Biochimica et Biophysica Acta xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

A molecular simulation protocol to avoid sampling redundancy and discover new states[☆]

Marco Bacci, Andreas Vitalis [*], Amedeo Caflisch [**]

University of Zurich, Department of Biochemistry, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 18 June 2014
Received in revised form 21 August 2014
Accepted 25 August 2014
Available online xxxx

Keywords:
Enhanced sampling
Molecular dynamics
Biomolecules
Simulation convergence
Transition path
Scalable algorithm

ABSTRACT

Background: For biomacromolecules or their assemblies, experimental knowledge is often restricted to specific states. Ambiguity pervades simulations of these complex systems because there is no prior knowledge of relevant phase space domains, and sampling recurrence is difficult to achieve. In molecular dynamics methods, ruggedness of the free energy surface exacerbates this problem by slowing down the unbiased exploration of phase space. Sampling is inefficient if dwell times in metastable states are large.
Methods: We suggest a heuristic algorithm to terminate and reseed trajectories run in multiple copies in parallel. It uses a recent method to order snapshots, which provides notions of "interesting" and "unique" for individual simulations. We define criteria to guide the reseeding of runs from more "interesting" points if they sample overlapping regions of phase space.
Results: Using a pedagogical example and an α-helical peptide, the approach is demonstrated to amplify the rate of exploration of phase space and to discover metastable states not found by conventional sampling schemes. Evidence is provided that accurate kinetics and pathways can be extracted from the simulations.
Conclusions: The method, termed PIGS for Progress Index Guided Sampling, proceeds in unsupervised fashion, is scalable, and benefits synergistically from larger numbers of replicas. Results confirm that the underlying ideas are appropriate and sufficient to enhance sampling.
General Significance: In molecular simulations, errors caused by not exploring relevant domains in phase space are always unquantifiable and can be arbitrarily large. Our protocol adds to the toolkit available to researchers in reducing these types of errors.

1. Introduction

The microscopic foundation of life sciences is an appreciation of biomolecules as complex systems. Chemical reactions, binding and assembly phenomena, and conformational transitions can all span a broad range of time and length scales. The desire to understand these processes has produced an unprecedented amount of data obtained using many different techniques. Among these, computer simulations, while marred by the caveat that they model rather than project a physical reality, have been a useful tool due to the level of resolution they offer [1,2]. Indeed, recent work has convincingly demonstrated the appropriateness and potential accuracy of the underlying physical models [3]. Molecular dynamics (MD) simulations [4], the topic of this special issue, are the most common type of computer simulations used for biomolecules. They propagate suitable equations of motion numerically. Auxiliary constructs, such as thermostats, may be required to produce well-defined, thermodynamic ensembles, most often the canonical (NVT) or isothermal-isobaric (NPT) ones.

A single, continuous MD trajectory is often expected to yield correct equilibrium statistics and realistic dynamics, although this is far from a trivial issue, in particular with respect to the numerical discretization [5]. Faithful dynamics can cause undesired precision problems if the underlying rates are low, i.e., simulations that are too short may provide limited information carrying large biases toward the initial state. If a system cannot traverse all relevant, metastable states on the simulation time scale, computed time averages will differ from correct canonical averages. Pragmatically, initial portions of simulations are discarded heuristically as "equilibration periods," and statements about simulation precision must be restricted to specific observables [1]. Due to the above, a canonical MD sampling approach (CS) will often utilize resources inefficiently, and this inefficiency has motivated the development of enhanced sampling methods over the past few decades [6–9].

It is well beyond the scope of this introduction to provide an overview of all enhanced sampling methods, and we apologize for inevitable omissions. In the following, we highlight conceptual aspects of different

Abbreviations: CLD, Cartesian Langevin dynamics; CS, canonical sampling; MC, Monte Carlo; MD, molecular dynamics; MSM, Markov state model; PCA, principal component analysis; PES, potential energy surface; PIGS, Progress Index-Guided Sampling; REX, replica exchange; RMSD, root mean square deviation

[☆] This article is part of a Special Issue entitled "Recent Developments of Molecular Dynamics."

