Biomolecular modeling III

Marcus Elstner and Tomáš Kubař

2016, December 13

Analysis of the simulation

Thermodynamic properties

- time averages of thermodynamic quantities
  – 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 to get heat capacity from a single simulation

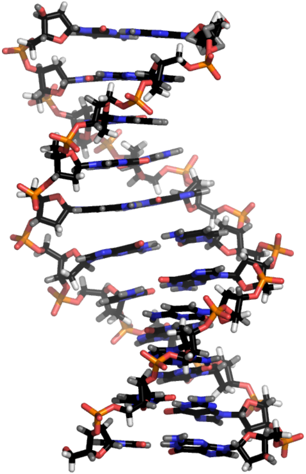
Structure – single molecule in solvent

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 $\vec{r}^{\,\mathrm{ref}}$

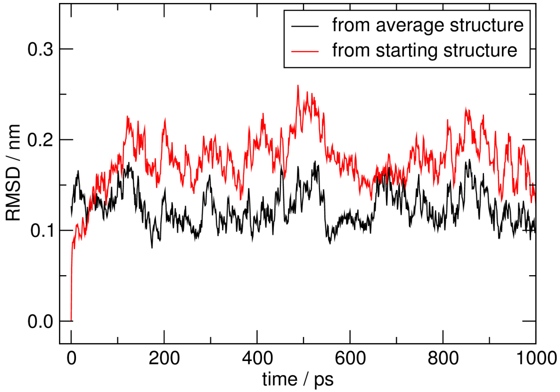
$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\vec{r}_i(t) - \vec{r}_i^{\,\mathrm{ref}}\right|^2}$$

