Enhanced Sampling Techniques

Contents

Chapter 1

Replica Exchange Solute Tempering Methods

Jaya Krishna Koneru and Korey Reid, and Paul Robustelli

1.1 Abstract

Accurate simulations of intrinsically disordered proteins and peptides predictive power and insight into experimental findings extending our predictive capabilities in molecular interactions and understanding biological mechanisms. Unlike equilibrium simulations of ordered proteins in their ground state, simulating disordered proteins, as well as rare allosteric effects in structured proteins, require long continuous simulations which may not be well sampled even out to 1 ms. Replica exchange molecular dynamics simulations with solvent scaling proves to be a powerful group of methods to sample dynamics with limited resources. REST2 and ssREST3 are promising methods in accelerating the sampling of the conformational ensemble, not ommiting the caveat: the simulation is only as good as the forcefield. In addition to discussing the background, we provide an example of REST2 and ssREST3 as it pertains to simulating the C-terminal domain of α-synuclein with and without a small molecular ligand and with an emphasis on assessing convergence of desired collective variables or observables.

Keywords Molecular dynamics, Replica Exchange, Collective variables, Solute-Scaling, α-synuclein

1.2 Introduction

Studying and critically understanding the underlying dynamics of biological systems at the atomistic scale can provide valuable insight into molecular mechanisms. Over many decades experimental techniques have been and continue to be developed in a collective effort to investigate processes at the molecular scale. However, the limitations of experimental techniques often are restricted from producing clear and relevant explaintions at the atomistic level. This is in large part to the ensemble characteristic of experimental measurements and lends to the difficulty of extracting conformational dynamics at various timescales, as well as extracting localized dynamics of say a biological system. Upon entry of molecular dynamics the usefullness was apparent, albiet limited by computational speed. As molecular dynamics engines[24, 5, 17, 3, 2, 20] and their accompanied forcefields[6, 12, 4, 9, 16, 7, 11] have matured, our ability to reach microsecond timescales has become routine. While the product of many decades of effort have resulted in our ability to simulate microseconds for a single biomolecule, vastly extending our comprehension of atomistic dynamics when paired with experimental results, these timescales are in fact beneath the threshold required to study rare events, e.g. allosteric transitions, or sample large degrees of freedom often accompanying IDPs.

When the desire is to study conformational dynamics it becomes a neccessity for long-timescale simulations on supercomputing systems[19, 18], for all other cases where statistical measures of an ensemble enhanced sampling techniques are available[8, 23, 14, 22, 25, 15, 13]. The phenomenolog-

ical observation desired dictates which simulation methodology will provide usefull information. One such method, Replica Exchange Solute Tempering or REST[10, 23, 25], has arrisen from Hamiltonian Replica Exchange Molecular Dynamics (HREMD) a derivative of Replica Exchange Molecular Dynamics (REMD). For each of these methods, multiple parallel simulations are conducted in parallel. These parallel simulations undergo exchanges at a designated interval. The relevance differences in each exchange method is discussed in the Introduction and Theory section, for a detailed account please refer to the original publications[21, 10, 23].

From Replica Exchange Molecular Dynamics[21], the hamiltonian representing the potential energy of the system can be written as sum of respective contributions, separated into protein-protein, protein-water and water-water:

$$E_n^{REMD}(X_n) = \lambda_n^{pp} E_{pp}(X_n) + \lambda_n^{pw} E_{pw}(X_n) + \lambda_n^{ww} E_{ww}(X_n) \tag{1.1}$$

λ_n^M is the scaling factor, where $M = \{pp, pw, ww\}$ which scales the corresponding energy term. For REST2[23], $\lambda_n^{ww} = 1$, $\lambda_n^{pp} = (\lambda_n^{pw})^2 = \lambda_n$, for simplicity the REST2 hamiltonian simplifies to:

$$E_n^{REST2}(X_n) = \lambda_n E_{pp}(X_n) + \sqrt{\lambda_n} E_{pw}(X_n) + E_{ww}(X_n) \tag{1.2}$$

where,

$$\lambda_n = \frac{\beta_n}{\beta_0} \tag{1.3}$$

and $\beta_n = \frac{1}{k_B T_n}$ for $n = \{0, 1, 2, \ldots, n_{replica}\}$.

Upon implementation of metropolis acceptance criteria with detail balance being satisfied, the acceptance probability between state n and $n+1$ is given by:

$$\begin{aligned}\Delta_{n,n+1} = (\beta_n - \beta_{n+1})\Big[&(E_{pp}(X_{n+1}) - E_{pp}(X_n)) \\ &+ \frac{\sqrt{\beta_0}}{\sqrt{\beta_n} + \sqrt{\beta_{n+1}}}(E_{pw}(X_{n+1}) - E_{pw}(X_n))\Big].\end{aligned} \tag{1.4}$$

By excluding the solvent-solvent interactions w.r.t. scaling, the contribution to the acceptance criteria contains less degrees of freedom relative to REMD resulting in better acceptance between replicas.

