Biomolecular modeling II

Marcus Elstner and Tomáš Kubař

2016, December 7

The complete equation

$$V(R^N) =$$

$$= \frac{1}{2}\sum_i k_i(r_i - r_i^0)^2 + \frac{1}{2}\sum_j k_j^\vartheta(\vartheta_j - \vartheta_j^0)^2 + \frac{1}{2}\sum_n V_n \cdot \cos\left[n\omega - \gamma_n\right]$$

$$+ \sum_i^N \sum_{j=i+1}^N \left\{ 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^6 \right) + \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}} \right\}$$

and get forces as derivatives with respect to atomic coordinates:

$$F_i^x = -\frac{\partial V}{\partial x_i}$$

Verlet integration method

$$\begin{aligned} r(t+\Delta t) &= 2\cdot r(t) - r(t-\Delta t) + \ddot{r}(t)\Delta t^2 \\ \ddot{r}(t) &= a(t) = \frac{F(t)}{m} = -\frac{1}{m}\frac{\partial V}{\partial r}(t) \end{aligned}$$

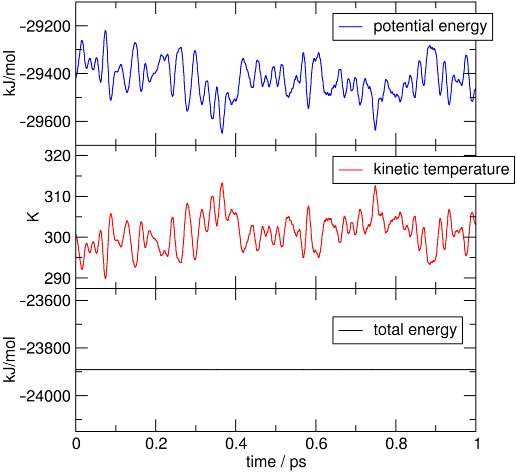
and choose an appropriate time step Δt

Temperature and pressure

what you simulate is what you would measure

Energies and temperature

Solution of equations of motion – conserves total / internal energy



Energies and temperature

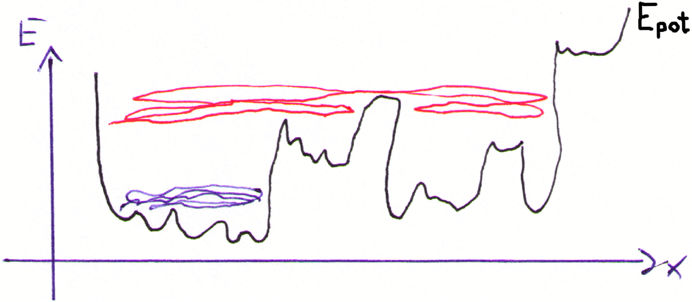
Solution of equations of motion – conserves total / internal energy

what we need – to control basic simulation parameters
– temperature and possibly pressure

significance of temperature
– determines which structures of the system are accessible
– different dynamics at high and at low temperatures

Energies and temperature

high E – multiple different structural 'classes' are reached
low E – restricted available structures



difference $E - E_{\text{pot}}$ corresponds to E_{kin} and temperature

Isolated system

- exchanges with surroundings neither energy (heat / work)
  nor matter (particles)
- total energy of system: $E = E_{\text{kin}} + E_{\text{pot}} = \text{const}$
- individually, E_{kin} and E_{pot} fluctuate in the course of time
  as they are being transformed into each other
- is what we get when using the Verlet method for a molecule

Isolated system

kinetic theory of gases → relation of E_kin and temperature:

$$\begin{aligned} \left\langle E_\text{kin} \right\rangle &= \frac{3}{2} NkT \\ \text{where } \left\langle E_\text{kin} \right\rangle &= \frac{1}{2} \sum_i m_i \left\langle v_i^2 \right\rangle \end{aligned}$$

'local' T – fluctuates in time; may differ between parts of system

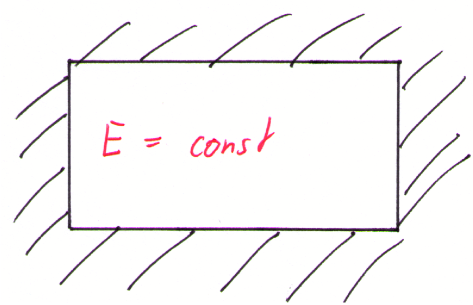
Isolated and closed system

experimental setup (a test tube with a sample)

