Analysing images of protein distributions using fluorescent microscopy

CS350 Data Science Project

Project Specification

Kim Ta 1907156

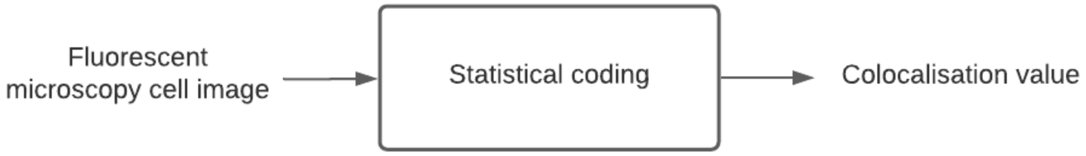
October 2021

Contents

1 Introduction

In this project, I aim to build an application for users to analyse the colocalisation of a cell distribution from an image they have uploaded into the software. This would be useful in determining whether two or more biomolecules are affiliated with the same cellular structures.[1]

Figure 1: The process context model



To understand the interaction between proteins, we use colocalisation as an indicator. The correlation between proteins help to determine the location of cellular structures of interest and identify features that they have in common, giving us a deeper understanding of cellular processes such as cell division and cancer. As a result, this information improves our understanding of how proteins respond to their environment and the effect of different interventions.

More recently, there is an increasing volume of data being produced. This is the result of automation in labs. Automation replaces repetitive manual work, improves the accuracy of data by removing human error and increases the reproducibility of an experiment. It opens up opportunities for novel research pipelines such as mass screening of proteins to test their suitability for cancer treatments.

This means we need to consider image scaling, this is resizing the digital image using geometric transformations with no loss of image quality. Scaling is important because some imaging software processes only work on 8-bit images, so downsizing is usually required.[2] Also, transformations can help with visualizing data.

[1] Michael Greenwood, Importance of Colocalization Studies, `https://bit.ly/3DuhJa0`

[2] Stephen J Royle, The Digital Cell: Cell Biology as a Data Science, Page 29

2 Background

In biology, we refer to colocalisation as being the observation of the spatial overlap between two(or more) cells and cellular components.[3] These cellular components will be identified though fluorescent microscopy, providing the images for data analysis.[4]

From the statistical point of view, we want to quantify the colocalisation in the cells. When we look at images of a cell, we can interpret the image as just a collection of pixels that can be organized in a 2D matrix, where each pixel is represented by a number at that location in the matrix [5].

We can use Pearson's Correlation Coefficient (PCC) to characterize the degree of overlap between two channels in a microscopy image, the equation we will use can be seen below [6]:

$$r_p = \frac{\sum(R_i - R_{avg})(G_i - G_{avg})}{\sqrt{\sum(R_i - R_{avg})^2 \sum(G_i - G_{avg})^2}} \tag{1}$$

With R_{avg} and G_{avg} as the averages of the R and G channel respectively and the summations with index i (pixel index) over all the image voxels. R_i and G_i are the Red and Green intensities of the pixel i. This will generate a range from 1 to -1, (1 being a perfect correlation).

PCC measures the pixel-by-pixel covariance in the signal levels of two images, and is independent of signal levels and signal offset. PCC can also be measured in two-color images without any form of pre-processing, making it both simple and unbiased. Alongside this, tools for quantifying PCC are provided in nearly all image analysis software packages making it very accessible.

[3] Wikipedia, Colocalization, `https://en.wikipedia.org/wiki/Colocalization`

[4] Fluorescent Labeling, `https://bit.ly/3BN1xQV`

[5] Stephen J Royle, The Digital Cell: Cell Biology as a Data Science, Book

[6] ColocalizationTheory, Scientific Volume Imaging, `https://svi.nl/ColocalizationTheory`

Another metric we can use to quantify the degree of colocalisation between the fluorophores is Mander's Overlap Coefficient (MOC):[7]

$$\frac{\sum_i (R_i \times G_i)}{\sqrt{\sum_i R_i^2 \times \sum_i G_i^2}} \tag{2}$$

MOC is more intuitive for measuring colocalisation compared to PCC, its more useful for data that are poorly suited to the simple linear model that underlies PCC and is more appropriate for 3D analysis of colocalisation. The major drawback is it is complicated. It needs to reliably identify background levels in an image and thus identify labelled structures.

