Best practices for biomolecular simulation setup : v0.0.1

Rommie E. Amaro[1*], John D. Chodera[2*], David L. Mobley[3*], Antonia Mey[4*], Julien Michel[4*]

***For correspondence:**
email1@example.com (REA);
john.chodera@choderalab.org (JDC); dmobley@mobleylab.org (DLM); email1@example.com (AM);
email1@example.com (JM)

[1]Institution 1; [2]Institution 2; [3]Departments of Pharmaceutical Sciences and Chemistry, University of California, Irvine; [4]Institution 4

Abstract Simulations of biomolecular systems are complex and often demanding but can provide great insights and valuable quantitative predictions. However, these simulations require considerable human expertise to set up well, and key aspects of system preparation and setup are often overlooked for a variety of reasons. Thus, simulations of the "same" biomolecular system prepared by different individuals or different software tools may yield substantially different results due to different choices made in system preparation, such as different handling of crystallographic waters, different treatment of missing residues or loops in a protein, or different choices for a ligand or receptor's protonation state. In some cases, human error can also play a role, with a user perhaps forgetting to check for disulfide bonds. Here, we identify the issues we believe are most important for users and/or software tools to consider when preparing biomolecular systems for simulation and present a checklist for users to follow to ensure they have handled the most crucial or troublesome aspects of system preparation, with recommendations for how to handle each step in the process when best practices have been established by the literature. We also present a detailed explanation of the items in our checklist, including why we make the recommendations we do. Additionally, because a useful checklist must necessarily be brief, we provide a discussion of other important factors to consider which did not make our major checklist. We believe this work provides a starting point to help reduce the number of simulations run which are flawed due to setup failures, and we hope the community will contribute to updating and improving this document to that end.

Goals

Produce a brief checklist of most critical and most overlooked items that practitioners can follow and potentially reviewers can use. Highlight issues where further study (or references we don?t know about yet) are needed to clarify what should be done; solicit input from the community on how best to handle issues. Guideline on how to report the checklist addressed items in a publication: Note that typical methods sections should provide the information in this checklist. (How to write a methods section by John Chodera: http://klogw.org/2016/07/28/how-to-write-a-methods-section/).

We are not trying to enforce; we are trying to inform.

Scope

Relatively simple biomolecular simulations of soluble proteins that may include small molecule ligands: May not be able to include nucleic acid recommendations initially because of lack of expertise; Toni can potentially cover lipids, but it may be best to have a separate membrane protein doc that follows this. Eventually to include cofactors as part of the consideration

Does NOT include: Generation of parameters, such as for nonstandard nonstandard residues (as opposed to standard nonstandard residues where literature and/or library parameters may already be available), or for complicated lipids or cofactors? AND these will be coupled to force field choice to some extent. Simulation protocol such as minimization, cutoffs, etc.

Checklist

Step 0: Know what you want to simulate

Protein

1. Sequence of the protein – is it the protein you wanted to simulate? (Assays, crystal structure, intended) And what conditions do you want to simulate (are they the structural conditions)? Oligomeric state (how many chains do we need to simulate?) (Chodera/Amaro)
2. Structure selection (Mey)
    - X-ray structural data
    - Dealing with NMR structural ensembles
    - Other sources of structural data: e.g. cryo-EM
3. Disulfide bonds depending on reducing/oxidizing environments (Amaro)
4. Post-translational modifications (Chodera)
5. Model in missing residues and loops (Chodera/Amaro tag-team)
    - Schrodinger tools model in short loops, but truncate long loops with proton caps (?)
    - UCSF Modeller terrible for loops without care; Rosetta Model works well ? (JDC says)
    - (But Amaro reports success stories with Schrodinger tools if one knows when to use them and does not try to push the tool beyond the limit of reasonable application - eg loops < 10-12 residues, otherwise need templated homology models)
6. Protonation states and tautomers: (Mobley to draft; maybe eventually rope in someone to comment on constant pH simulations – Jordi Juarez in Michel group has expertise here)
    - ProPKA - but see Beckstein ProPkaTraj (https://github.com/Becksteinlab/propkatraj)
    - MCCE
    - Early stages of constant pH (e.g. Case, McCammon, Roitberg, Shen, Roux, Chodera)
7. Crystal waters (Amaro, Mey)
8. Metal ions (JDC)

Ligands

1. Select protonation state/tautomer (Chodera)
2. Select the correct ligand binding mode (Greg Warren, Bob Tolbert, Mobley)
    - Is this the right ligand?
    - Does the density support?
    - Is that only because of crystal contacts?
    - How do you put your new ligand in? (Get help from Amaro)
    - ...docking, shape overlay, etc
    - Run dynamics of ligands?

Counterions/water

1. No counterions, minimal counterions, or physiological ionic strength? (Mobley)

Other things to think about that didn't make the checklist

Here, we focus on items which are also important part of system preparation, but which are less frequent causes of critical failures or are not as often or easily overlooked.

Proteins

Termini: Build them, or cap? Depends on length; if too long can contribute to timescales. Write some guidelines into how to know.

Ligands

Find/create parameters (Mobley) - mention but not totally in scope.

Handle covalently bound cofactors/adducts: Not in our scope but mention that it is a research topic which uses special treatment; refer to how. (As distinct from ones which have parameters available).

Counterions/water

(minor) Select simulation box size (Mey) – important, but first thing to check in analysis.

The selection of models not in our scope but will mention (Mobley), including: Water model; Ion model.

(minor) Add solvent then ions, or ions then solvent? Pre-equilibrate ion-water mixture? (Amaro) Determine electrostatic potential around molecule and place ions at minima (default AMBER approach does this in solvent, replacing some solvent with ions)

Detailed explanation of checklist items

Protein

Ligands

Counterions/water

Acknowledgments