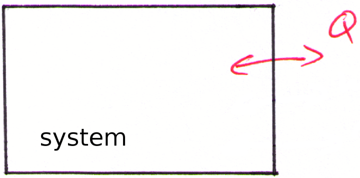
– usually in thermodynamic equilibrium with the surroundings

– temperature of system = temperature of suroundings

isolated system



closed system



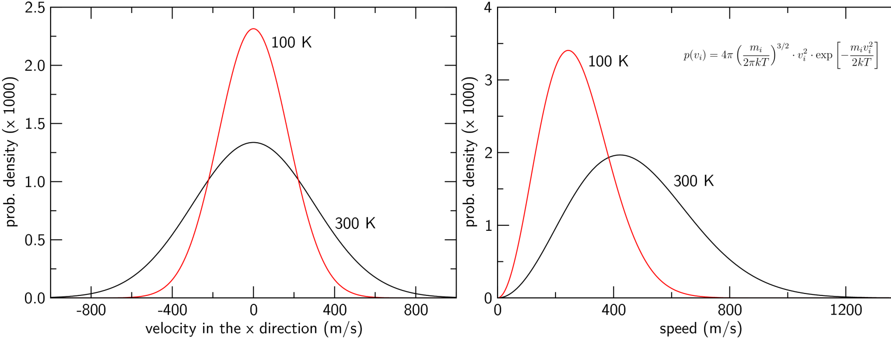
environment (bath)

Closed system

- thermal contact of system with surroundings
- exchange of energy in the form of heat
  until the temperature of surroundings is reached
- canonical ensemble
- velocity / speed of atoms – Maxwell–Boltzmann distribution

Canonical ensemble

Maxwell–Boltzmann distribution of velocity / speed (N_2, IG)



Naïve thermostat – scaling of velocities

in a Verlet MD simulation – ‘instantaneous temperature’ T deviates from the target T_{ref} (of bath = the surroundings)

$$T(t) = \frac{2}{3}\frac{E_{\mathrm{kin}}(t)}{Nk} \neq T_{\mathrm{ref}}$$

$T(t)$ – another name for E_{kin} determined by velocities

simple idea – scale the velocities by a certain factor λ:

$$\begin{aligned} T_{\mathrm{ref}} &= \frac{1}{\frac{3}{2}Nk} \cdot \frac{1}{2} \sum_i m_i \left(\lambda \cdot v_i\right)^2 = \\ &= \lambda^2 \cdot \frac{1}{\frac{3}{2}Nk} \cdot \frac{1}{2} \sum_i m_i v_i^2 = \lambda^2 \cdot T \end{aligned}$$

Naïve thermostat – scaling of velocities

scaling of all velocities by $\lambda = \sqrt{T_\mathrm{ref}/T} \;\rightarrow\; T_\mathrm{ref}$ reached exactly

- rescaling the velocities affects the 'natural' way of evolution of the system
- velocities – not sure if the distribution is correct (M–B)
- importantly, system does not sample any canonical ensemble – very important because every thermodynamic quantity A is obtained as an average:

$$\langle A \rangle = \frac{1}{Z} \int \rho(\vec{r}) \cdot A(\vec{r}) \, \mathrm{d}\vec{r}$$

- if sampling is wrong → wrong density ρ → wrong averages

Berendsen thermostat

How to avoid the drastic changes to the dynamics?
adjust velocities more smoothly, in the direction of T_{ref}

- temperature changes between two time steps according to

$$\Delta T = \frac{\Delta t}{\tau}\left(T_{\mathrm{ref}} - T\right)$$

- rate of change of T (due to the change of velocities) is proportional to the deviation of actual T from T_{ref}
- constant of proportionality – relaxation time τ

