A Two-Fold Structural Classification Method for Determining the Accurate Ensemble of Protein Structures

Pan Tan\textsuperscript{1,2}, Zuyue Fu\textsuperscript{2,3}, Loukas Petridis\textsuperscript{4,5}, Shuo Qian\textsuperscript{4}, Delin You\textsuperscript{6}, Dongqing Wei\textsuperscript{6}, Jinglai Li\textsuperscript{2,7} and Liang Hong\textsuperscript{1,2,*}

\textsuperscript{1} School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China.
\textsuperscript{2} Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China.
\textsuperscript{3} Zhiyuan College, Shanghai Jiao Tong University, Shanghai 200240, China.
\textsuperscript{4} Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States.
\textsuperscript{5} Department of Biochemistry, Cellular & Molecular Biology, The University of Tennessee, Knoxville, Tennessee 37996, United States.
\textsuperscript{6} School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China.
\textsuperscript{7} Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZL, UK.

Received 22 May 2018; Accepted (in revised version) 27 August 2018

Abstract. Atomic-level structural characterization of flexible proteins, such as intrinsically disordered proteins and multi-domain proteins connected by flexible linkers, is challenging as they possess distinct conformations in physiological conditions. Significant efforts have been made to develop integrated approaches by combining small angle neutron/X-ray scattering experiments with molecular simulations to reveal the distinct atomic structures and the corresponding populations for these flexible proteins. One widely used method, the basis-set supported ensemble method, classifies the simulation-generated protein conformations into a set of structural basis and then derives the corresponding populations by fitting to the experimental data. This method makes an implicit assumption that protein conformations of similar structures have similar small angle scattering profiles. The present work demonstrates that, for various protein systems ranging from compact globular proteins and flexible multi-domain proteins through to intrinsically disordered proteins, this method provides inaccurate assessment of the structural ensemble of the protein molecules due to the breakdown of the assumption made. To alleviate this problem, a two-fold-clustering...
method is developed to cluster the simulation-generated protein structures using information on both 3D structure and scattering profiles. As benchmarked by both simulation and experimental results, this new method yields much more accurate populations of structural basis of protein molecules.

PACS: 62-07

Key words: Protein structures, statistical data analysis, Monte Carlo, cluster analysis.

1 Introduction

A central task in molecular biochemistry and molecular biophysics is to determine the atomic structure of proteins at physiological conditions. Although X-ray crystallography and NMR can provide high-resolution atomic-level structure of bio-macromolecules, they are limited by either the availability of crystalline samples or the size of the macromolecules. These high-resolution techniques can be complemented by low-resolution ones, such as cryo-electron microscopy, mass spectroscopy and small angle scattering (SAS). SAS, either with X-ray or neutron, has the advantage of measuring protein structures in physiological conditions [3, 5, 12, 14, 25, 26, 31, 33]. However, SAS is inherently limited because the three-dimensional real-space structural information of a biomacromolecule is reduced to a one-dimensional scattering profile in reciprocal space, resulting in loss of information and the difficulty of converting the SAS intensity to a 3D structure [11, 15, 23, 24, 29, 31–33]. Ab initio method and the Bayesian refinement method [28] have been developed to re-construct low-resolution representations of the biomacromolecules by modeling the experimental SAS data using spatially packed spheres of subnanometer size [30], which, however, lack the atomic-level, or even secondary-structure-level information.

Recently, there have been significant efforts to develop integrated approaches by combining small-angle scattering experiments with molecular simulations to derive the atomic-level structures of bio-macromolecules [3, 4, 6, 11, 16, 20, 21, 24, 28, 32, 33]. These approaches roughly fall into two categories: one is to search for the protein conformations from existing structural candidates, pre-generated from molecular simulation using standard force fields, which best fits to the experimental SAS data [11, 21, 22, 24, 32, 33]; while the other one is to apply biased potentials to drive the simulation towards the protein conformations in better agreement with experiment [4, 6, 16]. The present work focuses on the discussion of the first type of approaches. It becomes increasingly clear that flexible biomolecules, such as multi-domain proteins linked through flexible linkers and intrinsically disordered proteins, possess multiple conformations in the physiological conditions [14, 33]. Revealing the populations of distinct conformations of such protein system in solution is of great importance towards the understanding of the enzymatic mechanism. This inspires the development of the ensemble-based SAS approaches, such as the ensemble optimization method [3], the basis-set supported ensemble method [26, 33],
the SAS module in an integrative platform by taking into account the structural
constraints in the bio macromolecules [11], and the minimal ensemble search method [24].
The ensemble-based SAS approaches present a significant advancement in interpretation
of flexible protein conformations as compared to methods which find a single protein
structure fitting best to the experimental SAS data [14].

