

Biomolecular modeling I

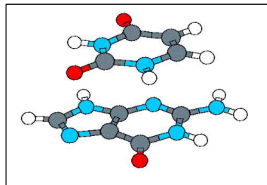
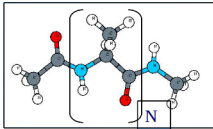
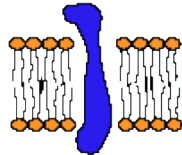
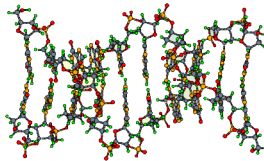
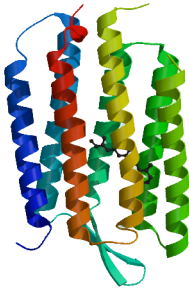
Marcus Elstner and Tomáš Kubař

2016, December 6

Biomolecular structure

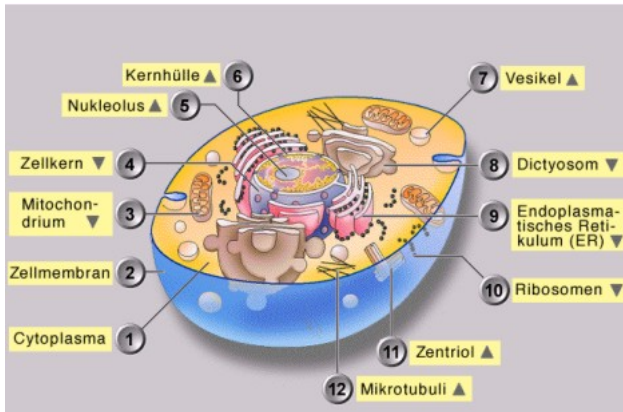
Structural elements of life

Biomolecules – proteins, nucleic acids, lipids, carbohydrates ...



Biomolecular structure

Biomolecules $\rightarrow$ biomolecular complexes $\rightarrow$ aggregates $\rightarrow \dots \rightarrow$
organelles $\rightarrow$ a cell



Biomolecular function

Biophysical processes

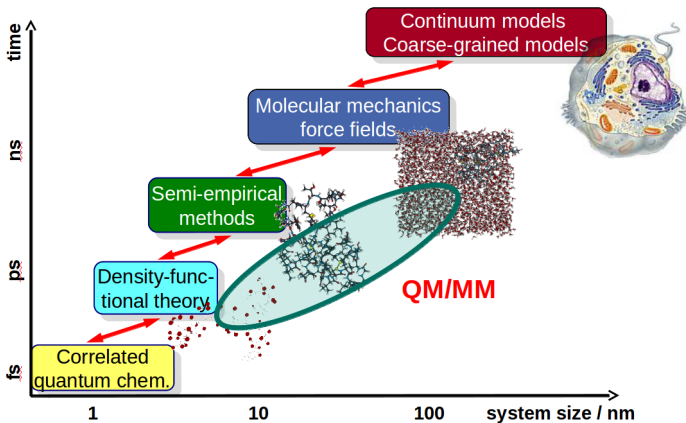
- Bioenergetics – reception, transformation and utilization of energy
- Catalysis – synthesis of substances, metabolism
- Transport – exchange of small molecules, ions up to proteins with the surroundings
- Sensing – detection / recognition / binding in presence of a stimulus

Theoretical / computational biophysics

Molecular modeling

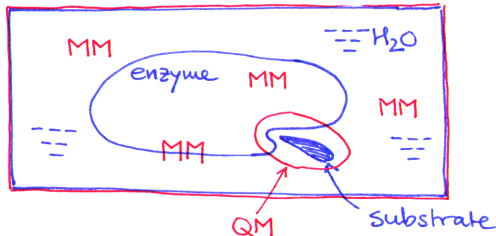
- Focus: understanding on the atomic scale
- The **structure** and **dynamics** determine the **properties** and **function** of biological molecules
- Prediction of experimentally relevant data
- Molecular design of materials with desired properties
- The way to the goal – in general:
solution of quantum mechanical many-body problem
- QM MB approach would be too inefficient
→ apply more approximative methods

Methods



Hybrid method – QM/MM

- Quantum chemistry (QM)
 - bonds created/broken
 - computationally costly
 - DFT or ab initio, up to 100 atoms
 - semi-empirical, up to 1000 atoms
- Molecular mechanics (MM)
 - efficient for up to 100,000 atoms
 - generally – structural properties
- Hybrid QM/MM
 - chemical reactions
 - etc.



