A Machine Learning Approach to Develop an EMT Classifier

Sanjana Srinivasan, Walter F Lenoir IV

Introduction/Background

Pancreatic ductal adenocarcinoma (PDAC) is one of the deadliest cancers, characterized by a median survival time of 6 months from time of diagnosis and a five year survival of 4%. The current course of treatment is typically gemcitabine, which does not shrink the tumor, but works to slow down tumor growth, and thus does not cure the patient [1]. Several studies have attempted to identify the subtypes of PDAC, the most widely recognized of which categorize into classical - which consists mostly of epithelial cells, quasimesenchymal - consisting of higher expression of mesenchyme genes, and exocrine-like, which demonstrates high expression of tumor cell derived digestive enzyme genes [1].

Research in our lab is focused on identifying targetable functional genetic vulnerabilities in PDAC that can be leveraged in drug development in a subtype specific manner. Thus, our models are limited to patient derived xenografts and cell lines. The exocrine-like subtype has been shown to be poorly characterized in these preclinical models owing to a large portion of the signal being driven by surrounding normal and stromal cells [1]. Thus, we mostly work with models falling into the epithelial or mesenchymal class.

Epithelial-mesenchymal transition (EMT) is a developmental process in cancer that enables cells to break away from their rigid structure and spread to distant sites. Epithelial cells adhere to one another forming a sheet-like architecture and have an apical-basal polarity. As cells undergo EMT, they tend to lose this architecture and polarity and take on a spindle-like morphology and become more motile [2]. EMT is a common feature in pancreatic cancer, and patients with tumors undergoing EMT face poorer prognosis, and disease free survival than those with less evidence of EMT. As the name EMT suggests, there exists a continuous spectrum, where tumors can be at varying stages of the transition [3]. Currently, quantification of EMT is done via biomarker staining, functional markers and RNA-seq. However, this does not allow for cell-by-cell quantification of EMT. Due to this, it is not possible to objectively measure *how* mesenchymal the line is – that is, the extent of the transition from epithelial to mesenchymal across different cell lines. Thus, once determination is done by eye if a sample appears to be mesenchymal, it is sequenced for confirmation. This tends to be both expensive and inefficient since several of the sequenced samples might not be as mesenchymal as would be required for further experimentation. By implementing a cell-by-cell tracking system for EMT, we would be able to track EMT along with growth of the cell-line, be able to identify sub clonal populations with distinct mesenchymal appearance, and ultimately, help understand tumor heterogeneity.

Hypothesis/Goal

We propose developing a machine learning model based on distinct morphological features of epithelial and mesenchymal cells that would actively track EMT in real time. We aim to develop a robust computational framework to identify and quantify the degree of EMT through analysis of cell lines using image analysis techniques and Matlab Machine Learning Apps to classify cells as Mesenchymal or Epithelial.

Approach and Methods

Approach - We attempted to classify images using two approaches. The first approach attempted to classify individual images as either completely mesenchymal or completely epithelial. The second approach was to classify individual cell groups (designated by the region prop function) as either mesenchymal or epithelial. We used metrics from our resulting data to create a model in the MATLAB classification learner application. Our training data was a set of 5 bright field images from one mesenchymal cell line and 9 bright field images from two epithelial cell lines. The training set consisted of a total of three publicly available cell lines with previously characterized and validated morphology. The test data consisted of seven images from each of three internally generated lines that have been characterized as epithelial, quasi-mesenchymal and mesenchymal through RNA-seq, which will serve as our “ground truth”.

Methods

Image Segmentation & Region Properties - We attempted to segment our images and take region properties of our resulting masks using tools from in class. These tools included edge masks, image closing, opening, dilation, erosion, etc. While these tools were successful with segmenting cells from background, they were insufficient in marking individual cells within cell colonies. Watershed methods and Ilastik additionally yielded insufficient results, as they did not accurately depict individual cell colonies [4].

