# Analysing images of protein distributions using fluorescent microscopy

CS350 Data Science Project

## Progress Report

Kim Ta 1907156

November 2021

# Contents

# 1 Introduction

Microscopy is the use of microscopes to view samples and objects that cannot be seen by the naked eye, such as cells. It uses visible light to illuminate and produce a magnified image of a sample. Therefore, images are created from transmitted light. In biological history, microscopy has played an important role in tissue analysis, studying atomic structures and studying the role of a protein within a cell.

In this project, fluorescent microscopic images are used to analyse the colocalisation between proteins, where images from the blood of patients with COVID-19 will be considered.

Fluorescence microscopy produces images of molecules that have been tagged with a fluorescent dye (known as fluorophores - added to the sample before being imaged). Compared to the conventional microscope, a fluorescence microscope uses a much higher light intensity source to excite the fluorophores, causing them to give off light with lower energy and of longer wavelength, this produces the magnified image. Therefore, images are created from reflected light. [1]

In biology, we refer to colocalisation as the spatial overlap between two(or more) cells and cellular components. [2] These cellular components will be identified through fluorescent microscopy, providing the images for data analysis. [3]

The colocalisation between proteins gives 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.

A fluorescent microscopic image can contain more than one cell, we usually want to measure a specific subset of that image for a particular purpose, e.g. only one cell from the image without the background noise. This is known as the region of interest (or ROI).[4]

[1] Ryding, S., “Fluorescence Microscopy vs.Light Microscopy”. https://www.news-medical.net/life-sciences/Fluorescence-Microscopy-vs-Light-Microscopy.aspx. Published: 27/11/2018. Accessed:27/11/2021.

[2] Spring, K R., “Colocalization of Fluorophores in Confocal Microscopy”. https://www.olympus-lifescience.com/en/microscope-resource/primer/techniques/confocal/applications/colocalization/. Accessed: 26/11/2021.

[3] PromoCell., “Fluorescent Labeling”. https://bit.ly/3BN1xQV. Accessed: 26/11/2021.

[4] Techopedia, “Region of Interest (ROI)”. https://www.techopedia.com/definition/339/region-of-interest-roi. Published: 24/04/2013. Accessed: 27/11/2021.

# 2 Background

## 2.1 COVID-19

Recently, new data regarding COVID-19[5] and its effects on blood clotting have surfaced, containing fluorescent microscopic images from blood samples. It has more data samples and a greater variety of shapes/cells than the previous dataset, making it more concrete and specific.[6]

In medical imaging, region of interest (ROI) selection is commonly used to identify the volume or size of specific components in an image. There exist mixtures of methods and algorithms for calculating and automating this. We would like to understand, compare and test these methods against our data.

ROI can be used to outline the shape of the structures in the image, helping to calculate the area (size) of said structure(s). A particular example would be examining plasma cells in the blood. Once the ROI is identified, we can calculate the percentage coverage of the cells to determine a clots size and its severity.

The automation of ROI is very useful in the medical industry. Take the initial wave of COVID-19, automated screening of blood could identify patients with severe blood clotting so that they get prioritised treatment. It saves time from visually assessing countless images when doctors are needed elsewhere (i.e. overfilled hospitals). It gives quantitative reassurance to help doctors come to a diagnosis or provide an initial assessment for when specialists are not available.

## 2.2 Data

SARS-CoV-2 (or more commonly known as COVID-19) is characterised by an illness of short duration, it can be rapidly progressive and in need of urgent care. It can result in excessive blood clotting, where clots can end up travelling to vital organs. Recently, a new COVID-19 phenotype has been observed in recovering patients known as Long COVID (or PASC: Post-Acute Sequelae SARS-CoV-2 infection). Symptoms include muscle weakness, shortness of breath, sleep difficulties, anxiety and depression. These can persist for as much as 6 months (or longer) after the acute infection.

There have been results showing significant failure in fibrinolytic processes during COVID-19 and with long-COVID symptoms, the plasma in the proteins in both COVID-19 and long-COVID patients were greatly resistant to breaking down clots in the presence of trypsin (enzymes that help digest proteins). This was confirmed using fluorescent microscopy.

[5] Pretorius, E. et al., “Prevalence of readily detected amyloid blood clots in ‘unclotted’ Type 2 Diabetes Mellitus and COVID-19 plasma: a preliminary report”. https://cardiab.biomedcentral.com/articles/10.1186/s12933-020-01165-7. Published: 17/11/2020. Accessed: 02/11/2021.

[6] Pretorius, E. et al., “Persistent clotting protein pathology in Long COVID/ Post-Acute Sequelae of COVID-19 (PASC) is accompanied by increased levels of antiplasmin”. https://cardiab.biomedcentral.com/articles/10.1186/s12933-021-01359-7. Published: 23/08/2021. Accessed: 02/11/2021.