3 Problem Discussion

Throughout my research, I will try to answer the following:

- What is the spatial distribution of a given protein and how can we describe the interaction of two or more proteins as a function of their location in space?
- Along with; what governs the spatial organisation of structural elements of a biological cell?

The ideology of identifying patterns in cell biology will challenge how I think, as I will be applying my statistical knowledge towards a biological concept and combining it with my technical skills.

There already exists pieces of software such as Fiji and ImageJ [8], allowing users such as biologists to perform image analysis tasks interactively. For processing large numbers of images, script-based data analysis is superior as it scales up. To create my application, I will implement PCC in R on images and make a feature that allows the selection of regions of interest from raw images, both manually and automatically - with automation to be guided by an optimality condition. Ultimately, I will create an app with the specific aim to simplify the process of analysing colocalisation for biologists.

[7] Kenneth W.Dunn, Malgorzata M.Kamocka, John H.McDonald, A practical guide to evaluating colocalization in biological microscopy, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074624/

[8] Stephen J Royle, The Digital Cell: Cell Biology as a Data Science, Book

4 Objectives

The main focus of this project is to understand and analyse the distribution inside a protein cell by measuring colocalisation. I will approach this by comparing the mathematical and statistical properties of different colocalisation measures.

4.1 Goals

My primary technical goals (classified as 'Must do') for my application are to:

1. Easily upload images into the applications
2. Convert images into tangible data (matrices) using R
3. Allow manual selection of region of interest
4. Give an interpretation of colocalisation and of adjustment to the window through scenario studies
5. Analyse the colocalisation of an region of interest using Pearson's correlation coefficient
6. Use R shiny to build a web app that allows users to explore colocalisation measures conveniently

My secondary technical goals (classified as 'Should or Could do') are to:

1. Automate detecting the region of interest
2. Give an interpretation of the outputted data

Note: I am using the MoSCow Requirements classification to help prioritise my objectives to focus on the most important aspects of the project.

My personal goals are to:

1. Build my background knowledge on cell biology through intensive research
2. Use data analysis in industry for biomedical data science
3. Be proactive and manage my time well during this project

4.2 Possible extensions and future plans

Despite the project area being quite specialised, the project can further model how cells colocalise by creating predictions from previous data and statistical patterns. In the introduction, I briefly mentioned the idea of protein interactions linking to cellular processes. An expansion of the application could include helping to detect patterns in certain cells or giving inference to if certain health conditions could be present in the sample. Given the time constraints, this will not be explored in detail.

5 Project Management Methods

Organisation requires both good planning and flexibility. To complete the project to a high standard, I will follow a plan-driven method integrated with agile methodology during the coding stages.

I chose to do a more plan-driven method as it wouldn't be wise to force a SCRUM or XP programming style methodology onto the project (strict agile). This is because I have regular coursework's for other modules, making it more difficult to perform things like a SCRUM cycle in a short time period.

5.1 Research

During my project, I will be reading about how experiments are done, how data is stored and what type of information/data is important for cell biologists. With a good foundation, it will help me understand what output my target audience (a cell biologist) would want when looking at a protein cell, along with providing inspiration for additional features. Therefore I heavily focus on research during the start of my project.

5.2 Development

To introduce flexibility into my method, I will also be using an incremental software development method for the coding process. Feedback from my meetings with my supervisor and peers are very valuable and will be incorporated into my work at each stage of the project. To make sure my code is functioning well, regular acceptance tests will be performed for each phase. By separating my project into chucks, it allows documentation to be kept to date, as I will be able to constantly record my progress and look back at my specification and see what I am missing or what I want to improve.

For the development of the application itself, I will first be creating the R code in order to perform the objectives I have stated above. This will be the base code for the functionality of my app. After I have a working prototype, I can use R shiny and look at templates to work on the UI to make it easy to use and generally make it more accessible.

5.3 Testing

It is a good practice to regularly test code in order to make sure it is functioning. My incremental approach means that I will be testing every section of code I create with a sample of data, and cross referencing the output with results that I would have expected.This is known as unit testing, which is very beneficial as consistently testing small chunks of code means that I can reduce the time spent of debugging when combining my chunks of code together into one big code. This is because it becomes much easier to narrow down on where the errors are present.

I will separate the data given in two sets of data, let's call them trial data and testing data; one for creating the code, one for after the application is created. The data will roughly be split in the ratio 2:1 as a standard practice, the second set of data is used to remove bias when creating the application, so that I am not trying to fit the model to a specific set of data.

