Archit Chabbi, Robert Heeter
BIOE 446 Computational Modeling Lab
12 December 2023

Image Analysis for Quantification of mRNA Levels

Introduction

This project looks at applying image processing methods with MATLAB including filtering, binarization, and a watershed algorithm for segmentation with the goal of quantifying levels of fluorescent mRNA molecules in individual cells. Images of fluorescent RNA molecules were obtained using the smFISH method, while images of cell nuclei were obtained using nuclear staining (DAPI). The results from this processing pipeline are compared to a model of stochastic gene expression for transcriptional bursting using MATLAB's SimBiology library.

Methods

Image Processing

The image processing pipeline consists of the following steps: isolating the cell nuclei from the DAPI image, isolating the mRNA in each nucleus from the corresponding smFISH image, and calculating the total number of free and nascent mRNA molecules in each cell.

Isolating Cell Nuclei

The DAPI image from the experiment is used to identify the locations of the cell nuclei in the image. Firstly, the image is converted into a matrix where the matrix values represent the intensities at each pixel. Then, the background noise in the image is reduced using median filtering. Adaptive thresholding is then applied to binarize the image, after which any holes within the image are filled (**Figure 1**).

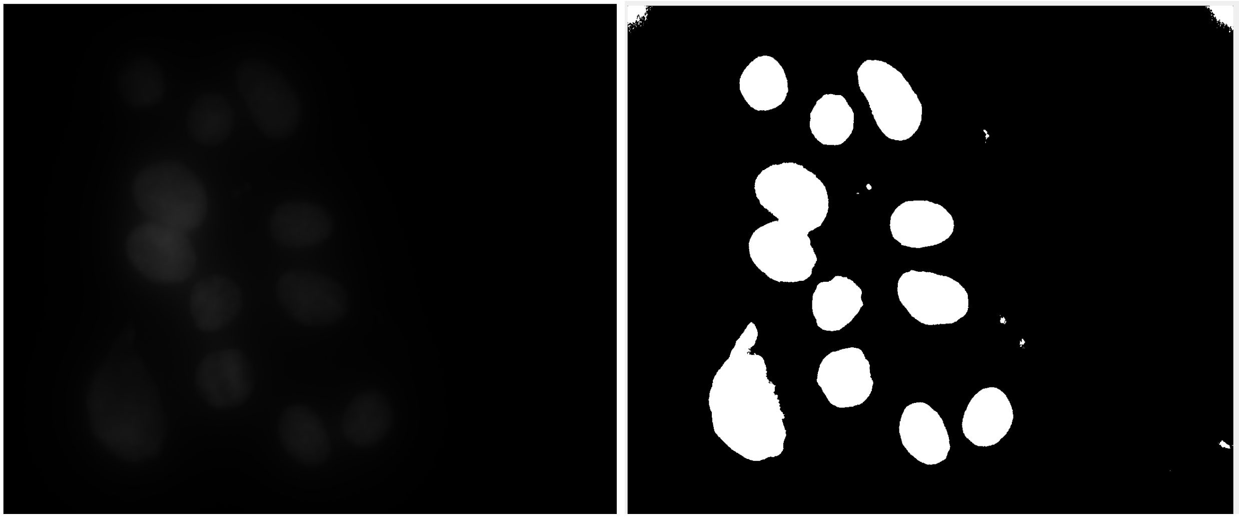


Figure 1. *Left*: original DAPI image. *Right*: DAPI image after background noise reduction, thresholding, binarization, and hole filling.

A distance transform is then applied to this binary image to aid in distinguishing the actual cells from the background. Local minima in the image are filtered out before applying a watershed to segment the image and obtain the cell boundaries (**Figure 2**).

