.keio.ac.jp

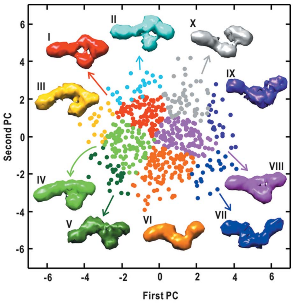
Received 9 February 2018
Accepted 18 July 2018

Edited by M. Yamamoto, RIKEN SPring-8 Center, Japan

Keywords: small-angle X-ray scattering; *ab initio* molecular modeling; multivariate analysis; low-resolution structure; protein structure.

Supporting information: this article has supporting information at journals.iucr.org/s

In structure analyses of proteins in solution by using small-angle X-ray scattering (SAXS), the molecular models are restored by using *ab initio* molecular modeling algorithms. There can be variation among restored models owing to the loss of phase information in the scattering profiles, averaging with regard to the orientation of proteins against the direction of the incident X-ray beam, and also conformational fluctuations. In many cases, a representative molecular model is obtained by averaging models restored in a number of *ab initio* calculations, which possibly provide nonrealistic models inconsistent with the biological and structural information about the target protein. Here, a protocol for classifying predicted models by multivariate analysis to select probable and realistic models is proposed. In the protocol, each structure model is represented as a point in a hyper-dimensional space describing the shape of the model. Principal component analysis followed by the clustering method is applied to visualize the distribution of the points in the hyper-dimensional space. Then, the classification provides an opportunity to exclude nonrealistic models. The feasibility of the protocol was examined through the application to the SAXS profiles of four proteins.



© 2018 International Union of Crystallography

1. Introduction

Small-angle X-ray scattering (SAXS) from proteins in solution provides their molecular weights, radii of gyration, pair-correlation functions and molecular shapes at low resolution (Glatter & Kratky, 1982; Svergun *et al.*, 2013). SAXS for solution specimens enables observations of the conformational changes of macromolecules upon physical and chemical stimuli. Recent developments in SAXS measurements using size-exclusion chromatography have further extended the capability for specimens displaying concentration-dependent aggregation (Watanabe & Inoko, 2009; Graewert *et al.*, 2015).

SAXS profiles lack the phase terms of the scattered waves. Furthermore, SAXS profiles are the average over both the orientation of macromolecules against the direction of the incident X-ray and the variation in the conformations of molecules during the exposure. Therefore, the information obtainable from the SAXS profile is insufficient to reconstruct the three-dimensional structural models of proteins. To discuss the low-resolution molecular structure of a protein from a SAXS profile, *ab initio* algorithms have been developed. The algorithms minimize the discrepancy between experimental and calculated scattering profiles under restraints to maintain a compactly interconnected configuration of beads or dummy residues (DRs) (Chacón *et al.*, 1998; Svergun, 1999; Svergun *et al.*, 2001; Franke & Svergun, 2009). In the last two decades, *ab initio* algorithms have contributed to the vast application of

https://doi.org/10.1107/S1600577518010342

SAXS for structural biology studies of proteins in solution (Jeffries & Svergun, 2015). In the *ab initio* calculations, various molecular shapes are restored rather than a unique shape owing to the lack of information in the SAXS profile. Then, a representative molecular model is obtained by averaging all models that are restored from a single scattering profile (Volkov & Svergun, 2003), and the effective resolution of the models is estimated by the Fourier shell correlation (Tuukkanen *et al.*, 2016). However, nonrealistic models would appear in a number of calculation trials, while some restored models can approximate the molecular shape of the protein. When a representative model is obtained by averaging all restored models, nonrealistic models have nontrivial influences to blur the details of realistic models. Therefore, it is necessary to select probable models before averaging. In addition, for good statistics, it would be better to select from more than a few hundred models rather than a few tens.

The aim of this study was objective extraction of the most probable and realistic models from a large number of models restored from SAXS profiles. Therefore, we proposed a protocol to classify and characterize those restored models by using multivariate analysis including principal component analysis (PCA) and *K*-means clustering. Then, the protocol suggests groups of plausible molecular models instead of only a model averaged over both the realistic and nonrealistic models, and provides an opportunity to discuss structures of proteins in solution in referring biochemical and other structural shows. Here, we describe the details of the theoretical background and examples of the application of this protocol to experimental data.

2. Computational method

In this study, we used models restored by the *GASBOR* program (Svergun *et al.*, 2001). Each model is composed of a number of DRs with a diameter of 3.8 Å. To classify a large number of restored models, the proposed protocol takes the following three steps: (i) superimposition of each model to a reference, (ii) voxelization of the superimposed models to be expressed as points in a hyper-dimensional space, and (iii) multivariate analysis for the classification of the models (Fig. 1). After the three steps, the averaged model for each class is calculated to visualize the representative molecular shapes. The detailed procedure and algorithm used in each step are described in the following sections.

2.1. Superimposition of models

At first, we superimpose all restored models to a randomly selected reference based on the idea used in the program *SUPCOMB* (Kozin & Svergun, 2001). For the superimposition, the center of gravity of each model is placed onto that of the reference. The model is then rotated to maximize the overlap with the reference using a rotation matrix determined by aligning the inertia axes of the model to those of the reference. The inertia tensor **I** of a model is defined as

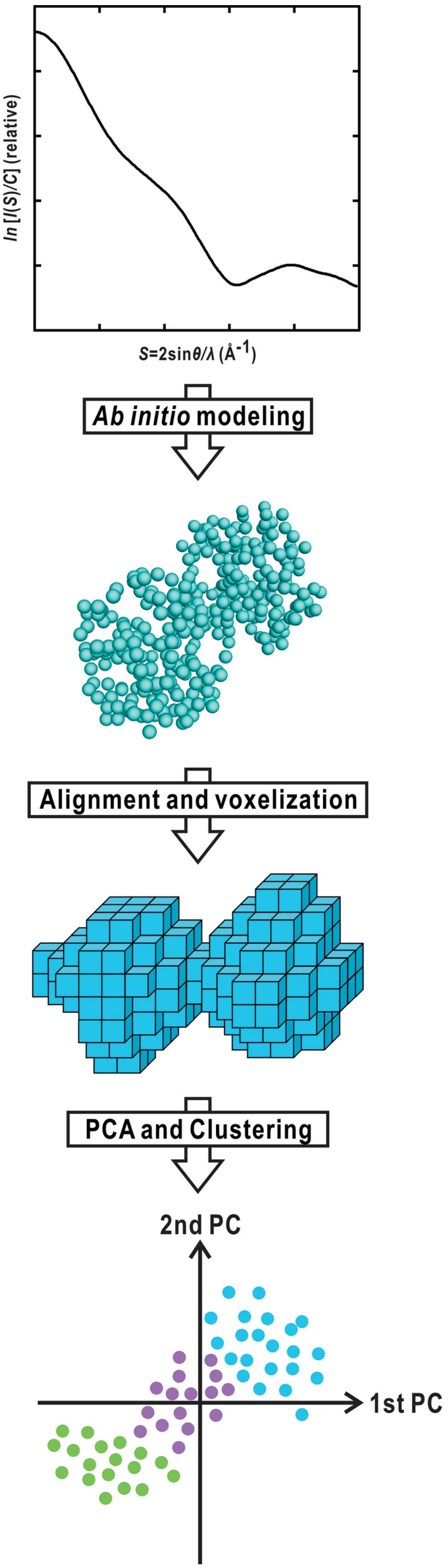


Figure 1
Schematic illustration of the multivariate analysis, which is applied to the molecular models restored by an *ab initio* molecular modeling algorithm from a single SAXS profile.