- 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

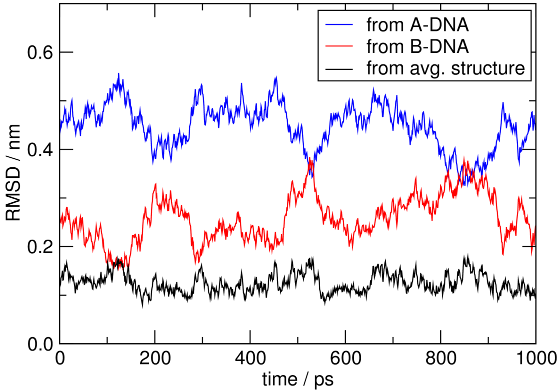
Root mean square deviation

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



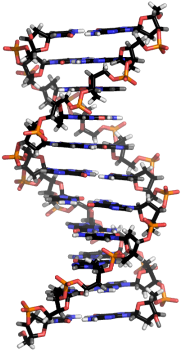
Root mean square deviation

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

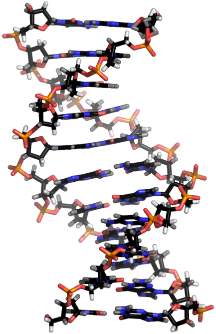


Root mean square deviation

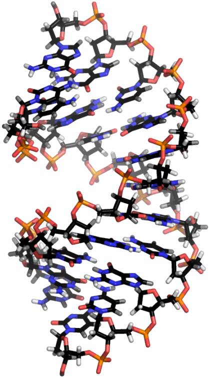
B-DNA



average structure



A-DNA



Magnitude of structural fluctuation

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

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

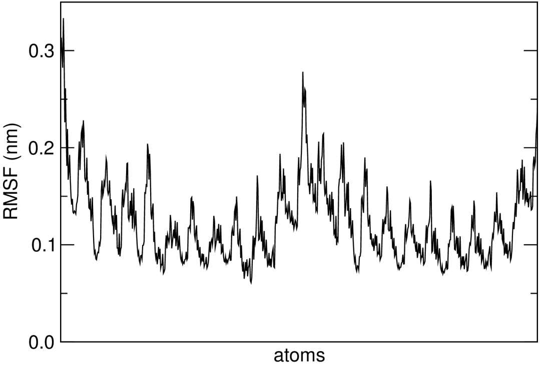
– may be converted to B-factor

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

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

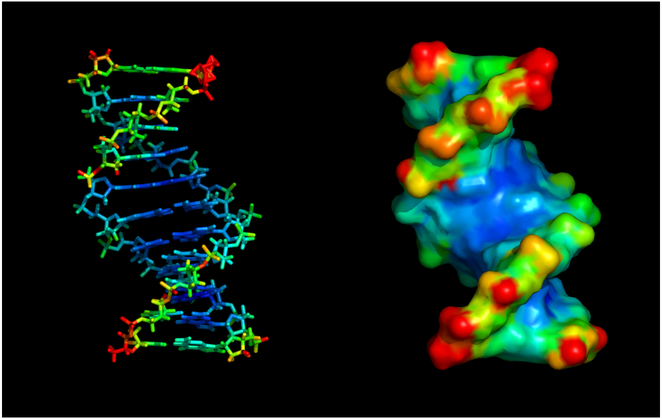
Root mean square fluctuation

RMSF of atomic positions in DNA oligonucleotide



Root mean square fluctuation

RMSF of atomic positions in DNA oligonucleotide

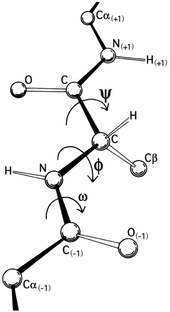


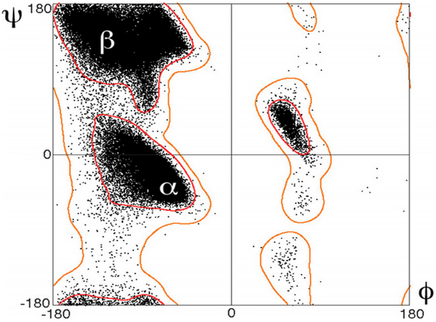
(blue $<$ green $<$ yellow $<$ red)

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

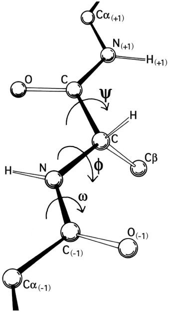


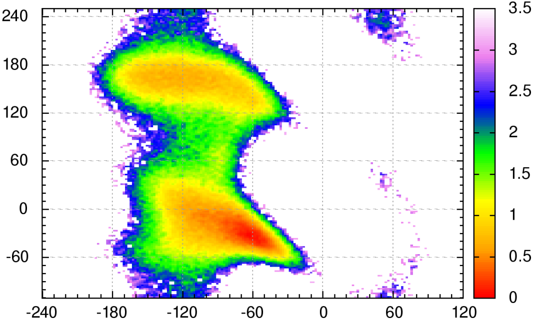


Structure of peptides and proteins

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

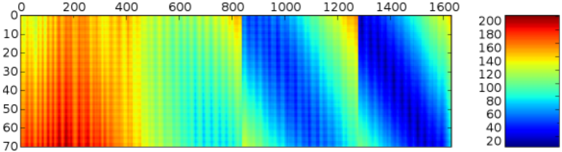


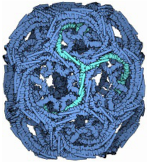


Structure of peptides and proteins

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

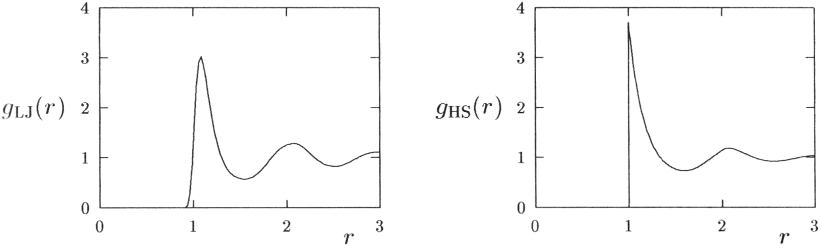
radial distribution functions

- 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



- $g(r)$ 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

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

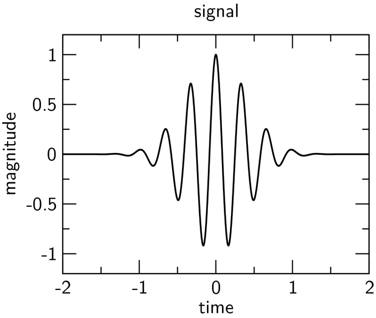
– quantifies the scattering of incoming radiation in the material
– measured in diffraction experiments (X-ray, neutron)

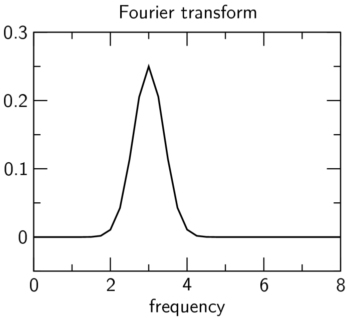
intermission: Fourier transformation

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

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

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





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 and y may exhibit correlation

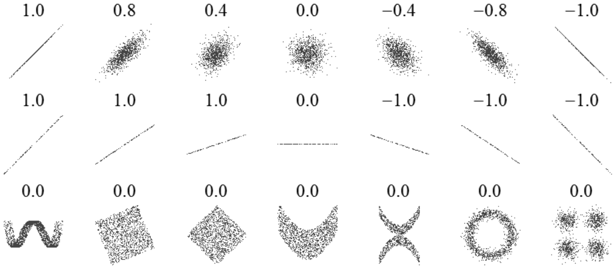
- indicates a relation of x and y, opposed to independence
- Pearson correlation coefficients
  - describe linear relationship between x and y
  - quantities fluctuate around mean values $\langle x\rangle$ and $\langle y\rangle$
  - consider only the fluctuating part
  - introduce correlation coefficient 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 corresponding 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,
averaged over all pairs of such time points,
normalized to take 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