An alternative method for image segmentation was tracking cells using image detection techniques [5]. We implemented this method by taking an edge mask of the normalized intensity of each of our images, and then taking the weighted centroid of the resulting edge mask. Edge mask techniques varied between 'sobel' and 'canny' methodologies. The resulting edge mask used, was decided by the optimal results of each individual unique image. This method yielded decent results, however could be further optimized. Each image had cell loss, and false cell identification, however this method was by far the best method we could implement. Once the cell mask and edge centroid mask was complete, we created a matrix of cell and image property metrics using the regionprops function. For our cell classification we selected 17 initial properties; Centroid, Weighted Centroid, Bounding Box, Mean intensity, Area, Perimeter, Eccentricity, EquivDiameter, Orientation, Euler Number, Convex Area, Extent, Area/Perimeter Ratio, Centroid Count per Mask, Centroid/Area, ConvexArea/Area, and Area/Centroid (Centroid/Area led to mostly zeroes). We decided to initially include centroid, bounding box, and weighted centroid as negative controls with our classification model. We found however, that these results significantly altered our classification learner, so they were omitted entirely in the development of the classification learner model. The last five properties were made to generate metrics that avoided image bias. For our image classification we selected 4 properties; Mean Connections (using bwconncomp), Minimum Centroid Distance, Mean Centroids Per Cell, Percent of Centroids in a Cell.

Model Selection - Model selection for our machine learning framework was performed using Classification Learner, a MATLAB machine learning app [6]. The Classification Learner app trains models to classify data in a supervised and automated manner training to search for the ideal classification model type. The parallel classifier training was applied and all models including decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, and ensemble classification were applied. As a first pass, we applied this method prior to feature selection to evaluate which methodologies maximize accuracy. Machine learning was performed on both cell classification - by combining all individuals cells with the image level classification in both the training and testing set, and on image classification - where each image was analyzed as one unit in both the training and testing set. Five-fold classification was used on all models, where the data was partitioned five times with one-fifth of the training data assigned to the testing set.

Feature Selection - Neighborhood Component Analysis (NCA) was performed using the training set to identify features that most contribute towards driving the classifier. Neighborhood component analysis (NCA) is a non-parametric supervised learning method for selecting features with the goal of maximizing prediction accuracy of any classification algorithm used. NCA computes utilizes leave-one-out classification by predicting class label of a single data point from the training set using the rest of the training set [7]. The average leave-one-out probability of correct classification is computed based on feature weights. NCA also utilizes lambda, a regularization parameter that maximizes the probability of correct classification by driving many of the feature weights to 0, thus identifying features that most contribute to accurate classification.

For the feature selection analysis, we first ran the NCA module within matlab using the default parameters to identify features with weights above zero. Following this, classification loss was evaluated using the training set and the testing set of images using the loss function in MATLAB. We next evaluated the optimal lambda value for our classifier that minimizes classification loss given our data. We used five-fold cross validation to partition our training data into five sets with one fifth of the data being assigned as a testing set in each of the five folds. Starting with lambda value of zero, we incrementally assessed the impact on classification loss at intervals of 1/(number of cells in training set) with an iteration limit of 30. The feature weights were standardized, with all the data being used to determine the best fit using the Stochastic Gradient Descent algorithm. This was computed for each of the five folds, and the average loss value across folds for each lambda value was computed. The lambda value that minimized the loss value was stored and NCA was rerun with the minimum lambda value and this features with weights more than zero were identified.

Model Selection with Features - Upon completion of the feature selection and narrowing down to our features of interest, we repeated the feature selection method using the parallel classifier training and all models. Metrics such as AUC value and precision were computed for all models and the best model based on these variables will be chosen.

Results

Image Classification

Image classification failed based on our four selected properties. Our training model had an accuracy of 79% but 12 out of 14 were selected to be epithelial. When the model was ran against our test data, 100% of tested images were labeled as mesenchymal, suggesting that there was significant differences between our image sets. For this reason, we limited our analysis to cell classification.