$$\mathbf{I} = \begin{pmatrix} \sum\limits_{i=1}^{N}(y_i^2+z_i^2) & -\sum\limits_{i=1}^{N}x_iy_i & -\sum\limits_{i=1}^{N}z_ix_i \\ -\sum\limits_{i=1}^{N}x_iy_i & \sum\limits_{i=1}^{N}(z_i^2+x_i^2) & -\sum\limits_{i=1}^{N}y_iz_i \\ -\sum\limits_{i=1}^{N}z_ix_i & -\sum\limits_{i=1}^{N}y_iz_i & \sum\limits_{i=1}^{N}(x_i^2+y_i^2) \end{pmatrix}, \qquad (1)$$

where (x_i, y_i, z_i) is the position of the ith DR in the model composed of N DRs. The diagonalization of matrix $\mathbf{I}$ gives the principal axis of the moment of inertia. In the superimposition, arbitrariness regarding eight possible orientations arises from the handedness and four sign combinations of the eigenvectors of the inertia axes. To determine the orientation of the target model against the reference, the similarity between the models is measured by using the normalized spatial discrepancy (Kozin & Svergun, 2001). We did not apply any further refinement procedures with respect to the relative orientation of the model against the reference.

After the superimposition, all models are set in a box composed of J voxels large enough to cover all models, and the number of DRs contained in each voxel is counted. Then, each model is expressed as a set of J voxels accompanying the number densities of DRs. For instance, the number density in the ith voxel is defined as ρ_i, and thus the model represents a point $(\rho_1, \rho_2, \ldots, \rho_J)$ in the J-dimensional space.

The voxel size L is defined so that the number density of DRs in each voxel becomes more than a few. For the average volume V of all models, L is calculated by the following equation,

$$L^3 J \simeq V. \qquad (2)$$

In this study, we used a voxel size L of 6 Å, although the voxel size would be varied depending on the dimensions of molecules. Finally, a Gaussian low-pass filter with a standard deviation of 4 Å is applied to each model to blur local variation in the shapes of the models, which would disrupt the following multivariate analyses.

2.2. Multivariate analysis

Prior to the classification, in order to visualize the distribution of models in a low-dimensional space with minimal loss of information, we apply PCA to the points representing models in the J-dimensional space. First, matrix P is calculated for n models comprising J voxels as

$$\mathrm{P} = \begin{pmatrix} \rho_{11}-\langle\rho_1\rangle & \rho_{12}-\langle\rho_2\rangle & \ldots & \rho_{1J}-\langle\rho_J\rangle \\ \rho_{21}-\langle\rho_1\rangle & \rho_{22}-\langle\rho_2\rangle & \ldots & \rho_{2J}-\langle\rho_J\rangle \\ \ldots & \ldots & \ldots & \ldots \\ \rho_{n1}-\langle\rho_1\rangle & \rho_{n2}-\langle\rho_2\rangle & \ldots & \rho_{nJ}-\langle\rho_J\rangle \end{pmatrix}, \qquad (3)$$

where ρ_{ij} is the number density of DRs at the jth voxel of the ith model, and $\langle\rho_j\rangle$ is the average for the jth voxel among the n models. To find a low-dimensional space suitable for the illustration of the distribution, the eigenvalues and eigenvectors of the variance-covariance matrix $\mathbf{D} = \mathbf{P}^t\mathbf{P}/n^2$ are calculated. The eigenvector with the largest eigenvalue represents the direction along which models are distributed with the maximum variance. When the H ($H \ll J$) largest eigenvalues have significant contribution to the total variance in the distribution of models in the J-dimensional space, the distribution of models can be characterized in the space spanned by the H eigenvectors. Then, the position of each model is given by its projection onto the H-dimensional space. In many cases, the plane spanned by the eigenvectors with the first and second highest eigenvalues is sufficient to characterize the distribution of models (see the *Results* section).

Table 1
Statistics in the structure analyses by the proposed protocol.

Parameters	PDI	P2	P1L1	LphyA
$S_{\max}$ (Å^{-1})	0.064	0.025	0.050	0.038
R_g (Å)†	33.4	48.4	23.5	57.1
$D_{\max}$ (Å)‡	127	188	100	181
Ambiguity score§	2.46	2.1	0.00	1.65
Protein stoichiometry	Monomer	Dimer	Dimer	Dimer
Number of amino acid residues / subunit	483	915	150	1072
Number of DRs / subunit	450	600	110	1000
Number of voxels used	7020	5940	1989	23800
Calculation time (min)¶	85	100	12	563

† Determined from Guinier's plot. ‡ Maximum dimension determined from the distance distribution function. § The ambiguity score was calculated by using *AMBIMETER* (Petoukhov & Svergun, 2015). ¶ Including the running time of parallelized *GASBOR* calculation.

The K-means clustering method (MacQueen, 1967) is used to classify the models in the H-dimensional space by minimizing the sum of the squared distances between the models and the centroids of the classes, defined as

$$T = \sum_{k=1}^{M} \sum_{u_{ik} \in k} \left(u_{ik} - \langle u_k \rangle\right)^2, \qquad (4)$$

where u_{ik} is the H-dimensional vector specifying the position of the ith model belonging to the kth class, and $\langle u_k \rangle$ is the centroid of the kth class among the M groups. Because classification depends on the distribution of initial centroids given randomly for assumed classes, the best clustering result displaying the minimum T was selected from 100 independent trials.

Finally, the molecular model representative of each class is provided by averaging the restored models within the class, and is then visualized as a three-dimensional map regarding the number density of DRs.

3. Experimental data and *ab initio* calculation

In this study, we examined the validity of the proposed protocol in the structure analyses for experimental SAXS profiles. We did not apply the protocol to calculated profiles from crystal structures, because the profiles lack experimental errors. SAXS profiles used here were obtained from the following four proteins (Table 1): protein disulfide isomerase (PDI) from thermophile fungi (Nakasako *et al.*, 2010), *Arabidopsis* phototropin2 in the dark state (P2) (Oide *et al.*, 2018), light-oxygen-voltage sensing domain 1 (LOV1) of *Arabidopsis* phototropin1 in the dark state (P1L1) (Nakasako *et al.*, 2004, 2008) and large phytochrome A of pea in the red-light-absorbing form (LphyA) (Nakasako *et al.*, 2005). The SAXS profiles were obtained at beamline BL40B2 or BL45XU of SPring-8, Japan. The pair-correlation function $P(r)$ from each SAXS profile was calculated by using the *GNOM* program (Svergun, 1992). The *AMBIMETER* program (Petoukhov & Svergun, 2015) was used to calculate the ambiguity score of each SAXS profile (Table 1).