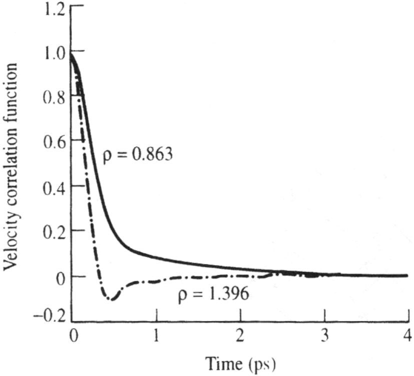
– usually averaged over all atoms i in the simulation

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

– 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:

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

- interesting observable quantities
- important to be able to calculate them from MD
- another way: Einstein relation for D

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

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

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:

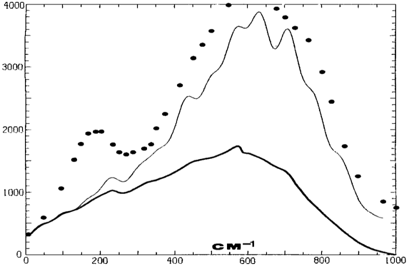
$$c_\mu(t) = \frac{\langle \vec{\mu}_{\text{tot}}(t) \cdot \vec{\mu}_{\text{tot}}(0) \rangle}{\langle \vec{\mu}_{\text{tot}}(0) \cdot \vec{\mu}_{\text{tot}}(0) \rangle}$$

– 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 $r_i \in \{x_i, y_i, z_i\}$

$$\begin{aligned} C_{ij} &= \left\langle (r_i - \langle r_i \rangle) \cdot (r_j - \langle r_j \rangle) \right\rangle_t \qquad \text{or} \\ C_{ij} &= \left\langle \sqrt{m_i}(r_i - \langle r_i \rangle) \cdot \sqrt{m_j}(r_j - \langle r_j \rangle) \right\rangle_t \end{aligned}$$

diagonalization →

eigenvalues – may be expressed as vibrational frequencies

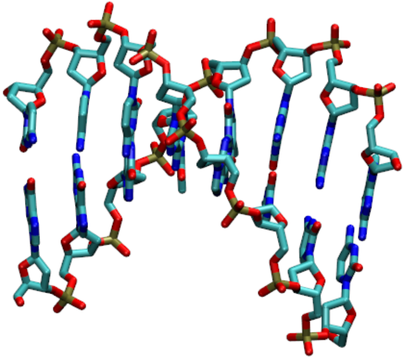
eigenvectors – principal or essential modes of motion

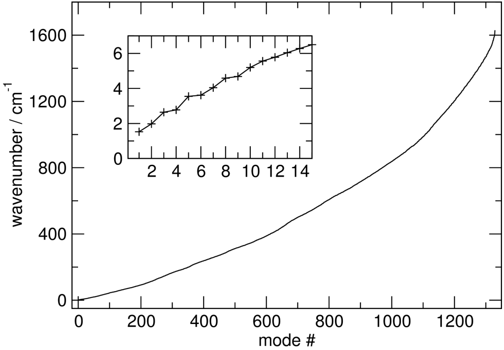
– analogy of normal modes of vibration

– 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

Coarse-grained models

United-atom force fields

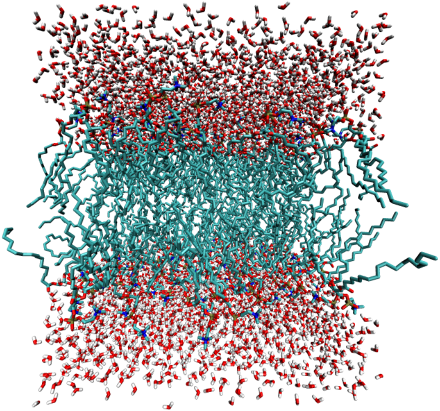
Early biomolecular force fields (e.g. Weiner 1984)
– united-atom approach
– hydrogen atoms considered as condensed to the heavy atom
– mass and charge represent such a group of atoms as a whole
– number of atoms reduced considerably relative to all-atom FF
– popular in the 1990's

This approach works very well for non-polar C–H bonds,
so a methyl group constituting of one united atom works good.