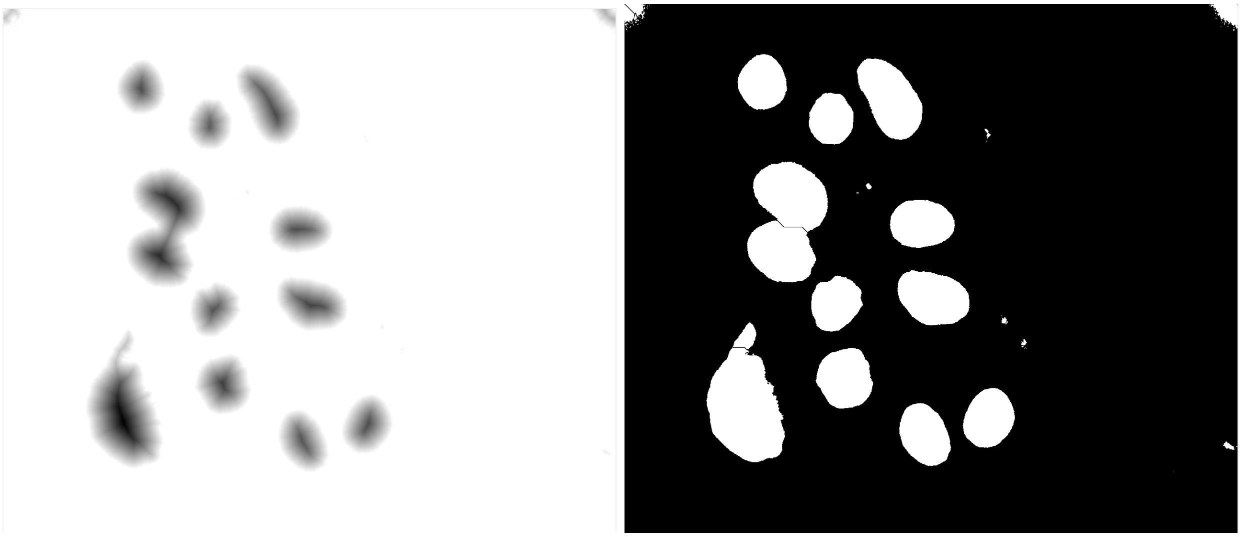


Figure 2. *Left*: image obtained after applying distance transformation to binary image. *Right*: final segmented image with cell boundaries highlighted.

After applying the watershed algorithm, the final image is separated into distinct objects. However, all objects in the image are selected, including the small white dots as well. To effectively isolate the cell nuclei, the areas of each of these objects are checked. As the individual cells will always have a much greater area than the other noise present, only objects with areas that exceed a certain threshold are identified as cells. Once each cell nucleus is identified, a mask is created for each nucleus that can be applied in the subsequent steps to isolate mRNA visible within the nuclear region.

Isolating mRNA in Nuclei

To isolate the mRNA for a particular nucleus, the original smFISH is preprocessed. Using multilevel thresholding, a threshold is obtained that is applied to binarize the smFISH image (**Figure 3**). Each nuclear mask created from the DAPI image is applied to the binarized image to isolate the corresponding region containing mRNA (**Figure 4**).