The low-resolution molecular models were restored by using the *GASBOR* program on a high-performance computing cluster composed of 576 cores of Intel Xeon CPU X5690 (3.47 GHz per core). The discrepancy between the experimental ($I_{\rm exp}$) and calculated ($I_{\rm model}$) scattering profiles was monitored by using

$$\chi^2 = \frac{1}{N_{\rm D}-1}\sum_{j=1}^{N_{\rm D}}\left[\frac{I_{\rm exp}(S_j) - CI_{\rm model}(S_j)}{\sigma(S_j)}\right]^2, \qquad (5)$$

where $N_{\rm D}$ is the number of data points, C is a scale factor and $\sigma(S_j)$ is the statistical error of $I_{\rm exp}(S_j)$ at the scattering vector S_j.

The oligomeric state of a specimen protein is determined by the size exclusion chromatography used in specimen preparation and the zero-angle scattering intensity in SAXS. This structural information is used as a constraint in the *GASBOR* calculation. In this study, twofold symmetry was assumed in the calculation for dimeric P2, P1L1 and LphyA. Even when proteins are composed of more than four identical subunits, we can conduct *GASBOR* calculations under all possible symmetry. Then, the classification protocol would be applied to all models or a part of them displaying the acceptable χ^2 values.

The net electron density of a protein molecule contributing to SAXS depends on both the contrast of the electron density to solvent region (Ibel & Stuhrmann, 1975) and the conformational fluctuations within the molecule. Therefore, the optimum number of DRs in the *GASBOR* calculation can differ from that of the amino acid sequence. Models composed of an optimum number of DRs are expected to give the smallest χ^2 value as an average. Here, the search for the optimum number was conducted by decreasing the number with a step of 25 from the number in the amino acid sequence of the target protein (Table 1). For a given number of DRs, the *GASBOR* calculations were independently conducted 20 times.

With the optimum number of DRs, 576 *GASBOR* calculations were independently carried out. The total computational time for the *GASBOR* calculation for each protein is listed in Table 1. The proposed protocol was applied to the restored 576 models, and then, in this study, the models were classified into ten classes. The averaged model of each class was compared with the crystal structures or homology models. For the assessment of restored models for PDI, a crystal structure of yeast PDI (Tian *et al.*, 2006) [the accession code in the protein data bank (PDB) is 2b5e] was used. For the restored molecular models of P2, we superimposed crystal structures of *Arabidopsis* phot2 LOV1 (Nakasako *et al.*, 2008) (PDB accession code: 2z6d), *Arabidopsis* phot2 LOV2 (Christie *et al.*, 2012) (PDB accession code: 4eep) and a homology model of the phot2 kinase domain (Oide *et al.*, 2018) made from a cAMP-dependent protein kinase (Akamine *et al.*, 2003) (PDB accession code: 1j3h). The restored models of P1L1 were compared with a crystal structure of P1L1 (Nakasako *et al.*, 2008) (PDB accession code: 2z6c). In the case of LphyA, we compared the SAXS models with a dimeric bacterial phytochrome (Bellini & Papiz, 2012) (PDB accession code: 4gw9), and a dimeric C-terminal part of tyrosine kinase (Childers *et al.*, 2014) (PDB accession code: 4q20), which were homology models of the light-receiving fragment, and the histidine kinase-like domain of LphyA, respectively. Scattering profiles of the atomic models fitted to the SAXS models were calculated by using the *CRYSOL* program (Svergun *et al.*, 1995) to compare with experimental profiles.

Table 2
Statistics in classification of *ab initio* models.

	PDI	P2	P1L1	LphyA
Number of voxels	7020	5940	1989	23800
Cluster	Number of models / population (%) / average χ^2†			
I	93 / 16.1 / 3.3	14 / 3.6 / 2.1	75 / 13.4 / 3.6	116 / 20.7 / 5.6
II	23 / 4.0 / 3.4	66 / 17.2 / 1.8	164 / 29.3 / 3.5	110 / 19.6 / 5.6
III	20 / 3.5 / 3.2	85 / 22.1 / 1.7	77 / 13.8 / 3.5	19 / 3.4 / 5.6
IV	91 / 15.8 / 3.4	64 / 16.7 / 1.7	2 / 0.4 / 4.8	23 / 4.1 / 5.5
V	25 / 4.3 / 3.3	62 / 16.1 / 1.7	46 / 8.2 / 3.9	23 / 4.1 / 5.6
VI	104 / 18.1 / 3.4	15 / 3.9 / 1.9	79 / 14.1 / 3.9	75 / 13.4 / 5.5
VII	23 / 4.0 / 3.4	8 / 2.1 / 1.6	9 / 1.6 / 4.6	61 / 10.9 / 5.6
VIII	120 / 20.8 / 3.4	32 / 8.3 / 1.8	50 / 8.9 / 3.5	77 / 13.8 / 5.6
IX	22 / 3.8 / 3.6	10 / 2.6 / 1.5	43 / 7.7 / 3.5	19 / 3.4 / 5.7
X	55 / 9.5 / 3.3	28 / 7.3 / 1.9	15 / 2.7 / 4.1	37 / 6.6 / 5.6

† The value was averaged over the χ^2 values of models in comparison with an experimental profile.

4. Results

In this section, we described the results of the classification of structural models restored from experimental SAXS profiles of the four proteins by the proposed protocol. For each protein, first we showed averaged models of ten classes separated by the K-means clustering after PCA. Then, we selected the probable and realistic models by referring biochemical data, the crystal structures of whole or domains of the protein, and the similarity of calculated SAXS profiles of the model to the observed one.

4.1. Molecular shape of PDI

The molecular models of PDI were restored in variety as expected from the large ambiguity score [Fig. 2(*a*), Tables 1 and 2]. The first and second principal components (PCs) accounted for 18% of the total variance of the distribution in the 7020-dimensional space (see Fig. S1 of the supporting information). Although the differences in the χ^2 values were small among the ten classes distributed in a circular area in the plane spanned by the two PCs (Table 2), the averaged models of the ten classes displayed substantial differences in the shapes [Figs. 2(*b*) and 2(*c*)].

The averaged models were then compared with the crystal structure of yeast PDI, which comprises four thioredoxin-fold domains, *a*, *b*, *b′* and *a′*, arranged in a J-shape (Fig. 2*a*). Normal mode analysis of yeast PDI predicts that domains *a*, *b* and *b′* arranged in a triangular shape collectively rotate against the *a′* domain, which is connected to the *b′* domain by a flexible loop (Nakasako *et al.*, 2010). Therefore, the fitness of the *a*–*b*–*b′* region to a part of a molecular model is one of the important factors to determine which models are probable. In addition, rearrangement of the *a′* domain from the crystal structure would be necessary in the fitting. Then, we constructed atomic models as illustrated in Fig. 2(*c*).