[*] Corresponding author. Tel.: +41 44 635 5597.

[**] Corresponding author. Tel.: +41 44 635 5521.

E-mail addresses: a.vitalis@bioc.uzh.ch (A. Vitalis), caflisch@bioc.uzh.ch (A. Caflisch).

http://dx.doi.org/10.1016/j.bbagen.2014.08.013
0304-4165/© 2014 Elsevier B.V. All rights reserved.

Please cite this article as: M. Bacci, et al., A molecular simulation protocol to avoid sampling redundancy and discover new states, Biochim. Biophys. Acta (2014), http://dx.doi.org/10.1016/j.bbagen.2014.08.013

classes of methods. Better MD integrators and in particular multiple time step methods can be viewed as advanced sampling methods as they allow larger (average) time steps to be used [5,10]. The same is true for efficient protocols for constrained dynamics [11,12] or mixed numerical and analytical schemes [13]. Dynamics can also be altered by scaling masses without introducing configurational bias [14].

We next want to mention knowledge-based approaches of two different kinds. The Rosetta [15] and similar methods essentially utilize database-derived, conformational biases in conjunction with hybrid sampling protocols to explore large regions of putatively relevant phase space. For polypeptides, this can be useful if the primary goal is to identify possible states of interest. Systematic coarse-graining [16,17] is a general strategy to achieve better sampling by increasing the ratio of simulation and CPU times and often also by smoothing the ruggedness of the potential energy surface (PES). One of its main challenges lies in maintaining protocols for subsequent fine-graining that could then yield physically realistic ensembles at the resolution of interest. Both ideas have been coupled to the replica exchange (REX) method discussed below [18,19]. We mention this aspect to highlight the difficulty in delineating classes of enhanced sampling methods, which often combine existing elements and ideas in innovative ways.

Rather than the smoothing incurred by coarse-graining, direct modifications to the PES can be used to control populations and exploration rates also for fine-grained systems. Two highly successful examples are umbrella sampling [20] and metadynamics [21] (or more generally, flat histogram methods) [22], both of which are usually meant to extract an estimate of an unbiased probability distribution or a density of states along a reaction coordinate. Steered MD simulations [23] use force rather than a potential, but can be viewed as comparable. All methods are excellent for generating potentials of mean force [24] that may even allow the prediction of coarse-grained dynamics. The major caveats of this class of methods are as follows. First, choice of reaction coordinates may not be straightforward. Second, there is no control over directions orthogonal to the chosen reaction coordinates [25]. This is a fundamental problem of low-dimensional projections, viz., different, kinetically and geometrically distal states mapping to the same value of the reaction coordinate. Third, the trajectories obtained are of limited scope because populations, microscopic rates, and pathways are all altered by the modifications to the PES. Detailed kinetic information is lost, and the effective, statistical weight of a large fraction of the data may be negligible. In practice, for large systems, the snapshot-based reweighting to recover equilibrium ensembles is too noisy [26]. This issue underlies similar approaches such as the accelerated molecular dynamics method [27].

Among multicanonical techniques, the most prominent one is the REX method [28,29], and most often temperature is used to globally scale the PES. By careful algorithm design, and by partitioning the data by temperature, each subset of the data can be analyzed as a *bona fide* canonical ensemble. The idea of the method is that excursions into higher temperatures facilitate barrier crossings, despite the absence of a dedicated geometric coordinate [30]. When following data at constant temperature, the perturbations incurred by the REX method consist entirely of swapping in compatible structures from neighboring conditions. It is precisely this design that gives the method its broad appeal [31] but also limits its price-to-performance ratio because the spacing and number of conditions to use cannot be considered free choices [32–34]. If data are required only at a single condition, multiple independent runs may utilize resources more efficiently. Due to its widespread use and simplicity, we choose the REX method for comparison purposes in this contribution. Relative to CS, REX poses difficulties when inferring kinetics and pathways [35,36] because rigorously only those stretches of the trajectories in between swaps can be mined for this purpose [37].

