IMA Summer Program

Classical and Quantum Approaches in Molecular Modeling

Lecture 1: Introduction to Molecular Dynamics

Robert D. Skeel
Department of Computer Science, and of Mathematics
Purdue University

http://bionum.cs.purdue.edu/2007July23.pdf

Acknowledgments: V. Dadarlat, NIH

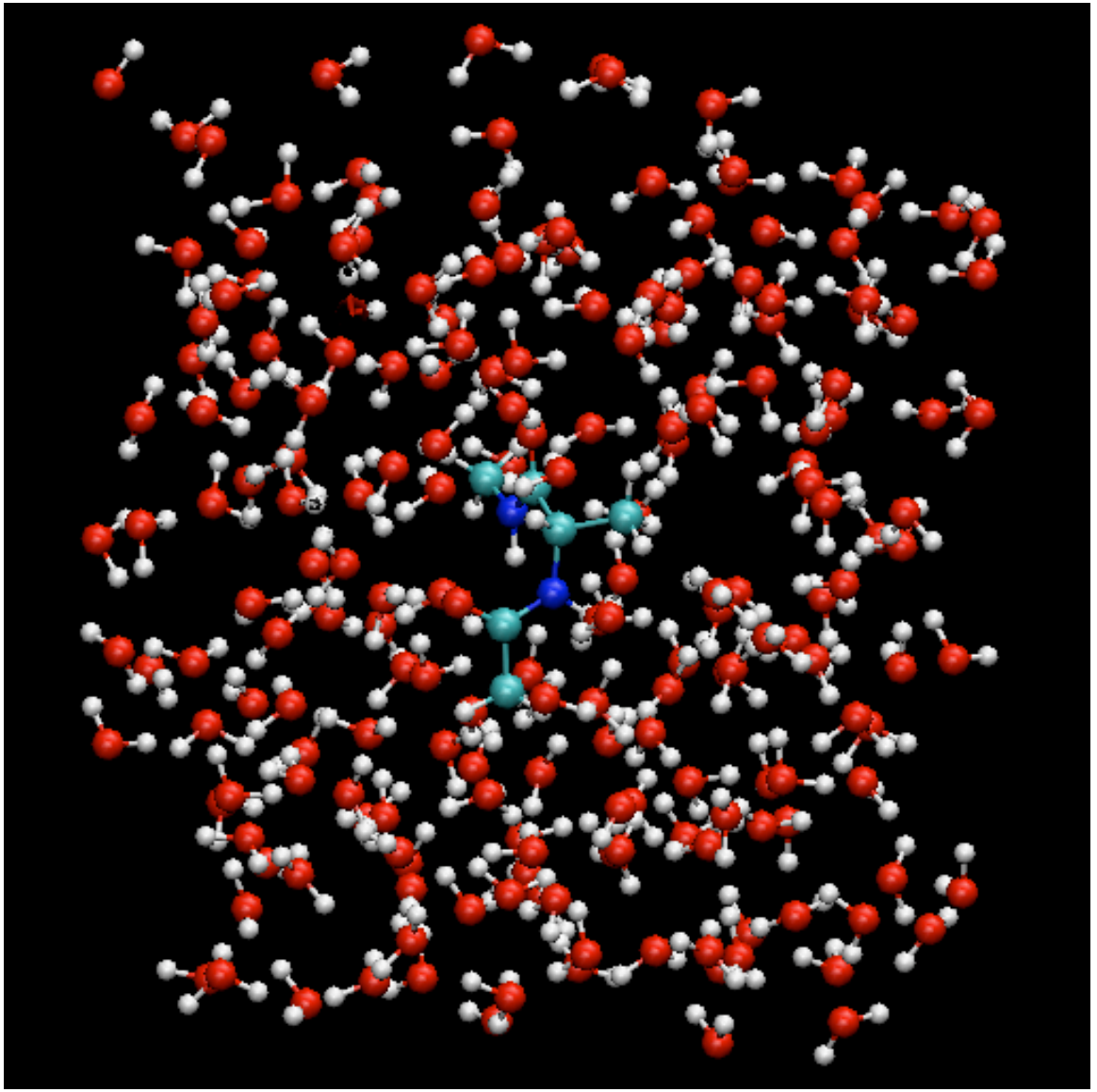
condensed matter
physics

physical chemistry

materials science
mechanical
engineering

...

molecular biophysics
structural biology



Twenty years ago at the IMA

Workshop on Atomic and Molecular Structure and Dynamics
week 5: July 13–17, 1987
Wilfred F. van Gunsteren

He expected most improvements to come from hardware rather than algorithms.

Since then, a 10 000-fold improvement in processor speed,
a factor of 20? typically for parallelism, and
a factor of 25?? typically for algorithms.

Possibilities for algorithms

Dramatic improvements in algorithms seem possible
—without radical innovations—by
deploying innovative ideas scattered in the literature,
improving them through analysis and abstraction, and
combining them.

Massive parallelization of algorithms is another opportunity
(but MPI is too low-level for regular use).

