First Annual Online Undergraduate Quantitative Biology Summer School UQ-Bio2021. Final Projects.

UQ-bio 2021 organization team

June 16, 2021

UQ-bio 2021 Final Projects Instructions

During the UQ-bio Summer School, students are required to test the computational skills. For this, the student must select between one of the three projects described below and complete the selected challenge by creating a report using a Python notebook. The students that satisfactorily complete this challenge will receive the certificate of completion.

Submission Instructions:

- The final report must be submitted as link to a Google Colab Notebook to the official q-bio email: engr-cbe_qbss@Mail.colostate.edu
- The entire team should submit one Colab Notebook.
- The Colab Notebook must include the name of the people in the team. If a participant in the team did not contribute to the project in a meaningful way, please do not list his name in the submission.

Deadlines:

- Progress will be checked every week during the meeting with the mentors.
- The final Colab Notebook must be send for evaluation by July 09, 2021.

Evaluation:

Projects will be evaluated based on:

- Quality and organization of the notebook
- Functionality
- Correctness of the response

List of contacts:

The following lists contain information about the teams members that you can contact for technical or scientific questions.

Project 1. Single-cell yeast response dynamics

Teams	Name	Contact
Project Leader	Zach Fox	zachrfox@gmail.com
Team 1a	Bryce Asay	aphlnbsba1@utah.gov
	Heidi Klumpe	hklumpe@caltech.edu
Team 1b	Kaan Öcal	kaan.ocal@ed.ac.uk
	Jafar Anafi	jafar.anafi@aims-cameroon.org

Project 2. Single-cell RNA FISH analysis.

Teams	Name	Contact
Project Leader	Huy Vo	Huy.Vo@colostate.edu
Team 2a	Michael May	michaelpmay@live.com
	Eric Ron	Eric.Ron@colostate.edu
Team 2b	Lisa Weber	llweber@rams.colostate.edu
	Linda Forero	Linda.Forero_Quintero@colostate.edu

Project 3. Single mRNA translation dynamics.

Teams	Name	Contact
Project Leader	Luis Aguilera	Luis.aguilera@colostate.edu
Team 3a	Keisha Cook	kcook7@tulane.edu
	Nava Leibovich	navaleib.research@gmail.com
	Joshua Cook	jcook10@rams.colostate.edu
Team 3b	Will Raymond	wsraymon@rams.colostate.edu
	Brooke Silagy	brooke.Silagy@colostate.edu

Project 1. Single-cell yeast response dynamics.

This project will look at synthetic gene regulatory mechanisms in yeast and quantify cellular responses to optogenetic inputs. This project is being organized by Dr. Zachary Fox using real data collected as part of his postdoctoral studies at Institut Pasteur in Paris.

1. Module 0 Tasks:
    - Create a Python notebook using Google Colab or Jupyter notebooks.
    - Import the needed modules and libraries.
2. Module 1 Tasks:
    - Load brightfield image and display in Colab notebook
    - Load fluorescent image and display in Colab notebook
    - Merge the two images to create a single image which overlays the fluorescence onto the brightfield image.
3. Module 2 Tasks:
    - Use the segmentation tool to segment the brightfield images
    - Use the tracking tool (TrackPy) to track the cells throughout the movie.
    - Download `cell_ids_1.npy` from (Link provided by the instructor.)
    - Plot number of cells in the field of view as a function of time
    - Fit to an exponential curve and extract the growth rate of the population
    - Plot a histogram of cell size
    - Plot a histogram of fluorescence across the population
    - Use a clustering algorithm to assign cells into 4 different bins. Compare these to the cells ids in `cell_ids_1.npy`.
4. Module 3 Tasks:
    - Simulate the stochastic model for the optogenetic system using the SSA.
    - Record and save solutions, plot histograms, plot mean over time, bleep blop
5. Module 4 Tasks:
    - Compare stuff from module 3 to module 4.
6. Module 5 Tasks:
    - Fit the distributions.

Project 2. Single-cell RNA FISH analysis.

Description

This project will look at single cell single molecule measurements for the number of RNA per cell in different experimental conditions. Model selection and parameter estimation are essential tasks in computational modeling of these data sets. Students will have a taste of the modeling workflow using analysis tools introduced throughout the modules (Summary Statistics, Finite State Projection, Maximum Likelihood Estimation, Markov Chain Monte Carlo). You will be given a set of mRNA counts at different measurement times that are drawn independently from an unknown stochastic biochemical process. You will then need to fit and decide between competing explanatory models of this data. After selecting the best model, you will draw MCMC samples from the posterior distribution of parameters for this model and provide predictions (with Bayesian confidence estimates) of physically relevant quantities such as mRNA half-life, mRNA synthesis rate, burst frequency. This project is being organized by Dr. Huy Vo.

Datasets

- **Imaging data**: Imaging data from Prof. Douglas Shepherd at Arizona State University. This dataset will be analyzed during Module 1.
- **mRNA count data**: Single-cell mRNA count data collected at different time points. These will be used for all modules except Module 1.

