Classifying Mechanical Environments of Cancer Cells

Ari Spiesberger, Chris Price

April 2020

1 Background

The morphology and migratory behavior of cells is tightly interwoven with the properties of the extracellular environment.[1] Cells interact with their environment through a complex series of both active and passive mechanical and chemical feedback loops. Identifying and differentiating these mechanisms is critical to understanding disease progressions such as cancer metastasis, which occurs via rapid cell migration *in vivo*. In a recent study, Wisdom *et al.* used 3D confocal microscopy to study the effect of extracellular matrix plasticity on the time evolution of cancer cells in the presence of various protein inhibitors.[3] They observe that these cells alter their environment through oscillatory protrusion formation (invadopodia) and corresponding mechanical plastic deformation of the surrounding medium, allowing for subsequent cell migration.

In this work, we will utilize the dataset collected by Wisdom *et al.* and investigate hypothesize that information about the time evolution of the cell shape can successfully classify the environment of the cell. A successful classification model can provide information about the microstructure of the local cell environment which is otherwise difficult to measure *in vitro*.

2 Overview of the Dataset

The initial dataset consists of 25 separate image series (assays), each 100-120 grayscale frames in length, where each frame is a particular 2D slice of the vertical focal plane taken at 10 minute intervals. Each assay contains 40-80 cells visible in the field of view for the duration of the experiment, and corresponds to a class of matrix plasticity (low, medium, or high) and protease inhibitor (inhibited, or uninhibited). This corresponds to roughly 12,000 freeze frames of individual cells per extracellular environment.

Additionally, we have access to cell tracking data calculated on the same set of experiments from an image analysis program. While we do not have a direct map from the cells in the raw images to the tracks in the analyzed data, we can use the relative position of the cell in the image to match image cells to cell tracks. The tracking data links unique cells through the time dimension. Rough measurements of the cell shape are also present in this dataset, such as the sphericity and ellipticity in 3D. While these metrics are poorly suited to describe cells which adopt highly anisotropic shapes with both positive and negative curvature, they retain some 3D information which is lost by viewing the 2D images in the focal plane.

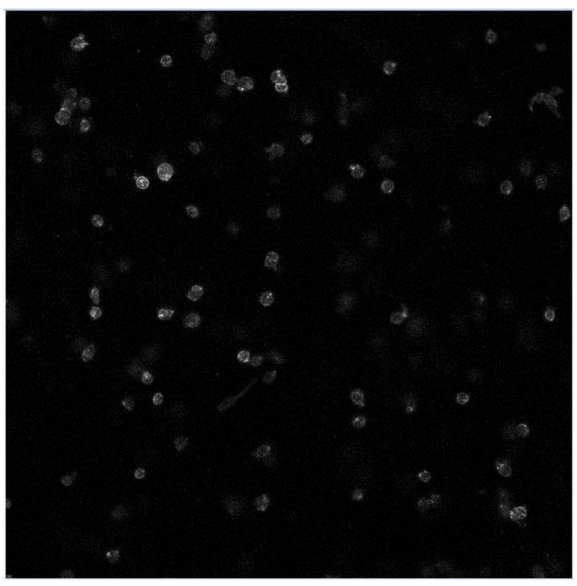

Figure 1: Sample raw image of cells at a single timestamp.

3 Proposed Work

3.1 Hypothesis

The morphology of cell shapes and the amplitude and frequency of the shape change over time can be used to classify the mechanical environment of the cell.

3.2 Project Steps

1. Transform the dataset from 25 time series of an ensemble of cells into individual time series for each cell, retaining the environmental label information.

2. Without using the time series information, classify the environment of each cell using a convolutional neural network (CNN). Analyze the error within individual cell tracks along with error across the dataset.

3. Incorporate time information by adding a recurrent neural network (RNN) or long-short term memory (LSTM) model to the output of the convolutional layers. Compare the classification accuracy of using the image lags to the classification of individual images.

4. "Doubtful", Classify the motion of the cells onto certain "Random Walk" type algorithms

4 Method Details

4.1 Transforming the Dataset

The initial phase of this project requires segmenting the raw images to transform each ensemble time series into many single-cell time series. Challenges and sources of error in this step include: cells drifting out of focus, cell mitosis (splitting in two), cells which are close together, and variable contrast ratios across different assays. Due to the relatively simple contrast within each image, however, we plan to try a thresholding procedure followed by a connected components algorithm to identify each cell object in the image frame. If this fails to distinguish a majority of the individual cells, we can switch to more sophisticated image segmentation algorithms, such as an autoencoder.[2] After identifying each cell, we will place the identified cell pixels at the center of a square, uniform background. We will scale the average pixel value of each cell to a fixed value to prevent contrast bias across assays from polluting the dataset.

4.2 Environment Classification

After transforming the dataset, the goal is to classify the cells by their mechanical environment and the presence / absence of certain chemical pathway inhibitors (drugs) related to cell migration. We have identified three methods of approaching this task, with increasing levels of complexity and expected accuracy:

4.2.1 RNN on Extracted Vectorized Data

We can utilize the extracted 3D shape metrics, organized by cell and timestamp, to train a classification model for the cell environment. We anticipate that the collection of these metrics averages out too much detail from the cells to be useful. However, this would be simple to implement and provide a baseline for both classification accuracy and the importance of 3D vs. 2D shape information.