An elucidation of pathways is of great interest for obtaining a mechanistic understanding of complex systems. Given a notion of two states to connect, techniques like transition path sampling [38], the nudged elastic band method [39], or the string method [40] are exceptionally powerful tools to understand pathway heterogeneity, predict net rates, etc. Imposing a preconceived geometry on a path ensemble may be beneficial [41,42]. If the system displays a separation of time scales, it may be possible to construct Markov state models (MSMs) [43,44], which coarse-grain phase space into a set of kinetically homogeneous or metastable states. The initial set of states is often inferred from long CS simulations or some of the enhanced sampling methods outlined above [45]. The generalization of transition path sampling approaches to include all states in a network, i.e., the combination of these two ideas [46], can potentially provide a comprehensive picture of the thermodynamics and kinetics of a complex system at a given condition and at the level of the resolution of the states of the MSM. In this contribution, we suggest a method that addresses two of the underlying goals, viz., obtaining realistic pathway information and achieving fast phase space coverage.

In order to preserve pathway information, it seems necessary to sample from an unaltered PES. A possible strategy is to guide sampling by simply restarting simulations from interesting points, a process we refer to as reseeding. In distributed computing, a fluctuation-based heuristic was suggested to monitor relevant transitions [47]. Trajectories are selectively reseeded from those points indicating that a relevant transition has occurred. Recent work has used kinetic reaction coordinates to guide sampling toward new states [48,49]. Here, we suggest a different heuristic that rewards uniqueness of the current sampling domain of individual trajectories. The scheme is scalable, unsupervised, and explicitly parallel. The decision about reseeding a given trajectory depends on the regions of phase space sampled by other trajectories. The notion of uniqueness as a guide makes it most similar to the recent WExplore method [50] that defines states by spatial discretization [51,52] to inform the reseeding procedure.

We term our approach PIGS (Progress Index-Guided Sampling) as it relies on an efficient ordering of a slice of simulation snapshots, the so-called progress index [53]. The remainder of the text is structured as follows. We first introduce the algorithm and simulation protocols. We then provide a detailed set of results evaluating the performance of the scheme on two systems, viz., a 1D model and the FS peptide [54] in implicit solvent. We are able to demonstrate that PIGS, while minimally invasive at the level of pathways, amplifies the rate of exploration, i.e., we detect several metastable states that are not reached by either REX or CS on the same time scale. By definition, PIGS ensembles are thermodynamically biased, and we do not consider refinement or reweighting here. This issue is, among others, discussed in the final section of this manuscript.

2. Materials and methods

In this section, we introduce the algorithm and describe the general setup and sampling protocols.

2.1. The PIGS algorithm

Consider a set of N_r molecular simulations (replicas) of exactly the same system and under the same conditions that are propagated by a given base sampling algorithm, e.g., CS. The stochastic sampling algorithms we use here are either Metropolis Monte Carlo or Langevin dynamics. We set a deterministic interval, f_p, for attempting to reseed up to N_r–N_p of the simulations with the final configuration (machine precision) of any of the N_p remaining replicas. The decision whether to reseed a replica or not relies on a heuristic that utilizes data from all replicas and is history-free, i.e., only data from the last f_p steps enter the analysis. The number of snapshots to use per replica for an interval of length f_p is constant and referred to as n_O throughout. Thus, the scheme is scalable and explicitly parallel. It is easy to recognize that independent runs using the base sampler are obtained if $N_r = N_p$ or if $N_r = 1$. As we will see, the heuristic is designed as an unsupervised

Please cite this article as: M. Bacci, et al., A molecular simulation protocol to avoid sampling redundancy and discover new states, Biochim. Biophys. Acta (2014), http://dx.doi.org/10.1016/j.bbagen.2014.08.013