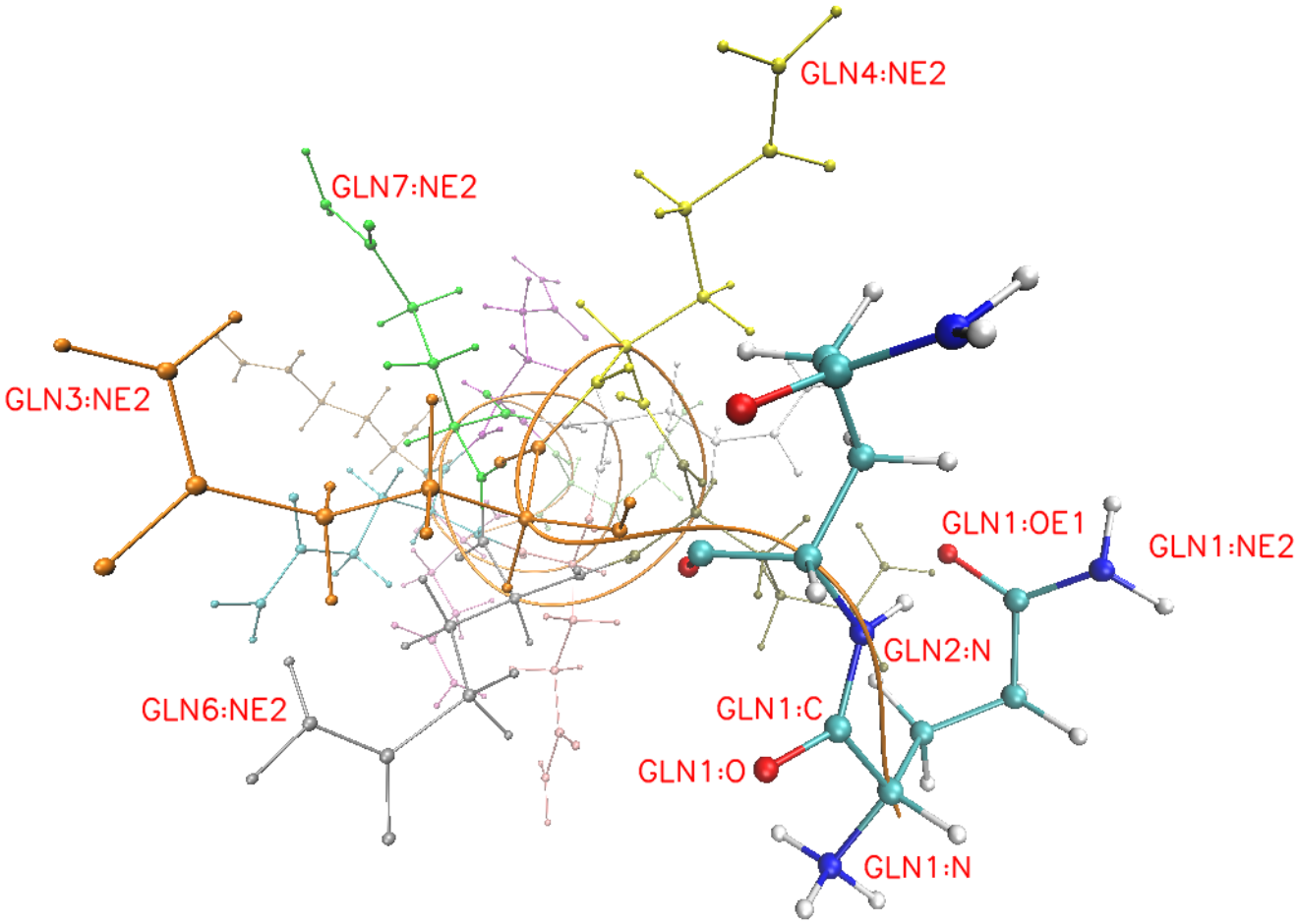
UNIVERSITEIT VAN AMSTERDAM

Analysis of the PolyGlutamine Q14 protein



Author:
Bas Châtel, 10246215
Date:
April 6, 2018

Analysis of the PolyGlutamine Q14 protein

Bas Châtel, 10246215
University of Amsterdam | Biomolecular Simulations

CONTENTS

Analysis of the PolyGlutamine Q14 protein

Bas Châtel, 10246215
University of Amsterdam | Biomolecular Simulations

Abstract— **Polyglutamine (polyQ) proteins have been linked to multiple neurodegenerative diseases. Severity and age onset of symptoms in these diseases are correlated with the length of these polyQ chains. In this paper we investigated the dynamics and conformations of the polyQ-14 protein using a mixture of VMD, GROMACS and Plumed analysis. These analysis yielded a total of four different possible conformations of the protein. Also a small analysis was performed to see how the system performs under high temperatures and a committor analysis was performed shedding light on the transition state between two different conformations.**

I. INTRODUCTION

A. Background of the system

Polyglutamine (polyQ) proteins have been linked to no less than nine neurodegenerative diseases ([1], [2], [3]). These diseases, like Huntington's disease and spinocerebellar ataxia, result from an abnormally increased number of residues in the corresponding gene product [4]. For example, the length of the polyQ region in the huntingtin protein is inversely proportional to both the severity as well as the age of onset of Huntington's disease symptoms [2]. These symptoms tend to occur when the number of glutamines exceeds a threshold of about 36 to 40 repeats.