Upon investigation, disordered proteins containing hydrophobic residues undergo conformational collapse with respect to scaling E^{pw} to higher effective temperatures. This outcome is unfavorable when attempting to capture a representative ensemble as hydrophobic collapse reduces the overall sampling of the proteins conformational ensemble. Zhang, Liu, and Chen 2023 provided a basis for biasing the scaling such that protein collapse is minimized or negated, with the hamiltonian:

$$E_n^{REST3}(X_n) = \lambda_n E_{pp}(X_n) + \sqrt{\lambda_n}\Big[E_{pw}^{elec} + \kappa_n E_{pw}^{lj}\Big](X_n) + E_{ww}(X_n). \tag{1.5}$$

By incorporating the scaling factor κ_n, where $n = \{1, 2, 3, \ldots, n_{replica}\}$, the dampening effect of λ is deminished. To accomplish this κ_n exist in

the range [1.0, 1.1]. Due to the method of implementation the base replica (unscaled topology), following conversion of the topologies to REST3 [25] formalism using combination rule 1 instead of 2, does not match the potential energy of the original topology. The implementation of REST3 by Zhang, Liu, and Chen targeted scaling of protein-water lj parameters by means of computing C6 and C12 parameters. The descrepancy between the unscaled, converted topology and the original topology led to inception of ssREST3, solvent-scaled REST3, where λ_n is still applied to the protein in the same fashion as REST2 [23] and solvent scaling is applied to the water oxygen via the well-depth, ϵ_{OW}, borrowing from the methods described in Best and Mittal 2010. To overcome the inherent mechanics within GROMACS[20], nonbonded overrides are implemented for water-water and water-ion. Additionally, one can override water-cosolute if scaling is not desired between the two. For ssREST3, κ_n is defined by the expression,

$$\kappa_n = \kappa_{low} * \exp\left(n * \frac{\log(\kappa_{high}/\kappa_{low})}{N_r - 1}\right) ; \quad 1.00 \leq \kappa_n \leq 1.10. \tag{1.6}$$

To understand where κ_n is applied, we start with an expression for the Lennard-Jones potential between the i^{th} protein atom and the water oxygen,

$$\sqrt{\lambda_n} \cdot \kappa_n \cdot E^{lj}_{i,OW} = \sqrt{\lambda_n} \cdot \kappa_n \cdot 4 \cdot \epsilon_{i,OW}\left[\left(\frac{\sigma_{i,OW}}{r}\right)^{12} - \left(\frac{\sigma_{i,OW}}{r}\right)^{6}\right] \tag{1.7}$$

The CHARMM and Amber forcefields both conform to the Lorentz-Berthelot rules, i.e. $\epsilon_{i,j} = \sqrt{\epsilon_i \cdot \epsilon_j}$. From Equation **??** we can refactor $\sqrt{\lambda_n} \cdot \kappa_n \cdot \epsilon_{i,OW}$ to clarify which forcefield parameters are scaled.

$$\epsilon^{scaled}_{i,OW} = \sqrt{\lambda_n} \cdot \kappa_n \cdot \epsilon_{i,OW} = \sqrt{(\lambda_n \cdot \epsilon_i) \cdot (\kappa_n^2 \cdot \epsilon_{OW})} \tag{1.8}$$

The sections that follow are the Materials, Methods and Notes sections. Within the materials sections we detail the required software and minimum hardware requirements to perform REST based simulations. The Methods sections contains a complete description with examples on how to create and produce REST2 and ssREST3 simulations following an example simulation involving α-synuclein and a small molecule. Additionally, we include a few examples of analysis to test for convergence. Lastly, we provide various Notes in the last section.

1.3 Materials

Spacial co-ordinates and force driving parameters for the simulations.