Twenty years from now

What is possible 20 years from now for the same accuracy compared to current practice?

integrators and fast force evaluation: factor of 5
sampling: factor of 10
coarse-graining: factor of 10

parallelism: factor of 50

processor speed: factor of 10

Misleading ideas

Structure determination is
the minimization of potential energy.

Molecular dynamics is
the calculation of a real trajectory.

Remedy:
Lecture 2: Statistical Mechanics and Molecular Dynamics

Outline

I. Equations of motion

II. Boundary effects

III. Initial conditions

IV. Computational tasks: thermodynamics and structure

V. Computational tasks: kinetics

VI. Practicalities

Atomistic models

x collection of N atomic positions $\vec{r}_i$,

M diagonal matrix of masses m_i,

v collection of velocities,

$p = Mv$ collection of momenta,

micro(scopic) state:

$$\Gamma = \begin{bmatrix} x \\ p \end{bmatrix}$$

possibly other variables, e.g., volume

Force field

Potential energy $U(x)$ is a sum of

$\mathcal{O}(N)$ few-body potentials for covalent bonded forces,

$\mathcal{O}(N^2)$ 2-body potentials for nonbonded forces,

which approximate quantum mechanics.

Nonbonded energy terms

1. Coulombic

$$\text{const}\frac{q_i q_j}{|\vec{r}_j - \vec{r}_i|}$$

where q_i, q_j are partial charges

2. London/dispersion (van der Waals)

$$-\text{const}_{ij}\frac{1}{|\vec{r}_j - \vec{r}_i|^6}$$

3. excluded volume

$$+\text{const}_{ij}\frac{1}{|\vec{r}_j - \vec{r}_i|^{12}}$$

Bonded energy terms

1. bond stretching for bond $i - j$:

$$\text{energy} = \text{const}(r_{ij} - \ell_0)^2.$$

2. angle bending for bonds $i - j - k$:

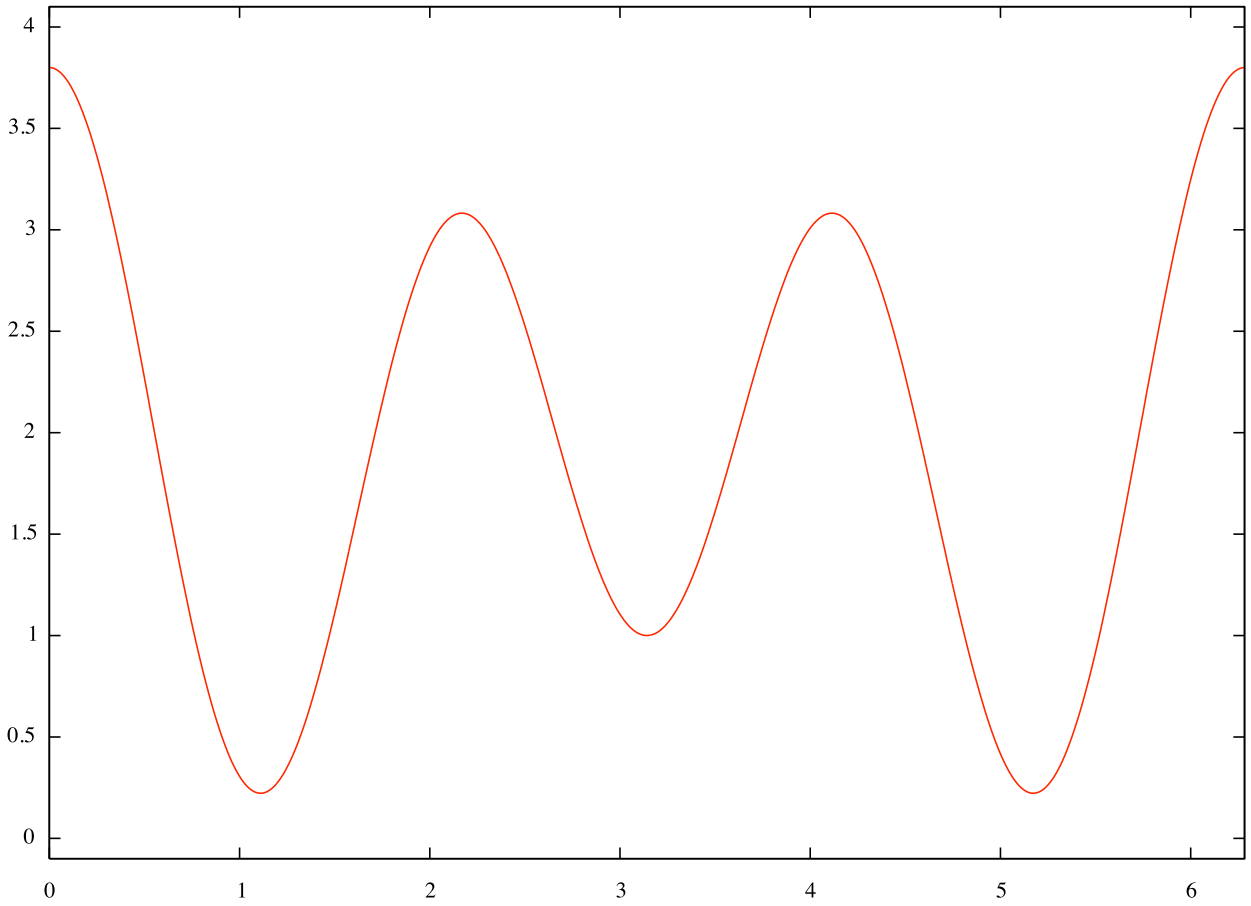
$$\text{energy} = \text{const}(\theta - \theta_0)^2$$

where θ is the angle between the two bonds.

