Luis Vollmers | Dr. Maria Reif | Prof. Martin Zacharias

Molecular Dynamics (SS2025)
Exercise 7

Potential of Mean Force of an Amyloid Layer Separation via Umbrella Sampling

Report Tasks

These simulations will take a considerable amount of time to complete. You should first test how long a single production run (for one umbrella window) takes, either on your own computer or in the CIP-Pool. Based on this, you can reduce the number of simulation steps to achieve a shorter runtime. For reference, we provide output files from full-length runs. In your analysis, please include both your own simulation results and the reference data (if your simulations used fewer steps). If you observe any differences between your results and the reference data, discuss them in your report.

1. Comment `02_simprep.sh` line by line and include it in the report (as a screenshot or similar).
2. Include plots of the histograms and the FES and discuss their appearance.
3. Report your result for ΔG_{bind} and compare it to the literature ($-50.5\,\text{kcal}\,\text{mol}^{-1}$).

Acknowledgement

This exercise was inspired by a tutorial offered by Justin Lemkul via his website[1]. The Lemkul Lab hosts it at the Virginia Tech University[2].

Introduction

This exercise explores the usefulness and applicability of free energy calculations, building on the previous exercise which focused on solvation free energy via alchemical transformations. Here, we estimate the free energy of binding by sampling two binding partners at various spatial distances. The larger binding partner is usually called the receptor, and the smaller one is the ligand. Receptor-ligand complexes are crucial in biomolecular signal transduction, enzymatic catalysis, and therapeutic efficacy.[3] Binding free energy, the energy change when a ligand binds to its receptor, is fundamental in understanding these processes. As with solvation free energy, there are many approaches to obtain the binding free energy, each with its strengths and weaknesses.
For this exercise, we analyse the Aβ(1-42) amyloid fibril associated with Alzheimer's disease using molecular dynamics (MD) simulations.[4] This amyloid fibril consists of five protofilaments

[1] `http://www.mdtutorials.com/gmx/umbrella/index.html`
[2] `https://www.thelemkullab.com/`
[3] `doi.org/10.1021/ct3008099`
[4] `doi.org/10.1073/pnas.0506723102`

with beta sheets interconnected via their side chains. The protofilaments are parts of the 42 amino acid-long peptides that give the amyloid its name.[5] In a series of simulations, the outer protofilament is pulled away from the remaining amyloid fibril and is extensively sampled along the distance coordinate, also known as a *collective variable*. Sampling along this collective variable is achieved using harmonic biasing potentials in a procedure known as umbrella sampling (US) simulations.[6] This method results in distance distributions that resemble an umbrella shape. The sampling data is then evaluated using an advanced algorithm to produce a free energy surface, from which the binding free energy is obtained.[7]
Lemkul and Revan precisely calculated the binding free energy (ΔG_{bind}) for this system, predicting a value of $-50.5\,\text{kcal}\,\text{mol}^{-1}$.[8] The goal of this exercise is to replicate this value using `GROMACS` built-in functionalities for conducting umbrella sampling simulations.

Theoretical Considerations

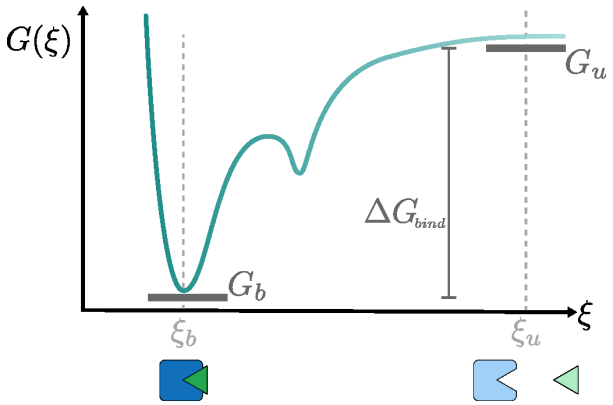

Figure 1: Visualisation of the binding process along the collective variable ξ. The free energy G approaches a minimum for the bound state, described by ξ_b. For further clarification, the bound and unbound complexes are depicted with geometrical shapes beneath the x-axis.

The binding free energy ΔG_{bind} is generally defined as the difference between the free energy of the bound state G_{b} and the unbound state G_{u}. Thus, the binding free energy represents the reversible work released during complex association. The free energies of the bound and unbound states are related to the average probability $\langle P \rangle$ of encountering these states in a given environment, such as an aqueous solution.[9]

$$\Delta G_{\text{bind}} = G_{\text{b}} - G_{\text{u}} = -k_{\text{B}} T \ln\left(\frac{\langle P_{\text{b}} \rangle}{\langle P_{\text{u}} \rangle}\right) \tag{1}$$

