CS612 - Algorithms in Bioinformatics

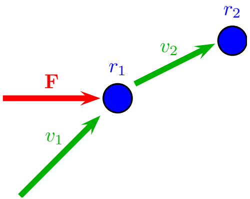
Protein Folding

March 27, 2023

Biomolecular Simulations using Molecular Dynamics (MD)

- A method that simulates the dynamics of molecules under physiological conditions
- Use physics to find the potential energy between and forces acting on all pairs of atoms.
- Move atoms to the next state.
- Repeat.

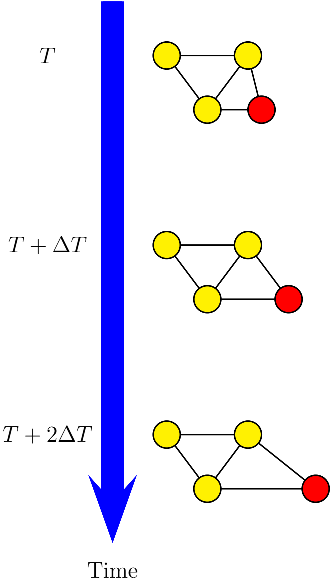
Using Newton's Second Law to Derive Equations



- $F = Ma = M * (dv/dt) = M * (d^2r/dt^2)$
- Or, with a small enough time interval Δt: $\Delta v = (F/M) * \Delta t \rightarrow v_2 = v_1 + (F/M)\Delta t$
- This is a second order differential equation:
- $r_2 = r_1 + v_2 dt = r_1 + v_1 dt + (F/M)dt^2$
- The new position, r_2 is determined by the old position, r_1 and the velocity v_2 over time Δt (which should be very small!).
- The above equation describes the changes in the positions of the atoms over time.

The process of MD

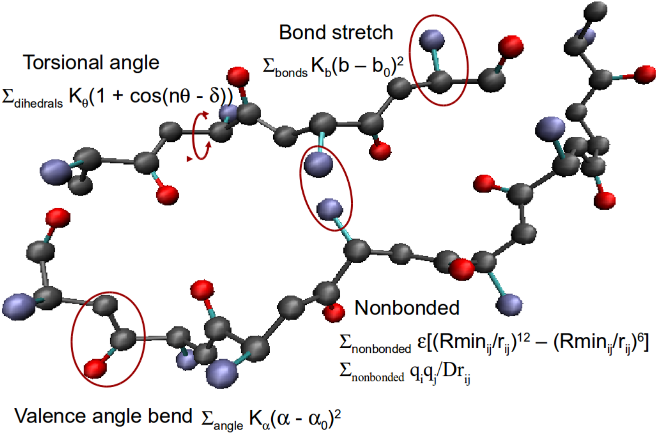
- The simulation is the numerical integration of the Newton equations over time
- Positions and velocities at time t →Positions and velocities at time t+dt
- Positions + velocities = trajectory.
- We get the initial positions and velocities as starting conditions
- Atom masses can be given as parameters (known experimentally)
- What about the force?



Connection Between Force and Energy

- $F = -dU/dr \rightarrow U = -\int Fdr = -1/2 * Mv^2$
- U = Potential energy (taken from the force field parameters)
- Gradient w.r.t. r – position vector, gives the force vector
- Energy is conserved, hence $\frac{1}{2} * \sum_{i=1}^{n} M_i v_i^2 + \sum E_{pot,i} = const$
- All the **equations** and the **adjusted parameters** that allow to describe quantitatively the energy of the chemical system are denoted force field.
- Note, that mixing equations and parameters from different systems always results in errors!
- Force field examples: CHARMM, AMBER, GROMACS etc.

Force Field Equations



Torsional angle

$\Sigma_{dihedrals} K_{\theta}(1 + \cos(n\theta - \delta))$

Bond stretch

$\Sigma_{bonds} K_{b}(b - b_{0})^{2}$

Nonbonded

$\Sigma_{nonbonded}\, \varepsilon[(Rmin_{ij}/r_{ij})^{12} - (Rmin_{ij}/r_{ij})^{6}]$