The present work mainly tests the basis-set supported ensemble method [26, 33],
which can be schematically illustrated by Fig. 1(a). Briefly, a large pool of candidates
of protein conformations, pre-generated using a simulation method, e.g., molecular dy-
namics simulations (MD) or Monte Carlo simulation, is clustered into a limited number
of sub ensembles, i.e., representative structures, based on structural similarity as quan-
tified by a structural metric. Here, three structural metrics are used: root-mean-squared
deviation (RMSD), distance root-mean-square (DRMS) [26] (see the supporting informa-
tion) and dihedral distance [7]. Then, SAS profiles are calculated for each sub ensemble
and used as a linear basis to fit against the experimental data to obtain the corresponding
populations. This method makes an implicit assumption that structurally similar con-
formations have similar SAS profiles, which can be invalid as reported by Refs. [15, 32].
Here, we demonstrate that, for various protein systems, this ensemble-search method
can lead to inaccurate assignment of protein structural ensemble because of the failure
of the assumption made. To overcome this deficiency, a two-fold-clustering method is
developed, which clusters the structural candidates using information on both structures
and SAS profiles. The ability of this new method in predicting a more accurate ensemble
of protein structures in solution is confirmed by both simulation and experimental tests.

2 Method

For simplicity, the basis-set supported ensemble method as shown in Fig. 1(a) is referred
as one-fold-clustering method herein. The two-fold-clustering method introduced here,
schematically illustrated in Fig. 1(b), contains four steps: (1). The original pool of protein
conformations, which are pre-generated by molecular simulations, is clustered into N sub
ensembles, using the algorithm developed in Ref. [8], based on their structural similarity,
as quantified by one of the three structural metrics (RMSD, DMSD or dihedral distance).
This step is the same as for the one-fold-clustering method; (2). Protein conformations
within each sub ensemble are further divided into several smaller clusters based on their
similarity in $I(q)$, quantified by,$$
\delta_{ij} = \frac{1}{L-1} \sum_s \frac{[I_i(q_s) - I_j(q_s)]^2}{\sigma^2(q_s)},
$$
where, $I_i(q_s)$ and $I_j(q_s)$ are the SAS intensities of the protein conformation $i$ and $j$, respectively,
$q_s$ is the scattering wave vector, $L$ is the total number of the scattering wave vectors,
and $\sigma^2(q_s)$ is the variance of $I(q_s)$ among all the conformations. Thus, the protein confor-
mations within each cluster are similar in both 3D structure and $I(q)$. (3). The population
Figure 1: (Color online) Schematic illustration for (a) one- and (b) two-fold-clustering methods. Ref. [33] also proposed a two-criteria method using both RMSD and $I(q)$ to cluster simulation-derived protein conformations to reveal the atomic structures of proteins. It is, however, different from the present work. It first clusters the simulation-derived protein conformations to small sub ensembles based on their structural similarity using RMSD, and then merges them to bigger clusters based on similarity in $I(q)$. Finally, these big clusters serve as linear basis to fit against the experimental $I(q)$ to obtain the corresponding populations. Within each of these big clusters, protein conformations can differ significantly despite $I(q)$ is similar. In other words, Ref. [33] obtains an ensemble of $I(q)$, from which one can hardly get the information on the populations of different representative structures. In contrast, this is the direct message delivered by the two-fold method proposed in present work.

of each cluster is obtained by fitting the experimental $I(q)$ to a set of linear basis composed by the scattering profiles calculated from each cluster. (4). The so-obtained populations of the clusters originating from the same sub ensemble are summed together to represent the population of that sub ensemble. Therefore, the two-fold-clustering method yields the same set of sub ensembles (representative structures) as the one-fold-clustering method does, while the corresponding populations could be different.
Figure 2: (Color online) Snapshots of the simulation systems, (a) IDP (b) lysozyme and (c) MerA. (d) SAXS profiles calculated based on the protein structures displayed in (a), (b) and (c) using software Sassena [19].

To test the performance of these two different methods, three proteins with distinct structural features are examined here: an intrinsically disordered protein (IDP, Tau267-312, PDB ID: 2MZ7), a multi-domain protein (mercuric ion reductase, MerA) and a globular single-domain protein (lysozyme). Typical structures of these three systems are displayed in Figs. 2(a) to (c), and the corresponding small angle X-ray scattering (SAXS) profiles, I(q), are presented in Fig. 2(d). We employ all-atom molecular dynamics simulation to generate 1000 conformations for each system (see the supporting information (SI) at www.global-sci.com/cicp.OA-2018-0140.SI for detailed simulation protocols), which are used as the initial pool of protein conformations for cluster analysis.

