Assignment 3

Lianne van Ruijven

Question 1

The essential difference between constraint force integration and umbrella is the determination of the coordination coefficients (in this case r, which is separation distance) for which the PMF is calculated. With constraint force integration, a number of constraint distances is chosen (for example 54). For each of these constraint distances, short constrained MD simulations are ran and used to calculate the Fmean. With umbrella sampling, the Fmean is calculated between a window of separation distances. This goes as follows: The MD simulations are run, then for every time step (for example every picosecond) in the simulation, the accompanying separation distance and energy are sampled.

Both methods assume that the number of particles (N) per volume (V) is constant and that the starting conformation of the system is in equilibrium.

Question 2

After adding water to system: em_sol_posre adjusts water molecules for potential energy minimum, resulting in minimum temperature. Md_sol_posre moves water and adds velocity (thus temperature) and equilibrates the whole system. Md_sol_eq moves protein and water and makes all data from the production run useful. If the production would run immediately, for the first samples the system may be still adapting to the environment (so the water molecules around the protein).

The difference between EM and MD is that the first does only take potential energy into account, while the second takes potential and kinetic energy in to account. EM finds a set of coordinates representing the minimum energy conformation for the given structure. So in EM, only the bonded (bonds, angles, dihedrals and impropers) and non-bonded (coulomb, van der waals) forces can be adjusted. MD takes potential and kinetic energy into account and can make bigger changes to the system.

Question 3

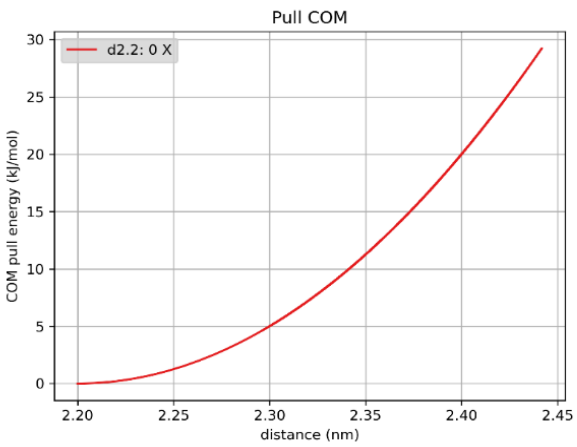

Figure 2. COM pull vs r for starting distance 2.2 nm

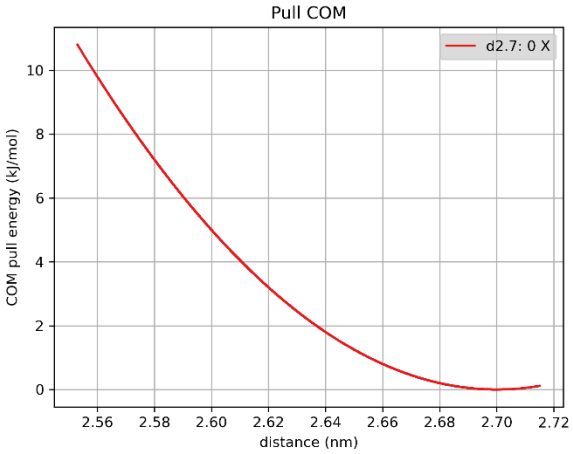

Figure 1. COM pull vs r for starting distance 2.7 nm

The COM pull energy has a relation with the distance between d_0 and d (COM pull energy = k_umbrella(d-d0)^2). Where d_0 is the starting distance of the umbrella sampling and d is the current distance in the simulation. This means, when the distance between d0 and d increases, the COM pull energy also increases.

For an infinitely long simulation, you would expect a parabolic shape. This is because in an infinite simulation, although the probability is low, the simulation will probably sample for all distances in the umbrella window. Currently, there are only samples for bigger (figure 1) or smaller (figure 2) distances than the starting distance.

The difference between the figures 1 and 2 is the slope of the graph. In figure 1, the COM pull energy goes upwards, so there is a high probability that the two proteins have a separation distance higher than 2.2 nm and a low probability of having a separation distance smaller than 2.2 nm. In figure 2, the COM pull energy goes downwards. So finding the two proteins having a separation distance smaller than 2.7 nm has a higher probability than finding a separation distance bigger than 2.7 nm.

