Biomolecular modeling II

Marcus Elstner and Tomáš Kubař

2015, December 16

System boundary and the solvent

Biomolecule in solution

typical MD simulations – molecular system in aqueous solution
task – make the system as small as possible (reduce cost)

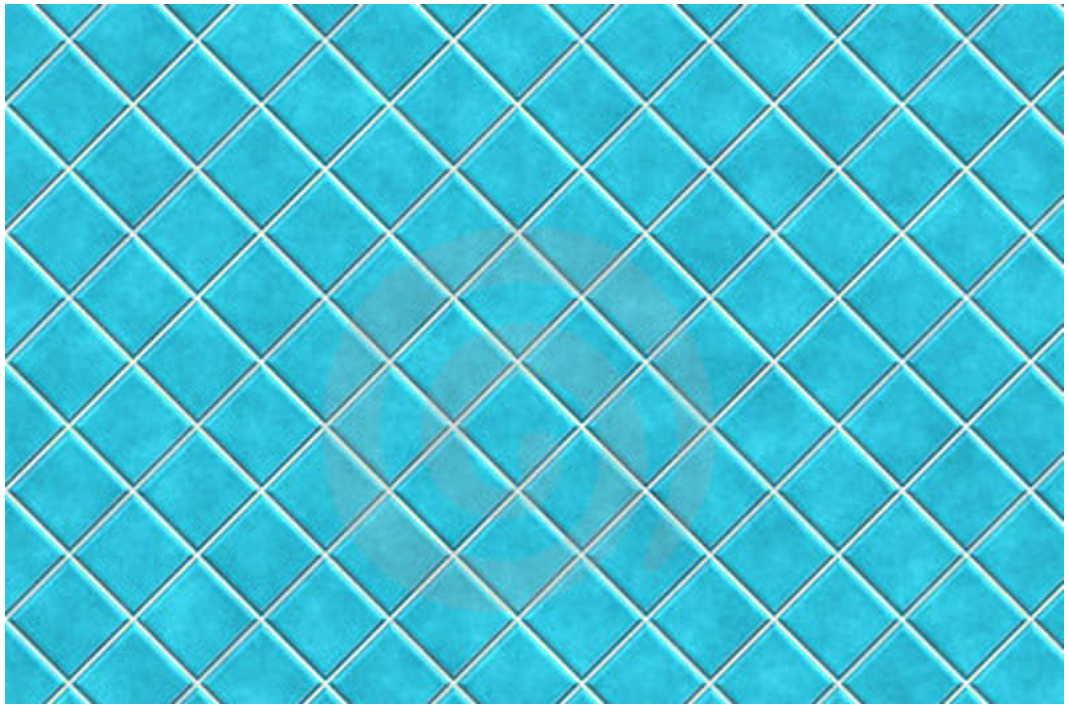
straightforward solution – single molecule of solute (protein, DNA)
with a smallest possible number of H_2O molecules
typical – several thousand H_2O molecules in a box $n \times n \times n$ nm

issue – everything is close to the surface,
while we are interested in a molecule in bulk solvent

Periodic boundary conditions

- elegant way to avoid these problems
- molecular system placed in a regular-shaped box
- the box is virtually replicated in all spatial directions

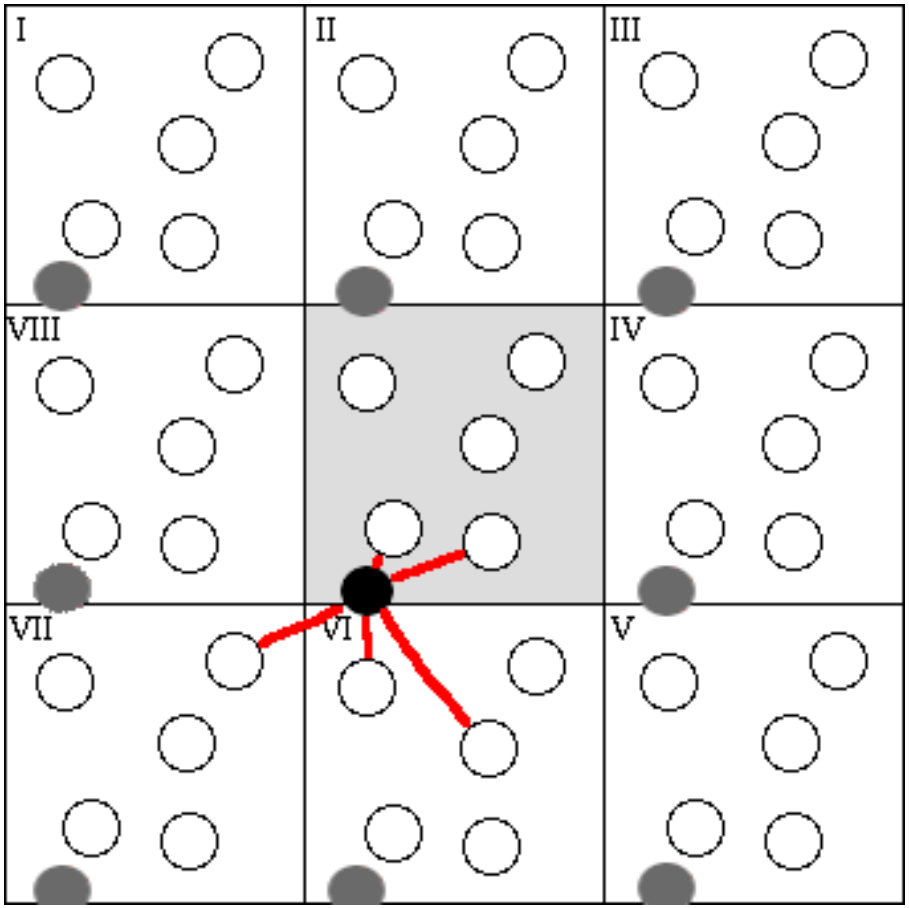
Periodic boundary conditions



Periodic boundary conditions

- elegant way to avoid these problems
- molecular system placed in a regular-shaped box
- the box is virtually replicated in all spatial directions
- positions (and velocities) of all particles are identical in all replicas, so that we can keep only one copy in the memory
- this way, the system is infinite – no surface!
- the atoms near the wall of the simulation cell interact with the atoms in the neighboring replica

Periodic boundary conditions

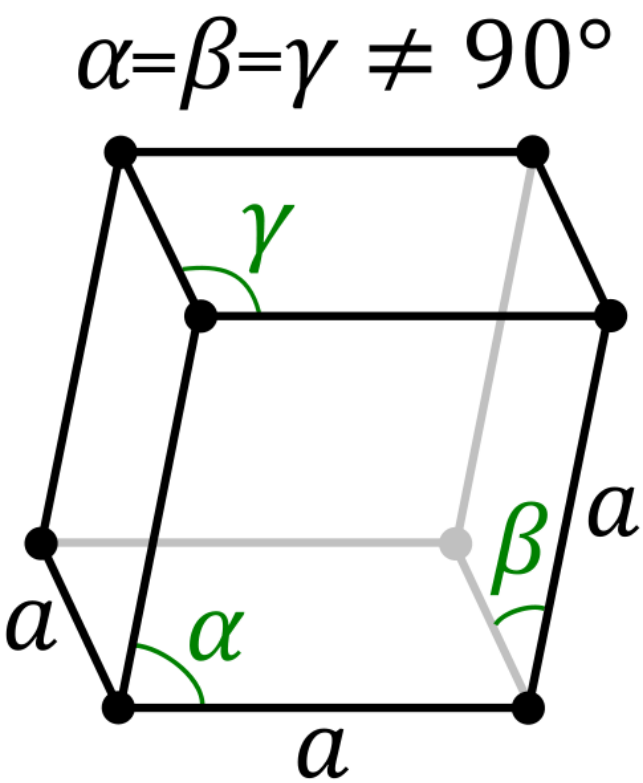


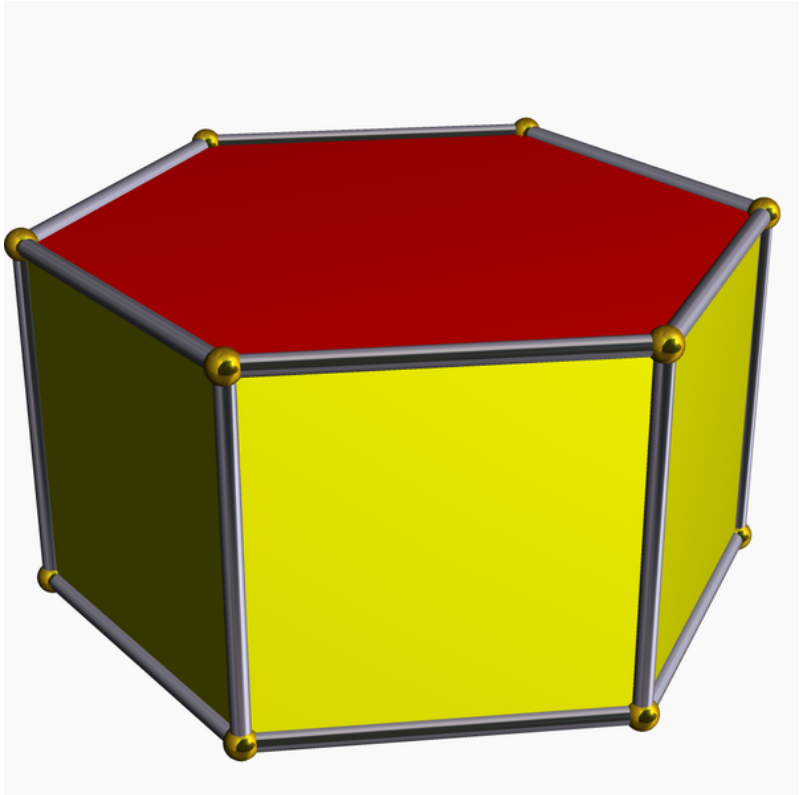
PBC – features

- small problem – artificial periodicity in the system (entropy ☹)
  – still much better than boundary with vacuum
- only coordinates of the unit cell are recorded
- atom that leaves the box enters it on the other side.
- carefull accounting of the interactions of atoms necessary!
  simplest – minimum image convention:
  an atom interacts with the nearest copy of every other
  – interaction with two different images of another atom,
  or even with another image of itself is avoided

PBC – box shape

may be simple – cubic or orthorhombic, parallelepiped (specially, rhombohedron), or hexagonal prism



