PLUMED-WESTPA: A Toolbox for Performing Weighted Ensemble Simulation Using PLUMED and WESTPA

Dhiman Ray

Department of Chemistry, University of California Irvine, Irvine CA 92697

E-mail:

Introduction

Weighted Ensemble (WE)[1,2] is a path sampling approach for studying long-timescale processes using molecular dynamics (MD)[3] and Monte-Carlo (MC) simulation. WE has gained significant popularity over the last decade due to its successful application primarily in the field of computational biophysics for studying biomolecular rare-events with varying degrees of complexity.[4] One of the biggest triumph of WE method has been the extensive sampling of the pathways of conformational opening in the SARS-CoV-2 spike protein, a system comprising of 0.5 million atoms and timescales in the range of seconds.[5] WE based approaches are now applied routinely to study processes like protein-folding,[6–8] conformational change, and ligand-unbinding,[9–11] either independently or in conjunction with other accelerated sampling schemes like Gaussian Accelerated Molecular Dynamics (GaMD)[12] or milestoning (see for example, Weighted Ensemble Milestoning (WEM),[13] Weighted Ensemble Milestoning with restraint release (WEM-RR),[14] and Markovian Weighted Ensemble Milestoning (M-WEM)[15]). WE and associated methods are implemented in the open source software WESTPA, which allowed its widespread application and quick development and growth.[16] A detailed discussion on the WE methods can be found in Ref 4 and more details on the WESTPA software can be obtained from the following references: 16, 17, 18.

The WE approach, similar to most other path sampling and enhanced sampling techniques, require the user to choose a handful of progress coordinates or collective variables (CV) along which the sampling is increased. The quality of the results obtained from the simulation can depend on the CV, making the choice of CV, a key step for successful simulations. A wide range of innovative and complicated CV's have been developed in recent days, many of which are based on Machine Learning based approaches like principal component analysis (PCA),[19] time-lagged independent component analysis (tICA),[20] Linear Discriminant Analysis (LDA),[21] Variational Autoencoders (VAE)[22,23] and Support Vector Machines.[24] Most of the novel CVs are implemented in PLUMED software,[25,26] which has played a significant role in the development and application of enhanced sampling methods in molecular

dynamics simulations. PLUMED consortium is also one of the first open science initiatives implemented in the molecular simulation community.[27] Another advantage of PLUMED is that it is compatible with many MD simulation software capable of performing both classical and ab-initio MD.

Here, we present a toolbox for combining PLUMED with WESTPA to perform weighted ensemble simulations using progress coordinates defined using PLUMED program. This will facilitate the use of newly developed advanced collective variables to be used as progress coordinate in WE simulation in WESTPA. In the following section we discuss the implementation and our outlook.

Theory of Weighted Ensemble

A single long MD trajectory, initiated from a specific starting condition, often remains trapped within the free energy minimum corresponding to the initial state due to the presence of high energy barriers. This causes a two fold problem: first, a large part of the configurational space remains unexplored and, second, a majority of the computational effort is wasted on oversampling the initial state minimum.

Weighted Ensemble approach avoids this problem by discretizing the conformational space between the starting and target state in multiple bins. A number of short trajectory segments are propagated from the starting state for a short time period, say δt. If some of those trajectories enter a new bin, they are split (i.e. new trajectories are initiated from the end point of the previous trajectory) to maintain a specific number of trajectories per bin. If the number of trajectories in a bin is more than that specified number, the trajectories are merged (i.e. the excess trajectories are discontinued). During the splitting and merging process the probabilistic weights of the trajectories are redistributed among the surviving trajectories in the bin to preserve the total probability to be one. In this way, the sampling is increased in under-sampled regions of the conformational space. This process is repeated

for multiple iterations until sufficient sampling or converged results are obtained. After a number of iterations, some very low weight trajectories reach the final state, from which kinetics can be calculated. Thus, WE approach allows for the direct evaluation of kinetic properties as no biasing force is applied on the system. In steady state weighted ensemble, the trajectories reaching the end state are recycled from the starting state to facilitate the attainment of a non-equilibrium steady state which leads to a quicker convergence of mean first passage time (MFPT) calculated based on the Hill relation.[28]

In WESTPA software, the trajectory splitting, merging, recycling etc. are performed by some well designed python and shell scripts included within the package. The trajectory segments belonging to each bin is propagated using standard MD simulation packages like AMBER,[29] GROMACS,[30] NAMD[31] etc. For calculating the value of the progress coordinate (the order parameter separating the initial and final state, along which bins are placed), usually, a script is prepared by the user which analyzes the trajectory files. We propose to employ PLUMED to compute the progress coordinate to seamlessly use complex reaction coordinates without having to implement them from scratch in an in house script.

