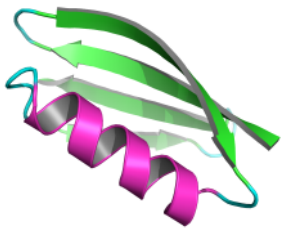
GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Faculty of Physics
Friedrich-Hund-Platz 1
37077 Goettingen

master-fp.physik.uni-goettingen.de

Advanced Lab Course

Main Research Area: Biophysics and Physics of Complex Systems
(M.Phy.1401/1402)



Experimental Manual BK.MDS

Molecular Dynamics Simulations of Proteins

BK.MDS
Version 2018.10.18/E

Max Planck Institute for Biophysical Chemistry
Greg Bubnis, Julian Brennecke, Gabor Nagy,
Matteo Aldeghi and Martin Werner

Summary

Molecular dynamics simulations are a method to describe the dynamics of macro-molecules such as proteins, lipids or synthetic polymers. The movement of every atom is calculated on a computer by numerically integrating Newton's equations of motion.

This experiment aims to introduce the fundamental methods to you. Therefore, you are going to simulate a short polymer chain, a "toy-model" that behaves similarly to a small protein. The steps necessary to set up a simulation system will be performed. Subsequently, you will be introduced to the analysis methods, e. g. the calculation of structural order parameters, interaction energies, and free energy profiles.

Exercises

This experiment is structured into four parts building on each other.

1. In the first part the 13-atom chain will be will be simulated, aiming at a fundamental understanding of the basic functionality of the molecular dynamics software package GROMACS. Based on these simulations, structural and thermodynamic properties of the system will be calculated.

   (a) At first, we simulate our (model) protein for 50 ns in a vacuum environment. Watch the trajectory: What conformational states do you observe?

   (b) Calculate and analyze the different energy terms that govern the chain's behavior. These include the bond stretching and bending energies as well as the Lennard-Jones (LJ) non-bonded interactions between atoms. Are trends (e.g. correlated changes) in the energy terms consistent with what you observed viewing the trajectory?

   (c) Calculate the radius of gyration and end-to-end distance of the chain using the GROMACS tools ("g_tools") `g_gyrate` and `g_dist`. Are these "order parameters" able to distinguish different conformational states?

   (d) As a next step, calculate the free energy profile $G(\xi)$ using radius of gyration as ξ and then with end-to-end distance as ξ.

   (e) Make a 2D plot of r_g versus L. Can you explain the bizarre shape of the distribution in terms of the conformations of the protein?

   (f) Also plot r_g versus potential energy. Approximate the change in potential energy between folded and unfolded states from this plot. Use this value and the change in free energy from the free energy profile $G(r_g)$ to estimate the change in entropy between the states. Is the increase/decrease consistent with your intuition? What do you expect to happen if we heat up or cool down the system?

   In what ways does this toy model represent (or not) the physics (hydrophobic effect, hydrogen bonding) of a real protein in water? How is the observed behavior of the toy system similar and different to a real protein?

2. In the second part you will simulate the model protein at different temperatures to see the effects on the free energy landscape. For this, you will make an analysis script that automates the analysis from the last section, allowing you to vary model/simulation parameters and quickly see the results.

   (a) First, you will simulate the system at 270 K and use this simulation to write/test your analysis script.

   (b) Next, simulate the system at 300, 330 and 400 K, and use the analysis script to compute free energy profiles for R_g. How do the free energy profiles change with temperature? How does the statistical uncertainty change with temperature?

3. In the third part you will vary the bending rigidity of the chain to see how this influences the folding/unfolding behavior

   (a) First, you will have to make copies of the topology (.top) file, and edit the bond angle stiffness parameters to have three different values: 10, 35, and 45 kJ/mol. After simulating and analyzing each of these different cases, how does the free energy landscape change?

   (b) Next, we will use another g_tool, `g_angle`, to look at the distribution of bond angles as a function of the bond angle stiffness. Is the compact state the same with k=10 kJ/mol and k=35 kJ/mol?

4. In the final exercise, we will change the parameters of our model protein so it behaves like an 'entropic spring' – a limiting case where the potential energy of a chain is always zero.

   (a) First, you will do a pen and paper exercise to explore the available configuration space of short (N=3,4,5) chains restricted to a 2D lattice.

   (b) Next, to simulate this scenario, you must modify the chain in two ways:

   - The LJ interactions will be weakened so they are repulsive spheres
   - The bond angle stiffness will be zeroed, allowing the chain to bend into any position so long as atoms do not overlap.