# 3 Aims and Objectives

To understand and analyse the distribution of proteins in a cell, we use the measure colocalisation. In this project, we apply colocalisation to particular regions of fluorescent microscopic images to do so.

The first approach involved investigating the different methods of identifying regions of interest. We started with learning the theory behind each method, this gave us multiple ways we could automate the ROI selection (if not already done). This was then followed by analysing the effects of the ROI application on colocalisation. The next aim is to:

- Find mathematical reasoning (working on the proof) for the effect on colocalisation of removing the blank space from an image.
- Determine the effects of; removing background noise (not space) from an image, and the misalignment of image channels.

Practical wise, our first steps were to upload the fluorescent images in R, perform ROI, and then analyse the colocalisation of images. Now, we need to:

- Implement seeded growing region (explained in section 4.1.2) as a method of automating ROI
- Use R shiny to build a web app that allows users to explore colocalisation (with seeded growing region as the automated ROI finder)
- Given fluorescent microscopic images of a blood sample, understand how blood clots can be detected and how this can be automated

# 4 Progress

## 4.1 ROI and colocalisation

Images from fluorescent microscopy will usually contain more information than needed (extra parts of the image that are not of interest). We are only interested in the sections of the image where the cells are present, or where the proteins spatial distributions overlap. The aim is to automate the region of interest selection, therefore different methods to do this will be explored below.

#### 4.1.1 Automated Threshold Detection Method (Costes)

Costes et al.[7] developed an approach that identifies the threshold value by:

- Measuring the Pearson Correlation Coefficient (PCC) for all pixels in the image
- Then measuring PCC again for the next lower red and green intensity values (or whatever two colour channels are being used)
- This process is repeated until pixel values are reached for which PCC drops to or below zero

This threshold value can be used to identify the background of an image based on the range of pixel values which returns a positive PCC value.

It is a robust and reproducible method that can be easily automated, both speeding up processing and eliminating user bias. It is effective for images with high signal-to-background ratios and struggles with images that have very high labelling density or large differences in the number of structures labelled with each probe.

#### 4.1.2 Seeded Growing Region Method

Region growing[8] is a method in which an initial set of small areas are iteratively merged according to a constraint. Applied to image analysis, we can identify a region of interest by using thresholding as a constraint.

Start by placing a set of seeds in the image to be segmented, where each seed could be a single pixel or a set of connected pixels. Then SRG grows these seeds into regions by successively adding neighbouring pixels to them. It finishes when every pixel in the image is assigned to one (and only one) region.

#### 4.1.3 Fuzzy Set Theory

The fuzzy set theory[9] is applied to image thresholding to partition the image space into meaningful regions. The method averages the grey level of a pixel by comparing it to the pixel's neighbours (given it's above the threshold) to smooth the image.

[7] Dunn, K., Kamocka, M. and McDonald, J. ”A practical guide to evaluating colocalization in biological microscopy.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074624/ Published: 5 Jan 2011, Accessed: 21 Nov 2021

[8] Fan, M., Lee, T., ”Variants of Seeded Region Growing”. https://bit.ly/3xb0gla. Published: Jun 2015, Accessed 21 Nov 2021

[9] L.K. Huang and M.J.J. Wang. Image thresholding by minimizing the measures of fuzziness. Pattern recognition, 28(1):41–51, Published: 1995, Accessed: 21 Nov 2021.

The measure of fuzziness usually indicates the degree of fuzziness. We can measure this via entropy:

$$E(A) = \frac{1}{nln2}\sum_i S(\mu_A(x_i)), \quad i = 1, 2, ....n.$$

Where $\mu_A(x_i)$ denotes the grade of possessing some brightness property $\mu_A$, which is the distance between the grey tone image and its nearest two-tone version.

This is based on Shannon's function:

$$S(\mu_A(x_i)) = -\mu_A(x_i)ln[\mu_A(x_i)] - [1 - \mu_A(x_i)]ln[1 - \mu_A(x_i)].$$

To measure the entropy of an image set $X$, we can extend these equations to a two-dimensional plane:

$$E(X) = \frac{1}{MNln2}\sum_m \sum_n S(\mu_x(x_{mn})) \quad with\, m = 0, 1, ..., M - 1\, and\, n = 0, 1, ..., N - 1$$

When we talk about the index of fuzziness, we refer to measuring the distance between the grey-level image and its crisp version (average amount of fuzziness). As for nonfuzziness, this takes the absolute difference between the crisp image and its complement (average amount of nonfuzziness)

There also exists an algorithm using fuzzy geometric properties, developed by Sankar K. Pal and Azeriel Rosenfeld. To choose the appropriate nonfuzzy threshold, compactness of fuzziness is minimised to obtain the fuzzy and nonfuzzy version of an ill-defined image. [10]