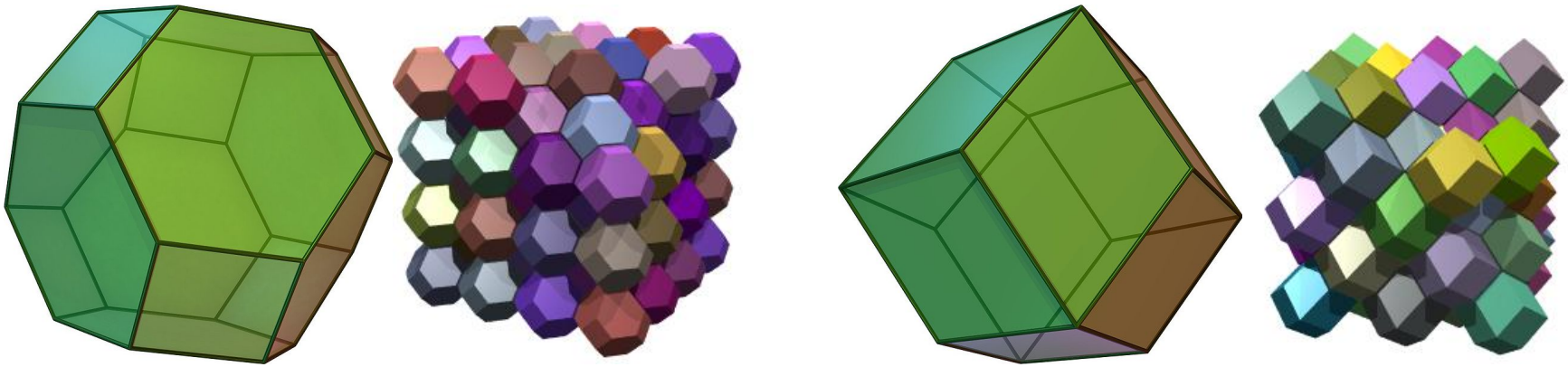


PBC – box shape

. . . but also more complicated

– truncated octahedral or rhombic dodecahedral

– quite complex equations for interactions & eqns of motion

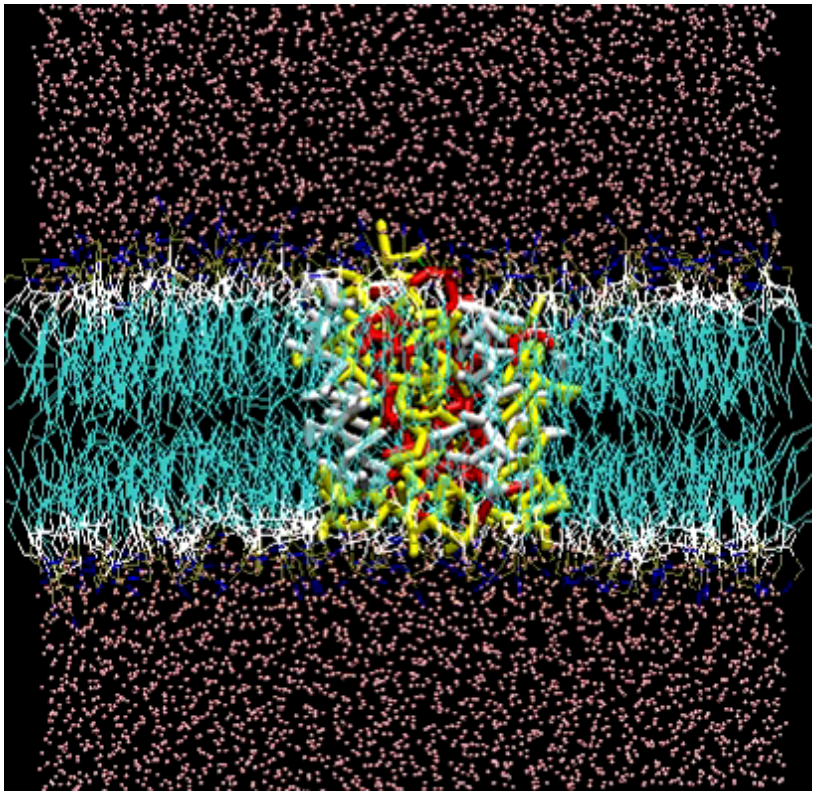


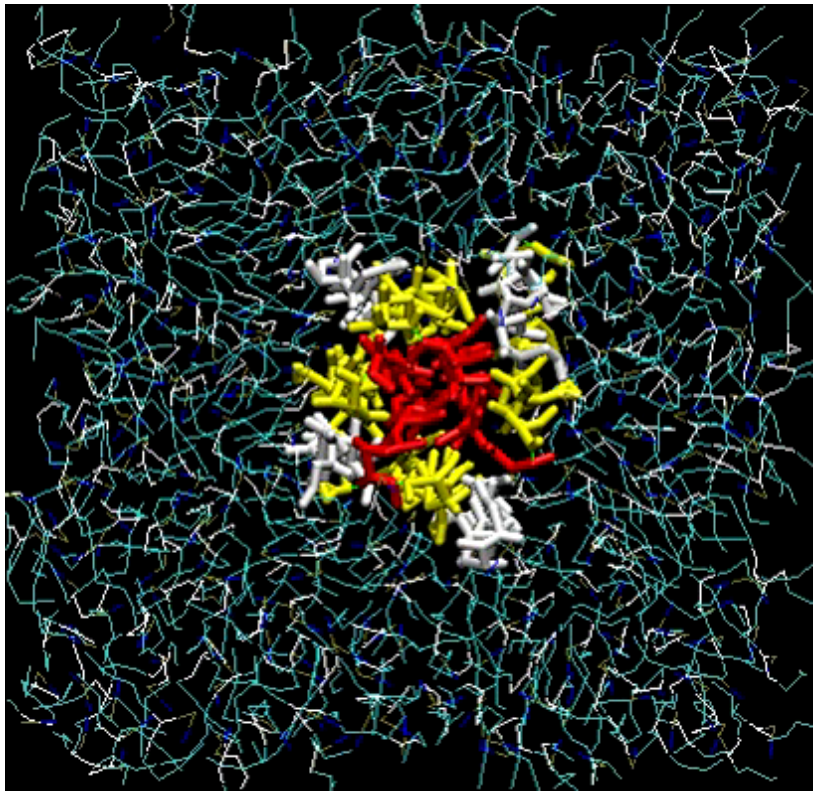
advantage for simulation of spherical objects (globular proteins)

– no corners far from the molecule filled with unnecessary H_2O

PBC – box shape

2D objects – phase interfaces, membrane systems
– usually treated in a slab geometry

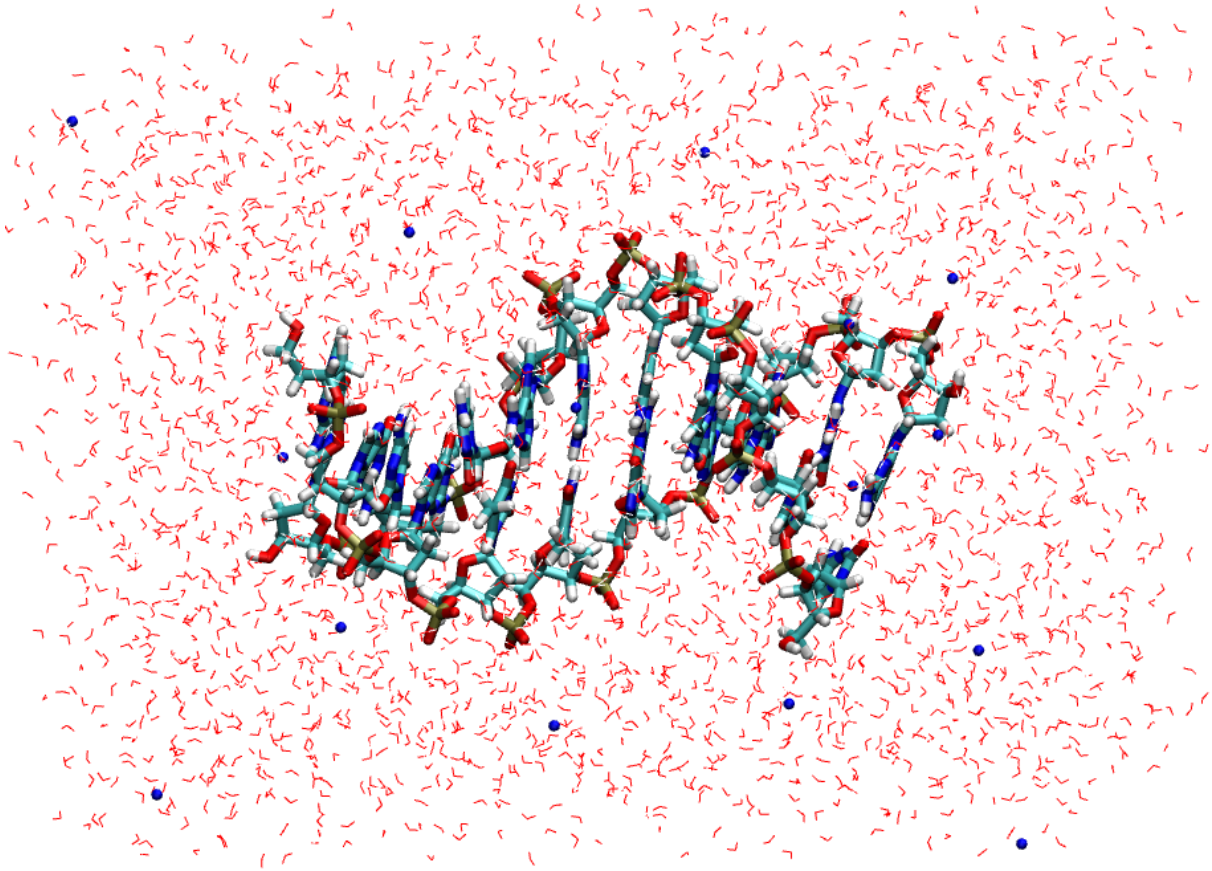




Water in biomolecular simulations

most simulations – something in aqueous solutions

H_2O – usually (many) thousands molecules



Water in biomolecular simulations

most simulations – something in aqueous solutions

H_2O – usually (many) thousand molecules

example – simulation of DNA decanucleotide:

- PBC box $3.9 \times 4.1 \times 5.6$ nm (smallest meaningful)
- 630 atoms in DNA, 8346 atoms in water and 18 Na^+
- concentration of DNA: 18 mmol/L – very high!
- of all pair interactions: 86 % are water–water,
  most of the others involve water

Water models

most interactions involve H_2O

$\rightarrow$ necessary to pay attention to its description

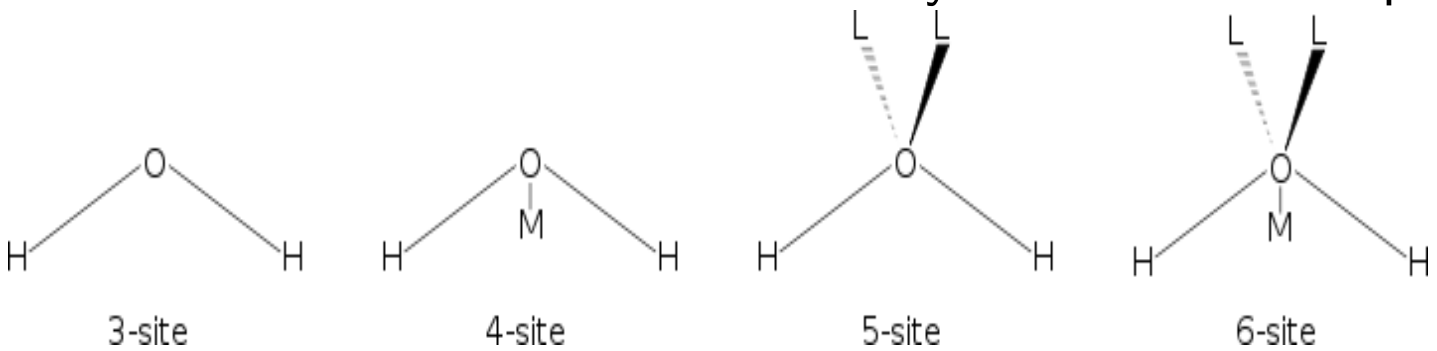
model of water must be simple enough (computational cost)
and accurate enough, at the same time

water models – usually rigid

– bond lengths and angles do not vary – constraints

molecule with three sites (atoms in this case), or up to six sites

– three atoms and virtual sites corresponding
to a 'center' of electron density or lone electron pairs



Water models

TIP3P (or SPC)

- most frequently used
- 3 atoms with 3 rigid bonds, charge on every atom ($-0.834/+0.417$)
- only the O possesses non-zero LJ parameters (optimization)

TIP4P

- negative charge placed on virtual site M rather than on the O
- electric field around the molecule described better

TIP5P

- 2 virtual sites L with negative charges near the O – lone pairs
- better description of directionality of H-bonding etc. (radial distribution function, temperature of highest density)

Non-bonded interactions

speeding up the number-crunching

Non-bonded interactions – why care?

- key to understand biomolecular structure and function
  - binding of a ligand
  - efficiency of a reaction
  - color of a chromophore
- two-body potentials $\rightarrow$ computational effort of $\mathcal{O}(N^2)$
  - good target of optimization
- solvent (H_2O) – crucial role, huge amount
  - efficient description needed

Non-bonded interactions

electrostatic interaction energy of two atoms with charges q_1 and q_2 on distance r:

$$E^{\mathrm{el}}(r) = \frac{1}{4\pi\varepsilon_0} \cdot \frac{q_1 \cdot q_2}{r}$$

Lennard-Jones interaction energy of two atoms:

$$E^{\mathrm{LJ}}(r) = 4E_0 \left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right)$$

Cut-off – simple idea

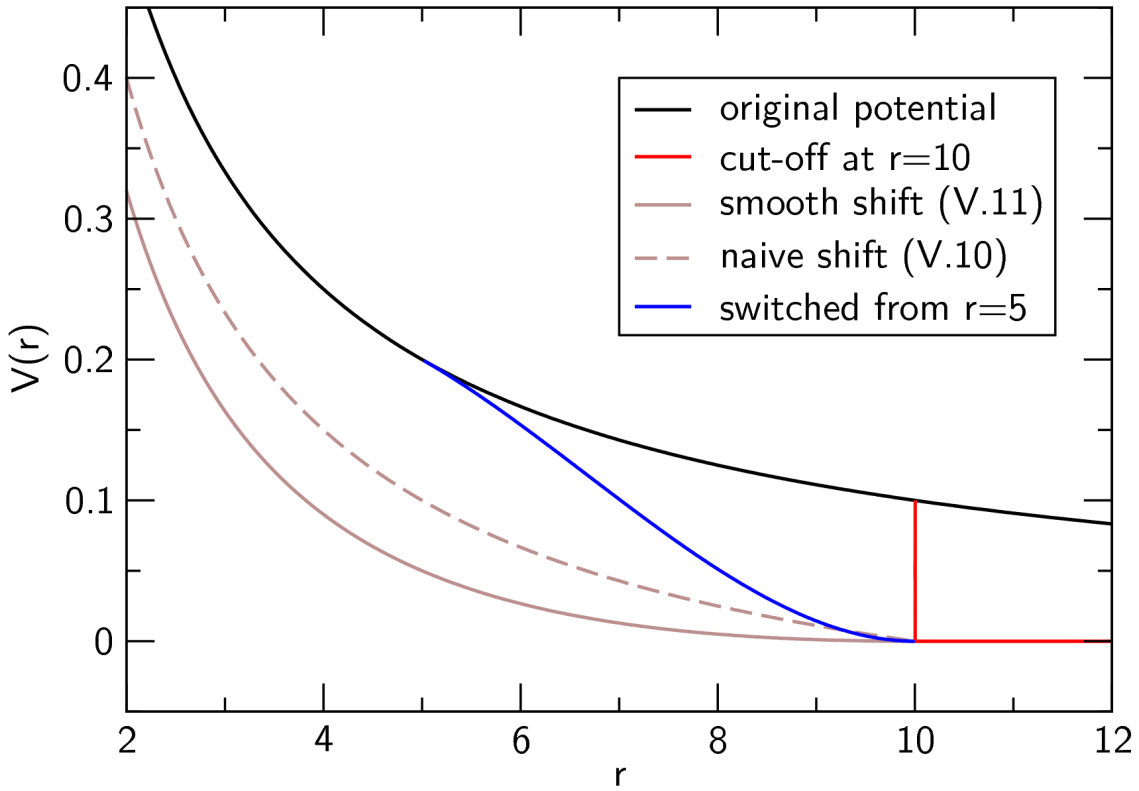
with PBC – infinite number of interaction pairs in principle,
but the interaction gets weaker with distance

simplest and crudest approach to limit the number of calculations
neglect interaction of atoms further apart than r_c – cut-off

very good for rapidly decaying LJ interaction $(1/r^6)$ ($r_c = 10$ Å)

not so good for slowly decaying electrostatics $(1/r)$
– sudden jump (discontinuity) of potential energy,
disaster for forces at the cut-off distance

