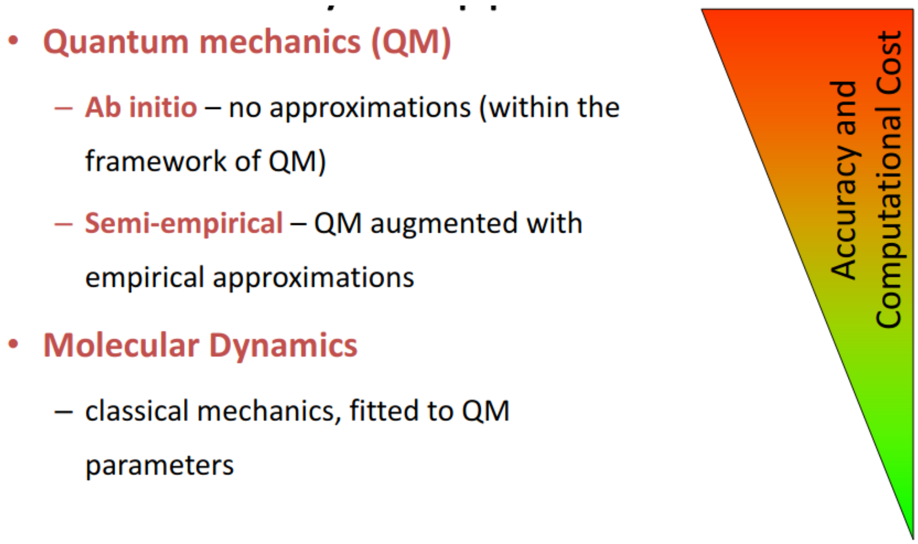
5 Prediction of protein structure

Introduction

Thermodynamic hypothesis: Proteins fold into the conformation of minimal free energy. By finding this minima we can solve the protein folding problem.

Proteins are flexible, and this is essential for function. This can be simulated in several ways, with different degrees of accuracy.



- **Quantum mechanics (QM)**
  - **Ab initio** – no approximations (within the framework of QM)
  - **Semi-empirical** – QM augmented with empirical approximations
- **Molecular Dynamics**
  - classical mechanics, fitted to QM parameters

Accuracy and Computational Cost

Classical mechanics

Is the easiest method computationally speaking. Models forces using hook's law (harmonic potential), doesn't take into account quantum effects and doesn't treat electrons explicitly. The time complexity of considering bond mediated interactions is $O(N)$. Considering spatial interactions, is $O(N^k)$, where k is the number of elements considered for the spatial interactions.

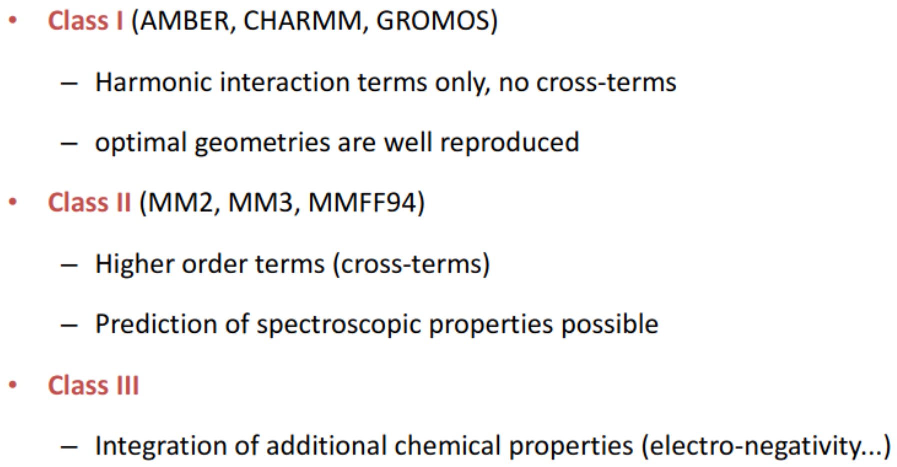
Force fields

A force field is a model whose objective is describing the energetic landscape of the different possible conformations of a molecule. For that purpose a force field describes the interactions through different analytical forms, and contains also knowledge-derived parameters that make possible this modelization and rules to associate parameters to atom types. They can make use of different treatments of the analytical forms (classical treatment, ML derived...) Parametrization can be:

- Based on quantum-mechanics calculations
  - Exotic geometries are accesible

  - Critically depends on choice of modeling method
  - inexpensive
- Based on experimental data
  - Limited by the scarcity of good experimental data
  - Only natural geometries are accesible

Some examples of Force fields often used for protein structure prediction are AMBER, CHARMM and GROMOS.