Two Approaches of Implementation

We demonstrate two alternate implementations:

Using PLUMED in on-the-fly mode

In the "on-the-fly" mode, the progress coordinate is calculated during the simulation using PLUMED and printed in the COLVAR file. The `progress_coord.py` script, then, converts it into a format readable by the WESTPA software.

This approach is beneficial is relatively small (<100,000 atom) systems and CPU based hardware, because PLUMED reaction coordinate calculation has to be done in the CPU core. Multiple communication between CPU and GPU may slow down the dynamics. One

advantage of on the fly technique is that we do not need to save the dcd file frequently, we can just save it at the end of each iteration. But we can print the collective variable much more frequently using the PLUMED program.

Using PLUMED in post-analysis mode

In the post-analysis mode, the PLUMED process is not invoked during the simulation within one segment. After the simulation is propagated for one iteration, `plumed driver` is used to recover the collective variable. This approach is advantageous for very large system simulated using GPU based hardware, because PLUMED does not need to exchange coordinate information between GPU and CPU during the propagation of simulation. But the coordinate value is only available at the same (or lower) frequency at which the dcd frames are saved. This may lead to saving very large trajectory files; additional scripts may be necessary to process these files afterwards to reduce their size.

Requirements

Software

- PLUMED 2.7+
- WESTPA 2.0
- GROMACS 2018+

Installation

Installation of PLUMED, GROMACS and patching PLUMED with GROMACS:
`https://birdlet.github.io/trash/plumed_gromacs_install.html`
(Also GROMACS pre-patched with PLUMED can be installed from conda:
`https://github.com/plumed/conda`)

Installing WESTPA 2.0 :

https://github.com/westpa/westpa/wiki/WESTPA-Quick-Installation

Prerequisites

A basic understanding of both WESTPA toolkit and PLUMED plugin is necessary to use the PLUMED-WESTPA toolbox. Furthermore, practical knowledge of Weighted Ensemble, Collective Variables, and stochastic trajectories is necessary.

Outline of the Example Implementation

We used the simple system of Na+ and Cl- association in Generalized Born Implicit Solvent (GBIS) model[32] to illustrate our protocol of integrating PLUMED with WESTPA. We use PLUMED 2.7 patched with GROMACS 2018.6[30] software, but our approach can be generalized to any simulation package that can be patched with PLUMED.

The basics of running WE simulation can be obtained from previous tutorials of WE. Particularly the following tutorial will be useful for this specific example:

https://github.com/westpa/westpa/wiki/Na--Cl--Association-with-Gromacs-2018.2.

Although this tutorial is for an explicit solvent Na-Cl system, the workflow of the weighted ensemble simulation part is very similar once the solvation and equilibration steps are ignored. The implicit solvent NaCl association tutorial is modeled after the following WESTPA 1.0 tutorial:

https://github.com/westpa/westpa_tutorials/tree/main/other_tutorials/nacl_gmx_gb.

Appropriate modifications are made to make the PLUMED-WESTPA toolbox example compatible with WESTPA 2.0, the latest release.

We provide a brief overview of the file structure and user guide of our example below.

File Organization and User Instructions

- For both on-the-fly and post-analysis mode, the input files are provided in the `nacl_gb` directory.

- The `gromacs_config` directory contains all the input files necessary for propagating the simulation in GROMACS.

    1. The topology and parameters are included in `nacl.top` and `nacl.tpr` files respectively. These input files are for GROMACS software and specific for the Na-Cl association process. The user has to prepare these files separately for their specific system. Instruction on preparing these files can be obtained from the above-mentioned tutorial.

    2. The `md-genvel.mdp` file is the simulation input file for the first WE iteration, where random velocities are generated according to the specified temperature. In cases where the WE simulation is initiated from the end point of a previous equilibration simulation, this step may not be necessary. We refer the user to WESTPA tutorials from Ref 17 for examples of both kind. We are starting from a PDB file without any minimization or equilibration due to the simplified nature of our system. So we need to separately generate velocities.

    3. The `md-continue.mdp` is the GROMACS input file for the subsequent WE iterations. This will propagate one segment in one bin. The "RAND" word is used where a random seed is supplied to make sure that each trajectory follow different path because of the stochastic integrator. Use of a stochastic integrator like Langevin dynamics is necessary for WE simulation.

    4. The `plumed.dat` is the PLUMED input file for computing the distance between the Na+ and Cl- ions. Here the user should implement the appropriate collective variable for their system.

- The input file for WESTPA simulation is the `west.cfg` file, which contains the bin definitions, number of trajectory segments per bin, total number of iterations etc. This should be modified by the user according to their system. Usually more iterations are required if the results are not converged. The sampling also improves with increasing number of segments per bin but so does the computational cost.

- The `bstates` directory contains the starting coordinate in the PDB format and a python script to calculate the progress coordinate value at the starting point.