As the larger repeats cause symptoms it is imperative to first create a stable knowledge base about the normal, smaller polyQ repeats that do not inflict symptoms, as molecular mechanisms at an atomic level are not yet fully understood [5]. This way we can later elaborate on the larger proteins and gain understanding about differences between the polyglutamine chains that exceed the threshold and those who do not. That is why this paper will discuss the polyQ-14 protein which consists of 14 repeated glutamine aminoacids.

B. System characteristics

Figure 1 shows the polyQ-14 protein in different colorations. The first two residues are depicted using a CPK representation method, showing the different atoms as spheres and the van der Waals forces as cylinders. The atoms in these first two residues are colored based on the kind of atom. Only the oxygen and nitrogen atoms in the first glutamine amino acid are labeled, as well as the connecting nitrogen atom from the second glutamine residue. The remaining glutamine amino acids are also depicted in CPK but colored based on the residue ID. Finally a newRibbon representation depicts the backbone, accentuating the helical property of the whole protein. The figure clearly visualizes that the polyQ-14 protein is of a helical nature with its nitrogen atoms at the outer rim, leaving a possibility to form hydrogen with its aqueous surroundings.

C. Research questions

In this paper, various questions regarding the polyglutamine protein are addressed. These questions are asked to accomplish an extensive overview of both the structural as the behavioral traits of the protein. These questions are as follows:

1) How many different conformations can there be found in the polyQ14 protein?
   - How do these conformations compare to one another?
2) How likely are these conformations to occur?
3) How does the protein react under extreme conditions (400 Kelvin)?
4) Is it possible to obtain a free energy profile from the reaction coordinates?
   - How does this look?

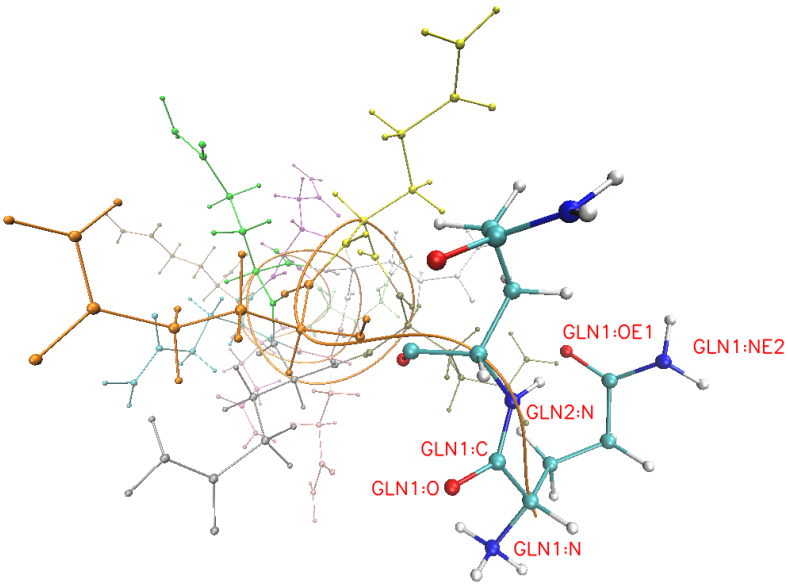


Fig. 1: PolyQ-14 showing residue 1 and 2 in CPK mode, revealing the bond between two glutamine amino acids. Other repeats are individually colored by residuID. Finally the backbone of the helix is shown using a newRibbon representation, showing the protein's helical property.

TABLE I: A list of settings used for the simulations.

Setting	**Values**
Length of the Simulations	100 ns
Temperature	Equilibration at 298K & simulations at 310K or 400K
Pressure	-5.64850e+01 bar
Potential	-8.52244e+04 mV
Integration Algorithm	Integrated MD
Time Steps	50.000.000 (1 step = 2 femtoseconds)
Force Field	AMBER99SBILDN
Water Model	TIP3P
Settings for non-bonded interactions and neighbor-lists	Neighbor-list Cut-off at 1.1 nm Cut-off updated every 10 time-steps Particle Mesh Ewald (PME) for electrostatics
Preparation Details	Protein placed in periodic box, 1nm between protein & boundary
Collective variables	RMSD, helical bonds, Distance (N-terminus $\Rightarrow$ C-terminus)
Hill height (kJ/mol) and width (in units of the collective variable)	Hill width: RMSD 0.5 nm Hill height: 0.01 kJ/mol
Deposition Rate	Time interval between hills: 500

D. Approach and expectations

To figure out the amount of conformations, a mixture of visual VMD analysis, RMSD (Root Mean Square Deviation), temperature, free energy and potential calculations will be done using GROMACS and PLUMED ([6], [7], [8], [9]).