Question 4

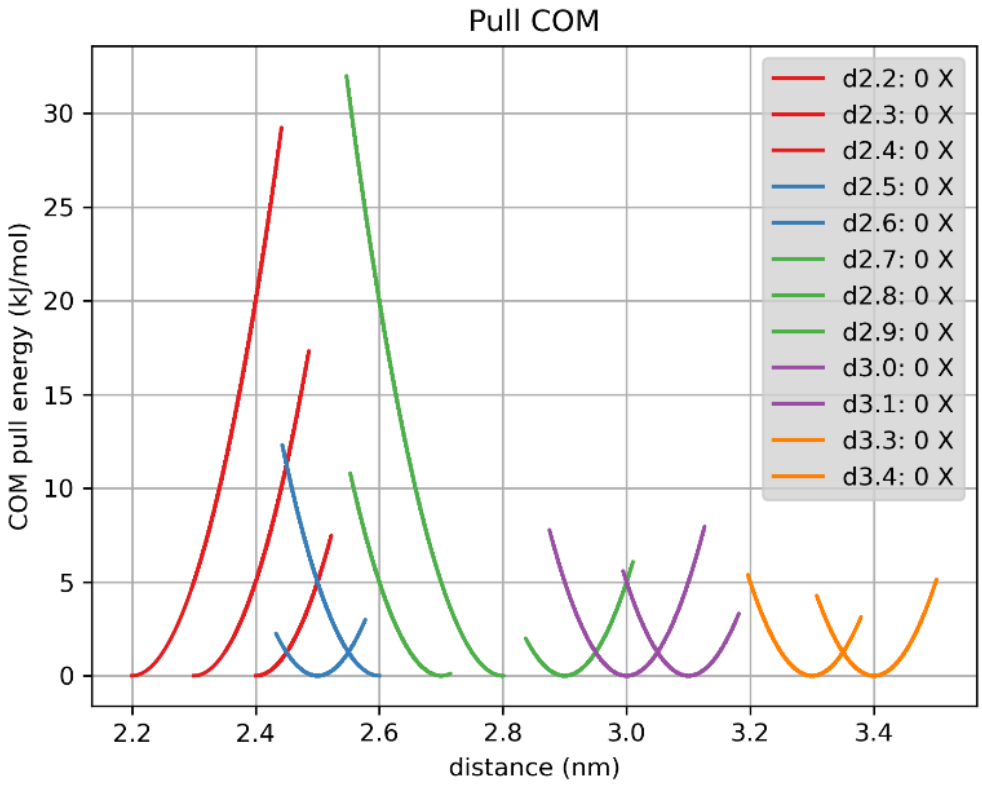

Figure 3. COM pull energy vs distances

The most favorable conformation lies at the free energy minimum. For all graphs in figure 3 with upward and downwards slopes applies: the simulation is drawn to one side of the starting distance. Starting the simulation with a distance between 2.2 and 2.4 nm, the conformation is favorable to be in a bigger separation distance. Starting the simulation with a distance between 2.6 and 2.8 nm, the conformation is more favorable to have a smaller separation distance. So the conformation is most favorable to have a separation distance at 2.5 nm, which means here lies the free energy minimum.

Question 5

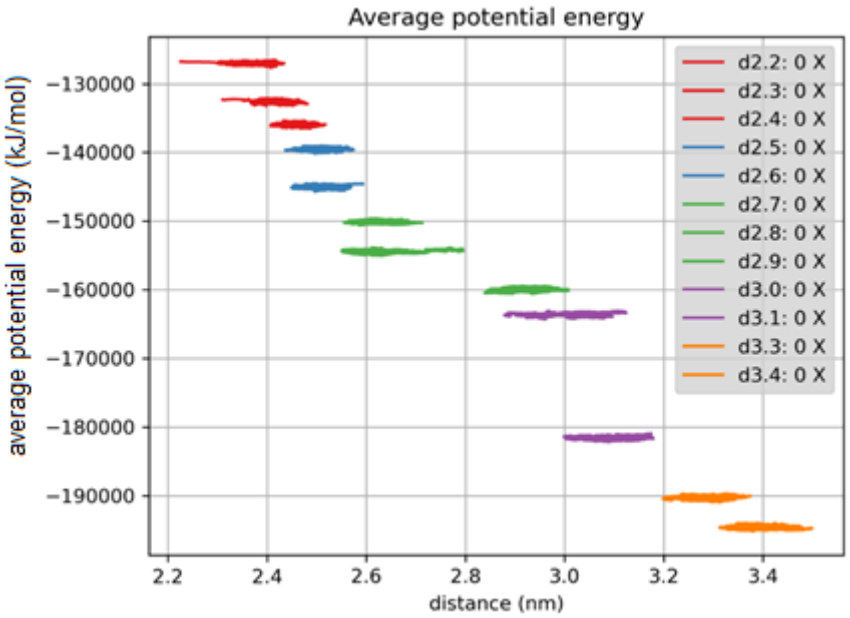

Figure 4. Average potential energy during 12 simulations (d2.2 till d3.4). The average potential energy decreases as the distance increases.