3 Results
3.1 Simulation test

In the simulation tests, the averaged $I(q)$ calculated from the snapshots of the protein molecule in the first one fifth of the MD trajectory (200 conformations) for each system are taken as the target SAXS data, and it will be modeled by the one- and two-fold-clustering methods. As seen in Figs. 3(a) to (c), both methods yield good fits to the simulation-generated target SAXS data. To quantify the agreement between the fitted and the simulation-generated $I(q)$, a score parameter is defined as [14,32]:

$$
\chi^2 = \frac{1}{L-1} \sum_s \left[ \sum_{i=1}^N P_i \ast I_i(q_s) - I_{tar}(q_s) \right]^2 \sigma^2(q_s),
$$

(3.1)
Figure 3: (Color online) Comparison of target SAXS profiles with those derived from one- and two-fold-clustering methods for (a) IDP, (b) lysozyme and (c) MerA. The error bar represents one standard deviation.

where \( P_i \) and \( I_i(q) \) are the fitted population and small angle scattering profiles for Sub Ensemble \( i \), \( I_{tar}(q) \) is the target SAXS profile, and \( \sigma(q) \) is the standard deviation among all \( I_i(q) \) in the target trajectory. The values of \( \chi^2 \) are displayed in Table 1, showing that one-fold method agrees better with the target SAXS profiles.

Table 1: Scoring parameters, \( \chi^2 \), the cutoffs of RMSD, DRMS and dihedral distance, and the ratio between \( \gamma_{one-fold} \) and \( \gamma_{two-fold} \). The analysis is based on SAXS profiles. Detailed procedures can be found in SI.

<table>
<thead>
<tr>
<th>Protein</th>
<th>( \chi^2_{one} )</th>
<th>( \chi^2_{two} )</th>
<th>( \gamma_{one}/\gamma_{two} )</th>
<th>RMSD</th>
<th>DRMS</th>
<th>Dihedral distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>lysozyme</td>
<td>1.27291</td>
<td>1.61756</td>
<td>5.5264</td>
<td>0.71003</td>
<td>1.2889</td>
<td>8.2081</td>
</tr>
<tr>
<td>IDP</td>
<td>0.34998</td>
<td>1.9347</td>
<td>4.1128</td>
<td>0.30538</td>
<td>1.4863</td>
<td>16.7</td>
</tr>
<tr>
<td>MerA</td>
<td>0.3793</td>
<td>2.0482</td>
<td>13.7346</td>
<td>0.2303</td>
<td>2.5459</td>
<td>3.4194</td>
</tr>
</tbody>
</table>

| | cutoff | | | | | |
| lysozyme | 1.65 Å | 1 Å | 0.15 |
| IDP | 11 Å | 9 Å | 0.52 |
| MerA | 6.5 Å | 4 Å | 0.16 |
Figure 4: (Color online) Simulation tests on the performance of one-fold and two-fold-clustering methods. Comparison of the true populations of sub ensembles constituting the target MD trajectories with the populations derived from one-fold and two-fold methods by fitting to the target \( I(q) \) for (a) IDP, (b) lysozyme and (c) MerA. (d) Relative errors, \( \gamma \) defined in Eq. (3.2) resulting from one- and two-fold methods and (e) the ratio of \( \gamma_{\text{one-fold}}/\gamma_{\text{two-fold}} \) for different systems. A Monte Carlo procedure is used to derive the populations of sub ensembles using the one- and two-fold methods, details of which can be found in supporting information. All the results presented in this figure are obtained by using RMSD as the structural metric. Using other structural metrics will provide qualitative similar results (see Table 1).

Although both methods provide seemingly good fits to the target data (Fig. 3), the resulting populations of sub ensembles differ dramatically (see Figs. 4(a) to (c)). For comparison, the true populations of sub ensembles constituting the target MD trajectories are also plotted, and are found to be in much better agreement with those derived from the two-fold-clustering method. (The true populations of sub ensembles are obtained by directly examining each snapshot in the target MD trajectory, and the detailed procedure for how to determine them is presented in Fig. S4 in SI.) To quantify the accuracy of these two methods in predicting the populations of protein conformations, an error parameter, \( \gamma \), is defined:

\[
\gamma = \sqrt{\frac{\sum_{i=1}^{N} (P_i - P^*_i)^2}{\sum_{i=1}^{N} (P^*_i)^2}},
\]

