A Machine Learning Approach to Cell Classification

MoHan Zhang[a], Cindy Tan[b] and Dhananjay Bhaskar[b]
Supervisors: Dr. Leah Edelstein-Keshet[b], and Dr. Calvin Roskelley[c]
[a]Department of Mathematics, University of British Columbia
[b]Faculty of Applied Science, University of British Columbia
[c]Department of Cellular and Physiological Sciences, University of British Columbia

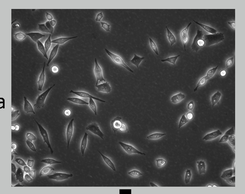
Introduction

The precise regulatory mechanism that governs cell shape, size and polarity is not well understood. To facilitate a systematic investigation of cell morphology, we have developed tools to identify cells from live imaging data, quantify cell geometry and automatically classify cells using unsupervised machine learning. This poster illustrates our methodology.

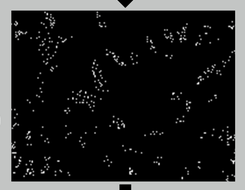
Step 1: Image Processing

We obtained 149 correct segmentations from 20 *in-vitro* pancreatic cancer images acquired by the Roskelley lab:

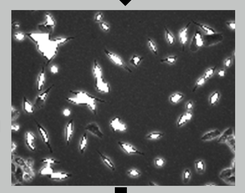
MIA PaCa-2 Cell Line
In vitro Model for Pancreatic Carcinoma
Phase-Contrast Microscopy



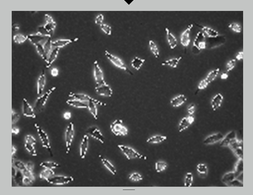
Edge Detection
Sobel–Feldman Derivative Filter
Grayscale Binarization (Thresholding)



Foreground Binary Mask
Mathematical Morphology
Erosion, Dilation, Opening and Closing



Segmentation
Distance Transform
Marker-Based Watershed Algorithm



Out of 149 segmented cells, four distinct morphologies were identified, namely circular, elliptical, elongated and cells with one or more protrusions.

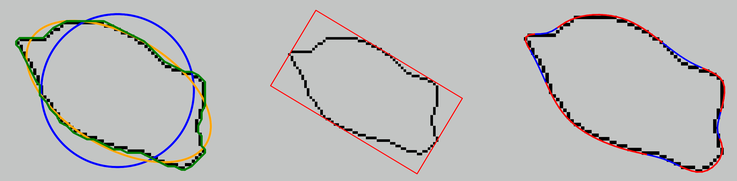
To validate our methodology, 63 cells were manually selected for feature extraction (Step 2); with 15, 16, 20 and 12 cells exhibiting circular, elliptical, elongated and protrusive morphology respectively.

Step 2: Feature Extraction

Consider an arbitrary geometry $f(\theta) = 0$ parametrized by M features, $\theta = (\theta_1, ..., \theta_M)^T$. To fit this geometry to a set of boundary points $(x_i, y_i)_{i=1}^N$, we solve the following optimization problem:

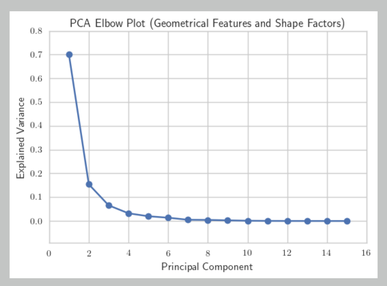
$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} r_i^2(\theta)$$

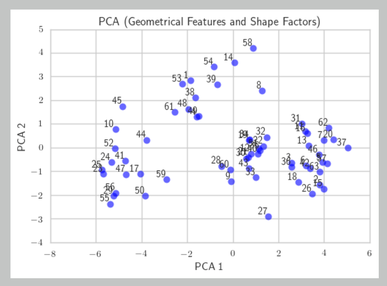
where r_i is the orthogonal distance between boundary point (x_i, y_i) and shape $f(\theta) = 0$.



Step 3: Dimensionality Reduction

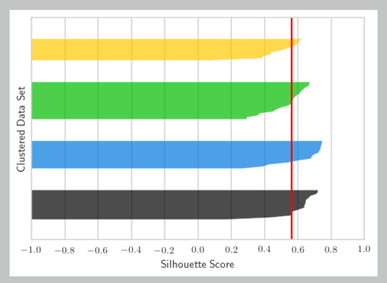
Linear combination of correlated features (Principal Component Analysis) project high dimensional feature vectors to 2-D space for clustering:

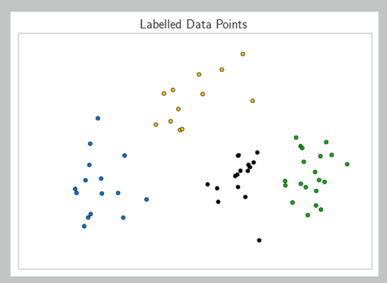




Step 4: Classification

Silhouette score computed using k-means determines number of clusters:

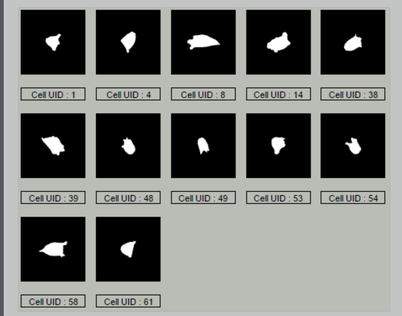




Step 5: Validation

Black cluster labels correspond to elliptical morphology

Green cluster labels correspond to elongated morphology



Yellow cluster labels correspond to protrusive morphology

Blue cluster labels correspond to circular morphology

Future Work

- Compute additional boundary features and quantify cell shape symmetry
- Develop new methods to identify clusters in higher dimensions
- Incorporate motion-based features from time-lapse microscopy
- Identify cells that morph during time-lapse microscopy

References

1. Amin et al., 2015. *Medical Signals and Sensors* 5, 49-58.
2. Rousseeuw, 1987. *Computational and Applied Math* 20, 53-65.
3. Sommer and Gerlich, 2013. *Cell Science* 126, 5529-5539.
4. Sun and Luo, 2009. *Microscopy* 233, 326-330.