Question 6

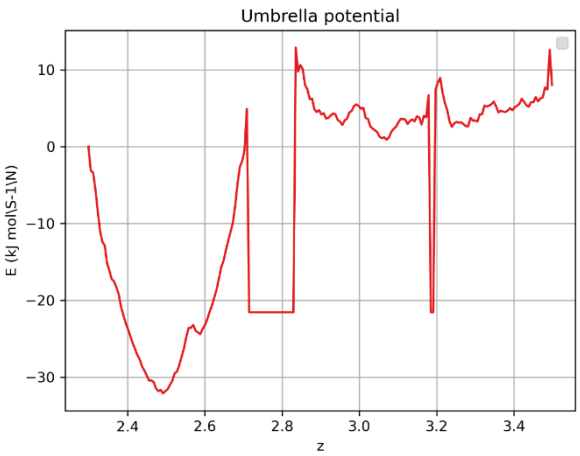

Figure 6. PMF plot. The umbrella potential vs separation distance z (nm)

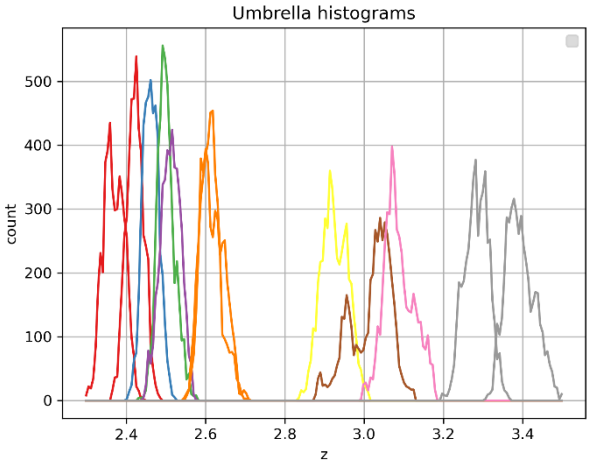

Figure 5. Histogram. Sample count vs separation distance z (nm)

In figure 5, there is a gap between 2.7 and 2.8 nm. So no samples in this range. This results in the PMF plot in the straight drop of the line between 2.7 and 2.8 nm. This is due to the energy barrier at 2.7 nm. The simulation is more favorable to move towards a distance smaller than 2.7, closer to the free energy minimum.

In figure 5, there are smaller and higher parabolas with more overlap between 2.4 and 2.6 than between 2.8 and 3.4. This is due to the energy free minimum at 2.5 nm (see figure 6). So the number of times the simulation ends up between 2.4 and 2.6 is higher. Between 2.8 and 3.4, the simulation

does not have an attraction to a certain distance because there is an energy plateau. This results in the parabolas being more wide, lower and less dense.

Question 7

I choose to combine my output with the output from an other student. So she gave me her simulation results for d2.7, 2.8, 3.1 and 3.3. This improved my PMF plot. It decreased the gap around 2.8 and removed the gap around 3.2.

When comparing my improved results with two other students that used the first solution (so thy rerun the production simulations with different starting points), they don't have any gaps anymore. This can be because they had the opportunity to change the starting positions and thus be more variable in the starting positions.

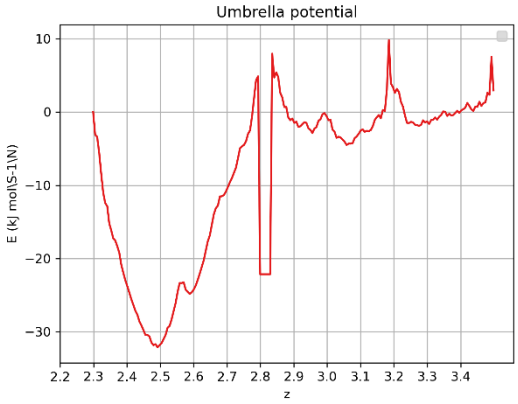

Figure 10. Improved PMF plot by combining original output with the output of an other student. Umbrella potential vs distance z (nm).