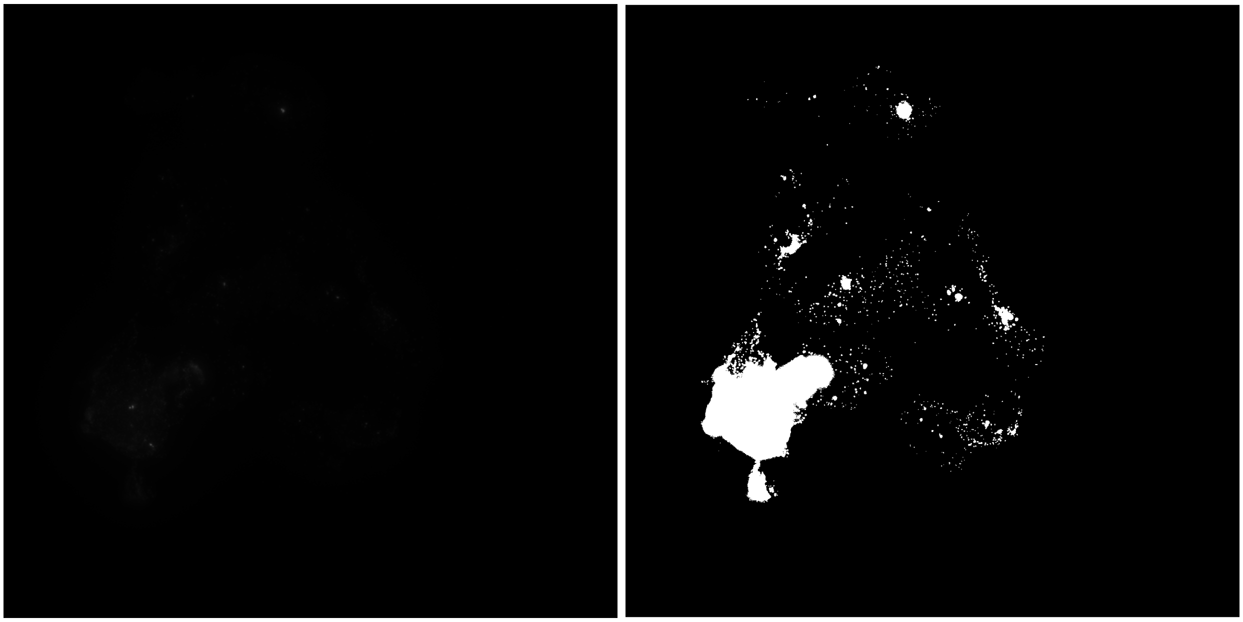


Figure 3. *Left*: original smFISH image. *Right*: binary image with white spots corresponding to mRNA molecules.

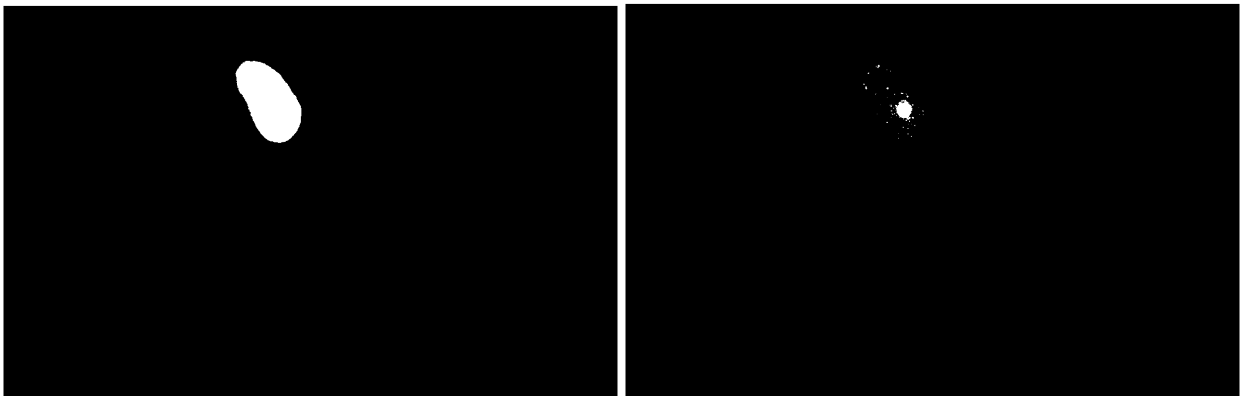


Figure 4. *Left*: isolated nucleus from segmented DAPI image. *Right*: corresponding region in binary smFISH image after applying the respective nuclear mask.

Calculating Total and Nascent mRNA Molecules

The areas of each individual bright spot in the binary smFISH image are then multiplied by the corresponding pixel intensity of the spot, measured by applying the nuclear mask to the original smFISH image. This is done to obtain the total intensity for the spot as a proxy for mRNA levels. To determine the intensity value for a single mRNA molecule, the distribution of all of the total intensities for all spots in all of the cells in the image is visualized. As shown in **Figure 5** for this particular example, there are distinct peaks at each separate intensity level less than 0.1, with the peak at the smallest intensity corresponding to that of a single mRNA molecule. The bin for this peak ranges from 0.017 to 0.02, so a single mRNA molecule is estimated to have an intensity of 0.02 within this image.