In molecular simulations, the initial positions of particles play a crucial role in determining the outcome and accuracy of the simulation. The Protein Data Bank (PDB) files provide the necessary three-dimensional atomic coordinates, where each atom is represented as a discrete particle in space. One might wonder how an arrangement of such particles in a 3-dimensional framework can emulate the behavior of complex biomolecules like proteins or small ligands. This is accomplished through the application of molecular mechanics principles, which utilize force-field parameters to define interaction potentials governing the behavior of the atoms. Force-field parameters are sets of fundamental constants and functions derived from quantum mechanical calculations, molecular dynamics, and empirical data, allowing for the accurate representation of interatomic forces. These parameters include bonded interactions (such as bond stretching, angle bending, and torsional terms) and non-bonded interactions (such as van der Waals forces and electrostatic interactions). By optimizing these parameters with the help of experimental data and quantum mechanical properties, the simulations can reproduce macroscopic observables and dynamical behaviors of the molecular system under study. Various force fields have been developed for use in molecular simulations, tailored for both explicit and implicit solvent models. Explicit solvent force fields, which simulate the environment around the biomolecule by representing individual solvent molecules, include widely-used parameter sets such as AMBER99SB-ILDN*, AMBER99SB-*disp*, CHARMM36m. In this study, we will employ the AMBER99SB-*disp* force field, as it has been shown to accurately reproduce experimental ensembles and conformational properties of intrinsically disordered proteins (IDPs). The intrinsically disordered protein (IDP) α-synuclein, which is implicated in the pathogenesis of Parkinson's disease, will serve as the model system for this study. Specifically, we will focus on the C-terminal region of α-synuclein, consisting of 20 amino acid residues, which will be subjected to molecular dynamics simulations. This region is of particular interest due to its role in interaction with small molecules such as Fasudil. The use of a refined force field like AMBER99SB-*disp* will enable us to explore the conformational landscape and dynamic interaction behavior of α-synuclein with high fidelity, providing insights into its structural properties and interactions that contribute to develope more potent small molecules.

Simulation software and protocols.

For all molecular dynamics (MD) simulations, we employ GROMACS 2023.5, integrated with PLUMED 2.8.0 to enhance the simulation capabilities. GROMACS (Groningen Machine for Chemical Simulations) is a highly versatile and widely-used molecular dynamics simulation package that provides the computational tools necessary for simulating a broad spectrum of molecular systems, ranging from small organic compounds to large and

complex biomolecular assemblies such as proteins, nucleic acids, and lipid bilayers. Its high efficiency and optimized algorithms allow for large-scale simulations with detailed atomic resolution, making it suitable for studying the dynamic behavior of biomolecules over extended timescales.
To further enrich the simulation protocols, PLUMED 2.8.0 is employed as a plugin for GROMACS. PLUMED is a sophisticated software package that expands the functionalities of traditional MD simulations by providing advanced sampling techniques and collective variable (CV) analysis tools. These enhanced sampling methods include metadynamics, umbrella sampling, and other free energy calculation techniques, which allow for an efficient exploration of the free energy landscape and conformational space of biomolecular systems. The use of PLUMED enables the investigation of rare events and slow conformational changes that are often inaccessible through conventional MD simulations.
For trajectory visualization and molecular modeling, software tools such as Visual Molecular Dynamics (VMD) and PyMOL are employed. These programs facilitate the analysis of MD trajectories by allowing the user to visually inspect the time-dependent behavior of the system, including changes in secondary and tertiary structures, interactions between molecules, and solvent effects. VMD, in particular, offers extensive functionalities for trajectory analysis, such as calculating root mean square deviation (RMSD), root mean square fluctuations (RMSF), and hydrogen bonding patterns, while PyMOL provides high-quality molecular graphics for generating publication-ready images.
In addition to visualization, quantitative trajectory analysis is conducted using MDTraj, a Python-based library designed for analyzing molecular dynamics trajectories. MDTraj provides robust methods for calculating various structural properties, such as distances, angles, dihedral angles, contact maps, and clustering of conformational states. These analyses enable the extraction of meaningful data from the simulation trajectories, aiding in the interpretation of structural dynamics and the evaluation of molecular interactions.
To ensure the reliability and reproducibility of the simulation results, the convergence of the molecular dynamics simulations is assessed through statistical methods such as block averaging. The Pyblock Python library is utilized for this purpose, implementing block averaging techniques to estimate statistical uncertainties and evaluate the convergence behavior of calculated observables. This method helps confirm that the simulation has adequately sampled the relevant conformational space and that the results are statistically significant.

