Analysis of the simulation

Marcus Elstner and Tomáš Kubař

2019, June 6

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:

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

– elegant way from a single simulation to heat capacity

General note on averaging

simulated 2 MD trajectories → two sets of 1000 values of A each
perform averaging of A separately → $\mu_1 \pm \sigma_1$ and $\mu_2 \pm \sigma_2$

- how to average over the whole ensemble over 2000 values?
- $\mu = \frac{1}{2}(\mu_1 + \mu_2)$
- what about the std. deviation σ? ~~$\sigma = \frac{1}{2}(\sigma_1 + \sigma_2)$~~
- hint: make use of $\sigma^2 = \langle A^2 \rangle - \langle A \rangle^2$
- solution: for each set, perform averaging of A as well as A^2,
  then it is safe to average the averages
  $\langle A^2 \rangle = \frac{1}{2}\left(\langle A^2 \rangle_1 + \langle A^2 \rangle_2\right)$
  which leads to σ

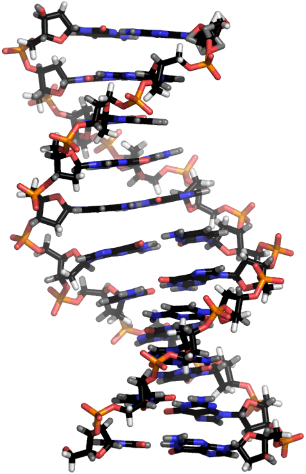
Single molecule in solvent

concentrating on the dissolved molecule
– protein, DNA,. . .

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

$$\vec{r}_i = \frac{1}{N} \sum_{n=0}^{N} \vec{r}_i^{\,(n)}$$

– clear, simple, often reasonable



Average structure

Possible problems:

- 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. . .
- freely rotatable single bonds – CH_3
  - all 3 hydrogens collapse to a single point
  - no problem – ignore hydrogens
- 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 $\vec{r}^{\,\mathrm{ref}}$

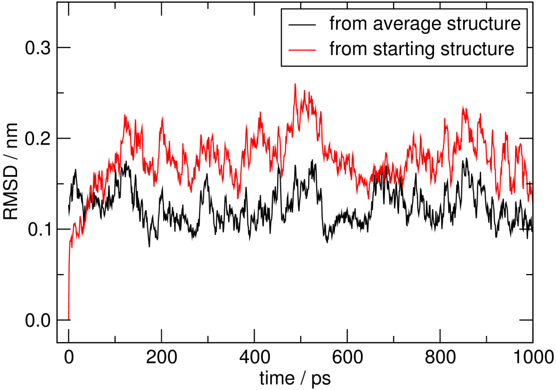
$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\vec{r}_i(t) - \vec{r}_i^{\,\mathrm{ref}}\right|^2}$$

- 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

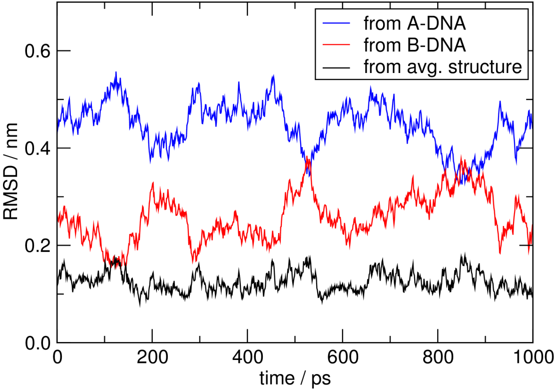
Root mean square deviation

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



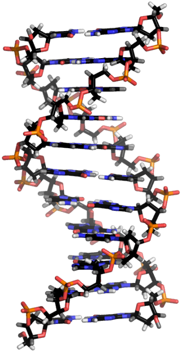
Root mean square deviation

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

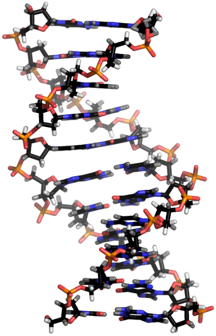


Root mean square deviation

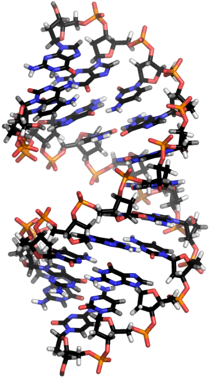
B-DNA



average structure



A-DNA



Magnitude of structural fluctuation

root mean square fluctuation (RMSF)

of position of every single atom averaged along MD trajectory

(T – no. of structures in trajectory; $\langle\vec{r}_i\rangle$ – averaged structure)

