Best practices for preparing protein structures for simulation

Dan Arlow

06/12/2012

We run MD simulations to try to determine the dynamical behavior of a chemical system. In order for your simulations to be meaningful interrogations of natural phenomena, you must take care to ensure that your chemical system is a faithful representation of the natural system, at least as much as possible within the constraints of our simulation technology. There are many ways in which mistakes or bad choices during the process of preparing a protein for simulation can lead to less useful simulation results. Here are the steps I typically take when preparing a chemical system to try to produce maximally realistic simulations:

1. Run the PDB file through PROPKA and manually examine each aspartate, glutamate, and histidine to decide on protonation states. Solvent-accessible aspartate and glutamate residues will inevitably be charged; completely buried aspartate and glutamate residues should most likely be protonated. Most histidine residues should be singly-protonated on the epsilon nitrogen (i.e. neutral) if the simulation corresponds to pH 7.0. (It is rare that there is clear evidence that a neutral histidine should instead be protonated on the delta nitrogen; in most cases you should just go with the epsilon protonation because it is slightly more energetically favorable than the delta protonation.) Some histidines may turn out to be doubly protonated (i.e. charged) – evidence for a charged histidine can come from biophysical measurements such as NMR spectra, or from the crystal structure itself, e.g. the histidine residue is unambiguously donating two hydrogen bonds or is making a strong interaction with a negatively-charged residue.
2. Run the PDB file through DOWSER (installed in the garden) to find probable internal structural water molecules.
3. Load the PDB file into Maestro.
4. Download the relevant protein sequence from [UniProt](http://www.uniprot.org/) and look for any discrepancies with the sequence annotated in the PDB file. Identify any loops not resolved in the crystal structure. An easy way to do this is to write out a FASTA file of the sequence from the loaded PDB (load the Prime panel and click “load sequence from workspace” and then File > save sequence as...) and then load the two sequences into [ClustalW](http://www.ebi.ac.uk/Tools/msa/clustalw2/) to make an alignment. Also, look for any post-translational modifications to annotated in UniProt and decide about whether to include them in your model.
5. Decide about modeling in missing loops: missing loops shorter than 10 residues should almost always be modeled back into the structure; missing loops longer than 10 residues can be difficult to model correctly and bad modeling choices can potentially bias simulation results. Loops not modeled into crystal structures are assumed to be “disordered” i.e. mobile within the crystal lattice, either completely flopping around or possibly occupying a small number of discrete conformations. If you decide to model in a longer disordered loop, you should take care to model it in an extended conformation that both does not place any strain on the part of the protein model resolved in the crystal structure.
6. Build a complete model of the protein including hydrogens. If you don’t have to build in any loops you can do this via the protein prep wizard; otherwise you’ll have to use Prime > Structure Prediction. In Prime make sure to “retain rotamers of conserved residues” so that Prime doesn’t take any liberties with the crystal structure coordinates and only builds in missing residues. Modeled in loops should subsequently be optimized with the Prime > Refine Loops tool. Again, make sure that prime only modifies precisely the modeled in residues and leaves the resolved residues alone, unless you have good reason to think that there’s something wrong with the crystal structure.
7. Always cap protein termini with neutral groups in the prepwizard (acetyl and methylamide) unless you are simulating a full-length sequence and you know the termini are charged.
8. Check to make sure that Maestro correctly recognized all of the disulfides in the protein.
9. Maestro by default protonates histidines on the delta nitrogen, but epsilon-protonated histidine is marginally more stable than delta-protonated histidine. You can mutate all your histidines back to epsilon-protonated in one step by selecting them using the selection GUI (or typing “workspaceselectionreplace res.ptype HIS” in the command input area) and right-clicking on one of the “H”s in the sequence viewer and selecting “mutate residue > HIE”.
10. Adjust any other protonation states (from step 1) as necessary, either by right-clicking the residue in the sequence viewer and selecting “mutate residue” or by using the interactive protassign wizard (available from the prepwizard window)
11. Make sure any ligands in the protein have the right protonation state(s) and topologies.
12. Load the DOWSER waters PDB file and merge it with your protein’s entry. Delete DOWSER waters that overlap with crystallographic waters, if necessary.