The probabilities given in equation (1) can be defined with respect to an arbitrary *collective variable* ξ along which the binding can be described. For ligand binding phenomena, a classical choice for ξ is the centre-of-mass (COM) distance with the receptor, denoted as r_{com}. Although, any observable can be chosen for ξ, e.g. dihedral angles or RMSDs. The average probability of encountering a certain r_{com} distance can be obtained only via the partition function, Z_{NPT}.

$$\langle P(r_{\text{com}}) \rangle = \frac{\int_{\mathbf{X}} d\mathbf{X}\, \delta(r_{\text{com}} - R_{\text{com}}(\mathbf{X})) e^{-\beta H(\mathbf{X})}}{Z_{\text{NPT}}} \tag{2}$$

[5] doi.org/10.1080/13506129.2020.1835263
[6] doi.org/10.1016/0021-9991(77)90121-8
[7] https://doi.org/10.1002/jcc.540130812
[8] https://doi.org/10.1021/jp9110794
[9] https://doi.org/10.1073/pnas.0409005102

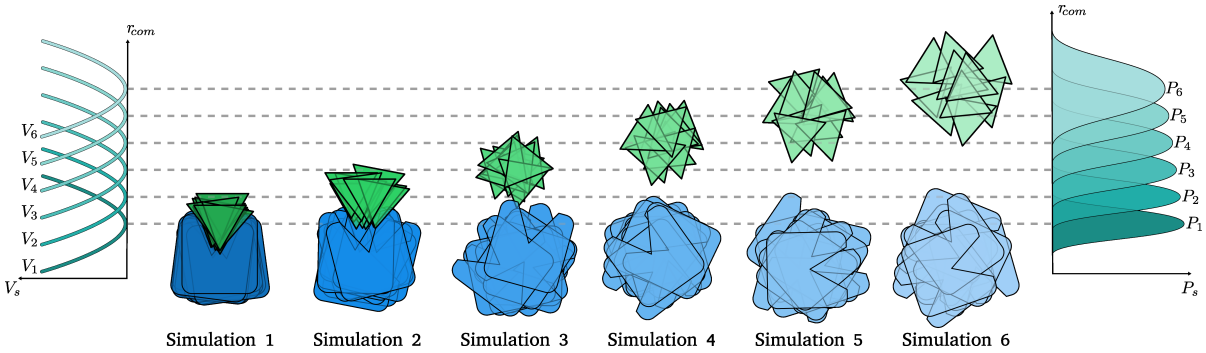

Figure 2: Conceptual visualisation of umbrella sampling for an arbitrary receptor-ligand complex. The harmonic potentials on the left keep the ligand (green triangle) restrained at a certain distance from the receptor (blue rectangle). The sampling is visualised via superimposed copies of the respective molecules in the six simulation windows. The resulting COM distances are shown in the graph on the right as probability distributions.

The function $R_{\text{com}}(\mathbf{X})$ evaluates the COM of the configuration $\mathbf{X}$ and is compared to an arbitrary r_{com} value via the Dirac delta function, $\delta(...)$. Integration occurs over all configurations of the system, $\int_{\mathbf{X}}$, causing the dependency on $\mathbf{X}$ to vanish for $\langle P(r_{\text{com}})\rangle$. The bound and unbound states can now be defined with respect to the reaction coordinate: beyond a certain COM distance threshold, the intermolecular interactions cease; thus, the binding partners are considered fully separated.
Unfortunately, this equation cannot obtain ΔG_{bind} from simulation since the system's partition function is practically unobtainable. Mainly, the intermolecular attraction is so high for tightly bound complexes that dissociation events would never be sampled within feasible MD simulation times. Therefore, a different approach is used, employing harmonic potentials along the COM distance to enhance sampling across all regions of the binding/unbinding process. The exact *stratification strategy* may vary, but typically, for binding processes, multiple simulations (s) are conducted, each with a biasing potential V_s resembling a harmonic spring potential. This potential is defined by the spring constant k_s and the equilibrium COM distance r^s_{com}.

$$V_s(r_{\text{com}}) = k_s \left(r^s_{\text{com}} - r_{\text{com}}\right)^2 \tag{3}$$

Each biased simulation results in a biased probability distribution, P^s_{com}, that preferentially samples r_{com} values close to the reference distance r^s_{com} (see figure 2). However, it can be shown mathematically that the expression for P^s_{com} still depends on the partition function Z_{NPT} and the unbiased probability distribution $\langle P(r_{\text{com}})\rangle$.[10]

$$\langle P_s(r_{\text{com}})\rangle = \frac{Z_{\text{NPT}}}{Z^s_{\text{NPT}}} \langle P(r_{\text{com}})\rangle e^{-\beta V_s(r_{\text{com}})} \tag{4}$$

This equation can be rearranged for the unbiased probability as follows:

$$\langle P(r_{\text{com}})\rangle = \langle P_s(r_{\text{com}})\rangle \cdot \frac{\langle e^{-\beta V_s(r_{\text{com}})}\rangle}{e^{-\beta V_s(r_{\text{com}})}} \tag{5}$$