1.4 Methods

Implementing REST(2 or 3) using the GROMACS[20] molecular dynamics engine requires installation of PLUMED2 plugin version$\geq$ 2.8 and patching of GROMACS source code before compiling mdrun. Please note that PLUMED2 can be used alongside the AMBER MD engine versionge18 and NAMD version$\geq$ 2.12 at a significant cost to performance. For a simple walkthrough of this installation procedure for GROMACS please refer to Note **??**. The first step in the ssREST3 implementation is to generate the a processed topology file your solvated protein of interest using *-pp* option in *gmx grompp*, note any position restraints contanted in an mdp file will be contained in the processed topology file.

```
gmx grompp -f *.mdp -c *.gro -r *.gro -p topol.top -o
    *.tpr -pp
```

This topology files will be used as to implement REST2 scaling procedure using PLUMED2.8 where one will supply the processed topology file and the corresponding λ value as such :

```
plumed partial_tempering   λn < processed.top >
    scaled.top
```

After the topology files are generated for the respective replicates, the oxygen atom of the solvent, for example OW_{tip4pd} which is the water oxygen atom of amber99sb-*disp* force field, will have ϵ scaled by a factor of κ_n^2 to satisfy the scaling condition, Eqs. **??** and **??**. If κ_n is set to 1.0 across all replicas the topologies conform to the REST2 convention. Along with the scaling of the water oxygen, we recover the interactions between the solvent molecules, and between solvent and ions, thus targeting only the protein-water LJ potential. This is accomplished by adding three non-bonded overrides to the *[nonbonded]* section of the GROMACS topology file as shown in Table **??**. Once these changes are made to the topology, we are ready to simulate ssREST3. Using the below command all the replicates are run in parallel and the the conformational exchange between the replicas is set for every 800 steps which equates to 1.6 ps.

```
gmx  grompp -f *.mdp -c *.gro -r *.gro -p scaled.top
    -o scaled.tpr
gmx mdrun -s scaled.tpr -multi <replica folders> -
    replex 800 -deffnm replica -plumed plumed.dat
```

At the start of the simulations it is best practice to check the acceptance ratios between replicas to ensure that the scaling is not too aggressive nor too weak. A minimum of 20% acceptance ratio is recommended for the simulations to be considered valid.

To observe if the solvent scaling is preventing the collapse in the conformations of an IDP we use α-synuclein in absence and presence of a ligand "Fasudil".

1.5 Notes

1. Basic installation of plumed and gromacs with openmpi preinstalled on a Linux system with the gcc compiler in a bash environment. This series of commands assumes you have the variable $MPICXX set. Further gains in performance can be reached by utilizing multiple NVIDIA GPUs and the CUDA compiler, nvcc, and the cude runtime. Detailed descriptions on compiling GROMACS and PLUMED2 can be found at http://gromacs.org and http:/plumed.org, respectively.

```
#!/bin/bash
INSTALL_ROOT=$HOME/opt
mkdir src
cd src
git clone https://github.com/plumed/plumed2.git
git clone https://github.com/gromacs/gromacs.git
cd gromacs
GMX_version=2024.3
git checkout -b v$GMX_version$
cd ../plumed2
./configure --prefix=$HOME/opt --enable-modules=all
    CXX="$MPICXX" CXXFLAGS="-O3 -axSSE2,AVX"
make
make check
make install
echo 'export PLUMED_KERNEL="/usr/share/lib/
    libplumedKernel.so"' >> ~/.bashrc
echo 'export PATH=$HOME/opt/bin:$PATH' >> ~/.bashrc
echo 'export PLUMED_ROOT=$HOME/opt'
source ~/.bashrc
cd ../gromacs
plumed patch -e gromacs-${GMX_version}
mkdir build
cd build
cmake .. -DGMX_THREAD_MPI=OFF -DGMX_MPI=ON -
    DGMX_BUILD_OWN_FFTW=ON -DREGRESSIONTEST_DOWNLOAD=
    ON -DCMAKE_INSTALL_PREFIX=$HOME/opt
make
make check
make install
```

Table 1.1: Table showing the differences and similarities of ϵ scaling between the REST2 and ssREST3 methods. In case of ssREST3 the water ϵ gets scaled along with the solute ϵ by a factor if κ^2 where as solvent parameters are not scaled during REST2.

Method	$T_{max}(K)$	λ	ϵ_{CA}	κ	ϵ_{OW}
–	300	1.0	0.359824	1.0	0.998989
REST2	450	0.666667	0.239883	1.0	0.998989
ssREST3	450	0.666667	0.239883	1.1	1.20878

Table 1.2: Table with $(\sigma_{ij}, \epsilon_{ij})$ showing the solvent interactions with itself and ions remain unaffected by the scaling factor κ_n.

Atom types	OW_{tip4pd}	$NA^+_{C22^*}$	$CL^-_{C22^*}$
OW_{tip4pd}	(0.3165, 9.98989)	(0.279746, 0.442754)	(0.360484, 0.791812)

1.6 Acknowledgements

I thannk blah blah for the blah blah to the blah blah, edit this for sure.

References

[1] Robert B. Best and Jeetain Mittal. "Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse". In: *The Journal of Physical Chemistry B* 114 (46 Nov. 2010), pp. 14916–14923.