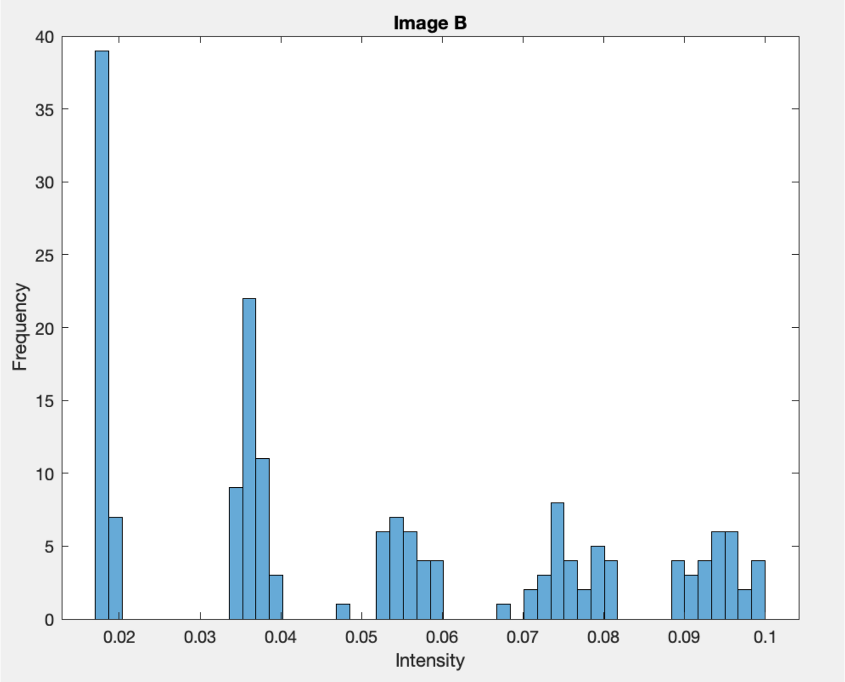
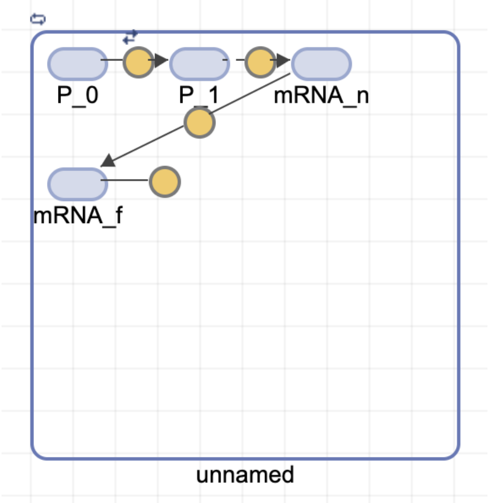


Figure 5. Distribution of mRNA intensities across all cells for image B.

To calculate the total number of mRNA molecules in each cell, the total intensities of all spots in each cell are added together and divided by the estimated intensity for a single mRNA molecule. To determine the number of mRNA molecules located at the transcription site within each nucleus, the total intensities are first examined. If any one or two spots have a much greater intensity than the remaining (exceeding a determined threshold), then those two spots are considered to be transcription sites. The number of mRNA molecules located at these sites is then calculated by dividing the total intensity at the sites by the intensity of a single mRNA molecule (estimated previously). The number of free mRNA molecules is calculated by subtracting the transcriptional mRNA from the total mRNA in the image. The final outputs of the image processing pipeline for a pair of DAPI/smFISH images are the total number of detected nuclei, total number of free mRNA molecules in the image, and the total number of mRNA molecules located at transcription sites in the image.

Modeling of Stochastic Gene Expression

To create a stochastic model of transcriptional bursting following the reactions in **Figure 6**, MATLAB's SimBiology library was used. A degradation rate of 0.05/min and transcription rate of 21/min were assumed, and the values of the release rate and production rate were calculated from the approximate ratio of free to nascent mRNA molecules as found from image analysis (see above and below). The values for the forward and reverse reaction rates were estimated using the variance in nascent mRNA from image analysis as well; a 1000-run stochastic ensemble simulation was performed to find forward and reverse reaction rates that would produce a similar variance to the experimental results in the image processing pipeline. A summary of the final parameters and variables used for the simulation is shown in **Table 1**.