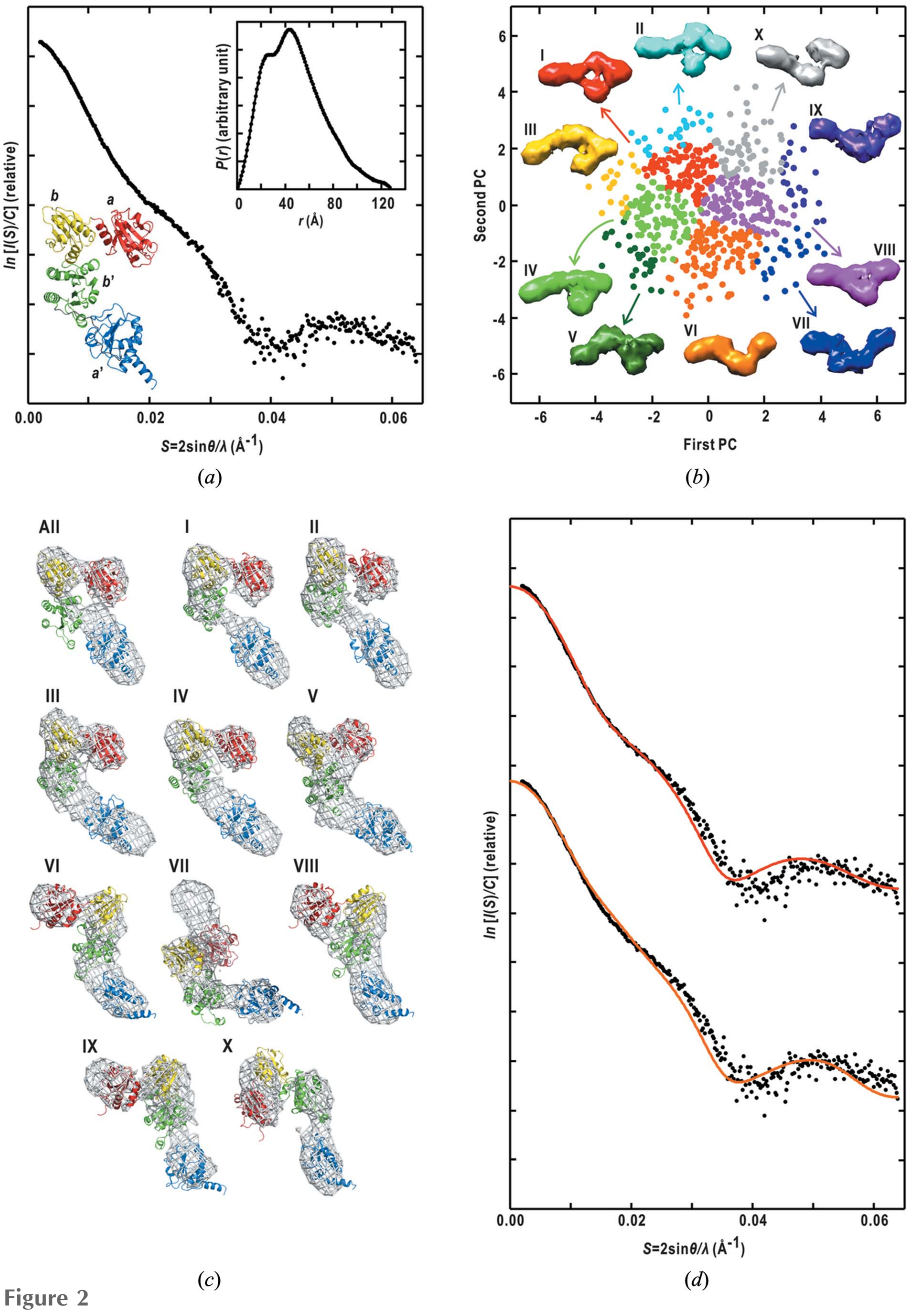


Figure 2
(*a*) Experimental SAXS profile of PDI and the *P*(*r*) function (inset). The crystal structure model of the yeast PDI is shown in the inset. The *a*, *b*, *b′* and *a′* domains are colored red, yellow, green and blue, respectively. (*b*) Distribution of the restored models in the plane spanned by the first and second PCs. The colors of dots indicate classes I–X after the classification. The averaged shape of molecular models in each class is shown. The averaged model of each class is drawn by using the *Chimera* program suite (Pettersen *et al.*, 2004; Goddard *et al.*, 2007). (*c*) Superimposition of the *a*, *b*, *b′* and *a′* domains in the crystal structure of yeast PDI onto the averaged models of all without classification, and of ten classes. In each superimposition, while the *a*–*b*–*b′* region was treated as a rigid body, the *a′* domain was placed at a position different from that in the crystal structure. The viewing directions of the models in panel (*c*) are the same as those in (*b*). (*d*) Comparison of the scattering profiles calculated from fitted atomic models of class I (upper) and VI (lower) in panel (*c*) with the experimental one (black dots).

The molecular shape averaged for all restored models had very small density at the position of the *b′* domain. The *a*–*b*–*b′* region of the crystal structure fit well with the averaged shapes of classes I and II, but to a lesser extent with those of classes III, IV and V. In these five classes, the *a′* domains are located at positions different from the crystal structure, suggesting the positional flexibility of the *a′* domain against the *a*–*b*–*b′* region. Models of classes VI–X were inconsistent with the triangular shape of the *a*–*b*–*b′* region. As a result, most of models in classes I–IV were accidentally characterized by the negative values of the first PC and the positive values of the second PC. It is interesting that the sign of the PC values would be a case-dependently good indicator for the correctness of models (Fig. 2*b*).

Fig. 2(*d*) shows theoretical scattering profiles calculated for two atomic models prepared for classes I and VI (Fig. 2*c*). Inspecting the χ^2 and radius of gyration (R_g) values (Table 3), the models were divided into a group composed of classes I–V with the negative first PC values and the other of classes VI, VIII and IX with the positive first PC values. A scattering profile of class I representing the former group was better than those of the latter regarding the similarity to the experimental profile in the small-angle region of $S < 0.025$ Å^{-1} (Fig. 2*d*).

Based on the plausibility of the models in real space and the similarity of SAXS profiles in reciprocal space, classes I–V were likely to simulate the molecular structure of PDI in solution. Thus, in the case of PDI, the classification protocol contributed to better selection of the probable molecular shapes rather than the ordinary method of averaging all models.

4.2. Molecular shape of P2

P2 is composed of two blue-light-receiving LOV domains (LOV1 and LOV2) and one kinase domain, and forms a dimer at LOV1 in solution (Oide *et al.*, 2018) (Fig. 3*a*). Therefore, the twofold symmetry was assumed in the *GASBOR* calculation for P2 (Fig. 3*a*). The ambiguity score of the SAXS profile was comparable with those of PDI (Table 1).