   With the modified topology, simulate the chain and compute the free energy as a function of end-to-end distance. Is a Hookean (harmonic) spring a reasonable model for the computed free energy? How can the chain act as a spring if it does not store potential energy?

Experimental setup and equipment

In the experiment we will provide you with all the necessary equipment. The software will be run using a Linux OS. Please make sure you know the most basic commands for Linux. We are going to use GROMACS and PyMol. If you want to use them on your own, you can Download both for free. Download GROMACS from `http://www.gromacs.org` or PyMol from `http://www.pymol.org`.

Questions for your preparation

Please make sure to answer the following questions in preparation for the experiment and include them in your protocol. (Do **NOT** answer them in a question - answer style in your protocol!)

1. Which interactions are relevant on atomistic scales?
2. What is a hydrogen bond?
3. What is the hydrophobic effect?
4. What is a protein?
5. What is the primary, secondary and tertiary structure?
6. Which simplifications are used in MD simulations to simulate complex molecules?
7. Which additional simplifications are we making here by simulating a simple chain of atoms?
8. What is a force field in MD?
9. What is an order parameter, how is it useful (or not) for describing protein dynamics?
10. In the funnel theory of protein folding, what is at the bottom of the funnel and what are its (relevant) thermodynamic properties?

Theory

Proteins

Proteins are macromolecules, present in all lifeforms and performing a variety of tasks. Examples of these tasks are transport of ions or small molecules (e.g. oxygen, ammonium), conversion between different forms of energy, and catalysis of chemical reactions.

Proteins are composed of 20 different amino acids. Most of the proteins have a well defined three dimensional structure which is found in a process called folding.

The common chemical structure of every amino acid is a backbone consisting of one nitrogen and two carbon atoms. Connected to the first carbon atom is a side chain which differs between the amino acids and determines its chemical properties. The hydrogen on the nitrogen and the oxygen on the second carbon are important atoms for the hydrogen-bond network which stabilizes the spatial structure of the protein.

The structure of a protein can be split into three elements:

- The sequence of amino acids is called the primary structure of the protein. The only interaction on this level is a covalent binding between neighboring amino acids. The primary structure of a protein is encoded by its genes. This genetic information is stored in the DNA/RNA of the organism.
- Segments of the amino acids form so called secondary structure. The most frequently observed secondary structure elements are α-helices and β-sheets. Both of these structures are stabilized by hydrogen bonds between backbone atoms.
- The global structure of a protein is called tertiary structure. This structure is mainly stabilized by Coulomb interactions and hydrogen bonds. Additionally, tertiary structure can also be stabilized by the hydrophobic effect or disulphide bridges. The latter are covalent bonds between the side chains of Cysteine amino acids.

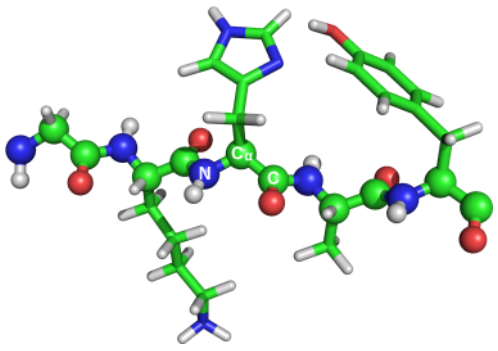


Figure 1: Example of five amino acid long chain. Backbone atoms are shown as spheres side chains are rods.

In addition, very complex tasks (such as the DNA replication) require the assembly of molecular machines consisting of several folded proteins and sometimes non-protein elements. The spatial configuration of parts in these assemblies is their quaternary structure. This structure is stabilized by the same interactions as the tertiary structure.

Protein Folding and Free Energy Landscapes

The efficient catalysis of chemical reactions, or the transport of specific molecules often require proteins to fold into well-defined, compact structures. To perform their biological functions, these "globular" proteins fold efficiently and spontaneously into their "native" compact states under typical cellular conditions. For these proteins, misfolded or unfolded states that are persistent can inhibit their proper function and give rise to different diseases. Some proteins need to interact with a large number if of different partner molecules, or require a high degree of structural flexibility to preform their biological function. These "disordered" proteins have many possible conformational states, with no well-defined native state. Larger proteins with complex biological functions often have both globular and disordered regions. These "domains" can often fold and perform their part of function, even if the rest of the protein is not present.