#### 4.1.4 Colocr Package

This R package[11] measures the colocalisation of two proteins inside the cell (using Pearson's correlation coefficient). It automates the regions of interest selection and relies on different algorithms from the imager package. To find the region of interest, structures in a grey-scale image are selected through:

[10] Rosenfeld, A., Pal, S.K., "Image enhancement and thresholding by optimixation of fuzzy compactness". https://www.cb.uu.se/ joakim/course/fuzzy/vt10/CE1/PalRosenfeld.pdf, Published: 1987, Accessed 21 Nov 2021

[11] Ahmed, M., Lai, T.H., Kim, D.R., "colocr: an R package for conducting colocalization analysis on fluresence microscopy images", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6612416/, Published: 4 Jul 2019, Accessed: 21 Now 2021

- Thresholding - excludes the pixels below a certain value (manually set).
- Grow and shrink test - for whether a number of pixels outward and inward belong to the structure.
- Then fill and clean can include and exclude gaps in the structure.

### 4.2 Blank Space

For this project, we use region of interest selection to aid in measuring the colocalisation in a protein (between the green and blue channels). It is important to understand how changes in the image affects the results, these include:

- Removing the black space.
- Cropping the image so that low levels of colour intensities are not included.
- Misaligning two colour channels of the same image.

This section will be analysing the effects of removing blank space.
It is assumed that increasing the blank space in images increases the amount of space both colour channels are not in, therefore increasing colocalization.

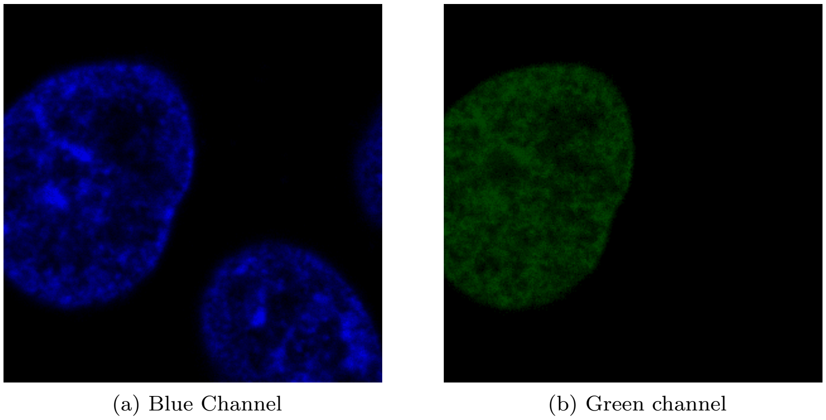
So by applying ROI to an image, we are potentially decreasing the amount of blank space we are measuring colocalisation on and therefore decreasing the Pearson's correlation value.

For example, here is a fluorescent microscopic image of some histones (the two colour channels tag each protein). The original image gives a Pearson's correlation value (PCC) of -0.1271 (4 d.p.). For the image with extra black space added, Pearson's correlation value is -0.0419 (4 d.p.). For these images, we can conclude that increasing the amount of black space in an image increases Pearson's correlation value, therefore increasing the value of colocalisation measured. (code in the appendix)
We are currently trying to mathematically prove this, the progress of the proof is attached in the appendix. (section 8.1)

## 5 Project management

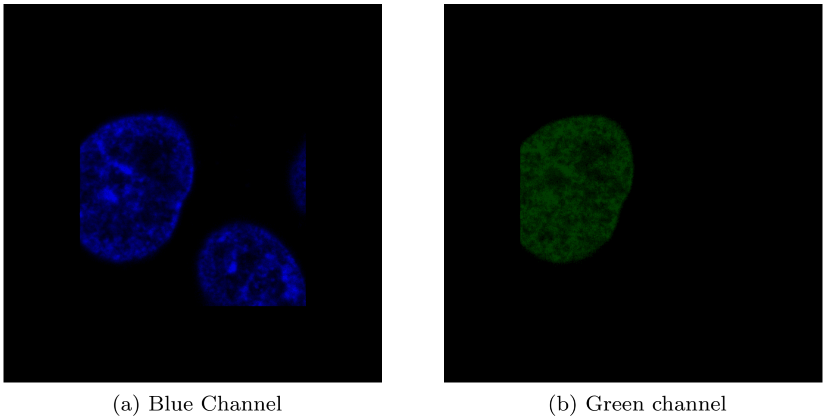
### 5.1 Planning

The project is organised into two sections; researching and programming. These sections are worked on simultaneously to allow more engagement with the studied concepts by applying the methods from the literature to the data.


