Development of an automated quantification algorithm for determining fluorescence distribution in yeast cells (September 1, 2015 to August 31, 2016)

Participants:

Student	Faculty Advisor	Department	University
Wen-Chieh Hsieh Hsieh14@purdue.edu	R. Claudio Aguilar claudio@purdue.edu	Biological Sciences	Purdue
Leqi Liu lliu03@brynmawr.edu	Jia Tao taoj@tcnj.edu	Computer Science	Bryn Mawr college

Details of dates and attendance of team meetings:

S.No	Dates	Venue	Participants
1	Dec 8 2015	Purdue University, IN	Wen-Chieh Hsieh Leqi Liu (online)
2	Dec 18 – 20 2015	Bryn Mawr college, PA	Wen-Chieh Hsieh Leqi Liu

In addition, the students virtually meet bi-weekly during the school year and weekly during the summer.

Details of money spent:

Grant awarded: $2000

Money spent: $ 813.81

Money remaining: $ 1186.19

Presentation:

Wen-Chieh Hsieh, Leqi Liu, Jia Tao and Claudio Aguilar. Development of an automatic quantification algorithm for determining fluorescence distribution in yeast cells. (2015, NSF site visit)

Project Introduction

The initial step of creating effective therapeutic approaches is to understand the disease causing mechanism in detail, which in many cases involves acquiring and analyzing massive amounts of biological data. However, conventional human-involved analysis methods present serious disadvantages when processing high-information content image data, such as user-introduced error/biases and low throughput. For instance, examining protein subcellular localization typically requires quantifying hundreds of cells per experiment through visual examining and manually outlining areas from images collected by microscopy. The procedure is not only laborious but also imprecise that would restrict the efficiency of data analysis process, and more importantly, the accuracy of the interpretation after the data analysis.

The objective of this student-initiated proposal is ***to develop an automated quantification algorithm to extract and analyze quantitative biological information related to endocytosis of membrane proteins*** *to facilitate data analysis and reduce user-introduced errors.*

Endocytosis is one of the key processes for regulating signaling pathways that are crucial for maintaining normal cellular functions and tissue development. For example, internalization of DLL1 protein is the triggering step of the Notch signaling pathway that when misregulated causes multiple developmental diseases and cancer. Therefore, understanding the regulatory mechanism of membrane protein internalization is a necessary step prior to discovering therapeutic approaches to those diseases. Budding yeast is a powerful model system to study endocytosis. Here we use a green-fluorescent-protein (GFP) tagged membrane protein to determine the level of endocytosis via quantifying the cellular distribution of fluorescence signal collected through fluorescence microscopy.

Working strategy:

Stage 1*:* Isolate single cells from microscopic images with multiple cells

Yeasts are unicellular model organisms that are routinely used to study distribution of biological products, such as proteins. In a microscopic image, there could be multiple yeast cells (Figure 1).

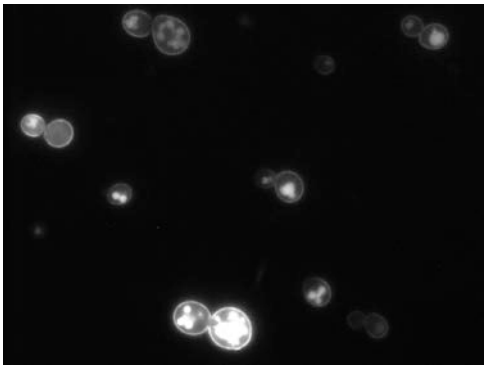

Figure 1: Microscopic image with multiple yeast cells.

For each cell, the distribution of protein can be different. To precisely obtain biological data from each individual cell, identify the cell boundaries from the background, and further isolate each cell through acquiring the coordinates from the original image. We use “R” as our programming language to implement the algorithm for image analysis.

Based on the pixel intensity of each image, we apply a customized serial number of thresholds to the image, and generate a set of filtered images. Then, an ideal filtered image will be automatically selected, which most

of the background has been filtered out. Based on the pattern we found, we set rules for the selection:

1. The filtered images are grouped into sequence sets of images that have the same amount of pixel clusters and each pixel cluster of the set of the images corresponds to the same cell component.

2. The first image from the set which includes the largest number of filtered images will be the ideal cut off image.

The resulting image selectively preserves the signal from the cells.