Since the polyQ14 protein is relatively short, it is expected to exhibit a lower amount of conformations than larger structures. In my expectation we will find around 3 different conformations, with one in particular that will be its resting state (i.e. the free energy profile will be at its lowest here), the second would be when the whole protein is unraveled and the third would be somewhere in between.

II. METHODS

A. Brief explanation of chosen approach

In this experiment 5 different MD simulations for polyQ monomers with 14 repeats were done for 100ns at body temperature (311K or 37.85°C) using the HPC/Carbon Cluster supercomputer. Also one 100ns simulation was conducted at 400 Kelvin (126.85°C). These proteins were placed in a periodic dodecahedron box with at least 1nm between the protein and the boundaries. The Amber force field, AMBER ff99SB, was used with a TIP3P water box to provide an explicit simulation of the solvent and for electrostatics the Particle Mesh Ewald (PME) was used.

The simulations were then analyzed using VMD for visual inspection and trajectory analysis. RMSD with respect to starting structures, number of helical hydrogen bonds and distance between the C-terminus and N-terminus as a function of time for each trajectory were calculated using GROMACS.

Furthermore 10x 20ns committor analysis were conducted on one of the 5 simulations at body temperature to single out a transition state. Each simulation started 10ns before a transition and with random velocities. After this an RMSD analysis using the stable state descriptions as described earlier was performed to determine in which state they ended. A conformation was considered stable if the system remained at least 10ns in the same state.

Lastly, using the description of stable states and potential reaction coordinates for the transition between states, a 50ns metadynamics simulation was performed to examine the free energy profile of the system with respect to the helix RMSD. With the RMSD collective variable as part of the reaction coordinate we can see if this is sufficient to obtain a free energy profile.

B. Computational details

A list of all the settings as used in the simulations can be seen in table I.

III. RESULTS

In this section results are given, and partly analyzed, in four different sections.

- A visual analysis will be given to gain quick understanding about the system.
- An MD analysis will be done, creating some more insight about dynamics.
- A committor analysis is conducted to determine a possible transition state.
- A meta dynamics analysis is finally done to test the plausibility of the test conducted earlier.

A. Visual analysis

Before starting an analysis of any kind, it is useful to first conduct a visual analysis of the protein. This is to see that upon creating the protein, it's box and water molecules are indeed in place. Also a rudimentary view of its different conformations can be ascertained by looking at the movements of the protein itself by running the different runs in VMD. At this stage, no statistical data is created but only a quick overview of possibilities can be given.

Upon this visual inspection, first the whole protein with its surrounding water molecules can be seen (see figure 2). By looking at this, we can see that the loading of the protein itself and filling up the box with water has been successful.

Upon further inspection, four different conformations can be distinguished (see figures 3A, 3B, 3C and 3D). Figure 3A, is the starting conformation and is helical by nature. Then as time passes, the bonds forming the helices start to dissolve

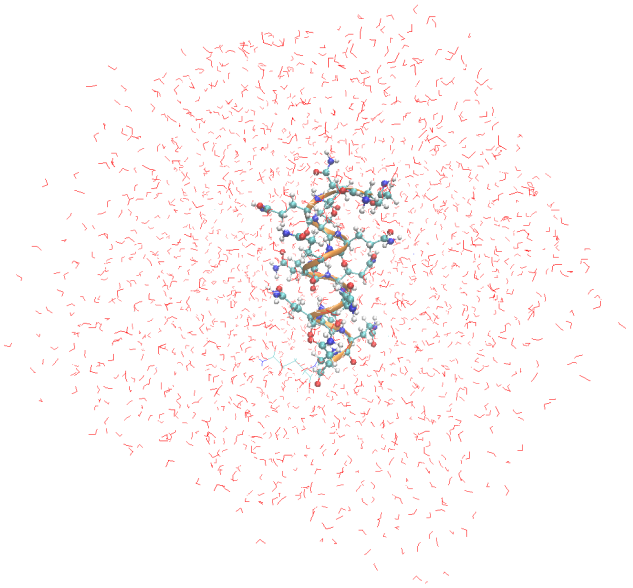


Fig. 2: All water molecules plus polyQ-14 protein, showing that the loading protein and filling of the box with water has been successful.

