MD SIMULATIONS: HOMEWORK QUESTIONS

Answers:

1. Describe and name interaction potentials needed to simulate (at atomistic level):
   - ideal gas
     - *no potentials except excluded volume which is modeled in most of all-atom MD simulation force-fields by Lennard-Jones non-bonded potential*
   - liquid of charged/uncharged spherical particles with excluded volume
     - *as above plus Coulomb electrostatic potential if particles are charged; no bonded terms as these are spherical particles*
   - biomolecular systems like solvated DNA/protein
     - *full all-atom Hamiltonian terms covered in Lecture 1; also slide 6 of Lecture 2. Particularly, these include bonded potentials describing biomolecule/solvent connectivity and flexibility (bond, angular, dihedral angle terms), and non-bonded interactions modeling short- and long-range vdW and electrostatic interactions, respectively (LJ and Coulomb).*

2. Name the reasons (as many as you can) why .pdb structures from Protein Data Bank may be bad inputs for MD simulations? Consider X-ray crystal and solution NMR structures as examples. What approaches are typically used to make those structures usable for MD simulations ?
     - *Crystal structures often suffer from crystal packing effects leading to significant deviations from what is observed in a "relaxed" solvated phase; as to the .pdb file format, it is often non-readable as is to the most of MD simulation packages; .pdb files must be "massaged" to strip unnecessary experimental information and to make correspondence to the topology file(s) of the used force-field; finally, there may be lots missing unresolved atoms/flexible parts and those must be added to the structure to be simulated; there are straightforward automated methods to do this.*
     - *NMR structures obtained from solution state experiments, while not suffering from packed environment, heavily rely on optimization strategy which usually results in the degeneracy of obtained solutions (more than one structure for same biomolecule); some of the structures from a set of structures may be quite far from, while others may be close to the reality. Other techniques (like SAXS) may be needed to further score conflicting NMR structures.*
     - *For small-to-medium size biomolecules (~<30-50 kDa) series of short minimization and equilibration steps should be enough to relax experimental structure for it to be usable for productive MD simulations.*

3. What are periodic boundary conditions (PBC) and why they are commonly used in all-atom MD simulations? When analyzing MD simulation trajectory produced with the use of PBC, what must be necessarily taken into account ?

- *PBC address a number of important issues in MD simulations: 1. PBC allow to emulate large statistical system (~Avogadro number, 10^23 particles) by simulating feasible ~10^6 atoms in a periodic unit cell; 2. PBC effectively solves surface effects problem by simply eliminating the surface of the simulated system.*
- *There are different PBCs which can be used, but most commonly used are cubic/rectangular PBCs for its obvious advantages in treating long-range Coulomb interactions (exceeding the size of primary simulation box), and the ease of post-processing generated trajectory. The latter includes, in particular, proper handling of inter-particle separations due to particle's re-entering of the simulation box in the course on simulation.*

4. What statistical ensembles can be simulated with MD ? Which of them naturally follows from the first principles of classical MD ? Does it matter what simulation ensemble you choose and why ?

    - *In general, all statistical ensembles can be simulated in MD such as micro-canonical (NVE), and different canonical (NVT, NPT) ensembles. Though beyond the scope of the set of lectures – for you information – even more exotic ensembles such as μVT (constant chemical potential), constant pH simulations, or grand canonical ensembles (varying # particles) can be simulated as well.*
    - *Micro-canonical (NVE) ensemble follows naturally from Newtonian dynamics.*
    - *Different ensembles correspond to different set of fixed thermodynamic parameters like V, P, T, E. Direct (or closer) correspondence to experiment may require simulation certain ensemble.*

5. What is ergodic hypothesis and what is the main assumption behind it? Why it is of primary importance in MD simulations ? Provide examples of systems for which ergodic hypothesis is not valid; how useful may be MD simulations of such systems?

    - *Ergodic hypothesis is the central assumption which basically tells that ensemble average in statistical mechanics (where we do not have time) is equivalent to the time average in MD simulations (with no explicit microstate's weighing – be it Boltzmann for canonical ensemble, or equiprobability assumption in the micro-canonical ensemble); This immediately implies that the force-field used to generate MD trajectory is "correct", i.e. it generates a set of microstates (snapshots) corresponding to sampling different regions of the system's phase space with correct probabilities.*
    - *If the phase space of the system is not continuous, i.e. it has inaccessible regions for MD simulations (even long ones), ergodic hypothesis is in principle not valid. Examples of such systems are glasses and IDPs. For IDPs, for example, a biomolecular force-filed developed for "normal" proteins (exhibiting a single major free energy minimum) may capture only (small) part of the overall picture, usually a "folded" sub-state of the IDP. So, such MD simulations may be still useful for limited IDP studies. Of course, no kinetics or transitions between sub-states can be reasonably addressed.*

6. Why free energy, rather than potential energy (interaction potentials), governs overall behavior of the system simulated with MD? Can free energy be measured in MD simulations and why?