$\Sigma_{nonbonded}\, q_{i}q_{j}/Dr_{ij}$

Valence angle bend $\Sigma_{angle} K_{\alpha}(\alpha - \alpha_{0})^{2}$

Force Field Equations

$$U =$$

$$\sum_{bonds} K_b(b - b_0)^2 +$$ Bonds

$$\sum_{angles} K_\alpha(\alpha - \alpha_0)^2 +$$ Angles

$$\sum_{torsion} \frac{V_n}{2}(1 + \cos[n\theta - \delta]) +$$ Dihedrals

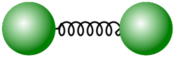
$$\sum_{i,j} \frac{q_i q_j}{\epsilon r_{ij}} +$$ Electrostatic

$$\sum_{i,j} \varepsilon[(\frac{Rmin_{ij}}{r_{ij}})^{12} - (\frac{Rmin_{ij}}{r_{ij}})^6]$$ Van der Waals (VdW)

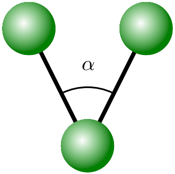
Force Field Equations

- Bonds, angles, dihedrals – Bonded terms
- Electrostatic, VdW – Non-bonded terms (calculated only for atoms at least 4 bonds apart)
- Other terms may appear as well
- The constants are taken from the force-field parameter files

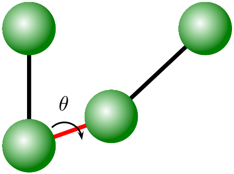
Bonded Terms



$K_b(b - b_0)^2$
Streching

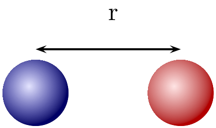


$K_\alpha(\alpha - \alpha_0)^2$
Bending



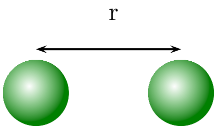
$\frac{V_n}{2}(1 + \cos[n\theta - \delta])$
Torsion

Non-Bonded Terms



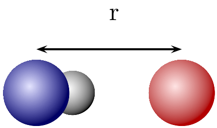
$$\frac{q_i q_j}{\epsilon r_{ij}}$$

Electrostatic



$$\varepsilon[(\frac{Rmin_{ij}}{r_{ij}})^{12} - (\frac{Rmin_{ij}}{r_{ij}})^{6}]$$

VdW

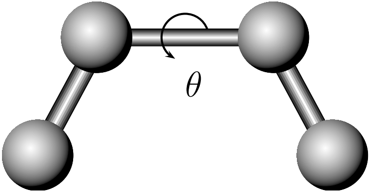


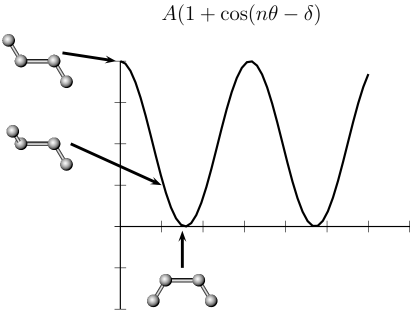
$$\varepsilon[(\frac{C_{ij}}{r_{ij}})^{12} - (\frac{D_{ij}}{r_{ij}})^{10}]$$

H-bond (optional)

Torsion Energy

$$E = \sum_{torsion} \frac{V_n}{2}(1 + \cos[n\theta - \delta])$$





V_n controls the amplitude of the curve
n controls its periodicity
δ shifts the entire curve along the rotation angle axis (θ).

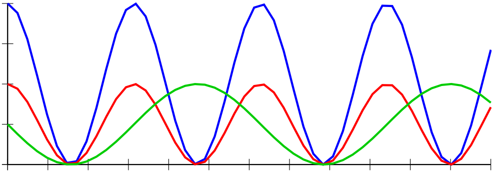
The parameters are determined from curve fitting.
Unique parameters for torsional rotation are assigned to each bonded quartet of atoms based on their types (e.g. C-C-C-C, C-O-C-N, H-C-C-H, etc.)

Torsion Energy Parameters

$A = 2.0, n = 2.0, \delta = 0.0^\circ$

$A(1 + \cos(n\theta - \delta))$ $A = 1.0, n = 2.0, \delta = 0.0^\circ$

$A = 1.0, n = 1.0, \delta = 90.0^\circ$



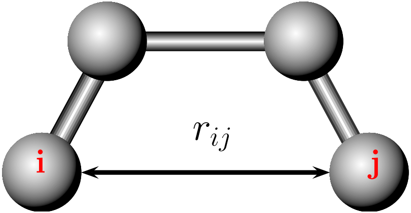
A is the amplitude.

n reflects the type symmetry in the dihedral angle.

