JOURNAL OF SYNCHROTRON RADIATION

Volume 25 (2018)

Supporting information for article:

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

Mao Oide, Yuki Sekiguchi, Asahi Fukuda, Koji Okajima, Tomotaka Oroguchi and Masayoshi Nakasako

S1. Supporting information 1

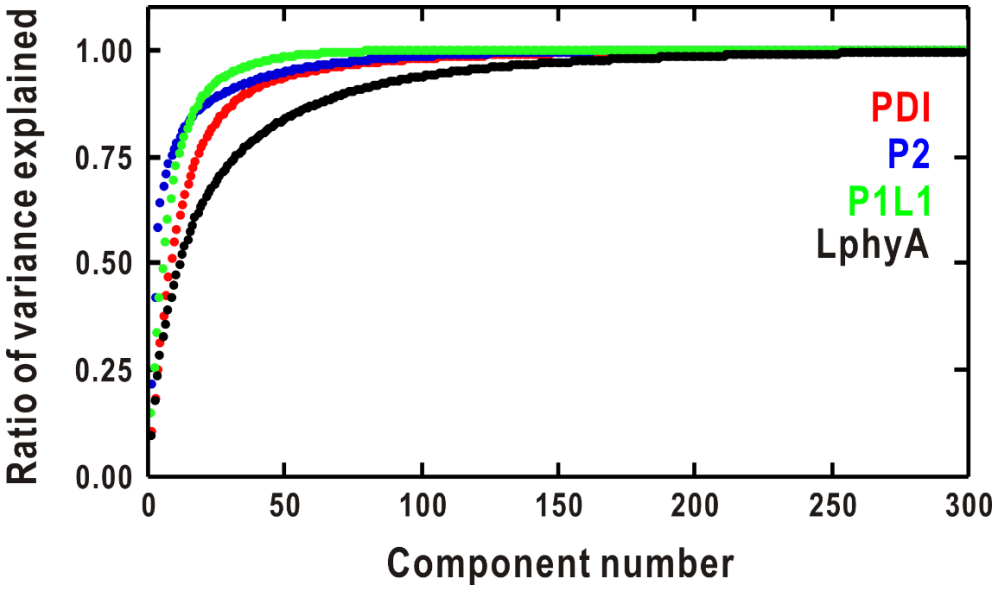


Figure S1 Cumulative contribution of principal components to the total variance of the molecular models in the principal component analysis.

S2. Supporting Information 2

In order to examine whether the classification of a large number of models is advantageous to select correct models, we conducted a series of classification for restored PDI models by the DAMCLUST program (Petoukhov *et al*., 2012), which usually treats 20 restored models.

First, we prepared three groups, each of which was composed of 20 models independently and randomly selected from the restored 576 models (Fig. S2(a)). The classification divided the models into four or five classes in each group. Classes assignable as correct models (classes 1-2 of group1, classes 1-3 of group 2, and classes 1-2 of group3) were minor. Most of models were classified into classes, which displayed averaged molecular shapes inconsistent with the crystal structure. For instance, any clusters of dummy residues assignable to four thioredoxin domains were missed in classes 3-5 of group 1, and classes 4-5 of group 2, and classes 3-4 of group 3, and then the averaged models for these classes had large lobes at the center. This result indicated that the classification and/or alignment procedures of 20 models were inappropriate for the three groups.

Next, we tried to classify group 4 in Fig. S2(b), which comprised randomly selected 100 models. The models were divided into eight classes. The difference in the

number of classes between groups 4 and the previous three (groups 1-3 in Fig. S2(a)) implies that a set of 100 models reflects the possible variation of molecular models better than a set of 20 models. The three minor classes 1-3 displayed molecular structure similar to the crystal structure. In contrast, class 8 composed of 88 models display little structural characteristics of PDI, probably because of the incorrect classification and/or misalignment before averaging. It should be noted that the computational time of DUMCLUST calculation for group 4 was about twenty hours on a CPU of Intel Pentium CPU G3220 (3.00 GHz), while the proposed protocol could process the classification of 576 models within one hour. Therefore, it was impossible to apply DAMCLUST to 576 models, because of the computational cost.