Cell Classification

Model Selection - Running our classification learner with 14/17 (not including centroid and bounding box properties) of the initial properties yielded an ensemble bagged tree model with a selection accuracy of 70.8%, and a resulting ROC curve of 0.67 (see figure 1 below). The classification learner was ran with a 5-fold classification in order to avoid overfitting. When ran against the test data, the mesenchymal cell line PATC53, yielded the highest percentage of mesenchymal cells. PATC124, a mixed cell line yielded the lowest percentage of mesenchymal cell lines. Table 1 below, displays the classification of the first pass test results.

Table 1: Counts of Epithelial vs Mesenchymal Cells within the Testing Set

Cell Line	Epithelial Count	Mesenchymal Count
PATC53 (Mesenchymal)	280	143
PATC69 (Epithelial)	610	205
PATC124 (Mixed)	227	59

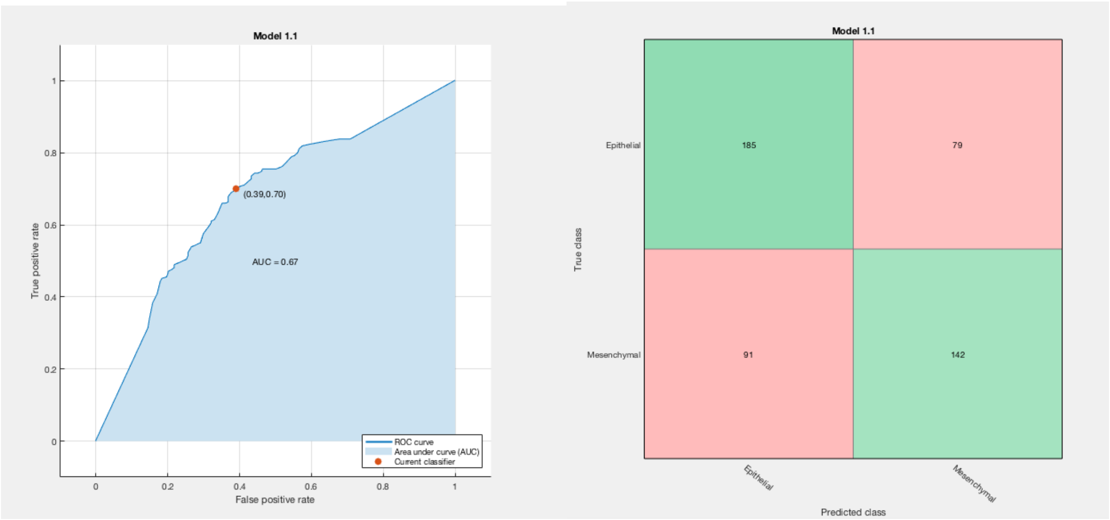

Figure 1: ROC Curve of initial classifier using five-fold cross validation of the training set only. Resulting AUC was 0.67, the true positive rate within epithelial and mesenchymal cells were 70% and 61% respectively, demonstrating poor stratification between the two classes when all features were applied to the model.

Feature Selection - The initial run through of the NCA algorithm using the default parameters identified area, orientation, perimeter and mean area per nucleus were identified as features of interest. However, the area and perimeter are subject to a certain degree of bias based on image resolution. Using five-fold cross validation, we identified that a lambda value of 0.0023 minimized the classification loss at a value of 0.2597. The association between lambda value and the classification loss is presented Figure 2. Following this, upon rerunning the NCA algorithm with a lambda value of 0.0023, the four features identified with feature weights above zero were mean intensity, mean centroids, area to perimeter ratio, mean area per nucleus. While this was not a parameter inputted, it is of note that these features are less likely to be impacted by image bias.