3. Consider bonds $i - j - k - l$.
   With the 3 bond lengths and 2 bond angles fixed,
   rotation about middle bond $j - k$ remains possible.
   Torsion (aka dihedral) angle φ
   is clockwise rotation of $i - j$ about the $k - j$ axis
   needed to minimize the distance from i to l.

$$\text{energy} = \sum_n \frac{1}{2} V_n (1 + \cos(n\varphi - \varphi_n)),$$

   where, for example, the sum might be over $n = 2, 3$. A typical potential plotted against φ ...



4. Miscellaneous: improper dihedral, CMAP correction.

Equations of motion

$F(x) = -\nabla U(x)$ collection of forces.

Equations are a Hamiltonian system,

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = M^{-1}p(t), \quad \frac{\mathrm{d}}{\mathrm{d}t}p(t) = F(x(t)),$$

with Hamiltonian $H(x,p) = \frac{1}{2}p^{\mathsf{T}}M^{-1}p + U(x)$.

A numerical integrator

The velocity Verlet scheme is

$$\begin{aligned} x^{n+1} &= x^n + \Delta t M^{-1} p^n + \frac{1}{2}\Delta t^2 M^{-1} F(x^n), \\ p^{n+1} &= p^n + \frac{1}{2}\Delta t (F(x^n) + F(x^{n+1})). \end{aligned}$$

Equivalent to the truncated Störmer and the leapfrog method.

An ancient method.

Discovered, not invented.

Alternative formulation

$$M\frac{x^{n+1}-2x^n+x^{n-1}}{\Delta t^2}=F(x^n)$$

$$v^n=\frac{x^{n+1}-x^{n-1}}{2\Delta t}$$

Trajectory error $\propto \Delta t^2 \mathrm{e}^{t/\tau}$.

$\tau \approx 50$ periods of fastest mode.

More later.

($\Delta t \approx \frac{1}{10}$ period.)

Refined models

employ a more complicated force field.

- polarizable forces
- bond-breaking force fields, e.g. REAXX
- quantum mechanics / molecular mechanics (QM/MM)

Coarse-grained models

use fewer degrees of freedom.

- constraints (remove highest frequency motions)
- implicit solvent
- reduced models (bottom-up coarse graining)
- continuum models (top-down coarse graining)

Outline

I. Equations of motion

II. Boundary effects

III. Initial conditions

IV. Computational tasks: thermodynamics and structure

V. Computational tasks: kinetics

VI. Practicalities

Modeling the surroundings

if thermal contact, temperature is specified
if mechanical contact, pressure is specified
if material contact, chemical potential is specified
    for each species

In the thermodynamic limit $N \to \infty$,
    boundary effects diminish as $\mathcal{O}(N^{-1/2})$.

modeling an isolated system ...

Spherical boundary restraints

Soft spherical wall centered at origin:
To $U(x)$ add a term

$$\frac{1}{4}k(\max\{0, |\vec{r}_i| - \text{radius}\})^4$$

for every atom i.

Seldom used in practice.

Instead, to avoid artificial boundary effects, use ...

Periodic boundaries

Simulation box is replicated infinitely often.

Forces are sums over infinitely many images.

Sums of Coulombic forces are not well defined.

Need to use a reasonable limiting process.

One such process leads to the Ewald sum,

which has the special property of being continuous as a function of x.

Be assured: no computational penalty.

Modeling thermal contact

A method easy to implement:
Instead of spherical restraints,
identify atoms i in the outermost layer
and harmonically restrain them to their initial positions,

add $\frac{1}{2}k|\vec{r}_i - \vec{r}_{i,0}|^2$ to $U(x)$,

and thermostat them stochastically.

Stochastically?

Wiener processes

A standard Wiener process $W(t)$, $t \geq 0$, is a family of Gaussian random variables, fully characterized by their expectations

$$E[W(t)] = 0$$

and covariances

$$E[W(s)W(t)] = \min\{s, t\}.$$

Stochastic thermostating

Add $-\gamma_i \frac{\mathrm{d}}{\mathrm{d}t}\vec{r}_i + \sqrt{2k_\mathrm{B}T\gamma_i m_i}\frac{\mathrm{d}}{\mathrm{d}t}\vec{W}_i(t)$ to $\vec{F}_i(x)$ where

k_B is Boltzmann's constant, T is temperature,

γ_i are damping constants (how to choose?),

$\vec{W}_i(t)$ are independent standard Wiener processes.