6 Timetable

I have used a Gantt chart to timetable my project. This will be my schedule for two terms, it includes my assignments for other modules to help me manage my time and the workloads. [See appendix for Gantt chart images]

In addition, every Thursday I will be having a meeting with my project supervisor. My timetable includes the areas that I want to work on (but not limited to) each week and what category of work it goes under, the choice of not setting an exact date gives me the flexibility to work around my schedule when necessary. In terms of dependencies, each coding task will depend on the previous one, which is why they don't overlap, and will require testing at every section.

Each schedule was made on lucid chart, where I can adapt and change it where necessary, It also mimics a calendar that can be printed and for a physical copy.

6.1 Potential Risks

Creating unrealistic and crammed timelines would result in deadlines not being met. To overcome this, I created a Gantt chart with sufficient time for each task and time to test out the code created at each stage. I used milestones to help mark my progress. I also took into account other commitments and deadlines.

A event that may disrupt my schedule is illness or being unfit to work, so there is some leeway time added in the schedule. I have also prioritised my tasks, so that I can concentrate on the essential parts in this situation. I specifically used the MoSCow requirements classification to choose what tasks I will prioritise, this gives me the flexibility to add more features to my application if I want to (given I have the time and resources) without compromising the schedule.

7 Resources

For research and literature, I will be basing my background information on online resources, books and articles - referenced in the footnote of pages. The dataset we will be using is fluorescent microscopy cell images (only consisting of the two colours red and blue) provided by collaborators from Warwick Medical School and online repositories.

The following technologies will be used to help create my application:

- Git – version control
- GitHub – GUI for git
- R – for statistical computing and graphics to clean, analyse, and graph my data
- RStudio - I will use tools implemented R-packages spatstat (on Cran) and colocr to quantify how proteins interact
- R Markdown – to organise my code, and to keep my results reproducible
- R Shiny – to create the app

7.1 Risks

My first risk would be the data set of images becoming unavailable. The resulting impact of this would mean I will not be able to test my solution or guarantee accuracy. As a contingency, if I am unable to source these images from another location, I will use images from free online data sets such as IEEEDataPort. This may not be as reliable but the statistical analysis used in the application does not directly depend images used in testing.

I also have to consider all the software and tools I will be using. For R and RStuido, I will be using the latest programming versions as it is better for security and contains more accessible packages. A limitation would be that some programs may become depreciated and fall out of use, making it harder to access the software to develop.

I would like to mainly keep my documents on a personal hard drive for security purposes, however, a big risk is a hard-drive failure. In this event and I do not have a backup source, I will lose all my data. So firstly, I should always have some sort of back-up data. My way around this would be using a GitHub repository, my code is then always backed up and can still be read despite a hard drive failure. This version control also lets me revert changes in the event I break my code beyond repair. Also, an advantage of using this is for collaboration purposes as is that it is very accessible. In bigger projects, it is useful for sharing data and for other researchers to validate or build on your ideas[9].