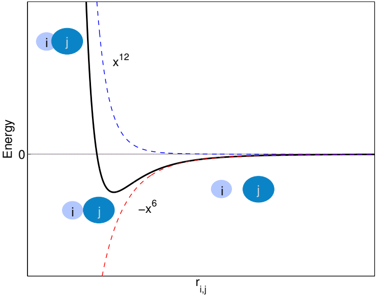
δ used to synchronize the torsional potential to the initial rotameric state of the molecule

Non-Bonded Energy Parameters

$$E = \sum_{i,j}\left(\frac{q_i q_j}{\epsilon r_{ij}} + \varepsilon\left[\left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}}\right)^{6}\right]\right)$$



$$\frac{Rmin_{ij}}{r_{ij}^{12}} - \frac{Rmin_{ij}}{r_{ij}^{6}}$$



The 12^{th} power term is the repulsion
The 6^{th} power term is the attraction
q_i is the partial charge of atom i
$Rmin_{ij}$ determines the well depth
ϵ is the dielectric constant

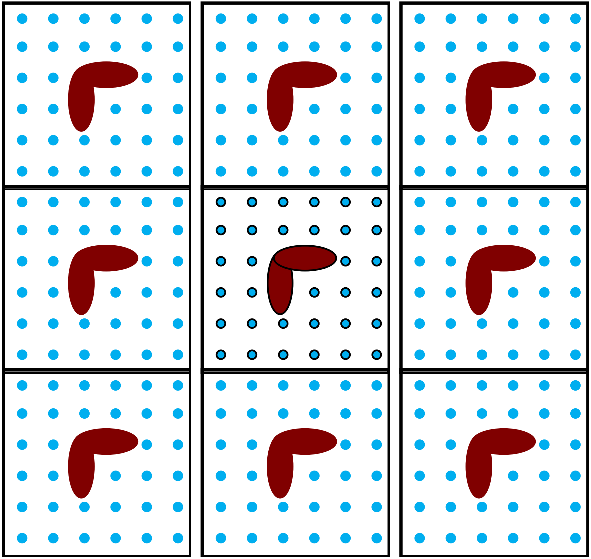
Solvation Models

- No solvent – constant dielectric.
- Continuum – referring to the solvent as a bulk. No explicit representation of atoms (saving time).
- Explicit – representing each water molecule explicitly (accurate, but expensive).
- Mixed – mixing two models (for example: explicit + continuum. To save time).

Periodic Boundary Conditions

- Problem: Only a small number of molecules can be simulated and the molecules at the surface experience different forces than those at the inner side.
- The simulation box is replicated infinitely in three dimensions (to integrate the boundaries of the box).
- When the molecule moves, the images move in the same fashion.
- The assumption is that the behavior of the infinitely replicated box is the same as a macroscopic system.

Periodic Boundary Conditions



A sample MD protocol

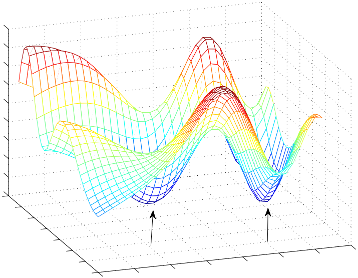
- Read the force fields data and parameters.
- Read the coordinates and the solvent molecules.
- Slightly minimize the coordinates (the created model may contain collisions), a few SD steps followed by some ABNR steps.
- Warm to the desired temperature (assign initial velocities).
- Equilibrate the system.
- Start the dynamics and save the trajectories every 1ps (trajectory=the collection of structures at any given time step).

Why is Minimization Required?

- Most of the coordinates are obtained using X-ray diffraction or NMR.
- Those methods do not map the hydrogen atoms of the system.
- Those are added later using modeling programs, which are not 100% accurate.
- Minimization is therefore required to resolve the clashes that may “blow up” the energy function.

