Biomolecular modeling III

Marcus Elstner and Tomáš Kubař

2019, December 10

Non-bonded interactions

speeding up the number-crunching

Non-bonded interactions – why care?

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

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

- key to understand biomolecular structure and function
  - binding of a ligand
  - efficiency of a reaction
  - color of a chromophore
- main contribution to the computational cost
  - good target of optimization

Cut-off – simple idea

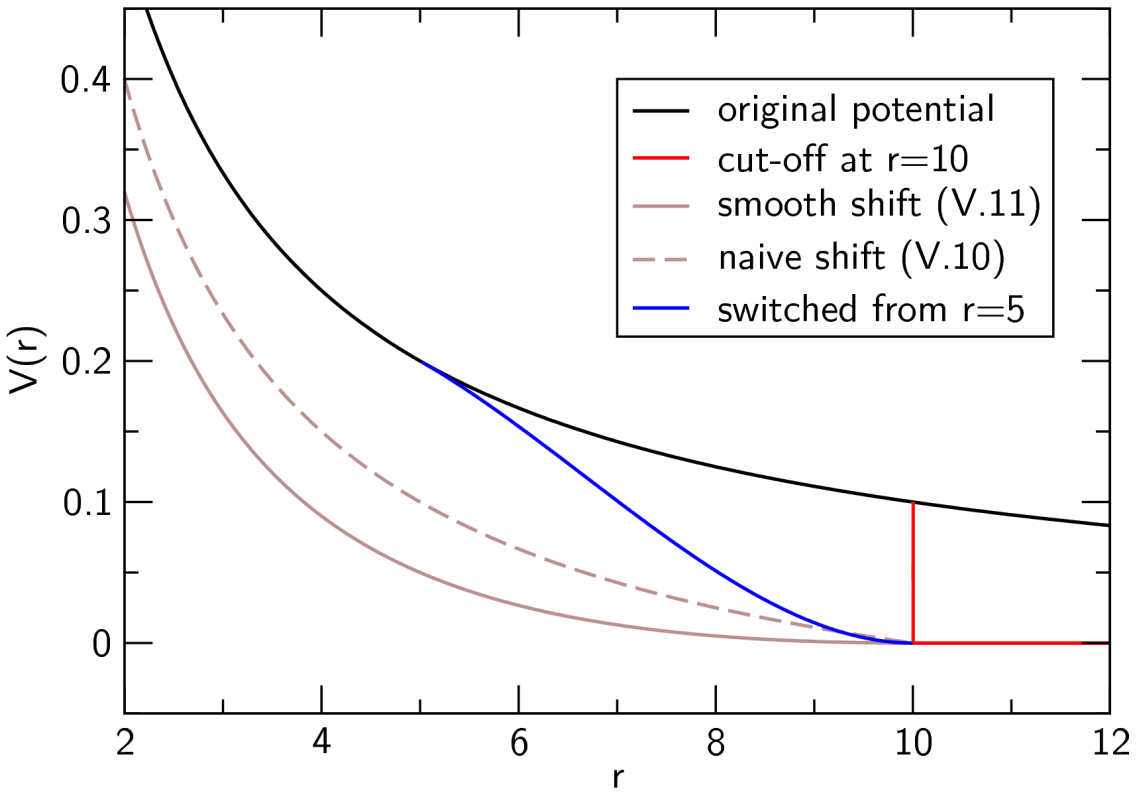
with PBC – infinite number of interaction pairs in principle,
but the interaction gets weaker with distance

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



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

the infinite system is periodic – a trick may be applied:
Ewald summation method, or even better
particle–mesh Ewald method PME

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

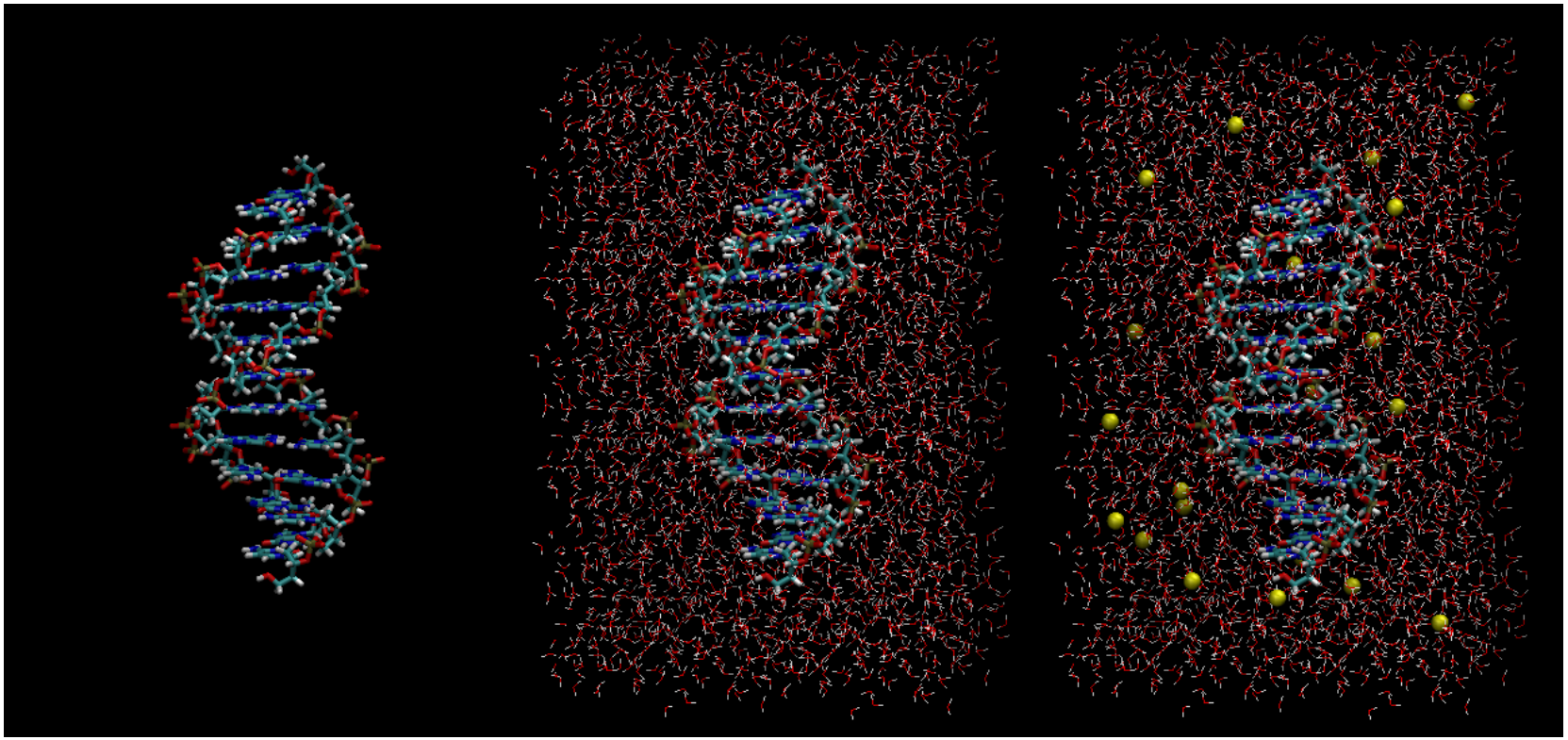
1. build the initial structure
2. bring the system into thermodynamic equilibrium
3. do the productive simulation
4. analyze the trajectory