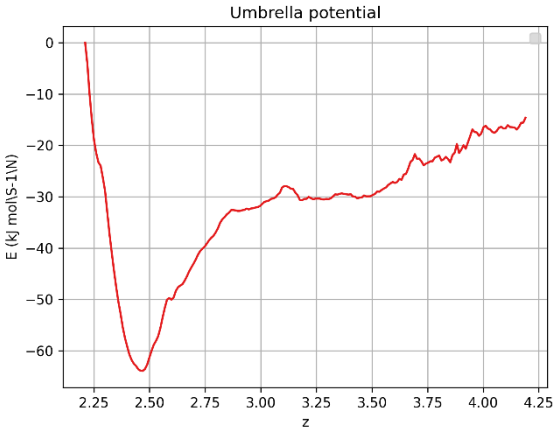

Figure 9. Improved PMF plot of other student. Plot is improved by rerunning certain simulations.

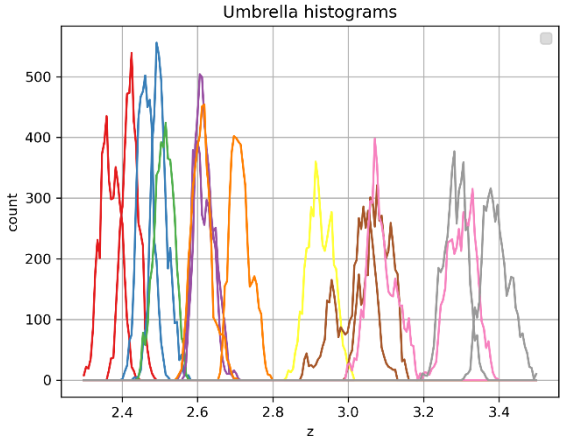

Figure 8. Improved histogram by combining original output with the output of an other student. Sample count vs distance z (nm)

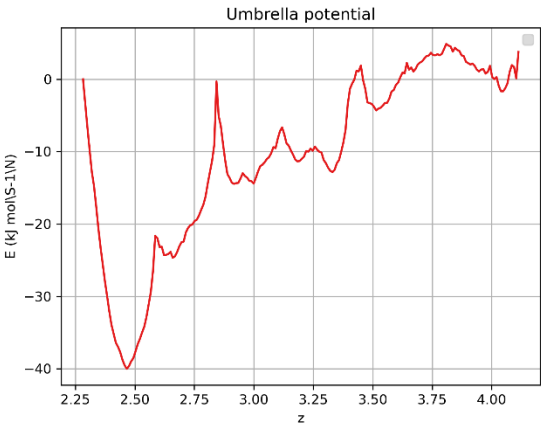

Figure 7 Improved PMF plot of other student. Plot is improved by rerunning certain simulations.

Question 8

In case of ideal sampling, I do not expect any differences between my PMF graph and the graph in the paper of May et al. But there is no case of ideal sampling for my umbrella sampling method. I have 12 umbrella windows and within each window 5000 samples. These samples are not equally distributed along the x axis. May has (around 20) samples for 54 equally distributed distances along the x axis.

During the umbrella sampling, there are distances with a low sample count (see fig. 8). In my PMF graph, the PMF value spikes up or down at these areas with low sample count (see fig 10, around distances 2.55, 2.8 and 3.2). So in my PMF graph, the average PMF for each distance is not calculated over the same amount of simulations and therefore the areas with a low sample count have a biased average PMF. While in the paper, because of the equally distributed sample distances, the average calculated at each distances is done with the same amount of samples. This is the reason why my PMF graph looks rather bumpy and spiky around the plateau and the PMF graph from the paper looks pretty smooth.

Question 9

Downward slope (2.2 till 2.5 nm): Proteins are bounded and there is a steric clash.

Global minimum (2.5 nm): Proteins have direct physical contact AND water mediated interactions. Highest probability to find the proteins in this state.

Upward slope (2.5 till 2.8 nm): When increasing the distance, there is more water molecules between the proteins. So less physical contact and more water mediated interactions.

Plateau (2.8 till 2.4 nm): Proteins are totally unbound. No energy is pulling the two proteins together. If the proteins meet, they will be likely to bind due to the hydrophobic surfaces of the proteins.

Question 10

When the two proteins are bound, their beta sheets are physically close to each other. The backbone hydrogen bonds in the beta sheets are important to bind the two proteins as one complex.