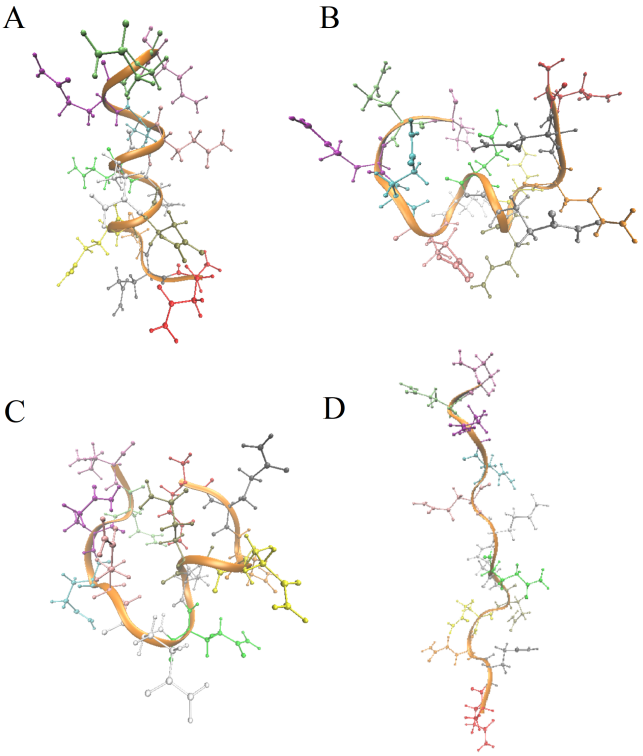


Fig. 3: A). PolyQ-14 starting structure. B). The protein is starting to unravel, losing its helical property at both ends. C). Total unraveling, helical property has dissipated and only circular shape remains. D). Total loss of binding with itself, protein is straight. To highlight its helical property, the backbone was colored orange. A CPK representation was used, coloring each repeat differently to highlight the structure by showing its atoms (spheres), and the van der Waals forces (cylinders).

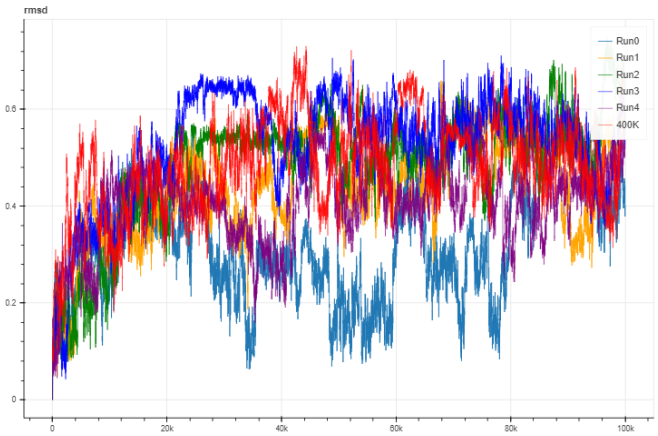


Fig. 4: A plot of the RMSD of all five runs. Since these runs do not really give us a clear picture of the different conformations, a frequency analysis was done on all variables to extract the most visited values.

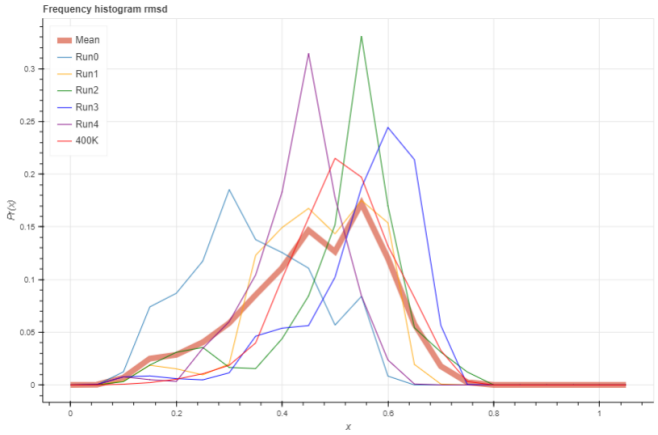


Fig. 5: A relative frequency plot created from RMSD values for all five standard runs, with a mean plotted in deep orange. Here we can see that there are three distinct RMSD values that the runs visit most often, with a hint of a fourth (purple line), indicating three or four conformations that are most visited.