In such an image, adjacent pixels within a certain area are clustered into individual components. Each component corresponds to a single cell in the original image. However, some of the components have larger area than the actual cells, because the pixels at the halo of cells are included, especially when the cells are bright (Figure 2). Therefore, we introduce few steps to identify the actual boundary from the crude cells.

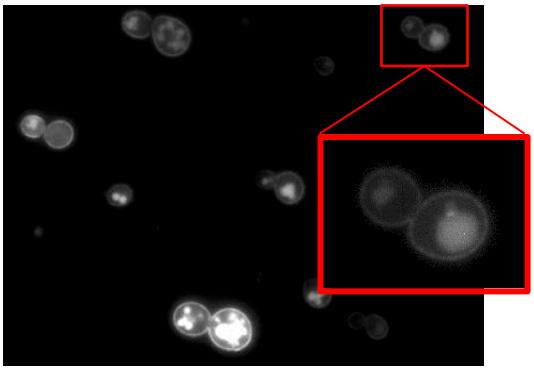

Figure 2: Isolate single cells from the cut-off image.

The strategy is to filter out the halo by applying a set of thresholds to the crude cells. After obtaining a series of filtered images, the total intensity of each image is calculated. The values of the intensities are plotted into a curve corresponding to their serial numbers. Since the pixel intensities of the dots in the halo are within a small range, they would be filtered out within a range of thresholds. The ideal thresholds could be identified by observing the inflection points of the curves. To manifest the inflection points, we convert the curve into peaks by plotting the fourth derivative of the curve.

Throughout the process, the coordinates and the pixel intensities of each cell can be retrieved from the original image. Therefore, the accurate total pixel intensity of each cell can be collected (figure 3).

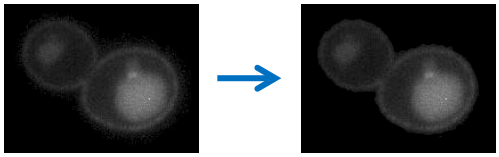

Figure 3: Precisely define cell boundary and total intensity.

***Stage 2*:** Analyze target protein distribution in each single cell

Since proteins fulfill their functions at specific locations, tracking their distribution is an important way to study their activities. For example, in our case, the target membrane protein internalizes when the protein is not in the action. Therefore, precisely quantify the fraction of protein being internalized could help us to understand the level of activity.

Therefore, we separate the source of signal into two parts: cell surface (*e.g.* cell membrane) and intracellular part. Since the area of the intracellular part of a cell is proportional to the area of the whole cell, we “shrink” the cell by a factor equal to the

average membrane thickness to remove the area of cell surface. Therefore, we can obtain the total pixel intensity of the intracellular part by summing up all the pixel values (Figure 4).

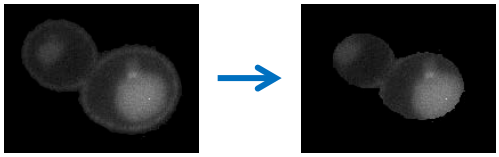

Figure 4: Remove cell membrane and define internal intensity.

Due to auto-fluorescence and the GFP on the target protein illuminating the surrounding area, some background signals exist in the image. Considering these background signals may lead to over-estimate the protein distribution within the cell, we need to filter out the backgrounds. Currently, we are plotting the histograms of the pixel intensity values within the cells to define the backgrounds and we are polishing the steps. After obtaining the background signals of each cell, we can extract the intracellular signal contribution of the target protein (Figure 5).

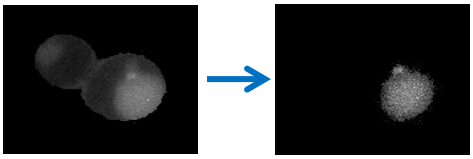

Figure 5: Define intracellular component and background intensity.

The obtained pixel intensities (total cell, intracellular part, background) can fit to an equation that calculates the distribution of the target protein with an equation:

$$Intracelluar\ localization\ = \frac{Intracellular - Background}{Total - Background}$$

Our algorithm enables us to systematically analyze sets of microscopy images in an automatic manner, that would not only increase the accuracy of the results and also reduce human operation time.

Future direction: In the future, after fully develop the algorithm, we would like to build graphical user interface of our program so that more users can easily use it. Meanwhile, we are also considering using machine learning to future improve the accuracy of identifying the cell membrane and background signal of a given image.