Project Plan
Vinay Kakkar
CS4821 - MSci Final Year Project

Structural Bioinformatics Framework using a MapReduce formalism

Supervised By: Hugh P. Shanahan
Department of Computer Science
Royal Holloway, University of London

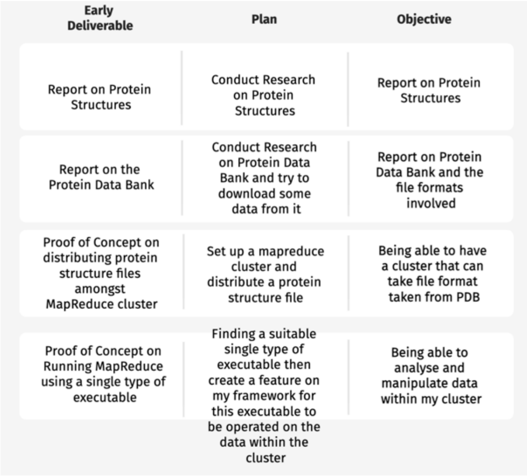
I. Abstract

This project plan will outline the topic “Structural Bioinformatics Framework using a MapReduce formalism” showing how I will conduct research and practical, which will help in the development in structural bioinformatics applications. The need for structural bioinformatics is to try create/use emerging technology in order to try manipulate and analyse biological macromolecular data as a result to solve problems in biology and generate new knowledge for example a new drug discovery. In order to do this, we will need to tackle some of the issues at hand for example the data sets can be exponentially big in this case we would want to use a MapReduce formalism on these data sets making them easier to work with at a biology standpoint. MapReduce is a popular framework for data-intensive distributed computing of batch jobs. It focuses on applying transformations to sets of data records, and allow the details of distributed execution, network communication and fault tolerance to be handled by the MapReduce framework [4]. Apache spark is an open-source powerful distributed querying and processing engine, that allows the user to read, transform and aggregate data [5]. Thus, it allows for fast, general-purpose, in-memory, iterative computing for large scale data processing whilst providing high fault tolerance and high scalability by introducing the resilient distributed dataset abstraction [1]. Specifically, this framework will be used to help with various data types provided from PDB and Alpha fold with features such as filtering for similarities, finding missing links between protein cell data sets etc. A data set is mapped into a collection of (key value) pairs. The (key, value) pairs can be manipulated. The result is a reduction over all pairs with the same key. Spark provides high-level APIs for Java, Scala, R and the Python language – pyspark it also provides interfaces for SQL, machine learning, and data graphs [6]. Using this I will be able to achieve creating a framework for MapReduce formalism. The general approach to MapReduce is the following: A mapping step that using the input of a data set it associates them with an output key. Followed by a shuffling step that groups result with the same output key. Finally, a reducing step that processes groups of results with the same output key [2]. Not only will conducting this research help me improve my knowledge of programming in python alongside the use of the spark API and implementing clustering strategies it will also let me indulge into the biology sector. I will need to conduct research on the protein data bank and specifically investigate on the biological macromolecular data I will be working on. At the end of this project using bioinformatics as a key science I hope to help both biology sector with the ease of working with big data and computer science sector with the development/knowledge of applications such as pyspark.

II. Early Deliverables

I will provide some deliverables which I will be conducting within my first term these will include also please refer to figure 1:

1. Report - Protein Structures
2. Report - The Protein Data Bank and the file formats used in accessing it.
3. Proof of Concept - distributing protein structure files amongst MapReduce cluster.
4. Proof of Concept - Running MapReduce using a single type of executable for analysing protein structures.



Early Deliverable	Plan	Objective
Report on Protein Structures	Conduct Research on Protein Structures	Report on Protein Structures
Report on the Protein Data Bank	Conduct Research on Protein Data Bank and try to download some data from it	Report on Protein Data Bank and the file formats involved
Proof of Concept on distributing protein structure files amongst MapReduce cluster	Set up a mapreduce cluster and distribute a protein structure file	Being able to have a cluster that can take file format taken from PDB
Proof of Concept on Running MapReduce using a single type of executable	Finding a suitable single type of executable then create a feature on my framework for this executable to be operated on the data within the cluster	Being able to analyse and manipulate data within my cluster

Figure 1 Early Deliverables

As my topic is related to a part of the biology sector, I will need to dwell into some research of Protein Structures and The Protein Data Bank as understanding the data I am working will help me determine if the output from my framework is correct but more importantly if it has the potential of being helpful to the biology department. More specifically Research on the Protein Structures will help me understand the data also help me create my single type of executable and research on the Protein data Bank will help me implement the data into my cluster. Thus, I will need to produce two proofs of concepts within pyspark which can demonstrate that working with such data is possible i.e., that I am able to analyse or manipulate the data provided from Protein Data Bank more specifically using a MapReduce cluster.