A substitution of a polar O–H group by a single particle
would be very crude (without any correction terms in FF)
→ only non-polar hydrogens are usually condensed with heavy

United-atom force fields

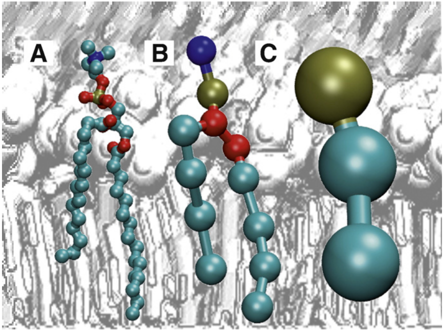
– still used e.g. to describe lipids, where each CH_2 is a united atom



– simulation of a DOPC bilayer in water – Berger FF for the lipid

from the website of Rainer Böckmann

United-atom and coarse-grained force fields



(A) united-atom, (B) specific and (C) generic coarse-grained

from Marrink et al., Biochim. Biophys. Acta 2009

Coarse-grained models

Coarse graining – an advanced and sophisticated approach to reduce the computational expense of simulations

The same idea – reduction of the number of particles
Considered are particles composed of several atoms – beads
The number of inter-particle interactions decreases, reducing the computational expense largely.

The necessary parameters of the force field are often obtained by fitting to all-atom force fields.

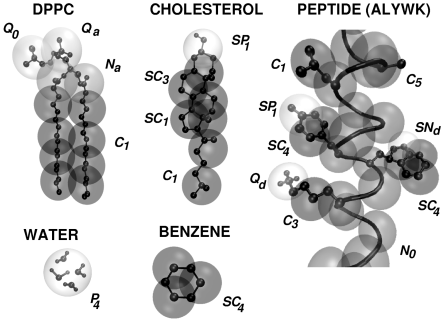
Coarse-grained models

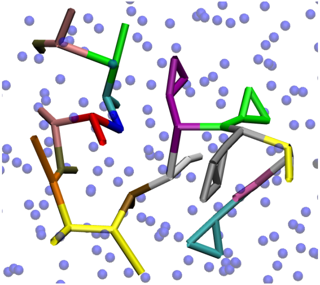
Every bead usually represents several atoms,
and a molecule is composed of several beads.
For the solvent, there is e.g. a 'water bead'
composed of four H_2O molecules.

Note that some of the transferability of all-atom FF is lost
– e.g. secondary structure of proteins is fixed with Martini FF
Also, hydrogen bonding cannot be described with beads!
solution – compensation with Lennard-Jones contributions

Such CG force fields are particularly useful for simulations
of large-scale conformational transitions, which involve
exceedingly large molecular systems, excessive time scales,
or both.

Martini force field

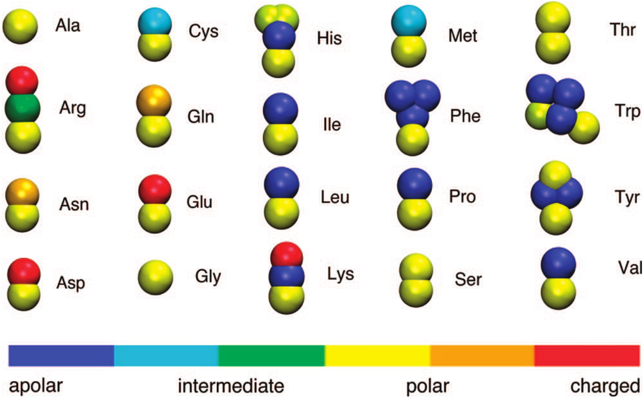




left – mapping of beads onto molecular fragments with Martini FF
– 3 to 4 heavy atoms compose one bead ('4-to-1 mapping')
– mass of beads – 72 u (= 4 H_2O), or 45 u in ring structures
right – a solvated peptide with Martini

from the Martini website

Martini force field



The CG force field Martini – amino acids

from Monticelli et al., J. Chem. Theory Comput. 2008

Acceleration of the simulation

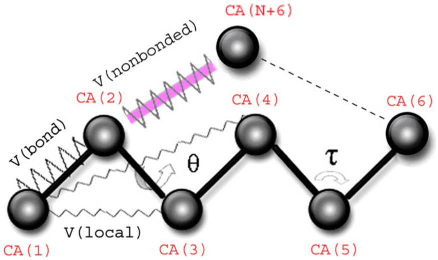
Why does a coarse-grained simulation run faster?

- smaller number of particles → fewer interactions
- long integration time step due to large masses of beads
  – 25 fs with Martini (i.e. 100 fs effectively, see below)
- FF often constructed for use with faster simulation algorithms
  – e.g. cut-off for electrostatics with Martini
- smaller number of DoF → smoother free energy surfaces
  → fewer barriers → acceleration of all processes
  (by a factor of 3 to 8 for Martini, but not uniformly!
  – factor of 4 for acceleration of diffusion in water)

“. . . length and time scales that are 2 to 3 orders of magnitude larger compared to atomistic simulations, providing a bridge between the atomistic and the mesoscopic scale.”

Coarse-grained models

Another example – Vamm force field for proteins,
where every amino acid is represented by a single bead at C-α.



from Korkut & Hendrickson 2009