Common Minimization Protocols

- First order algorithms: Steepest descent, Conjugated gradient
- Second order algorithms: Newton-Raphson, Adopted basis Newton Raphson (ABNR)



Steepest Descent

- This is the simplest minimization method:
- The first directional derivative (gradient) of the potential is calculated and displacement is added to every coordinate in the opposite direction (the direction of the force).
- The step is increased if the new conformation has a lower energy.
- Advantages: Simple and fast.
- Disadvantages: Inaccurate, usually does not converge

Conjugated Gradient

- Uses first derivative information + information from previous steps – the weighted average of the current gradient and the previous step direction.
- The weight factor is calculated from the ratio of the previous and current steps.
- This method converges much better than SD.

Newton-Raphson's Algorithm

- Uses both first derivative (slope) and second (curvature) information.
- In the one-dimensional case: $x_{k+1} = x_k + \frac{F'(x_k)}{F''(x_k)}$
- In the multi-dimensional case – much more complicated (calculates the inverse of a hessian [curvature] matrix at each step)
- Advantage: Accurate and converges well.
- Disadvantage: Computationally expensive, for convergence, should start near a minimum.

Adopted Basis Newton-Raphson's Algorithm (ABNR)

- An adaptation of the NR method that is especially suitable for large systems.
- Instead of using a full matrix, it uses a basis that represents the subspace in which the system made the most progress in the past.
- Advantage: Second derivative information, convergence, faster than the regular NR method.
- Disadvantages: Still quite expensive, less accurate than NR.

Assignment of Initial Velocities

- At the beginning the only information available is the desired temperature.
- Initial velocities are assigned randomly according to the Maxwell-Bolzmann distribution:

$$P(v)dv = 4\pi(\frac{m}{2\pi k_B T})^{\frac{3}{2}} v^2 e^{\frac{-mv^2}{2k_B T}}$$

- P(v) - the probability of finding a molecule with velocity between v and dv.
- Note that:
  1. The velocity has x,y,z components.
  2. The velocities exhibit a gaussian distribution

Bond and Angle Constraints (SHAKE Algorithm)

- Constrain some bond lengths and/or angles to fixed values using a restraining force G_i.

$$m_i a_i = F_i + G_i$$

- Solve the equations once with no constraint force.
- Determine the magnitude of the force (using lagrange multipliers) and correct the positions accordingly.
- Iteratively adjust the positions of the atoms until the constraints are satisfied.

Equilibrating the System

- Velocity distribution may change during simulation, especially if the system is far from equilibrium.
- Perform a simulation, scaling the velocities occasionally to reach the desired temperature.
- The system is at equilibrium if:
    - The quantities fluctuate around an average value.
    - The average remains constant over time.

The Verlet Integration Method

Taylor expansion about r(t):

$$
\begin{aligned}
r(t+\delta t) &= r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \ldots \\
r(t-\delta t) &= r(t) - v(t)\delta t + \frac{1}{2}a(t)]\delta t^2 + \ldots
\end{aligned}
$$

Adding the two terms gives a velocity independent term:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2$$

The odd terms go, so the error is the order of magnitude of δt^4, the next term

The Verlet Integration Method

Velocities can be calculated via the derivation method:

$$v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t}$$

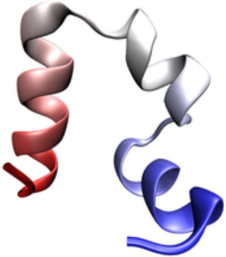
Here the error is of order δt^2.
Note – the time interval δt is in the order of 1fs. (10^{-15}s)

The Verlet Algorithm

1. Start with $r(t)$ and $r(t - \delta t)$
2. Calculate a(t) from the Newton equation: $a(t) = f_i(t)/m_i$.
3. Calculate $r(t + \delta t)$ according to the aforementioned equation.
4. Calculate $v(t)$.
5. Replace $r(t - \delta t)$ with $r(t)$ and $r(t)$ with $r(t + \delta t)$.
6. Repeat as desired.

Case Study – Vilin Headpiece Simulation

- Folds very fast – 4-5ms
- A mutant folds in under 1ms.
- Folding process characterized in all-atom explicit solvent simulation.