- The `westpa_scripts` directory contains the necessary shell and python scripts for weighted ensemble simulation. Particularly noteworthy are the following files:

    1. `progress_coord.py` calculates the progress coordinate from the COLVAR file created by PLUMED.

    2. `get_pcoord.sh`, is used to evaluate the value of the progress coordinate at the beginning of the first iteration using `plumed driver` from the pdb file of the starting configuration.

    3. `runseg.sh` performs the propagation of weighted ensemble trajectory. It links the necessary input files from other directories before executing the `gmx grompp` and `gmx mdrun` commands. Each segment of the WE simulation should begin from its parent trajectory from the previous segment. So the output `parent.gro` is used as a starting point for the segment.

       This is by far the most important script in this directory as it dictates both the simulation propagation and progress coordinate calculation.

       On-the-fly mode: In the on-the-fly mode the plumed is called during the GROMACS mdrun command:

       ```
       $GMX mdrun -s seg.tpr -o seg.trr -c seg.gro -e seg.edr -cpo seg.cpt -g seg.log -plumed plumed.dat
       ```

Then the progress coordinate is calculated from the COLVAR output file of PLUMED using `progress_coord.py`:

```
python $WEST_SIM_ROOT/westpa_scripts/progress_coord.py > $WEST_PCOORD_RETURN
```

Post-analysis mode: In this case GROMACS is propagated without PLUMED:

```
$GMX mdrun -s seg.tpr -o seg.trr -c seg.gro -e seg.edr -cpo seg.cpt -g
seg.log
```

But the PLUMED is used to extrac the CV by analyzing the trajectory:

```
plumed driver --plumed plumed.dat --mf_trr seg.trr
python $WEST_SIM_ROOT/westpa_scripts/progress_coord.py > $WEST_PCOORD_RETURN
```

- For running the simulation, user needs to first initiate the system by executing `init.sh` which will output the number of WE bins, total number of segments etc. Then the WE simulation can be propagated by executing `run.sh`. The scripts are quite self explanatory so a detailed discussion is not included here.

  These scripts need to be modified in case they are to be submitted in slurm or PBS queue system. For example check `runwe.slurm` in

  `https://github.com/westpa/westpa_tutorials/blob/main/basic_nacl/runwe.slurm`

Discussions and Conclusions

In this tutorial, we demonstrate the PLUMED-WESTPA toolbox which can be used to design complex progress coordinates using PLUMED software for the use in weighted simulation in WESTPA toolkit. It should be noted that PLUMED-WESTPA is not a software that one can install; this is just a template which can be used to design simulation protocols interesting to the user.

The performance reduction (or increase) of WESTPA simulation with the integration of PLUMED has not been tested yet, although we wish to do that in future. Technically

the performance reduction should be negligible in the post-analysis mode as long as system is large enough where propagating molecular dynamics trajectories is the performance bottleneck, and not the trajectory cloning, merging or recycling. In the post-analysis mode PLUMED does not interact with the simulation engine during the propagation of the segment. So the simulation can be as efficient as possible by the availability of computing resources.

The current framework only shows the implementation of the very simple inter-atomic distance based CV. But, the implementation can be generalized to other CVs with increased complexities. The use of advanced collective variables has transformed the field of enhanced sampling simulation over the past decade. Starting from path collective variables to the more recent machine learning based CVs obtained using Harmonic Linear Discriminant Analysis (HLDA),[21] Deep-LDA,[33] Spectral Gap Optimization of Order Parameters (SGOOP),[34] tICA,[20] PCA[19] etc have been useful in identifying the key degrees of freedom in complex transformations in biomolecular and material systems. Using our toolbox, such novel CVs can be directly applied for propagating weighted ensemble trajectories without re-implementing them in separate codes.

Software availability

The PLUMED-WESTPA Toolbox is available from the following GitHub repository:

`https://github.com/dhimanray/PLUMED-WESTPA`

This repo and the manuscript will be updated to add additional features when the need arises.

References

(1) Huber, G.; Kim, S. Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophysical Journal* **1996**, *70*, 97–110.

(2) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of Chemical Physics* **2010**, *132*, 054107.

(3) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **2002**, *9*, 646–652.

(4) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annual Review of Biophysics* **2017**, *46*, 43–57.

(5) Sztain, T. A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat. Chem.* **2021**, *13*, 963–968.

(6) Abdul-Wahid, B.; Feng, H.; Rajan, D.; Costaouec, R.; Darve, E.; Thain, D.; Izaguirre, J. A. AWE-WQ: Fast-Forwarding Molecular Dynamics Using the Accelerated Weighted Ensemble. *Journal of Chemical Information and Modeling* **2014**, *54*, 3033–3043.

