Project plan

Project title:

Implementing the string method with swarms of trajectories using the newly developed Gromacs API.

March 2020

Student: Marko Petrović

Supervisor: Pr. Lucie Delemotte

Co-Supervisor: PhD. candidate Oliver Fleetwood

Version: 2

Contents

1 Introduction

Welcome to the plan specifying the work on my masters thesis: *Implementing the string method with swarms of trajectories using the newly developed Gromacs API.* Where an application will be constructed in the Python3 programming language to implement the string method and provide an easier to use user interface for researchers.

2 Objectives

The main objective is to construct an easy to use interface for the Gromacs molecular dynamics simulation software, focused on the string method.

2.1 Setup

Install Gromacs and it's API locally on my personal computer. Get simulations to run locally and on calculation cluster using Gromacs API

2.2 Create implementation

Construct an implementation capable of iterating through the running of multiple simulations and doing subsequent refinement calculations for the following iteration.

2.3 Create user interface

2.3.1 Results presentation

My implementation should be able to create visualisations of collective variables for use in articles and reports.

2.3.2 User experience

This might not actually be an objective but rather a goal. The implementation should have a graphical user interface for setting up and initiating simulation runs.

3 Background

Simulating molecular and atomic systems at atomic resolution is a practice known as molecular dynamics (MD). A MD simulation is run one time step at a time. In each time step, the forces acting on each atom is calculated and used for calculating the atoms acceleration and subsequent movement for that time step. As the relative speed of the atoms and molecules is quite large compared to both the volume of the simulated system and distance between atoms and molecules, each time step is limited to 1-2fs. MD can be used for very large systems, beyond 50000 atoms. The combination of very large systems and very short time steps results in huge amounts of calculations for simulations spanning a only few μs setting very narrow limits of which biological processes can be simulated.

Trying to use conventional MD simulations to analyse conformational changes in proteins and other macro-molecules would not be feasible, *"Conformational changes in large biomolecules are complex and slow processes taking place on timescales that are beyond the reach of brute force molecular dynamics simulations."*[1] The string method with swarms of trajectories can be used to alleviate the calculational load by setting a best guess path from one state to the next. Using snapshots equidistantly placed along that path as virtual initial states (VIS) the string method runs several short MD simulations for each VIS, a swarm of trajectories. The assumption made is that the general drift in each swarm would be towards a conformation of lower energy. Each VIS is updated[2] according to a mean of its swarm of trajectories and then also reparametrised to keep the VISs equidistant along the path.[1]

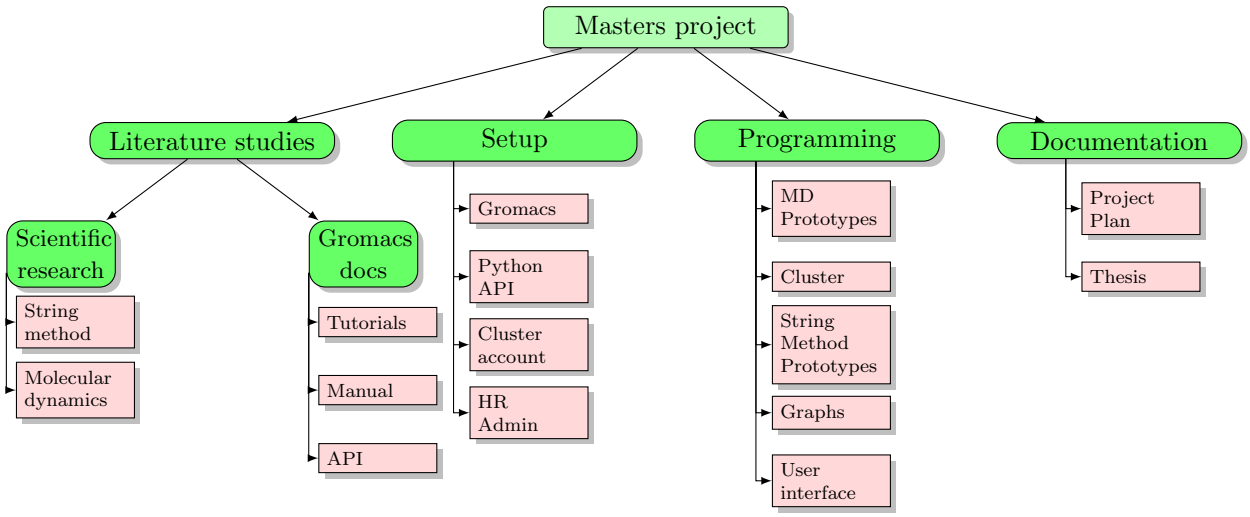
Gromacs is mainly a software package created to run one singular simulation at a time. In the string method, many iterations of many simulations need to be run with some calculations between iterations in