(a) Blue Channel

(b) Green channel

Figure 1: Original Image


(a) Blue Channel

(b) Green channel

Figure 2: Image with added blank space

A realistic timeline helps manage the workload so that other deadlines could be (and were) met, and will be practical to keep for the rest of the project - with minor adjustments. Organising helped improve workflow and productivity through setting goals in the style of milestones, which were achievable and showed the current state of the project. Milestone one has been completed, and milestone two is currently being completed (refer to appendix).

As part of risk management, tasks were spaced out to include leeway time, in case of unforeseen problems. This allowed for adjustments without falling behind on the project and ended up being useful for problems that did occur ( next section).

## 5.2 Obstacles and adjustments

The original aim of the project was to automate ROI and colocalisation selection. The research revealed various packages to perform these tasks have recently been made.

Instead, the focus of the project was redirected towards analysing these newly created packages. We start by looking at what methods were used to calculate the ROI, followed by analysing how existing packages for ROI have been implemented. This has also resulted in an adjustment of the schedule for term 2 (new Gantt chart in the appendix - section 8.2).

The other adjustment was pushing back the programming task of “implementing new features on how to calculate the ROI” and alternatively testing the already existing packages.

## 5.3 Next steps

We have specialised the project to focus more on the fluorescent imaging of proteins involved in COVID-19. It was noticed that there is the potential use of ROI calculations in imaging plasma cells to detect blood clots automatically. Blood clots from images are usually identified visually, this being automated may help doctors more accurately diagnose whether a patients blood is clotting, and to only prescribe medication when necessary. Alongside this, it would be useful to model COVID-19 cases and their effects, to show the severity of this virus, and how this is affected by blood clotting from previous COVID-19 infection.

The Gantt chart attached in the appendix shows the timeline of what will be done next. So next term will start by implementing seeded growing region and running simulations on the data, followed by the other packages found such as ‘colocr’.

During the researching phase, the proof for blank space will also be continued.

# 6 Challenges

A major challenge is working on a project which is using data relating to a current ongoing problem. The knowledge surrounding resulting blood clots from COVID-19 is still limited and untested. Newly released data and analysis may not yet be replicated and reproduced by other audiences, questioning the reliability and correctness of some information sourced.

An unanticipated obstacle occurred whilst reading articles and researching papers. The research materials contained many terms relating to the medical industry and the analytical techniques used to draw conclusions. Although more time was spent researching than expected, this was beneficial to understanding the resources available and the type of analysis the target audience is interested in.

The mathematical challenge was and is trying to create a proof for the blank space problem stated in section 2.1. The aim is to provide mathematical calculations which can be used to prove the effects of removing blank space (background) to the correlation of an image. This work is still ongoing.

# 7 Ethical consent

The new data set analysed in this project is a collection of images obtained from fluorescent microscopy. Blood samples across patients with COVID-19 were compared against blood samples from patients without COVID-19 and with diabetes.

It does not require any ethical consent to use this data and it is publicly available for use. The data was collected with ethical clearance from the Health Research Ethics Committee (HREC) of Stellenbosch University. Reference: N19/03/043, project ID: 9521. Volunteers had the objectives of the experiment and the risks explained, and informed consent was obtained before blood collection.

# 8 Appendix

## 8.1 Blank Space Proof (unfinished)

The equation for Pearson's Correlation Coefficient is:

$$r_p = \frac{\sum(R_i - \overline{R})(G_i - \overline{G})}{\sqrt{\sum(R_i - \overline{R})^2 \sum(G_i - \overline{G})^2}} \tag{1}$$

We want to see if the same image with less background noise have the same colocalisation.

So we take the sample space $I$, $i \in I$ where $I = (R, G)$.
Let $I_0 = (0, 0)$, and $I_{\setminus 0} = (R, G)\setminus(0, 0)$
$I = I_0 \cap I_{\setminus 0}$

$$\widetilde{r}_{nom} = \sum_{i \in I_{\setminus 0}}^{n} R_i G_i - \overline{R}_{\setminus 0} \sum_{i \in I_{\setminus 0}}^{n} G_i - \overline{G}_{\setminus 0} \sum_{i \in I_{\setminus 0}}^{n} R_i + \sum_{i \in I_{\setminus 0}}^{n} \overline{R}_{\setminus 0}\overline{G}_{\setminus 0} \tag{2}$$

$$r_{nom} = \sum_{i \in I}^{n+1} R_i G_i - \overline{R} \sum_{i \in I}^{n+1} G_i - \overline{G} \sum_{i \in I}^{n+1} R_i + \sum_{i \in I}^{n+1} \overline{RG} \tag{3}$$