Berendsen thermostat

- velocities are scaled by λ:

$$\begin{aligned} T_{\text{new}} &= T + \Delta T = T + \frac{\Delta t}{\tau} \left(T_{\text{ref}} - T\right) \\ \lambda &= \sqrt{\frac{T_{\text{new}}}{T}} = \sqrt{1 + \frac{\Delta t}{\tau} \left(\frac{T_{\text{ref}}}{T} - 1 \right)} \end{aligned}$$

- usually: $\tau = 0.1 - 10$ ps
- T will fluctuate around the desired value T_{ref}
- problem – still does not generate correct canonical ensemble

Nosé–Hoover thermostat

- generates the correct canonical ensemble → ideal choice
- conceptionally and mathematically $>$ difficult to understand
- heat bath is treated not as an external element
  rather as an integral part of the system
  is assigned an additional DOF s with fictitious mass Q
- eqns of motion for this extended system $(3N + 1$ DOF$)$:

$$\begin{aligned} \ddot{r}_i &= \frac{F_i}{m_i} - s \cdot \dot{r}_i \\ \dot{s} &= \frac{1}{Q}\left(T - T_\text{ref}\right) \end{aligned}$$

Temperature and thermostats

fluctuation of temperature – desired property
for canonical ensemble – variance of 'inst. temperature' T:

$$\sigma_T^2 = \left\langle (T - \langle T \rangle)^2 \right\rangle = \left\langle T^2 \right\rangle - \left\langle T \right\rangle^2$$

and relative variance

$$\frac{\sigma_T^2}{\left\langle T \right\rangle^2} = \frac{2}{3N}$$

large number of atoms N: fluctuations $\rightarrow 0$
finite-sized systems: visible fluctuation of temperature
– is correct (feature of the canonical ensemble)

Introducing pressure

chemical reality – constant pressure rather than constant volume

goal – implement such conditions in simulations, too

How to calculate pressure? – first, calculate virial of force

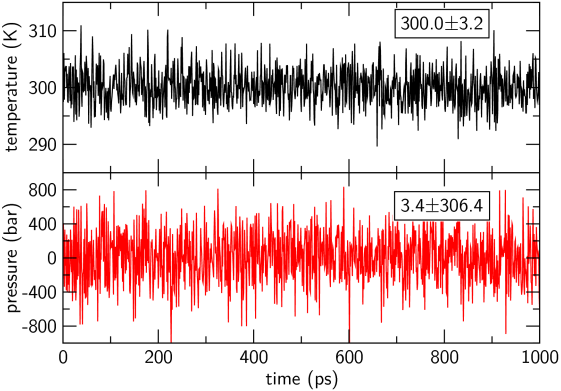
$$\Xi = -\frac{1}{2}\sum_{i<j} \vec{r}_{ij} \cdot \vec{F}_{ij}$$

($\vec{r}_{ij}$ distance of atoms i and j, $\vec{F}_{ij}$ – force between them)

$$P = \frac{2}{3V} \cdot (E_{\text{kin}} - \Xi) = \frac{2}{3V} \cdot \left(\frac{1}{2}\sum_i m_i \cdot |\vec{v}_i|^2 + \frac{1}{2}\sum_{i<j} \vec{r}_{ij} \cdot \vec{F}_{ij} \right)$$

Measuring pressure

T and P in an NPT simulation of a DNA oligomer in water ($T_{\text{ref}} = 300$ K, $P_{\text{ref}} = 1.0$ bar)



Controlling pressure

we can calculate the pressure
– so how do we maintain it at a desired value?

barostat – algorithm that is equivalent of a thermostat,
just that it varies volume of the box instead of velocities

alternatives are available:

- Berendsen barostat
  – direct rescaling of box volume
  – system coupled to a 'force / pressure bath' – piston
- Parrinello–Rahman barostat
  – extended-ensemble simulation
  – additional DOF for the piston

System boundary and the solvent

Biomolecule in solution

typical MD simulations – molecular system in aqueous solution
preferably – 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

