EXTENSIBLE AND SCALABLE ADAPTIVE SAMPLING TO FOLD PROTEINS ON SUPERCOMPUTERS

Allocation: NSF PRAC/5,200 Knh
PI: Shantenu Jha[1]
Collaborators: Cecilia Clementi[2], Eugen Hruska[2]

[1]Rutgers, The State University of New Jersey
[2]Rice University

EXECUTIVE SUMMARY

The Extensible Toolkit for Advanced Sampling and analYsis (ExTASY) is a software toolkit to effectively simulate protein folding on supercomputers. The use of adaptive sampling for proteins achieves a shorter time-to-solution than brute-force molecular dynamics (MD) but requires a more complex workflow. The research team has shown that ExTASY can effectively execute adaptive sampling and produce accurate simulations of protein folding and protein dynamics. The ExTASY package allows researchers to utilize and investigate different sampling strategies with great flexibility. The effective and scalable execution on supercomputers is ensured by RADICAL–Cybertools, a suite of Python modules that enables interoperability across high-performance computing machines.

RESEARCH CHALLENGE

The previous version of ExTASY was developed to reduce the complexity of adaptive sampling and was used by the research team to show in [1] that ExTASY can scale complex workflows on supercomputers. The next step for ExTASY was to demonstrate an end-to-end execution of adaptive sampling for reference proteins. By comparison with reference results, the research team confirmed that adaptive sampling delivers accurate results for protein folding and protein dynamics and that the team could investigate the achieved speed-up. For three proteins—Chignolin, BBA, and Villin, with 10, 28, and 35 residues, respectively—the protein folding and protein dynamics are well understood and are good model reference proteins to test the performance of ExTASY.

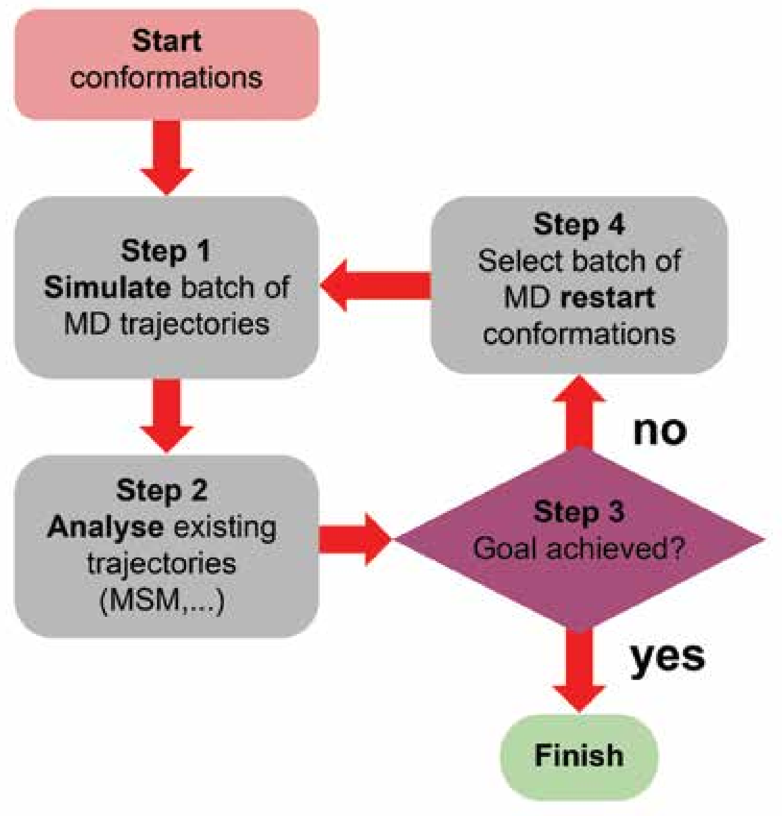

Figure 1: Adaptive sampling requires an iterative workflow with the individual steps requiring different parallelization. Step 1 comprises long, parallel MD simulations. In contrast, steps 2–4 require only a single node. To fold a protein, these steps have to be repeated hundreds of times. A robust and effective workflow management toolkit is essential to enable more researchers to execute adaptive sampling.