Cut-off – better alternatives



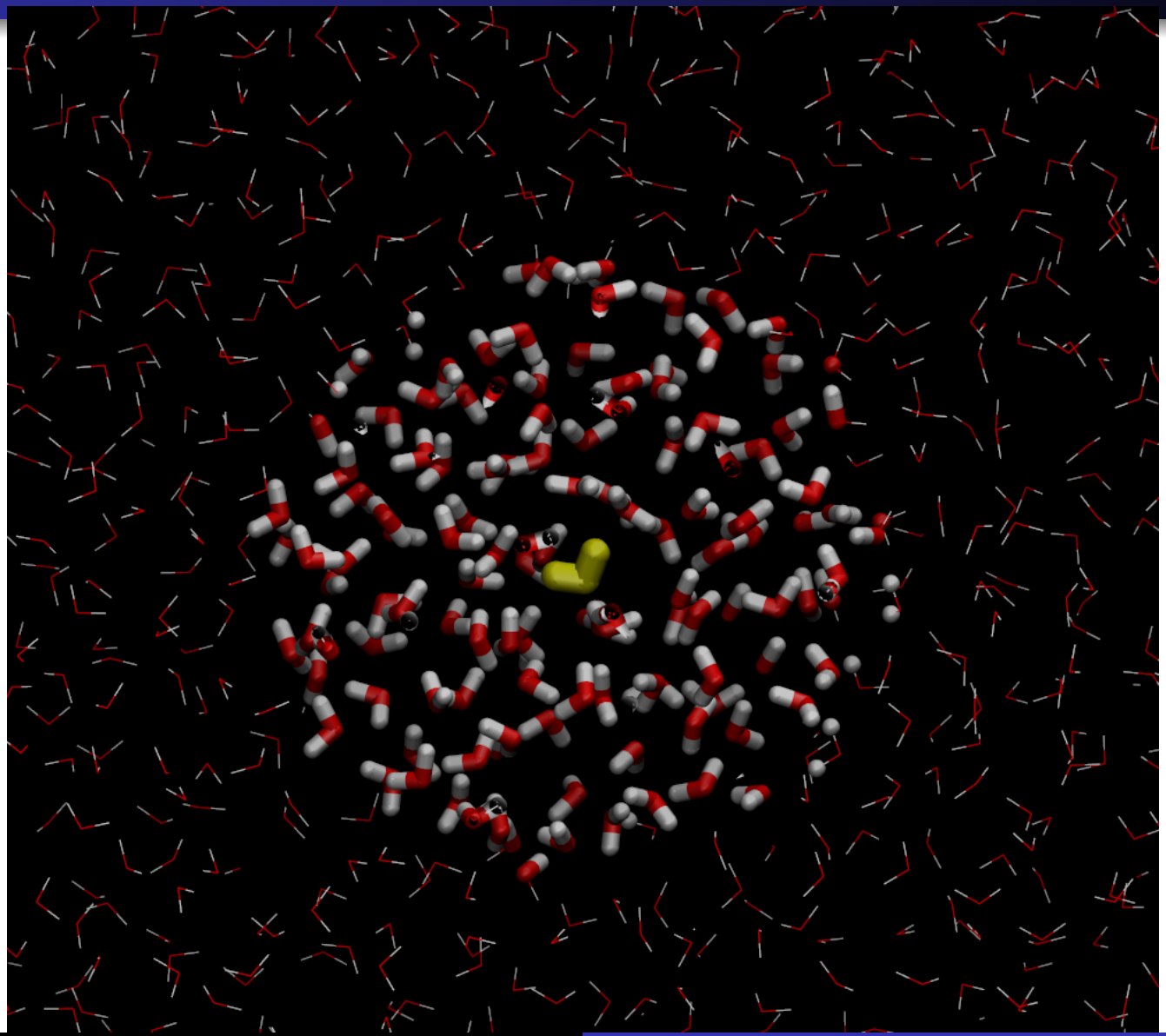
Neighbor lists

cut-off – we still have to calculate the distance for every two atoms
(to compare it with the cut-off distance)
$\rightarrow$ we do not win much yet – there are still $\mathcal{O}(N^2)$ distances

observation: pick an atom A.
the atoms that are within cut-off distance r_c around A,
remain within r_c for several consecutive steps of dynamics,
while no other atoms approach A that close

idea: maybe it is only necessary to calculate the interactions
between A and these close atoms – neighbors

Neighbor lists



Neighbor lists

what will we do? calculate the distances for every pair of atoms
less frequently, i.e. every 10 or 20 steps of dynamics, and
record the atoms within cut-off distance in a neighbor list

atom	how many?	list of neigboring atoms											
1	378	2191	408	1114	1802	262	872	649	805	1896	2683	114	189
2	403	1788	1624	1048	1745	2546	506	203	288	2618	1445	880	133
3	385	779	2869	800	2246	1252	570	454	1615	1656	1912	2395	152
4	399	367	2143	1392	1448	1460	1411	2921	2725	429	845	2601	181
5	406	1385	425	1178	2112	1689	1897	1650	1747	1028	1366	605	176
6	388	1748	130	2244	631	1677	1748	2566	303	552	562	1142	255
7	379	20	15	1322	196	1590	655	552	1401	2177	411	2904	236
8	395	888	1074	786	2132	1703	218	1846	337	1683	1917	2005	94
9	396	2433	934	1055	1518	2750	2534	1697	2006	769	2407	1478	123
10	381	2461	1910	459	2628	2523	1709	2069	1151	1710	2107	1909	13
11	400	1029	756	670	1592	612	676	1473	2859	392	986	155	265

then – calculate the interaction for each atom
only with for the atoms in the neighbor list – formally $\mathcal{O}(N)$

Accounting of all of the replicas

cut-off – often bad, e.g. with highly charged systems
(DNA, some proteins)

switching function – deforms the forces (slightly)
$\rightarrow$ e.g. artificial accumulation of ions around cut-off

only way – abandon the minimum image convention and cut-off
– sum up the long-range Coulomb interaction
between all the replicas of the simulation cell

Accounting of all of the replicas

infinite number of atoms – some tricks are needed:

Ewald summation method $\mathcal{O}(N^{\frac{3}{2}})$ or even particle–mesh Ewald method, $\mathcal{O}(N \cdot \log N)$

2 main contributions:

- 'real-space' – similar to the usual Coulomb law, but decreasing much quicker with distance
- 'reciprocal-space' – here are the tricks concentrated
  - atom charges artificially smeared (Gaussian densities)
  - Fourier transformation can sum up the interaction of all of the periodic images!

Ewald – realistic simulations of highly charged systems possible

Preparing an MD simulation

the procedures – briefly

Work plan

1. build the initial structure
2. bring the system into equilibrium
3. do the productive simulation
4. analyze the trajectory

Tools to build the structure

- do it yourself
- specific programs within simulation packages
- 'universal' visualization programs – VMD, Molden, Pymol

Tools to build the structure

- do it yourself
- specific programs within simulation packages
- 'universal' visualization programs – VMD, Molden, Pymol
- databases of biomolecular systems – PDB, NDB

Tools to build the structure

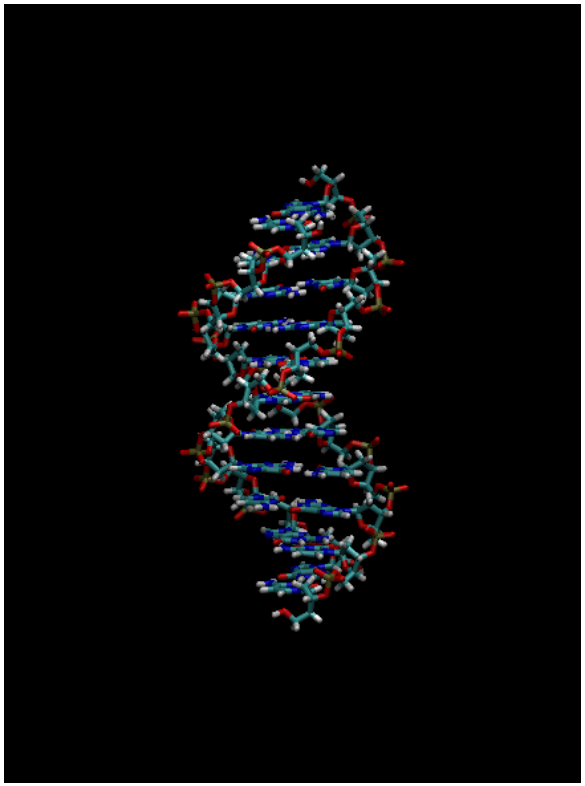
- do it yourself
- specific programs within simulation packages
- 'universal' visualization programs – VMD, Molden, Pymol
- databases of biomolecular systems – PDB, NDB
- specialized web services – Make-NA

Tools to build the structure

- do it yourself
- specific programs within simulation packages
- 'universal' visualization programs – VMD, Molden, Pymol
- databases of biomolecular systems – PDB, NDB
- specialized web services – Make-NA
- tools to create periodic box and hydrate system