Tools to build the structure

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



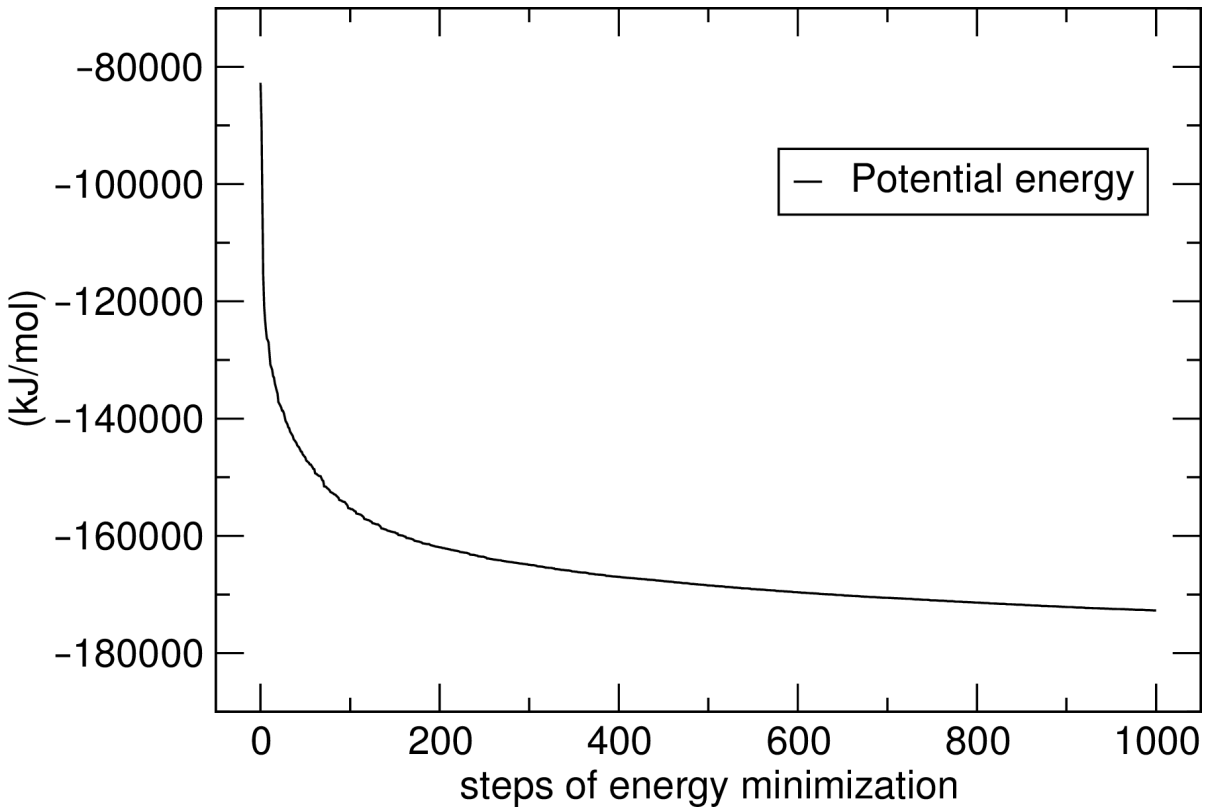
Why equilibrate?