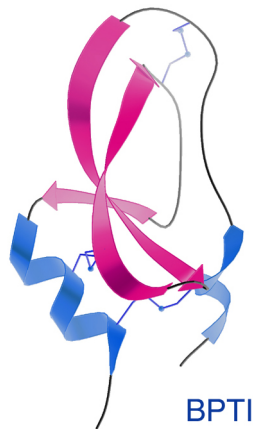
Timeline

- 1687 – Newton, equations of motion
- similar – Hooke, harmonic spring
- ...
- 1946 – molecular mechanics
- 1950s – useful computers
- 1959 – Alder & Wainwright, molecular dynamics of a fluid
- 1975 – molecular dynamics of a protein – Levitt & Warshel, Gelint & Karplus
- 1976 – QM/MM proposed, Levitt & Warshel
- 1990 – significant QM/MM work, Karplus

Timeline

1st simulation of protein dynamics

- BPTI, 58 AAs, in vacuo, 9.2 ps
- McCammon, Gelin & Karplus, Nature 1977
- starting point – crystal structure

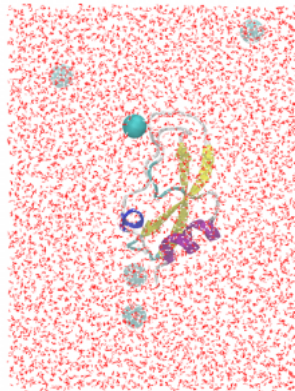


Timeline

simulation of protein in aqueous solution

- BPTI + water, 210 ps
- Levitt & Sharon, PNAS 1988

today's standard – 100,000 atoms, 100 ns



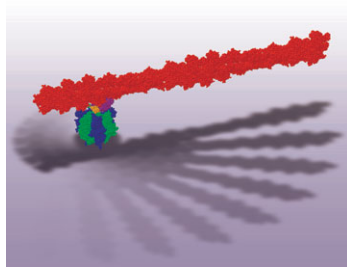
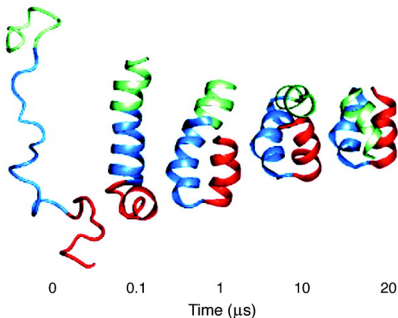
Contact with experimental reality

- X-ray and neutron diffraction
- STM/AFM imaging
- electronic / optical spectra
- electronic and nuclear magnetic resonance
- vibrational / IR spectra
- thermodynamic measurements – DSC, ITC

Applications

Structure and dynamics of complex biomolecular systems

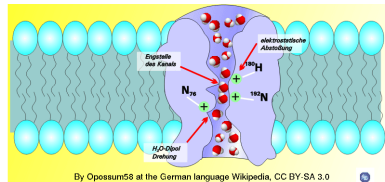
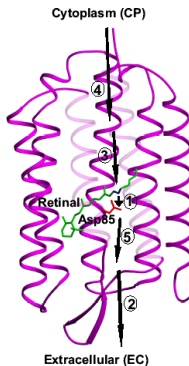
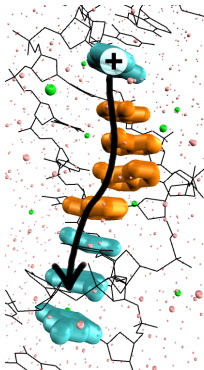
- protein folding
- protein–ligand interaction
- proteins as nanomachines



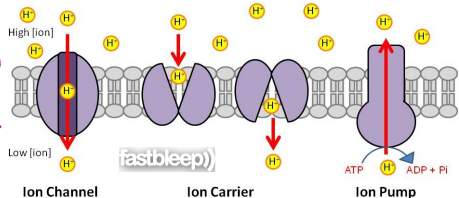
Applications

Transport

- electrons
- protons – across membrane
- water, ions, small molecules, ...