13. Optional: use MacroModel to minimize the protein hydrogens and water positions (i.e. minimize with constrained atoms “not atom.ele H and not res.ptype HOH”)
14. For soluble proteins: use the System Builder GUI in Maestro to build a solvated system for your protein, and you’re ready to go!
15. For GPCRs: align the protein so that it is properly matched to the membrane. Our convention has been that the center of the membrane is at (0,0,0) and that the extracellular side is in the +Z direction. (Polarity is important in case the membrane is asymmetric, e.g. if there is POPS in the intracellular leaflet.) A good way to do this is to download the PDB’s corresponding model from the [OPM database](http://opm.phar.umich.edu/) and using Maestro’s superimpose tool on some short snippet of sequence (if your GPCR has an entry in the OPM database) or using the Protein Structure Alignment tool if you are aligning your GPCR to a different GPCR from OPM.
16. Write out a Maestro file of your final (oriented) prepared structure for use in system building using saute3.py or its successor.
17. When constructing a solvated protein system with a lipid bilayer, it is important to include enough lipid bilayer so that the protein doesn’t “feel” the effects of its periodic images; I recommend building a large enough bilayer patch that no atom in the transmembrane region of the protein can get within 35 Å of any periodic image of this set of atoms, allowing for inevitable rotation of the protein around the membrane-normal axis. This separation allows for approximately 6 lipid diameters between periodic images and is designed to be longer than the lateral correlation length of a POPC bilayer. (As such, 35 Å is the default for -m in saute3.py.)
18. Furthermore, for any solvated system, I recommend building in enough water such that no atoms in the entire protein ever are able to get closer than 20 Å to atoms of the protein in another periodic image, even if the protein rotates as a rigid body. Rotation in membrane proteins is naturally restricted to be around the membrane normal axis but soluble proteins can tumble freely. This separation of 20 Å is designed to be long enough to allow for the decay of much of the electric field of a single charge in the dielectric of water even without NaCl. That way, surface charges on the protein have negligible electrostatic interactions with surface charges on periodic images of the protein.
19. Typically, sodium or chloride atoms are added by the system building tool to neutralize the net charge of the protein, and subsequently additional sodium and chloride atoms are added until the concentration is 150 mM, matching physiological NaCl concentration. Take care to at least consider what the buffer conditions of the biochemical experiment you’re trying to simulate.
20. The chemical system you have constructed will inevitably contain packing defects because all system building tools we use simply superimpose the protein with an equilibrated solvent/bilayer system and delete clashing solvent and lipid molecules. As such, NPT-ensemble “equilibration” is necessary before production simulation. (Note that for membrane systems, semi-isotropic barostatic coupling \*must\* be used to allow the aspect ratio of the box to equilibrate; for soluble systems, isotropic couple \*must\* be used.) We’ve conventionally used a protocol of applying 5 kcal mol-1 Å-2 position restraints to the non-hydrogen atoms of the protein (alternatively, only restraining Cα atoms is probably okay,) and tapering them off linearly over 50 ns, which simultaneously allows the box dimensions to equilibrate and the protein to relax. The initial stiffness of the restraints will result in the protein barely moving by the time that a typical chemical system has equilibrated its box dimensions; subsequently, the protein will begin to move. However, in the case of a membrane protein, position restraints will not allow the protein to tilt relative to the membrane normal direction, so when the protein is fully released from the restraints it may still not be properly matched to the hydrophobic thickness of the bilayer. In the future, I’d recommend exploring the possibility of using RMSD restraints on membrane proteins for equilibration instead of position restrains, because RMSD restrains will allow the protein to tilt while preserving its conformation.

Once you are finally running production simulation you must remain vigilant for artifacts that arise due to errors during system preparation such as incorrect protonation states, missing chemical groups, poor modeling of unresolved regions, or insufficient box dimensions. Localized unfolding of a region of the protein that is resolved in the crystal structure (and is not making crystallographic contacts in the structure) is most often a sign of an error during system preparation. After a large conformational change or a partial unfolding of the protein, make sure that the protein atoms are still sufficiently separated from periodic images by solvent (and bilayer if applicable) so as to not make strong periodic interactions.

Good luck, and may your simulations be realistic!