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

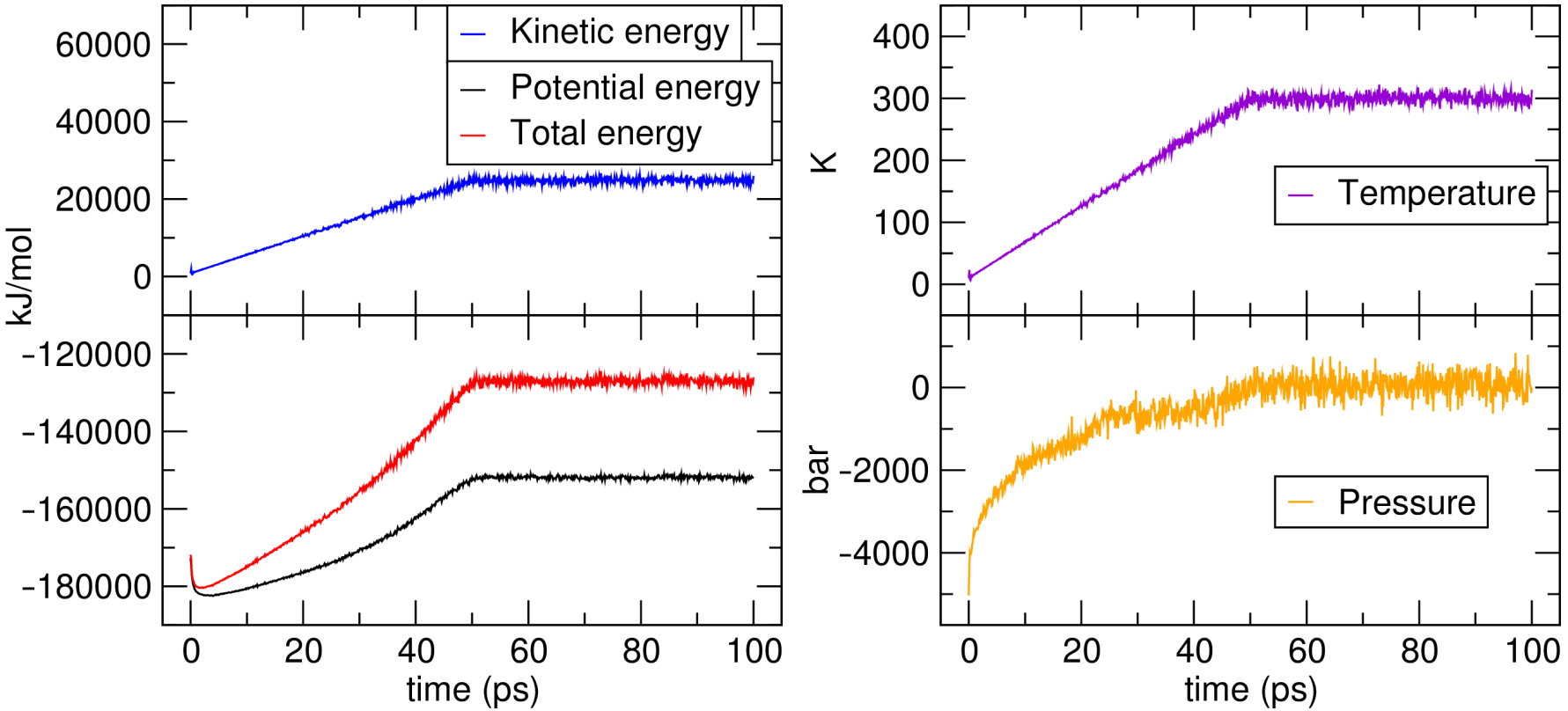
How to equilibrate

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

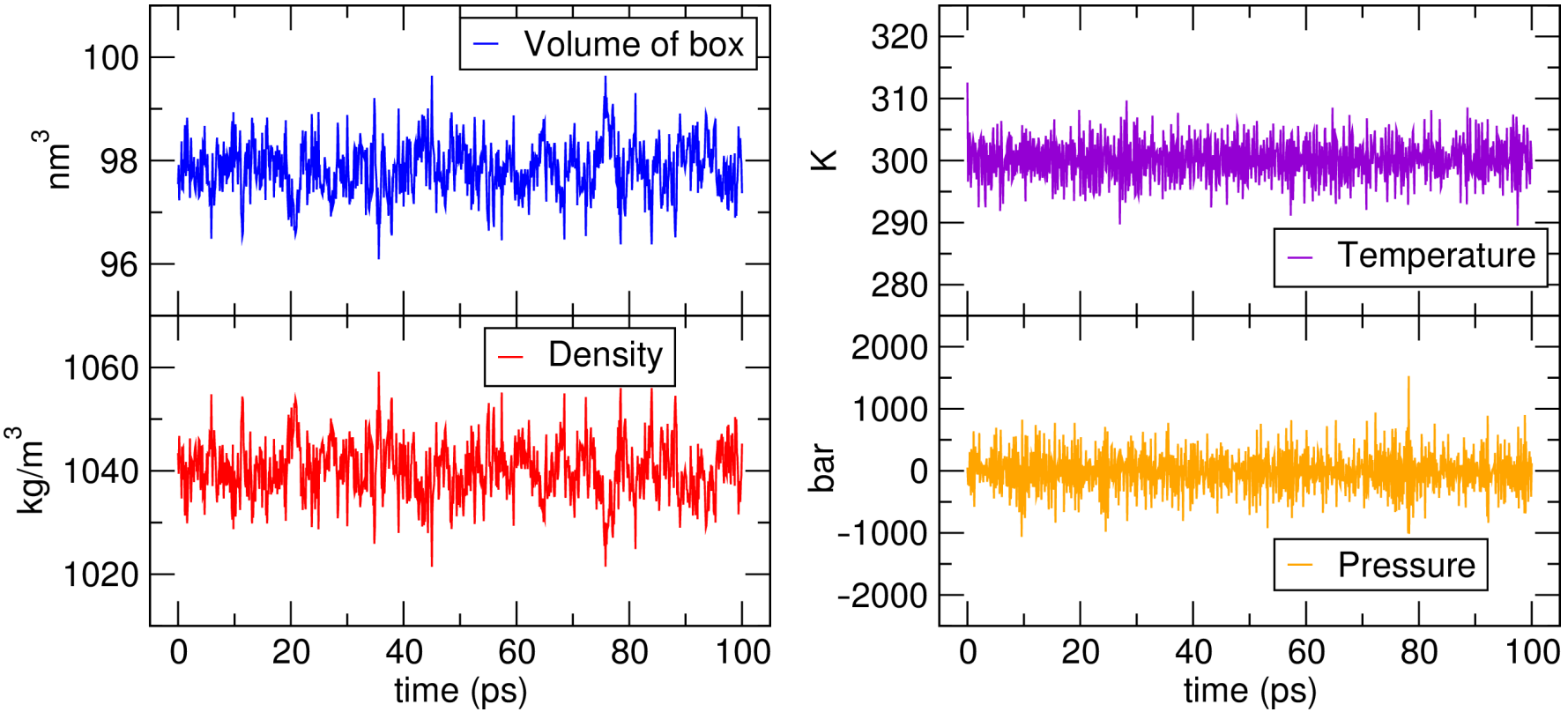


Thermalization



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