[2] B. R. Brooks et al. "CHARMM: The biomolecular simulation program". In: *Journal of Computational Chemistry* 30 (10 July 2009), pp. 1545–1614.

[3] Bernard R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: *Journal of Computational Chemistry* 4 (2 Jan. 1983), pp. 187–217.

[4] Wendy D Cornell et al. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules". In: *Journal of the American Chemical Society* 117 (19 May 1995), pp. 5179–5197.

[5] Andreas W. Götz et al. "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born". In: *Journal of Chemical Theory and Computation* 8 (5 May 2012), pp. 1542–1555.

[6] Jing Huang et al. "CHARMM36m: An improved force field for folded and intrinsically disordered proteins". In: *Nature Methods* 14 (1 2016), pp. 71–73.

[7] Sofie Jakobsen, Tristan Bereau, and Markus Meuwly. "Multipolar Force Fields and Their Effects on Solvent Dynamics around Simple Solutes". In: *The Journal of Physical Chemistry B* 119 (7 Feb. 2015), pp. 3034–3045.

[8] Kuo Hao Lee and Jianhan Chen. "Multiscale enhanced sampling of intrinsically disordered protein conformations". In: *Journal of Computational Chemistry* 37 (6 Mar. 2016), pp. 550–557.

[9] Kresten Lindorff-Larsen et al. "Improved side-chain torsion potentials for the Amber ff99SB protein force field". In: *Proteins: Structure, Function, and Bioinformatics* 78 (8 June 2010), NA–NA.

[10] Pu Liu et al. "Replica exchange with solute tempering: A method for sampling biological systems in explicit water". In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (39 Sept. 2005), pp. 13749–13754.

[11] Stefano Piana et al. "Development of a Force Field for the Simulation of Single-Chain Proteins and Protein–Protein Complexes". In: *Journal of Chemical Theory and Computation* 16 (4 Apr. 2020), pp. 2494–2507.

[12] Elizabeth A. Ploetz et al. "Kirkwood-Buff-Derived Force Field for Peptides and Proteins: Philosophy and Development of KBFF20". In: *Journal of Chemical Theory and Computation* 17 (5 May 2021), pp. 2964–2990.

[13] Arushi Prakash et al. "Biasing Smarter, Not Harder, by Partitioning Collective Variables into Families in Parallel Bias Metadynamics". In: *Journal of Chemical Theory and Computation* 14 (10 Oct. 2018), pp. 4985–4990.

[14] Ruxi Qi et al. "Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example". In: vol. 1777. 2018, pp. 101–119.

[15] Dhiman Ray and Michele Parrinello. "Kinetics from Metadynamics: Principles, Applications, and Outlook". In: *Journal of Chemical Theory and Computation* 19 (17 Sept. 2023), pp. 5649–5670.

[16] Paul Robustelli, Stefano Piana, and David E. Shaw. "Developing a molecular dynamics force field for both folded and disordered protein states". In: *Proceedings of the National Academy of Sciences of the United States of America* 115 (21 May 2018), E4758–E4766.

[17] Romelia Salomon-Ferrer et al. "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald". In: *Journal of Chemical Theory and Computation* 9 (9 Sept. 2013), pp. 3878–3888.

[18] David E. Shaw et al. "Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer". In: *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*. Vol. 2015-January. IEEE, Nov. 2014, pp. 41–53.

[19] David E. Shaw et al. "Millisecond-scale molecular dynamics simulations on Anton". In: *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. ACM, Nov. 2009, pp. 1–11.

[20] David Van Der Spoel et al. "GROMACS: fast, flexible, and free." In: *Journal of computational chemistry* 26 (16 Dec. 2005), pp. 1701–18.

[21] Yuji Sugita and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314 (1-2 Nov. 1999), pp. 141–151.

[22] Andreas Vitalis and Rohit V. Pappu. "Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules". In: vol. 5. Elsevier, Jan. 2009, pp. 49–76.

[23] Lingle Wang, Richard A. Friesner, and B. J. Berne. "Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)". In: *The Journal of Physical Chemistry B* 115 (30 Aug. 2011), pp. 9431–9438.

[24] Paul K. Weiner and Peter A. Kollman. "AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions". In: *Journal of Computational Chemistry* 2.3 (1981), pp. 287–303.

[25] Yumeng Zhang, Xiaorong Liu, and Jianhan Chen. "Re-Balancing Replica Exchange with Solute Tempering for Sampling Dynamic Protein Conformations". In: *Journal of Chemical Theory and Computation* 19 (5 Mar. 2023), pp. 1602–1614.