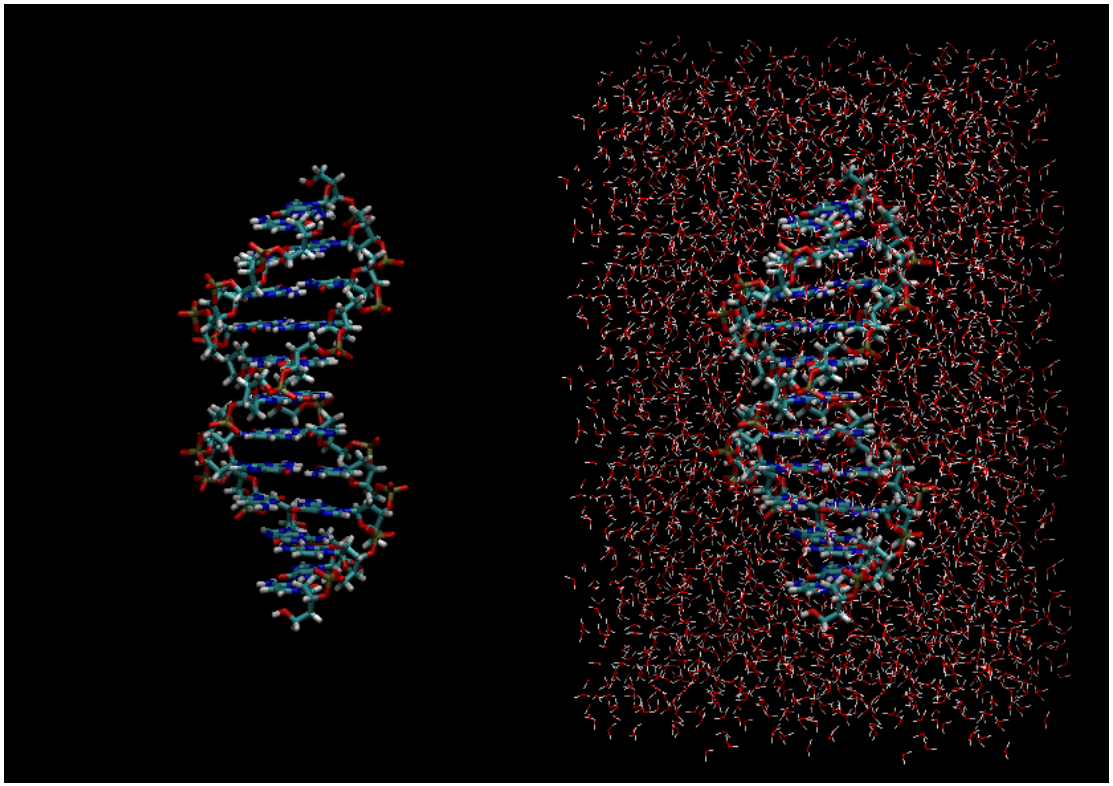
Tools to build the structure

build the solute, solvate it and add counterions



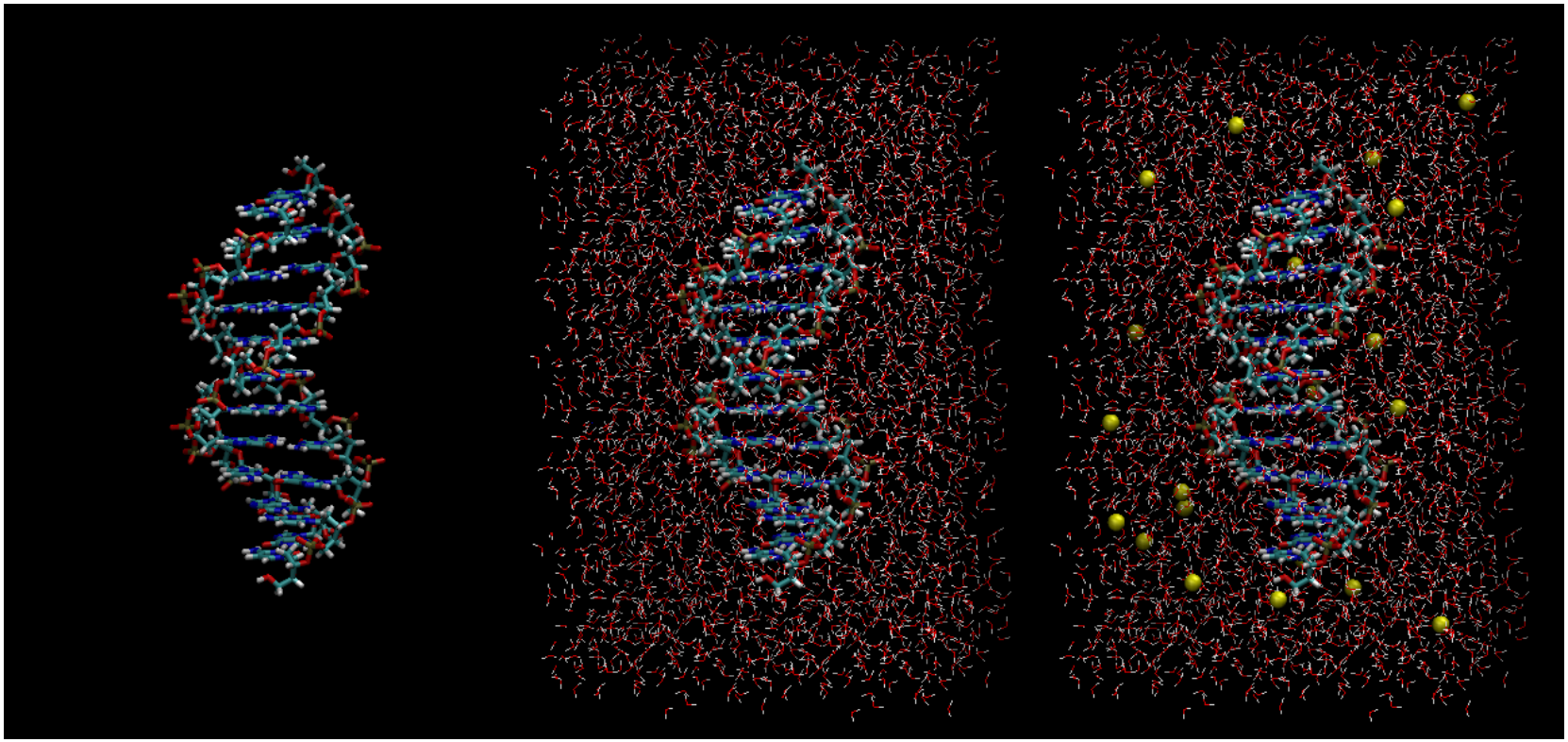
Tools to build the structure

build the solute, solvate it and add counterions



Tools to build the structure

build the solute, solvate it and add counterions



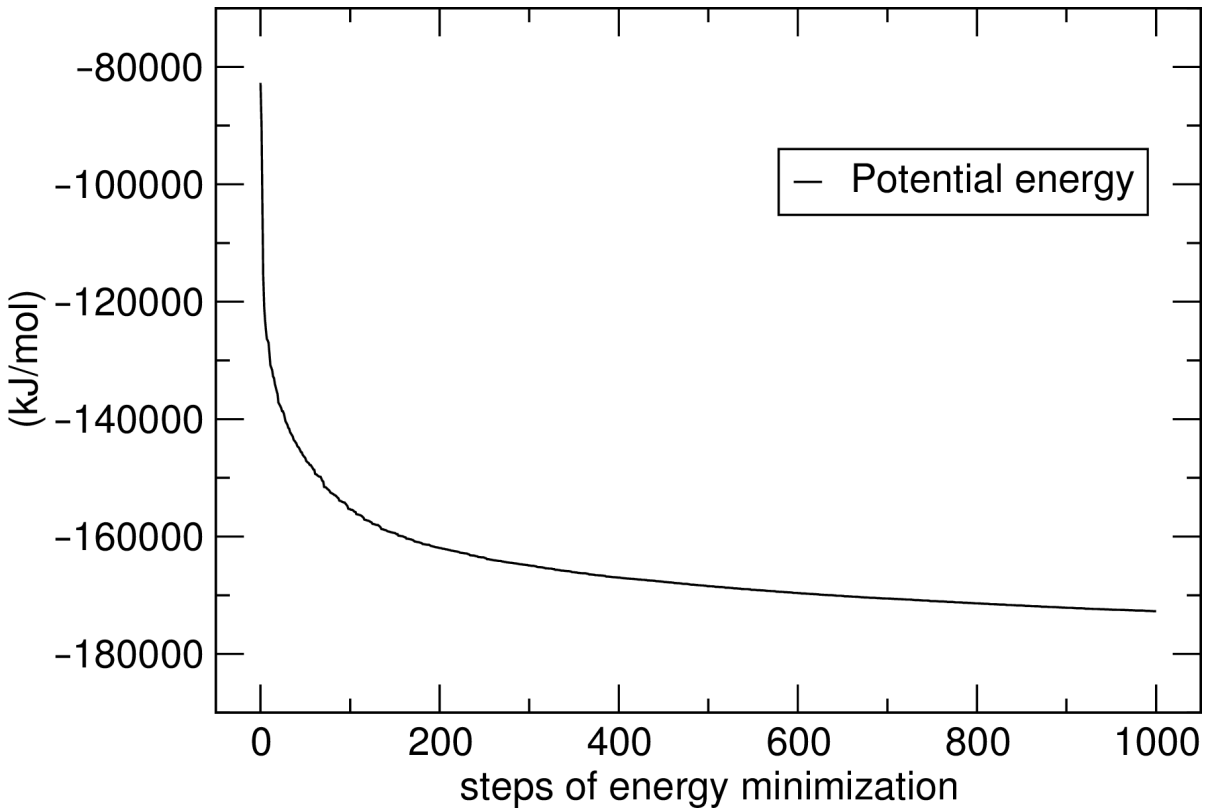
Why equilibrate?

- the initial structure may have high potential energy – dangerous – remove 'close contacts'
- often, static structure available – velocities missing
- often, structure resolved at different conditions (xtal)
- structure of solvent artificially regular – entropy wrong

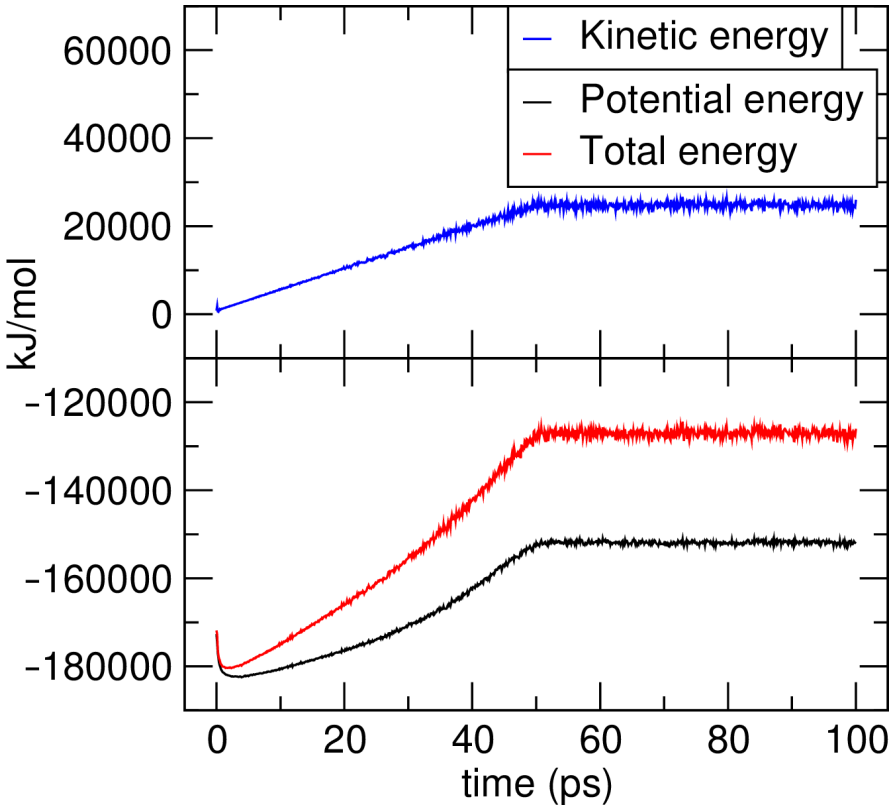
How to equilibrate

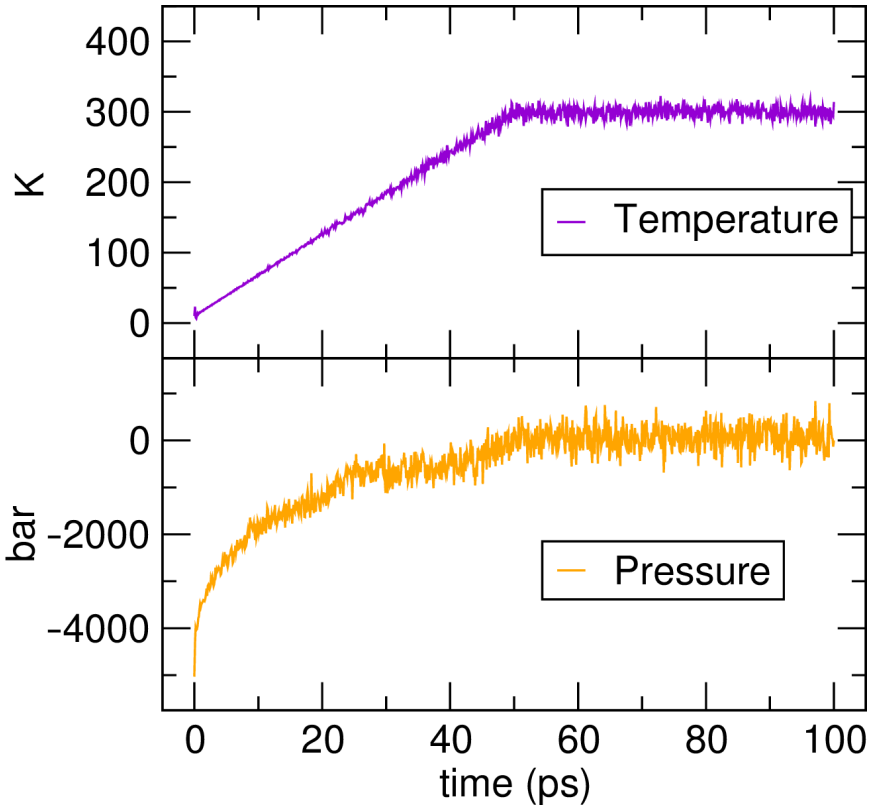
1. short optimization of structure – remove 'bad contacts'
2. assignment of velocities – randomly, at some (low) T
3. thermalization – heating the system up to the desired T, possibly gradually, with a thermostat – NVT simulation
4. simulation with the same setup as the production – probably NPT, with correct thermostat and barostat

Short optimization



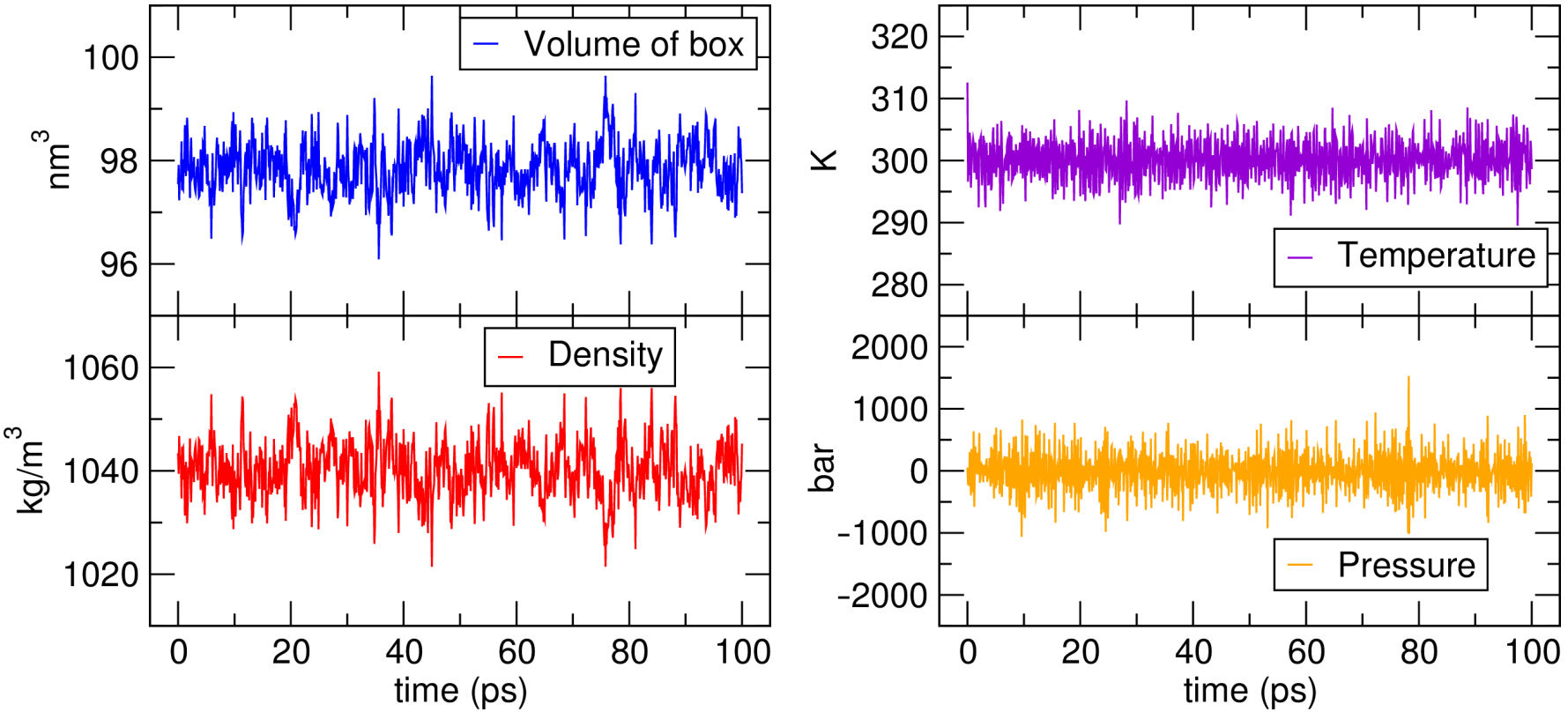
Thermalization





last 40 ps: $T = 300 \pm 7$ K, $p = 64 \pm 266$ bar

Equilibration



last 40 ps: $T = 300 \pm 3$ K, $p = -11 \pm 331$ bar

What comes then?

Productive simulation
- easy ☺

Analysis of the trajectory
- let us see...