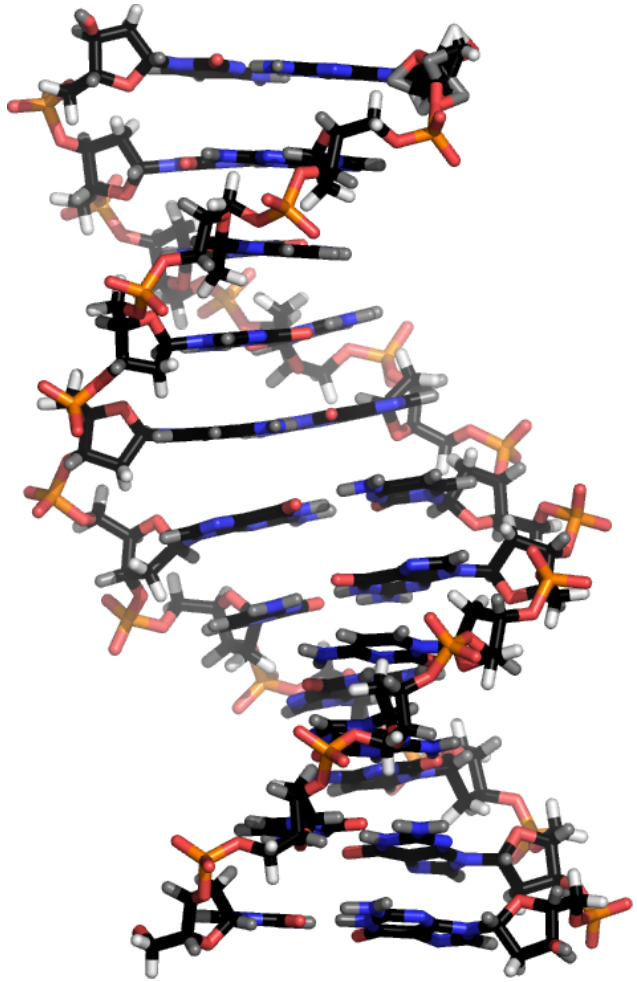
Structure – single molecule in solvent

concentrating 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 $\vec{r}^{\text{ref}}$

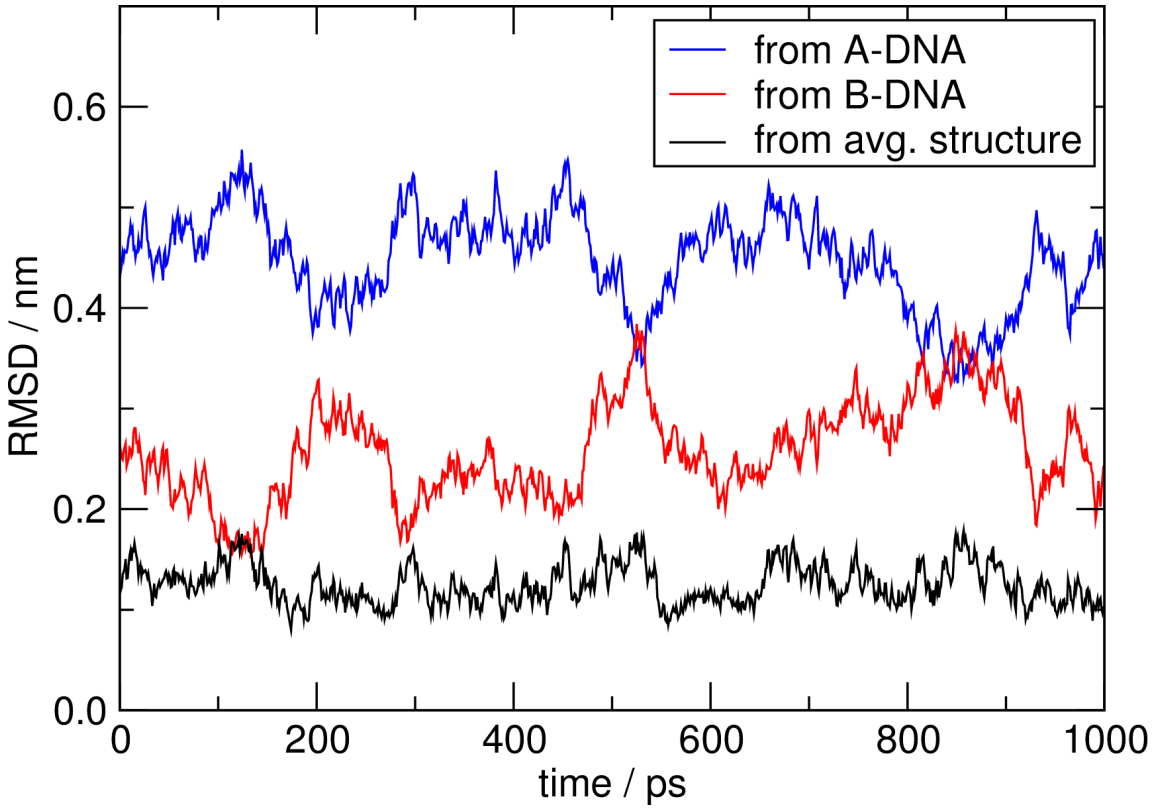
$$\text{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\vec{r}_i(t) - \vec{r}_i^{\text{ref}}\right|^2}$$