order to set the initial conditions for the next iteration. A lot of the calculations are similar or even the same and can easily be automated. With a good structure of automation the researcher workload can be focused not on repetitive tasks but rather interpretations and analysis.

Computer software ranges from games, through creative aids to scientific calculations and record keeping. As novel scientific software often is created by the researcher in need of it, the user interface design is commonly kept as simple as possible[3]. In the case of Gromacs, the main user interface is through the command line. Over time, Gromacs has grown and evolved updating it's calculation algorithm but also adding an application programmers interface (API)[4]. As Gromacs has evolved and grown, so has it's user base. With a wider user base, a command line interface for the program (and usually needing scripting skills for the operating system) creates a skill barrier for new users. The end result of one MD simulation is a trajectory file, this file then needs interpreting. Depending on the purpose of the simulation, the interpreting can be a 3D visualised model or movie, an atom to atom distance or binding angles.

In order to automate the iterations of MD runs and VIS updates a software needs to be created. If this software can lower the computer skill threshold for users and also automate some of the interpretations of simulation, the process of researching protein state transitions using MD will be significantly simplified.

4 Work breakdown structure



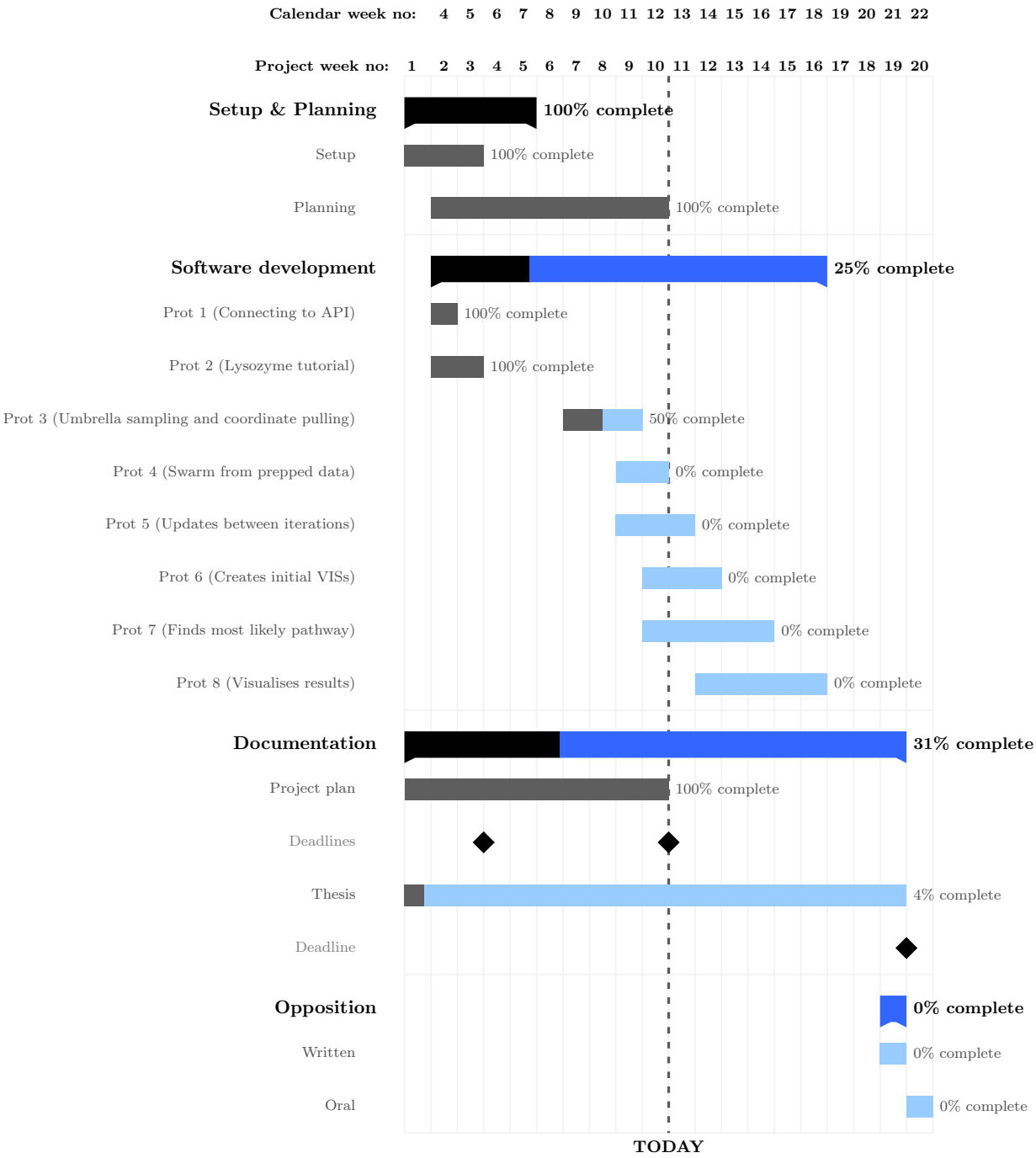
5 Milestones

The milestones of the project in intended order of completion:

1. Prepared Gromacs simulations can be launched from Python.
2. The entire Lysozyme Gromacs tutorial is automated from a Python program.
3. Initial project plan finished.
4. Running MD simulations on cluster.
5. Given a set path of initial VISs, the software can do one iteration of MD simulations.
6. The software can use the results from one iteration of the String method to prepare the next.
7. With hardcoded collective variables, a path of VISs is calculated.
8. The software can iteratively improve upon a path to produce a final most likely transition pathway.