Analysis of the simulation

Thermodynamic properties

- time averages of thermodynamic quantites
  – correspond to ensemble averages (ergodic theorem)
- some quantities – evaluated directly

$$U = \langle E \rangle_t$$

- fluctuations – may determine interesting properties:
  isochoric heat capacity:

$$C_V = \left(\frac{\partial U}{\partial T} \right)_V = \frac{\sigma_E^2}{k_B T^2} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2}$$

  – elegant way from single simulation to heat capacity

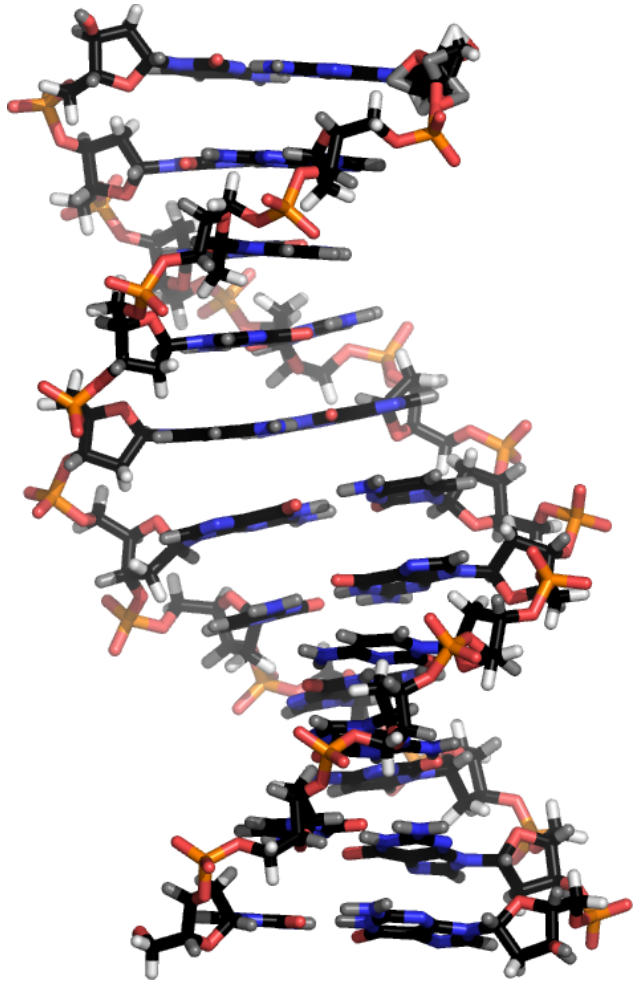
Structure – single molecule in solvent

concetrating on the dissolved molecule
- protein, DNA,...

average structure
- arithmetic mean of coordinates from snapshots along MD trajectory

$$\vec{r}_i = \frac{1}{N} \sum_{n=1}^{N} \vec{r}_i^{(n)}$$

- clear, simple, often reasonable



Average structure

Possible problems:

- freely rotatable single bonds – CH_3
  - all 3 hydrogens collapse to a single point
  - no problem – ignore hydrogens
- rotation of the entire molecule – no big issue
  - RMSD fitting of every snapshot to the starting structure

    what is RMSD? see on the next slide...
- molecule does not oscillate around a single structure
  - several available minima of free energy
  - possibly averaging over multiple sections of trajectory

Dynamic information

root mean square deviation (RMSD)
of structure in time t
from a suitable reference structure $\vec{r}^{\,\mathrm{ref}}$

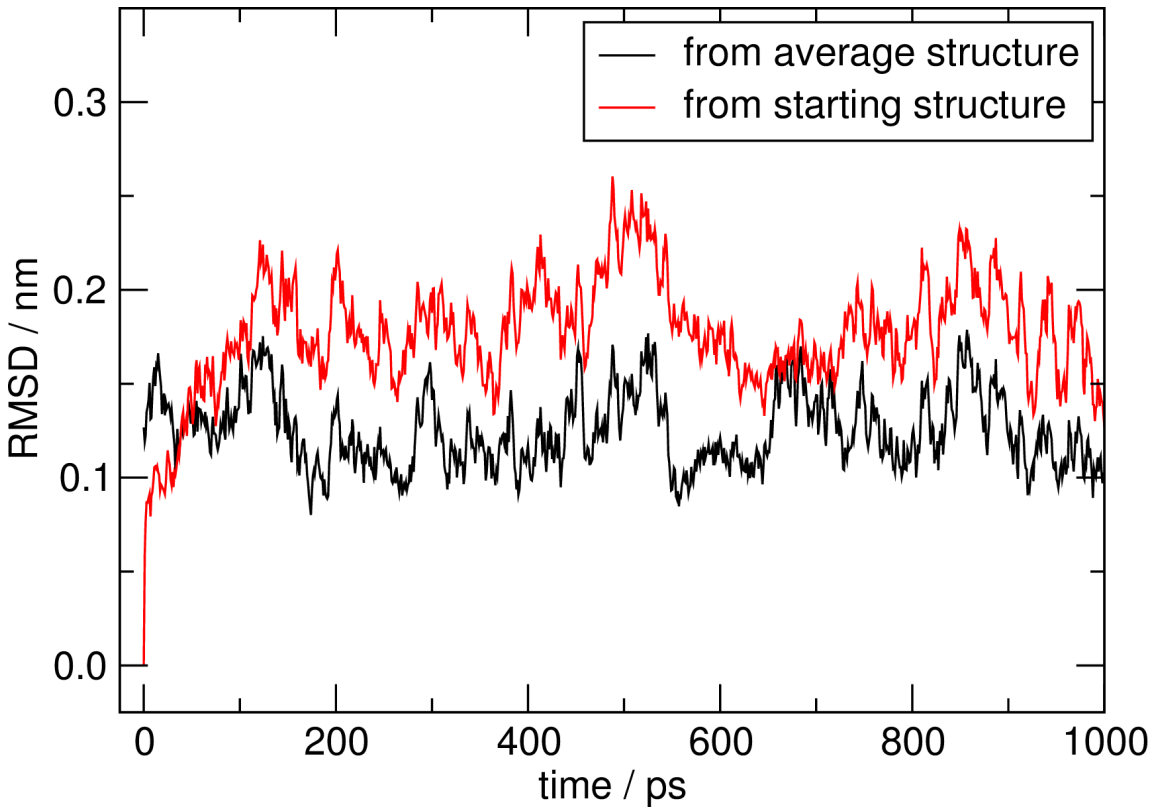
$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\vec{r}_i(t) - \vec{r}_i^{\,\mathrm{ref}}\right|^2}$$

- follows the development of structure in time
- reference structure – starting or average geometry
- also possible – comparison with another geometry of interest
  DNA: A- and B-like; proteins: α-helix and extended β

RMSD fitting – finding such a translation + rotation
that minimizes the RMSD from the reference structure

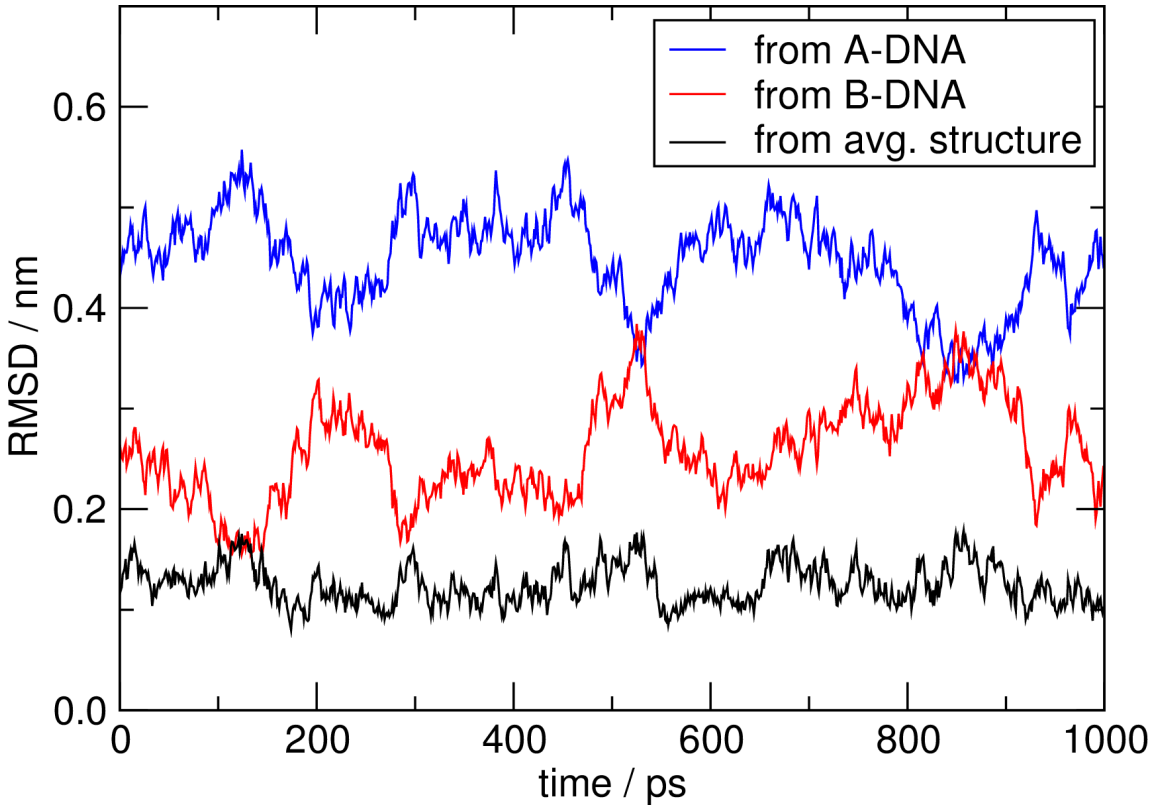
Root mean square deviation

RMSD of non-hydrogen atoms of a DNA oligonucleotide from given geometries



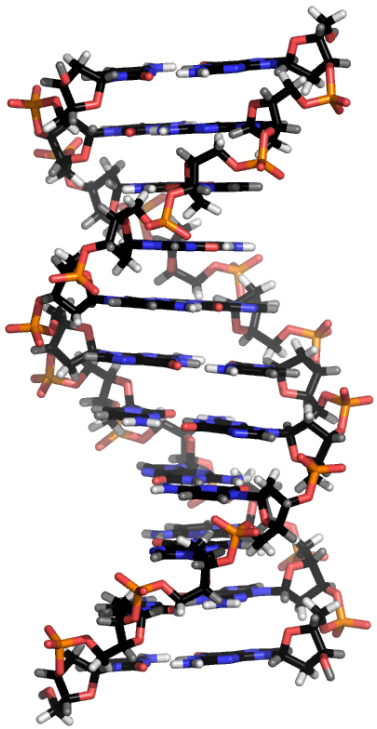
Root mean square deviation

RMSD of non-hydrogen atoms of a DNA oligonucleotide from given geometries

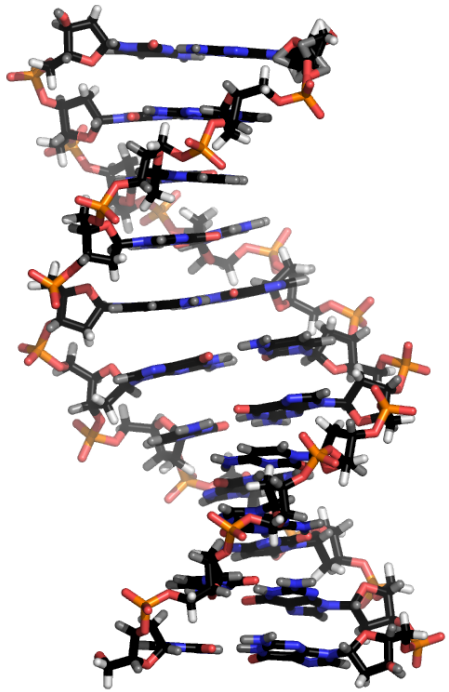


Root mean square deviation

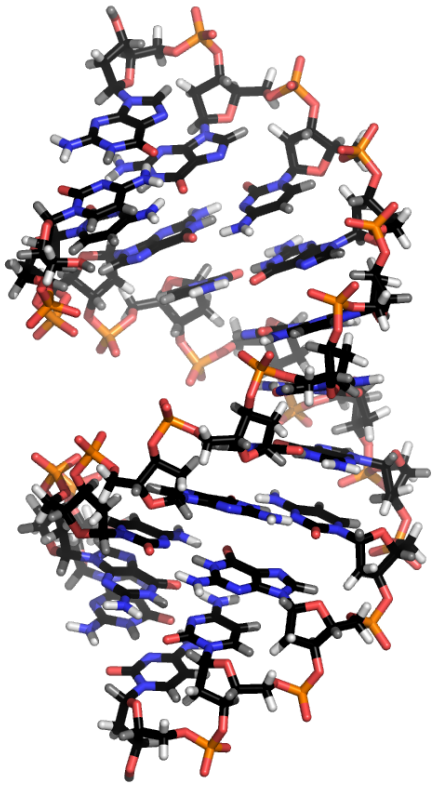
B-DNA



average structure



A-DNA



Magnitude of structural fluctuation

root mean square fluctuation (RMSF)
of position of every single atom
averaged along MD trajectory

$$\mathrm{RMSF}_i = \sqrt{\left\langle \left| \vec{r}_i - \langle \vec{r}_i \rangle \right|^2 \right\rangle}$$