- 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

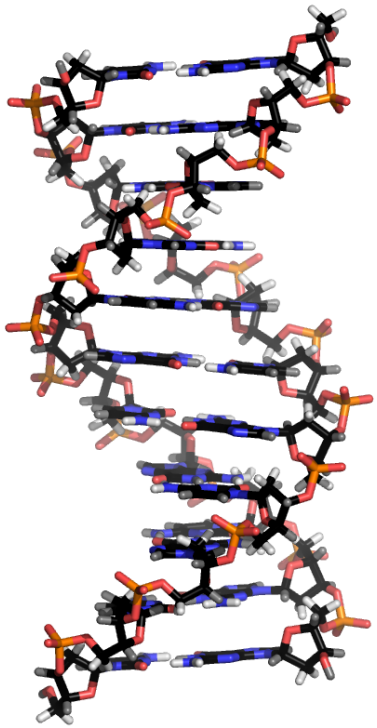
Root mean square deviation

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

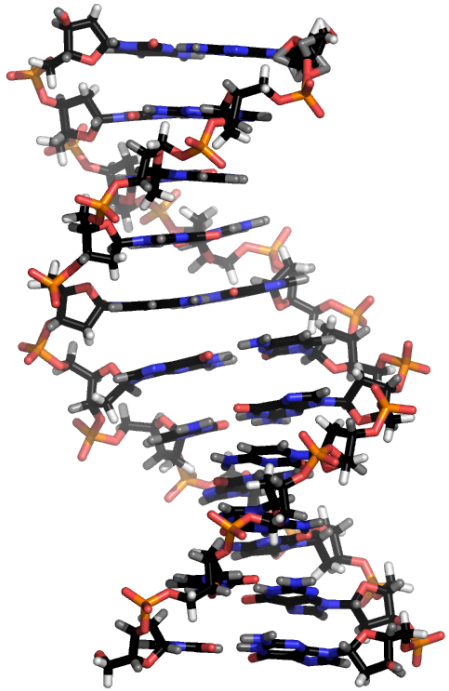


Root mean square deviation

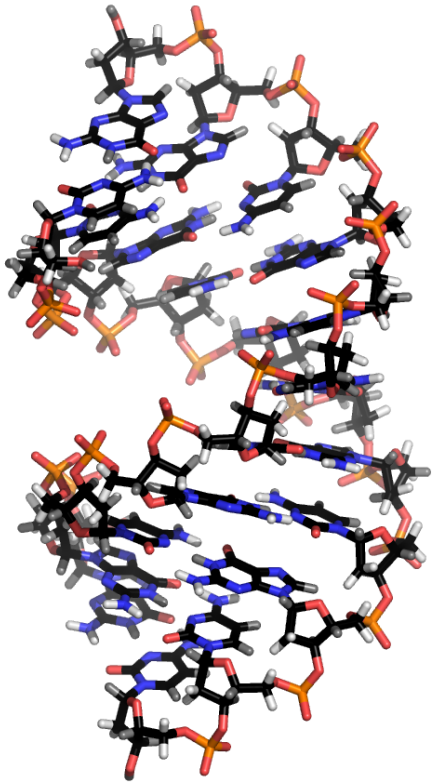
B-DNA



average structure



A-DNA



Magnitude of structural fluctuation