Instructions

1. Module 0 Tasks:
    - Learn how to load comma-separated CSV files by either Pandas or Numpy.
    - Learn how to use Python to probe the basic information about the data: how many single-cell measurements are there? What is the dimensionality of each observation? You can compare your code output with what you get from eyeballing the data, but you must provide the code.
2. Module 1 Tasks (co-managed by Dr. Linda S. Forero-Quintero):
    - Load the .TIFF files onto Colab notebooks.
    - Produce mRNA counts per cell for each gene. You need to produce three sets of results corresponding to different settings for the image processing tool.
    - Appreciate the uncertainty due to different settings of the image analysis process.
    - Save these outputs into an appropriate file on your Google Drive.
3. Module 2 Tasks:

- Plot the histograms of the observed mRNA counts in the data.
- Compute the summary statistics of the data: mean, variance of the mRNA counts, the Fano factor (defined as the ratio between the variance and the mean). Plot these statistics vs time for all biological replicates.
- Appreciate the variability of summary statistics computed from finite data.

4. Module 3 Tasks:

   - Code the stoichiometry matrix and propensity functions for all of the three competing models.
   - Find some physically plausible parameters for these models (you can consult a book such as *Cell Biology by the Numbers* or use my suggested parameters. Note that these are not yet the real parameters to fit the data, and you will need to optimize these parameters to fit the data after Module 5. Then draw Monte Carlo trajectories of the three models from these data, using either PySB or your own implementation of the Gillespie algorithm.

5. Module 4 Tasks:

   - Modify the provided skeleton code of the FSP to appropriately compute the distribution of mRNA counts in the three models.
   - Write a function to compute the log-likelihood given the model and the parameter vector. You will need to compute the probability of the single-cell observations given the parameter and model using FSP.

6. Module 5 Tasks:

   - Before doing any fitting, specify a prior for all parameters in each model. You can use the 'guess' parameters used in Module 3. The prior distributions need to be properly dispersed to represent our lack of information about the true process. *You must not look at the data or the fit when specifying this prior.*
   - Use maximum likelihood estimation to identify the best fit parameters for the three competing models. The likelihood could be computed either by analytical solution (when available), or by the finite state projection (your own implementation from Module 4).
   - Use MCMC (your own implementation or via a well-established library such as PyStan) to draw samples from the posterior distribution of the parameters for the best model selected above. Use the prior you specify above.
   - From the drawn samples (appropriately truncated to discard the 'warm-up' period and tested for non-convergent issues) provide posterior prediction for these quantities of interest: burst size, burst frequency, mRNA degradation rate, mRNA synsthesis rate.

- Compute model-predicted mean, variance of mRNA counts and the model-predicted Fano factor. Overlay this with the plot of data-derived summary statistics you made in module 2. Comment on how close/different they are.
- OPTIONAL: could you come up with a better model than the three models provided? Explain.

Project 3. Single mRNA translation dynamics

Description of the project:

During the last years, important advances in molecular biology and fluorescence microscopy have allowed the quantification of single-molecule gene expression. These modern technologies generate large amounts of data that require computational and mathematical analyses to interpret these datasets correctly. In this project, we will develop Python codes to analyze live-cell super-resolution fluorescence microscopy data to quantify the translation dynamics on single molecules of mRNA for different mRNA species. This project aims to give the student new computational skill required for image processing, modeling biological phenomena, and estimating biophysical parameters. This project is being run by Dr. Luis Aguilera using real and synthetic data collected from his collaborations with the Stasevich and Munsky labs at Colorado State University.

Instructions:

1. Module 0 Tasks:

    - Get in contact with the project leader, learning assistants, and students in your group. Please make sure join in the discussion forums on slack, check out the tutorials and lectures.
    - Get familiarized with the Python language and learn how to get the best of Google Colab or Jupyter notebooks.
    - Learn how to import Python modules that are needed to manipulate microscope images. Do some research to find additional modules that may help you with image processing.
    - Check the concepts of module dependencies, module version, module compatibility, and the official package installer for Python (pip).

2. Module 1 Tasks:

    - Use the image processing techniques learned in Module 1 to analyze single-molecule microscope data and perform the following procedures:
        i) cell-segmentation,
        ii) single-molecule tracking,
        iii) intensity calculation.
    - Evaluate how accurate are your analysis by comparing your results with the dataset provided by the instructor.

3. Module 2 Tasks:

    - Use the provided synthetic data to calculate important statistics to describe the data tendency (mean, std, auto-correlation).

- Use the provided synthetic data and implement a machine learning approach to classify spots based on mobility automatically.
- Determine the accuracy of your implementation using the true data frame provided by the instructor.
- Explore different approaches such as architecture types, classifiers, and data normalization to achieve higher accuracy.
- Consider the limitations of the teams final/best approaches.

4. Module 3 Tasks:

   - Install rSNAPsim in a Colab (https://pypi.org/project/rsnapsim/).
   - Using rSNAPsim perform stochastic simulations to reproduce the single-molecule translation dynamics of a selected gene.
   - Generate some summary statistics of the translation dynamics for the selected gene, including intensity distribution, mean intensity, auto-correlation dynamics.
   - Estimate important biophysical parameters for the simulated gene, including initiation and elongation rates.

5. Module 4 Tasks:

   - Calculate the following biophysical parameters from the synthetic data:
     i) ribosome elongation rate,
     ii) ribosome initiation rate,
     iii) diffusion constant.
   - Implement a Metropolis–Hastings algorithm to calculate the confidence interval in your estimated parameters.

6. Module 5 Tasks:

   - Work with your team to prepare the final Colab notebook.
   - Work with your team to create a final presentation with your project.