4.2.2 CNN on Single Cell Images

Adding more features to the classification requires utilizing our transformed dataset. Ignoring time information, we can use a convolutional neural net to classify each cell image by its environment. This model will not be aware of any time series information, but will seek to build the classification model from the frozen single-cell snapshots. We can asses the accuracy of the classification across each cell, as well as the uniformity of classification among images which belong to the same track. For the multi-class classifier, we plan to use a cross-entropy loss with a softmax function to give probability estimates for each classifier.

4.2.3 CNN + RNN or LSTM on Cell Image Time Series

Taking inspiration from [4], we can incorporate the image time series information into our model by taking the output of our convolutional layers, flattening it, and creating several lags. This exploits the time translational symmetry of each cell track and informs the model about the variance of the cell features in time. Computationally, we expect that this model could be demanding in memory requirements, so we plan to use the nodes on the Chestnut computing cluster (Chris has access) if memory issues are encountered. The loss function will be the same as the previous method, enabling direct comparison and interpretation of the information gained by adding the time axis.

5 Broad Challenges

We have set up a github project to collaboratively develop these goals. Chris will develop the data wrangling and segmentation, and Ari will work on the convolutional interpretation and bench marking of the data. The implementation of the time series and experimental development will be tackled jointly between Ari and Chris.

Data wrangling is certainly a challenge. The realities of imaging cells in 3D over multiple hours and on different days introduces noise to the data. Furthermore, cells may come into and out of focus during the sampling period, so that each cell track does not span the entire time series. To address this, we will need to prune the dataset to only include cells we have certainly tracked properly. This means that in whatever cell tracking software we have built, we will need to dropout any cell track which does not contain a minimum of consecutive timestamps, or where the tracked object is above a maximum total size (likely to be 2 cells bunched together).

Another problem involves class balance in training the classifier. Part of the data normalization process will be ensuring that the populations of each class in the training set is equally represented. Depending on the severity of previous data cuts, we can achieve this by randomly dropping data from overrepresented classes, or augmenting underrepresented classes by adding 90° rotated images.

The next problem is that the effects that certain environments or drugs may have on cell morphology may be very subtle. This is why we propose to use full raw image data instead of software built data. However in this situation we have the possibility of strong over-fitting. We will implement dropout layers and down-sampling to develop the most robust model. However this will only be effective up to a point in a multi-classification problem.

One very difficult challenge we expect to deal with is potential over-fitting. This may occur due to the subtlety of differences between our environments. We hope that the introduction of raw image data will be able to overcome this. Potentially, using dropout methods, skip-layers, and perhaps L2 penalties we will be able to overcome over-fitting.

5.1 The Overall Idea/Plan

We hope to develop a robust technique for any data of the class of which we describe. To do so we will proceed by testing and working with a number of different methods

The first task is to accurately track and segment cells

- Perform Bench marking to measure experimental accuracy, the following are the bench marking tools we will use
  1Using a standard LSTM on the vectorized data, we will get some measure of initial accuracy. Standard Activation, Cross-Entropy Loss with soft max. Lags will be whatever gives reasonable test accuracy..
  2Using Method proposed by [4] to build CNN $\rightarrow$ RNN on vectorized data to get standard classification accuracy.
  3 Using the original down-sampled images through standard CNN to predict classifier. 3 Convolution Layers, ReLU activation, Cross Entropy Loss. We will include modern updates on the image classification problem including skip-layers, and the dropout method.

- Develop Proposed Methods, to increase accuracy overall. As mentioned in 4.2.3, we propose to improve on previously developed methods from [4], and create a stronger classifier for data in the form we present. To improve accuracy we will work with different set-ups on our models. Varying Activation, Layer-size and depth. Using different LSTMs, (lags, and modifications).
  1 We will track accuracy of our model with both training accuracy and with a confusion matrix. We will observe training time, and any steps we make in accuracy or confusion. Simply, we aim to be as accurate overall.

References

[CS14] Guillaume Charras and Erik Sahai. "Physical influences of the extracellular environment on cell migration". In: *Nature Reviews Molecular Cell Biology* 15.12 (Dec. 2014), pp. 813–824. ISSN: 14710080. DOI: 10.1038/nrm3897.

[WYL17] Chunlai Wang, Bin Yang, and Yiwen Liao. "Unsupervised image segmentation using convolutional autoencoder with total variation regularization as preprocessing". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Institute of Electrical and Electronics Engineers Inc., June 2017, pp. 1877–1881. ISBN: 9781509041176. DOI: 10.1109/ICASSP.2017.7952482.

[Wis+18] Katrina M. Wisdom et al. "Matrix mechanical plasticity regulates cancer cell migration through confining microenvironments". In: *Nature Communications* 9.1 (Dec. 2018), pp. 1–13. ISSN: 20411723. DOI: 10.1038/s41467-018-06641-z.

[KBM19] Jacob Kimmel, Andrew Brack, and Wallace Marshall. "Deep convolutional and recurrent neural networks for cell motility discrimination and prediction". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019). ISSN: 15579964. DOI: 10.1109/TCBB.2019.2919307.