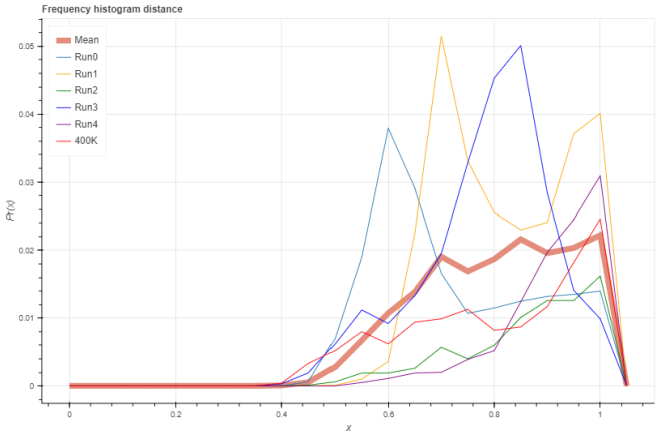


Fig. 6: A relative frequency plot created from distance values for all five standard runs, with a mean plotted in deep orange. As can be seen in figure 5 the different runs reach a total of four peaks. Indicating the existence of four preferred conformations.

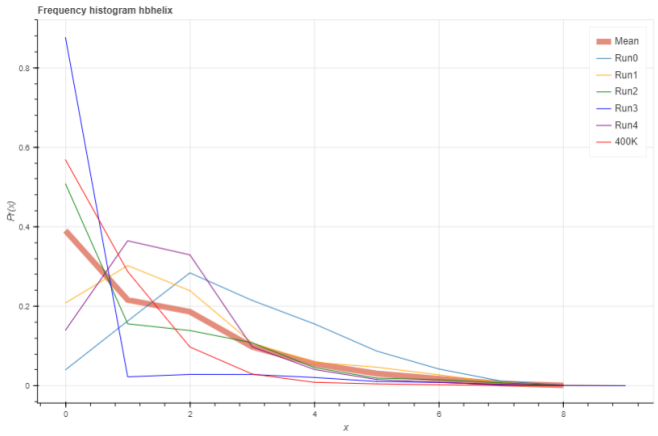


Fig. 7: A frequency plot created from hbhelix values for all five standard runs, with a mean plotted in deep orange. The number of hydrogen bonds are depicted versus the relative frequency is plotted showing that less hydrogen bonds are more frequent than the larger amounts, indicating that the helical property is mostly diminished during the runs.

forming the second conformation (fig 3B). Here the two ends of the protein have straightened up, with two coils in the middle. After this, the protein continues straightening up and the third conformation emerges (fig 3C). All helical properties have been disbanded and only a straight protein remains. Finally, the protein curls up again creating this circular form (fig 3D). Even though no analysis was conducted in this stage, a crude understanding of the proteins mechanics is obtained. These findings roughly fall in line with the expectations.

B. MD simulations

MD simulations yielded information about three different reaction coordinates for the transitions between the stable states.

1) RMSD: In figure 4 we see the course of the RMSD over time. Here we can determine five different RMSD values in which the system remains for at least 10ns at the same value. These values are 0.2, 0.3, 0.4, 0.54 and 0.6. However,

because of the large amount of noise in the system it is hard to determine values that the system visits the most, even at a singular run level. For this purpose it was chosen to analyze the system in a frequency plot (fig 5), counting the amount of times the systems visits a value and normalizing that with respect to the total amount of data points. This yields a clearer overview of the amount of conformations. As can be seen in figure 5, three clear peaks can be seen. With a fourth close to another. This is an indication that the system has four preferred states which it often visits and in which in would remain for a certain time period.

2) Distance: Upon looking at the relative frequency plot (fig 6) we can see the same trend in that can be seen in figure 5. Four different peaks can be distinguished in the graph, and when looking at the mean three small peaks can still be distinguished, indicating three to four preferred conformational states that the protein is drawn to.

3) Hbhelix: As a third option the number of hydrogen bonds within the helix was monitored. Again a frequency analysis was done on this data to explore whether the system has any particular values it visits more often. In figure 7 we see the frequency of 0 to 10 hydrogen bonds for each run. We can see that the highest frequency of amount of bonds can be found in the lower range. This indicates that the protein is most often in a state in which the helical hydrogen bonds are absent, meaning an unfolding of the protein is happening.

4) 400K: Finally the analysis was also conducted for a 400 Kelvin temperature (figures 4, 5, 6 and 7, the red lines). In this simulation, when compared to the 310 Kelvin runs the 400K simulation does not really exhibit a different behavioral pattern. The foremost difference is that because of the higher temperature, the system seems to have less stable moments causing it to never be in the same state for more than 10ns.