III. Timeline

Ideally, my plan will be to focus on implementation during term one and focus on the analysis, fine-tuning and evaluation in term two. In the case where I am on track, I will also leave some time to further expand my ideas this can be in the form of another report or more features to my program.

2.1 Term 1

Week 1: · Study papers and books about Protein Structures.
Week 2: · Study papers and books about Protein Data Bank.
Week 3-4: ·Read and complete tutorials on spark and python - pyspark.
Week 5-6: · Implement the code for interpreting data structure provided by protein data bank.
Week 7-8: · Analysing and manipulating a single type of executable to be run on my framework.
Week 9: · Fine-tune and optimise program.
Week 10-11: · Prepare for interim report and presentation.

2.2 Term 2

Week 1-2: · Augment the approach further by letting the user input a map and reduce step.
Week 3-4: · Implement visualisations and guidelines the user to understand the map/reduce executables the users is allowed to do.
Week 5: · Give the user the option to be able to return data in different formats.
Week 6-7: · Implement framework to be able to work on other data banks such as alpha fold which will include different inputs, functions, and calculations.
Week 8-9: · Re-evaluate final project results and extend to relaxations, time permitting.
Week 10-11: · Prepare for final report and viva.

IV. Risks and Mitigations

Whilst conducting my project there is some risks to consider. I will list out some risks and some actions I can conduct to try mitigating them. I will list the Risks down in order of Importance and Likelihood so that the first risk is least important, and has a low likelihood and the last risk is the most importance and has the most likelihood.

Delays in Timeline

Timelines is an important aspect for a successful project more importantly it outlines a structure on when tasks should be completed by and ensures that if stuck to the time scale we will have a finished and polished product at the end. In the case where some tasks longer than expected it can raise the issue of not being able to complete the project in time. To mitigate this, I have added in a week in the timeline as a safety net which I can fall back on. Furthermore, I will be assessing the time scale after every task so in the case where multiple tasks has drawn me back more than the safety net the timescale will always altered and adjusted so that I will still have a finished/working product.

Uneven Balance Between Report/Code

This project will be a mixture of a report and a program. They are both important aspects where if one is lacking it can affect the entire project. This can easily happen for example being tunnel minded on some code either trying to make it work or going beyond on the features where I end up using time I should be spending on the report. This case can also be flipped where spending too much time on the report over time I should be spending on the code could mean that the finished program is not up to standards or set out to be. To mitigate this, I will ensure when working on either code or reports I will be aware that the other will also need time spent on. I can do this by looking at my time scale and ensure that I have available time to add more features to the code or delve deeper into some research. In the case where I am stuck on some code or research, I will make sure I continue with the other specifically considering an easier or simpler solution for the time being until I get more available time to go back and investigate.

Using PySpark over Hadoop

Working on this project there are two main programs I can use Hadoop and spark both have features that would allow me to conduct this project. There might be a point in this project where I want to code a feature in spark which is difficult to implement but can be easily done on Hadoop. Hadoop is older and has more support online where spark is newer. I decided to pick spark as it can be used as a API on python – pyspark which should make some challenges easier due to my experience in python. I have made sure to complete pyspark tutorials so I should be aware of the capabilities it holds and will be able to make my program accordingly. If in the case where I am still stuck on a feature, I can consult with my supervisor to potentially look at another route of implementing such feature.

Data validity and data from different sources

For majority of my project, I will be working with data sets that are provided from different sources this can be either protein data bank or alpha fold. I will need to make sure that the data is valid as not valid data will hinder the essence of my project as it will be hard to see the beneficial standpoint of the program. These data sets can be taken from different sources so I will need to make sure that my program will be able to handle most variations of data types as inputs. To make sure this won't cause an issue I will need to create some proof of concepts that will show that my program will work with such data sets. I have also set some time for research on the protein data bank to make sure that the data is valid and understand the file types etc.

Limitations of hardware

This project will require me to work with big data this means that the data sets ie protein structures are too large and complex to be handled by traditional data-processing application software. Using spark to handle data from the Protein data bank will require loads of hardware specifications more specifically memory. This is an issue as trying to use all the data from the protein data bank will take too long or even crash my hardware. To combat this, I will ensure that spark is capped to about 75% of my computer's memory leaving the rest of general application and operating system whilst at the same time I will be working with a snippet of the data so that I can work on my program in a more efficient manner i.e. when running experiments or tests getting a faster and easier output will help me develop my program quicker and efficiently.

Computational Efficiency

Again, this project requires work to be conducted on big data. This means that if functions or solutions to sub-problems are not coded in a efficient manor it can have an exponential effect on hardware specification and the time it takes to provide a output. This means that it will risk losing its advantages aspect when analysing protein structures at a biology stand point as the current set system will be better at doing so or just as good. To combat this from the start of the project I will be evaluating the theoretical running time and estimating the output for a single type of executable, on top of measuring practical running time/baselines of the code throughout the project.

