1. Molecular Dynamics Simulations

Martin Almlöf, Martin Andér, Sinisa Bjelic, Jens Carlsson,
Hugo Gutiérrez de Terán, Martin Nervall, Fredrik Österberg and Stefan Trobro

18th August 2005

Contents

1 Molecular Dynamics: theory

Molecular Dynamics (MD) can be defined as a technique that provides a bridge between the average *positions* of atoms in static structures (xray or NMR structure, or homology model) and the measurable *thermodynamic properties* of biological interest.
The interactions of all atoms in our system are described by classical mechanics, through the use of a *Force Field*. Under the scheme of a force field, the *potential energy* of an atom (*U*) is obtained from the sum of all its bonded and non bonded interactions.

- The bonded interactions are applicable to bond and angle vibration and bond (torsion) rotation, plus an extra term for the so called improper torsions. This total bonded energy can be modeled as harmonic potentials (for bonds, angles and impropers) and periodic torsion angle potentials, according to equation 1:

$$U_{\text{bonded}} = \sum_{\text{bonds}} K_r(r-r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta-\theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1+\cos(n\phi-\gamma)] + \sum_{\text{impropers}} K_\xi(\xi-\xi_{eq})^2 \quad (1)$$

  The values of all the constants and equilibrium values in eq 1 are calibrated against experiments or quantum mechanical calculations

- In addition to the bonded interactions, pair interactions between non-bonded atoms are included in the potential energy function as electrostatic and van der Waals interactions, according to eq 2.

$$U_{\text{non-bonded}} = \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right] + \sum_{i<j} \left[\frac{q_i q_j}{4\pi\epsilon_0 R_{ij}} \right] \quad (2)$$

  The van der Waals expression (first summation in eq 2) has a repulsive term ($\frac{A_{ij}}{R_{ij}^{12}}$) which models short range repulsions, while the second term ($\frac{B_{ij}}{R_{ij}^{6}}$) models induced dipole-dipole interactions; The electrostatic equation, corresponding to the second summation in eq 2, is the Coulomb interaction for two point charges. Partial charges are obtained mainly by quantum mechanical calculations sometimes adjusted to reproduce dipoles and hydrogen bonds.

It is important to note that the atoms that form bonds, angles and torsions are excluded from the non bonded interactions (torsions only partially) and that the number of pair interactions scale as N^2 while the bonded only scale to N.
An intuitive illustration of the mathematical and chemical meaning of each term of the force field equations is given in figure 1.

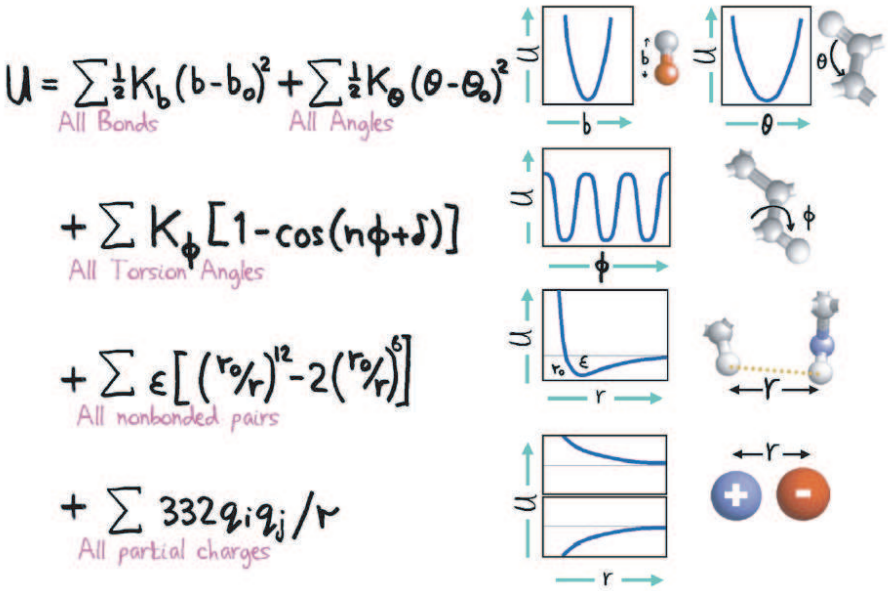

Figure 1: the meaning of each term in the force field equation

The force field allows the atoms to be treated as point masses with forces acting between them governed by the particle positions. This is where Newtonian classical mechanics comes in to play, (see 3 and 4).