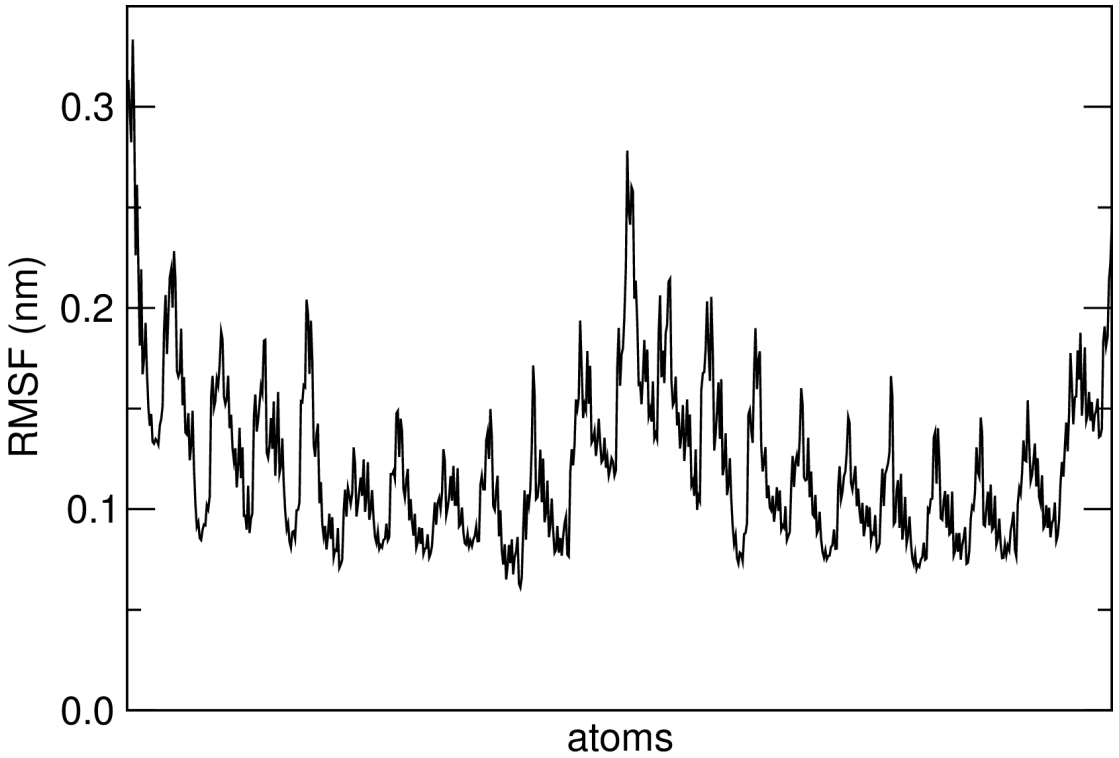
– may be converted to B-factor

$$B_i = \frac{8}{3}\pi^2 \cdot \mathrm{RMSF}_i^2$$

– observable in diffraction experiments (X-ray. . .)
– contained in structure files deposited in the PDB
– comparison of simulation with X-ray may be difficult

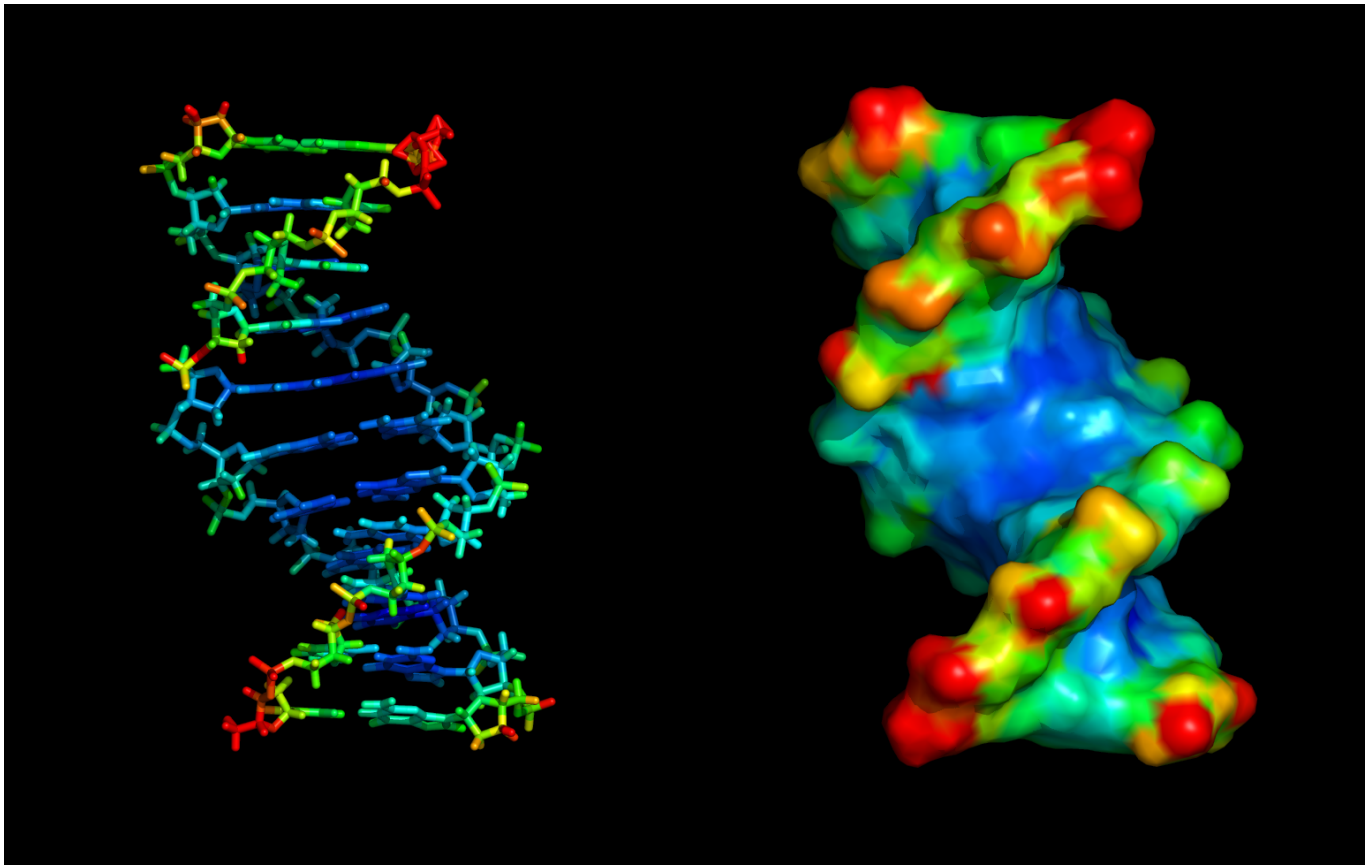
Root mean square fluctuation

RMSF of atomic positions in DNA oligonucleotide



Root mean square fluctuation

RMSF of atomic positions in DNA oligonucleotide

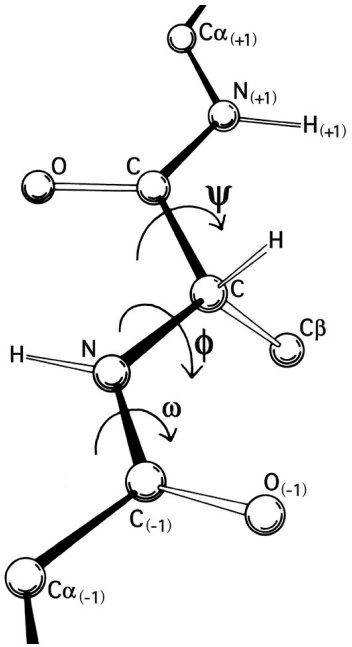


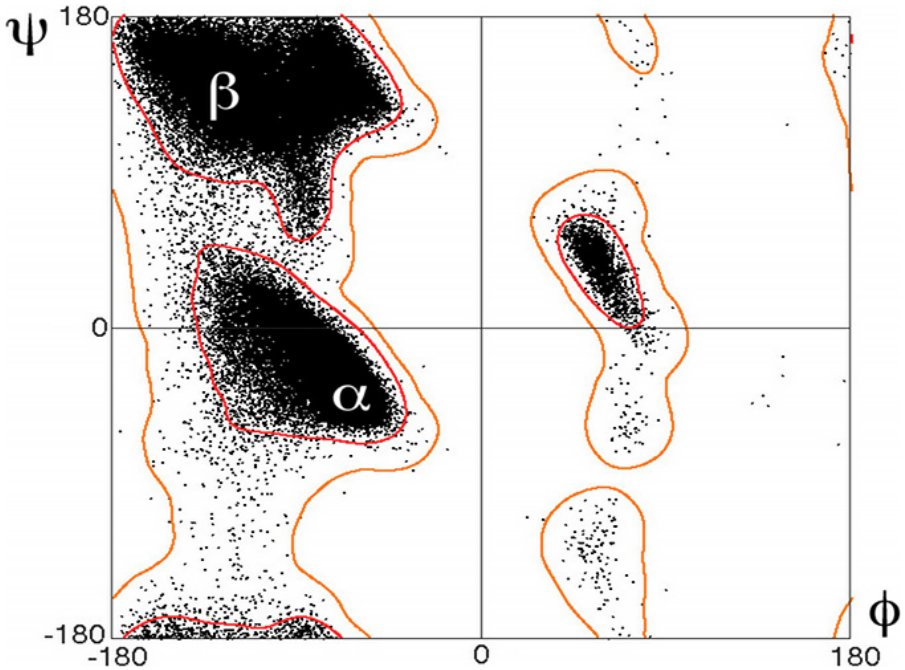
(blue $<$ green $<$ yellow $<$ red)

Structure of peptides and proteins

Ramachandran plot

- 2D histogram of dihedrals ϕ and ψ along the backbone
- different regions correspond to various second. structures
- may be generated easily in simulation software packages

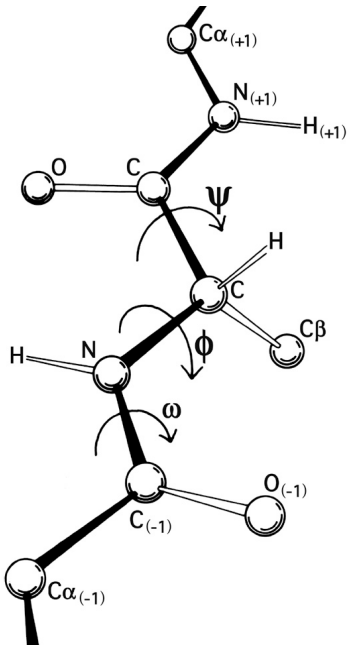


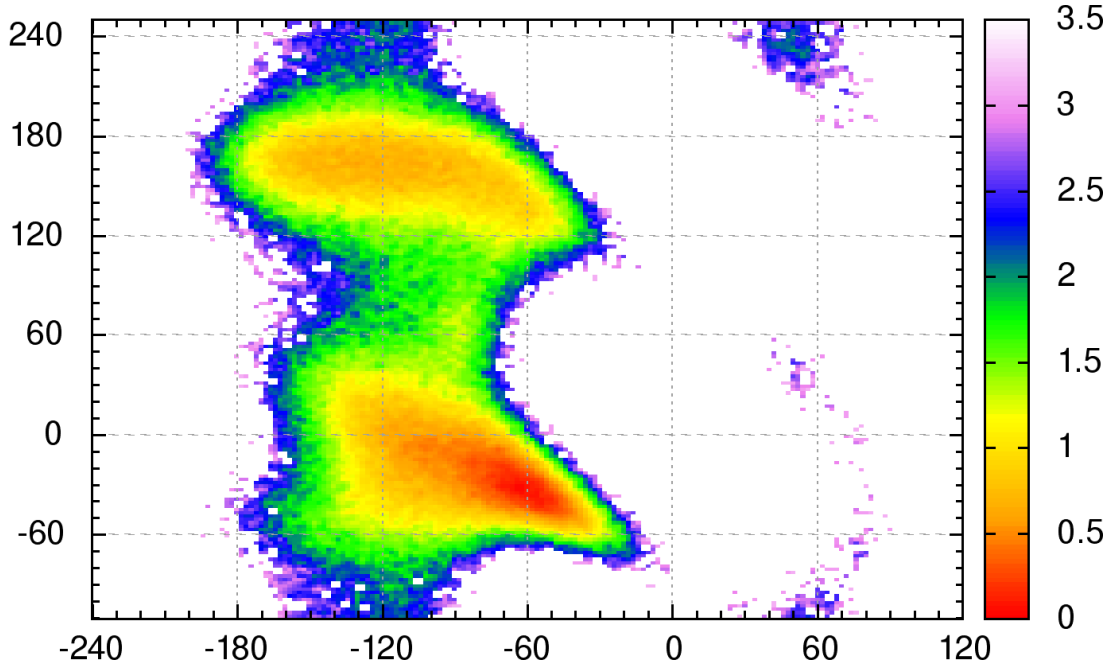


Structure of peptides and proteins

Ramachandran plot

- 2D histogram of dihedrals ϕ and ψ along the backbone
- different regions correspond to various second. structures
- may be generated easily in simulation software packages

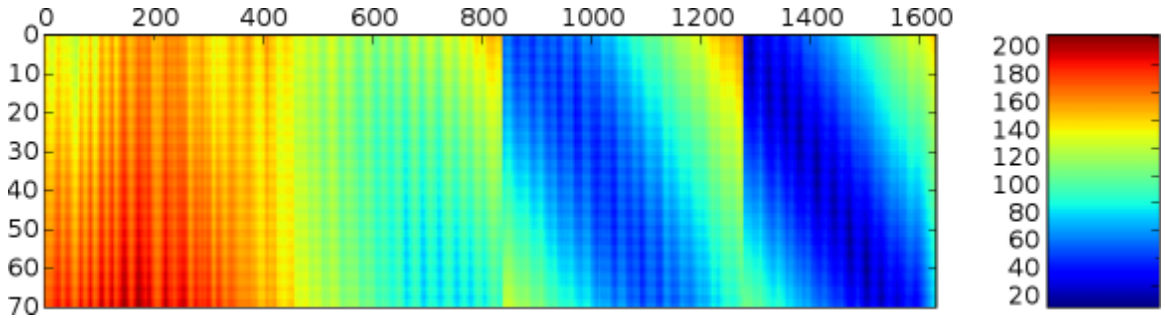


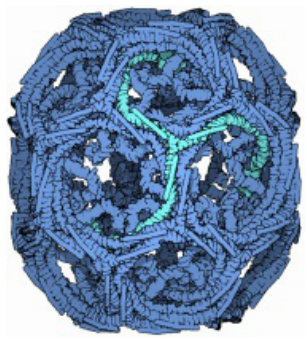


Structure of peptides and proteins

Distance matrix

- distances of amino-acid residues, represented e.g. by centers of mass or by C^{α} atoms
- either time-dependent or averaged over trajectory
- bioinformatics





distance matrix between two chains (horiz. and vertical axes)
shows contacts between secondary structure elements