More discretely, add

$$-\gamma_i \frac{\vec{r}_i^{\,n+1} - \vec{r}_i^{\,n-1}}{2\Delta t} + \sqrt{2k_\mathrm{B}T\gamma_i m_i}\frac{\vec{W}_i(t^{n+1/2}) - \vec{W}_i(t^{n-1/2})}{\Delta t}$$

to $\vec{F}_i(x^n)$.

We have

$$\vec{W}_i(t^{n+1/2}) - \vec{W}_i(t^{n-1/2}) = \sqrt{\Delta t}\vec{Z}_i^n$$

where $\vec{Z}_i^n$ are independent standard Gaussian random numbers.

Outline

I. Equations of motion

II. Boundary effects

III. Initial conditions

IV. Computational tasks: thermodynamics and structure

V. Computational tasks: kinetics

VI. Practicalities

Initial conditions

The microstate of a system is largely unknown,
so initial values $\Gamma(0)$ are chosen at random
    from some prescribed distribution
    with probability density function (p.d.f.) $\rho(\Gamma)$,
...consistent with macroscopic observables such as T.

The appropriate distribution is known mathematically
    (under certain hypotheses).

For a system in thermal contact with its surroundings—the canonical ensemble, use the Boltzmann-Gibbs distribution:

$$\rho(\Gamma) = \mathrm{e}^{-H(\Gamma)/k_{\mathrm{B}}T} \Big/ \int \mathrm{e}^{-H(\Gamma)/k_{\mathrm{B}}T} \mathrm{d}\Gamma.$$

Probability depends on energy:

factor of 10 $\Leftrightarrow$ a difference of 1.4 kcal/mol,

factor of 10^5 $\Leftrightarrow$ a difference of 7.0 kcal/mol

(at physiological temperature).

Due to random initial values,

one must study an *ensemble* of systems.

Equilibration

A random microstate can be obtained for the canonical ensemble by performing a long episode of Langevin dynamics:

$$M\frac{\mathrm{d}}{\mathrm{d}t^2}x = F(x) - CM\frac{\mathrm{d}}{\mathrm{d}t}x + (2k_{\mathrm{B}}TCM)^{1/2}\frac{\mathrm{d}}{\mathrm{d}t}W(t)$$

where

C is a diagonal matrix of damping constants and

$W(t)$ is a set of $3N$ independent standard Wiener processes.

(Definitely nonphysical.)

Outline

I. Equations of motion

II. Boundary effects

III. Initial conditions

IV. Computational tasks: thermodynamics and structure

V. Computational tasks: kinetics

VI. Practicalities

Thermodynamics and structure

- Most quantities of interest are defined only in terms of a stationary distribution and the purpose of the simulation is to sample configuration space.

- In some cases, kinetic quantities are of interest and realistic Newtonian dynamics must be used.

In the former case, only $\rho(\Gamma)$ is needed
and equations of motion are not essential.

An observable is a ρ-weighted average of some function of the microstate:

$$\langle A(\Gamma)\rangle = \int A(\Gamma)\rho(\Gamma)\mathrm{d}\Gamma,$$

e.g., if $U^{\mathrm{int}}(x)$ is energy exclusive of outside contributions, $\langle U^{\mathrm{int}}(x)\rangle$ is the internal energy.
This might be calculated as

$$\langle A(\Gamma)\rangle \approx \frac{1}{N_{\mathrm{trials}}} \sum_{\nu=1}^{N_{\mathrm{trials}}} A(\Gamma_{(\nu)}),$$

which requires random sampling of phase space.

Representative tasks

- thermodynamics
  e.g., pressure vs. temperature

- structure
  e.g., radial distribution function

- energetics
  e.g., free energy differences, potentials of mean force

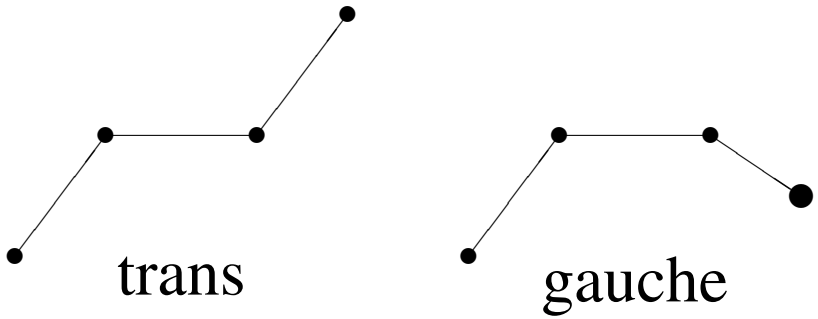
Structure

The meaning of this term ranges

from geometry:
configuration of a system
    modulo uniform translation and rotation

to topology:
bonding patterns. e.g., for proteins
(1) primary structure: covalently bonded sequence of amino acids
(2) secondary structure: backbone hydrogen bonds
    (strong noncovalent associations) N—H $\cdots$ O==C
(3) tertiary structure: other hydrogen bonds, salt bridges, . . .