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{t=0}^{T}\left|\vec{r}_i(t) - \langle\vec{r}_i\rangle\right|^2}$$

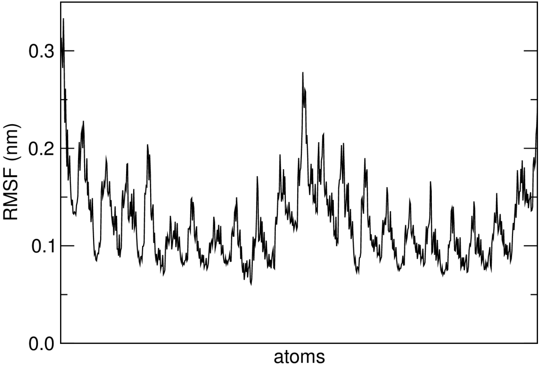
– may be converted to B-factor:

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

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

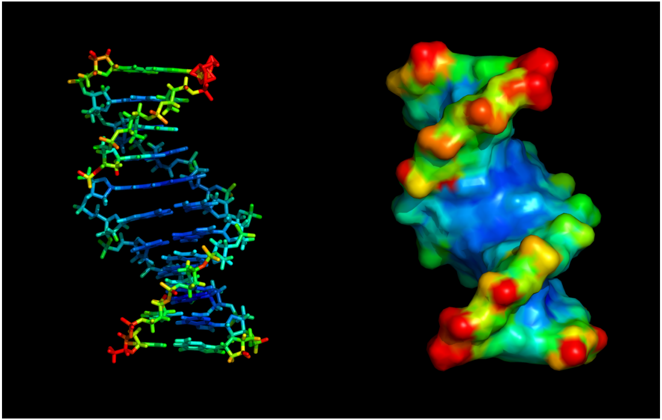
Root mean square fluctuation

RMSF of atomic positions in DNA oligonucleotide



Root mean square fluctuation

RMSF of atomic positions in DNA oligonucleotide



(blue < green < yellow < red)

Structure of double-helical nucleic acids

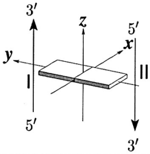
PDB ID 1EHZ
phenylalanine tRNA from *S. cerevisiæ*

downloaded from http://x3dna.org

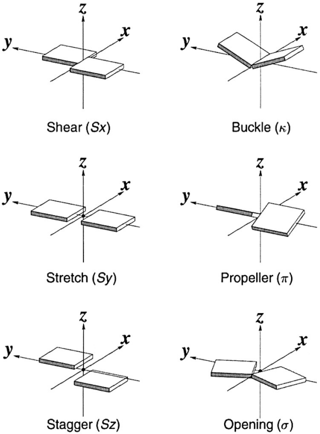
Structure of double-helical nucleic acids

Helical parameters

bases within a pair



Coordinate frame



Shear (Sx)

Buckle (κ)

Stretch (Sy)

Propeller (π)

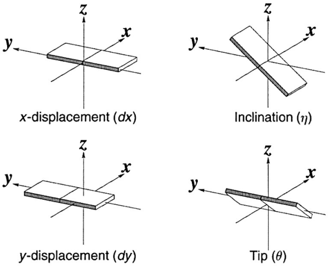
Stagger (Sz)

Opening (σ)

Structure of double-helical nucleic acids

Helical parameters

pair in a helix



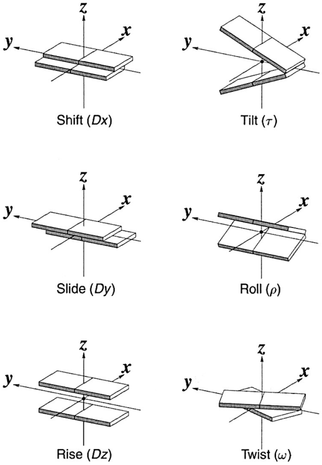
x-displacement (dx)

Inclination (η)

y-displacement (dy)

Tip (θ)

two pairs relative



Shift (Dx)

Tilt (τ)

Slide (Dy)

Roll (ρ)

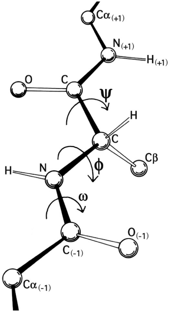
Rise (Dz)