PDB ID 1XI4, clathrin cage lattice, April 2007 Molecule of the Month

http://www2.warwick.ac.uk/fac/sci/moac/people/students/peter_cock/python/protein_contact_map

Structure of fluids

example – pure argon or water – different situation
– many molecules, which are all equally important

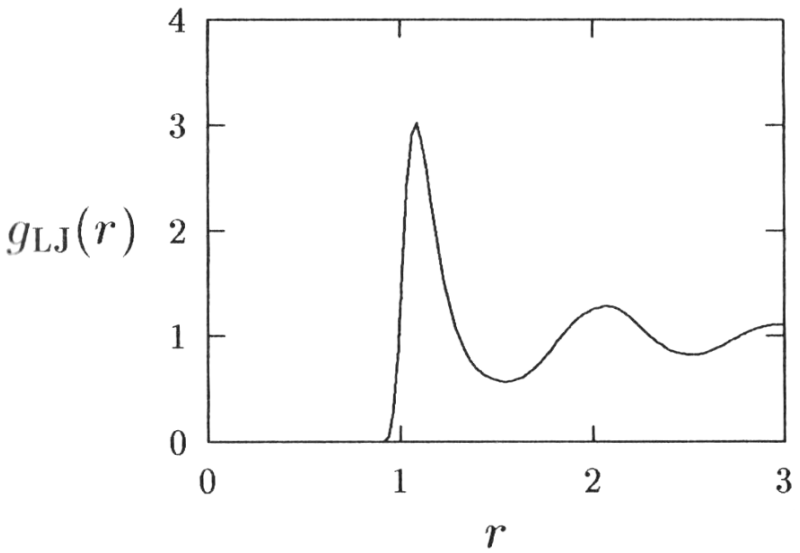
radial distribution functions

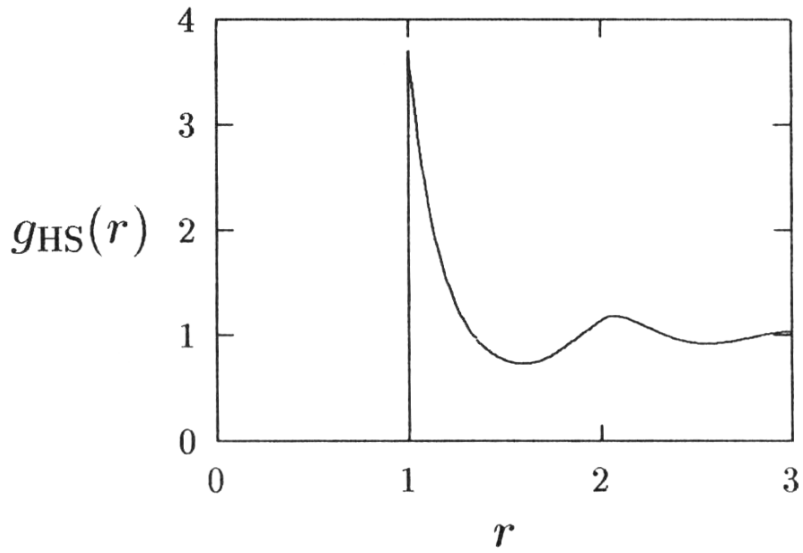
- describe how the molecular density varies
  as a function of the distance from one particular molecule
- spherical shell of thickness δr at a distance r: $\delta V \approx 4\pi r^2 \cdot \delta r$
- count the number of molecules in this shell: n
- divide by δV to obtain a 'local density' at distance r
- pair distribution function
  – probability to find a molecule in distance r from ref. mol.

$$g(r) = \frac{n/\delta V}{\rho} = \frac{n}{4\pi r^2 \cdot \delta r} \cdot \frac{1}{\rho}$$

Pair distribution function

Lennard-Jones fluid near the triple point and hard-sphere fluid





reprinted from Nezbeda, Kolafa and Kotrla 1998

Pair distribution function

- $g(r)$ vanishes on short distances – molecules cannot intersect
- high peak – van der Walls radius, closest-contact distance
  (even though hard spheres do not have any attraction!)
  – much more likely to find this distance in LJ or HS than in IG
- longer distances – a few shallow minima and maxima,
  converges to unity – uniform probability as in IG

Fourier transform of $g(r)$ – structure factor S
– quantifies the scattering of incoming radiation in the material
– measured in diffraction experiments (X-ray, neutron)

$$S(\vec{q}) = \frac{1}{N} \left\langle \sum_j \sum_k \exp\left[-i \cdot \vec{q} \cdot (\vec{r}_j - \vec{r}_k)\right] \right\rangle$$

Pair distribution function

Importance – not only information about the structure

calculation of thermodynamic properties possible

using potential energy $u(r)$ and force $f(r)$ of a molecule pair

corrections to the IG values of total energy and pressure (EOS!):

$$E - \frac{3}{2} N k_B T = 2\pi N \rho \int_0^\infty r^2 \cdot u(r) \cdot g(r)\, dr$$

$$P - \rho\, k_B T = -\frac{2\pi}{3} \rho^2 \int_0^\infty r^3 \cdot f(r) \cdot g(r)\, dr$$

(as long as pairwise additivity of forces can be assumed)

Equilibration

- 'preliminary' simulation to reach the termodynamic equilibrium
- goal – stable thermodynamics properties (no drift)
- usually – E_{pot}, T, p, in NPT also ρ
  – evaluated by program readily and written to output
- structure – has to be taken care of, too
- start – often artificially regular (crystal-like) structure,
  which should be washed out during equilibration

Structural parameters

- translational order – Verlet's order parameter

$$\lambda = \frac{\lambda_x + \lambda_y + \lambda_z}{3}, \qquad \lambda_x = \frac{1}{N}\sum_{i=1}^{N} \cos\left[\frac{4\pi x_i}{a}\right] \quad \text{etc.}$$

  a – edge of the unit cell
  ideal crystal: $\lambda = 1$
  disordered structure: λ fluctuates around 0

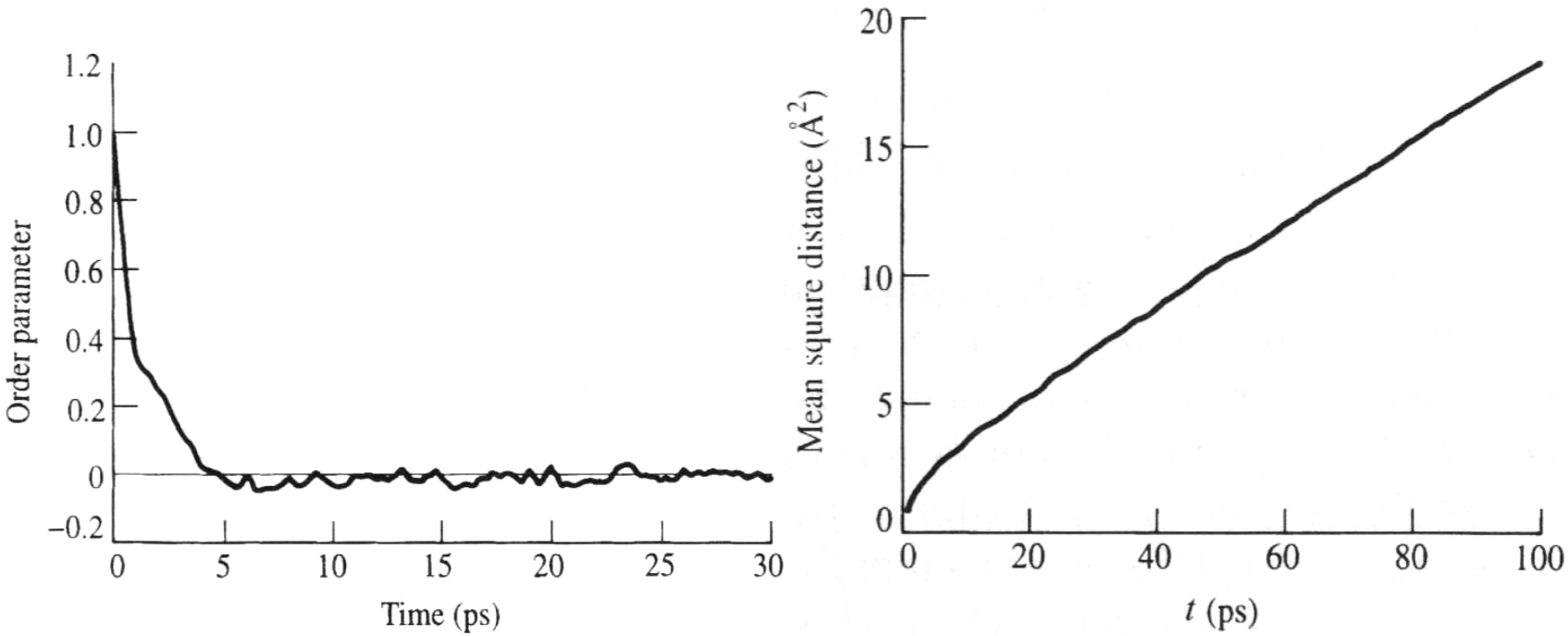
- mean squared displacement from initial position

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N} |\vec{r}_i(t) - \vec{r}_i(0)|^2$$

  should increase gradually in fluid with no specific structure
  would oscillate about a mean value in a solid

Structural parameters

equilibration of liquid argon followed by λ and MSD



Reprinted from Leach: Molecular Modelling

Correlation functions

two physical quantities x and y may exhibit correlation

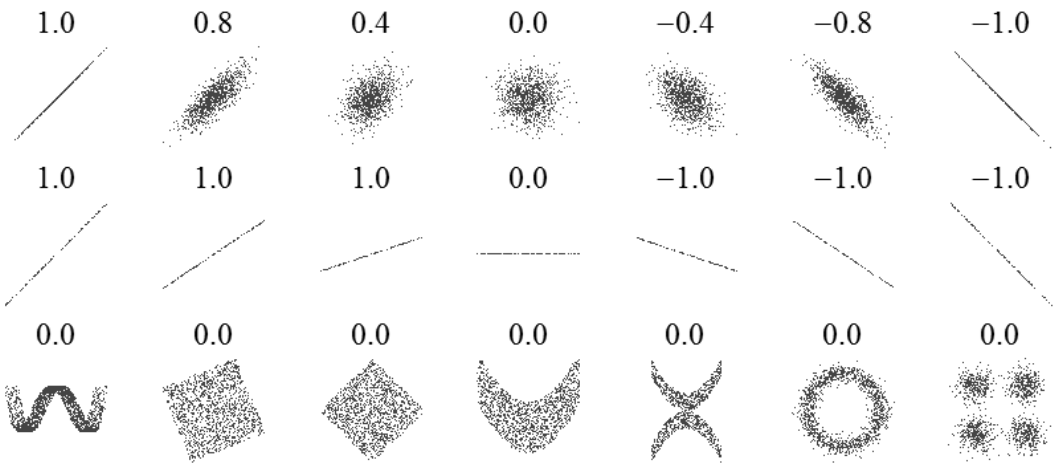
- indicates a relation of x and y, opposed to independence
- Pearson correlation coefficients
  - describe linear relationship between x and y
  - quantities fluctuate around mean values $\langle x \rangle$ and $\langle y \rangle$
  - consider only the fluctuating part
  - introduce correlation coefficient ρ_{xy}

$$\rho_{xy} = \frac{\langle (x - \langle x \rangle) \cdot (y - \langle y \rangle) \rangle}{\sqrt{\langle (x - \langle x \rangle)^2 \rangle \cdot \langle (y - \langle y \rangle)^2 \rangle}} = \frac{\mathrm{cov}(x, y)}{\sigma_x \cdot \sigma_y}$$

$\mathrm{cov}(x, y)$: covariance of x and y

Correlation functions

(not necessarily linear) correlation of two quantities and the coresponding correlation coefficients



Downloaded from Wikipedia

Correlation functions

MD – values of a quantity x as a function of time

possible – at some point in time, the value of x is correlated with the value of x at an earlier time point

– described by autocorrelation function (ACF)

$$c_x(t) = \frac{\langle x(t) \cdot x(0) \rangle}{\langle x(0) \cdot x(0) \rangle} = \frac{\int x(t')\, x(t'+t)\, \mathrm{d}t'}{\int x^2(t')\, \mathrm{d}t'}$$

– correlation of the same property x at two time points separated by t, normalized to takes values between -1 and 1