$$\vec{F}_{\mathrm{i}} = -\frac{\partial U_{\mathrm{i}}}{\partial r_{\mathrm{i}}} \tag{3}$$

$$\vec{F}_{\mathrm{i}} = m\vec{a}_{\mathrm{i}} \tag{4}$$

The time evolution of the system can be calculated according to a numerical algorithm. One such integration is the leap-frog version of Verlet's algorithm:
For every atom i:

- calculate $\vec{F}_{\mathrm{i}}(t)$
- update velocity

$$\vec{v}_{\mathrm{i}}\left(t + \frac{\Delta t}{2}\right) = \vec{v}_{\mathrm{i}}\left(t - \frac{\Delta t}{2}\right) + \frac{\vec{F}_{\mathrm{i}}(t)}{m_{\mathrm{i}}} \cdot \Delta t$$

- update position

$$r_{\mathrm{i}}(t + \Delta t) = r_{\mathrm{i}}(t) + \vec{v}_{\mathrm{i}}\left(t + \frac{\Delta t}{2}\right) \cdot \Delta t$$

The goal of MD is to generate an ensemble of structures and energies that reproduces thermal equilibrium. These ensembles can then be used to calculate thermodynamic properties. A few points and recommendations are worth mentioning:

- Starting coordinates are taken from xray structures or homology models. We need to generate a *topology file*, which combines the information contained in the initial PDB file (initial positions of the atoms) and the information contained in the force field for each atom.

- The system will be modeled as spherical and centered on the chemical group of interest. Water molecules are added before the simulation to fill vacant positions and restraints are used to reproduce bulk water density and polarization near the system boundary. Atoms outside the system boundary are conformationally restrained to initial positions.

- Charges close to the boundary should be neutralized because of the inability to solvate them.

- To keep the temperature fixed around a certain value (**temperature** keyword) we must couple the MD simulation to a thermal bath (see **bath_coupling** keyword).

- Starting velocities are randomly assigned (see **random_seed** keyword) from a Maxwell-Boltzman distribution for relevant mass and temperature. The velocity distribution is used to determine the temperature of the system and all velocities are then scaled to the target temperature repeatedly during the simulation. This information is stored in the so called restart files (extension `.re`), that allow us to continue a calculation from the last step or to restart a job if a crash occurs.

- Only non-bonded interactions involving atoms *inside* the system boundary are calculated.

- To reduce the number of pair interactions, several approximations are introduced:

  - For every atom *i* there is a cutoff (keyword **cutoff**) distance for the treatment of its non bonded interactions: every possible pair *i,j* inside the cutoff is periodically tabulated according to a user defined interval of time steps.
  - Beyond the cutoff the electrostatic interactions are approximated through the local reaction field (keyword **lrf**) approximation, in which a fourth order series expansion of the electric field, E, due to all atoms outside the cutoff is calculated. The force acting on an atom *i* due to the electric field is obtained by:

$$\vec{F}_{\mathrm{i}} = \vec{E}_{\mathrm{i}} \cdot q_{\mathrm{i}}$$

  - All van der Waals forces outside the cutoff are ignored.

  It is important to note than in any free energy calculation the atoms which energy will be calculated (so called Q atoms in our programs, *i.e.* the ligand) do not have any of these approximations, and they explicitly see *every* other atom within the sphere of simulation. The complete list of those atoms must be specified in a separate file, *e.g.* **lie.fep**. Detailed information about the variations on this file will be given in the next practical.

- The leap-frog algorithm calculates the velocities and positions after a given *time step* jump (Δt, keyword **stepsize**). One decides the number of time steps, *n*, (keyword **steps**) to be performed, and the total time of the simulation will be:

$$t_{\text{total}} = n \cdot \Delta t$$

2 Methods: Prepare and run the MD simulations

In this practical you will see how to set up a MD simulation, in which the surrounding energies of a ligand, complexed with a protein, will be monitored.
Two parts are needed in order to run a MD simulation:

- **Topology**: A file must be created combining the positions of the atoms and the parameters of the forcefield for every atom. The resulting file contains information about atom properties, connectivity, amino acid sequence, solvation model, etc ...
- **Input file**: It specifies the values of any variable parameter in the MD scheme: temperature, time step, number of steps, cutoff for non-bonded interactions etc ...