C. Committor analysis

A committor analysis of candidate transition state in run 2 was performed. This run and point was chosen for its characteristic change of RMSD values from one steady state to another (see figure

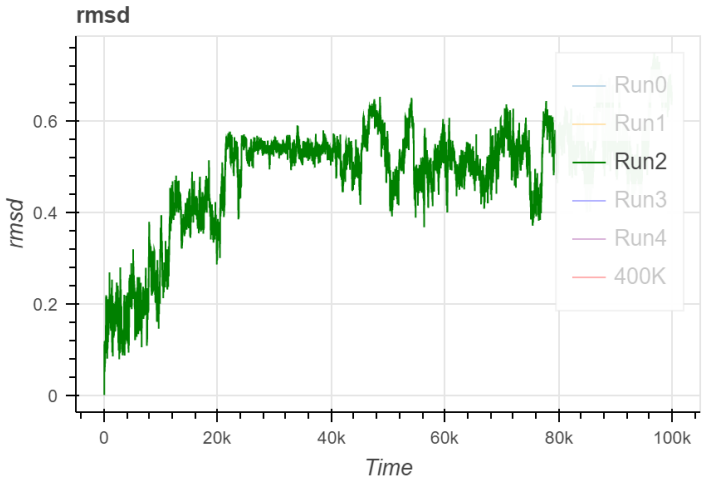


Fig. 8: Time lapse of run two with respect to the RMSD. At t=13K a clear leveling off of the RMSD occurs for about 10ns after which a switch is made at t=23K where another leveling off of the RMSD occurs until t=48K. This indicates a transition state of two steady states following each other.

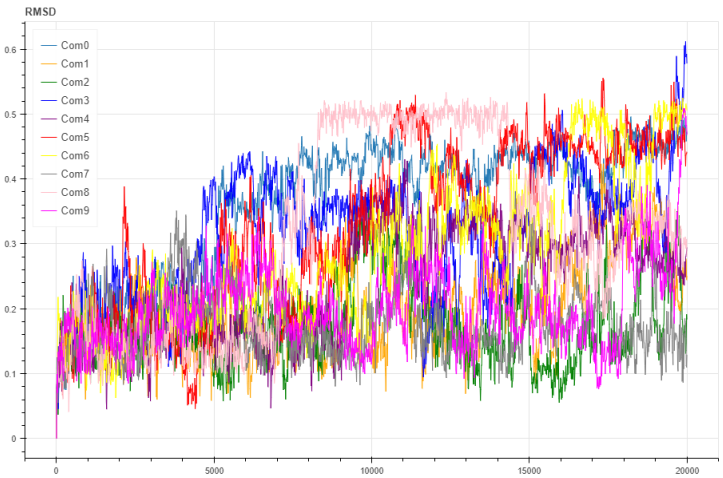


Fig. 9: Graph showing RMSD for all committor simulations. Even though there is a lot of noise, band forming can already be observed

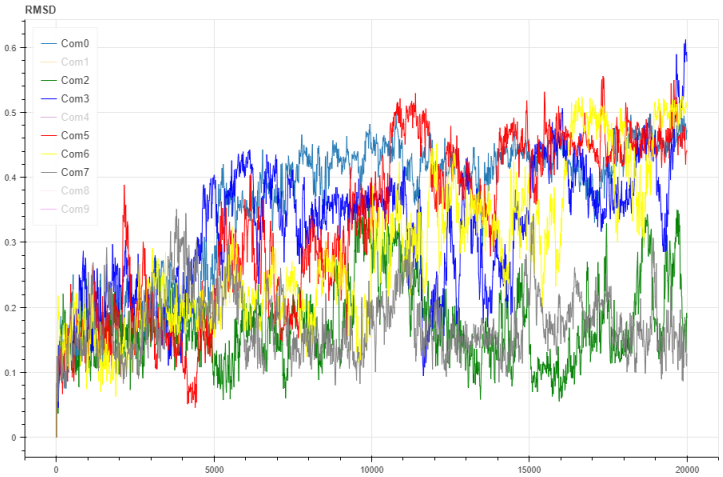


Fig. 10: A selection in the committor simulations reveals a bifurcation within the RMSD, indicating that there are two different conformations present.

8). At t=13K a clear leveling off of the RMSD occurs for about 10ns after which a switch is made at t=23K where another leveling off of the RMSD occurs until t=48K. When examining the data in figure 9, we can see the RMSD for all committor simulations. Even though there is a lot of noise, band forming can already be observed. But it is after looking more selectively at the different runs that we can clearly see a bifurcation in the RMSD trajectory at 0.15ns and 0.45ns. This is shown in figure 10 where two different RMSD values clearly take up the bulk of the simulation, indicating that there is, indeed, a transition state at the chosen point. Notable is that upon closer examination of the different runs, a third band can be seen (figure 11). Committor run 4 and 8 show a flattening of RMSD values around 0.3nm over time at the end of their simulations. This is an indication that there might be a third conformational state within this transition state.