Autocorrelation of velocity

autocorrelation function – quantifies 'memory' of the system, or how quickly the system 'forgets' its previous state

velocity autocorrelation function

– tells how closely the velocities of atoms at time t resemble those at time 0

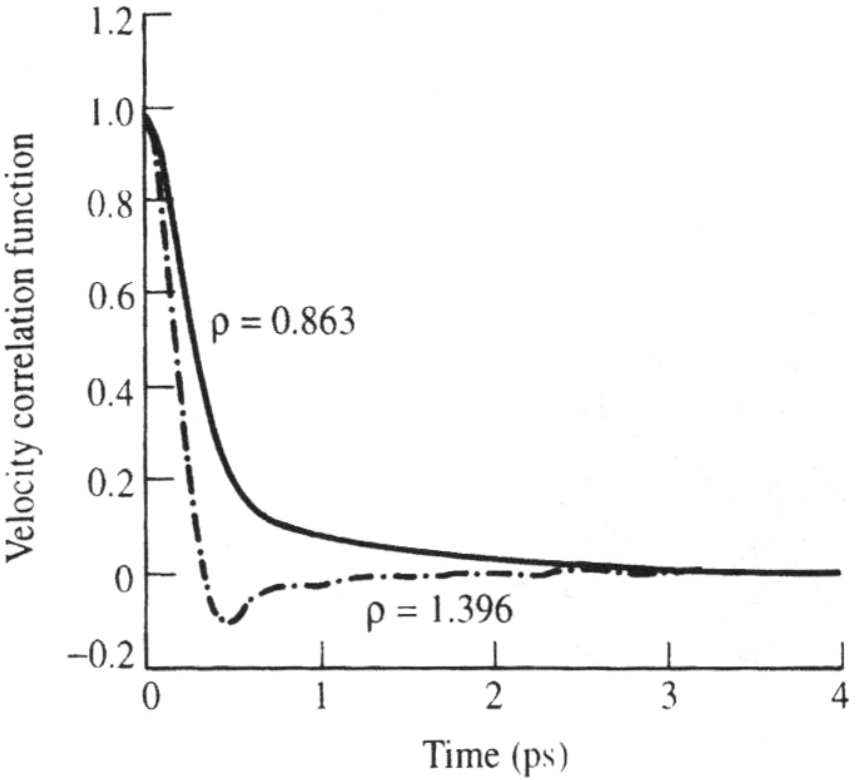
– usually averaged over all atoms i in the simulation

$$c_v(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{\langle \vec{v}_i(t) \cdot \vec{v}_i(0) \rangle}{\langle \vec{v}_i(0) \cdot \vec{v}_i(0) \rangle}$$

– typical ACF starts at 1 in $t = 0$ and decreases afterwards

Autocorrelation of velocity

ACF of velocity in simulations of liquid argon (densities in g·cm^{-3})



Reprinted from Leach: Molecular Modelling

lower ρ – gradual decay to 0
higher ρ – ACF comes faster to 0
– even becomes negative briefly
– 'cage' structure of the liquid
– one of the most interesting achievements of early simulations

Autocorrelation of velocity

time needed to lose the autocorrelation whatsoever
– correlation time or relaxation time:

$$\tau_v = \int_0^\infty c_v(t)\,\mathrm{d}t$$

may help to resolve certain statistical issues:
when averaging over time the properties of system,
it is necessary to take uncorrelated values
if the property is dynamical (related to v),
we can take values of the property separated by τ_v

Autocorrelation of velocity

connection between velocity ACF and transport properties

– Green–Kubo relation for self-diffusion coefficient D:

$$D = \frac{1}{3} \int_0^\infty \langle \vec{v}_i(t) \cdot \vec{v}_i(0) \rangle_i \, \mathrm{d}t$$

– interesting observable quantities

– important to be able to calculate them from MD

– another way: Einstein relation for D using the MSD

$$D = \frac{1}{6} \lim_{t \to \infty} \frac{\left\langle |\vec{r}_i(t) - \vec{r}_i(0)|^2 \right\rangle_i}{t}$$

NB: Fick's laws of diffusion $J = -D \frac{\partial \phi}{\partial x}$, $\frac{\partial \phi}{\partial t} = D \frac{\partial^2 \phi}{\partial x^2}$

Autocorrelation of dipole moment

velocity – property of a single atom; contrary to that –
– some quantities need to be evaluated for whole system

total dipole moment:

$$\vec{\mu}_{\text{tot}}(t) = \sum_{i=1}^{N} \vec{\mu}_i(t)$$

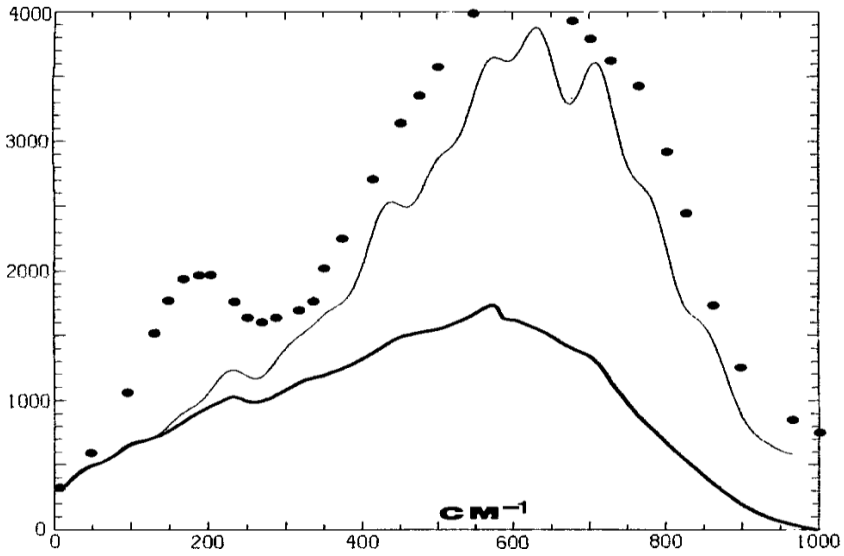
ACF of total dipole moment:

$$c_\mu(t) = \frac{\langle \vec{\mu}_{\text{tot}}(t) \cdot \vec{\mu}_{\text{tot}}(0) \rangle}{\langle \vec{\mu}_{\text{tot}}(0) \cdot \vec{\mu}_{\text{tot}}(0) \rangle}$$

– related to the vibrational spectrum of the sample
– IR spectrum may be obtained as Fourier transform of dipolar ACF

Autocorrelation of dipole moment

IR spectra for liquid water from simulations



thick – classical MD,
thin – quantum correction,
black dots – experiment

B. Guillot, J. Phys. Chem. 1991

no sharp peaks at well-defined frequencies (as in gas phase)

rather – continuous bands – liquid absorbs frequencies in a broad interval

frequencies – equivalent to the rate of change of total dipole moment

Principal component analysis

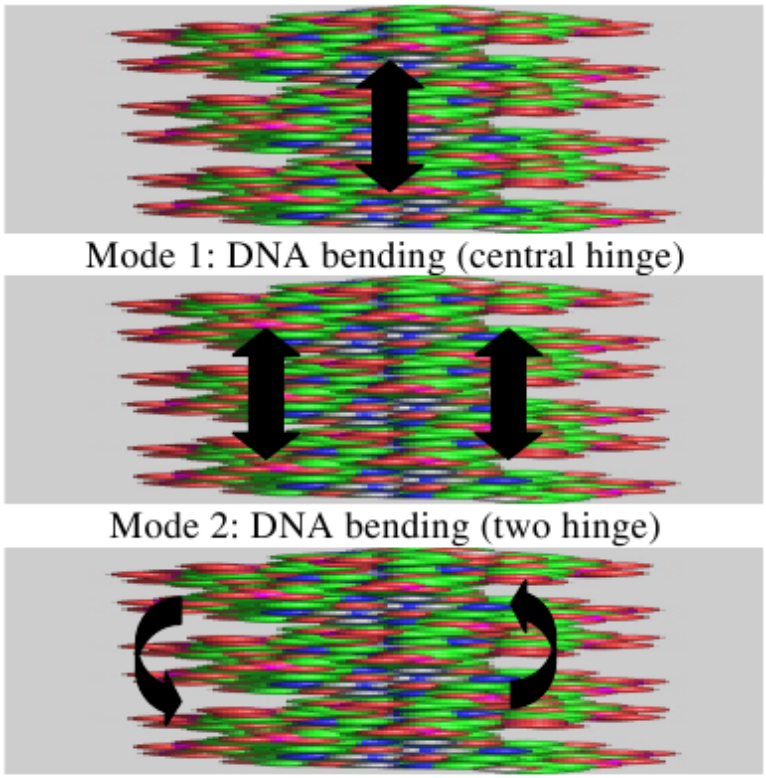
covariance analysis on the atomic coordinates along MD trajectory
= principal component analysis (PCA), or essential dynamics
$3N$-dim. covariance matrix C of atomic coordinates $r_i \in \{x_i, y_i, z_i\}$

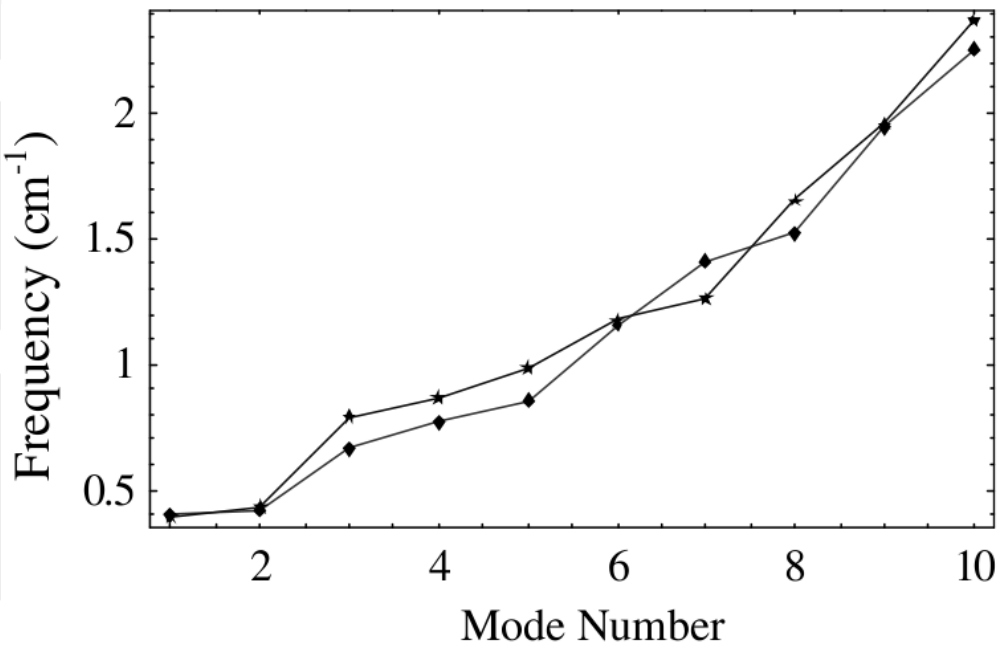
$$C_{ij} = \langle (r_i - \langle r_i \rangle) \cdot (r_j - \langle r_j \rangle) \rangle_t \quad \text{or}$$
$$C_{ij} = \left\langle \sqrt{m_i}(r_i - \langle r_i \rangle) \cdot \sqrt{m_j}(r_j - \langle r_j \rangle) \right\rangle_t$$

diagonalization $\rightarrow$
eigenvalues – may be expressed as vibrational frequencies
eigenvectors – principal or essential modes of motion
– analogy of normal modes of vibration
– first few – global, collective motions, many atoms involved

Principal component analysis

example – PCA of double-stranded DNA, 3 lowest eigenvectors and corresponding frequencies (2 different DNA sequences)





Reprinted from S. A. Harris, J. Phys. Condens. Matter 2007

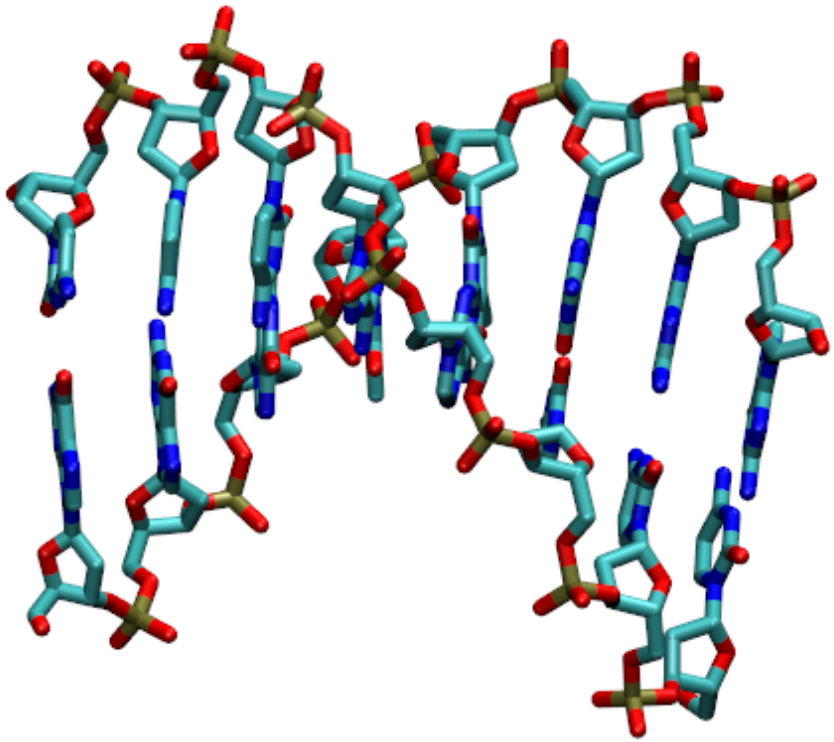
Principal component analysis

DNA – the modes are the same as expected for a flexible rod
- 2 bending modes around axes perpendicular to the principal axis of the DNA, and a twisting mode

PCA – gives an idea of what the modes of motion look like
- additionally – basis for thermodynamic calculations
  - vibrational frequencies may lead to configurational entropy

Principal component analysis

DNA octamer, eigenvectors 1, 2 and 3



Fourier transform

FT describes which frequencies are present in a function (of time)
– decomposes $f(t)$ into a 'sum' of periodic oscillatory functions

$$F(\omega) = \int_{-\infty}^{\infty} f(t) \cdot \exp\left[-\mathrm{i}\,\omega t\right] \mathrm{d}t$$

note that $\exp\left[-\mathrm{i}\,\omega t\right] = \cos\left[\omega t\right] - \mathrm{i}\sin\left[\omega t\right]$