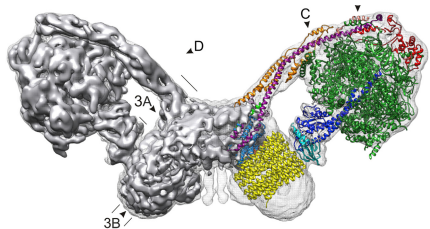
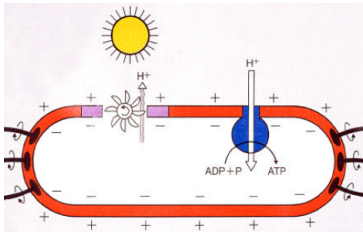
By Opossum58 at the German language Wikipedia, CC BY-SA 3.0



Applications

Enzymes

- catalyzed chemical reactions
- energy conversion (light, chem., mech., gradients)
- bioenergetics



Challenges

- system size limited to ca. 100,000 atoms
- time scales limited to few microseconds
- accuracy of description
 - bonded interaction (vibration, rotation)
 - non-bonded interaction (charge-transfer, polarizability)
- excited electronic states
- quantum character of movement of nuclei

Some can be tackled with

- further development of available methods
- combination of various methods
- optimization of algorithms
 - parallelization, $\mathcal{O}(N)$ – linear scaling

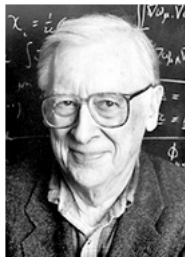
Nobel prizes for computational chemistry

The Nobel Prize in Chemistry 1998



Walter Kohn

Prize share: 1/2



John A. Pople

Prize share: 1/2

The Nobel Prize in Chemistry 1998 was divided equally between
Walter Kohn *"for his development of the density-functional theory"*

Nobel prizes for computational chemistry

The Nobel Prize in Chemistry 2013



© Harvard University
Martin Karplus



Photo: © S. Fisch
Michael Levitt



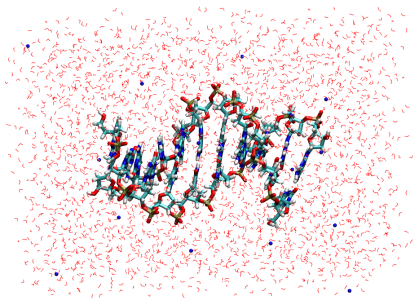
Photo: Wikimedia
Commons
Arieh Warshel

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale*

Biomolecular simulation

Elementary body – **atom**

Usually – one molecule/complex of interest (e.g. protein, NA)