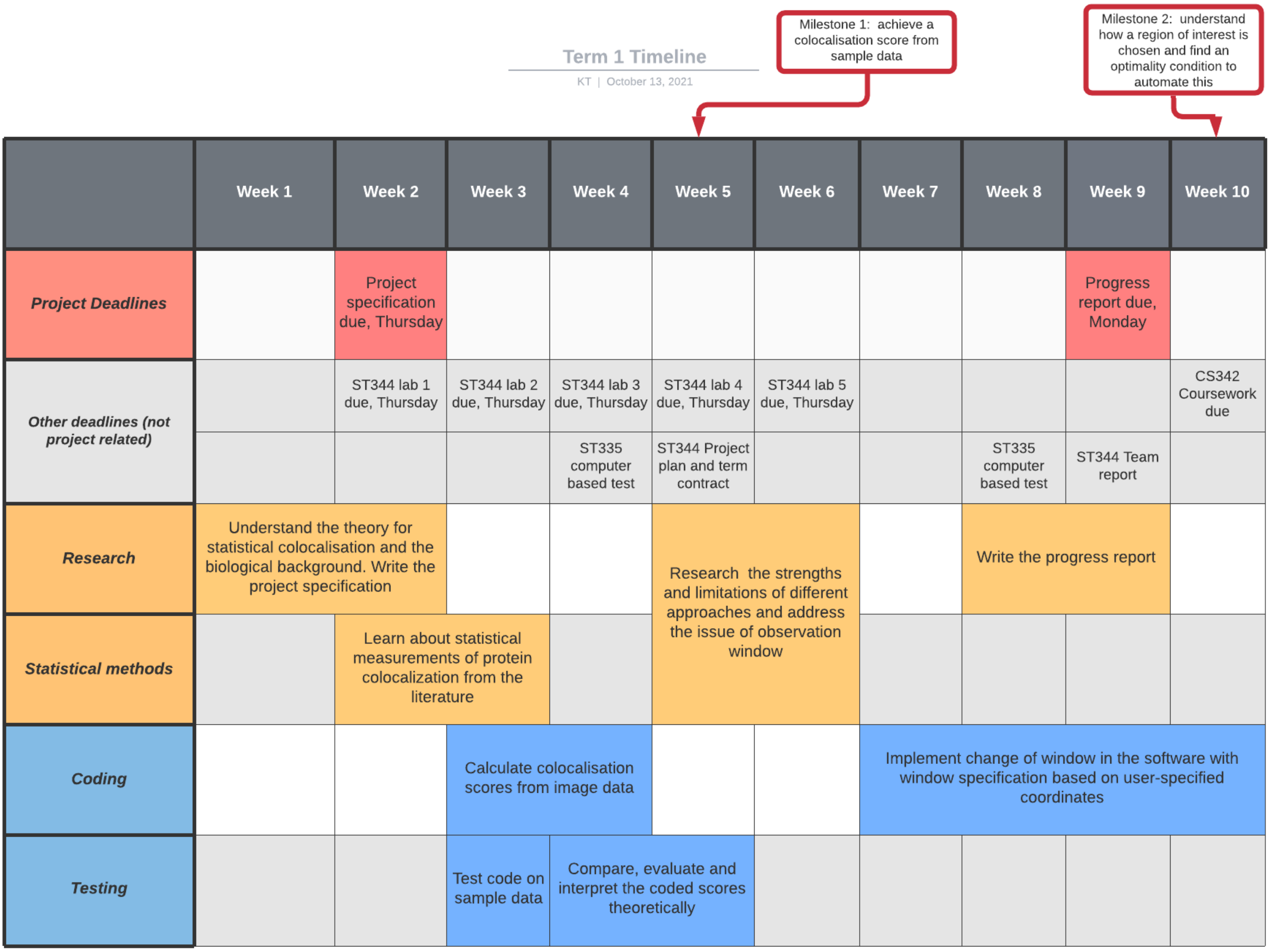
8 Legal, social and ethical considerations

There are no such issues to consider as I will only be working with data from the Warwick Medical School and online repositories, so I will just need access and permission. I am assuming that the data has been collected using the appropriate practices and to protocol requirements.