Seemingly, not much is achieved by performing multiple simulations with biasing potentials, as $\langle P(r_{\text{com}})\rangle$ and the unbiased partition function remain unknown. The partition function of the biased simulation would also need to be calculated. Nevertheless, Ferrenberg and Swendsen[11]

[10] https://doi.org/10.1002/jcc.540130812
[11] doi.org/10.1103/PhysRevLett.61.2635

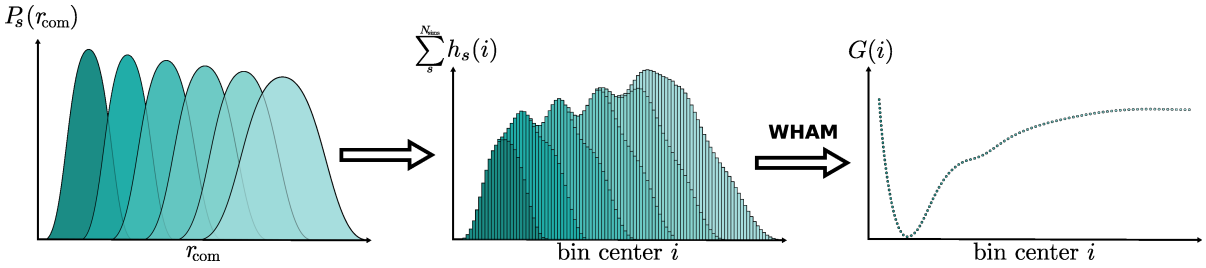

Figure 3: Depiction of data generation and the WHAM algorithm. The COM distances are collected from the N_{sims} simulations. The binned histograms are summed, and then the WHAM equations are applied iteratively until the unbiased probability distribution $\langle P(i)\rangle$ emerges, from which the free energy surface $G(i)$ can be readily obtained.

developed the weighted histogram analysis method (WHAM),[12] which employs similar mathematical techniques to those implemented in the BAR method from the previous exercise.[13] WHAM operates using histograms of the collective variable; instead of continuous values of r_{com}, only certain histogram bins with their center values i are considered.
The lengthy and complex derivation of the WHAM equations is beyond the scope of this tutorial. However, the resulting two equations will be discussed in detail. The objective is to accurately estimate the unbiased, discretised probability distribution $\langle P(i)\rangle$. Sampled COM distances from each simulation are collected and binned into a histogram with N_{bins}, chosen by the user. Consequently, the resulting histogram $h_s(i)$ from simulation s should contain counts mostly close to the reference distance given by the corresponding harmonic potential V_s. The sum of all simulation histograms is then divided by a self-consistent term, determined iteratively (see also figure 3).[14]

$$\langle P(i)\rangle = \frac{\sum_{s}^{N_{\mathrm{sims}}} h_s(i)}{\sum_{s}^{N_{\mathrm{sims}}} n_s \cdot \exp\left(-\beta(V_s(i) - f_s)\right)} \tag{6}$$

$$f_s = -k_B T \ln\left(\sum_{i}^{N_{\mathrm{bins}}} \langle P(i)\rangle \cdot \exp\left(-\beta V_s(i)\right)\right) \tag{7}$$

The so-called free energy constants f_s are initialised to zero and approach convergence after a certain number of iterations.[15] These constants correspond to the free energy of the biased simulation s, but they do not relate directly to the free energy of binding. To obtain ΔG_{bind}, the unbiased probability $\langle P(i)\rangle$ is used to calculate a free energy surface (FES) $G(i)$.[16]

$$G(i) = -k_B T \ln\left(\frac{\langle P(i)\rangle}{\langle P_{\mathrm{max}}\rangle}\right) \tag{8}$$

Dividing by the maximum of the probability distribution results in $G(i)$ having its minimum exactly at zero. This is common practice since the free energy surface can only be calculated up to a constant, and free energy differences are usually of interest.

[12] https://doi.org/10.1016/0010-4655(95)00053-I
[13] https://doi.org/10.1016/0021-9991(76)90078-4
[14] doi.org/10.1021/ct100494z
[15] https://doi.org/10.1002/jcc.540130812
[16] https://doi.org/10.1021/ct100494z