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

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

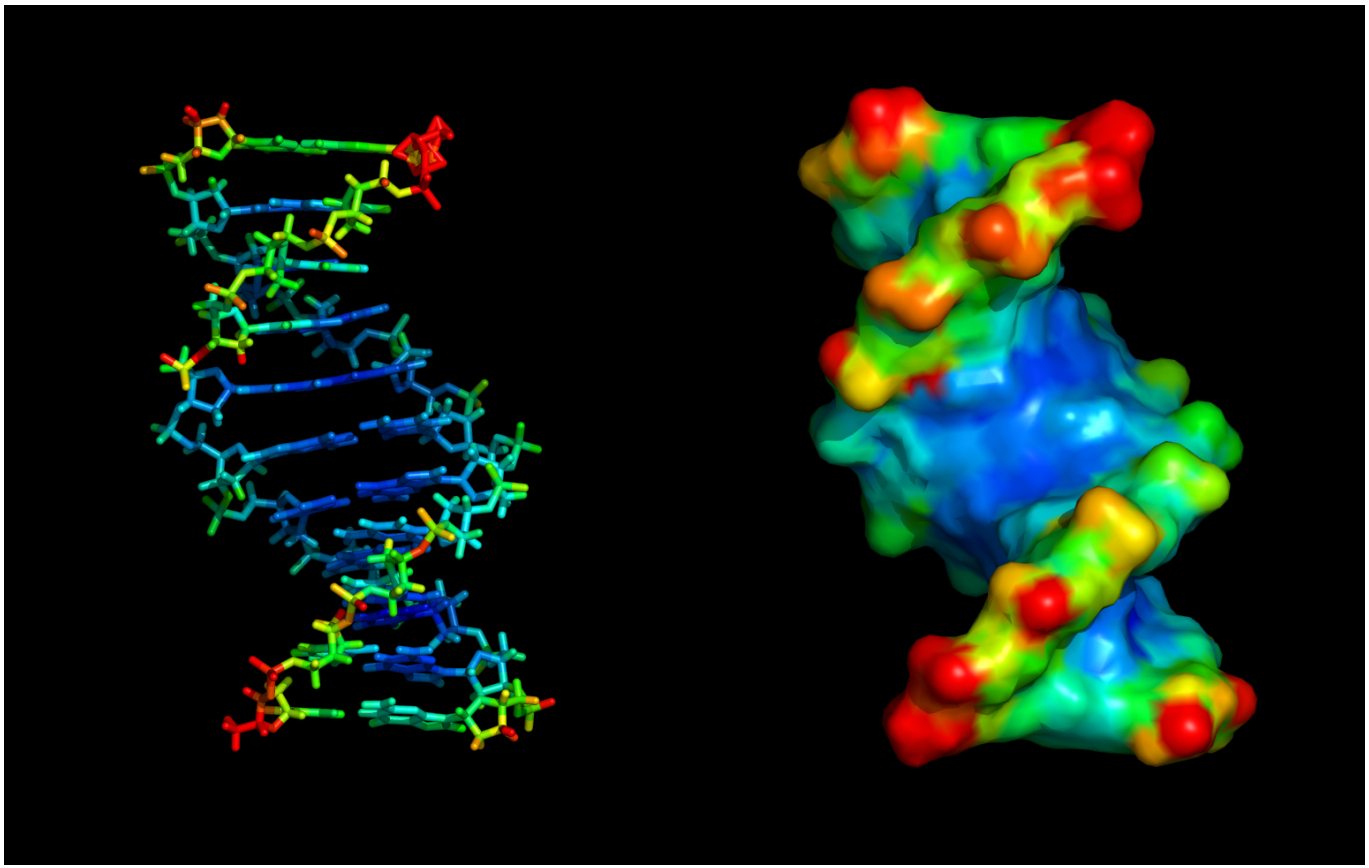
– may be converted to B-factor

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

– 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

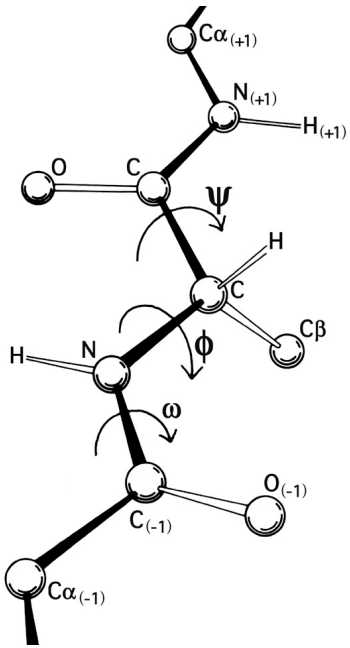


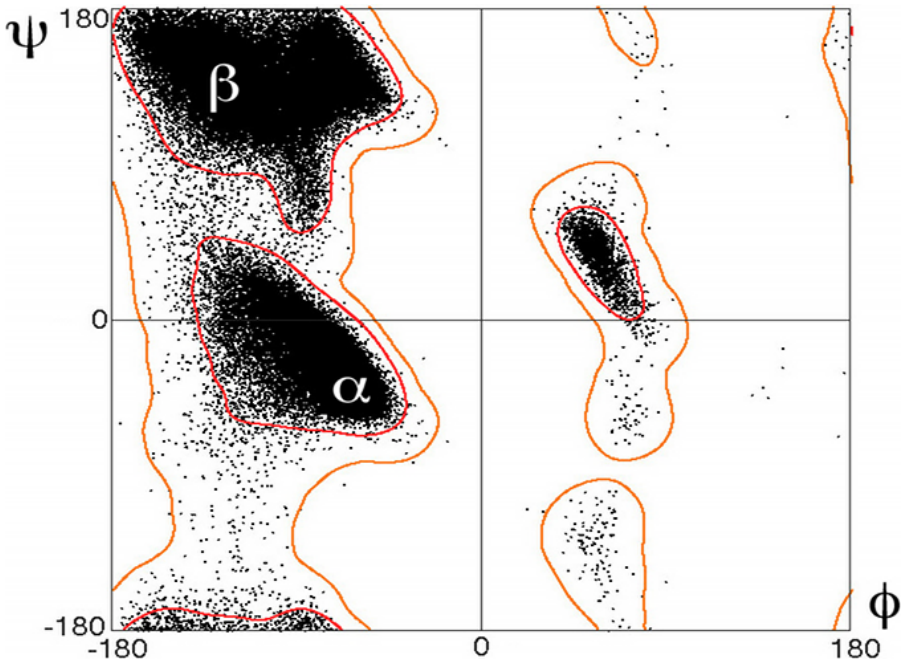
(blue $<$ green $<$ yellow $<$ red)

