Bayesian Computation in Big Spaces - the Case of Biomolecular Simulation

David Wild (Warwick, Statistics)

Approximately 50 years ago, Nobel Laureate Christian Anfinsen and colleagues demonstrated that protein molecules can fold into their three-dimensional 'native state' reversibly, leading to the view that these structures represented the global minimum of a rugged funnel-like 'energy landscape' [1], and it was long assumed that this native structure was the protein's functional conformation. However, in recent years there has been increased interest in intrinsically disordered proteins (IDPs), that is, proteins which do not have a well-defined fixed structure at near-physiological conditions, and which may, for example, only take on a well-defined structure upon binding [3]. This interest is because it is now understood that IDPs are significantly more common and important than previously thought, comprising a large fraction of eukaryotic proteins, and playing key roles in a range of biological processes, including cellular signalling, molecular recognition, and transcriptional regulation [6, 9, 15, 17]. There is also increasing evidence that IDPs play a major role in important diseases such as Parkinson's, Alzheimer's, type II diabetes and cancer [2, 7, 8]. The development of detailed atomistic descriptions of IDPs is a longstanding challenge in structural biology [16]. Since the structural diversity and flexibility inherent to IDPs makes their study at atomic resolution by experiment alone difficult, computer simulations currently offer perhaps the best route to understanding their biological function.

The *de facto* standard algorithm for general configurational phase space exploration is replica exchange molecular dynamics (REMD) [14]. A set of canonical MD trajectories are run with each 'replica' using a different temperature. Periodically, the swapping of conformations for two replicas is proposed and is accepted using the standard Metropolis-Hastings Monte Carlo (MC) acceptance criteria, and the high temperature replicas ensure that the system can escape from energy wells. However, the study of intrinsically disordered proteins is a field which is ripe for new simulation tools, since, for these proteins, it is not clear what the initial configurations should be for REMD *a priori*. IDPs adopt a number of different conformations, which can interchange dynamically, giving rise to a set of conformations, called an ensemble. Nested sampling (NS) is a recent Bayesian sampling algorithm, specifically designed to sample high dimensional spaces [11, 12]. The algorithm is ideally suited for the sampling of atomistic systems, and particularly proteins as, for these systems, the dimension of the phase space is large, accessible conformations (at temperatures of interest) are located in exponentially small regions of phase space and, although the nested sampling procedure is *athermal*, a single nested sampling simulation can be used to estimate thermodynamic observables at any temperature. Nested Sampling, with its top down approach to exploring phase space, offers a particularly beneficial tool for the study of the thermodynamics of intrinsically disordered proteins.

In our previous work, we have adapted the nested sampling algorithm to work within an MD framework , using a recently proposed variant in which each atom is given velocities, and system specific MC moves are not required [5]. The velocities are then used to evolve sample points using *Galilean sampling*, a novel exploration procedure, rather than using the standard Hamiltonian or canonical exploration [13]. Bouchard-Côté, *et al.* [4] have recently described a similar method for sampling high dimensional parameter spaces, referred to as the 'bouncy particle sampler'.

In this project, we propose to further develop the Galiliean nested sampling algorithm within an MD framework, and use it to study the equilibrium distributions of IDPs of biological interest. In the original description of this algorithm, Skilling suggested that certain choices of a 'semimetric', matrix, which encourages movement in some directions while discouraging it in others, could be used to improve Galilean exploration [13]. Work by previous students has demonstrated that an appropriate semimetric can improve the region of conformational space explored by the Galilean Monte Carlo algorithm. This will be essential for the study of biomolecular systems, since certain degrees of freedom, such as the stretching of covalent bonds, are very highly constrained, whereas others, such as the dihedral angles, are not very constrained at all. It is clear that the magnitude of velocities in the highly constrained directions should be smaller than those in other directions. A possible solution is offered by recent work by Mones *et al.*[10], which has developed a new set of preconditioners, based on a decomposition of the Hessian of molecular mechanical terms with arbitrary form, which results in a positive definite matrix. The construction of the sparse preconditioner matrix then requires only the computation of the gradient of the corresponding molecular mechanical terms that are available in popular force field-based program packages such as AMBER. A Python implementation of these preconditioners with several potential forms of nonbonded terms is available within the Atomic Simulation Environment (ASE) (`https://gitlab.com/molet/ase`).

Experience of Python and or C/C++ will be essential and experience of shell script and FORTRAN programming (or an ability to understand them) and an interest in biomolecular systems would be an advantage

for the project.

References

[1] CB Anfinsen. Principles that govern the protein folding chains. *Science*, 181:233–230, 1973.

[2] M Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins: regulation and disease. *Current opinion in structural biology*, 21(3):432–440, 2011.

[3] Christopher M Baker and Robert B Best. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *WIREs Comput. Mol. Sci.*, 4:182–198, 2014.

[4] Alexandre Bouchard-Côté, Sebastian J Vollmer, and Arnaud Doucet. The bouncy particle sampler: A non-reversible rejection-free markov chain monte carlo method. *arXiv preprint arXiv:1510.02451*, 2015.

[5] Nikolas S Burkoff, Robert JN Baldock, Csilla Várnai, David L Wild, and Gábor Csányi. Exploiting molecular dynamics in nested sampling simulations of small peptides. *Computer Physics Communications*, 201:8–18, 2016.

[6] A Keith Dunker, Celeste J Brown, J David Lawson, Lilia M Iakoucheva, and Zoran Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–6582, 2002.

[7] A Keith Dunker, J David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, Juan Ausio, Mark S Nissen, Raymond Reeves, ChulHerr Kang, Charles R Kissinger, Robert W Bailey, Michael D Griswold, Wah Chiu, Ethan C Garner, and Zoran Obradovic. Intrinsically disordered protein. *J. Mol. Graphics Modell.*, 19(1):26–59, 2001.

[8] A Keith Dunker, Israel Silman, Vladimir N Uversky, and Joel L Sussman. Function and structure of inherently disordered proteins. *Current opinion in structural biology*, 18(6):756–764, 2008.

[9] H Jane Dyson and Peter E Wright. Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology*, 12(1):54–60, 2002.

[10] Letif Mones, Christoph Ortner, and Gábor Csányi. Preconditioners for the geometry optimisation and saddle point search of molecular systems. *Scientific reports*, 8(1):1–11, 2018.

[11] John Skilling. Nested sampling. *AIP Conf. Proc.*, 735:395–405, 2004.

[12] John Skilling. Nested Sampling for General Bayesian Computation. *J. Bayesian Anal.*, 1(4):833–859, 2006.

[13] John Skilling. Bayesian computation in big spaces-nested sampling and galilean monte carlo. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1443, pages 145–156, Melville, NY, USA, 2012. AIP Publishing.

[14] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1):141–151, 1999.

[15] Peter Tompa. Intrinsically unstructured proteins. *Trends in biochemical sciences*, 27(10):527–533, 2002.

[16] Peter Tompa. Unstructural biology coming of age. *Current opinion in structural biology*, 21(3):419–425, 2011.

[17] Vladimir N Uversky and Anthony L Fink. Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1698(2):131–153, 2004.