Setup and Simulations

The computational implementation of the described theoretical background is completely facilitated by the GROMACS software. However, the specifics are still quite complicated. The steps can be categorised into obtaining equilibrated molecular coordinates and topologies, generation of initial configurations along the r_{com} coordinate via a pulling simulation and eventually, the umbrella sampling simulations, which can be subcategorised in equilibration and production simulations. The latter step can take several days. Therefore, it is recommended that these simulations be started as soon as possible. Afterwards follows the analysis which is described in its individual section. Should the analysis reveal missing bins in the histograms or generally deviating free energy results, additional sampling can be conducted in regions with insufficient sampling, so-called post-sampling. The analysis can then be rerun, incorporating the new data, hopefully leading to a better result.
The system of interest is the Aβ(1-42) amyloid fibril consisting of five peptides, and the process that shall be investigated is the dissociation of the outer protofilament. The protein complex has the PDB identifier 2BEG and can be downloaded from the PDB databank or the online repository (recommended). Apart from the original PDB file, the other necessary files entail the TOP, GRO and MDP files. Both the TOP and GRO files can be created by GROMACS' toolkit, but the MDP files need to be downloaded from the online repository. Most US-specific differences are concentrated in the MDP file. However, this time also, the TOP file will require slight alteration, and the GRO file will be elongated to better encompass the separation instead of its usual cube shape.
The GROMACS input files are generated in the usual way, firstly converting the PDB to a GRO file, specifying the box and then solvating it, followed by ion generation. The commands can be found below. The resulting file that can be forwarded for minimisation is called `solv_ions.gro`.

```
gmx pdb2gmx -f 2BEG_model1_capped.pdb -ignh -ter -o complex.gro # Choose GROMOS96 53A6; SPC; all
    N-termini: None; all C-Termini COO-
gmx editconf -f complex.gro -o newbox.gro -center 3.280 2.181 2.4775 -box 6.560 4.362 12 # box
    size complies with minimum image convention and extend of pulling
gmx solvate -cp newbox.gro -cs spc216.gro -o solv.gro -p topol.top # solvate with SPC water
gmx grompp -f ions.mdp -c solv.gro -p topol.top -o ions.tpr # generate dummy TPR for ion
    generation
gmx genion -s ions.tpr -o solv_ions.gro -p topol.top -pname NA -nname CL -neutral -conc 0.1 #
    generate NaCl with c=0.1M
```

The resulting input GRO file can be checked via `vmd solv_ions.gro`. The water should already indicate the correct box dimensions. However, the box can be explicitly visualised by clicking (`Extensions` → `Tk Console`), which should open a new window. Insert the command `pbc box` and press enter. The resulting view should match the one in 4.

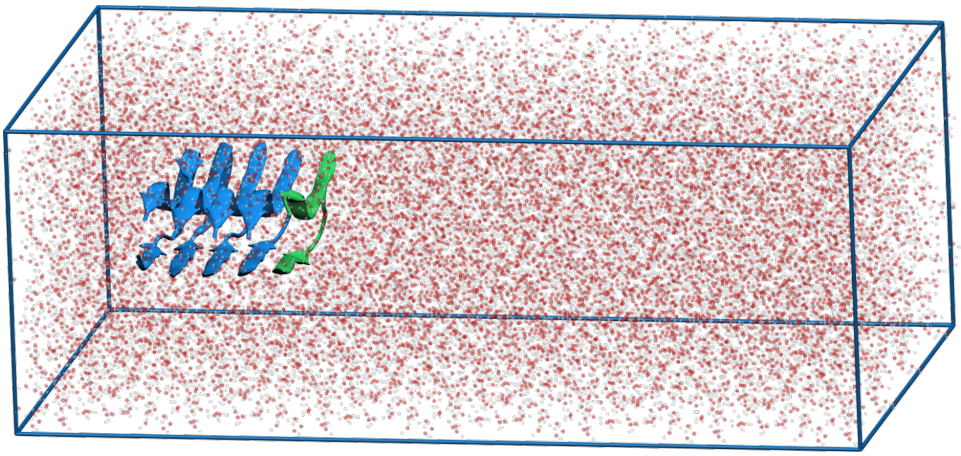

Figure 4: Example image of what the simulation box should look like for this tutorial. The cartoon represents the amyloid, and the colour scheme corresponds to figure 2.

Another issue needs to be addressed after the correct simulation box setup. The second chain of the amyloid fibril, `Chain B`, needs to be fixed in space, which can be realised by so-called positional restraints that bind each atom to its initial XYZ coordinates. Positional restraints differ from distance restraints because they fix the absolute positions, while distance restraints only restrain coordinates relative to each other. The positional restraints are activated by adding the appropriate lines to the `topol_Protein_chain_B.itp` file (`itp` stands for 'include topology'). The lines to be added are shown below. The keyword `POSRES_B` is critical since the `MDP` file uses this keyword to implement the positional restraints algorithmically.

```
#ifdef POSRES_B
#include "posre_Protein_chain_B.itp"
#endif
```

This last manual step concludes the input file generation. Before a pulling simulation is conducted to obtain configurations along the COM distance coordinate, the system must be minimised and equilibrated. The `MDP` files can be downloaded from the online repository and used to execute the commands as usual.

```
gmx grompp -f em.mdp -c solv_ions.gro -p topol.top -o em.tpr
gmx mdrun -v -deffnm em
gmx grompp -f npt.mdp -c em.gro -p topol.top -r em.gro -o npt.tpr
gmx mdrun -deffnm npt
```