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

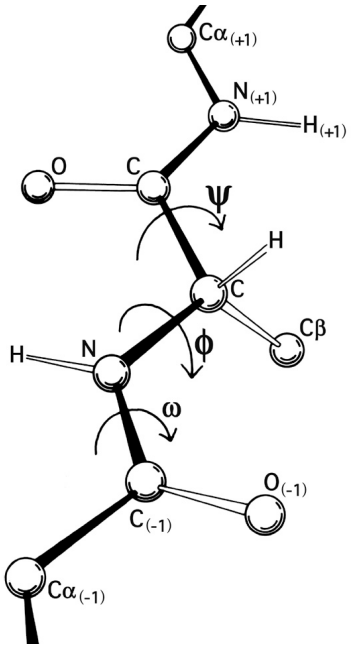


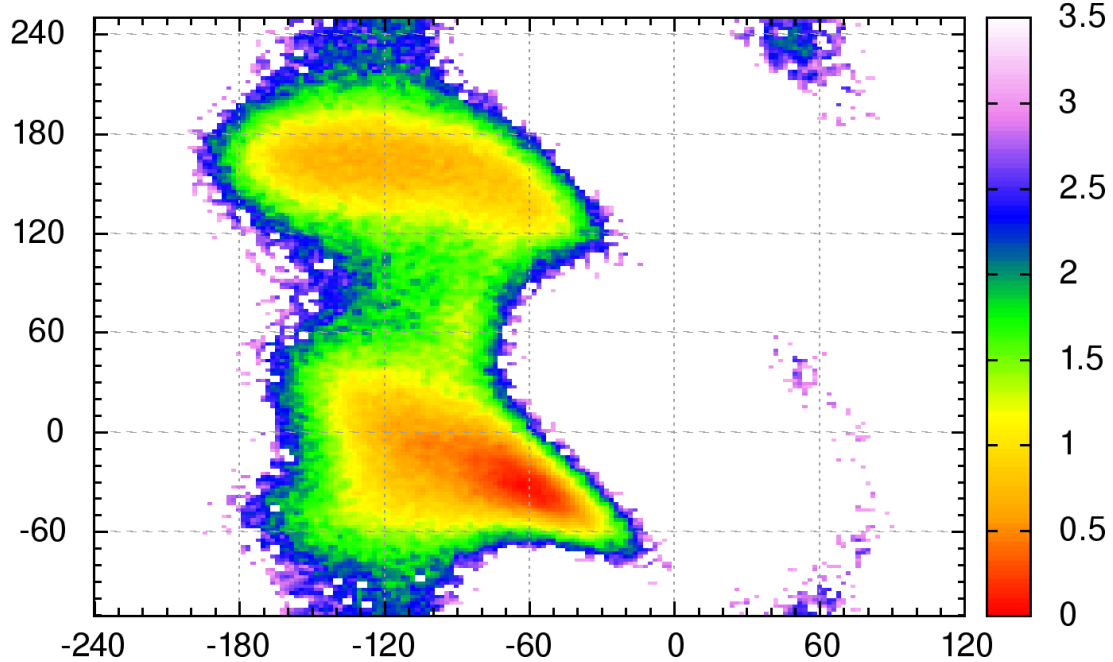


Structure of peptides and proteins

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

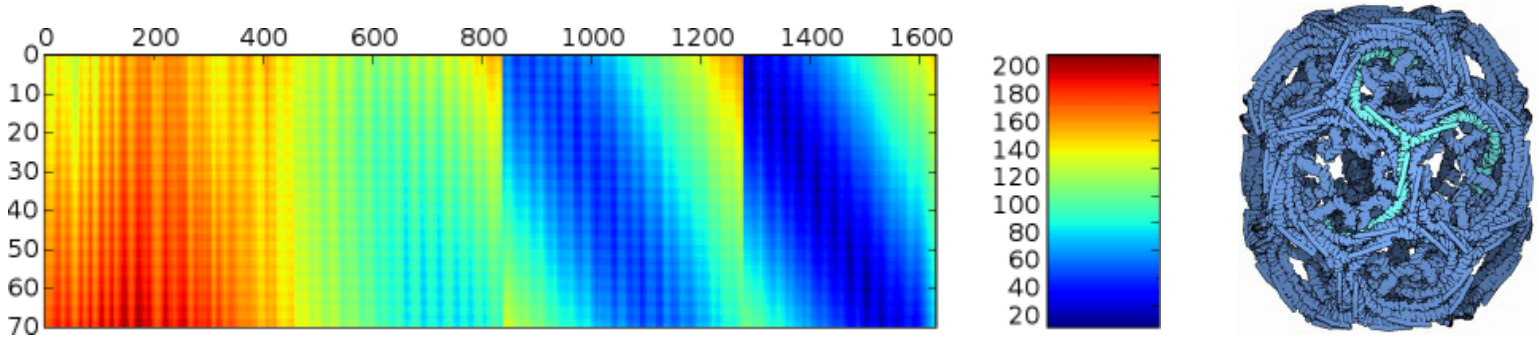




Structure of peptides and proteins

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

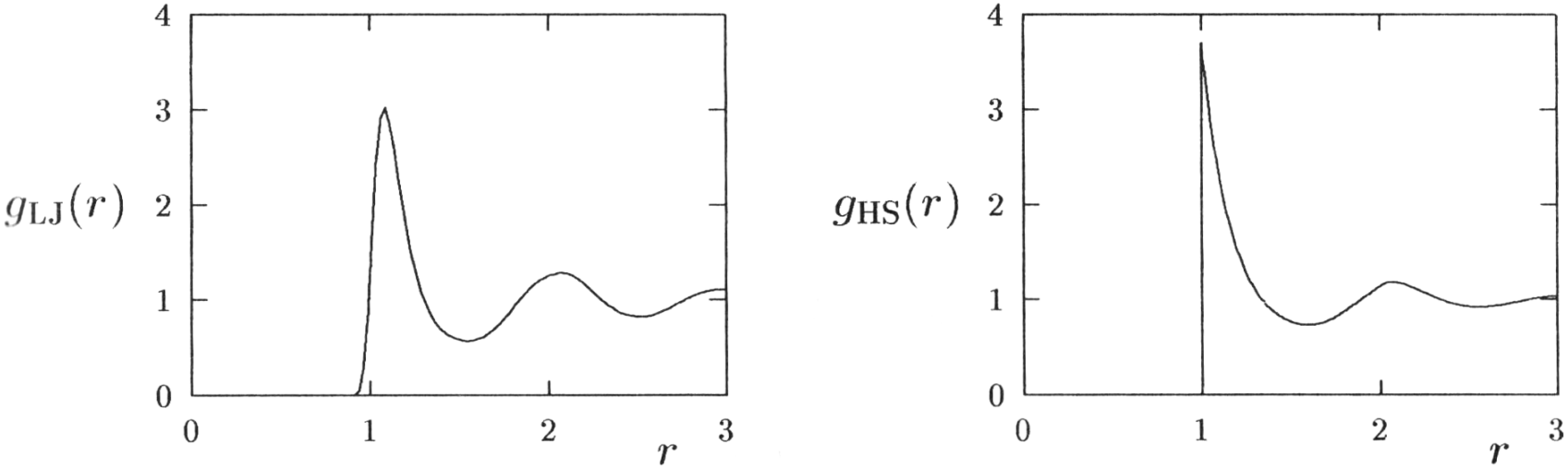
radial distribution functions

- 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



- $g(r)$ vanishes on short distances – molecules cannot intersect
- high peak – van der Waals 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