D. Metadynamics simulation

A metadynamics simulation using the RMSD reaction coordinate was performed. In figure 12 we can see the free energy in the system with respect to the RMSD. Since the RMSD is the root mean standard deviation of the molecule compared to the beginning state, a RMSD value of 0 means that the system is in its starting position. We can also see that in this starting position (RMSD=0) the free energy is at its lowest. This means that the protein has a greatest tendency towards the helical structure shown in figure 3. If more steady conformations exist in the system, we should be able to observe another energy minimum within the graph. As this is not the case, we can state that the system only has one true steady state, and judging from the previously mentioned results and this metadynamics simulation there is a plurality of preferred states but no true steady state. This can also be seen in the flattening of the free energy at $\pm$ -20 kJ.

IV. DISCUSSION

As the results are analysed within the result section itself, this section will be a short summary of the results.

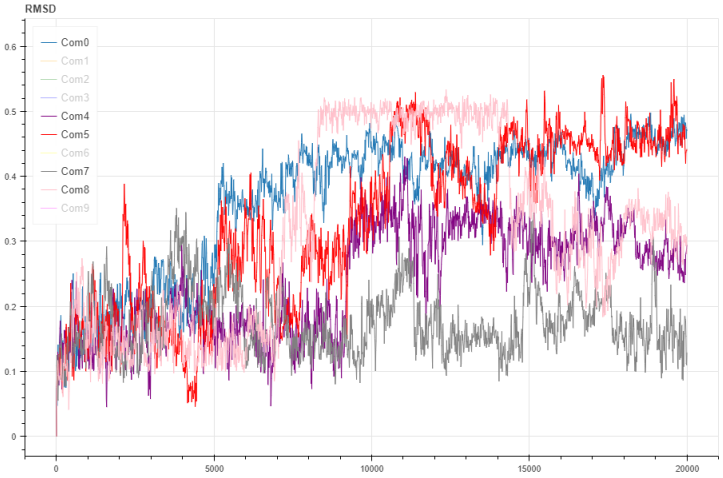


Fig. 11: Upon further analysis, a third branching in the RMSD arises indicating that this transition point might be the start of a three way conformational branching point. Pay attention to the pink run, which remains within the same range for about 7ns.

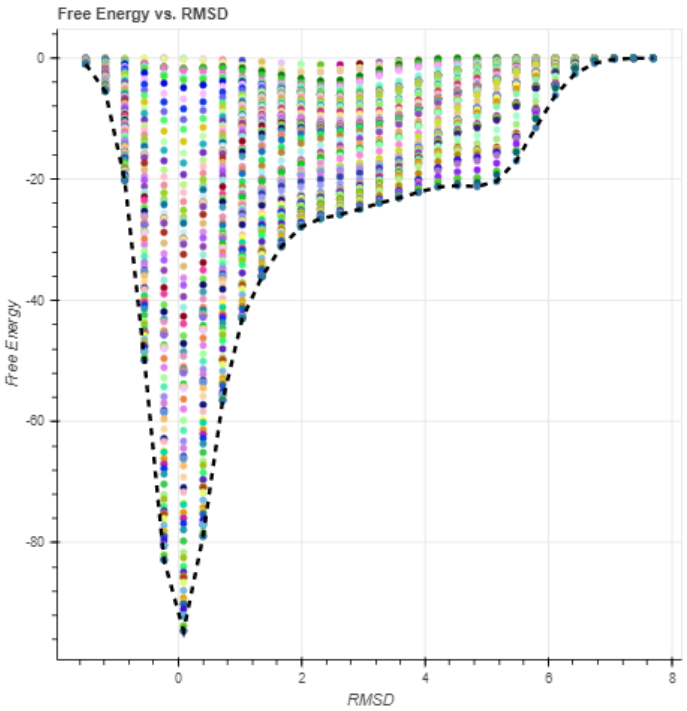


Fig. 12: The free energy with respect to the RMSD was calculated, showing that at RMSD 0 (starting conformation) the free energy is at its lowest. Showing that the system is at its optimal conformation.

At first the visual analysis yielded a foursome of possible conformational states (helical, semi-helical, straight and circular). These findings formed the premise of an increasing evidence base supporting the expectations.

In the MD simulation analysis frequency his-

tograms and timelapse analysis of RMSD, distance and amount of helical hydrogen bonds also supported these findings. At four different values in these variables peaks arose in the frequency histograms both in the RMSD as the distance plots. The amount of helical hydrogen bonds diminished over time, supporting the notion that the protein was unwinding as time passed.

The committor analysis found a transition state into a threefold of different states. 20% of the runs stayed at a similar RMSD value while 40% of the runs visited RMSD values of around .25ns and 40% visited RMSD values around .5ns. This means that indeed a transitional state has been found.

The metadynamics simulation yielded a free energy profile, indicating that the alpha-helical starting structure has the most favorable conformation as its free energy is the lowest. Upon deviating from this starting structure, the free energy increased but leveled off between RMSD values 1.5 and 4.5ns, meaning that other conformational states are energetically feasible as well, but less favorable.