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

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

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

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

Figure 6. Stochastic transcriptional bursting SimBiology model and reactions.

Parameter or Variable	Value
Overall mean free mRNA (*mRNA_f*)	479 (#)
Overall mean nascent mRNA (*mRNA_n*)	7169 (#)
Degradation rate (*gamma_1*)	0.05 min^{-1}
Transcription rate (*k_tsc*)	21 min^{-1}
Release rate (*k_release*)	0.0033 min^{-1}
Production rate (*k_prod*)	23.9505 min^{-1}
Forward reaction rate (*k_on*)	0.0500 min^{-1}
Reverse reaction rate (*k_off*)	-0.0062 min^{-1}

Table 1. Stochastic gene expression simulation parameters/variables and associated values.

Results

Image Processing

The results of the image processing pipeline for all images are shown in **Table 2**. **Figure 7** shows the distribution of mRNA intensity levels for each image. The mean, variance, coefficient of variation, and correlation coefficient for both the free and nascent mRNA per cell are displayed in **Table 3**. **Figure 8** shows the corresponding scatter plots used to calculate the correlation coefficient for the nascent mRNA as a function of free mRNA per cell for each image.

Image	Cell Count	Free mRNA	Nascent mRNA	Total mRNA
A	5	224.3435	2235.115	2459.458
B	12	4421.238	46211.04	50632.27
C	24	6464.81	173856.2	180321
D	56	34932.91	377532.2	412465.1
E	52	21530.24	466012.9	487543.2
F	47	26312.45	339235.9	365548.3

Table 2. Results of the image processing pipeline for each image pair. Free, nascent, and total mRNA values are the number of mRNA molecules present.

Image	Free mRNA			Nascent mRNA			CC
	Mean	Variance	CV	Mean	Variance	CV	
A	44.8687	3788.377	1.371776	447.023	656871.6	1.813053	0.50
B	368.4365	76785.96	0.7521047	3850.92	128599300	2.944794	-0.17
C	269.3671	238412.5	1.812675	7244.006	35384580	0.8211608	-0.50
D	623.802	1317763	1.840229	6741.646	72428440	1.262376	-0.22
E	414.043	431566.8	1.586641	8961.787	119654400	1.22059	-0.10
F	559.8393	259162.7	0.9093327	7217.784	88615370	1.304219	-0.22

Table 3. Statistics on mRNA molecules per cell. CV is the coefficient of variation and CC is the correlation coefficient for the free mRNA as a function of nascent mRNA.