Overhead with working with Big Data

Working with Bigdata can be time-consuming specifically trying to understand that the program is working correctly this is all dependent on how big of a data set I will be working with. For example it can be overwhelming if the data set is very big as the output will show a lot making it hard to analyse the output and judge the success of it. To try and minimise this from the outset I will ensure to follow good practices when working with spark this can include stating with small data sets and ensuring I understand the basics which I have kept a week dedicated to completing spark tutorials

Overhead of Biology Knowledge

This project steps into the field of Biology, I understand the dissertation topic presented but there are multiple parts of biology that will arise whilst working on this project more specifically the data I am working with might cause some issues. For example, not being able to understand the input or output of the data. This will make it hard to specify to the program how to handle the data correctly and it will be hard to ensure that the program is handling the data correctly from looking at the output. To combat this, I have set time to research into the protein structures which is the data I will be working with, and I have set time for research into the protein data bank which is where the data is stored and will be providing me with this data ie the file formats. Should any issues arise I have been recommended some books and papers to read into which explains more into these two aspects. I will also consult my questions with my supervisor in the case where books and papers are not enough.

V. Planning and Time-Scales

For term 1 most my objectives are based around completing the early deliverables. In order to complete them I have had to set up objectives with dates in mind so that I can keep on track to having a finished framework and final report by the end of Term 2. The first two objectives involve reading and writing a report which should both be completed before the specified date. I will continue reading throughout this project but the first two are vital to understand for this project. This will set me up to then start on my Proof of Concepts as it will give me the background knowledge to implement and experiment with data from the protein data bank on to a cluster. I have added a fine tuning and safety net week for any drawbacks or hiccups. You can see this in figure 2

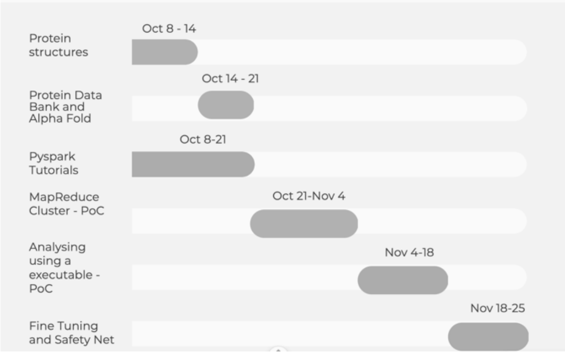


Figure 2 Term 1 Time Scale

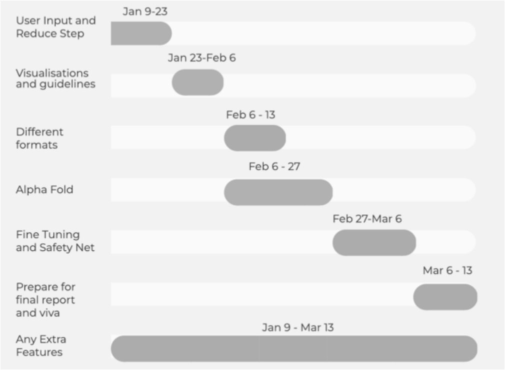


Figure 3 Term 2 Time Scale

With Term 2 the time scales refer to figure 3 the objectives are subject to change as apart of delays in time scale risks, one of the mitigations was to continually look at the time scale and asses and change the time scale if necessary in order to either add more features, conduct more research, or have to remove some things, which is in the case where I have been drawn back or stuck. However, they do represent what my final goals are and a what dates certain features need to implement and completed by. These objectives outline what needs to be completed on order for my final project to be completed. I have also added a week in for fine tuning and polishing of my reports and program. I have also added a slot for Extra Features which could be adding more computers to a cluster for managing bigger data sets or to add better User Interface to the program.

VI. Bibliography

[1] Guo, Runxin, Yi Zhao, Quan Zou, Xiaodong Fang, and Shaoliang Peng. 'Bioinformatics Applications on Apache Spark'. GigaScience 7, no. 8 (7 August 2018): giy098. https://doi.org/10.1093/gigascience/giy098.

[2] 'BigData with PySpark: MapReduce Primer'. Accessed 7 October 2022. https://nyu-cds.github.io/python-bigdata/02-mapreduce/.

[3] Zvelebil, Marketa J. Understanding Bioinformatics / Marketa Zvelebil & Jeremy O. Baum. New York: Garland Science/Taylor & Francis Group, Garland Science, Taylor & Francis Group, 2008.

[4] Elmeleegy, Khaled, Russell Sears, and Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein. 'MapReduce Online', n.d., 15.

[5] Drabas, Tomasz, and Denny Lee. Learning PySpark. Packt Publishing Ltd, 2017.

[6] Lovrić, Mario, José Manuel Molero, and Roman Kern. 'PySpark and RDKit: Moving towards Big Data in Cheminformatics'. Molecular Informatics 38, no. 6 (2019): 1800082. https://doi.org/10.1002/minf.201800082.