Journal of Molecular Graphics and Modelling 73 (2017) 179–190

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Combined Monte Carlo/torsion-angle molecular dynamics for ensemble modeling of proteins, nucleic acids and carbohydrates

Weihong Zhang[a], Steven C. Howell[b], David W. Wright[c], Andrew Heindel[d,1], Xiangyun Qiu[e], Jianhan Chen[a,*], Joseph E. Curtis[b,**]

[a] Kansas State University, Manhattan, KS, USA
[b] NIST Center for Neutron Research, 100 Bureau Drive, Gaithersburg, MD, USA
[c] Centre for Computational Science, Department of Chemistry, University College London, 10 Gordon St., London, UK
[d] James Madison University, Department of Chemistry & Biochemistry, Harrisonburg, VA, USA
[e] Department of Physics, The George Washington University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 13 October 2016
Received in revised form 19 January 2017
Accepted 17 February 2017
Available online 23 February 2017

MSC:
00-01
99-00

Keywords:
Monte Carlo
Torsion angle molecular dynamics
Structural biology
Small-angle scattering
SANS
SAXS

ABSTRACT

We describe a general method to use Monte Carlo simulation followed by torsion-angle molecular dynamics simulations to create ensembles of structures to model a wide variety of soft-matter biological systems. Our particular emphasis is focused on modeling low-resolution small-angle scattering and reflectivity structural data. We provide examples of this method applied to HIV-1 Gag protein and derived fragment proteins, TraI protein, linear B-DNA, a nucleosome core particle, and a glycosylated monoclonal antibody. This procedure will enable a large community of researchers to model low-resolution experimental data with greater accuracy by using robust physics based simulation and sampling methods which are a significant improvement over traditional methods used to interpret such data.

Published by Elsevier Inc.

1. Introduction

Molecular simulation encompasses a powerful and comprehensive set of methods to provide atomistic insight into a diverse set of material, chemical, and biochemical systems. The type of problem that we aim to address is the modeling of low-resolution experimental structural biology data from small-angle scattering (SAS) using either neutrons (SANS) or X-rays (SAXS). Many structural biology problems involve intrinsically disordered proteins and nucleic acids and flexible multi-domain complexes that often adopt a large range of conformations in solution. The use of molecular dynamics simulation to study such systems is often a challenging if not intractable task as many systems sample broad time and length scales that are beyond the current capabilities of algorithms and hardware. Thus a large number of structural biology problems exist where low-resolution experimental data are often modeled using analytical [1,2] and dummy-ball models [3,4]. While these methods are fast, robust, widely used and are commensurate with the idea that a low-resolution model adequately reflects low-resolution data, they are fundamentally limited by not using atomistic information.

Chemical bonding, topology and interactions are key factors to accurately model experimental data. In recent years, several groups have applied existing or developed new methods to model SAS experimental data of biological systems [5–9]. To different degrees, these methods incorporate atomistic aspects into the modeling process. We have developed a set of algorithms to perform Monte Carlo (MC) sampling of backbone-dihedral angles that can quickly generate ensembles of proteins and/or single-stranded nucleic acids to model low-resolution scattering and reflectivity data [10,11]. Recently we have implemented a MC algorithm to simulate B-DNA that utilizes an intermediate worm-like chain coarse

* Corresponding author.
** Corresponding author.
E-mail addresses: jianhanc@ksu.edu (J. Chen), joseph.curtis@nist.gov (J.E. Curtis).
[1] Current address: Drug Product Formulation Technologies, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA.

http://dx.doi.org/10.1016/j.jmgm.2017.02.010
1093-3263/Published by Elsevier Inc.

grain representation that is mapped back to all-atom coordinates [12]. These all-atom models are created and defined by mature force fields, are structurally robust and directly available for further advanced simulation methods such as accelerated molecular dynamics [13], replica-exchange [14] and reverse MC methods [15] as needed. While powerful, advanced MD simulation methods can have limited utility based on the size of the system and/or available computational resources.

MC sampling is fast and numerically inexpensive since the number of degrees of freedom are dramatically reduced. Typical systems are well sampled in minutes using commodity single CPU hardware. Such algorithms inherently sample a rugged energy landscape and thus there are limitations to these methods. Additionally, sequential moves can be highly correlated, and the basic move-set is not designed for internal or concerted moves, such as internal loops between two rigid domains. Nonetheless this approach has been successfully used to model a large number of important structural biology problems [16–21]. The combination of MC sampling with advanced simulation methods is largely not used, thus there remains a need to develop additional sampling procedures to improve the quality and robustness of such simulations. These procedures' will help realize the full-potential of fast atomistic modeling of scattering data.