Conformations



- clusters of configurations/structures
- better still, regions of configuration space such that transitions between them are rare
- more conveniently, dihedral angle ranges

Free energy of binding

Consider a protein in a *dilute* solution of ligands.
The free energy of binding ΔG is defined by the relation

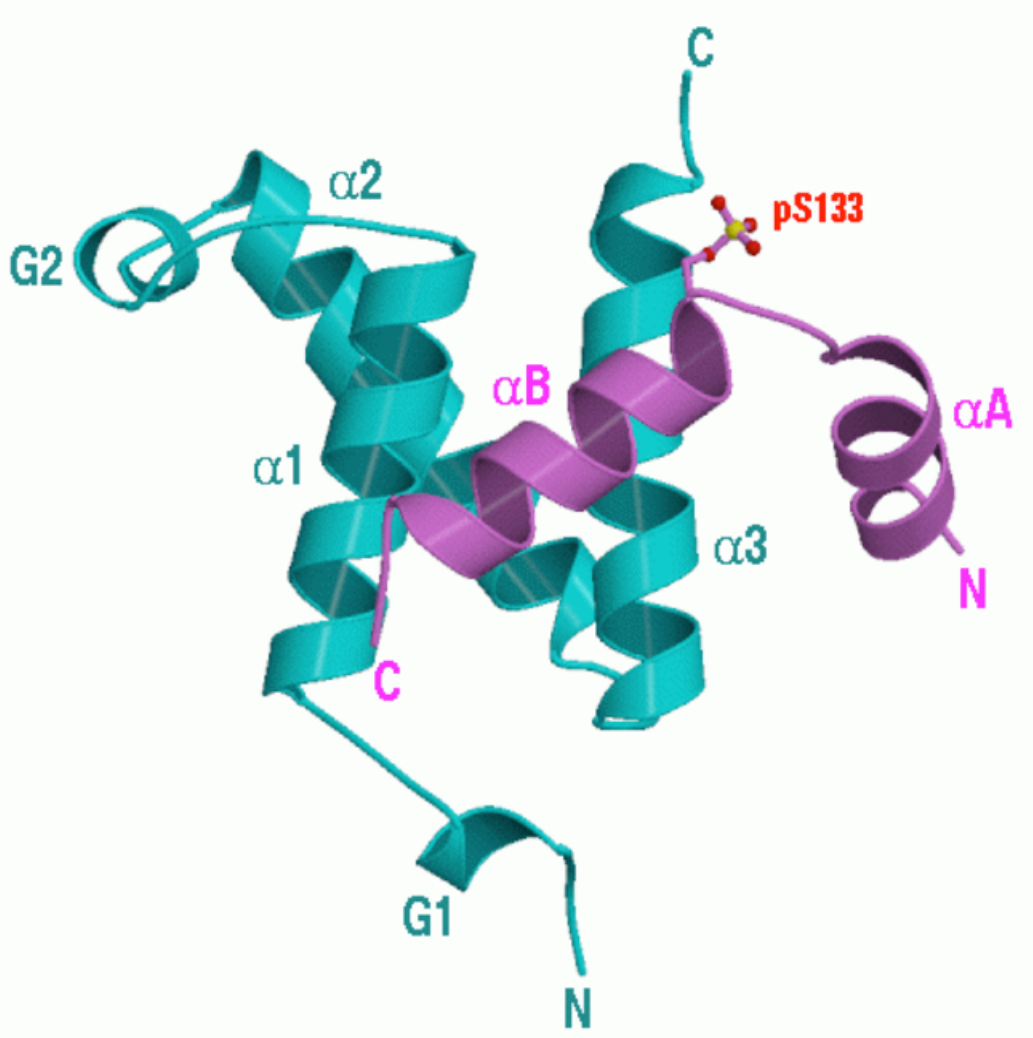
$$\frac{\text{Pr(a ligand is bound to protein)}}{\text{Pr(no ligand is bound to protein)}} = c\mathrm{e}^{-\Delta G/k_{\mathrm{B}}T}$$

where c is the ligand concentration in mol/liter.

It can be modeled with a single protein and ligand in solution,

e.g., ...

