Image Analysis for Quantification of mRNA Levels

For this project, you will be tasked with developing an image processing workflow in MATLAB to quantify levels of mRNA in individual cells and compare the results to theoretical modeling based on the framework learned in our lab on stochastic gene expression modeling. smFISH is an analytical method that uses fluorescently labels individual RNA molecules such that the number can be determined from fluorescent signaling. Figure 1 contains an example output of a smFISH experiment.

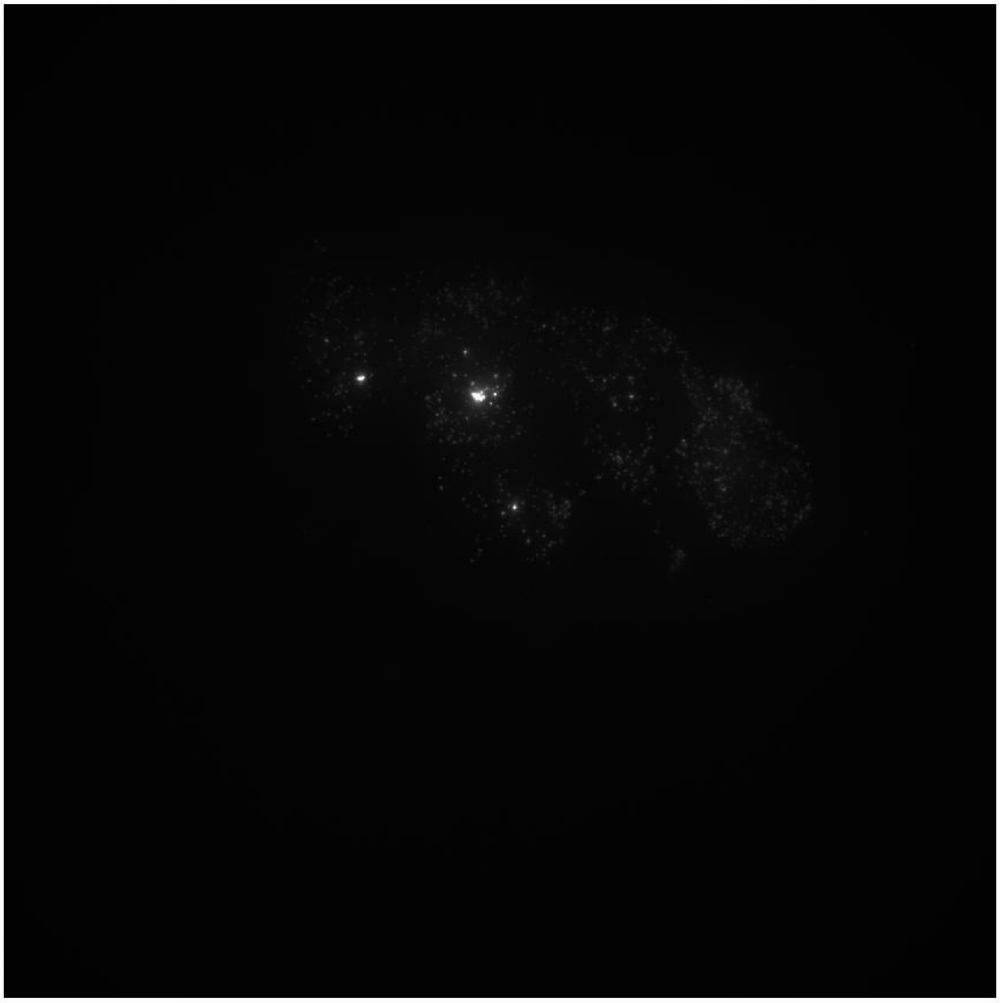

Figure 1: Example image from a single smFISH experiment with contrast adjusted.

In Figure 1, we can see that mRNA molecules are fluorescently labeled and are visualized as the bright spots in the image. The image above shows a group of cells being analyzed, so we also need to use data from a nuclear staining (DAPI) experiment to determine where the individual cells are located. Figure 2 shows the corresponding nuclear staining image to the smFISH experiment in Figure 1.