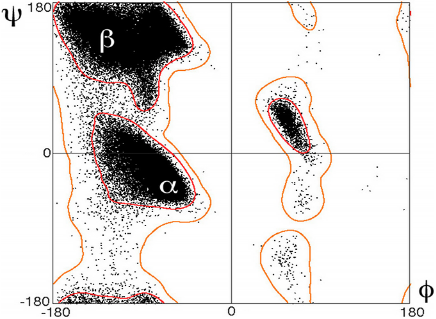
Twist (ω)

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

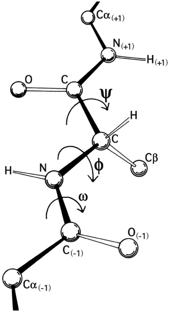


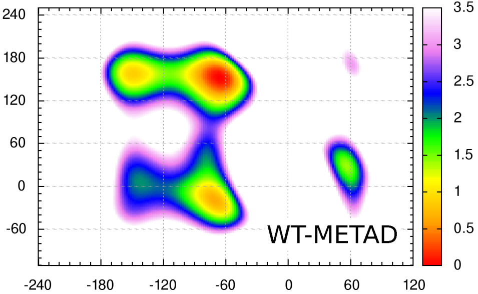


Structure of peptides and proteins

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



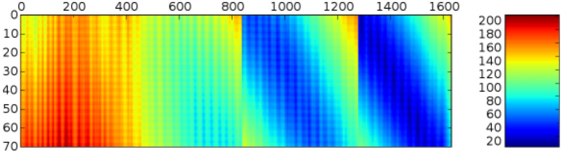


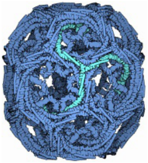
color codes Gibbs free energy in kcal/mol

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

Structure of fluids

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

radial distribution functions

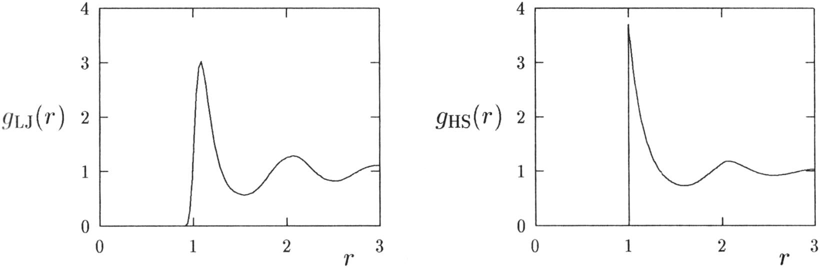
- pair distribution function

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

– probability to find a molecule in distance r from ref. mol.
– division by the macroscopic density – normalization

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 do not 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

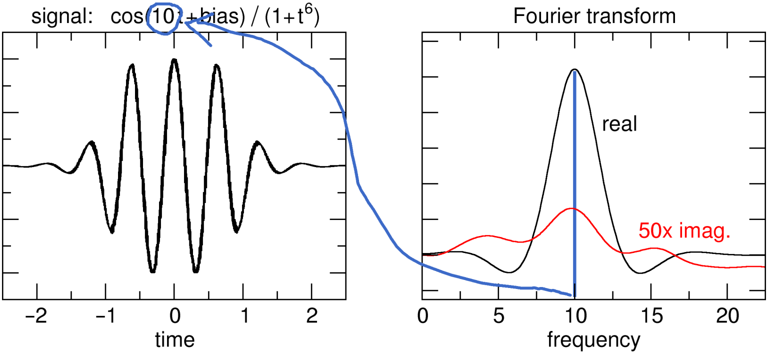
Fourier transform of $g(r)$ – structure factor S
– quantifies the scattering of incoming radiation in the material
– measured in diffraction experiments (X-ray, neutron)

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

intermission: Fourier transformation

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

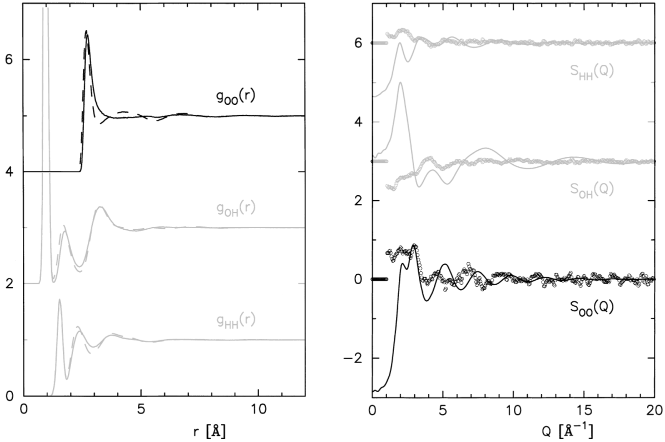
$$F(\omega) = \int_{-\infty}^{\infty} f(t) \cdot \exp[-\mathrm{i}\,\omega t]\,\mathrm{d}t \qquad \exp[-\mathrm{i}\,\omega t] = \cos[\omega t] - \mathrm{i}\sin[\omega t]$$