The optimum number of dummy residues determined for the calculation showed a discrepancy from the actual number of amino acid residues (Table 1). This discrepancy would be attributed to the flexible regions in P2 and the electron density contrast between P2 and the buffer solution. P2 comprises three functional domains (524 residues), two flexible regions connecting the domains (219 residues) and the N- and C-terminal tails (172 residues) [inset of Fig. 3(*a*)]. Firstly, the flexible regions have a smaller contribution to the scattering intensity than the three functional domains, due to their low density contrast in ensemble average. Secondly, P2 was suspended in buffer containing 0.5 *M* NaCl and 10% (*w*/*v*) glycerol (Oide *et al.*, 2018) to avoid non-specific aggregation, and then the electron density contrast of P2 against the buffer is decreased to 88% of that against pure water. These two factors probably decrease in the net scattering density of P2 contributing to the scattering intensity.

The restored models were distributed roughly in a T-shape (Fig. 3*b*) on the plane spanned by the first and second PCs, which accounted for 42% of the total variance in the 5940-dimensional space (Fig. S1). Approximately 80% of the restored models were S-shaped sticks as classes I–VI, while those in the other classes were approximated as straight sticks [Fig. 3(*c*) and Table 2]. Therefore, the averaged model appeared as an S-shape. The S-shaped stick models were distributed in the region of negative values of the second PC, while the straight stick models were distributed in the positive region.

For the S-shaped models, the LOV1 dimer fitted well with the central region, and then LOV2 and kinase models were superimposed onto the elbow and the edge, respectively (Oide *et al.*, 2018). In the stick-shaped models, similar arrangements of the atomic models were possible, but the central regions were somewhat narrow to place the LOV1 dimer. In reciprocal space, the SAXS profiles of all atomic models in Fig. 3(*c*) displayed little differences with the experimental profile as indicated by the χ^2 values [Fig. 3(*d*) and Table 3]. As a result, in the case of P2, the fitness of the atomic model of LOV1 dimer was essential to select models of classes I–VI as the most probable molecular structure of P2.

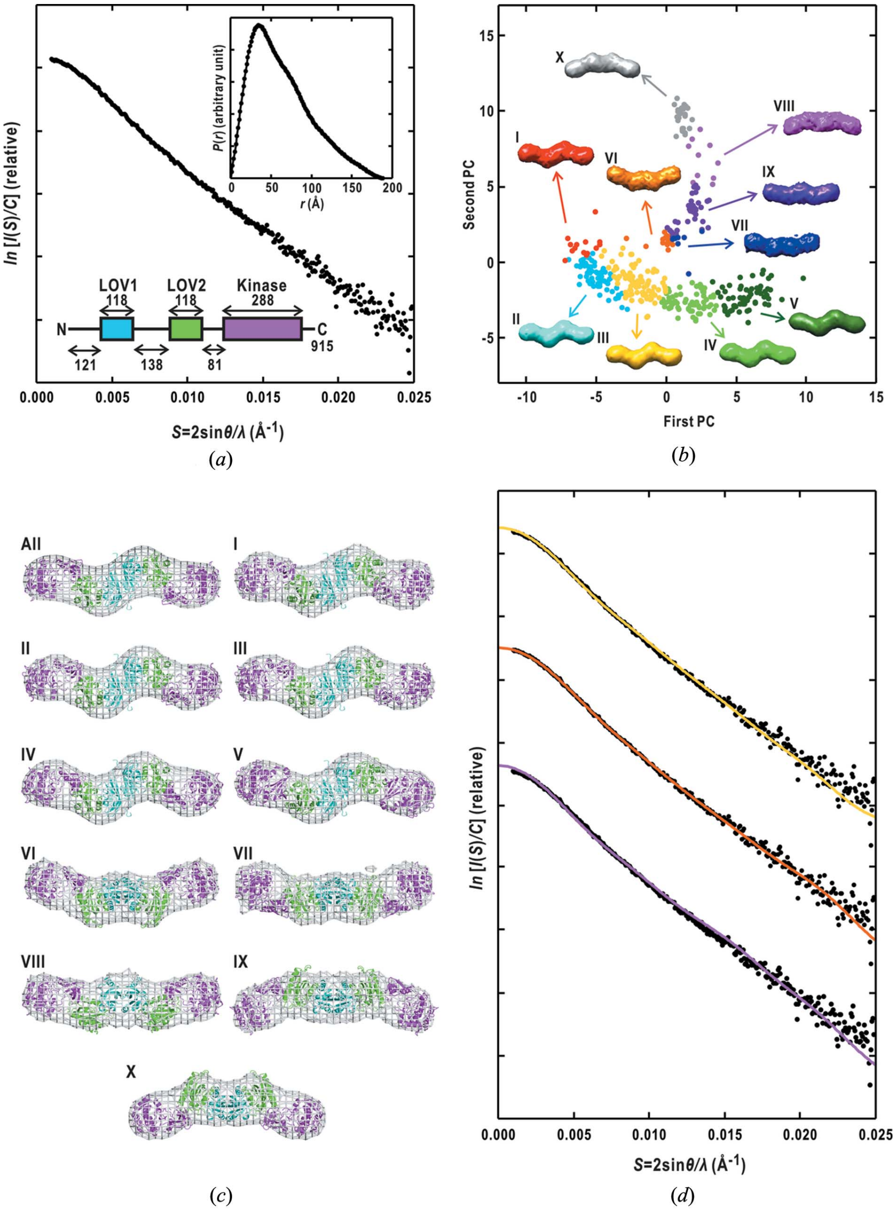


Figure 3
(*a*) Experimental SAXS profile of P2, the *P*(*r*) function (inset in the upper right), and schematic illustration of the arrangement of LOV1, LOV2 and kinase domain in the amino acid sequence of P2 (lower left). (*b*) Distribution of restored models in the plane spanned by the first and second PCs as illustrated according to the manner in Fig. 2(*b*). (*c*) Superimposition of the crystal structures of the phot2 LOV1 dimer (cyan-colored ribbon model), phot2 LOV2 (green) and the homology model of the kinase domain (magenta) onto the averaged models of all without classification, and of ten classes. The viewing directions of the models in panel (*c*) are the same as those in (*b*). (*d*) Comparison of the scattering profiles calculated from fitted atomic models of classes III (upper), VI (middle) and IX (lower) in panel (*c*) with the experimental one (dots).

4.3. Molecular shape of P1L1

For P1L1, the *GASBOR* calculations were carried out under the constraint of the twofold symmetry [Fig. 4(*a*), Tables 1

Table 3
χ^2 and R_g values of atomic models fitted to predicted molecular shapes.

	χ^2 / R_g (Å)	
Cluster	PDI	P2
Average	– / –	2.1 / 52.5
I	58 / 33.9	1.8 / 52.9
II	104 / 32.8	1.9 / 53.2
III	49 / 35.5	1.9 / 53.2
IV	75 / 34.0	2.0 / 52.9
V	82 / 35.5	2.2 / 53.4
VI	192 / 35.7	1.6 / 51.1
VII	– / –	2.3 / 53.2
VIII	180 / 35.1	2.1 / 53.2
IX	225 / 35.9	3.2 / 55.2
X	– / –	1.6 / 51.4

and 2], because the LOV1 domain forms a dimer both in solution and in a crystal (Nakasako *et al.*, 2004, 2008) (Fig. 4*a*). The PCs with the first and second largest eigenvalues accounted for 24% of the total variance in the distribution of the models in the 1989-dimensional space [Fig. S1 and Fig. 4(*b*)]. The SAXS profile of P1L1 displayed an ambiguity score of almost zero, suggesting a small variation among the restored models (Table 1).