We will prepare topology and input files for the system of study, cytochrome P450cam complexed with the inhibitor camphane. The P450 enzymes, which receive this name because they absorb light maximally at 450 nm when they are complexed with CO, catalyze the hydroxylation of unactivated alkanes. This chemical reaction is used in the liver to increase the solubility of foreign substances (xenobiotic compounds), facilitating their detoxification. P450 enzymes are also involved in the synthesis of steroids and fatty acids, and in the metabolic activation of powerful carcinogens. The particular enzyme known as P450cam catalyzes specifically the hydroxylation of camphor.
The system will be warmed up and equilibrated through a series of short MD runs. All

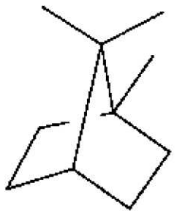

Figure 2: chemical structure of camphor

MD simulations will be done with the modified version of the forcefield OPLSAA, as implemented in the software `Q`. Visual analysis will be done with the molecular modeling program `PyMOL`.

2.1 Topology generation

- Login in to your personal account and go to the working directory:

- Start your PC, and **PRESS F12 BEFORE WINDOWS STARTS**.
- Log in to your personal account
- Open a UNIX terminal by right button clicking on the Desktop, and selecting "open terminal" from the pull down menu.
- Change to our working directory:
  `cd MD/p450cam`

- PDB preparation:
  You will find the original PDB file for the P450cam-CMA complex, under the name `6cpp.pdb`. However this file needed some editing before using it as input for MD with our program `Q`. Here is a list of the relevant steps that were done to generate the suitable file `complex.pdb`, which is also in your working directory.

  - You dont't need to go through these steps, but please take 5 minutes and try to understand these modifications! Open and compare the two pdb files `6cpp.pdb` and `complex.pdb`
  - Only the ATOM and HETATM lines have been retrieved (i.e., the coordinates of the atoms).
  - A "GAP" line is introduced between molecules, instead the default TER used by PDB convention.
  - Only one water molecule has been conserved, which is important for the stability of the heme group, as we will see later on. For this water molecule, the hydrogens were manually added so that the orientation is optimal for hydrogen bonding to the heme group.
  - The terminal oxygen of the last protein residue has been deleted, as it is not implemented in the `Q` residue libraries (it will not be part of our sphere of simulation)
  - Ionizable residues (aspartic, glutamic, lysine or arginine): in order to ensure a correct solvation of any charged group present in the system, only those ionizable residues located deep in the sphere of simulation will be considered as charged. The OPLS residue library file (`Qoplsaa.lib`) contains both the charged (ASP, GLU, LYS, ARG) and neutral (ASH, GLH, LYN, ARN) version for these residues, so for a given residue changing the default residue name to the charged version name on the pdb will solve the problem. After visual inspection, all ionizable residues within approximately 16 Å of the C_1 atom of camphane were set to their ionized state, except Asp297 and Arg240. Here is a list of the charged residues in the sphere:

    * ASP: 88, 173, 242
    * GLU: 357
    * ARG: 103, 177, 290
    * LYS 169, 188

- Note that the pdb has no hydrogens: they will be automaticaly added in the next step.

- Solvating the system and generating the topology:
  The edited pdb file, named `complex.pdb` can be now processed by the module `Qprep5` of `Q` in order to solvate the simulation sphere with TIP3P waters and write the topology file, which contains all the necessary force field parameters. For this step we need some files, stored in the directory `FF_Q/`:

  - A library file for each molecule in the complex, with relevant information about the *atom names*, *atom types*, *partial charges* and *bonds* present in the molecule. For this particular case you have the `Qoplsaa.lib` file, with information about all the protein residues plus the water (TIP3 water model), `heme.lib` and `cma.lib`, with information about the heme group and the ligand, respectively.
  - A parameter file, with all the molecular mechanics parameters needed for simulate this system: *Van der Waals, bond stretching, angle bending, torsion and improper angle* parameters. This file is called `Qoplsaa.prm`
  - You can take a look at the files `../FF_Q/Qoplsaa.lib` and `../FF_Q/Qoplsaa.prm` to get an impression of how a force field is implemented.

  Start the program `Qprep5` by typing its name in the UNIX shell. The program is interactive, so it is waiting for your instructions. Type:

  - `readlib ../FF_Q/Qoplsaa.lib`
  - `readlib ../FF_Q/heme.lib`
  - `readlib ../FF_Q/cma.lib`
  - `readprm ../FF_Q/Qoplsaa.prm`
  - `readpdb complex.pdb`
  - `addbond 5480 6383 y` (to form the bond between the Fe atom and the S of Cys 348; "y" indicates that we accept the 2.2 Å bond lenght)
  - `boundary sphere 407:C1 18` (the centre of the 18 Å radius sphere of simulation will be the central atom of the ligand, indicated by residue:atom_name)
  - `solvate 407:C1 18 1 HOH` (solvate with TIP3P waters the sphere of simulation)
  - `maketop "cma compex topology"` (make the topology, giving a title for it)
  - `writetop cma.top`
  - `writepdb cma_top.pdb y` (we will have a pdb of the starting coordinates of the system; "y" indicates write GAP between molecules)
  - `quit`

We have generated the topology file, `cma.top`, which will be the file used as input for the MD simulation.
Alternatively you can write all `Qprep5` commands in a separate file (*e.g.* `maketop.inp` and run `Qprep5 < maketop.inp > maketop.log`, so you can carefully analyze the output given by the program.