[9] Stephen J Royle, The Digital Cell: Cell Biology as a Data Science, Book

9 Appendix

Figure 2: Gantt chart: Term 1



Term 1 Timeline

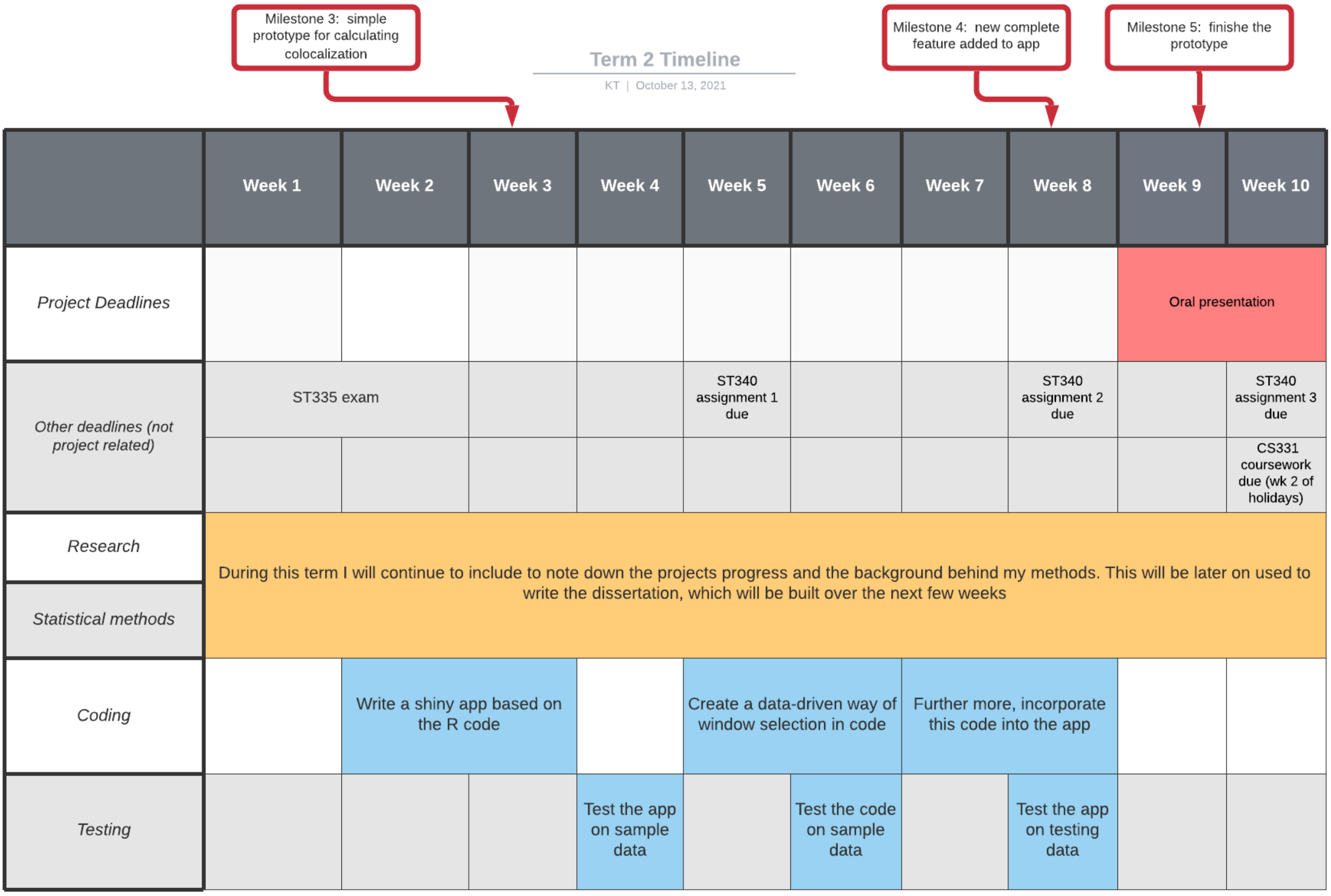
KT | October 13, 2021

Milestone 1: achieve a colocalisation score from sample data

Milestone 2: understand how a region of interest is chosen and find an optimality condition to automate this

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Project Deadlines		Project specification due, Thursday							Progress report due, Monday	
Other deadlines (not project related)		ST344 lab 1 due, Thursday	ST344 lab 2 due, Thursday	ST344 lab 3 due, Thursday	ST344 lab 4 due, Thursday	ST344 lab 5 due, Thursday				CS342 Coursework due
				ST335 computer based test	ST344 Project plan and term contract			ST335 computer based test	ST344 Team report	
Research	Understand the theory for statistical colocalisation and the biological background. Write the project specification				Research the strengths and limitations of different approaches and address the issue of observation window			Write the progress report		
Statistical methods		Learn about statistical measurements of protein colocalization from the literature								
Coding			Calculate colocalisation scores from image data				Implement change of window in the software with window specification based on user-specified coordinates			
Testing			Test code on sample data	Compare, evaluate and interpret the coded scores theoretically						

Figure 3: Gantt chart: Term 2



Term 2 Timeline

KT | October 13, 2021

Milestone 3: simple prototype for calculating colocalization

Milestone 4: new complete feature added to app

Milestone 5: finishe the prototype

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10
Project Deadlines									Oral presentation	
Other deadlines (not project related)	ST335 exam				ST340 assignment 1 due			ST340 assignment 2 due		ST340 assignment 3 due
										CS331 coursework due (wk 2 of holidays)
Research	During this term I will continue to include to note down the projects progress and the background behind my methods. This will be later on used to write the dissertation, which will be built over the next few weeks									
Statistical methods										
Coding		Write a shiny app based on the R code			Create a data-driven way of window selection in code		Further more, incorporate this code into the app			
Testing				Test the app on sample data		Test the code on sample data		Test the app on testing data		