And Note:

$$\overline{R} = \frac{n}{n+1}\overline{R}_{\setminus 0} \tag{4}$$

This is the same for G

We can say that these two expressions are equal as the sum of $R_i$ and $G_i$ points will be the same, since the only difference between the two ranges is adding a zero

$$\sum_{i \in I_{\setminus 0}}^{n} R_i G_i = \sum_{i \in I}^{n+1} R_i G_i \tag{5}$$

Likewise for the below, also simplifying the expression for ease of reason

$$\sum_{i \in I_{\setminus 0}}^{n} R_i = \sum_{i \in I}^{n+1} R_i = x \tag{6}$$

And

$$\sum_{i \in I_{\setminus 0}}^{n} G_i = \sum_{i \in I}^{n+1} G_i = y \tag{7}$$

Also a standard formuala:

$$n\overline{R_{\setminus 0}} = \sum_{i \in I_{\setminus 0}}^{n} R_i \tag{8}$$

and same with the green.

So now
$r_{nom}$
$= \sum_{i\in I}^{n+1} R_iG_i - \overline{R}y - \overline{G}x + \sum_{i\in I}^{n+1} \overline{RG}$ (6), (7)
$= \sum_{i\in I}^{n+1} R_iG_i - \overline{R}y - \overline{G}x + (n+1)\overline{RG}$ (summing over constant)
$= \sum_{i\in I}^{n+1} R_iG_i - \frac{n}{n+1}\overline{R}_{\setminus 0}y - \frac{n}{n+1}\overline{G}_{\setminus 0}x + (n+1)\frac{n}{n+1}\overline{R}_{\setminus 0}\frac{n}{n+1}\overline{G}_{\setminus 0}$ (4)
$= \sum_{i\in I}^{n+1} R_iG_i + \frac{n}{n+1}(-\overline{R}_{\setminus 0}y - \overline{G}_{\setminus 0}x + n\overline{R}_{\setminus 0}\overline{G}_{\setminus 0})$
$= \sum_{i\in I}^{n+1} R_iG_i + \frac{n}{n+1}(\widetilde{r}_{nom} - \sum_{i\in I_{\setminus 0}}^{n} R_iG_i)$ (2)
$= \frac{\sum_{i\in I}^{n+1} R_iG_i}{n+1} + \frac{n}{n+1}\widetilde{r}_{nom}$ (5)

Next we have to look at the denominator, we can start with: $\sum(R_i - \overline{R})^2$

For $i \in I_{\setminus 0}$: $\sum^n R_i^2 - 2\overline{R}_{\setminus 0}\sum^n R_i + \sum^n \overline{R}_{\setminus 0}^2$
$= \sum^n R_i^2 - 2\overline{R}_{\setminus 0}x + n\overline{R}_{\setminus 0}^2$ , (4), (6)
$= \sum^n R_i^2 - 2n\overline{R}_{\setminus 0}^2 + n\overline{R}_{\setminus 0}^2$, (6)
$= \sum^n R_i^2 - n\overline{R}_{\setminus 0}^2$

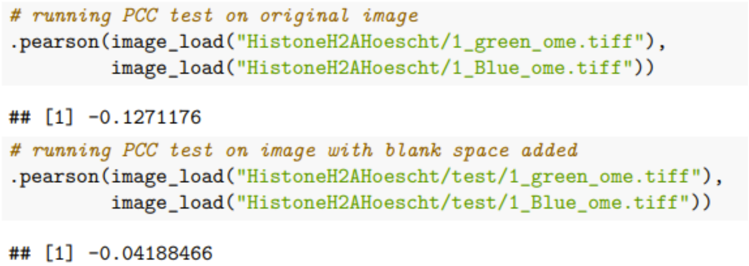
And:
For $i \in I$: $\sum^{n+1} R_i^2 - 2\overline{R}\sum^{n+1} R_i + \sum^{n+1} \overline{R}^2$
$= \sum^{n+1} R_i^2 - 2\frac{n}{n+1}\overline{R}_{\setminus 0}x + (n+1)(\frac{n}{n+1})^2\overline{R}_{\setminus 0}^2$ , (4) ,(6)
$= \sum^{n+1} R_i^2 - \frac{2n^2\overline{R}_{\setminus 0}^2 + n^2\overline{R}_{\setminus 0}^2}{n+1}$
$= \sum^{n+1} R_i^2 - \frac{n^2\overline{R}_{\setminus 0}^2}{n+1}$

Therefore
$\sum_{i\in I}(R_i - \overline{R})^2 = \sum^n R_i^2 + \frac{n}{n+1}(\sum_{i\in I_{\setminus 0}}(R_i - \overline{R})^2 - \sum^n R_i^2)$

$\quad = \frac{\sum_{i\in I} R_i^2}{n+1} + \frac{n}{n+1}\sum_{i\in I_{\setminus 0}}(R_i - \overline{R})^2$
likewise for G.