Pair distribution function

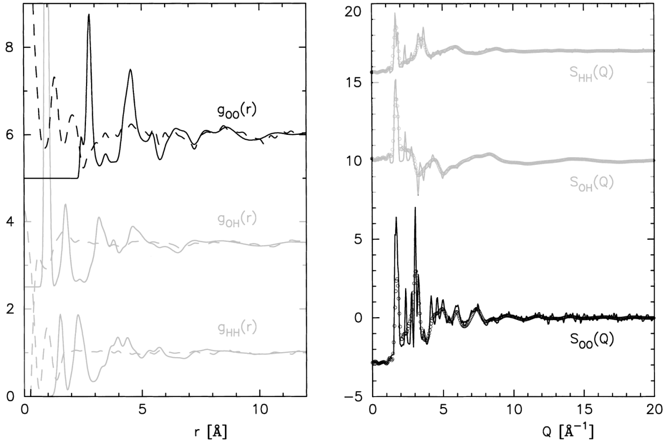
$g(r)$ and $S(q)$ of water

(Soper, Chemical Physics 2000)



Pair distribution function

$g(r)$ and $S(q)$ of ice Ih at 220 K and 1 bar (Soper, Chemical Physics 2000)



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!):

$$\begin{aligned} E - \frac{3}{2} N k_\mathrm{B} T &= 2\pi N \rho \int_0^\infty r^2 \cdot u(r) \cdot g(r)\,\mathrm{d}r \\ P - \rho\, k_\mathrm{B} T &= -\frac{2\pi}{3} \rho^2 \int_0^\infty r^3 \cdot f(r) \cdot g(r)\,\mathrm{d}r \end{aligned}$$

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

Correlation functions

two physical quantities $x(t)$ and $y(t)$ may exhibit correlation

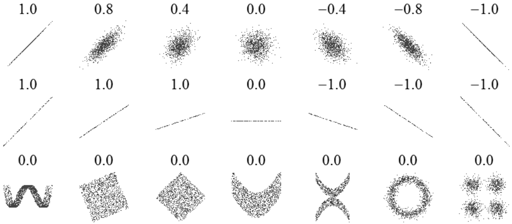
- indicates a relation of x and y, opposed to independence
- quantification – several kinds of correlation functions
- Pearson correlation coefficients
  - describe linear relationship between x and y
  - quantities fluctuate around mean values $\langle x \rangle$ and $\langle y \rangle$
  - consider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

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

$\operatorname{cov}(x, y)$: 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

- at some point in time, the value of x may be 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

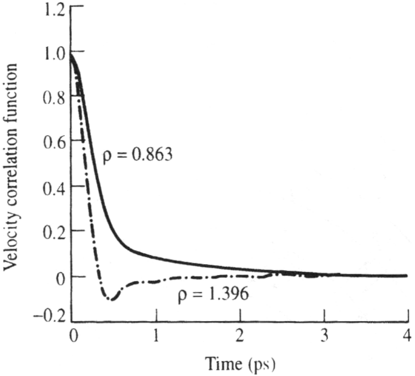
- tells how closely the velocities of atoms at time t resemble those at time 0
- usually averaged over all atoms i in the simulation

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

- 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$\cdot$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:

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

- interesting observable quantities
- important to be able to calculate them from MD
- there is yet another way from simulation to D
  - Einstein relation for D using mean squared displacement

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

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

Autocorrelation of dipole moment

total dipole moment:

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

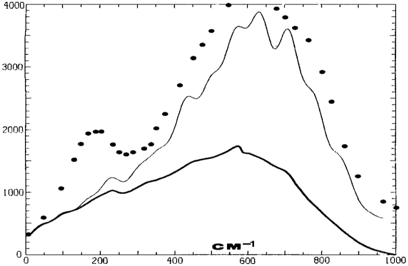
ACF of total dipole moment:

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

– related to the vibrational spectrum of the sample
– IR spectrum may be obtained as Fourier transform of $c_\mu(t)$

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 $r_i \in \{x_i, y_i, z_i\}$