KIX and pKID:



Potentials of mean force

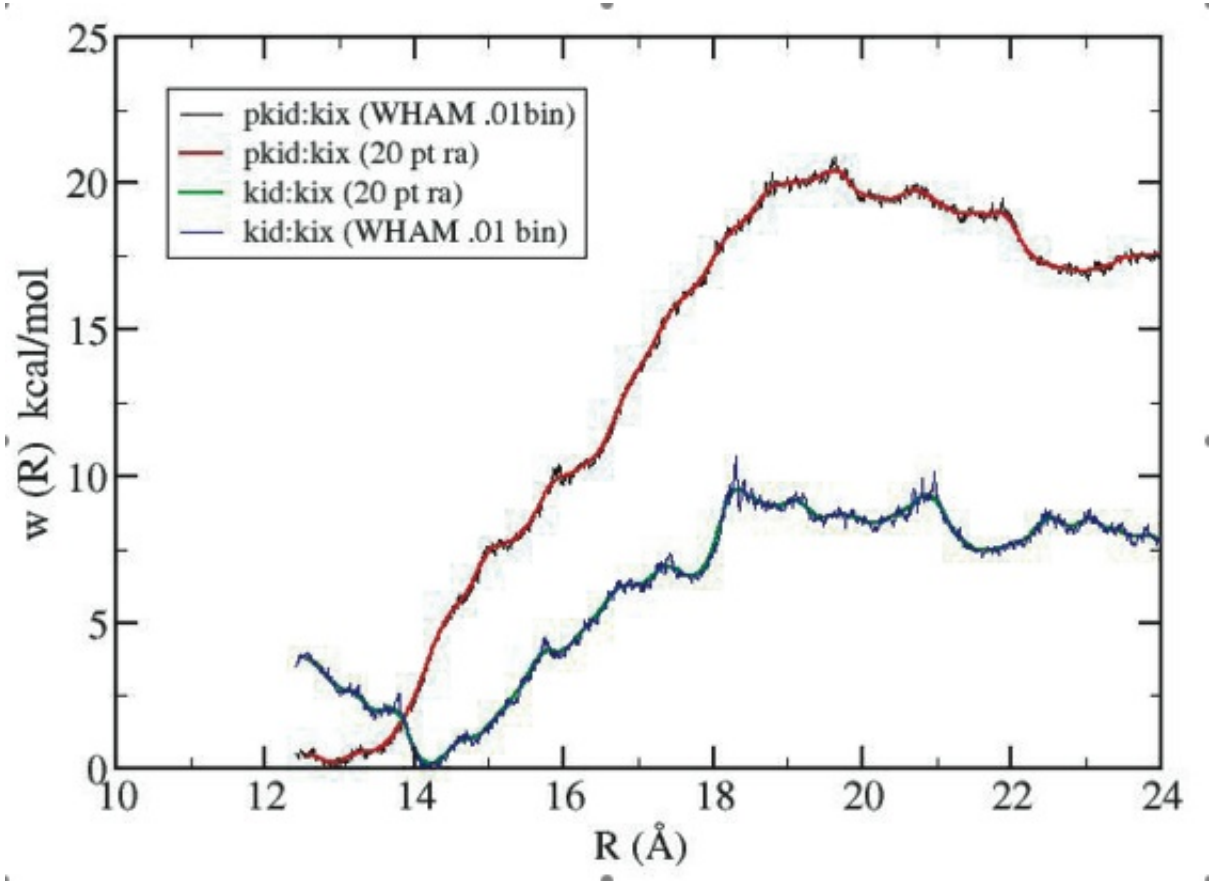
Let $R = \xi(x)$ be a reaction coordinate,
e.g., distance between the centers of mass of 2 molecules.
Let $\rho_\xi(R)$ is the p.d.f for $\xi(x)$ with x taken to be random.

$$\rho_\xi(R) = \int\int \delta(\xi(x) - R)\rho(x,p)\mathrm{d}x\mathrm{d}p.$$

The potential of mean force $w(R)$ is defined by

$$\mathrm{e}^{-w(R)/k_\mathrm{B}T} = \mathrm{const}\,\rho_\xi(R).$$

Example. With $R = \xi(x)$ the distance between centers of mass of KIX and pKID (or KID), the potential of mean force is



Outline

I. Equations of motion

II. Boundary effects

III. Initial conditions

IV. Computational tasks: thermodynamics and structure

V. Computational tasks: kinetics

VI. Practicalities

Kinetics

In principle, this requires an ensemble of, say, 2 to 20 000 realistic trajectories with random initial conditions:

$$\Phi_t(\Gamma_{(\nu)}), \quad \nu = 1, 2, \ldots, N_{\text{trials}}.$$

where $\Phi_t(\Gamma)$ denotes the t-flow of the dynamics (phase space trajectory with initial value Γ).

Representative tasks

- short animations
  e.g., impact of projectile on material (bullets move 15Å/ps)
- time correlation functions
- transition paths
- transition rates / conformational dynamics

Coping with chaos

Motion is chaotic and trajectories are swamped with error.

Compensated by the fact that initial values are unknown.

Times can be lengthened by shadowing arguments.

However, for very long times, only time correlation functions (and time averages) may be computable.

Bottom line:
formulate questions in terms of computable quantities.