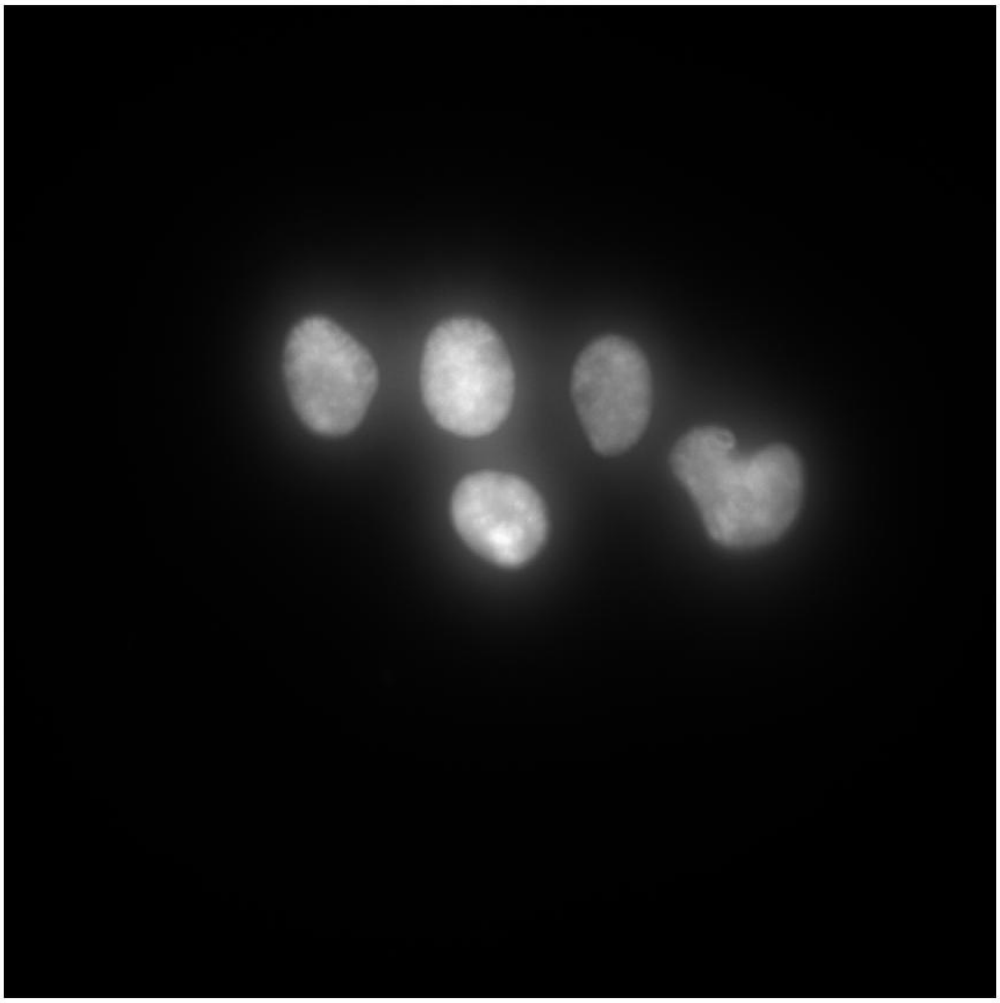

Figure 2: Nuclear staining image corresponding to smFISH experiment in Figure 1 with contrast adjusted.

The data from these two experiments can be combined to determine the number of mRNA molecules present in each nucleus. From here we can obtain the distribution of mRNA molecules in each cell, and then compare our results to a model of stochastic gene expression to determine the parameters relating to transcriptional bursting. For this project, prepare a detailed report containing the following sections:

1. Introduction: Briefly describe the project and the ultimate goals of this investigation.
2. Methods: Describe the methods used in the image processing code developed and the theoretical modeling of stochastic gene expression used. Images of intermediate results, e.g. segmentation can be used here to illustrate your pipeline.
3. Results: Just include the final results here, i.e. the distribution of mRNA levels per cell, the predicted on and off rate for transcriptional bursting, and the results of model simulation.
4. Discussion: Discuss your results commenting on how the theoretical predictions compare to the experimental data and assess the robustness of your image processing pipeline.

5. Conclusions: Briefly state the important conclusions from your investigation.

Dataset

On canvas there is a zip file called image_processing_data that contains DAPI and smFISH images from six different experiments. Each pair of images contains multiple cells, which you will need to isolate and process individually. The images are labeled by which go together, i.e. the images for experiment A are labeled DAPI_A and mRNA_A. Of the six pairs of images, you should start with experiments A-C as they will be more straightforward than the other three. Experiments D-F are more difficult due to cell crowing, so once you have developed an algorithm for processing A-C attempt to modify this algorithm to also work for experiments D-F.

Image Processing

To quantify the amount of mRNA molecules in the nucleus of each cell, you will need to use MATLAB's image processing toolbox to analyze the DAPI staining and smFISH images for each group of cells. The mRNA molecules in the smFISH images can be in two different states: free mRNA where transcription is complete (molecules diffuse in the nucleus and can exit to cytoplasm) and nascent mRNA where they are still being transcribed. You will be using this images to quantify the total amount of mRNA within each nucleus and also determine the amounts of free and nascent mRNA. The amount of nascent mRNA can be thought to be proportional a current transcription rate from the gene.