- Class I (AMBER, CHARMM, GROMOS)
  - Harmonic interaction terms only, no cross-terms
  - optimal geometries are well reproduced
- Class II (MM2, MM3, MMFF94)
  - Higher order terms (cross-terms)
  - Prediction of spectroscopic properties possible
- Class III
  - Integration of additional chemical properties (electro-negativity...)

Example 1: ECEPP/2

- Models bond lengths and angles as stiff, so they are constraints and not parameters.
- The conformational flexibility is derived from different dihedral angles.
- Van der waals interactions are modeled using Lennard-Jones-potential
- Electrostatics using Coulomb
- Torsions using cosine-term
- H-Bonds using 10-12 potential are treated as non-directional electrostatic interactions

Example 2: AMBER (Assisted Model Building with Energy Refinement)

Five contributions:

- Bond lengths ($E_{stretch}$)
- Bond angles (E_{bend})
- Dihedrals (E_{tors})
- Van-der-Waals (E_{vdW})
- Electrostatics (E_{ES})

$E = E_{stretch} + E_{bend} + E_{tors} + E_{vdW} + E_{ES}$

It also uses additional constraints to keep planar systems (aromatic rings) planar.

It uses 54 different atom types, i.e: - C - sp^2 carbon in carbonyl group - CT - Aliphatic spˆ3 carbon …

Class II Force Fields

- Make use of Anharmonic potentials (i.e Morse potential, more accurate than harmonic potential for describing bond energecits)
- Include cross terms, describing coupling between different terms (e.g. bond lengths and angles)

Modeling of particle motion

Can be modeled using classical mechanics in function of position velocity and acceleration: Three newton laws of motion.

Acceleration: can be computed derived from forces and masses thanks to the equation:

$a = F/m$

Force can be derived,as commented, from the energy thanks to the equation:

$F(r) = -\nabla E(r)$

where $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$

Then this information can be used for creating a system of linear equations, which can be numerically solve yielding a simulated molecular dynamics.

Trajectory

Trajectory: the path that an object with mass in motion follows through space as a function of time.

Trajectories can be found integrating the equations of motion.

Phase space

The positions of the elements of the system is not sufficient to describe the state of a dynamic system.

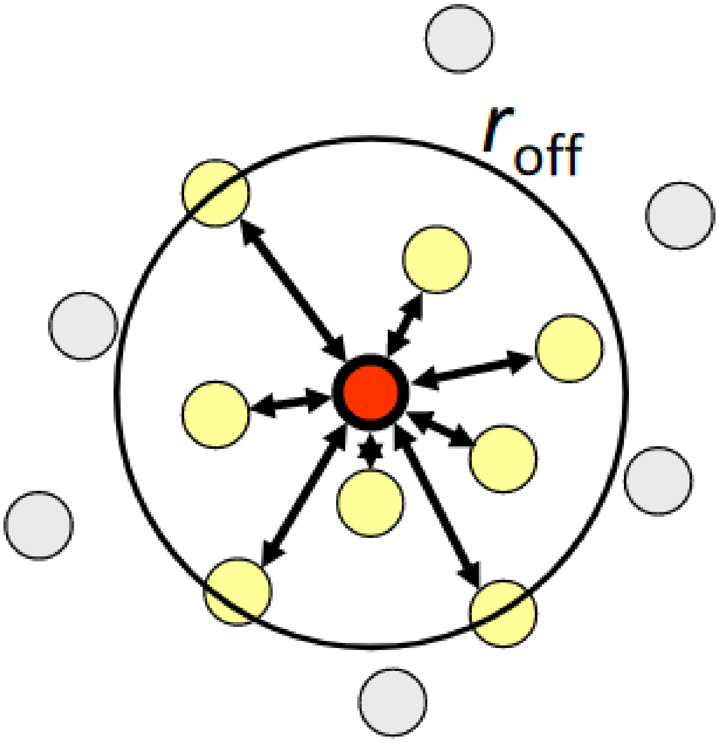
Phase space: Positions + momentums of the elements of the system describing its state.