The ten classes were divided roughly into two groups with regard to the overall shapes [Table 2, Figs. 4(*b*) and 4(*c*)]. Averaged shapes of classes I–VII were ellipsoid, while bulges appeared at the edges of the averaged shapes in VIII–X. Little correlation was found between the two groups and their PC values, in contrast to the cases of PDI (Fig. 2*b*) and P2 (Fig. 3*b*). The averaged model for all restored models was dominated by classes I–VII, which accounted for more than 80% of all restored models (Table 2). Models in classes I–III tended to be more densely distributed in the PC plane. The crystal structure of the P1L1 dimer was superimposable onto the central region of the averaged shapes of all classes as well as the averaged shape for all models (Fig. 4*c*). Both edges of the molecular shapes inconsistent with the crystal structure model would correspond to the N- and C-terminal regions of P1L1, which were missed in the electron density map obtained by the crystal structure analysis (Nakasako *et al.*, 2008).

The SAXS profile calculated from the crystal structure deviated largely from the experimental one (Fig. 4*a*), probably because of the lack of the N-terminal and C-terminal regions, which were invisible in the electron density map of the crystal structure analysis (Nakasako *et al.*, 2008). The major ellipsoid shapes of classes I–V would be more suitable to approximate the crystal structure than the minor shapes with nonrealistic bulges in classes VIII–X.

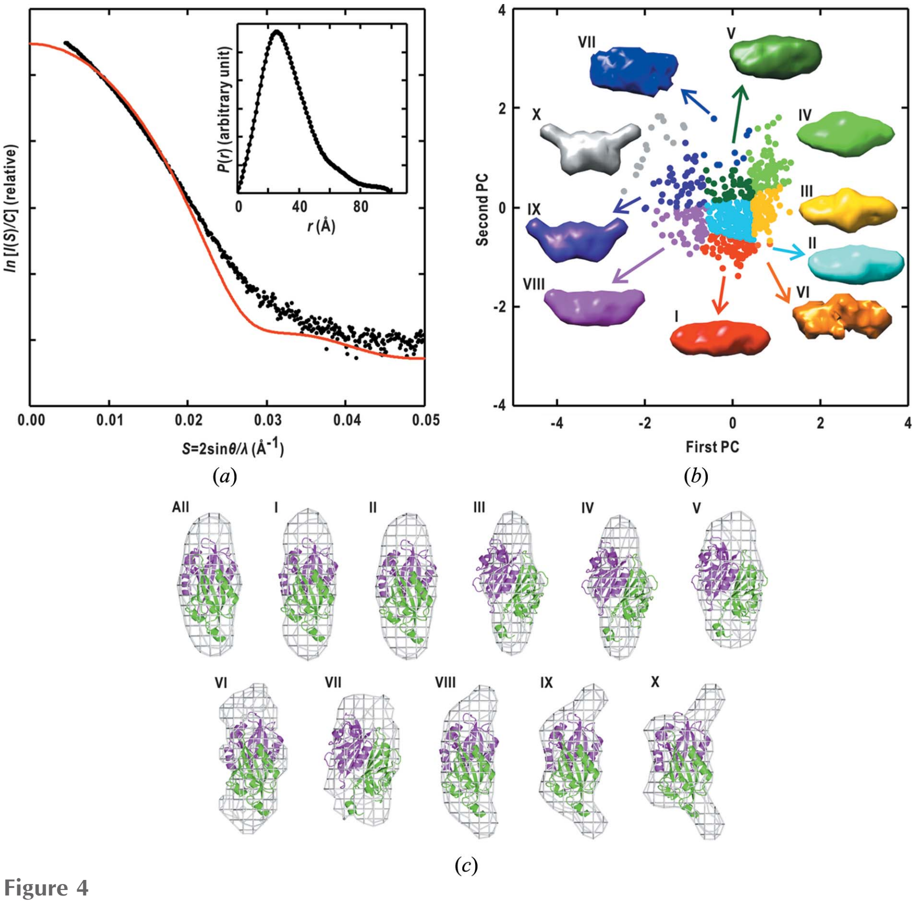


Figure 4
(*a*) Experimental SAXS profile of P1L1 (black dots) compared with that calculated from a crystal structure of phot1 LOV1 dimer (red line). The inset is the $P(r)$ function. (*b*) Distribution of restored models in the plane spanned by the first and second PCs illustrated according to the manner in Fig. 2(*b*). (*c*) Superimposition of the crystal structure of P1L1 onto the averaged models of all without classification, and of ten classes. The viewing directions of the models in panel (*c*) are the same as those in (*b*).

4.4. Molecular shape of LphyA

LphyA is composed of one light-receiving fragment and a kinase domain, and forms a dimer in solution (Nakasako *et al.*, 2005). The molecular shape of the red-light-absorbing form of LphyA was reexamined by using the proposed protocol for *GASBOR* models restored from the SAXS profile under the assumption of the twofold symmetry (Fig. 5*a*). The ambiguity score of the SAXS profile was smaller than those obtained for PDI and P1L2K (Table 1).

The restored molecular models were classified roughly into two groups on the plane spanned by the first and second PCs, accounting for 18% of the total variance in the 23800-dimensional space [Fig. S1 and Fig. 5(*b*)]. Models of classes I–VI were composed of a pair of twisted rods, whereas those of VII–X were approximated as two oblate ellipsoids contacting at their edge regions. Although classes VII–X tended to be distributed in the region of positive PC values, it is difficult to clearly separate the models into the two groups only by the PC values.

For all classes, the crystal structures of the dimeric bacteriophytochrome and the dimeric C-terminal part of tyrosine kinase were difficult to be simultaneously superimposed onto the averaged model of any class (Fig. 5*c*). In the previous study, we selected VII–X models based on the biochemical evidence demonstrating that the LphyA dimer is associated only with the kinase domains (Nakasako *et al.*, 2005). Although critical shows were unavailable to select classes suitable to approximate the molecular structure of LphyA, the classification contributed, at least, to separate the two types of possible molecular models.

5. Discussion

In the present study, we proposed a protocol to classify *ab initio* models restored from a SAXS profile by using multivariate analysis. The protocol illustrates the differences among the molecular shapes of the classified models and provides an opportunity to examine which molecular shapes are more probable. Here, we discuss the benefits and future improvements of the proposed protocol.