So far, we have got the mathematical working out to obtain the PCC for the whole sample space in terms of variables without 0 in the sample space. Our next step is trying to compare these values to prove that one is larger than the other (at the very least prove they can be different).

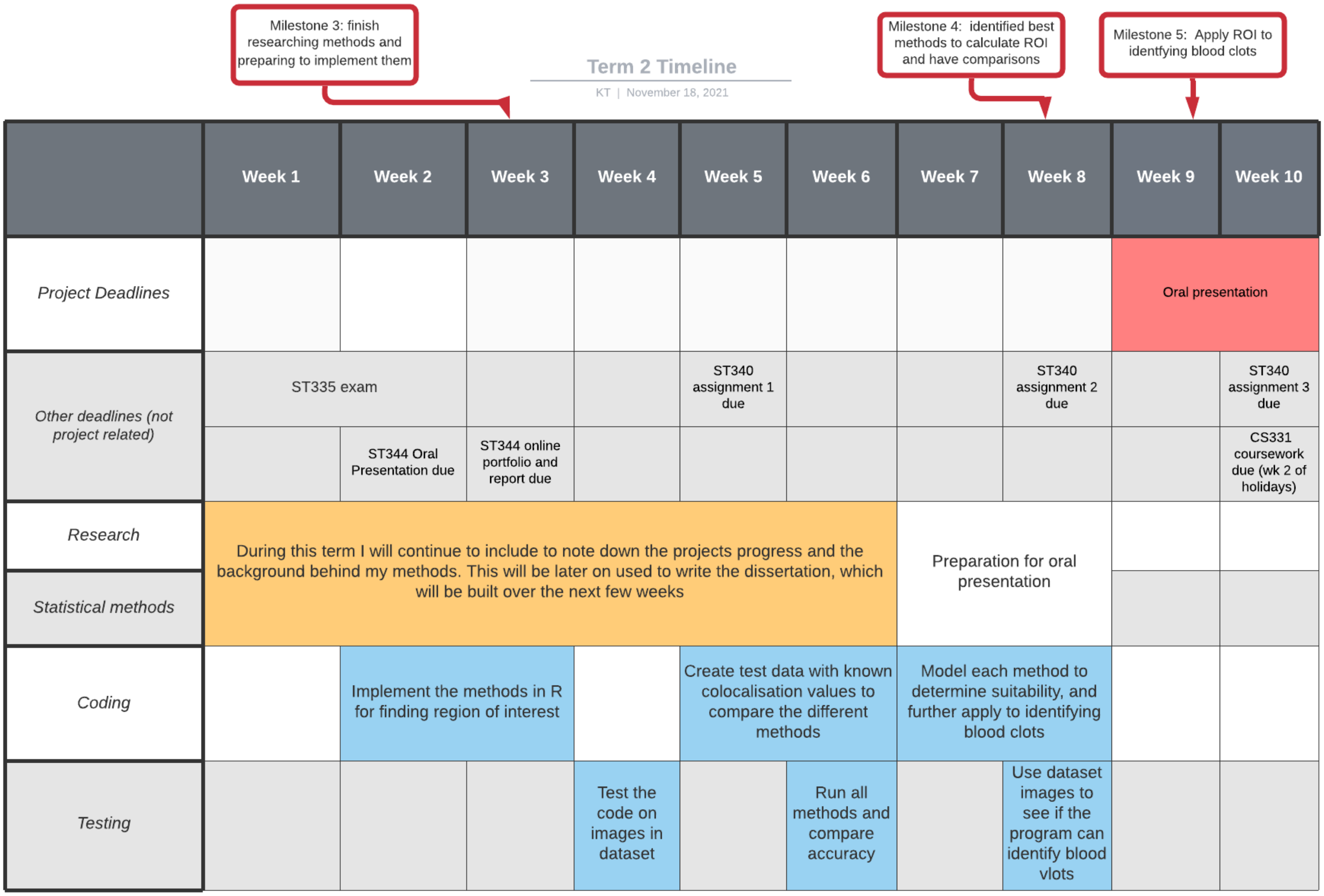
Figure 3: Comparing colocalisation on image values



```
# running PCC test on original image
.pearson(image_load("HistoneH2AHoescht/1_green_ome.tiff"),
         image_load("HistoneH2AHoescht/1_Blue_ome.tiff"))

## [1] -0.1271176

# running PCC test on image with blank space added
.pearson(image_load("HistoneH2AHoescht/test/1_green_ome.tiff"),
         image_load("HistoneH2AHoescht/test/1_Blue_ome.tiff"))

## [1] -0.04188466
```