The pulling simulation serves the purpose of generating configurations which will be used as starting frames for the individual umbrella sampling simulations. Thus, `GROMACS`' pulling functionality is utilised to steer the outer protofilament away from the remaining amyloid fibril in one simulation from which individual snapshots are saved within an approximately linear spacing along the COM distance coordinate (see figure 5). In order for `GROMACS` to pull a certain atom group, an appropriate index file must be created by the `gmx make_ndx` command that contains the indices of the atom groups to be separated. In this case, the atom groups of the two outermost protofilaments are saved to the `index.ndx` files with the names `chain_A` and `chain_B`. They correspond to the residues one to 27 and 28 to 54. The command `gmx make_ndx` can seem a bit complex for beginners. Therefore, the correct command to generate the index file can be seen below.

```
gmx make_ndx -f npt.gro
...
 > r 1-27
 > name 19 Chain_A
```

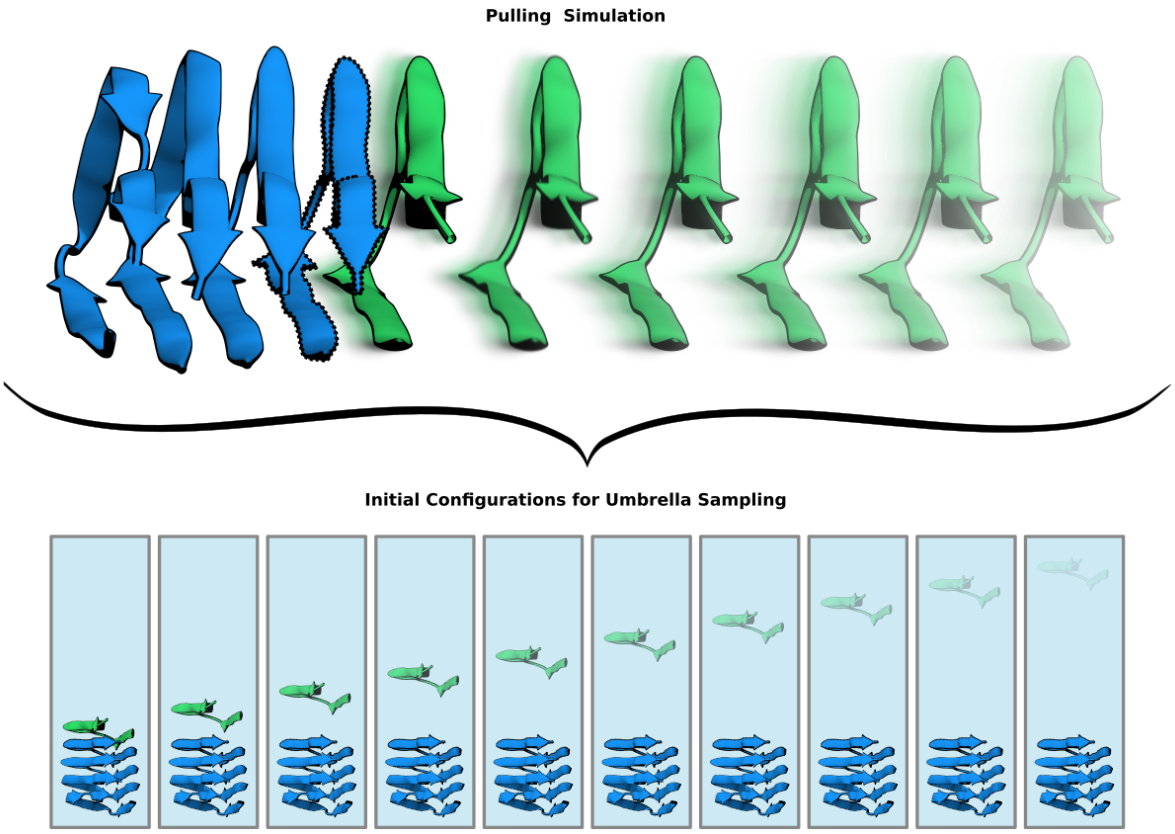

Figure 5: Concept visualisation of the pulling simulation. Along the COM distance coordinate, the outermost protofilament (green) is pulled away from the fibril (blue). The second outermost protofilament is positionally restraint (pronounced outline). Along the reaction pathway, configurations are saved for the umbrella sampling.

```
> r 28-54
> name 20 Chain_B
> q
```

The validity of `index.ndx` can be checked by opening the file via `VIM` and going to the end of the file (by typing shift + G). The non-standard index groups named `Chain_A` and `Chain_B` should be there now. `GROMACS`' pulling functionality is completely parametrised via the `md_pull.mdp` file (downloadable from the online repository). Below is a brief overview of the keywords that can and must be specified for successful pull simulations in table 1.