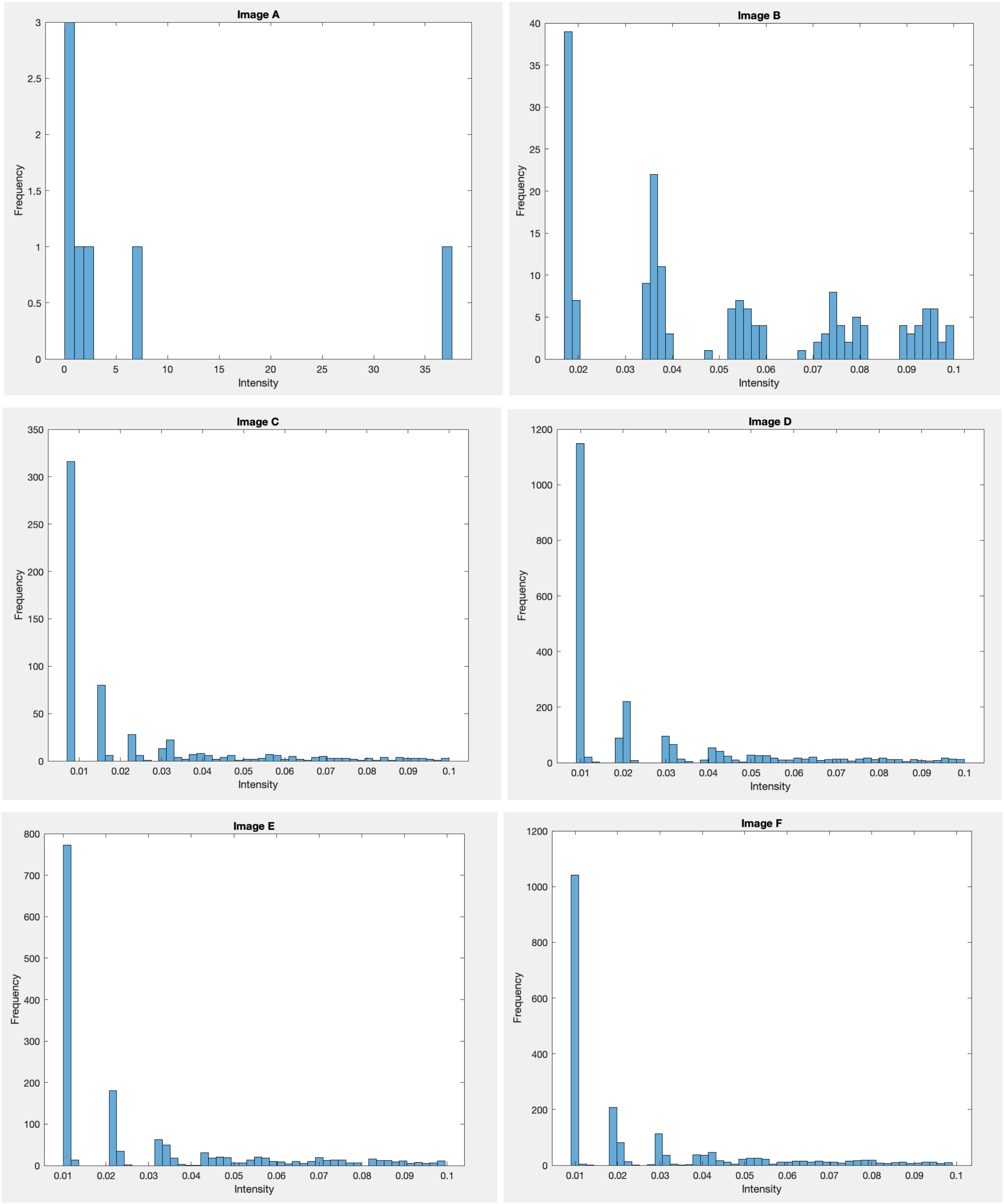


Figure 7. Distribution of total mRNA intensities across all images. Each histogram is labeled with its corresponding image.