(3.2)

where \( P_i \) and \( P^*_i \) are the fitted and true populations for a given Sub Ensemble \( i \), respectively, and \( N \) is the total number of sub ensembles. Thus, \( \gamma \) quantifies the deviation between the fitted and the true populations, i.e., smaller \( \gamma \) indicates better agreement to the true populations. As seen in Figs. 4(d), 4(e) and Table 1, \( \gamma_{\text{one-fold}} \) is many folds larger than \( \gamma_{\text{two-fold}} \), i.e., the populations derived from the two-fold method is much closer to the true ones, although \( I(q) \) derived from the one-fold method agrees better with the target SAS results, i.e., smaller \( \chi^2 \) (see Table 1). Hence, determining the 3D protein struc-
tures simply based on examining the value of $\chi^2$ can sometimes be misleading. These findings are independent of which portion of the MD trajectory is used to represent the target $I(q)$ (see Fig. S1 in SI), independent of the size, compactness, secondary and tertiary structures of the system studied, independent of the choice of the structural metric, and independent of whether SAXS or SANS profiles (see Table S1) are considered.

The improved performance of the two-fold-clustering method arises from not assuming that the structurally proximate protein conformations possess similar SAS profiles, an assumption which can be invalid. This is evident by Figs. 5(a) to (c), which show no strong correlation between the structural difference (e.g., RMSD) $\delta_{ij}$ (difference between $I(q)$ as defined in Eq. (2.1)) for every two conformations of a given protein system. Similar findings have been also reported by others [15, 32] and should be generally applicable to bio-macromolecules as the test is conducted herein over a wide range of systems, independent on the choice of the structural metric (Figs. S2 and S3 in SI) and independent on whether small angle X-ray or neutron scattering is considered (Fig. S3 in SI).

### 3.2 Experimental test

In addition to the simulation test, we also examined the performance of the two methods using the experimental SAXS data of MerA [14]. In order to get sufficient sampling of protein conformations to have better agreement with SAXS data, here we conducted gas phase simulation of MerA (see more details in SI), the protein structures generated from which were used for cluster analysis to fit SAXS data. As seen in Fig. 6(a), the two methods both provide good fits to the experimental $I(q)$, with the one-fold method presenting a slightly smaller $\chi^2$, but yield vastly different populations of representative structures, i.e., sub ensembles (see Fig. 6(b)). The two-fold method predicts two highly populated protein structures, while the one-fold method gives one dominant structure with the other two much less populated conformations. More importantly, the dominant protein structure identified by the one-fold method does not agree with any of the two structures revealed by the two-fold method.

MerA possesses metallochaperone-like N-terminal domains (NMerA), which are tethered to the homodimeric catalytic core by 30 residue linkers [1, 2] (Fig. 8). NMerA contains a pair of cysteine residues in a GMTCXXC sequence motif that is conserved in those soft metal ion trafficking proteins sharing the common $\beta\alpha\beta\alpha\beta$ structural fold [18]. MerA binding and processing of Hg(II) is schematically illustrated in Fig. 8. Hg(II) binds first to a pair of cysteines (C11 and C14) in NMerA, these being the most solvent-accessible cysteines in the protein [17, 18]. The ion is then transferred to another pair of cysteines (C561’ and C562’) located on the mobile C-terminal segment of the core. The mobile segment then moves the C-terminal Hg(II)-bound cysteines from the surface to the interior of the protein, from which the ions is further delivered to the buried, active-site cysteine pair (C135 and C140) where the reduction occurs [9, 10]. In vivo experiments showed that the presence of NMerA domains significantly enhances cell survival in the presence of Hg(II) [17], in which the high-affinity Hg(II)-chelating NMerA domains are
Figure 5: Comparison of RMSD and the difference in SAXS profiles, $\delta_{ij}$ as defined by Eq. (2.1) for every two protein conformations for (a) IDP, (b) lysozyme and (c) merA. We found that the difference between SAS profiles for two protein conformations has no strong correlation with their structural difference. These findings are independent on which structural metric used (Figs. S2 and S3 in SI) and independent on whether small angle X-ray or neutron scattering is considered (Fig. S3 in SI).

...essential for rapid acquisition, localization and transfer of Hg(II) to the core for reduction and detoxification [17]. Ref. [13], by combining small angle scattering, coarse-grained simulation and all-atom molecular dynamics simulation, revealed that MerA adopts a highly compact structure in solution, where the NMerA domains are leashed by the linkers to bound strongly with the core domain, being close to the C-terminal Hg(II)-binding cysteines for rapid delivery of mercury ions. It is further confirmed by neutron spin echo...
measurements, which showed no appreciable inter-domain dynamics being present between the NMerA and core domains [13].