## 8.2 Gantt Chart

Figure 4: Gantt Chart: Term 2



Milestone 3: finish researching methods and preparing to implement them

Term 2 Timeline

KT | November 18, 2021

Milestone 4: identified best methods to calculate ROI and have comparisons

Milestone 5: Apply ROI to identfying blood clots

| | 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 |
| | | ST344 Oral Presentation due | ST344 online portfolio and report 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 | | | | | | Preparation for oral presentation | | | |
| *Statistical methods* | | | | | | | | | | |
| *Coding* | | Implement the methods in R for finding region of interest | | | Create test data with known colocalisation values to compare the different methods | | Model each method to determine suitability, and further apply to identifying blood clots | | | |
| *Testing* | | | | Test the code on images in dataset | | Run all methods and compare accuracy | | Use dataset images to see if the program can identify blood vlots | | |

## 8.3 Specification

# Analysing images of protein distributions using fluorescent microscopy

CS350 Data Science Project

## Project Specification

Kim Ta 1907156

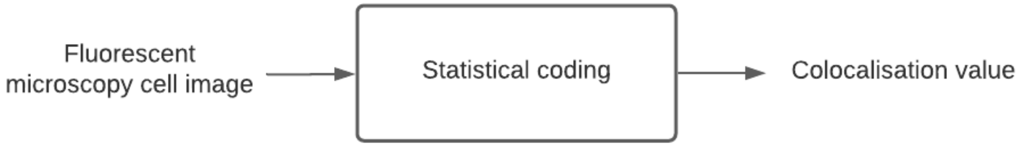
October 2021

# 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].

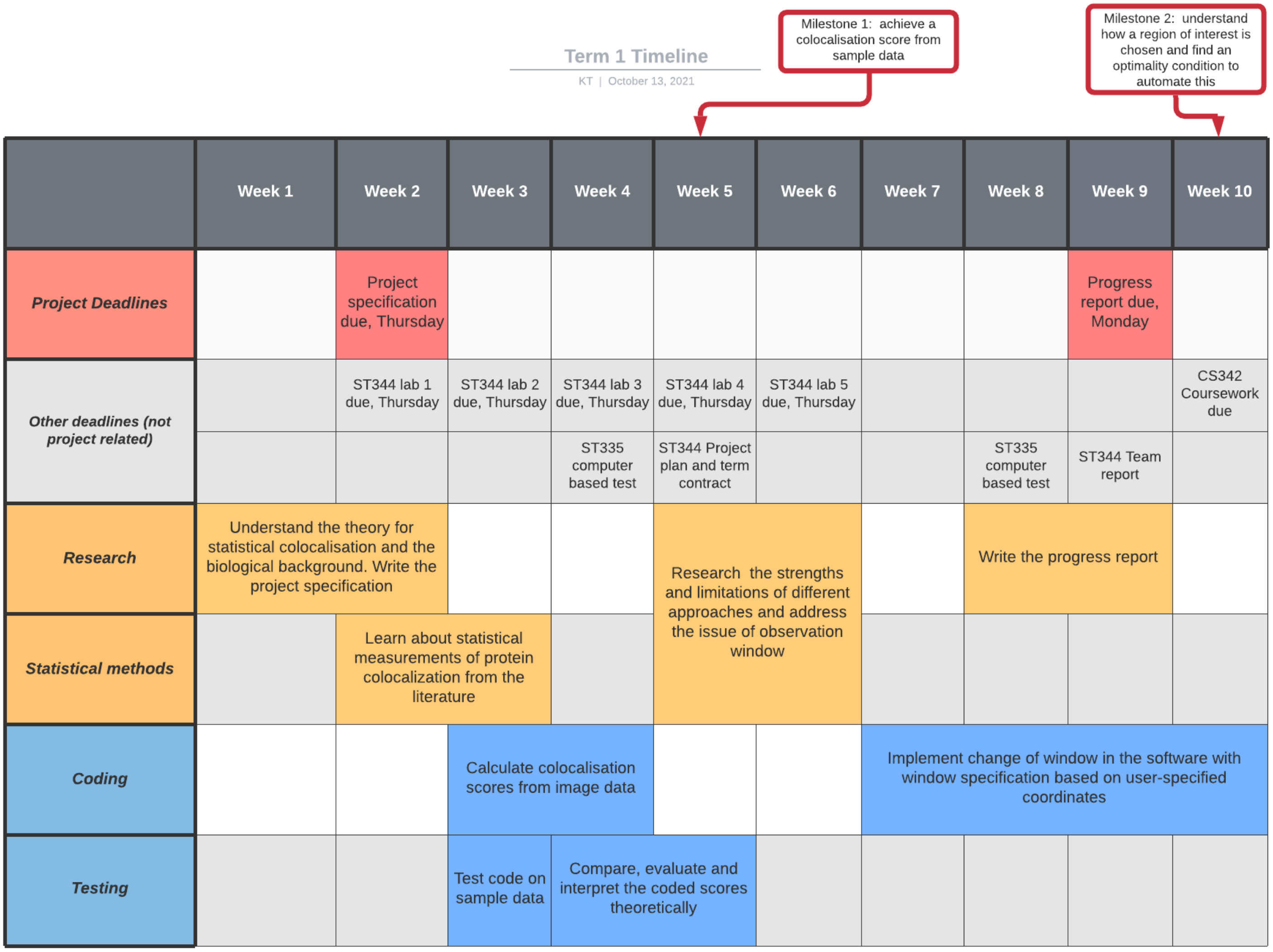
## 8 Legal, social and ethical considerations