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

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

diagonalization →

eigenvalues – may be expressed as quasi-harmonic frequencies

eigenvectors – principal or essential modes of motion

– analogy of normal modes of vibration

– first few – largest eigenvalues, lowest frequencies

– global, collective motions, many atoms involved

intermission: diagonalization of a matrix

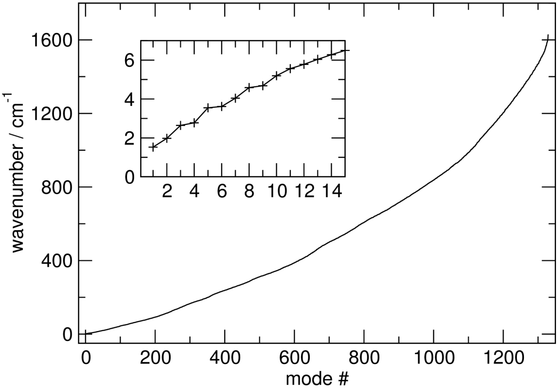
- is a process of finding
  eigenvectors a with corresponding eigenvalues α
  of a square $(n \times n)$ matrix $\mathcal{A}$:

  $$\mathcal{A} \cdot a = \alpha \cdot a$$

  (here: eigenvectors are column vectors)
- why the name?
  if eigenvectors are arranged into a matrix $\mathbf{a} = (a_1 a_2 \ldots a_n)$
  then $\mathbf{a}^{-1} \cdot \mathcal{A} \cdot \mathbf{a}$ is a diagonal matrix
- symmetric or Hermitian matrix → all eigenvalues are real
- computational cost of diagonalization: $\mathcal{O}(n^3)$

Principal component analysis

Double-stranded DNA oligonucleotide – lowest frequencies

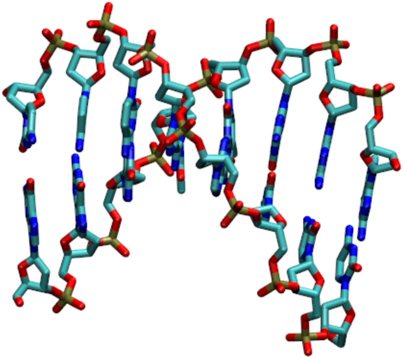


10 ns simulation of a double-helical DNA 11-nucleotide

691 atoms, of which 445 non-hydrogen → 1329 vib. modes

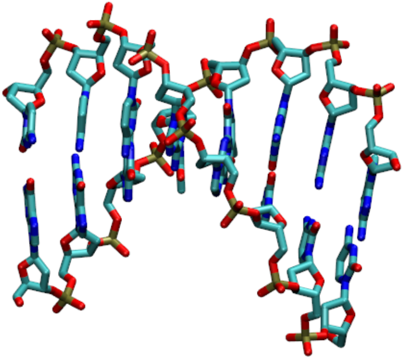
Principal component analysis

DNA octamer, eigenvector 1



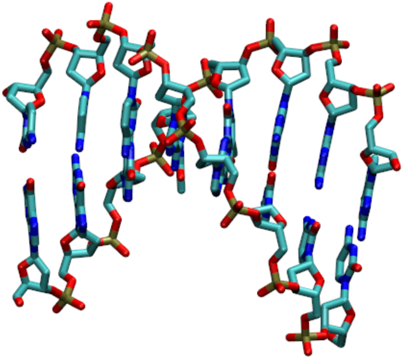
Principal component analysis

DNA octamer, eigenvector 2



Principal component analysis

DNA octamer, eigenvector 3



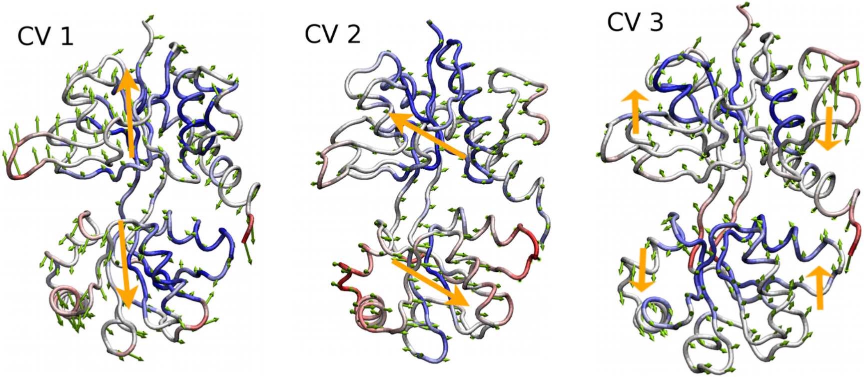
Principal component analysis

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

Binding domain of a glutamate receptor protein



clamshell twisting rocking

courtesy Tino Wolter