You will need to complete the following steps (before completing these steps, use MATLAB's `im2double()` command to convert the images to the double format). Note that his is just a suggestion you are welcome to explore alternative ways to achieve the same goals as well as other iamge processing (e.g. filtering etc):

1. Using the DAPI staining images, create a mask that can be used to determine the locations of each nucleus. Here you will want to use MATLAB tools to reduce background noise in the image, binarize the image such that the nuclei are separate from the background, and apply a watershed transformation to separate any overlapping/touching nuclei. Useful function here are `imbinarize()`, `adaptthresh()`, `medfilt2()`, `imfill()`, `bwdist()`, `imextendedmin()`, `imimposemin()`, and `watershed()`. The demo in the beginning of the first lab period will help you with the watershed algorithm. Success in this step will results in a mask that can isolate the fluorescence from the smFISH images to the nucleus and a list of each nucleus from the watershed results.
2. Iterate through each nucleus and count the number of bright spots (mRNA molecules) by repeating the following steps on each cell identified in the previous step:
    a. Determine which of the identified objects from part 1 correspond to cell nuclei and which are random objects based on size and/or shape (e.g. random objects will be much smaller). Use `regionprops()` to get variety of object characteristic. Perhaps you can spot outliers by looking at the distribution of some of these.
    b. For each identified cell nucleus, apply a mask that will isolate a single cell nucleus for further analysis. See `bwlabel()` command
    c. On the isolated nucleus apply a mask that will isolate the mRNA signal from the background. Useful functions here are `multithresh()` and `imbinarize()`.
    d. Record the number, pixel area, and mean intensity for all the bright spots within the nucleus. Here, you will want to use the `regionprops()` function with the name value

arguments 'Area' and 'MeanIntensity'. Calculate the total intensity of each spot my taking a product of the area and average pixel intensity.

3. The florescence signal is in arbitrary units. To convert this value to mRNA counts we need to determine the value of total intensity that corresponds to a single mRNA molecule. Since not all of the spots we see are single mRNA you expect that some of them will contain2, 3 etc mRNA and, therefore, will be about 2x or 3x more in intensity. Plot the distribution of all the total intensity values for the cells analyzed in previous steps, focusing on the smaller values in the histogram (<= 0.1). You should see several distinct peaks that correspond to one, two, three, etc. mRNA molecule. Hint: you may need to adjust the number of histogram bins.
4. Using the determined intensity for a single mRNA molecule, determine the number of mRNA molecules in each cell.
5. If a nucleus has a single (or perhaps two if DNA replication has occurred), extremely bright spot(s), this corresponds to the site of transcription and the mRNA molecules located here are nascent mRNA. Use the intensity corresponding to a single mRNA molecule you determined in step 3 to determine how many mRNA molecules are located at transcription sites.

The final output of your image processing pipeline for a single pair of images should include: 1) the number of cells in the image, 2) the number of free mRNA molecules in the image, and 3) the amount of mRNA located at the site of transcription.

Statistics

To compare the experimental results to theoretical predictions, you will need to calculate some statistical measurements from your data. Calculate the mean, variance, and CV for the number of total mRNA molecules (free and nascent mRNA) among all cells in the field of view. Additionally, determine the correlation coefficient between free mRNA molecules as a function of nascent mRNA molecules by using matlab `corr` command. Show this data as a scatter plot.

Modeling of Stochastic Gene Expression

The final step of this project is to modify the model you used in Lab 5 to interpret the results of your image analysis and predict transcriptional bursting parameters. Your model will consist of the following reactions.

$$P_0 \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} P_1$$

$$P_1 \xrightarrow{k_{tsc}} P_1 + mRNA_n$$

$$mRNA_n \xrightarrow{k_{release}} mRNA_f$$

$$mRNA_f \xrightarrow{\gamma_1} \emptyset$$

Here, P_0 is the promoter in the off state, P_1 is the promoter in the on state, $mRNA_n$ is nascent mRNA (still being transcribed), and $mRNA_f$ is the free mRNA. Your goal is to use this model to determine k_{on}, k_{off}, and the release rate $k_{release}$ that gives you the distribution of nascent to free mRNA you obtained from your image analysis pipeline, as well as predict the distribution of protein given your estimated release

rate. To do this, assume that all other parameters are known: $\gamma_1 = 0.05\ minutes^{-1}, k_{tsc} = 21\ minutes^{-1}$. Determine the release rate, $k_{release}$ based on the ratio of nascent and free mRNA recovered from your image analysis, $\frac{\overline{mRNA_f}}{\overline{mRNA_n}} = \frac{k_{release}}{\gamma_1}$. Now with the determined release rate, the mean mRNA production rate can be determined from the mean nascent mRNA concentration using the relationship $\overline{mRNA_n} = \frac{\overline{k_{prod}}}{k_{release}}$. Use the experimentally determine value of $\overline{mRNA_n}$ and your estimated value of $k_{release}$ to determine the ratio of $k_{on}:k_{off}$ using the expression for average production rate $\overline{k_{prod}} = k_{tsc}\frac{k_{on}}{k_{on}+k_{off}}$. With the determined ratio, you can determine the values of k_{on} and k_{off} by manipulatin the value of k_{on} to match the observed variance. To do this, pick an arbitrary value of k_{on}, calculate k_{off} based on the ratio, and determine the predicted variance in nascent mRNA by running an ensemble of 1000 iterations. Iterate until you get a k_{on} value that yields similar variance to what was observed experimentally. Once you have determined these unknown parameters, build a model of this system in SimBiology and use the model analyzer GUI or the sbiosteadystate() function to determine values for each species. Complete 1000 iterations of the model using the ensemble section of the GUI or the sbioensemblerun() command with the determined values for k_{on}, k_{off}, and $k_{release}$. Calculate the predicted mean, variance, and CV of total mRNA and compare to the experimental results.