- Visual analysis:
  Typing the command
  `pymol topology.pml`
  you will get a nice view of the complex.

  – Take a look at the location of the ligand, in the centre of the sphere, and the dimensions of the sphere of simulation. In practice, only atoms inside the sphere will be taken into account in the MD simulation. The definition of the size of the sphere must be a compromise between accurate description of the long-range interactions and computational effort.

2.2 Input file and running the MD simulation in `Qdyn5`

We have done half of the work. Now we need to prepare the input files for the MD simulation, with the specifications about the MD conditions.
Any MD simulation with explicit solvent is divided in two phases:

- Equilibration phase: The system must be equilibrated, since we start from a frozen image of the complex (*i.e.* crystallographic coordinates) solvated with a predefined grid of waters.

- Production phase: This is the part of interest, which will be later analyzed by extracting the information about the energies and other properties of interest.

In this practical we will run the equilibration of the complex P450cam-CMA starting from the topology just generated. There will be 5 blocks of equilibration, given by the input files `eq1.inp` to `eq5.inp`. The variables of each block are outlined in the following table:

Starting file	Input file	Temp (K)	Bath coupling (fs)	Δ t (fs)	*n* steps	Force constant ($\frac{\text{kcal}}{\text{mol}}\dot{\text{A}}^2$)
complex.top	eq1.inp	1	0.2	0.2	100	200
eq1.re	eq2.inp	50	10	1	500	50
eq2.re	eq3.inp	150	10	1	500	25
eq3.re	eq4.inp	300	10	1	500	10
eq4.re	eq5.inp	300	10	1	500	2

- `eq1`: It is similar to energy minimization of the solvent and hydrogens of the solute: in this 0.02 ps run, we use a short time step and a strong coupling to the thermal bath at 1 K temperature, and heavy solute atoms restrained to their starting positions.

- `eq2 - eq4`: The system is gradually heated to 50, 150 and 300 K during 0.5 ps on each temperature point, with the time step increased and coupling constant relaxed. The restraints on heavy solute atoms are gradually relaxed.
- `eq5`: It is very similar to a production phase, but still maintaining a weak constraint on heavy solute atoms.

A closer look to the last input file (`eq5.inp`) will tell us exactly which keywords are used on the MD simulation of this example:

.....

```
[MD]
steps                500
stepsize             1
temperature          300.0
bath_coupling        10
random_seed          1231
shake_solute         off
lrf                  on

[sphere]
shell_force          10
shell_radius         0.80

[intervals]
output               50

[files]
topology             cma.top
restart              eq4.re
final                eq5.re
fep                  lie.fep

[sequence_restraints]
1 6486 2.0 0
```

.....

Each of the titles in brackets starts a section. We are thus defining basic simulation data (section [MD]), settings for the sphere of simulation (section [sphere]), intervals for saving data and updating non-bonded interactions (section [intervals]), file names for input and output ([files]) and restraining sequences of atoms (section [sequence_restraints]). In this last section, the line refers to the restraint on every heavy atom in the solute.
Make sure that you understand the meaning of every parameter in the input file. Now you can launch the job. This can be easily done by running the shell script `run_Q.sh`. Type:

`tcsh run_Q.sh qdyn eq`
and wait 5 minutes until the simulation has finished.

3 Analysis of the results

- Open the last log file `eq5.log` with your favourite text editor (`emacs, vi`). The file has a first part in which the input data and options are read, and a second part in which the information from the trajectory is prediodically stored.
  - Check that all input parameters are right and that your system is effectively neutral (look for the Total charge of system).
  - How often (how many time steps) is the output written?
  - Is the temperature effectively kept constant?
  - Take a look at the register of the surrounding energies of the ligand along the trajectory. This is an example of an average property calculated through MD.
- You can also have a graphical feeling of the MD simulation corresponding to the "production phase" (eq5), opening the trajectory file on `PyMOL` by typing:
  `pymol trajectory.pse`
  Once the file has been loaded click on the arrows on the lower-right corner of the graphical interface to play the trajectory, and you will see a small video of your run.
  - Is the system stable?