Molecular dynamics simulations (MDS)

Simulation of the dynamics of molecular systems based on force fields and associated equations of motion. e.g Derivatives of AMBER terms.

They can make use of cut-off radii: Compute interactions only with molecules in a ball centered on the particle with a given radius. but:

They introduce discontinuities (solved using solution-shifting functions or switching functions)



Algorithms used for integrating equations of motion

- Motion of particles is coupled (many-body problem)
- Analytical integration not possible $\implies$ Approximated by numerical integration
- Assumption: terms expressed as Taylor-series expansion

Taylor-expansion

$f(x + \Delta X) = \sum \frac{1}{v!} f^v(x) \Delta x^v$

This is an infinite power series, but are usually approximated by terminating series.

Verlet integration

Uses information of current and preceeding timestep to compute position and forces for the next time step. At time t, t r(t) and r(t-Δ t are known)

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

After the computation of the next discrete position, iteration starts again.

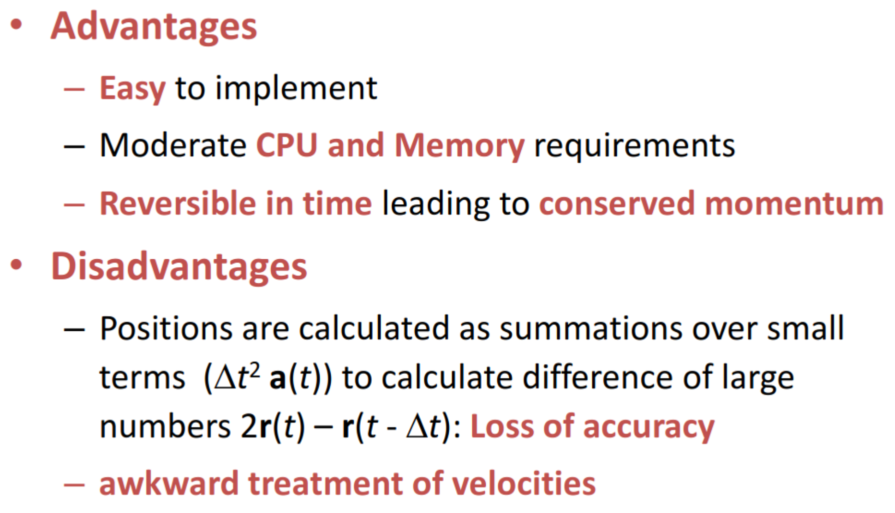
It can be modified to be used with velocities:

$v(t) = \frac{1}{2\Delta t}\left(r(t + \Delta t - r(t - \Delta t))\right)$

Where also an initial condition is necessary because $t = 0$ is not defined

Time step:

- if too short the simulation takes unnecessarily long
- if too long the forces becomes too large and the simulation "explodes"



- **Advantages**
  - **Easy** to implement
  - Moderate **CPU and Memory** requirements
  - **Reversible in time** leading to **conserved momentum**
- **Disadvantages**
  - Positions are calculated as summations over small terms $(\Delta t^2\ \mathbf{a}(t))$ to calculate difference of large numbers $2\mathbf{r}(t) - \mathbf{r}(t - \Delta t)$: **Loss of accuracy**
  - **awkward treatment of velocities**

Simulation of water

Can be treated explicitly. But simulations become very expensive.

Simulations depending on T

https://sites.engineering.ucsb.edu/~shell/che210d/Advanced_molecular_dynamics.pdf

If T is constant: the canonical NVT-ensemble treats pressure and energy as variables, and the different temperatures describe the tendence of the system to fold or not. If we want to consider a system where the temperature is not constant, we have to consider the changes in the molecular kinetics produced by this changes. Several different strategies:

- Velocity rescaling: At each time step, the new temperature is derived from the kinetic energy. This doesn't capture the correct kinetic energy fluctuations.
- Berendsen thermostat: Similar to Velocity rescaling, but there's an additional timescale for temperature rescaling. It suffers from the same problems as velocity rescaling.
- Andersen thermostat: It also models random colisions between particles, and particles and the heat bath, in function of a colision frequency (τ), where kinetic energy is transfered. It produces the correct canonical ensemble, but true molecular kinetics are not represented.
- Other thermostats: More involved and accurate approaches.