9. The software can set up and run a complete transition pathway generation from known start and end states, and preselected collective variables.

10. The software can create graphs of collective variable and transition path iterations.

6 Time plan



7 Specification

As the objective of the project is to construct software the specification will be focused on the software.

7.1 Must

- Run Gromacs MD simulations for the user.
- Implement the string method with swarms of trajectories using Gromacs and the Gromacs API.
- Track and log the collective variables through a series of iterations of the string method.
- Update/create new topology files between runs.
- Work on both personal(UNIX/LINUX) computer and calculation cluster.

7.2 Should

- Prepare for the string method from pdb files and defined collective variables.
- Visualise the pathway in the collective variable space (graphs).

7.3 Could

- Identify collective variables.
- Automate 3D visualisations.
- Record movies.
- Assist in selecting MD parameters.

7.4 Won't

A graphical user interface was temporarily considered but would not be easily adaptable to how simulations are usually run. Although with some extra planning and considerations would be feasible.

8 Stakeholder analysis

Currently identified stakeholders and their interests in and influence over the project.

Stakeholder	Role	Potential benefits/interests
Marko Petrović	Thesis student	• Creating an application that will ease future research and contribute to own software portfolio. • Finish thesis project with a passing result. • Expand professional network.
Lucie Delemotte & Oliver Fleetwood	Supervisor & Co-supervisor	• Support the production of a useful tool for future research. • Further research using molecular dynamics and the string method using swarms of trajectories.

Patrik Ståhl	Examiner	• Ensure the quality of the thesis work and student learning. • Make sure the project runs smoothly and on schedule.
KTH	Title issuer & Thesis publisher	• Produce a good quality Masters graduate. • Publish a qualitative Masters thesis.
The gromacs team & their python API developers	Secondary stakeholder	• Recieve early feedback on a relatively unexplored feature in their software. • Improve usability of their software.
The MD community	Secondary Stakeholder	• Easier implementation of the string method.