As shown by Table 2 and Fig. 7, Sub Ensemble 1 (Red in Fig. 7) is the most extended conformation, in which its NMerA domains are detached significantly from the catalytic core (marked in red in Table 2), thus contradicting the findings of Ref. [13]. This unlikely conformation is highly populated (66%) as revealed by the one-fold method, but has zero weight as predicted by the two-fold method. Moreover, the most dominant structure identified by the two-fold method, i.e., Sub Ensemble 2 (Blue in Fig. 7), is the most compact conformation among the three sub ensembles and has the shortest distances between C11 and C516’ and between C11’ and C561, i.e., characteristic distances measuring the capability of the linker to leash the NMerA domain around the core C-terminal Hg(II)-binding cysteines, which is in better agreement with the finding of Ref. [13]. How-

Table 2: Structural characteristics of MerA conformations revealed by one- and two-fold-clustering methods by modeling the experimental SAXS data [14]. The values inside brackets represent one standard deviation. Values marked in red indicate that the conformation is highly extended.

<table>
<thead>
<tr>
<th>sub-ensemble ID</th>
<th>population</th>
<th>Minimum distance between each of the two NmerAs and the core domain ( Å)</th>
<th>C11-C516’ / C11’-C516 ( Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>one-fold</td>
<td>1</td>
<td>66%</td>
<td>14(6)/12(5)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>21%</td>
<td>7 (3)/5 (3)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>13%</td>
<td>10 (4)/9 (4)</td>
</tr>
<tr>
<td>two-fold</td>
<td>1</td>
<td>0%</td>
<td>14(6)/12(5)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>53%</td>
<td>7(3)/5 (3)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>47%</td>
<td>10 (4)/9 (4)</td>
</tr>
</tbody>
</table>

Figure 6: (Color online) The experimental test on the performance of one- and two-fold-clustering methods in the case of MerA. (a) Comparisons of I(q) of MerA obtained from the SAXS experiment with that derived from one- and two-fold methods using RMSD as the structural metric for structural clustering. (b) Populations of sub ensembles derived from one-fold (black), two-fold methods (red).
Figure 7: Structural representatives for Sub Ensemble 1 (red), Sub Ensemble 2 (blue) and Sub Ensemble 3 (orange) obtained through cluster analysis on the simulation-generated protein conformations. To generate a sufficiently broad pool of protein conformations for further clustering to have a reasonably quantitative agreement with SAXS experimental results, we performed a gas-phase MD for MerA at 500 K [14] (see more details in the supporting information). As the experimental SAXS data were collected from a protein solution, water molecules will significantly contribute to the scattering signals. To account for the solvent contribution, the SAXS profile calculated from MD conformations is using the software FoXS [27], which assumes an implicit solvent layer.

Figure 8: A schematic picture of MerA where the crucial residues for the enzymatic function are labeled.
ever, this compact conformation weights much less as predicted by the one-fold method. Hence, the experimental test also favors the protein conformational ensemble revealed by the two-fold clustering method.

4 Conclusion

By studying three proteins with different flexibility - an intrinsically disordered protein, a compact single-domain globular protein and a multi-domain protein, the present work demonstrates that the basis-supported ensemble method can give rise to inaccurate assessment of the populations of protein conformations in solution when modeling the small angle scattering experimental data. This arises from the failure of the assumption made that structurally similar protein conformations have similar small angle scattering profiles. Here, a two-fold clustering method is developed, which classifies the simulation-derived protein conformations based on similarity on both 3D structure and scattering profiles. As confirmed by both simulation and experimental results, this new method provides exact the same set of representative protein structures as the basis-supported ensemble method does, but much more accurate populations.

Author contribution

P. Tan did the simulation, analyzed the data and wrote the manuscript. Z. Fu did the simulation and analyzed the data. ‡They contribute equally to the work. S. Qian wrote the manuscript. L. Petridis and J. Li designed the work and wrote the manuscript. L. Hong conceived the research, designed the work and wrote the manuscript.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NO. 11504231 and NO. 31630002). The SJTU group acknowledges support from Center for High Performance Computing (HPC) facility at Shanghai Jiao Tong University. LP was supported by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U. S. Department of Energy. The MD trajectory of IDP was provided by Prof. Haifeng Chen in School of Life Science and Biotechnology of SJTU.

References


[14] Richard Ledwidge, Baoyu Hong, Volker Dotsch, and Susan M Miller. Nmera of tn 501 mer-