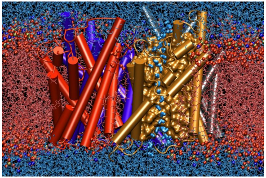


http://www.ks.uiuc.edu

Case Study – Water Transport in Aquaporins

- Membrane water channels that play critical roles in controlling the water contents of cells.
- The pores are impermeable to charged species, such as protons, a remarkable property that is critical for the conservation of membrane's electrochemical potential, but paradoxical since protons can usually be transfered readily through water molecules

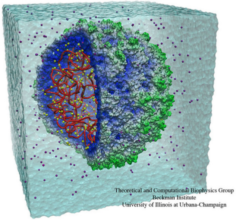
http://www.ks.uiuc.edu



Water molecules passing the channel are forced, by the protein's electrostatic forces, to flip at the center of the channel, breaking the alternative donor-acceptor arrangement that is necessary for proton translocation

Case Study – Simulating an Entire Virus

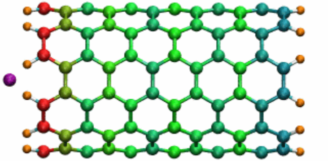
- viruses contain two components: the capsid (a protein shell), and a genome.
- MD shows the assembly and disassembly of several viruses as part of the virus life cycle.
- STMV (Satellite tobacco mosaic virus) particle consists of 60 identical copies of a single protein that make up the viral capsid (coating), and a 1063 nucleotide single stranded RNA genome which codes for the capsid and one other protein of unknown function.


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

http://www.ks.uiuc.edu

Case Study – Potassium in a Carbon Nanotube

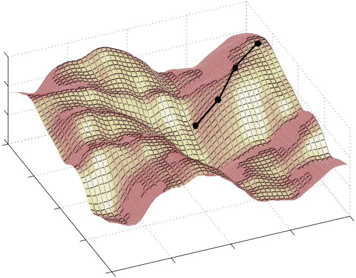
- The ion is attracted by the potential well and begins to oscillate.
- During the simulation, the ion finished two complete oscillation cycles with a frequency of 0.43 THz.
- The motion of the ion naturally drags the electrons of the SWNT to oscillate at the same frequency.
- The carbon atoms are colored according to their induced charges (red: negative; blue: positive).



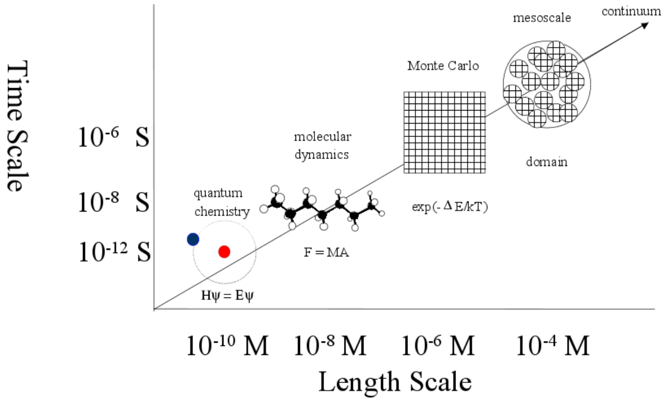
http://www.ks.uiuc.edu

MD Shortcomings – Expensive!

- Small integration time step (10^{-15} sec).
- Complex interactions between atoms in the molecule.
- Simulating 1ns of a medium sized protein (300+ amino acids, approx. 100,000 atoms incl. solvent) requires millions of calculations per step X 1,000,000 steps.
- Must use distributed computing to scale up to reasonable sized systems.



Time Scales for Simulations



MD Variants – Replica Exchange MD (REMD)

- Exhanced sampling relative to standard MD.
- Multiple replicas of the same system are run at different temperatures.
- This allows to overcome energy barriers on the potential energy surface.
- Every period of time (at least 1ps) replicas are exchanged among close-by temperatures.

MD Variants – Targeted MD (TMD)

- A subset of the atoms is guided towards a final target structure using a steering force.
- The steering force is assigned for each atom using the gradient of the following potential:
  $U_{TMD} = \frac{1}{2}\frac{k}{N}\left[RMSD(t) - RMSD^*(t)\right]^2$
- $RMSD(t)$ is the least RMSD of the current coordinates with the target coordinates at time t.
- $RMSD^*(t)$ evolves linearly from the initial RMSD at the first TMD step to the final RMSD at the last TMD step.
- The spring constant k is scaled down by the number N of targeted atoms.

MD Variants – Steered MD (SMD)

- The basic idea is to apply an external force to one or more atoms, which we refer to as SMD atoms.
- Another group of atoms may be held fixed.
- This enables to study the behaviour of your protein under various conditions.
- Examples – (un)folding and binding events that do not happen under MD time scales.