In this report we discuss the combination of MC sampling with torsion-angle molecular dynamics (TAMD) simulations [22] that overcome some of the limitations of the rugged MC sampling methods. TAMD samples molecular configurations in torsion space and allows convenient specification of rigid domains and flexible degrees of freedoms consistent with sampling of MC trajectories. Using TAMD allows large time steps when propagating the equations of motion [22]. In TAMD, the molecule is represented by a branched tree structure consists of clusters of rigid-body atoms linked by hinges. Importantly, forces arising from the traditional Cartesian force field are projected along the internal coordinates in TAMD [23], thus allowing one to benefit from the significant improvement and accuracy of physics-based force fields. Using TAMD greatly increases sampling efficiency over traditional MD simulations and is widely used in NMR structure refinement [24,25].

Details of subsampling of MC ensembles and the performance of the improved algorithm are explored with several representative examples covering systems containing not only protein but also nucleic acid and carbohydrate components. In the development of the algorithm several implicit solvent models and constrained simulation variables greatly affected the efficacy of the resulting algorithm. The combination of MC and TAMD will enable the generation of robust models to interpret SAS, neutron and X-ray reflectivity, NMR, electron microscopy and other experimental data.

2. Methods

Seven systems are used in this study to illustrate the utility of the combined MC/TAMD modeling protocol. Four protein only systems of differing complexity are studied: full-length human immunodeficiency virus type 1 (HIV-1) Gag protein, two truncated Gag protein constructs (A-B and A-tail) and a TraI protein fragment. Two example nucleic acid systems are highlighted, a linear 60 bp B-DNA molecule and nucleosome core particle (NCP), the latter of which combines protein and DNA elements. The application of the methodology to carbohydrate containing systems is represented using a glycosylated and truncated human monomer IgA1 antibody molecule (PTerm455). Systems were prepared for MC simulations using CHARMM27 [26–29] and CHARMM36 force fields [30–33]. A summary of model systems is shown in Table 1.

Table 1
Model systems, flexible residues and experimental radius of gyration (RGYR) values. Errors are reported as ±1 standard deviation.

Name	Flexible residues	RGYR (Å)
HIV-1 Gag	123–143, 277–281, 354–373, 378–389, 408–412	34 ± 1 [16]
A-B (HIV-1 Gag 1–276)	123–143	29 ± 1 [16]
A-tail (HIV-1 Gag 1-143)	123–143	no data
TraI (381–858)	574–576, 790–803	36.9 ± 0.6 [34]
Linear B-DNA	chain A &B: 1–60	no data
NCP	chain A &B: 1–31, 117–147	no data
Pterm455	N-linked glycans	no data

2.1. Model proteins

Three of the protein systems used in this study were derived from the full-length human immunodeficiency virus type 1 (HIV-1) Gag protein. Full-length HIV-1 Gag protein can be divided into five globular domains, specifically, MA domain (residues 1–122), N-terminal domain of CA (residues 144–276), C-terminal domain of CA (residues 282–353), p2 "spacer" (residues 374–377), and NC domain (residues 390–431). The construction of structural model of the full-length HIV-1 Gag protein has been previously described [16]. An A-B system (first two domains of full-length HIV-1 Gag protein, residues 1–276) and an A-tail system (first domain with a flexible linker of full-length HIV-1 Gag protein, residues 1–143) were built from the full-length model. A fourth protein model taken from a fragment of TraI protein consisting of residues 381-858 was constructed as previously described [34]. This fragment contains three globular domains spanning residues 381–573, residues 577–789 and residues 804–849, respectively connected by regions of flexible amino-acids. Initial models were energy minimized for 2000 steps following by 10 ps of vacuum MD simulation using NAMD [35] and the CHARMM27 force field prior to use in the MC simulations.

2.2. Model linear B-DNA and NCP

A 60 bp linear DNA model was generated using psfgen, a plug-in of VMD [36], and based on a model of a random sequence generated by the 3D-DART DNA structure modeling server [37]. The initial NCP model was generated using psfgen and based on the PDB-ID 1KX5 X-ray structure of the NCP solved at 1.9 Å resolution [38]. Using NAMD with the CHARMM36 force field, the DNA models were prepared for MC simulations by independently performing 2000 energy minimization steps followed by 200 vacuum MD steps (0.2 ps) then another 2000 energy minimization steps. Note that the CHARMM36 force field more accurately represents the experimentally measured distribution of BI to BII DNA, a behavior not modeled well by the CHARMM27 force field.

2.3. Model protein – carbohydrate system

The initial model structure was built using coordinates from a Fc domain crystal structure (PDB: 1OW0) [39] and glycan coordinates from a previous solution modelling study [40] as described in [41]. The composition of the bianntenary N-glycan incorporated in the model was (NeuAc)2(Gal)2(GlcNAc)2(Fuc)1(Man)3(GlcNAc)2. The PTerm455 structure was prepared for simulation using the glycan reader component of CHARMM-GUI [42,43] using the CHARMM36 forcefield. The PTerm455 construct contains the full antibody structure truncated at residue 455 in the heavy chain (removing the flexible tailpiece region and a pair of N-linked glycans). The remaining structure contains two N-linked glycans and the Fc domain. Currently there is no MC move-set available to sample