Simulation vs. reality

One molecule instead of many

Tiny volume of $\approx 10^{-21}$ L
instead of $\approx 10^{-5}$ L

Dynamics – short time scale
of max. $\approx 10^{-5}$ s

Biomolecular simulation

Each atom – x, y, z coordinates

“A protein is a set of coordinates.” (Gromacs, A. P. Heiner)

```
Peptide in lipid+water
48609
  1LYS      N      1      4.360      4.040      8.207      0.2882      0.4041 -0.5575
  1LYS     H1      2      4.416      4.119      8.178      0.4151      0.4652 -0.1555
  1LYS     H2      3      4.340      4.037      8.306      0.8750     -1.7473 -0.4570
  1LYS     H3      4      4.407      3.954      8.185      0.3515      0.2061      0.2987
  1LYS      CA      5      4.231      4.037      8.136      0.0777      0.2898     -0.1753
  1LYS      HA      6      4.162      4.112      8.174     -0.6060     -0.8009      0.7924
  1LYS      CB      7      4.262      4.069      7.990     -0.3012      0.3768     -0.2373
  1LYS     HB1      8      4.360      4.025      7.969      0.6025      1.3517      1.6736
  1LYS     HB2      9      4.300      4.171      7.998      0.1892      0.2300     -0.6695
  1LYS      CG     10      4.161      4.049      7.877     -0.5841      0.0286      0.0762
  1LYS     HG1     11      4.056      4.067      7.900     -0.9362     -1.1078     -0.5669
  1LYS     HG2     12      4.148      3.942      7.863     -3.1278      0.1078      1.2506
  1LYS      CD     13      4.196      4.123      7.749      0.0459     -1.0686     -0.3967
  1LYS     HD1     14      4.298      4.095      7.721     -0.3753     -3.6647      0.4016
  1LYS     HD2     15      4.205      4.228      7.778      3.4358     -1.4671      0.3786
  1LYS      CE     16      4.088      4.101      7.644     -0.3622      0.1377     -0.2469
  1LYS     HE1     17      3.992      4.138      7.679     -1.1725     -0.1480     -2.0367
  1LYS     HE2     18      4.073      3.994      7.628     -3.0282      0.2507      0.9872
  1LYS      NZ     19      4.124      4.174      7.521     -0.0992      0.0204     -0.2407
  1LYS     HZ1     20      4.056      4.156      7.449     -2.5018      1.3804      1.5513
  1LYS     HZ2     21      4.118      4.275      7.528     -1.2171     -0.0196     -0.4614
```

Simulation vs. reality

Why should we want to simulate molecular systems?

Experiment – the molecule has its genuine properties

Simulation – we need a model to describe the interactions of atoms
– the quality of the model is decisive

Advantage of simulation – structure on atomic level defined

Structure → function

Combination of experiment and simulation – added value

Molecular mechanics

classical description of molecules

Motivation

- To investigate the **function** of biomolecules, we need to characterize its **structure** and **dynamics**.
- We will look how the molecules are moving
 - **Molecular Dynamics**
- For this, we need to calculate the **forces** on atoms and the **energy** of the system
- Energy from quantum mechanics / quantum chemistry seemingly easy $E = \langle \Psi | \hat{H} | \Psi \rangle$ but not quite possible for large molecular systems

Motivation

$$\left[-\frac{\hbar^2}{2m_e\lambda'^2} \nabla'^2 - \frac{e^2}{4\pi\epsilon_0\lambda'} \right] \phi' = \mathcal{E}\phi'$$

$$[ij|kl] = \int d\mathbf{x}_1 d\mathbf{x}_2 \chi_i^*(\mathbf{x}_1) \chi_j(\mathbf{x}_1) r_{12}^{-1} \chi_k^*(\mathbf{x}_2) \chi_l(\mathbf{x}_2)$$

$$\begin{aligned} F_{\mu\nu} &= \int d\mathbf{r}_1 \phi_\mu^*(1) f(1) \phi_\nu(1) \\ &= \int d\mathbf{r}_1 \phi_\mu^*(1) h(1) \phi_\nu(1) + \sum_a^{N/2} \int d\mathbf{r}_1 \phi_\mu^*(1) [2J_a(1) - K_a(1)] \phi_\nu(1) \\ &= H_{\mu\nu}^{\text{core}} + \sum_a^{N/2} 2(\mu\nu|aa) - (\mu a|a\nu) \end{aligned}$$

$$|\Phi\rangle = c_0|\Psi_0\rangle + \sum_{\substack{ra \\ r < s}} c_a^r |\Psi_a^r\rangle + \sum_{\substack{a < b \\ r < s}} c_{ab}^{rs} |\Psi_{ab}^{rs}\rangle + \sum_{\substack{a < b < c \\ r < s < t}} c_{abc}^{rst} |\Psi_{abc}^{rst}\rangle + \dots$$

$$E_{(1)}^{(2)} = \sum_{a,b=1}^{N/2} \sum_{r,s=(N/2+1)}^K \frac{\langle ab|rs\rangle(2\langle rs|ab\rangle - \langle rs|ba\rangle)}{\epsilon_a + \epsilon_b - \epsilon_r - \epsilon_s}$$

Idea of molecular mechanics

- often – well localized bonding orbitals (organic molecules)
- **idea** – similar bonds have similar strength and properties
e.g. similar C–H σ -orbitals $\rightarrow$ all C–H bonds are ‘similar’
- possibly model with a harmonic spring? (simplest function)

$$E(x) = \frac{1}{2}k(x - x_0)^2$$
$$F(x) = -\frac{\partial E(x)}{\partial x} = -k(x - x_0)$$

- 2 parameters k and x_0 with defined meaning
– can be obtained from spectroscopy

Concept of (atom, bond. . .) type

- let us use harmonic springs for covalent bonds
- we do not want to parametrize k and x_0 each bond separately
- use just several sets (k, x_0) , for different **types** of bonds

Why can we expect such ‘unification’ to work?