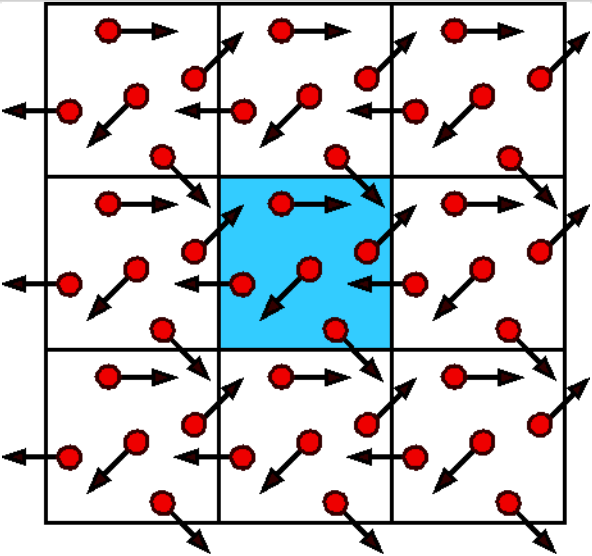
Periodic boundary conditions

- 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

- 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

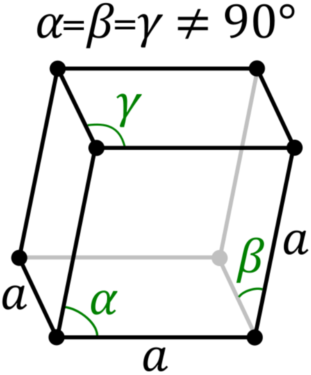
- only coordinates of the unit cell are recorded
- atom that leaves the box enters it on the other side.
- carefull accounting of the interactions of atoms necessary!

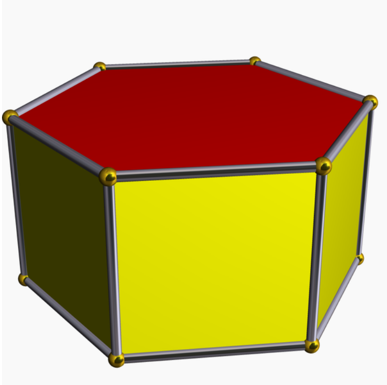
  simplest – minimum image convention:

  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

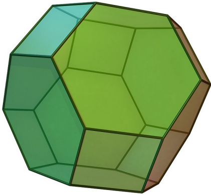


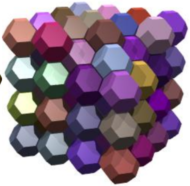


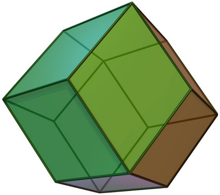
PBC – box shape

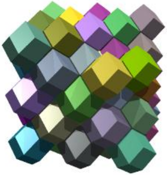
. . . but also more complicated

- truncated octahedral or rhombic dodecahedral
- quite complex equations for interactions & eqns of motion





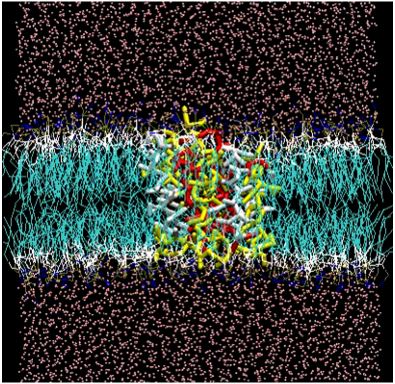


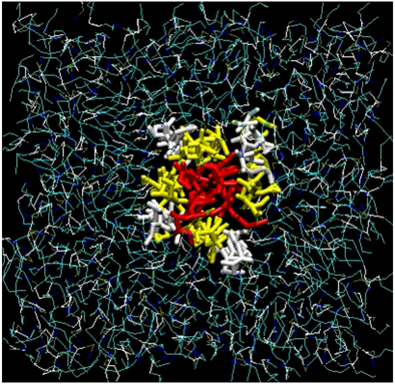


advantage for simulation of spherical objects (globular proteins)
– no corners far from the molecule filled with unnecessary H_2O

