Undergraduate Researcher: Daniel Puleri

Graduate Mentor: Charles Zhao

Faculty Advisor: Prof. Hang Lu

Affiliation: Georgia Institute of Technology

Topical Area: Food, Pharmaceutical, and Biotechnology

# Image alignment and segmentation algorithms for image-based screens of subtle features using *C. elegans*

## Introduction

The quantitative phenotyping of subtle traits is crucial to the genetic mapping of currently unexplored regions of the genome. By using high-throughput microfluidics and automated image processing, we show that we can acquire phenotypic features rapidly and quantitatively from a model system. With high-volume, high-information feature data, one can statistically map genomic regions with phenotypes; alternatively, one can search for genes of interest in a forward genetic screen. Forward genetics typically uses mutagenization to create an abundance of phenotypes and then determines which genetic mutations correlate with a change in phenotype by isolating the differing phenotypes and comparing changes in DNA. Determining the effect on one particular gene by comparing populations with different genetic backgrounds could shed light on epistatic effects—an important topic in modern biology. Finding which genes affect phenotypes, such as those for diseases, can then be used for future downstream medical research with more direct clinical applicability.

When studying small structures such as neurons, microscopy is a vital tool for observing phenotypes. The images that one gathers from conventional fluorescent microscopy are typically information-rich; however, the relevant data is obfuscated by noise and false positives.

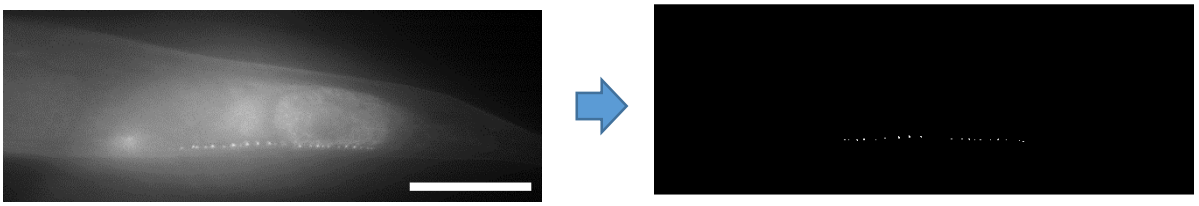


**Figure 1. The synapse detection algorithm** takes in a maximum projection of a set of images in the z-plane (left) and generates an estimate of the pixels in the image that are part of a synapse (right). Scale bar approximately 30μm.

Using automated image processing, one is able to phenotype large populations of organisms characterized via microscopy in an impartial and high throughput manner that allows human intervention to be minimized. I have used a combination of statistical and heuristic techniques to determine which features of an image are important enough for extraction. For example, we characterize features that would be hard for a human to determine, such as total synaptic intensity and major axis length—these quantitative phenotypes allow us to make more specific claims about genetic effects, when compared with more qualitative observations.

For our studies of the synaptic domain we use *Caenorhabditis elegans* (*C. elegans*), a soil nematode that is widely used in biology due to the breadth of knowledge about the species and its ease of handling. Using traditional techniques to handle and image *C. elegans* is both protracted and cumbersome. The Lu Lab uses microfluidic devices to augment the handling of *C. elegans* and enable high throughput imaging and sorting of worms on-chip[1]. By using microfluidic devices, we are able to gather the large amounts of data needed to perform statistical mapping of genes.

[1] Chung, K., Crane, M. M. & Lu, H. Automated on-chip rapid microscopy, phenotyping and sorting of C. elegans. *Nat. Methods* **5,** 637–643 (2008).

More specifically, we study the synaptic morphology of *C. elegans* which is visible due to the use of an integrated fluorescent protein (*Pmig-13::snb-1::yfp*) that localizes to the synapse. Synaptic morphology is an important trait that has been linked to diseases ranging from schizophrenia to autism spectrum disorder (ASD)[2][3].

**Background**

*Caenorhabditis elegans* is a non-parasitic nematode that has been used as a model organism in the laboratory since the 1960s. The organism is about 1mm in length and is transparent. *C. elegans* is also one of the only organisms to have its connectome fully mapped—a description of how its neurons are physically laid out.

The model organism *C. elegans* is well-suited to study the subtle phenotypes of multigenic effects such as in synaptic morphology—which may play an important role in neuronal disease such as autism. Current understanding of synaptic morphology and aging, given the large amount of quantitative data needed to characterize these traits, is relatively incomplete even in the oft-studied model organism *C. elegans*. The need to generate mutants, rapidly characterize worms, and sort populations using traditional methods is tedious and very labor intensive. Using a novel high-throughput microfluidic imaging platform developed in the Lu Lab, these concerns can be mitigated and automated quantitative characterizations of synaptic domains can be performed.

Overall, *C. elegans* is easy to work with due to its rapid mating and transparency under the microscope, which makes it ideal for usage in high magnification experiments. It is easy to house, has a rapid reproductive cycle, and can lay as many as 300 eggs per mother.