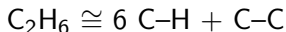
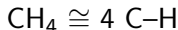
Concept of (atom, bond...) type

Spectroscopy

- every C–H bond: length 1.06–1.11 Å,
frequency ca. 3100 cm^{-1} , in any molecular environment

Thermochemistry

- heat of formation – roughly additive:



Concept of (atom, bond...) type

How to identify the **atom types**? – chemical ideas

i) hybridization

- different types for sp^3 carbon (4 bonds) and sp^2 C (3 bonds)
- different functions for bonds of types C–C, C=C and C≡C
- determine the parameters (k , x_0) with some selected molecules, typical for the binding situation
- example: use C₂H₆, C₂H₄, C₂H₂ and benzene for k , x_0

Concept of (atom, bond...) type

How to identify the **atom types**? – chemical ideas

ii) polarity

- an atom bonded to electronegative atom – electron deficient
→ affects its bonding to other atoms
- example: C–C bond in **O=CH–C**... is affected
and needs to be parametrized differently from apolar C–C
→ an atom type for carbonyl C introduced

Biomolecular force fields

– usually 20 types for C, 10 for N and 5 for O and H

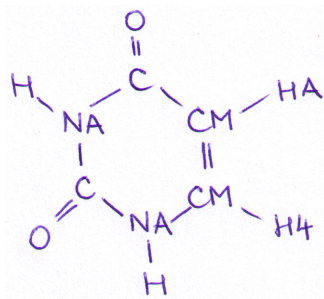
Concept of (atom, bond...) type

AMBER types for carbon:

- C - sp² C carbonyl group
- CA - sp² C pure aromatic (benzene)
- CB - sp² aromatic C, 5&6 membered ring junction
- CC - sp² aromatic C, 5 memb. ring HIS
- CK - sp² C 5 memb.ring in purines
- CM - sp² C pyrimidines in pos. 5 & 6
- CN - sp² C aromatic 5&6 memb.ring junct.(TRP)
- CQ - sp² C in 5 mem.ring of purines between 2 N
- CR - sp² arom as CQ but in HIS
- CT - **sp³** aliphatic C
- CV - sp² arom. 5 memb.ring w/1 N and 1 H (HIS)
- CW - sp² arom. 5 memb.ring w/1 N-H and 1 H (HIS)
- C* - sp² arom. 5 memb.ring w/1 subst. (TRP)

Concept of (atom, bond...) type

AMBER atom types in a molecule of uracil



Interactions between atoms

Bonded

- mediated by, and resulting directly from covalent bonds
- cover all of the quantum-mechanical phenomena between pairs of atoms with effective potentials
- harmonic springs between atoms (also angles and dihedrals)

Non-bonded

- longer-range interactions
 - charge–charge (Coulomb) and van der Waals (vdW)
- between molecules and distant parts of one molecule

Coulomb interaction

- idea – condense electrons in each atom with the nucleus
→ **effective atomic charge** $q_i = -Q_i + Z_i$:

$$E_{QQ} = \frac{1}{2} \sum_{ij} \frac{q_i \cdot q_j}{R_{ij}}$$

- needs to be defined for **every atom**
rather than atom type – this would be too crude
- from quantum-chemical calculations
of typical (bio)molecular fragments
 - amino acid residues and peptide bonds for proteins
 - nucleobases, sugars and phosphate groups for DNA/RNA

Coulomb interaction

How to calculate atomic charges?