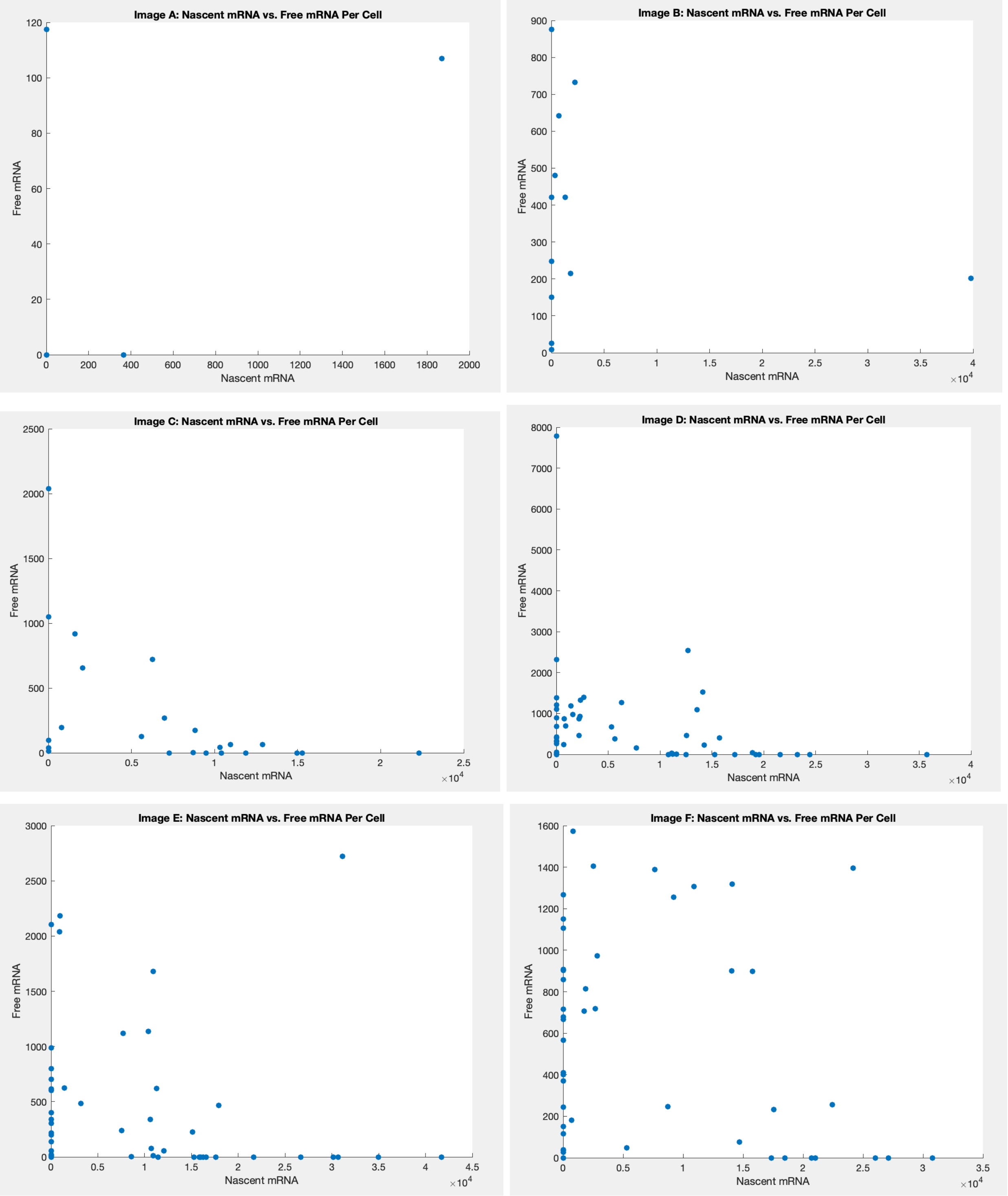


Figure 8. Scatter plots for the free and nascent mRNA levels per cell. Each point represents a single cell within the respective image.

Modeling of Stochastic Gene Expression

The predicted “on” rate (forward reaction rate) was estimated to be 0.05/min, while the predicted “off” rate (reverse reaction rate) was estimated to be -0.0062/min (see **Table 1**). These parameters were found to produce a nascent mRNA variance of roughly 320, which is significantly lower than the variance results from the image analysis pipeline (multiple orders of magnitude lower), but was the highest achievable nascent mRNA variance using the given parameters. This could be due to accumulated errors in our image processing pipeline. Several other statistics from the simulation are shown in **Table 4** below. Histograms of the frequency of both free and nascent mRNA molecules at the final *t=100 min* time point were also created to show the distribution of both species in **Figure 9**.