5.1. Benefits of the proposed protocol

The *ab initio* molecular modeling algorithms have been contributing to the estimation of molecular structures of proteins in solution. However, the algorithms occasionally yield nonrealistic molecular shapes as well as plausible ones in a number of calculations. Therefore, a simple average of all predicted molecular models include the structural features of nonrealistic models and subsequently blurs the details of probable molecular shapes. The proposed protocol using multivariate analysis provides an opportunity to separate probable molecular models from nonrealistic ones. The probable class displays a small structural variation, and then the averaged model has structural details better than that from all models with a large variation (Figs. 2–4). This result is consistent with the idea that the variation of *ab initio* models reflects the resolution of models (Tuukkanen *et al.*, 2016).

In order to examine whether the large number of models is advantageous to select correct models, we conducted a series of classifications of a smaller number of PDI models by the *DAMCLUST* program (Petoukhov *et al.*, 2012) (see §S2 and Fig. S2 of the supporting information). The results suggest that the classification of 20 models, which is usually treated by *DAMCLUST*, is insufficient to overview the possible variation of restored models and also difficult to avoid the contamination of incorrect models in averaging, particularly for SAXS profiles displaying a large ambiguity score such as PDI. To improve this and ensure the statistical significance of the selected models, the proposed protocol, which is applicable to several hundred restored models, was advantageous for proposing the most probable and realistic molecular models.

The classification of models is advantageous for discussing the arrangement of domains and/or the structural differences of proteins between the crystalline and solution conditions, when partial or whole structures of target or homology proteins are available at atomic resolutions. Major clues to select probable models are the fitness of the atomic models to the SAXS models and/or the similarity of scattering profiles calculated from the constructed atomic models as demonstrated (Figs. 2–4). As seen in the case of LphyA, when atomic models of some parts are inconsistent with the restored SAXS models, the protocol proposes possible candidates of molecular models. Biochemical evidence regarding the interaction of functional domains would be necessary to extract probable molecular shapes.

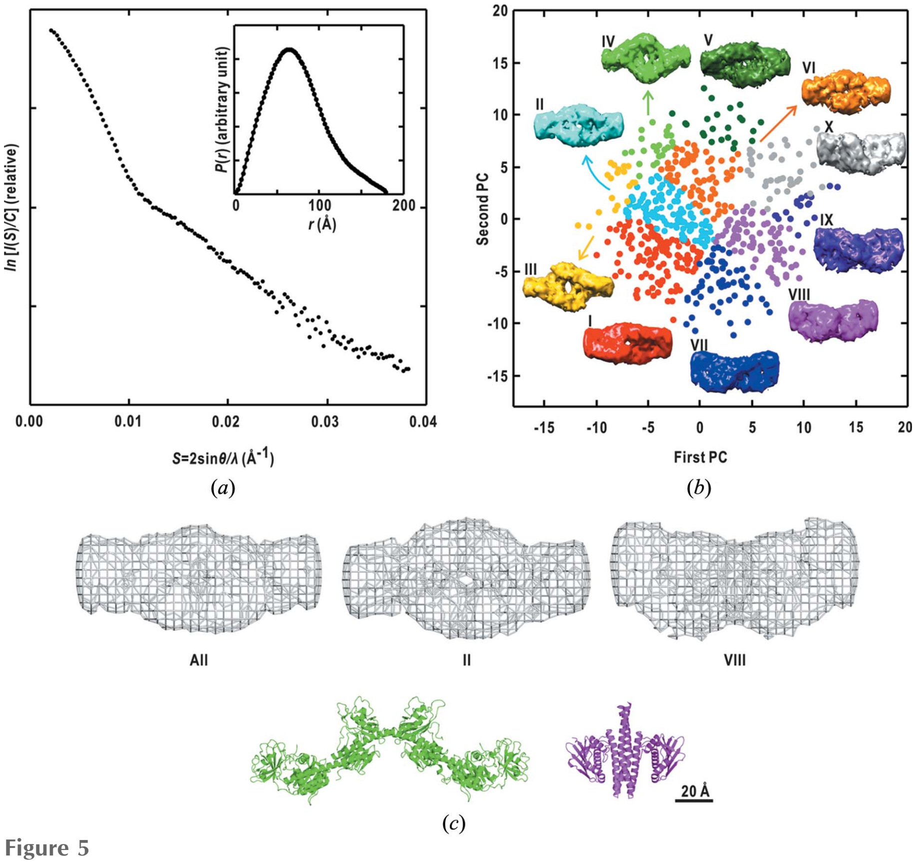


Figure 5
(*a*) Experimental SAXS profile of LphyA and the *P*(*r*) function (inset). (*b*) Distribution of restored models of LphyA in the plane spanned by the first and second PCs illustrated according to the manner described for Fig. 2(*b*). (*c*) Comparison of the averaged models of all without classification, and two representative models from classes II and VIII (upper panels) with the crystal structures of the light-receiving domain of the bacteriophytochrome (green-colored model) and the dimeric C-terminal part of tyrosine kinase (magenta) (lower panels). The viewing directions of the models in panel (*c*) are the same as those in (*b*).

At the end of this section, we consider the possibility that the protocol may help to illustrate conformational changes of a protein, *e.g.* the rearrangement of domains, under external physicochemical stimuli in solution, such as ligand binding, light irradiation and so on (Nakasako *et al.*, 2005, 2010; Takayama *et al.*, 2011; Okajima *et al.*, 2014; Oide *et al.*, 2016, 2018). By comparing the classified models before and after the stimuli, we would be able to trace ways of conformational changes in detail with the assistance of, for instance, molecular dynamics simulations (Oroguchi *et al.*, 2009).

5.2. Correlation between the ambiguity score and variation in model shapes

The variation in the molecular models provided by *GASBOR* would correlate with the ambiguity score of SAXS profiles. For the SAXS profile of P1L1 with a minimal ambiguity, the predicted molecular shapes are similar to each other with respect to size and shape, except for the nonrealistic bulges at the edges in some classes (Fig. 4). For SAXS profiles with large ambiguity scores, nonrealistic models tend to appear (Figs. 2 and 3). Indeed, there is a large variation among the *GASBOR* models for PDI, the SAXS profile of which showed a large ambiguity score (Fig. 2).

This correlation suggests that the ambiguity score is a useful indicator for the complementary application of different protocols for finding correct molecular models. For instance, for a SAXS profile displaying a small ambiguity score, the *DAMCLUST* program is advantageous for obtaining a probable molecular model without any *ad hoc* parameters (Fig. 4 and Table 1). On the other hand, when a SAXS profile displays a large ambiguity score, the proposed protocol allows us to extract models without blurring by nonrealistic models (Figs. 2, 3, 5 and Table 1). Since *DAMCLUST* requires a heavy computational cost for classifying a hundred molecular models (§S2 of the supporting information), the choice of the protocols based on the ambiguity score is efficient for finding the most probable molecular models.

5.3. Future improvements of the protocol

The proposed protocol could be improved in each of the two steps. First, for the superimposition of restored models, the present protocol relies on the alignment of the inertia axes of a target model to the reference. For instance, more fine alignment using the normalized spatial discrepancy as a target function would improve the quality and efficiency of the procedure in the superimposition of models.