Time correlation functions

(Unnormalized) time correlation function

$$\langle A(\Phi_t(\Gamma))B(\Gamma)\rangle = \int A(\Phi_t(\Gamma))B(\Gamma)\rho(\Gamma)\mathrm{d}\Gamma.$$

In principle, this might be calculated as

$$\langle A(\Phi_t(\Gamma))B(\Gamma)\rangle \approx \frac{1}{N_{\mathrm{trials}}}\sum_{\nu=1}^{N_{\mathrm{trials}}} A(\Phi_t(\Gamma_{(\nu)}))B(\Gamma_{(\nu)}).$$

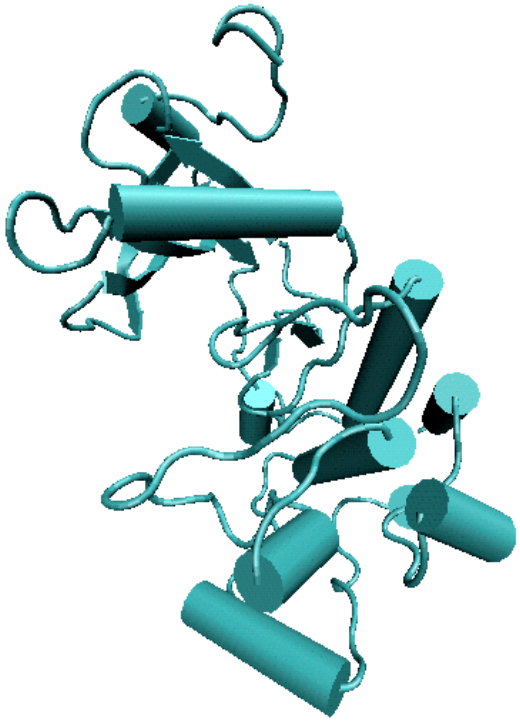
*Auto*correlation functions can be used to compute transport coefficients like diffusion, thermal conductivity, viscosities, e.g., velocity autocorrelation function $\rightarrow$ diffusion coefficient:

$$3D_i = \int_0^\infty \langle \vec{v}_i(t') \cdot \vec{v}_i(0)\rangle \mathrm{d}t'.$$

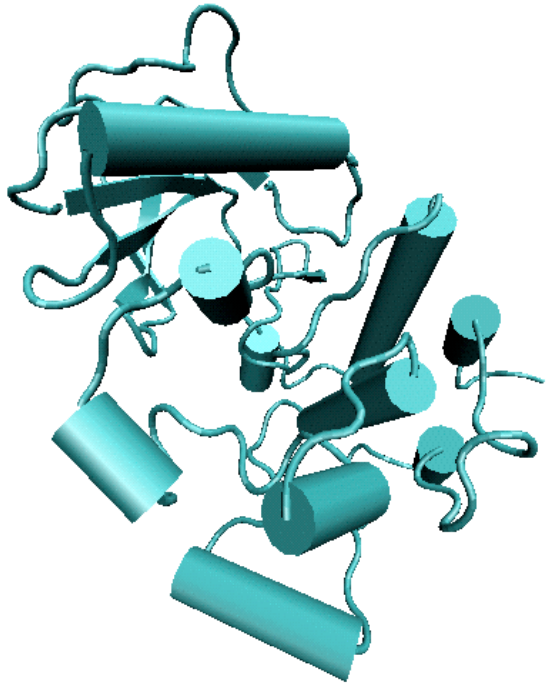
For greater efficiency, average over all indistinguishable atoms i.