The behavior of a protein is determined by its free-energy landscape. The dominant theoretical model for protein folding defines the native states of globular proteins as minima at the bottom of their deep, funnel-shaped free-energy landscapes. Disordered proteins in contrast have shallow and rugged free-energy landscapes with many, equally likely states. Is the landscape deep or shallow and smooth or rugged? How do factors such as temperature, amino acid sequence, pH, and many others change the shape of the landscape? To probe the features of these landscapes we combine the results of computer simulations with the knowledge of thermodynamics.

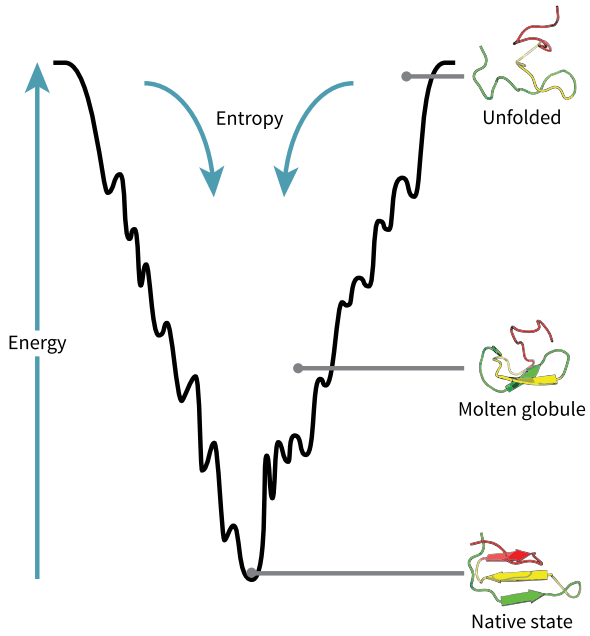


Figure 2: Schematic illustration of a protein folding funnel. Picture credit: Thomas Splettstoesser (www.scistyle.com) - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=28353539

The free energy difference between two states is defined as $\Delta G = \Delta H - T\Delta S$ (the enthalpy H and potential energy U are interchangeable for our purposes). ΔG is therefore a sensitive balance between U and S that dictates the relative stabilities of folded, partially folded, and unfolded states. As the formula suggests, the balance can be shifted by the temperature T, and influenced by interactions with other molecules such as ions, small

molecules, and other macromolecules. One scenario that could give rise to a funnel-shaped landscape is when the folded state has a combination of a low potential energy U, due to strong molecular interactions, and a low entropy S resulting from its well-defined shape. In contrast, the unfolded state has both a higher potential energy and entropy. In the absence of solvation effects, this scenario is plausible; however, for proteins under physiological conditions the solvent thermodynamics play a much stronger role and can not be neglected. Moreover, to disentangle the solvent, protein and interfacial contributions to the free energy "budget" is non-trivial and an active area of research.

The toy-model protein we simulate here, depicted in Fig. 3, does away with much of the complexity of a protein in a biological environment. For instance, it is not simulated in a solvated environment, but just in vacuum. In addition, the potential energy only has terms (described below) for van der Waals interactions and bond stretching/bending. Despite its simplicity, this model retains the essential features that give rise to a folding free energy landscape. The simplicity of the model also allows us to simulate many folding/unfolding transitions over the course of seconds on a desktop computer.

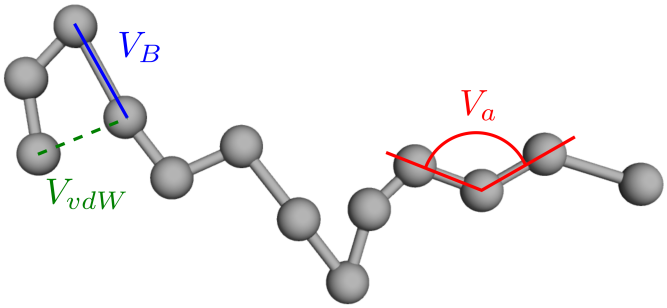


Figure 3: Illustration of the 13-atom model protein chain and the potential energy components of its force field. The functional forms of the van der Waals, bending and angle terms are defined below.