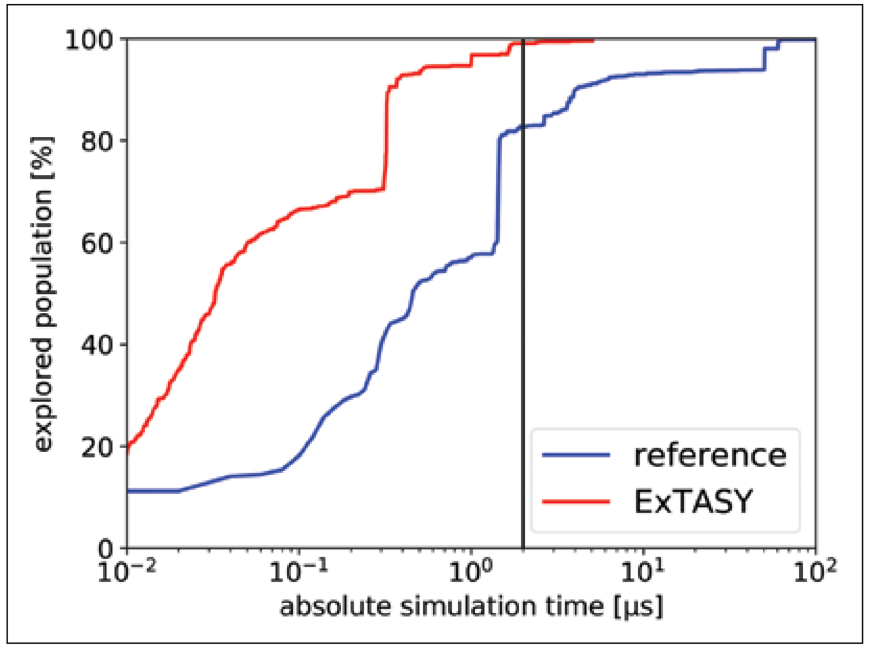

Figure 2: The brute-force MD simulation (blue) requires an order of magnitude longer time-to-solution than the adaptive sampling solution utilizing ExTASY (red). The more effective exploration strategy requires determining the restarting points adaptively during runtime. The ExTASY framework allows effective execution of these adaptive sampling strategies.

METHODS & CODES

The ExTASY framework uses RADICAL–Cybertools or, specifically, the Ensemble Toolkit and RADICAL–Pilot, which ensures scalability, extensibility, and ease of deployment on high-performance computing platforms. The building block capabilities of RADICAL–Cybertools greatly increase the flexibility of ExTASY to utilize different sampling strategies in an easy fashion. Additionally, RADICAL–Cybertools enables execution of a complex workflow without explicit resource management, which also ensures the maintainability of the ExTASY workflow.

The adaptive sampling of proteins is an iterative process where MD and analysis steps alternate (Fig. 1). In general, adaptive sampling strategies pick optimal restarting coordinates for the next iteration of MD. This enables more effective use of computational resources in undersampled areas. In this project, the researchers utilized two different adaptive sampling strategies, *cmacro* and *cmicro*, to first effectively fold the protein and then reach accurate protein dynamics. Both strategies generate Markov state models from all generated MD trajectories in step 2. In step 4, the *cmacro* strategy picks Markov states to effectively cross transition barriers. Once the folded state is found, the *cmicro* strategy can be used to increase the accuracy of protein dynamics. Further comparison of different adaptive sampling strategies is discussed in [2]. The ExTASY code is open source and is provided at https://github.com/ClementiGroup/ExTASY.

RESULTS & IMPACT

The new version of the ExTASY workflow folds proteins with a shorter time-to-solution than brute-force MD. Fig. 2 shows that adaptive sampling utilizing ExTASY is about one order of magnitude faster than brute-force MD. Additional results confirming the accuracy of protein folding and protein dynamics may be found in [3]. By utilizing different adaptive sampling strategies than in previous versions of ExTASY, the research team has shown that this workflow can be easily adapted to different exploration strategies.

WHY BLUE WATERS

Protein folding simulations require large numbers of GPU node-hours despite the speed-up achieved by ExTASY. Blue Waters is essential to deliver these computational resources. The investigated proteins are relatively small and undergo fast folding; larger proteins would require even larger computational resources.

PUBLICATIONS & DATA SETS

E. Hruska, V. Balasubramanian, J. R. Ossyra, S. Jha, and C. Clementi, "Extensible and scalable adaptive sampling on supercomputers," 2019, arXiv: 1907.06954.

E. Hruska, J. R. Abella, F. Nüske, L. E. Kavraki, and C. Clementi, "Quantitative comparison of adaptive sampling methods for protein dynamics," *J. Chem. Phys.*, vol. 149, no. 24, p. 244119, 2018, doi: 10.1063/1.5053582.

V. Balasubramanian *et al.*, "ExTASY: Scalable and flexible coupling of MD simulations and advanced sampling techniques," in *Proc. 2016 IEEE 12th Int. Conf. e-Science*, Baltimore, MD, U.S.A, Oct. 23–27, 2016, pp. 361–370.