Transition pathways

active kinase domain



inactive kinase domain



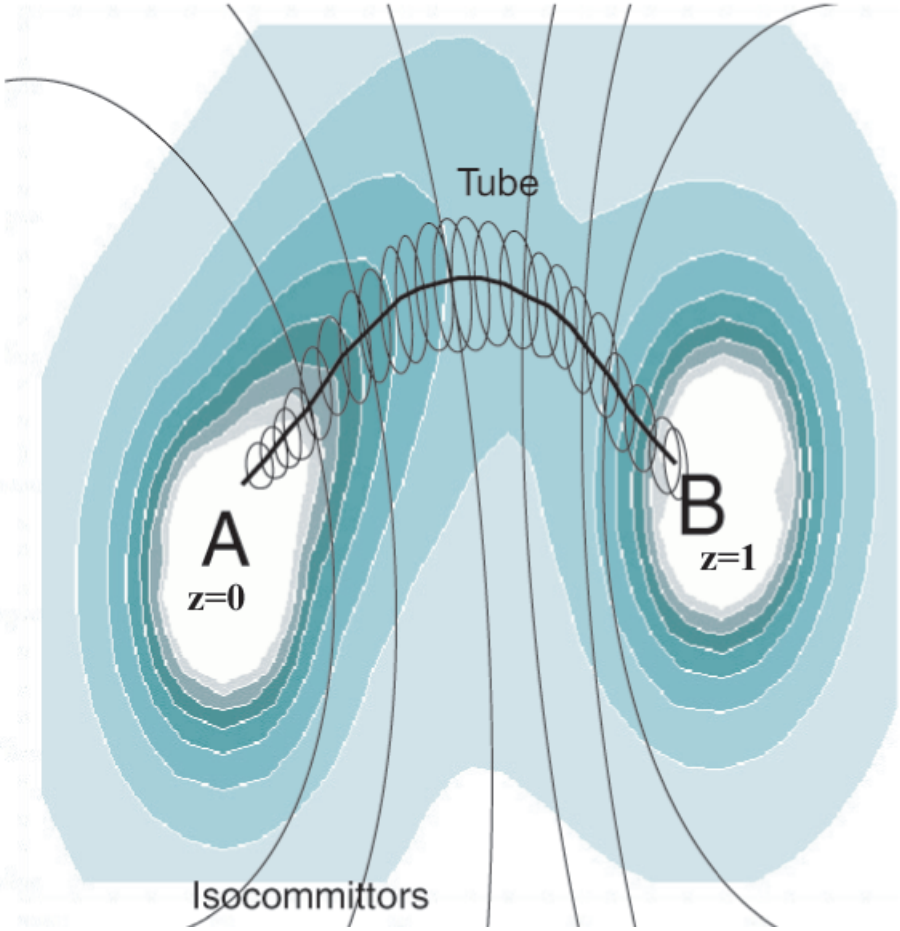
Defining a pathway

Problem is to calculate a "representative path"
from metastable state A in x-space to metastable state B.

committor function: $z(x) =$
$\Pr$(trajectory starting at x with random v reaches B before A).

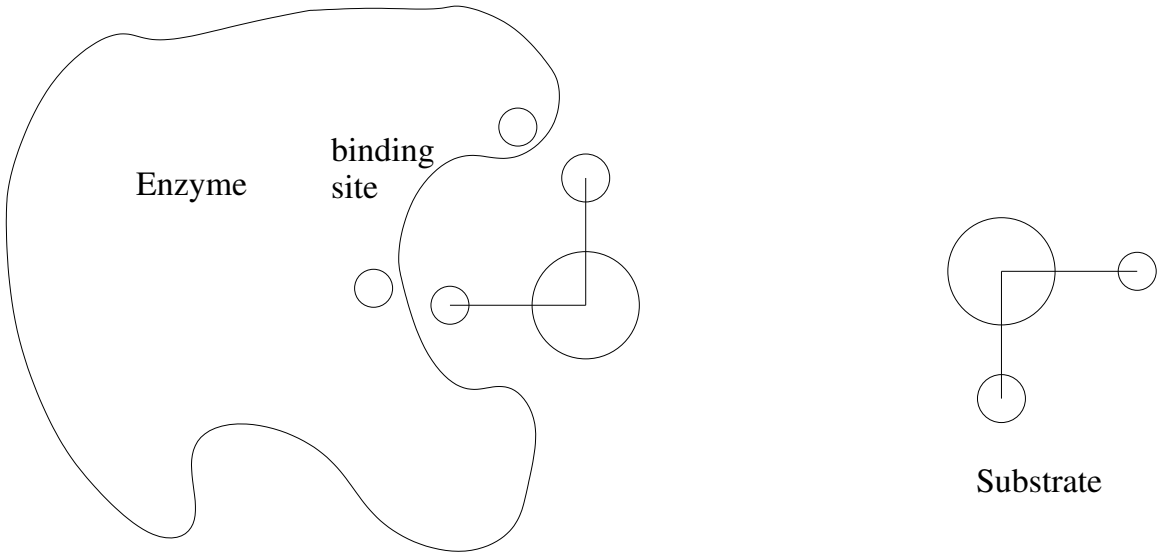
On each isocommittor consider
the distribution of crossing points from reactive trajectories.
Choose the "center" of this distribution to be representative.

This is illustrated in the following figure,



where shading indicates contours of potential energy, thin curves denote isocommittors, ellipses enclose concentrations of crossing points from reactive trajectories, and the thick curve is the center.

Diffusion-limited reactions



Problem: calculate rate constant k where

$$\begin{array}{c}\text{reaction rate}\\ \text{per unit volume}\end{array} = k \times \begin{array}{c}\text{substrate}\\ \text{concentration}\end{array} \times \begin{array}{c}\text{enzyme}\\ \text{concentration.}\end{array}$$

Outline

I. Equations of motion

II. Boundary effects

III. Initial conditions

IV. Computational tasks: thermodynamics and structure

V. Computational tasks: kinetics

VI. Practicalities

Practicalities

The practicalities of doing such calculations involve three steps:

structure building Setting up the input files is best done interactively with scripts and visual feedback.
visualization programs: RasMol, VMD, PyMOL, ...

simulation Generating dynamics or sampling trajectories is best done in background or remotely.
simulation programs: CHARMM, Amber, Gromacs, NAMD, LAMMPS, NWChem, Tinker, ...

analysis Analyzing trajectory data.

Simulation specifications

- Specify molecular system & surroundings
- Specify computational tasks
- Select computational model:
  uncontrolled approximations and error tolerances
  - internal forces
  - external forces, e.g., temperature and pressure control
  - dynamics (sampling or real)
- (Override defaults for performance parameters)
- Design simulation protocol

References

- M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, 1987,

- D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd edition, 2002.

- A. R. Leach, *Molecular Modelling: Principles and Applications*, 2nd edition, 2001,

- T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*, 2002,

- *Journal of Chemical Physics*