Based on the molecular dynamics simulatioins we want to calculate the free energy between the different conformational states. In an NVT ensemble at equilibrium, a system of N particles with $3N$ Cartesian coordinates will occupy a microstate i with coordinates $\mathbf{x}_i$ with a probability $p(\mathbf{x}_i) \propto e^{-\beta U(\mathbf{x}_i)}$, where U is the potential energy and $\beta = 1/(k_B T)$. This is the well known Boltzmann distribution. As a result, the ratio of probabilities of two configurations $\mathbf{x}_i$ and $\mathbf{x}_j$ is determined by their potential energies $p_i/p_j = e^{-\beta(U_i - U_j)}$.

These probabilities are additive, so if we consider a volume A of the configuration space, then $p_A \propto \int_A e^{-\beta U(\mathbf{x}_i)}$. An important result from this (not proven here) is that the free energy difference between two states A and B, defined as above is

$$\frac{p_B}{p_A} = e^{-\beta(G_B - G_A)}, \tag{1}$$

where G is the Gibbs free energy. (in our simulated system, the box volume is fixed so we can use Gibbs free energy interchangably with Helmholtz free energy). This can be re-arranged to a more useful form

$$\Delta G_{AB} = -k_B T \ln(p_B/p_A) \tag{2}$$

which means that the ratio of probabilities between two states, p_B/p_A, a quantity we can measure using molecular dynamics simulations, gives us information about the free energy between states. We are not restricted to two states, rather we can define successive discrete states (maybe 20 to 100) along some order parameter $\xi(\mathbf{x})$ – a scalar function, sometimes also called a "reaction coordinate" – and compute the free energy at each point along the path via

$$G(\xi) = -k_B T \ln(p(\xi)) - G_0 \tag{3}$$

where G_0 is an arbitrary additive constant. The challenge is therefore to find a good order parameter $\xi(\mathbf{x})$ that indicates the degree of folding and can distinguish different metastable states.

Two order parameters we will compute for this purpose are the radius of gyration R_g and end-to-end distance, defined here. The radius of gyration describes the extension of a chain. A high radius of gyration corresponds to an extended chain while a low radius of gyration is characteristic for a compact configuration. It is calculated by $R_g^2 = 1/N \sum_{k=1}^{N} (r_k - r_{mean})^2$. The end-to-end distance is the distance between the first and the last atom of a protein.

Molecular dynamics Simulation

Molecular dynamics (MD) simulations are used to simulate molecular systems of 10 to 10^6 atoms for nano- to microseconds. The exact solution of such a system is given by the time dependent schroedinger equation for the nuclei and electrons. However, solving the schroedinger equation is computationally demanding making it unfeasible for larger systems. To be able to compute larger systems, MD simulations make three approximations:

- The mass of the electron is much less than the mass of the atom core. Therefore, the electrons move much faster than the atom cores. The Born-Oppenheimer-Approximation takes this difference in speed into account and separates the motion of the electrons from the motion of the cores. The approximation describes electrons as moving while the atom cores are fixed. The influence of the electron dynamics on the atom core motion can be described by an effective potential which is only dependent on the atom core position.
- The effective potential is approximated by a sum of analytical functions (x^n, $\cos(x)$). All of the potentials together are called a force field. Each of the potential functions are based on the chemical structure (bonded interactions) or on the long range interactions (non-bonded). The parameters of the potentials are obtained from quantum chemical calculations or by comparison to experimental results. Here, the interactions in our model chain have the same functional forms as in a fully atomistic forcefield, but the parameters do not match any specific set of atoms.
- The potential gives rise to a force which is used to calculate the atom core motions. In a classical approximation the forces are integrated numerically using Newton's equations of motion. The integration is achieved using the leap-frog algorithm with the atom coordinates $r(t)$ and the atom velocities $v(t)$:

$$v\left(t + \frac{\Delta t}{2}\right) = v\left(t - \frac{\Delta t}{2}\right) + \frac{F(t)}{m}\Delta t$$

$$r\left(t+\Delta t\right)=r(t)+v\left(t+\frac{\Delta t}{2}\right)\Delta t$$

The integration step Δt is chosen to fit the fastest motion of the system. In proteins, the fastest motion are the hydrogen fluctuations which typically move on a timescale of 10^{-14} s. To account for this motion an integration step of 10^{-15} s will be used in most molecular dynamics simulations. However, we use a so called coarse grained system which unites multiple atoms into a single bead. This allows us to use an integration step of 10^{-14} s.