PBC – box shape

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

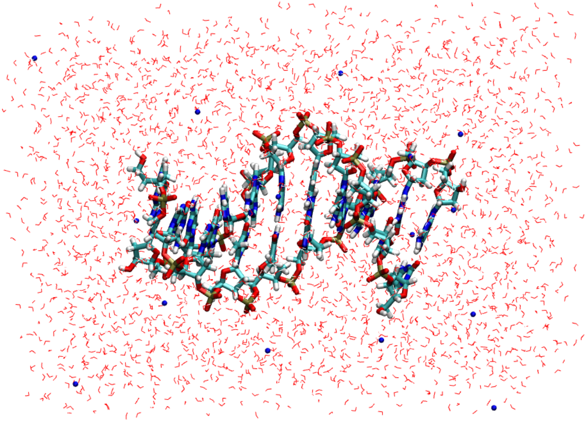




Water in biomolecular simulations

most simulations – something in aqueous solutions

H_2O – usually (many) thousands of molecules



Water in biomolecular simulations

most simulations – something in aqueous solutions

H_2O – usually (many) thousands of 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

→ necessary to pay attention to its description

model of water must be simple enough (computational cost)

and accurate enough, at the same time

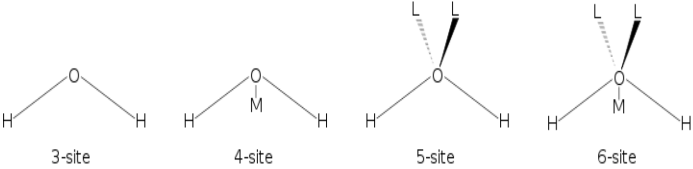
water models – usually rigid

– 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 → 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

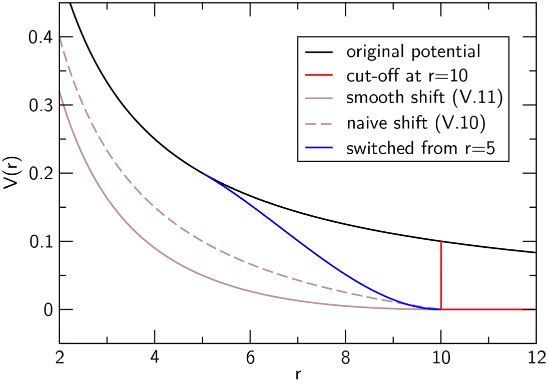
with PBC – infinite number of interaction pairs in principle,
but the interaction gets weaker with distance

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 $(1/r^6)$ $(r_c = 10\ \text{Å})$

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



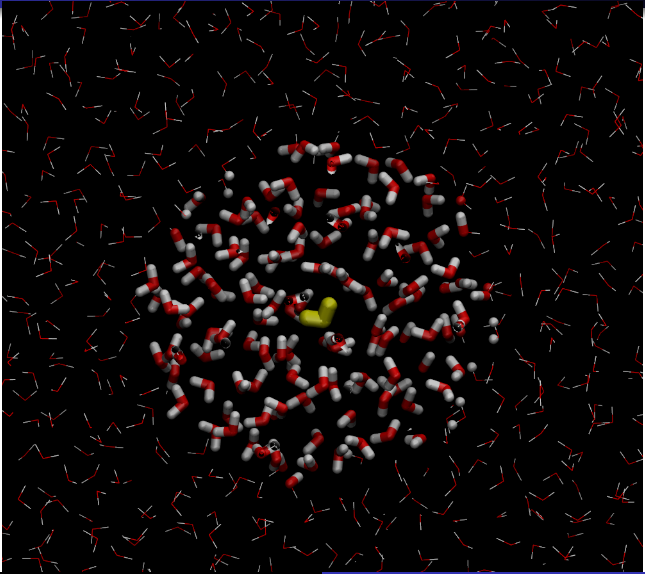
Neighbor lists

cut-off – we still have to calculate the distance for every two atoms
(to compare it with the cut-off distance)
→ 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

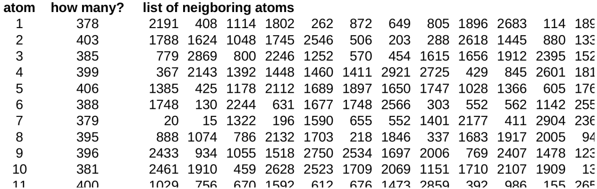
idea: maybe it is only necessary to calculate the interactions
between A and these close atoms – neighbors

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)
→ 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 $\mathcal{O}(N^{\frac{3}{2}})$ 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