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

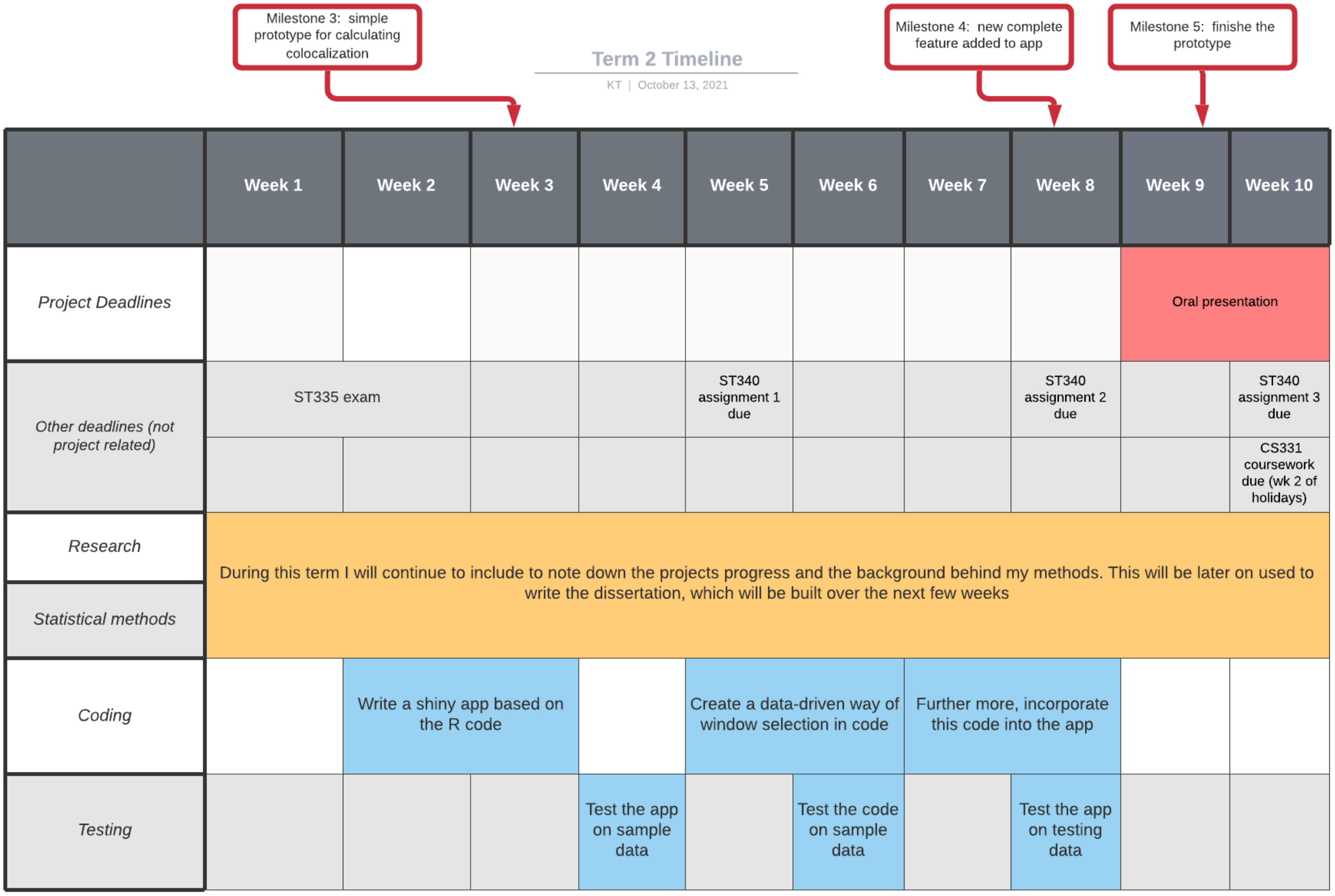
KT | October 13, 2021

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*<br>*Statistical methods* | 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 | | | | | | | | | |
| *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 | | |