Pair distribution function

Fourier transform of $g(r)$ – structure factor S

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

– quantifies the scattering of incoming radiation in the material
– measured in diffraction experiments (X-ray, neutron)

Principal component analysis

analysis of covariance/correlation of the atomic coordinates
= PCA a.k.a. essential dynamics

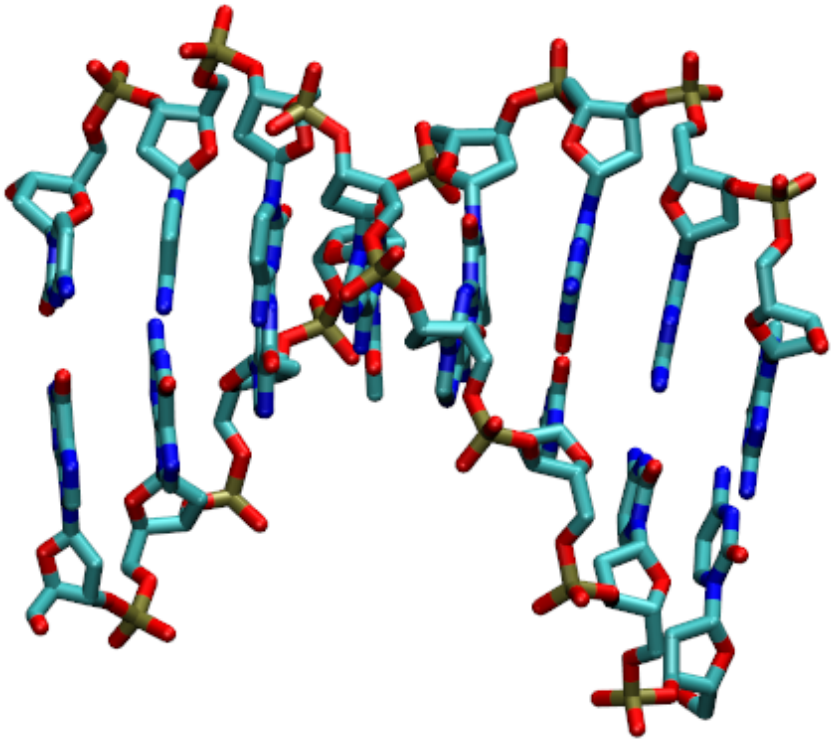
$3N$-dim. covariance matrix C of atomic coordinates $r_i \in \{x_i, y_i, z_i\}$

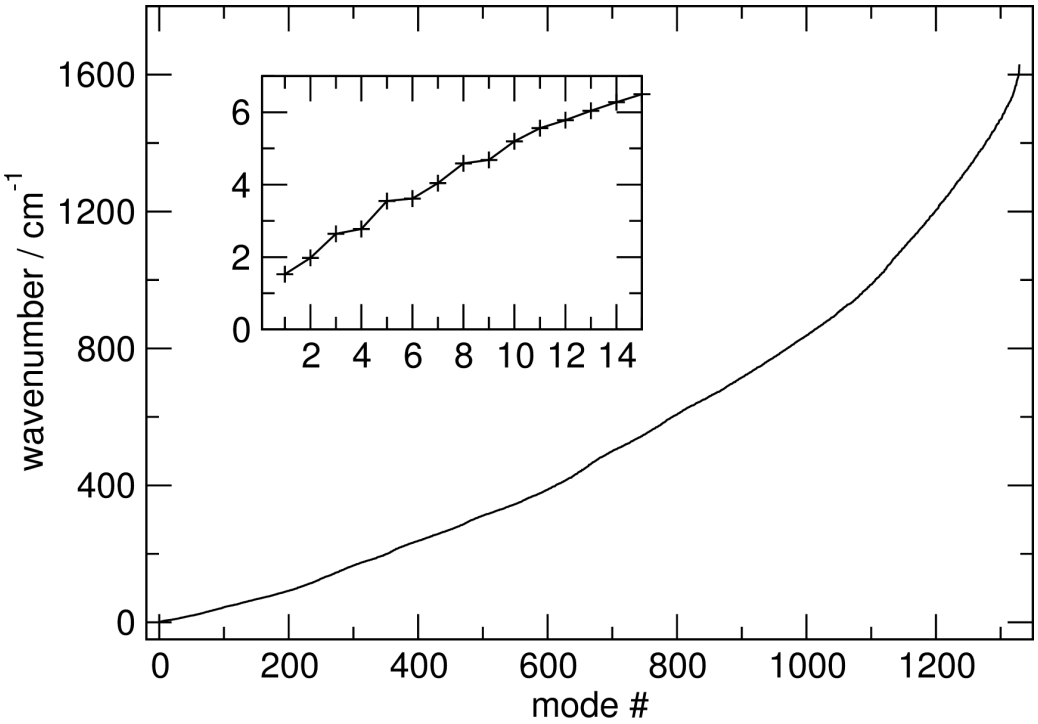
$$C_{ij} = \langle (r_i - \langle r_i \rangle) \cdot (r_j - \langle r_j \rangle) \rangle_t \quad \text{or}$$
$$C_{ij} = \left\langle \sqrt{m_i}(r_i - \langle r_i \rangle) \cdot \sqrt{m_j}(r_j - \langle r_j \rangle) \right\rangle_t$$

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 a double-stranded DNA octanucleotide, frequencies and 3 lowest eigenvectors





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