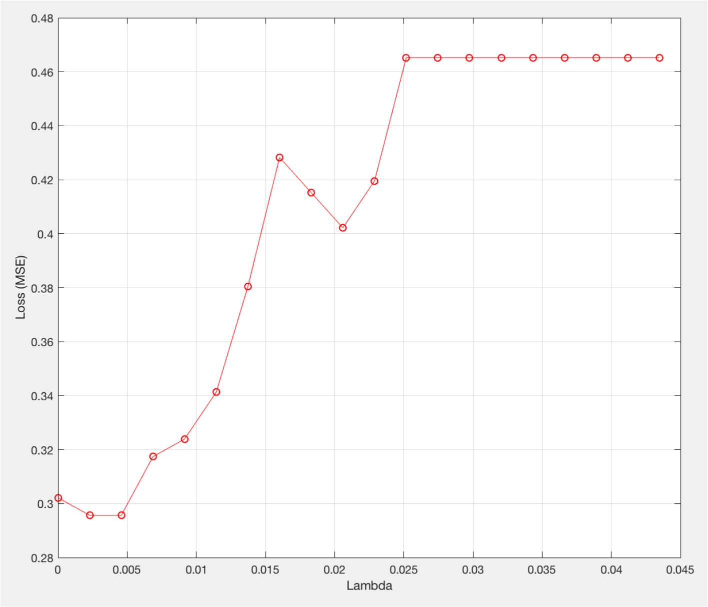

Figure 2: Evaluation of Lambda values and the corresponding classification loss demonstrates that this is minimized at a lambda value of 0.0023, which was applied to the NCA analysis to identify features with the largest weights suggesting they play a role in determining the accuracy of the class prediction.

Model Selection with Features - The classification learner was ran a second time with 4 features selected (see feature selection); Mean Intensity, Mean Centroid per Cell Mask, Area/Perimeter Ratio, and Area per Centroid. The selected model for the second pass was a quadratic support vector machine model with an selection accuracy of 72.4% using the training data (see figure 3). The classification learner was again ran with a 5-fold classification in order to avoid overfitting. When ran against the test data, there were increases of mesenchymal counts in PATC53, with increases of epithelial counts in PATC69 and PATC124 (see Table 2). This suggests that there was an overall improvement with classification of the PATC53 and PATC69 cell lines, and decreased performance of the classification of PATC124 cell line.