Second, we used PCA for the dimensional reduction from the hyper-dimensional space to visualize the distribution of the models in a low-dimensional space. Alternatively, adoption of the diffusion map method (Coifman *et al.*, 2005; Yoshidome *et al.*, 2015) might suggest a low-dimensional space to more appropriately describe the distribution of models. Although we classified molecular models into ten classes by the *K*-means clustering method, the number of classes is not limited to ten and can be varied by inspecting the variation of models from a trial classification. After the trial, the mean shift method (Fukunaga & Hostetler, 1975; Cheng, 1995) can be applied without the input of the number of classes.

Apart from the algorithms, the computational performance would be faster owing to the progress in computer technology. Although it still takes several hours for a large number of calculations (Table 1) in this study, on-the-fly SAXS structure analysis would be feasible for suggesting candidates of probable molecular models within a SAXS beam time at a synchrotron facility in near future.

6. Conclusion

We proposed a protocol to classify a large number of molecular models restored from SAXS profiles at a low computational cost. The protocol is advantageous for excluding nonrealistic models, which cause blurring of averaged models. This is particularly effective for extracting probable and realistic molecular models from SAXS profiles with a large ambiguity score. In addition, when atomic structures of parts of a protein and biochemical evidence on the interactions between them are available, the protocol allows us to discuss the arrangement of domains. Even when little structural information is available, the protocol suggests possible candidates of molecular shapes.

Acknowledgements

The authors thank Dr Katsuaki Inoue (Japan Synchrotron Radiation Research Institute / Diamond Light Source) and Dr Takaaki Hikima (RIKEN SPring-8 Center) for their help in the SAXS data collection at SPring-8. The SAXS experiments were performed at BL40B2 under the approval of the JASRI (proposal Nos. 2002B015, 2002B-0153, 2003A0106, 2003A0107, 2003B0141, 2003B-0142 and 2005B0380), and at BL45XU under the approval of RIKEN Harima Institute (proposal Nos. 2011007, 20130012, 20140018 and 20150021).

Funding information

The following funding is acknowledged: Ministry of Education, Culture, Sports, Science and Technology (grant Nos. jp15076210, jp20050030, jp1920402, jp22018027, jp22244054, jp23120525, jp24654140, jp25120725, jp15H01647, jp16H02218, jp17H05891 to MN; grant Nos. jp24113723, jp17K19209, jp17H04854 to TO; grant No. jp15K18559 to KO).

References

Akamine, P., Madhusudan, Wu, J., Xuong, N. H., Ten Eyck, L. F. & Taylor, S. S. (2003). *J. Mol. Biol.* **327**, 159–171.

Bellini, D. & Papiz, M. Z. (2012). *Structure*, **20**, 1436–1446.

Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.

Cheng, Y. (1995). *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 790–799.

Childers, W. S., Xu, Q., Mann, T. H., Mathews, I. I., Blair, J. A., Deacon, A. M. & Shapiro, L. (2014). *PLOS Biol.* **12**, e1001979.

Christie, J. M., Hitomi, K., Arvai, A. S., Hartfield, K. A., Mettlen, M., Pratt, A. J., Tainer, J. A. & Getzoff, E. D. (2012). *J. Biol. Chem.* **287**, 22295–22304.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. W. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 7426–7431.

Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342–346.

Fukunaga, K. & Hostetler, L. (1975). *IEEE Trans. Inf. Theory*, **21**, 32–40.

Glatter, O. & Kratky, O. (1982). *Small-Angle X-ray Scattering.* New York: Academic Press.

Goddard, T. D., Huang, C. C. & Ferrin, T. E. (2007). *J. Struct. Biol.* **157**, 281–287.

Graewert, M. A., Franke, D., Jeffries, C. M., Blanchet, C. E., Ruskule, D., Kuhle, K., Flieger, A., Schäfer, B., Tartsch, B., Meijers, R. & Svergun, D. I. (2015). *Sci. Rep.* **5**, 10734.

Ibel, K. & Stuhrmann, H. B. (1975). *J. Mol. Biol.* **93**, 255–265.

Jeffries, C. M. & Svergun, D. I. (2015). *Methods Mol. Biol.* **1261**, 277–301.

Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Cryst.* **34**, 33–41.

MacQueen, J. (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, *Statistics*, pp. 281–297. Berkeley: University of California Press.

Nakasako, M., Iwata, T., Inoue, K. & Tokutomi, S. (2005). *FEBS J.* **272**, 603–612.

Nakasako, M., Iwata, T., Matsuoka, D. & Tokutomi, S. (2004). *Biochemistry*, **43**, 14881–14890.

Nakasako, M., Maeno, A., Kurimoto, E., Harada, T., Yamaguchi, Y., Oka, T., Takayama, Y., Iwata, A. & Kato, K. (2010). *Biochemistry*, **49**, 6953–6962.

Nakasako, M., Zikihara, K., Matsuoka, D., Katsura, H. & Tokutomi, S. (2008). *J. Mol. Biol.* **381**, 718–733.

Oide, M., Okajima, K., Kashojiya, S., Takayama, Y., Oroguchi, T., Hikima, T., Yamamoto, M. & Nakasako, M. (2016). *J. Biol. Chem.* **291**, 19975–19984.

Oide, M., Okajima, K., Nakagami, H., Kato, T., Sekiguchi, Y., Oroguchi, T., Hikima, T., Yamamoto, M. & Nakasako, M. (2018). *J. Biol. Chem.* **293**, 963–972.

Okajima, K., Aihara, Y., Takayama, Y., Nakajima, M., Kashojiya, S., Hikima, T., Oroguchi, T., Kobayashi, A., Sekiguchi, Y., Yamamoto, M., Suzuki, T., Nagatani, A., Nakasako, M. & Tokutomi, S. (2014). *J. Biol. Chem.* **289**, 413–422.

Oroguchi, T., Hashimoto, H., Shimizu, T., Sato, M. & Ikeguchi, M. (2009). *Biophys. J.* **96**, 2808–2822.

Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.

Petoukhov, M. V. & Svergun, D. I. (2015). *Acta Cryst.* D**71**, 1051–1058.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.

Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.

Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.

Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

Svergun, D. I., Koch, M. H. J., Timmins, P. A. & May, R. P. (2013). *Small Angle X-ray and Neutron Scattering From Solutions of Biological Macromolecules.* Oxford University Press.

Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.

Takayama, Y., Nakasako, M., Okajima, K., Iwata, A., Kashojiya, S., Matsui, Y. & Tokutomi, S. (2011). *Biochemistry*, **50**, 1174–1183.

Tian, G., Xiang, S., Noiva, R., Lennarz, W. J. & Schindelin, H. (2006). *Cell*, **124**, 61–73.

Tuukkanen, A. T., Kleywegt, G. J. & Svergun, D. I. (2016). *IUCrJ*, **3**, 440–447.

Volkov, V. V. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 860–864.

Watanabe, Y. & Inoko, Y. (2009). *J. Chromatogr. A*, **1216**, 7461–7465.

Yoshidome, T., Oroguchi, T., Nakasako, M. & Ikeguchi, M. (2015). *Phys. Rev. E*, **92**, 032710.