**Methods/Techniques used**

The microfluidic device that is being used was designed by Dr. Adrianna San Miguel for the imaging of *C. elegans.* The device allows for the high volume imaging of worms in a “drive-through”-like process where worms are pushed up to the imaging channel and then directed through one of two exits based on phenotype. The camera is aimed to image worms in a tail-first orientation, while discarding worms in the incorrect orientation.

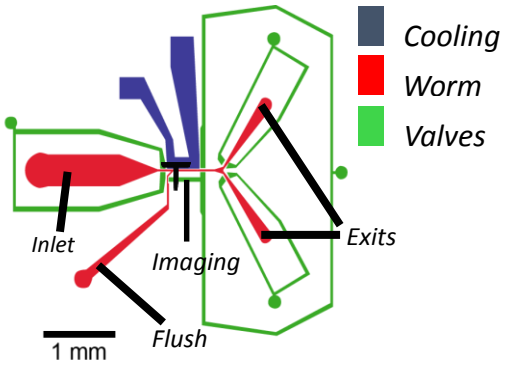



**Figure 2. A diagram of the microfluidic device used for imaging and sorting.**

The microfluidic devices are fabricated by curing polydimethylsiloxane (PDMS) over a silicon or polyurethane master. The features of the microfluidic devices arise from the detailed height variations of the silicon master. The devices are bored into to create the holes for the inlet, outlet, and valve tubes. Afterwards, the device is plasma bonded to a glass slide which acts as the transparent floor which rests upon the microscope objective.

Worms are imaged at regular points in their lifecycle using an upright fluorescent microscope at 40x.The images are then captured and analyzed in MATLAB®. To distinguish the synapses from background noise, local background removal and thresholding identify synapses and other bright objects. Synapses are then identified from other points of high intensity, mostly autofluorescent lipids, by using DBSCAN, a clustering algorithm. The clusters identified are then checked for their angle to only allow pixel groups in the horizontal dimension. Clusters in the same vicinity are then connected so that the longer axon can be

[2] Shinoda, Y., Sadakata, T. & Furuichi, T. Animal models of autism spectrum disorder (ASD): a synaptic-level approach to autistic-like behavior in mice. *Exp. Anim. Jpn. Assoc. Lab. Anim. Sci.* **62,** 71–78 (2013).

[3] Calahorro, F. & Ruiz-Rubio, M. Caenorhabditis elegans as an experimental tool for the study of complex neurological diseases: Parkinson’s disease, Alzheimer’s disease and autism spectrum disorder. *Invertebr. Neurosci. IN* **11,** 73–83 (2011).

reconstructed. The features of the worms are extracted and compared to other populations based on a feature set that characterizes, among other things: synapse size, shape, intensity, and domain length.

Due to the length of the worms, some images had to be taken off-axis to facilitate the entire worm in the image. Using a linear regression, the slope of the worm was determined and then it was aligned onto the horizontal axis. The centroid was used to center the worm so that all other processing could be done with a regular reference frame. Post-processing is then performed on the worm using robust principal component analysis (RPCA)[4].

**Results**

Linear regression to straighten images, regression to determine cluster linearity, and vertical line detection are some of the many heuristics we employ to “clean up” images so that they are more amenable to further processing and analysis. Moreover, I implemented an additional heuristic that connects separated clusters of synapses so that the synapses, if they are in a certain vicinity, can be reconnected to form the complete chain of synapses. Seemingly simple heuristics like this need to be used to correct slight defects that are not picked up by major pieces of our algorithm.

When analyzing images and comparing their 29 features to sort worms into respective populations, we quickly ran into a multiple comparisons problem. In addition, there are underlying issues such as inherent relations between image features that share similar properties like synaptic domain length and number of synapses. To resolve these issues, we propose the use of principal component analysis (PCA). Comparing strains by their strongest component, the component with the eigenvalue of the largest magnitude, we can more successfully segment strains into separate populations. However, this comes with some drawbacks. The PCA is sensitive to large data sets that may contain correspondingly large amounts of noise. To combat this, we have implemented a newer technique called robust principle component analysis (RPCA). Using this technique in a novel setting, we are able to mitigate the aforementioned issues and I was able to optimize it for use with our large datasets which helps yield which features are most important in distinguishing populations. This reduced the memory load from several hundred gigabytes to less than four—a number that is manageable by a consumer laptop.

We were then able to uncover that the genetic background of the worm does affect key features of synaptic morphology. Laboratory adapted wild type strains from different original natural environments showed differing effects on the particular phenotypes. This shows that synaptic morphology is not only controlled by specific alleles; but, is also affected, albeit subtly, by the organism’s underlying genetic background.

[4] Cand, E. J., et al. (2011). "Robust principal component analysis?" J. ACM 58(3): 1-37.