Capstone Proposal

Estevao Vieira

March 2017

1 Domain Background

Bright Yellow cells are a model cell line for studying plants. They are largely used in cell cycle analysis, such as morphology and differential gene expression. To estabilish relations between gene expression during the cell cycle and the plant's phenotype, we have to understand the expression pattern of the gene of interest.

While the technology needed for single cell expression analysis is not yet stabilished for plants, it is possible to determine this expression pattern by using populations of cells. For this analysis to be reliable, the population has to be majoritarily in the same cell cycle phase. For this, a cell cycle synchronization is commonly used[1].

2 Problem Statement

In biological researchs that involve events of the cell cycle or cell morphology, sometimes it is salutar to synchronize a population of cells to enable comparisons or descriptions[1]. To make any conclusions acceptable, one crucial step is that of the syncronization validation (i.e. showing that the cells were in fact well synchonized), for without sure that the procedure worked and the cells were synchronized, any later result is meaningless. For the validation, in the case of tobacco BY-2 cells, populations of cells have to be observed after 8 to 10 hours after the end of procedure and the proportion of cells in the mitotic stage has to be close to 70%. The present project aims at automating most of this process of validating by detecting if a given BY cell is in the mitotic stage.

3 Datasets and Inputs

- The dataset consists in approximately 5000 photos of single BY-2 cells, 1200 of which are labeled by their mitotic stage (interphase or mitosis) or are labeled as Unknown, Unclassifiable, Not a cell

- Original photos have been taken at GaTE lab at University of São Paulo, using DAPI stain, that colors genetic material in fluorescent blue, and have been cropped automatically around cells.
- The dataset photos are 100x100 pixels, RGB.
- The stages are not represented in the same proportion, such that interphase is more frequent in the dataset.
- The labels have been provided by the benchmark model
  - Interphase or Mitoses are labels that say something about the cell phase.
  - Unknown is a label given when answer is very unsure.
  - Unclassifiable labeled photos have cells broken apart in some part of the process, "dead cells".
  - Not a cell means that the photo has been wrongly cropped, such that there is nothing to classify.

4 Solution Statement

A solution to the present problem consists in automating part of the synchronization's validation by using a classifier based on supervised learning to identify cells currently undergoing cellular divison or interphasis.The cells can then be counted and the proportion of them in some part of the division stage (mitotic index) can be easily calculated. This process involves dumping away bad photos, a work also done by supervised learning.

5 Benchmark Model

In this domain, the process of classification is not commonly automated, such that the most widely used method is manual classification of cells via visual comparisson to standard images[1]. The benchmark to be considered is thus a typical scientist that works with cell morphology, and its performance can be approximated by the mean performance of three biologists of this kind(each making its classifications alone) via the evaluation metrics described below.

6 Evaluation Metrics

For evaluation of the model's performance, sets of images classified in consensus by a pool of biologists will be compared to the results from the model, and the performance will be given by the proximity between the mitotic index of both, calculated via mean squared error.

7 Project Design

7.1 Preprocessing

- Dump away cells classified as "Unknown".
- Transform photos from RGB to grayscale, since this loses very little information in this case and well reduces the dimensionality.
- Transform photos to the frequency domain via a Fast Fourier Transform in two dimensions. This may suit well our needs because the cells can be in any direction, so that shouldn't make difference on which one it is. In the frequency domain we get measures of pixel correlations at given distances, what seems much more in line with the classification task in hand.
- Apply PCA on the data, to reduce from the big 100x100 dimension space. To be worked later, a substitution of this step with ICA is also planned, and the best one chosen after some test.
- Split data into train and test in a stratified k-fold manner for validation.
- Encode labels

7.2 Classification

1. Separate photos into good photos (interphase, mitosis) and bad photos (Unclassifiable, Not a cell).
2. Make a list of ten possible classifiers to be cross validated for choosing the two best ones.
3. Grid search the two classifiers for the best parameters, and choose the best one.
4. Separate good photos into mitosis and interphasis.
5. Classify probabilistically the unlabeled part of the dataset.
6. Label as 1 every point fitted with more than 90% belief, and label 0 every point below 10%.
7. If more than 10% of the points are successfully classified as above, retrain the classifier and repeat steps 5 and 6.

References

[1] Kumagai-Sano F, Hayashi T, Toshio Sano T, Hasezawa S, Cell cycle synchronization of tobacco BY-2 cells, Nature Protocols 2006, Published online 11 January 200[1].