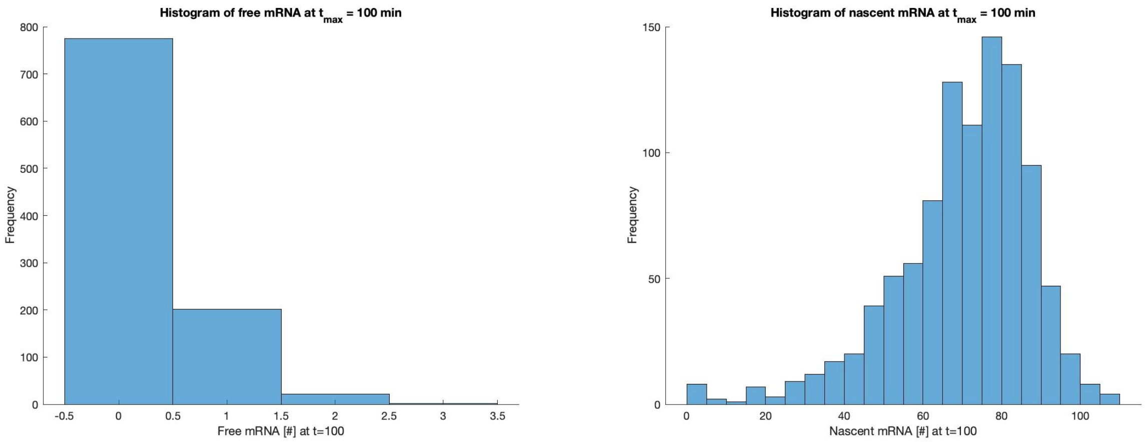


Figure 9. Histograms of free and nascent mRNA species at the end of the stochastic simulations.

Statistic	**Value**
Nascent mRNA variance (*mRNA_n_sim_var*)	320.2046 (#)
Total mRNA mean (*total_mRNA_sim_mean*)	69.2030 (#)
Total mRNA variance (*total_mRNA_sim_var*)	354.7265 (#)
Total mRNA coefficient of variation (*total_mRNA_sim_cv*)	0.2722 (27.22%)

Table 4. Calculated stochastic gene expression simulation statistics.

Discussion

Overall, the image processing pipeline does an adequate job of isolating the mRNA from each cell and calculating the relevant quantities. The correlation coefficient for the experimental data could possibly have higher error as it was expected to be closer to -1 for all of the images. However, this could be due to discrepancies in the exact methods used to calculate the mRNA per cell as the threshold for a single mRNA molecule was set simply by looking at the mRNA intensity distribution. Furthermore, it should be noted that not all of the cells had a “bright spot” that would normally be indicative of mRNA at a transcription site, resulting in several cells per image having free mRNA but no nascent mRNA. Implementing a more quantitative and precise method for estimating this value would have helped in obtaining more accurate measurements from the images.

The stochastic gene expression model requires a number of assumptions for determining the forward (on) and reverse (off) reaction rate constants; in particular, because it relies on the calculation of nascent mRNA variance from the image processing pipeline, it is highly sensitive to error. Improving the accuracy of mRNA measurements would thus improve the reliability of this model. A more robust method of determining the rate constants, perhaps with a simplified set of reactions, might also be appropriate. The limited data (mRNA levels) from image processing result in an underdetermined system, forcing assumptions to be made, such as the degradation rate and transcription rate parameter values. Lastly, the method of iterating across numerous pairs of forward (on) and reverse (off) reaction rate constants and performing a 1000-iteration ensemble simulation for each set to identify the optimal parameters is tedious and computationally expensive, and gives relatively low precision. A better optimization algorithm would produce a better set of simulation parameters.

Conclusion

In the image processing part of this investigation, techniques such as filtering, binarization, and watershedding were used to quantify the levels of mRNA in each cell in a pair of fluorescent DAPI and smFISH images. In general, there appears to be a weak, negative correlation between the nascent and free mRNA. This makes logical sense as if more of the total cellular mRNA is located at the transcription site, there will be less available freely throughout the cell. However, the overall accuracy and robustness of the image processing pipeline could be improved by more effectively isolating individual nuclei in each image and estimating the mRNA thresholds with a more quantitative method.

In the stochastic modeling part, the experimental results obtained from the image processing pipeline were used to develop a model for transcriptional bursting within a cell. The set of rate parameters that were calculated and estimated from the image data did not produce a well-representative model of transcriptional bursting. When comparing the experimental results to the theoretical predictions, the accuracy of the model itself is highly dependent on the calculated nascent mRNA variance. As this value itself likely has errors from inconsistencies in the image processing pipeline, the predictions from the model must be interpreted within the context of the experimental error. Thus, improving the accuracy of the pipeline would also improve the predictive capability of the stochastic model.

Code

See all image data, SimBiology models, and MATLAB code at github.com/robertheeter/BIOE446/tree/main/project. The *results/* folder contains all results from the project, including segmented DAPI and binary smFISH images after image processing. The *scripts/* folder contains all MATLAB scripts.