# 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].

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[1].

## 2 Problem Statement

For the synchronization 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. Two binary classifications have to be applied in sequence:

1. Photo classifier: Separate good photos from bad photos (this is made clearer from the dataset explained below)
2. Phase classifier: Among the good photos, separate dividing cells (mitosis) from resting cells (interphasis)

# 3 Datasets and Inputs

- The dataset consists in approximately 3000 photos of single BY-2 cells, labeled by their mitotic stage (interphase or mitosis) or 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.

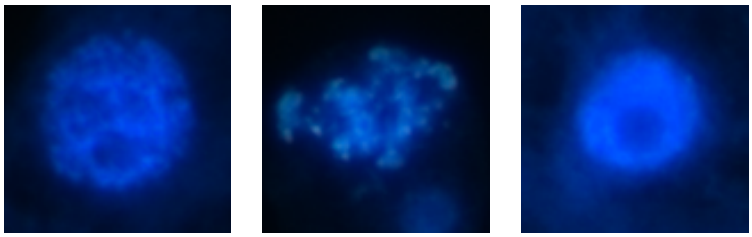

Fig 1: Two cells during mitosis in the left, and an interphasis cell on the right.

# 4 Solution Statement

Many traditional algorithms of supervised classification will be compared, such as Support Vector Machine and k-Nearest neighbors (others are listed in project design), given as inputs two features proven to be successful in texture analysis, i.e Histogram of Gradients (HoG) and Local Binary Patterns (LBP)[2], and for each classification step the best one will be chosen.

Both classifiers's (photo and phase) performance are measurable by comparing predicted to real labels, via the evaluation metrics explained below.

# 5 Benchmark Model

In this domain, the process of classification is usually by manual classification of cells via visual inspection and comparisson to standard images[1], and lacks a stabilished automated model. The benchmark below was chosen in contrast to random guessing to account for biases in the label distribution (such as the greater number of interphasis) and to learn with simple features, for an estimation of what can be gained in our case by the extraction of more complex features.

The benchmark model will then be a Support Vector Machine classifier acting upon simple features, i.e the fourier transform of the original photo (of the blue channel, which is the only important because of the stain). This benchmark model is expectedly better than random guessing, and will be generated by grid searching an Radial Basis Function kernel for C values [0.01,0.1,1,10,100] and $\gamma$ values [0.25,0.5,0.75,...,4] via 5-fold cross validation.

Both classification task (photo and phase) will have their own benchmark generated as above, each with their own parameters.

# 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 F-score (the harmonic mean between precision and recall) for the mitosis detection. This performance is well suited for our purposes, given that we both want a high detection rate for mitosis and do not want false positives.

# 7 Project Design

## 7.1 Preprocessing

- Throw away cells classified as "Unknown".
- Get only the blue part of RGB, since this loses very little information in this case and well reduces the dimensionality.
- Get the Histogram of Gradients and Local Binary Pattern of cells.
- Apply PCA on the data, to reduce the 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. Train in a 5-fold manner the following eight classifiers(with default parameters):
   - Random Forest
   - k-Nearest Neighbors
   - Support vector machine (Radial basis function)
   - Linear Support vector machine
   - Decision Tree
   - Logistic Regression
   - Gaussian Process Classifier
   - Naive Bayes
3. Grid search the two best classifiers for the best parameters, and choose the best of them.
4. Separate good photos into mitosis and interphasis.
5. Repeat steps 2 and 3

## 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].

[2] Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Huang, T. (2011, June). Large-scale image classification: fast feature extraction and svm training. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1689-1696). IEEE.