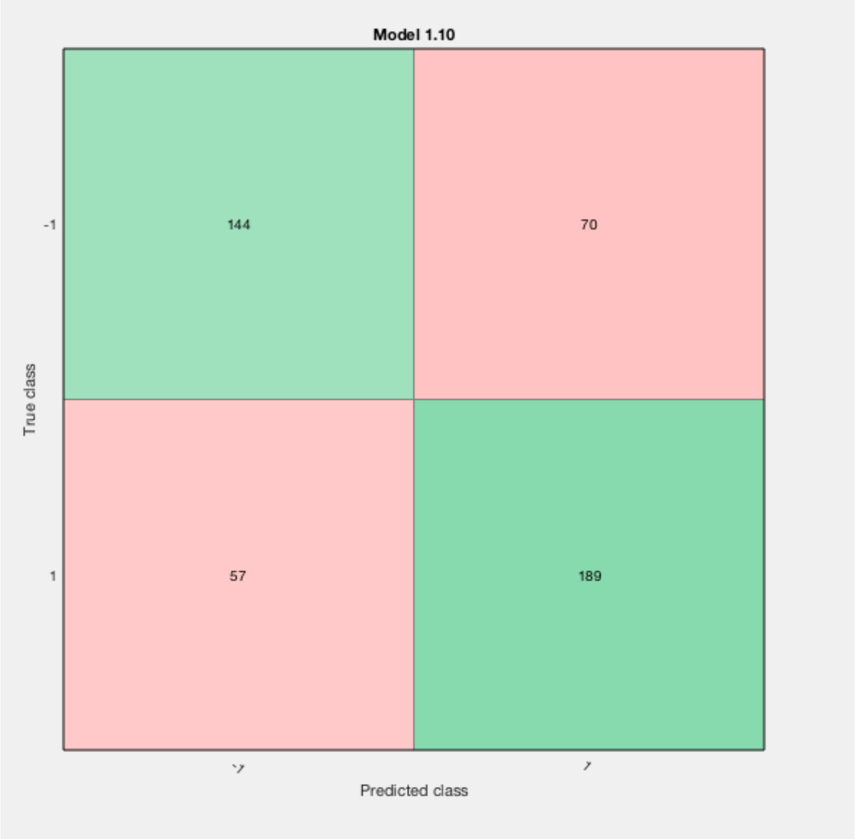

Figure 3: Table showing the concordance between true epithelial and mesenchymal classifications and predicted classes in the training set after limiting the model to features identified by NCA. Compared to figure 1, the true positive rate for epithelial and mesenchymal classification is 67% in epithelial cells and 77% in mesenchymal cells. -1 represents epithelial, while 1 represents mesenchymal.

Table 2: Improved Model Counts of Epithelial vs Mesenchymal Cells within the Testing Set

Cell Line	Epithelial Count	Mesenchymal Count
PATC53 (Mesenchymal)	253	170
PATC69 (Epithelial)	734	81
PATC124 (Mixed)	254	32

Table 2: Classification results of cell groupings in PATC53, PATC69, and PATC124 cell lines. Red indicates an increase in classification counts, and blue a decrease, when compared to the first classification learner.

Discussion

Our initial results suggest that classification and stratification of epithelial and mesenchymal cells using images is a feasible approach. Using only bright field images of commercially available cell lines with previously characterized morphology served as a good initial indicator of image properties that highlight the differences between epithelial and mesenchymal cells. The fact that we were able to show an improvement in the machine learning classification, which is especially seen to be the case in accurate classification of mesenchymal cells, suggests that feature selection to identify salient visual properties of these classes of cells is informative in delineating the two.

That being said, one limitation of this analysis is the quality of the images being analyzed. The strength of any machine learning model lies in the strength of the training data, which proved to be challenging in this analysis. Classification based on morphology alone is likely insufficient to train a model to predict the subtype. This is compounded by the difficulty we faced in segmenting our colonies into individual cells. Accuracy of image segmentation, and thus, of the model, can be improved by incorporating additional data such as nuclei staining and biomarker staining.

As previously mentioned, EMT is not a binary phenomenon but rather, occurs on a spectrum. Future directions include optimization of the machine learning process with the inclusion of more parameters that can serve as features. Enabling this would allow us to score EMT on a continuous scale based on the extent of mesenchymal cells present in different cell lines.

Division of Labor

Walter Frank Lenoir – Evaluated the classification learner and segmented the training data sets. Additionally, prepared code used for properties of (non-regionprops) cell and image classification (centroid distances, ratios, etc.). Combined and cleaned code.

Sanjana Srinivasan – Segmented images used for test data sets. Additionally spent time obtaining data. Wrote all code for and analyzed NCA results.

Even – Writing up the proposal & presentation preparation.

Works Cited

1. Collisson, E.A., et al., *Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy.* Nat Med, 2011. **17**(4): p. 500-3.
2. Dangi-Garimella, S., et al., *Epithelial-mesenchymal transition and pancreatic cancer progression*, in *Pancreatic Cancer and Tumor Microenvironment*, P.J. Grippo and H.G. Munshi, Editors. 2012: Trivandrum (India).
3. Moffitt, R.A., et al., *Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma.* Nat Genet, 2015. **47**(10): p. 1168-78.
4. C. Sommer, C.S., U. Köthe, F. A. Hamprecht, *ilastik: Interactive Learning and Segmentation Toolkit.* Eighth IEEE International Symposium on Biomedical Imaging (ISBI), 2011: p. 230-233
5. Choudhry, P., *High-Throughput Method for Automated Colony and Cell Counting by Digital Image Analysis Based on Edge Detection.* PLoS One, 2016. **11**(2): p. e0148469.
6. *MATLAB and Classification Learner Toolbox Release 2017a*. 2017, The MathWorks, Inc.: Natick, Massachusetts, United States.
7. Wei Yang, K.W., Wangmeng Zuo, *Neighborhood Component Feature Selection for High-Dimensional Data.* Journal of Computers, 2012. **7**(1): p. 161-168.