Table 1: Tabular description of parameters related to pulling simulations in `GROMACS`. On the left is the `MDP`-parameter with its corresponding value, and on the right is a brief description. For further details, refer to the `GROMACS` manual

`pull = yes`	All pull-related keywords are parsed by `gmx grompp`. All pull settings are ignored if set to `no` (default).
`pull-pbc-ref-prev-step-com = yes`	The reference distance for COM-Pulling is calculated from the periodic boundary state of the prior step. This option is required for newer GROMACS versions with `pull-group#-pbcatom`.
`pull_ngroups = 2`	Number of pull groups to be steered within the pull simulation. For distance pulling, two groups are required, while an angle would require three groups.
`pull_ncoords = 1`	Several pull coordinates can be steered simultaneously (useful for domain unfolding); however, only one pull coordinate is used here.
`pull_group#_name = Chain_X`	Numbers (#) and names of the corresponding pull groups as defined in the `index.ndx`. Two names must be provided with `pull_ngroups = 2` specified in this system. Here: `Chain_A` and `Chain_B`.
`pull-group#-pbcatom = 175`	The reference atom of both pull groups for determining the periodic boundary treatment for calculating the COM-distance between the pull groups. The atoms should be located in the centre of their respective group. In this case, Atom 175 should fulfil this criterion for both amyloid chains.
`pull_coord1_type = umbrella`	Apply a harmonic potential to the pull coordinate number 1 to significantly limit the phase space accessible to the two pull groups.
`pull_coord1_geometry = distance`	For pull coordinate number 1, apply the harmonic potential to the distance of the two pull groups.
`pull_coord1_groups = 1 2`	pull coordinate number 1 should apply to pull groups 1 and 2. This option becomes increasingly important, especially if several groups and coordinates are defined.

`pull_coord1_dim = N N Y`	Apply a bias only in the z-dimension. Thus, x and y are set to "no" (`N`) and z is set to "yes" `Y`.
`pull_coord1_start = yes`	the starting value of the pull coordinate number 1 is the value it assumes in the starting configuration.
`pull_coord1_rate = 0.01`	The rate at which pull coordinate number 1 increases in $\mathrm{nm\,ps^{-1}}$ over the course of a simulation. If the pull coordinate refers to an angle/dihedral, this rate would be given in $^\circ\,\mathrm{ps^{-1}}$.
`pull_coord1_k = 1000`	The force constant to parametrise the harmonic potential with, that was set in `pull_coord1_type` in $\mathrm{kJ/mol/nm^2}$.

Additionally, `md_pull.mdp` references the statement previously added to the `topol_Protein_chain_B.itp` file. The `define` keyword at the very beginning uses the same expression to activate the positional restraints of the `Chain B` (second outermost protofilament). Not activating the positional restraints might lead to deformation of the amyloid fibril since the intermolecular interactions between the individual protofilaments are so strong that individual residues would simply be towed away together with `Chain A`. From the `MDP` file, it becomes apparent that throughout the 500 ps trajectory, the protofilament is being pulled 5.0 nm away from the fibril. The keyword `nstxtcout`, which is set to 500 (every 1 ps), determines the frequency that the compressed trajectory `XTC` file is written to. This file will extract the starting configurations for the umbrella sampling. The bash script named `02_simprep.sh` is downloadable from the online repository and automatically takes care of the simulation and frame extraction with a distance spacing of around 0.2 nm. Include a commented version of this bash script as part of the report. Especially the `awk` command is quite complex. Nevertheless, any attempts to correctly comment it are appreciated. Once the run finishes, the success can be tested by plotting the contents of `pullf.xvg` which records the force required over the simulation time. It should look similar to figure 6.