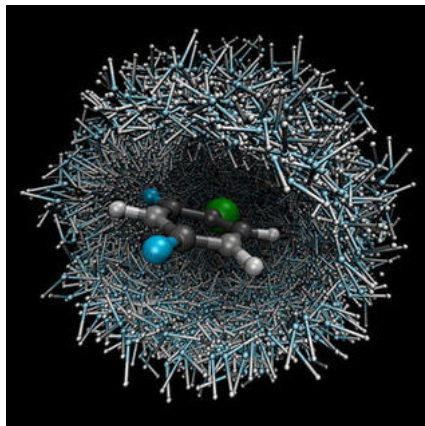
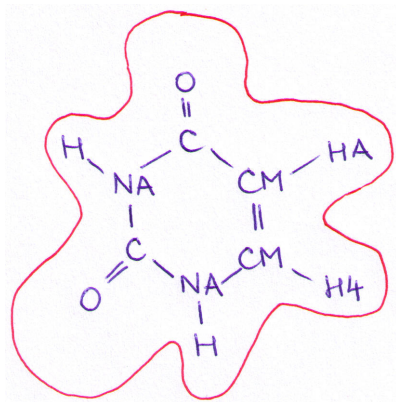
- popular – **potential-derived** charges:
 - 1 calculate the electron density in the molecule
 - 2 get **electrostatic potential** at surface of the molecule
 - 3 **fit** point electric charges on atoms to reproduce the ESP

Possible improvement – **polarizable force field**

- atomic polarizability α_i is assigned to every atom i
- external field induces atomic dipole $\vec{\mu}_i = \vec{\alpha}_i \cdot \vec{E}$

Coulomb interaction

surface of the uracil molecule



van der Waals interaction

Pauli repulsion

- electrons with the same spin avoid spatial overlap



- modeling:

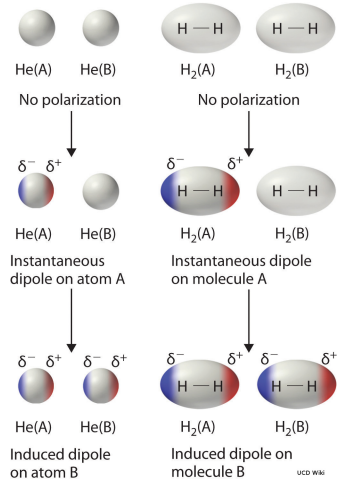
$$E_{\text{ex}} = \exp[a - b \cdot R_{ij}]$$

$$E_{\text{ex}} = \left(\frac{\sigma}{R_{ij}} \right)^{12}$$

van der Waals interaction

dispersion due to correlation

- correlation – between electrons irrespective of spin, retained on longer distances
- instantaneous dipole $\rightarrow$ induced dipole $\rightarrow$ interaction
- orientation of dipoles is correlated – **attractive** interaction
- R^{-6} -dependence, proportional to polarizabilities



van der Waals interaction

most common function: **Lennard-Jones 12-6** potential

$$V(r) = 4\varepsilon \left(\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right)$$

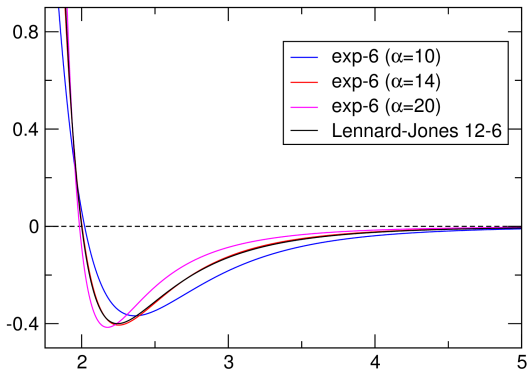
- 2 parameters – σ and ε

repulsive: $\exp \left[\alpha \left(1 - \frac{r}{\sigma} \right) \right]$ sometimes better than $\left(\frac{\sigma}{r} \right)^{12}$
→ **exp-6** potential

- may be a better choice for phase transitions
- e.g. MM water would not freeze below 0 °C with LJ 12-6

note: **phase transitions** are difficult to simulate generally

van der Waals interaction



van der Waals interaction

parametrization

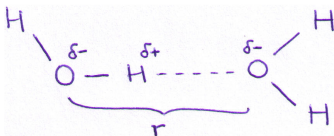
- challenging task in general
- fitting of params to experimental / quantum-chem. data
- e.g. relation to density and heat of vaporization
 - obvious in organic liquids – major interaction