V. CONCLUSION

The polyQ14 protein visited 4 states. State 1 is an alpha-helical structure, state 2 is an alpha helix with the four terminal residues on both sides frayed, state 3 is an entirely unfolded and straight chain and the fourth state is a circular state in which the chain is entirely unfolded but the two termini are drawn to one another. It is, however, important to state that these conformational states are visited for period longer than 10ns but do not really seem to be actual steady states. As the free energy analysis does not support the existence of another energetic minimum in the system, we can state that these four conformations are actually one steady state and three conformational states that have a preference in the system.

After visual and MD analysis a committor analysis was conducted to find a transition state. These simulations revealed a transition into two different conformational states, and possibly a third as different RMSD values are linked to different conformations. Committor analysis on these conformations indicated that 40% of the runs end up in an RMSD value of .5ns, 40% ends up with an RMSD value of .25ns and the remaining 20% end up with and RMSD value of .15ns. The range of .5ns RMSD however is over the course of time more prevalent than the .25ns RMSD range, indicating that this conformation is more dominant. These findings show that it is indeed a transition state.

A metadynamics simulation using RMSD was also conducted to obtain a free energy profile, showing a dominantly low energy at the alpha-helical starting structure. With higher RMSD values, the free energy quickly rises to a less negative state, but levels of between RMSD values 1.5 to 4.5ns. This supports the notion that there is one optimal conformational state, and other conformational states do exist, but are not as stable.

All in all, we can state that four different conformational state have been found, of which one (the alpha-helical starting structure) is the most dominant.

A. *Outlook*

Now that we have gain understanding in the mechanics and dynamics of the polyQ14 protein, some improvements in the analysis can still be made. The metadynamics analysis now used only the RMSD as a collective variable, but when paired with distance between the termini and the hydrogen bonds the analysis would have been more robust with a higher resolution.

Furthermore because of its small state, the polyQ14 protein does not have a lot of room to find energetic local minima because of its size. Larger structures would have more room to cover a larger state space, increasing the chance to find another local minimum. It would be interesting to expand our study to larger structures for the sake of this dynamical and mechanical point of view, but namely because literature discusses a greater pathological tendency amongst the larger structures as is discussed in the introduction. This study was done in parallel with others that analyzed polyQ proteins of different sizes (polyQ-12 till polyQ-50). So if all these results were pooled, a very complete overview could emerge about the emergence of pathological symptoms. Are there specific structural conformations present in polyQ proteins larger than 36 repeats that are not present

in the non-pathologically smaller proteins? What is the relation between the size of the protein and the amount of conformational states? These are questions that would be very interesting to address in the future.

Lastly an open path sampling simulation (OPS) on the proteins can also be interesting to perform. OPS allows you to set up many different kinds of move schemes using trajectories, enlarging the possibility to find another conformational state. It can find transition states from one stable state to another that occur too rarely to be observed on a computer timescale. To establish this it performs transition path sampling and transition interface sampling as well as committor analysis and flux calculations. So to have a more in-depth understanding of the dynamics this would be an interesting approach.

REFERENCES

[1] Gómez-Sicilia, À., Sikora, M., Cieplak, M., Carrión-Vázquez, M. (2015). An Exploration of the Universe of Polyglutamine Structures. *PLoS Comput. Biol.*.

[2] Lakhani, V. Vinal., Ding, F., Dokholyan, V. N. (2010). Polyglutamine Induced Misfolding of Huntingtin Exon1 is Modulated by the Flanking Sequences. *PLoS Comput. Biol.*.

[3] Suguya, K., Matsubara, S., Kagamihara, Y., Kawata, A., Hayashi, H. (2007). Polyglutamine Expansion Mutation Yields a Pathological Epitope Linked to Nucleation of Protein Aggregate: Determinant of Huntington's Disease Onset. *PLoS ONE*.

[4] Gatchel JR, Zoghbi HY (2005). Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet 6: 743–755.*

[5] Wen, J., Scoles, DR., Facelli, JC (2017). Molecular dynamics analysis of the aggregation propensity of polyglutamine segments. *PLoS ONE.*

[6] B. Hess and C. Kutzner and D. van der Spoel and E. Lindahl (2008). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation *J. Chem. Theory Comput. 4 pp. 435-447*

[7] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen (2005). GROMACS: Fast, Flexible and Free *J. Comp. Chem. 26 pp. 1701-1719*

[8] E. Lindahl and B. Hess and D. van der Spoel (2001). GROMACS 3.0: A package for molecular simulation and trajectory analysis *J. Mol. Mod. 7 pp. 306-317*

[9] H. J. C. Berendsen, D. van der Spoel and R. van Drunen (1995). GROMACS: A message-passing parallel molecular dynamics implementation *Comp. Phys. Comm. 91 pp. 43-56*