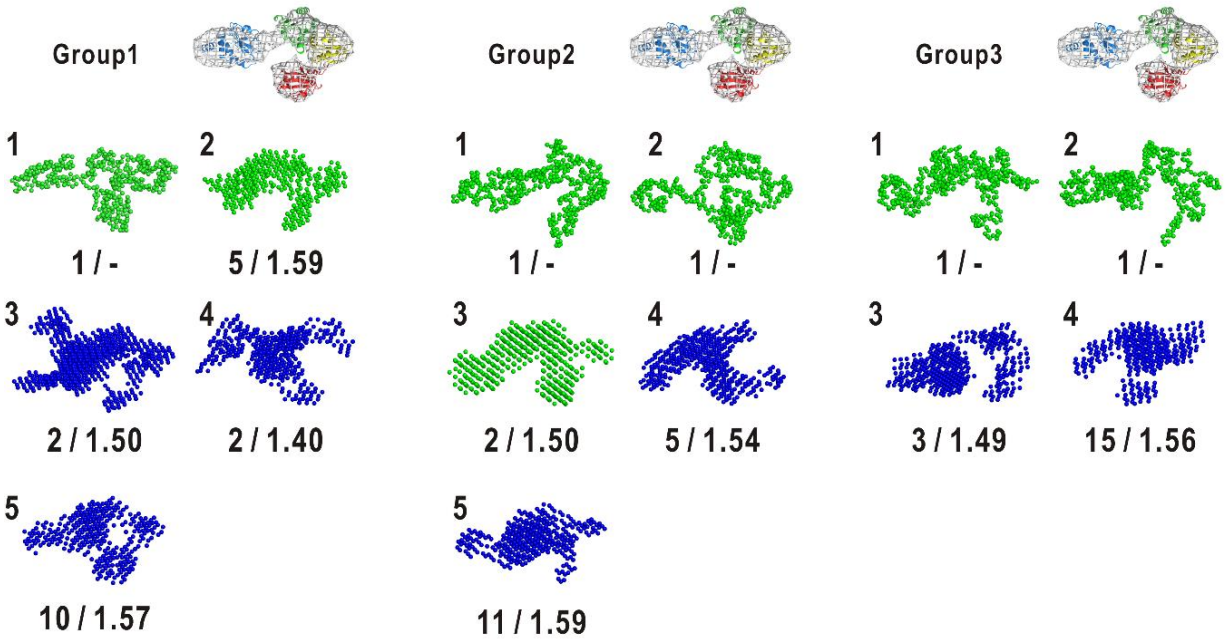
Finally, we examined whether DAMCLUST could correctly classify the models, which were already classified by the multivariate analysis. We made group 5 (Fig. S2(c)) by selecting a pair of models located around the centroids in each of 10 classes (see Fig. 2 in the main text). In contrast to groups 1-3, the models in group 5 were divided into seven classes. This difference in the number of classes would be attributed to the appropriate sampling, which represents the possible variation of restored models in group 5. However, the class 7 composed of ten models was inconsistent with the molecular shape of PDI. Only minor classes 1 and 5 approximate the molecular shape.

The classification of 20 models would be still difficult even when representative models were used.

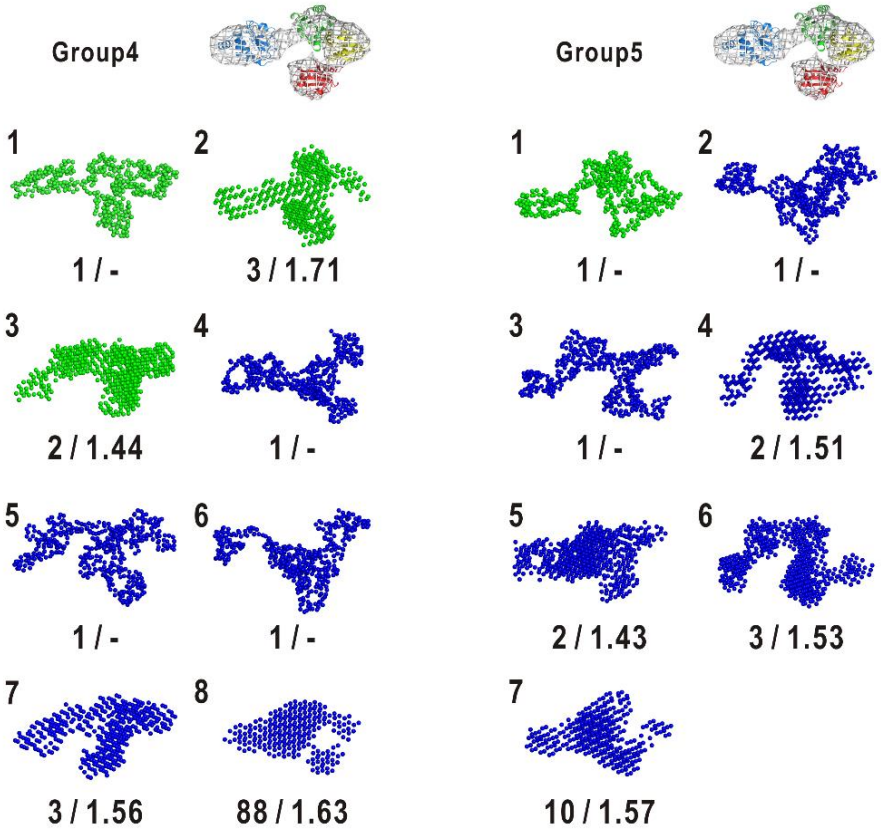
From these classification trials of different types of groups, we concluded that the proposed protocol was advantageous to get probable and realistic molecular models than DAMCLUST, particularly when the SAXS profile had a large ambiguity score as PDI.

Reference

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


Group1
1
1 / -
2
5 / 1.59
3
2 / 1.50
4
2 / 1.40
5
10 / 1.57
Group2
1
1 / -
2
1 / -
3
2 / 1.50
4
5 / 1.54
5
11 / 1.59
Group3
1
1 / -
2
1 / -
3
3 / 1.49
4
15 / 1.56

(a)


Group4
1
1 / -
2
3 / 1.71
3
2 / 1.44
4
1 / -
5
1 / -
6
1 / -
7
3 / 1.56
8
88 / 1.63
Group5
1
1 / -
2
1 / -
3
1 / -
4
2 / 1.51
5
2 / 1.43
6
3 / 1.53
7
10 / 1.57

(b) *(c)*

Figure S2 Classification of restored models of PDI in five groups. The averaged model in each class or single model is illustrated with the numbers of models classified and the NSD scores Illustrated are results for (a) three groups, each of which are composed of 20 randomly selected models, (b) a group containing randomly selected 100 models, and (c) a group of 20 models selected from the 10 classes after the classification by the proposed protocol. The model of class I obtained by the proposed protocol (see Fig. 2(b) in the main text) is illustrated in each panel. The models were colored in green when they are superimposable with the reference model, while those colored in blue are difficult.