Hydrogen bonding

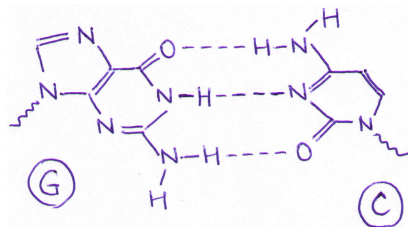
- crucial interaction in biomolecules
- interplay of several kinds of interactions
- typical binding energies: 20 kJ/mol
 - higher for strongly polarized or even charged molecules
 - or if there are several H-bonds (nucleobase pairs)
- early force fields – special potential functions for H-bonding
- modern force fields – no special treatment

Hydrogen bonding

H₂O dimer



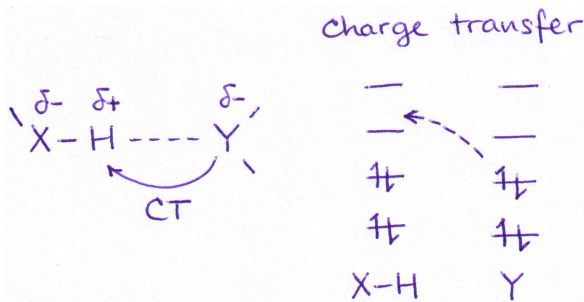
guanine:cytosine base pair



Hydrogen bonding

- Coulomb interaction is dominant
- vdW interaction
 - may become important, especially in weakly bound systems
 - crucial e.g. for angular dependence in $\text{H}_2\text{CO} \dots \text{H}_2\text{O}$ etc.
- **charge transfer** contribution
 - cannot be covered by force fields due to constant charges
 - may be included in other terms effectively

Hydrogen bonding



charge transfer – into the σ^* orbital

→ weakening of the X-H bond → red shift in the IR spectrum

Parametrization of bonded interactions

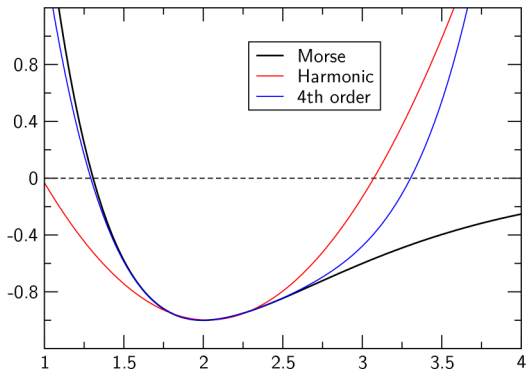
Bonds

- harmonic approximation (Taylor expansion up to 2nd order)

$$E(r) = \frac{1}{2}k(r - r_0)^2$$

- 2 parameters – equilibrium distance and force constant
- works in a narrow interval of distances
- often sufficient (vibrations are within the interval)
- if bonds are to be created or broken (chemistry)
 - another solution has to be sought
 - probably leave molecular mechanics ☹

Parametrization of bonded interactions



- for accurate vibration frequencies
 - quartic terms can be important to describe the curvature

Parametrization of bonded interactions

Angles

- harmonic approximation for the angle deformation

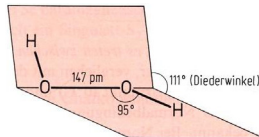
$$E_{\text{bend}}(\vartheta) = \frac{1}{2}k_{\vartheta}(\vartheta - \vartheta_0)^2$$

- 2 parameters needed – equilibrium angle and force constant
- from experiment (vib-rot spectra) or quantum chemistry