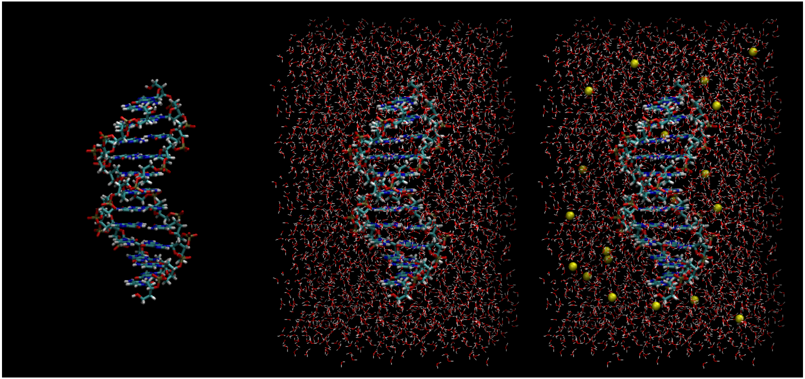
1. build the initial structure
2. bring the system into equilibrium
3. do the productive simulation
4. analyze the trajectory

Tools to build the structure

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



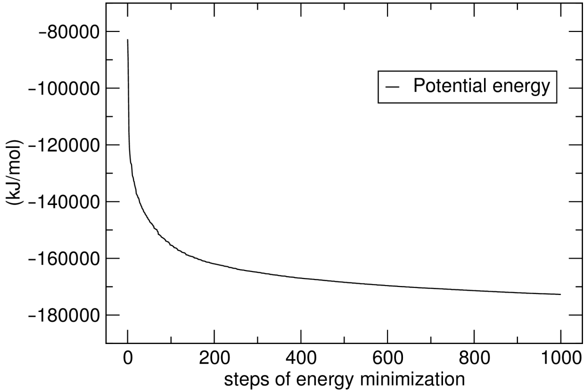
Why equilibrate?

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

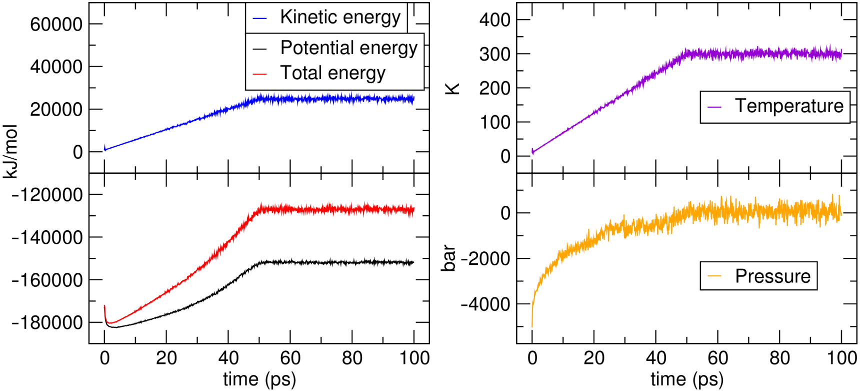
How to equilibrate

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

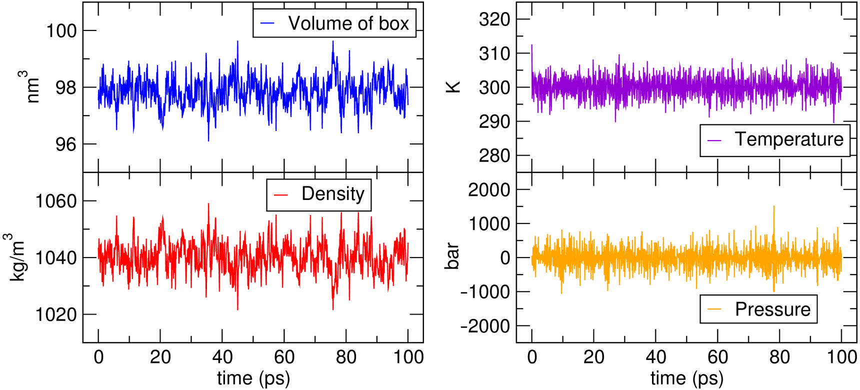


Thermalization



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. . .