MSc Bioinformatics with Systems Biology - Biocomputing II: Project Essay

Approach to the project

Interaction with the team

As this was a group work assignment, collaboration with other students formed a large part of the project. It was decided that we should communicate using different apps and websites (Github, Trello and Messanger) to ensure that we were in constant contact throughout the assignment, with initial conversation made in person before lectures, latterly by remote communication and in the run up to the deadline, in person to check each member had performed what was required.

We used Messanger for direct communication to organise meetings, Trello to created 'to do lists' and outlined what was required for each layer and finally Github was used to store all our codes and changes. I was not as engaged with the other two members of the group as they were with each other, as I directed my research to the project and fed back once I had achieved my goals. I attempted to participate in as many group interactions as possible, both face to face and remotely. This was due, I feel, to work commitments I had meaning I was less available to meet and also that I am based in south London, whilst the other members were centrally based. As a result I communicated remotely much more with the other members of the group, who preferred to meet and talk directly.

Overall project requirements

After reading the specification for the project together, we talked through the requirements needed for the project to succeed and later assigned project layers to team members. Together we made a list of tasks needed for each layer and how to structure them. We discussed what type of problems were likely to be encountered during the development of each layer and how we could solve them, discussing in depth, such that everyone would understand what needed to be done and brainstormed new ideas in order to improve our strategies before starting the project.

Requirements for my contribution

I was concerned over being in charge of the middle layer as it was the area that I felt least comfortable doing as it was not immediately apparent how to communicate between the database and the website layers. However, as the other members knew the layer they wanted to do, I decided to challenge myself and go for the middle layer.

Looking at the specification for the middle layer, I went through each part understanding what was needed from me and that I was the communication link between the database and the website. I knew that I had to create APIs as soon as possible and functions that would either retrieve data from the database and pass it on to the website or create data calculations such as the codon analysis and restriction enzyme recognition sites and position.

Performance of the development cycle

Overall as a group I believe we worked very well throughout the development cycle. To start, as we still had lectures together, we made sure that each of us understood what was required and together went through any problems that we might find. We would also shared any code that we had created and tested it to find any areas needing improvement.

During the term break we did not meet as much as before but still communicated through different platforms when new code was created or if we had any problems with the designing/development of the tasks.

My part of the development cycle relied upon spending time developing codes and trialing them before contributing the finished article to the group. Development was spent understanding areas that the codes needed to be developed on and later formulating codes that would satisfy these requirements. My cycle required understanding the format of the database and taking inputs from the website end, producing outputs. I began by researching various ways by which I could create my codes and later wrote and tested them. Once I had produced my codes, I worked with the website end to ensure that my codes were easy to understand and readily transferable and with the database end to make sure that the data that I was retrieving was present in the specified table and format.

The development process

I spent a lot of my time at the designing stage and only a shorter time in the development stage, in so doing I tested less by trial and error and postulated code through detailed research. Knowing what type of tasks I had to carry out, I investigated how to write functions and what type of coding to use. To do this, I went through my notes from Biocomputing I, II and Data Management as well as online forums and website to get a better understanding of my role and the type of codes available. I later put my notes together and created different codes that would extract both data from the database and perform calculations like the codon usage frequency. I learned how to use Biopython to retrieve the position of the restriction sites of the enzymes on the DNA sequences and took into account the enzymes products (sticky ends or blunt ends) when creating the codes.

Code testing

Upon creating code I tested it first to see if it worked instead of carrying on and only testing at the end. For every code that I wrote, I always tested it by using the print function and made sure that either my output was what I thought it would be or that it was on the right path. Sometimes it did happen that I needed up or downstream codes to perform a test, and if the code was not available I created a dummy version to ensure that my code produced the intended output.

Known issues

Throughout my coursework I had several issues accessing No Machine, and therefore the database, as I was not able to login from my laptop. Most of the codes that access the database were developed following the guidelines provided in the Data Management course but were tested quickly at the very end. Everything else works fine and should not cause any problems.

What worked and what did not – problems and solutions

Despite my rudimentary knowledge of coding, I was able to create good codes that worked. My functions on both the finding of the restriction enzyme binding sites and the calculations of the frequency of the codon usage worked well. Finding restriction sites for different enzymes, I decided to use regular expression to retrieve the site for BsuMI. As it creates blunt ends it was quite straightforward to find the position of the cut. For the other two enzymes (BamHI and EcoRI) I decided to retrieve the positions using BioPython, demonstrating my coding versatility.

I had problems while I was developing the codon frequency function codes, as it was a longer and more complex code. My main issue was in the calculation of frequency, as I was not able to make the ‘for loop’ to calculate the frequency for each codon used. Howeve, I was able to overcome this and produce a functional code.

I feel that there are areas of my work that could have worked better, especially the creation of the link between the webpage and the middle layer as it was something completely new to me. I had no issues querying the database for the different tasks but experienced some in linking them to the website end. I invested considerable time researching the best way to “read” the inputs from the website and give out the right output.

As a group we did interact with each other as much as possible and made sure that everything was going in the right direction. During the term break I had a period of time where I knew I could have not been available but I did make sure that I communicated it with the others beforehand. Overall, I am very happy with the way that we worked as a group and how we helped each other if there was a problem.

Alternative strategies

While I was planning my work, I did find different ways to produce the same results. In fact at the beginning, for the restriction enzyme sites, I created all three functions using regular expression and only later adopted BioPython for EcoRI and BamHI enzymes. This was done to demonstrate my ability to use both regular expression covered by the course and new BioPython methods I had researched. For the APIs that extracted data from the website I could have used XML::DOM, however, after discussing it with the person in charge of the website section we decided that the best way would be to connect them using only Python through functions.

Personal insight

After doing this project, I feel more confident about my computational skills and the ability to combine what was taught to me previously with new techniques that I have learned along the way. Both my Python and SQL skills have improved since the beginning of the module and I feel that if I continue to apply my skills I can develop them further. I especially enjoyed reading up and trying out BioPython as I think it is a really nice combination of the biology side learned during my undergraduate with my coding knowledge from this Masters. I now feel more confident in writing scripts without the need for in depth prior planning and am learn from my mistakes. This experience has showed me what a possible future in bioinformatics could look like.