Parametrization of bonded interactions

Dihedral angles

- describe the rotation around covalent bonds



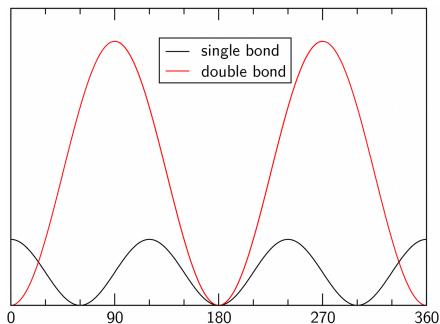
- defined by 4 atoms
- potential energy – periodic function of the dihedral angle:

$$E(\omega) = \sum_{n=1,2,3,4,6} V_n \cos [n \cdot \omega - \gamma_n]$$

- V_n – amplitude (barrier), n – periodicity, γ – phase shift

Parametrization of bonded interactions

Dihedral angles – example: C–C single and C=C double bonds



The complete equation

$$\begin{aligned}
 E(R^N) = & \\
 = & \frac{1}{2} \sum_i k_i (r_i - r_i^0)^2 + \frac{1}{2} \sum_j k_j^\vartheta (\vartheta_j - \vartheta_j^0)^2 + \frac{1}{2} \sum_n V_n \cdot \cos[n\omega - \gamma_n] \\
 + & \sum_i^N \sum_{j=i+1}^N \left\{ 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right) + \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}} \right\}
 \end{aligned}$$

Molecular dynamics simulation

how to get things moving

Equations of motion

$$m \cdot \ddot{r} = F$$

ordinary differential equations of 2nd order

- have to be solved numerically
- solution proceeds in discrete **steps** of length Δt
- numerical integration starts at time t_0 ,
where the **initial conditions** are specified
 - the positions r_0 and the velocities v_0
- calculations of forces at r_0 to get accelerations a_0
- then, an integrator calculates r and v at time $t_0 + \Delta t$
- accelerations $\rightarrow$ step $\rightarrow$ accelerations $\rightarrow$ step $\rightarrow \dots$

Verlet integration method

for the development of the method: take a virtual step in positive time and in 'negative' time,

and Taylor expansion up to 2nd order:

$$r(t + \Delta t) = r(t) + \dot{r}(t) \cdot \Delta t + \frac{1}{2} \ddot{r}(t) \cdot \Delta t^2$$

$$r(t - \Delta t) = r(t) - \dot{r}(t) \cdot \Delta t + \frac{1}{2} \ddot{r}(t) \cdot \Delta t^2$$

add both equations – eliminate the velocity $\dot{r}$:

$$r(t + \Delta t) = 2 \cdot r(t) - r(t - \Delta t) + \ddot{r}(t) \Delta t^2$$

$$\ddot{r}(t) = a(t) = \frac{F(t)}{m} = -\frac{1}{m} \frac{\partial V}{\partial r}(t)$$

Verlet integration method

another but equivalent formulation – velocity Verlet

$$\begin{aligned}r(t + \Delta t) &= r(t) + v(t) \cdot \Delta t + \frac{1}{2}a(t) \cdot \Delta t^2 \\v(t + \Delta t) &= v(t) + \frac{1}{2}(a(t) + a(t + \Delta t)) \cdot \Delta t\end{aligned}$$

yet another – Leap-frog

$$\begin{aligned}v(t + \frac{1}{2}\Delta t) &= v(t - \frac{1}{2}\Delta t) + a(t) \cdot \Delta t \\r(t + \Delta t) &= r(t) + v(t + \frac{1}{2}\Delta t) \cdot \Delta t\end{aligned}$$

both: better numerical precision than Verlet normal form

Δt – crucial parameter

Let us say: we want to obtain a trajectory over a time interval T

- we perform M steps
- we have to evaluate the forces on atoms $M = T/\Delta t$ times

Computational cost of the calculation of forces

- major computational effort
- determines how many steps we can afford to make

Δt – crucial parameter

- we neglect contributions in Δt^3 and higher orders
→ error per step in the order of Δt^3
- keep the step short → make the error small
but need too many steps to simulate certain time T
- make the step long → cut computational cost
but increase the error and decrease stability
- compromise needed

Δt – crucial parameter

- fastest motion – hydrogen atoms, period around 10 fs
- rule of thumb – stable integration with $\Delta t \leq \frac{1}{10}$ fastest period
- practically, Δt of 1 fs is used (2 fs with special treatment)
- 1M calculations of forces needed for a trajectory of 1 ns
- large systems – multi-ns simulations routinely, μ s possible