(7) Adhikari, U.; Mostofian, B.; Copperman, J.; Subramanian, S. R.; Petersen, A. A.; Zuckerman, D. M. Computational Estimation of Microsecond to Second Atomistic Folding Times. *Journal of the American Chemical Society* **2019**, *141*, 6519–6526.

(8) Suárez, E.; Lettieri, S.; Zwier, M. C.; Stringer, C. A.; Subramanian, S. R.; Chong, L. T.; Zuckerman, D. M. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *Journal of Chemical Theory and Computation* **2014**, *10*, 2658–2667.

(9) Dickson, A.; Lotz, S. D. Ligand Release Pathways Obtained with WExplore: Residence Times and Mechanisms. *Journal of Physical Chemistry B* **2016**, *120*, 5377–5385.

(10) Dickson, A.; Lotz, S. D. Multiple Ligand Unbinding Pathways and Ligand-Induced Destabilization Revealed by WExplore. *Biophysical Journal* **2017**, *112*, 620–629.

(11) Donyapour, N.; Roussey, N. M.; Dickson, A. REVO: Resampling of ensembles by variation optimization. *Journal of Chemical Physics* **2019**, *150*, 244112.

(12) Ahn, S.-H.; Ojha, A. A.; Amaro, R. E.; McCammon, J. A. Gaussian-Accelerated Molecular Dynamics with the Weighted Ensemble Method: A Hybrid Method Improves Thermodynamic and Kinetic Sampling. *Journal of Chemical Theory and Computation* **2021**, *17*, 7938–7951.

(13) Ray, D.; Andricioaei, I. Weighted ensemble milestoning (WEM): A combined approach for rare event simulations. *The Journal of Chemical Physics* **2020**, *152*, 234114.

(14) Ray, D.; Gokey, T.; Mobley, D. L.; Andricioaei, I. Kinetics and free energy of ligand dissociation using weighted ensemble milestoning. *The Journal of Chemical Physics* **2020**, *153*, 154117.

(15) Ray, D.; Stone, S. E.; Andricioaei, I. Markovian Weighted Ensemble Milestoning (M-WEM): Long-Time Kinetics from Short Trajectories. *Journal of Chemical Theory and Computation* **2021**, acs.jctc.1c00803.

(16) Zwier, M. C.; Adelman, J. L.; Kaus, J. W.; Pratt, A. J.; Wong, K. F.; Rego, N. B.; Suárez, E.; Lettieri, S.; Wang, D. W.; Grabe, M.; Zuckerman, D. M.; Chong, L. T. WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis. *Journal of Chemical Theory and Computation* **2015**, *11*, 800–809.

(17) Bogetti, A. T.; Mostofian, B.; Dickson, A.; Pratt, A.; Saglam, A. S.; Harrison, P. O.; Adelman, J. L.; Dudek, M.; Torrillo, P. A.; DeGrave, A. J.; Adhikari, U.; Zwier, M. C.; Zuckerman, D. M.; Chong, L. T. A Suite of Tutorials for the WESTPA Rare-Events Sampling Software [Article v1.0]. *Living Journal of Computational Molecular Science* **2019**, *1*.

(18) Russo, J. D. et al. WESTPA 2.0: High-performance upgrades for weighted ensemble simulations and analysis of longer-timescale applications. *bioRxiv* **2021**, 2021.12.05.471280.

(19) Jolliffe, I. *Principal Component Analysis*, 2nd ed.; Springer-Verlag: New York, 2002; p 488.

(20) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; Fabritiis, G. D.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102.

(21) Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. *Journal of Physical Chemistry Letters* **2018**, *9*, 2776–2781.

(22) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *The Journal of Chemical Physics* **2018**, *149*, 072301.

(23) Bandyopadhyay, S.; Mondal, J. A deep autoencoder framework for discovery of metastable ensembles in biomacromolecules. *The Journal of Chemical Physics* **2021**, *155*, 114106.

(24) Grazioli, G.; Andricioaei, I. Advances in milestoning. II. Calculating time-correlation functions from milestoning using stochastic path integrals. *The Journal of Chemical Physics* **2018**, *149*, 084104.

(25) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009**, *180*, 1961–1972.

(26) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **2014**, *185*, 604–613.

(27) Bonomi, M. et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods 2019 16:8* **2019**, *16*, 670–673.

(28) Zuckerman, D. M. *Statistical Physics of Biomolecules: An Introduction*; 2010; p 334.

(29) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005**, *26*, 1668–1688.

(30) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindah, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(31) Phillips, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **2020**, *153*, 044130.

(32) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 383–394.

(33) Rizzi, V.; Bonati, L.; Ansari, N.; Parrinello, M. The role of water in host-guest interaction. *Nature Communications 2021 12:1* **2021**, *12*, 1–7.

(34) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113*, 2839–2844.