Table 1: Stakeholders and interests

9 Business case

This project should provide a useful tool for simplifying the implementation of the string method as well as component procedures such as umbrella sampling and coordinate pulling as well as simplified analysis of MD simulations. This would primarily result in some reduction in work time for researchers. As far as environmental, societal and ethical matters are concerned the direct results of this project does not have any significant negative effects. Indirect effects would be through some research being somewhat less labour intensive and thus mainly reducing costs.

10 SWOT analysis

Strengths	Weaknesses
• Student has multiple applicable backgrounds. • Python based makes programming relatively easy to initiate and the finished product accessible and understandable. • Team of very knowledgeable and helpful coworkers provide support.	• Student has overambitious curriculum, working and taking extra courses during thesis project. • Easily distracted student. • As the base simulations are very configurable, any simplification created in the application may need to balance simplicity against configurability. This will result in the application also being complex in order to be a reusable tool rather than a demo.

Opportunities	Threats
• Further development of the software and expansion to other aspects of using Gromacs. • Increased popularity of Gromacs, which several stakeholders have interest in, as preferred molecular dynamics software. • Possibility to affect the development of the Gromacs API.	• The developed application software becomes too specific. • Software documentation doesn't get prioritised and though useful the software is rendered incomprehensible. • Delays due to Covid-19 as an infection of either primary stakeholder may cause significant bottlenecks.

Table 2: SWOT analysis

11 Risk evaluation assessment

Analysis of the risks identified in the SWOT analysis in 2.

Event	Prob.	Effect	Precaution
Overambitious curriculum	High	The project is delayed and deadlines are missed.	Some adaptation of the planning has been made as the other activities are far fewer and less intense during the second half of the project, shifting project workload to the latter half of the project.
Unfocused student	Medium	Delays due to student doing irrelevant or non related activities.	The initial plan was to make certain that student spent most time in an environment with passive surveillance (colleagues) and few distractions. As this is currently not an option ideas about live streaming the work is being explored.
Application too specific	Low	The software is not applicable to similar but not identical tasks.	Keeping this risk in mind as to not let the application make too specific assumptions during programming and testing the application on several test cases.
Software documentation	Medium	Difficulty and frustration for end users	A first step of internal documentation of the source code is implemented but this risk as of yet not adequately addressed.
Covid-19	Total & Low	Both mild as communication, at times, not being as effective as is usually expected and work totally halting due to infection.	Practicing social distancing as well as readjusting the home to a better working environment.

Table 3: SWOT Risk evaluation

The ethical aspects of the risk evaluation assessment are quite limited. The software in and of it self can not cause any harm and does not handle sensitive data and the development is as well.

12 Summary

The purpose of this project is to develop complimentary software to the molecular dynamics simulations application Gromacs implementing it's recently made available python based python API. The developed software will be meant to increase automation of the string method using swarms of trajectories and to some extent lower the complexity of use. The project is mainly performed on the students own personal computer, using freely available software such as Emacs and atom for writing source code, overleaf and LaTeX for writing reports, and VMD (free for academics) for system visualisation in 3D. Zulip, Zoom and email are the primary tools used for communication.

References

[1] Pan, A.C., Sezer, D. and Roux, B. 2008. Finding transition pathways using the string method with swarms of trajectories. The Journal of Physical Chemistry. B 112(11), pp. 3432–3440.

[2] Lemkul, J.A., Bevan, D.R. 2010. Assessing the Stability of Alzheimer's Amyloid Protofibrils Using Molecular Dynamics. The Journal of Physical Chemistry. B 114(4), pp. 1652-1660.

[3] Howinson, J., Herbsleb, J.D., 2011. Scientific Software Production: Incentives and Collaboration. CSCW '11, pp. 513-522.

[4] Gromacs [Internet]. [place unknown]: [date unknown]. About Gromacs; [date unknown][Updated 2018 Sep 24; Cited 2020 Mar 4]. Available from: http://www.gromacs.org/About_Gromacs