- *Because MD is a "practical analogue" of the statistical mechanics (operating with ensembles of microstates), free energy – a negative log of the number of microstates – is of central importance. Mathematically, requirement of the minimum of the free energy (at equilibrium) translates in the simultaneous requirements of the minimum internal energy and maximum of the entropy. These two conditions may compete with each other, so higher energy states with larger entropy may be more preferable than lower energy states with lower entropy. This of course can not be captured by considering interaction energies alone.*
- *Absolute free energy – log of all accessible microstates – can not be measured in MD simulations for obvious reason that not a single MD simulation can cover all microstates; Even in infinitely long simulations, the very fact that models used in MD are approximations to the nature implies absolute free energies can not be measured in principle.*

7. Why Free Energy Perturbation (FEP) method can often NOT be straightforwardly used to compute even small free energy differences? Provide your thoughts.

   - *While mathematically exact, FEP requires that target states to be effectively sampled in the unperturbed, reference state. This is of course not always the case, whether free energy differences between states are large or small. In general, it is hard to tell how "similar" reference and target states of particular system are; therefore, a safer approach is to perform FEP calculations in small steps, where reference and target states at each iterations are "close".*

8. What is Thermodynamic Integration method and what processes can be effectively studied with it ?

   - *TI is another free energy difference calculation method; it deals with defining thermodynamic cycle representing transition between states in terms of a number of alternative real or alchemical transformations; it is assumed that these alternative transformations are easier to compute, potentially with higher accuracy. Because these transformations are reversible, backward and forward calculations are often compared to address ergodicity of the system and the accuracy of free energy difference estimation.*
   - *Relative (and also absolute) free energies of solvation and binding are typical examples of TI applications.*

9. What is Umbrella Sampling and why it is often used in conjunction with Potential of Mean Force (PMF) calculations ?

   - *PMF is obtained by redefining partition function in terms of the order parameter, or reaction coordinate which best describes distinct states of interest. The very term "potential of mean force" follows from the fact that systems behavior is mapped on the order parameter space with all other effects to be effectively "integrated out".*
   - *If the path along reaction coordinate lies through significant free energy barriers, PMF can not be effectively calculated from MD simulations as the target (and intermediate) states are unreachable. Different techniques have been developed to alleviate this problem; umbrella sampling is a collective name for many of those techniques; The main idea behind umbrella sampling is that additional biasing potential defined in terms of the used order*

parameter is applied to a system to sample low probability states along the path connecting reference and target states; The unbiased (real) PMF is then obtained from such biased MD simulations by proper re-weighing contributions from different umbrella windows. This is typically done with WHAM (Weighed Histogram Analysis Method).

10. When modeling biological systems at coarse-grain (CG) resolutions, what is the main motivation for choosing certain resolution ? What strategy is typically used to develop accurate CG models (interaction parameters governing MD simulations of such systems) ? What criteria used to determine maximum simulation time step ?

    - *Resolution of CG models is motivated by the nature of the phenomena to be studied. For example, lower resolutions are used to study large-scale phenomena provided finer details are not important. CG models for MD simulations are based on the same principals as all-atom MD (Newtonian dynamics, different ensembles etc.); effective beads or particles represent a group or groups of atoms in all-atom model.*
    - *A typical approach to develop interaction parameters for CG models is a "bottom-up" approach when effective potentials between group of atoms are derived from smaller but fine-detailed all-atom models; these potentials are then tested in larger CG models and optimized as needed to reproduce properties of the underlying accurate atomistic resolution model(s).*
    - *Time step in CG simulations may be much larger that in all-atom models; a reasonable (and necessary) criteria for choosing time step for CG simulations is the conservation of energy (and other thermodynamic parameters defining simulating ensemble).*

11. What is the minimum input for MD simulation ? (list the files and mention information provided by these files)

    - *.pdb file – coordinates for initial structure; topology and parameter files – files from the used biomolecular force field package providing information on system connectivity and parameters for different types of interaction potentials; .psf – a concise file generated from topology and parameter files providing concise information on the system simulated.*