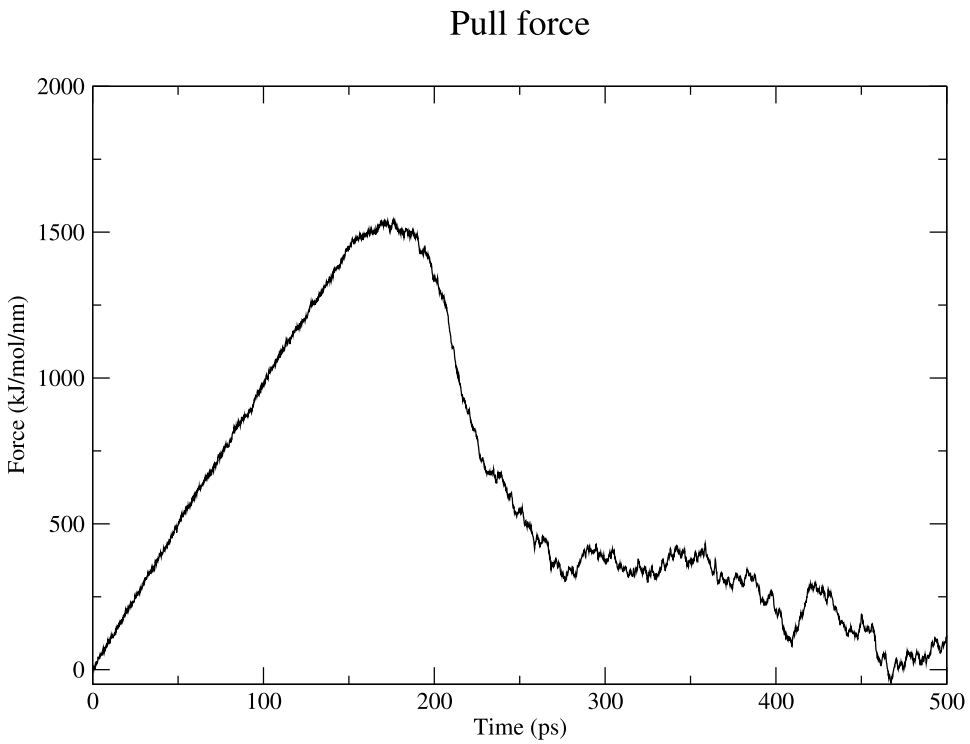

Figure 6: Force on the two pull groups over the course of the simulation.

Apart from the pulling simulation outputs, the script creates two auxiliary files (`dist.xvg` and `important_frames.txt`) and a couple of `GRO` files, e.g. `conf191.gro` , that correspond to the frames listed in `important_frames.txt` . These `GRO` files serve as the starting configurations for the umbrella windows. The 0.2 nm spacing can be verified by comparing the frames listed `important_frames.txt` with their corresponding distances in `dist.xvg` .
The umbrella sampling simulations rely on the same `GROMACS` parameters as the pull run. Comparing `md_pull.mdp` with `md_umbrella.mdp` or `npt_umbrella.mdp` reveals that only the `pull_coord1_rate` differs by being set to zero for the umbrella sampling. Thus, the molecules are not actively pulled apart; the distance is kept constant to its initial value by a harmonic potential. Each starting frame extracted by `02_simprep` requires a quick NPT equilibration followed by a production run over 500 ps. The individual simulations are sometimes referred to as umbrella windows. A code example for executing the runs can be found below.

```
gmx grompp -f npt_umbrella.mdp -c conf<f>.gro -p topol.top -r conf<f>.gro -n index.ndx -o npt<f>.
    tpr
gmx mdrun -v -deffnm npt<f>
gmx grompp -f md_umbrella.mdp -c npt<f>.gro -t npt<f>.cpt -p topol.top -r npt<f>.gro -n index.ndx
     -o umbrella<f>.tpr
gmx mdrun -v -deffnm umbrella<f>
```

Creating a bash script to automate the simulation runs is highly recommended.

Analysis

With the umbrella sampling successfully conducted, the equilibrated data can be analysed. Usually, ΔG_{bind} is calculated from the free energy surface (FES) obtained from MBAR or

WHAM. Herein, the convenient `gmx wham` tool already implemented in GROMACS is showcased.[17] The value of ΔG_{bind} is simply the difference between the highest and lowest values of the FES (see figure 1), as long as the values of the FES converge to a plateau at large COM distances. The input to be fed into `gmx wham` consists of two files to be created by the user. The first one contains line-wise entries of all TPR file names corresponding to the umbrella windows, and the other file includes the names of the `pullf.xvg` or `pullx.xvg` files respectively. These lists could be called, e.g. `tpr-files.dat`, and the contents are listed below for illustration.

```
umbrella45.tpr
umbrella123.tpr
...
umbrella487.tpr
```

Conducting the analysis is relatively simple since it only requires the user to execute the correct GROMACS command.

```
gmx wham -it tpr-files.dat -if pullf-files.dat -o -hist -unit kCal
```

The WHAM module opens each of the `umbrella*.tpr` and `umbrella*_pullf.xvg` files and runs the WHAM analysis on them. The two essential outputs from this program are `histo.xvg` and `profile.xvg` which contain the histogram data and the resulting FES profile against the COM distance coordinate. They can be visualised using `xmgrace`.

```
xmgrace profile.xvg
xmgrace -nxy histo.xvg
```

As exemplary results, a FES is shown in figure 7. The difference from the highest to the lowest point of this FES according to equation (1) yields a ΔG_{bind} of around $-37\,\text{kcal}\,\text{mol}^{-1}$. The difference to the published value of $-50.5\,\text{kcal}\,\text{mol}^{-1}$ is quite significant, however, there are also artefacts visible in the FES around 1 nm, 1.5 nm, 3.5 nm and 4.3 nm that result from insufficient histogram overlap.