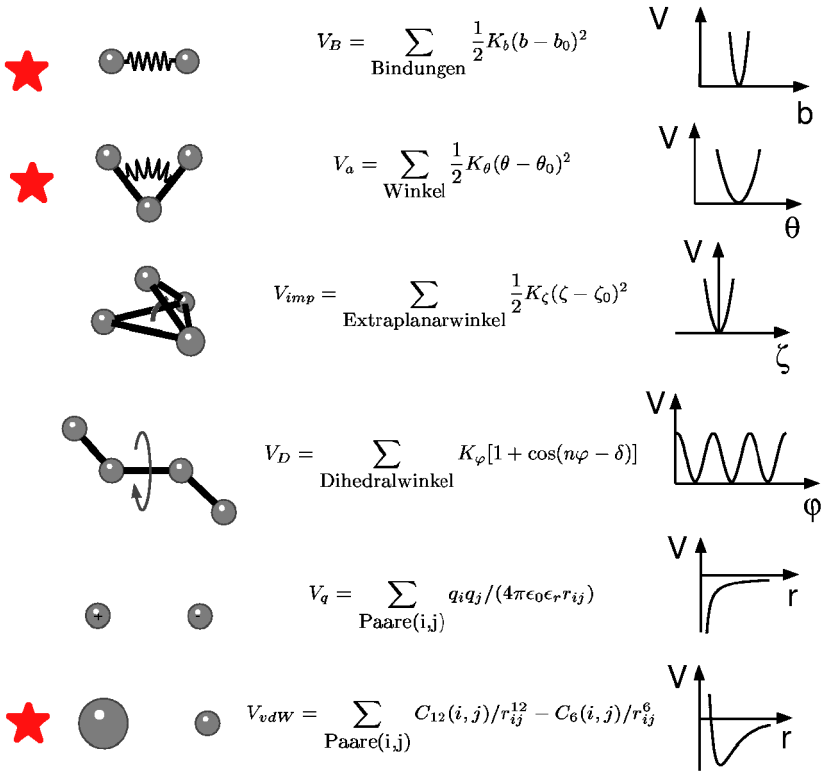


Figure 4: Visualization of the potential components of a typical atomistic force field. The present, simplified model uses only the terms indicated with red stars.

Literature and Links

- GROMACS User Manual (erhältlich auf der GROMACS-Homepage)
- J. M. Berg, J. L. Tymoczko und L. Stryer, *Biochemie*, Spektrum Akademischer Verlag, 2007
- P. W. Atkins und J. de Paula, *Physikalische Chemie*, Wiley-VCH, 2006
- Leopold PE; Montal M; Onuchic JN. "Protein folding funnels: a kinetic approach to the sequence-structure relationship". Proc. Natl. Acad. Sci. USA **89** 87215 (1992)
- W. F. van Gunsteren und H. J. C. Berendsen, "Molekül̈dynamik-Computersimulationen; Methodik, Anwendungen und Perspektiven in der Chemie", Angewandte Chemie **102**, 1020-1055 (1990)
- D. Frenkel and B. Smit, *Understanding Molecular Simulation*, Elsevier, 2002
- `http://www.gromacs.org` (GROMACS)
- `http://manual.gromacs.org/current/` (GROMACS Online Reference)
- `http://www.pymol.org` (PyMol)

Appendix

General usage of GROMACS

To start a simulation using GROMACS the following files are required:

- A file containing the starting coordinates of the system. This is called a `PDB` file (e.g., `startstructure.pdb`).
- A topology (`topol.top`) file containing all required force field parameters and the number of components (proteins, ions, water molecules).
- A simulation file (`md.mdp`) containing the run time parameters of the simulation. This are the time of the simulation, reference temperature and pressure, details for electrostatic calculations, etc.

When all the files are present a single starting file `topol.tpr` can be created from all these files by calling `grompp -f md.mdp -c startstructure.pdb -p topol.top -o topol.tpr`. Using this file the simulation can be started using: `mdrun -s topol.tpr -c finalstructure.pdb -v` The final structure (atom coordinates) of the simulation will be written to the file `finalstructure.pdb`. In addition further files will be written by GROMACS:

- `traj.trr`: Coordinates and velocities of every atom (only relevant for the a restart of the simulation)
- `traj.xtc`: Coordinates of every atom (used for the analyses)
- `ener.edr`: Different energy contributions (binary format)

By using `g_energy -f ener.edr -o energy.xvg` the different energy contributions from `ener.edr` are converted to a text file `energy.xvg`. The content of the file can be visualized using xmgrace, gnuplot, Excel or any other software.