Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2018-0140

Vol. **25**, No. 4, pp. 1010-1023
April 2019

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

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

[1] *School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China.*
[2] *Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China.*
[3] *Zhiyuan College, Shanghai Jiao Tong University, Shanghai 200240, China.*
[4] *Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States.*
[5] *Department of Biochemistry, Cellular & Molecular Biology, The University of Tennessee, Knoxville, Tennessee 37996, United States.*
[6] *School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China.*
[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

[*]Corresponding author. *Email addresses:* `hongl3liang@sjtu.edu.cn` (L. Hong), `tpan1039@sjtu.edu.cn` (P. Tan), `zuyuefu@sjtu.edu.cn` (Z. Fu), `petridisl@ornl.gov` (L. Petridis), `qians@ornl.gov` (S. Qian), `Jinglai.li@liverpool.ac.uk` (J. Li)

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 macroscopy, 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 sub nanometer 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],