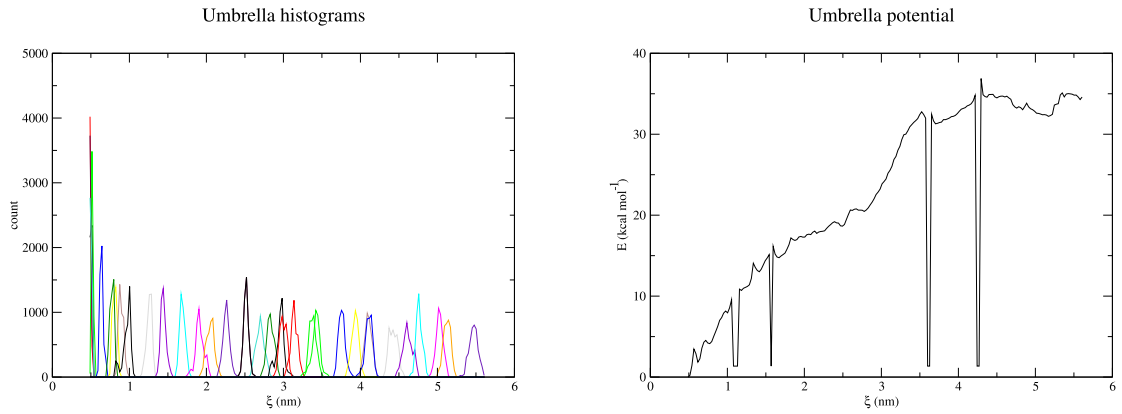

Figure 7: Distance probability distributions (histograms) for the different windows (left) and the corresponding FES along the COM distance coordinate (right).

Since each simulation is independent of all others, these artefacts can be solved by post-sampling the missing distance histograms. Inserting new umbrella simulations with starting configurations centred on the defective region should mitigate the inaccuracy. The file `dist.xvg` contains frames on the first column and distance on the second column so that the

[17] http://membrane.urmc.rochester.edu/?page_id=126

correct snapshot to start a post-sampling simulation can be found. Once the desired frame is known, equilibration and production simulation can be conducted, yielding `TPR` and `XVG` files to `tpr-files.dat` and `pullx-files.dat`. Repeating the WHAM analysis should result in an FES that has fewer defects. An exemplary FES that contains no artefacts anymore after successful post-sampling is shown in figure 8. Here, ΔG_{bind} is obtained with $40\,\text{kcal}\,\text{mol}^{-1}$, which is a bit better than before but still deviates with 20% from the literature value. There are many tools for quality control and optimisation of free energy calculations[18], but for the sake of the tutorial, such a value is more than sufficient.

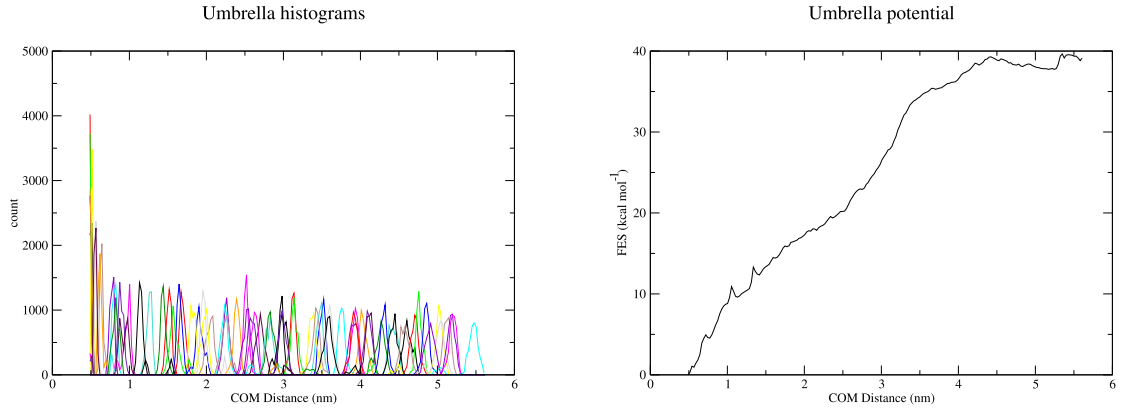

Figure 8: Free-energy profile (FES) along the reaction coordinate ξ (top panel) and corresponding distance probability distributions (histograms) for the different windows (bottom panel).

The report should include the histograms and the FES as plots, as well as the value that is obtained for ΔG_{bind}. A brief discussion of the plots and a comparison with the literature value should be included. Post-sampling is optional but recommended. Eventually, the FES, and thus the free energy estimate, can be further enhanced by collaborating with your fellow students to collect as many `TPR` and `PULL` files as possible. After adjusting `tpr-files.dat` and `pullx-files.dat` (or `pullf-files.dat`) WHAM would analyse a vast amount of data which could close the gap to the literature value.

[18] `https://doi.org/10.1007/s10822-015-9840-9`