Project Title:

Recognition and Classification of human Embryonic Stem (hES) Cells' Differentiation Level

Team members

Dennis Won (jwon2014), Jeong Soo Sim (digivice)

Project Purpose

1. Given an image of hES cell culture, use learning algorithms to classify whether the recognized cell colonies are differentiated or not (Binary classification)
2. Perform multi-class classification to determine the level of differentiation of the recognized cell colonies
3. Spot the particular sections in the recognized colonies where there remains undifferentiated cell colonies
4. Provide a graphical display on the cell images demonstrating which colonies or which parts of the colonies are undifferentiated, and thus utilizable for medical research

Introduction

While human embryonic stem (hES) cells impose a huge potential in biomedical research discoveries to cure numerous diseases (such as Alzheimer's diseases, spinal cord injury, heart disease, etc.), the difficulty to maintain the "undifferentiated" state of hES cells hinders researchers to perform accurate and efficient experiments on hES cells. Human embryonic stem (hES) cells must be monitored and cared for in order to maintain healthy, undifferentiated cultures. In order to maintain the undifferentiated state of hES cells, the cultures must be continuously made sure to have no lost or lacking nutrients and be free of unwanted differentiation factors. In addition, although a small amount of differentiation is normal and expected in stem cell cultures, the culture should be routinely cleaned up by manually removing, or "picking" differentiated areas because those differentiated areas within the colony starts the chain-like reaction to differentiate others parts of the cultures, spreading the differentiation[1].

Having no technology to systematically maintain the undifferentiated state of hES cell culture, the stem cell scientists today completely rely on human eyes, human hands, and human decision to i) keep the favorable environment for the cell culture to remain undifferentiated, and to ii) identifying and removing excess differentiation from hES cell cultures to maintain the healthy population of cells. Due to the subjective nature of human decision based on human eyes and hands, the maintenance of healthy, undifferentiated hES cell culture has become one of the most difficult things to do for those scientists[1].

This particular project aims to use Machine Learning technologies (supervised learning algorithms) to perform AI recognition and binary classification of healthy, undifferentiated stem cell colonies from hES cell images with their differentiation levels. There are two main parts to this project.

1. Train an SVM and Logistic Regression model to learn how to classify between differentiated and undifferentiated hES colony. The input features for SVM training includes: the closeness of the shape of the colony to a perfect circle, the defined-ness of the edge of the colony, the degree of color intensity distribution within the colony, size of the colony, texture of the colony, etc.
2. Train SVM and Logistic Regression model to learn how to classify the differentiatedness of the each individual superpixel within the hES colony in order to localize the undifferentiated, healthy section within the colony.

Part I. Extracting feature vectors from the training set of hES cell colony images

This project involves a heavy load of image processing work that needs to done prior to training and testing the learning classifier in order to extract features from the images. We aggregated a total of 536 images of hES stem cell colony provided from Wu Cardiovascular Stem Cell Laboratory at Stanford Medical Center. Of the 536 images, exactly 268 images were differentiated hES cell colony images (positive: +1) and exactly other half 268 images were undifferentiated, healthy hES cell colony images (negative: -1). These images were taken by the same microscope with the same size scale and the same color intensity. Figure 1 shows the sample images of a positive colony image (a), and a negative colony image (b).

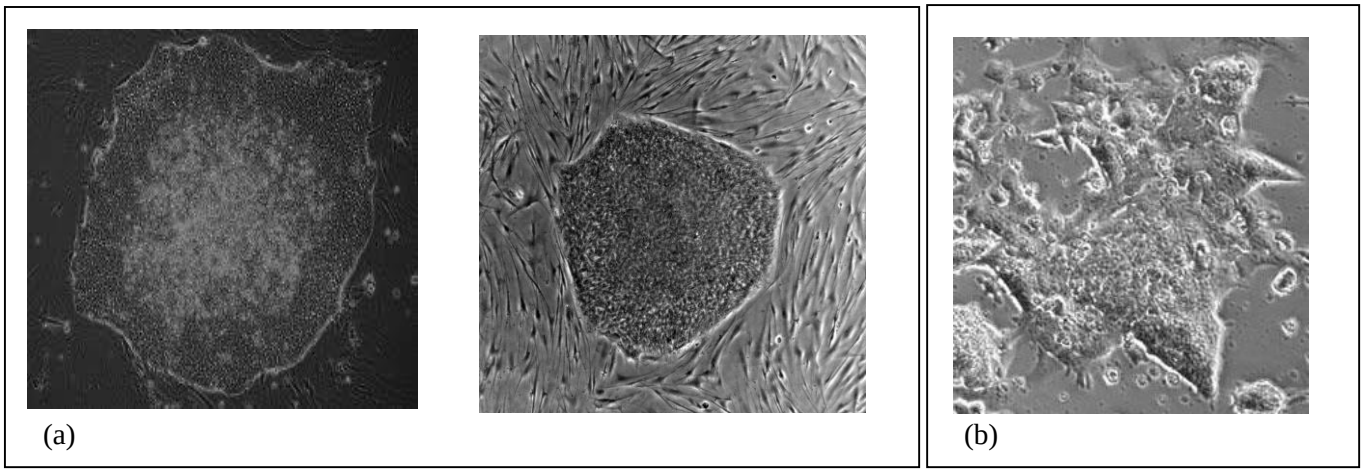


Figure 1. (a) Positive colony image for the differentiated colony. (b) Negative colony image for the undifferentiated colony. Note the morphological differences between the two.

I. Colony image feature extraction

OpenCV (Open Source Computer Vision Library) is a library of programming functions mainly aimed at real-time computer vision, developed by Intel. We used this OpenCV library to extract the following features of the hES colonies. Figure 2 shows the detection of the cell colony using the OpenCV edge and object detector, which also provides a variety of feature extraction methods. Figure 3 below displays the feature vector extracted from each hES cell colony image, along with the description of how individual features were measured and calculated.

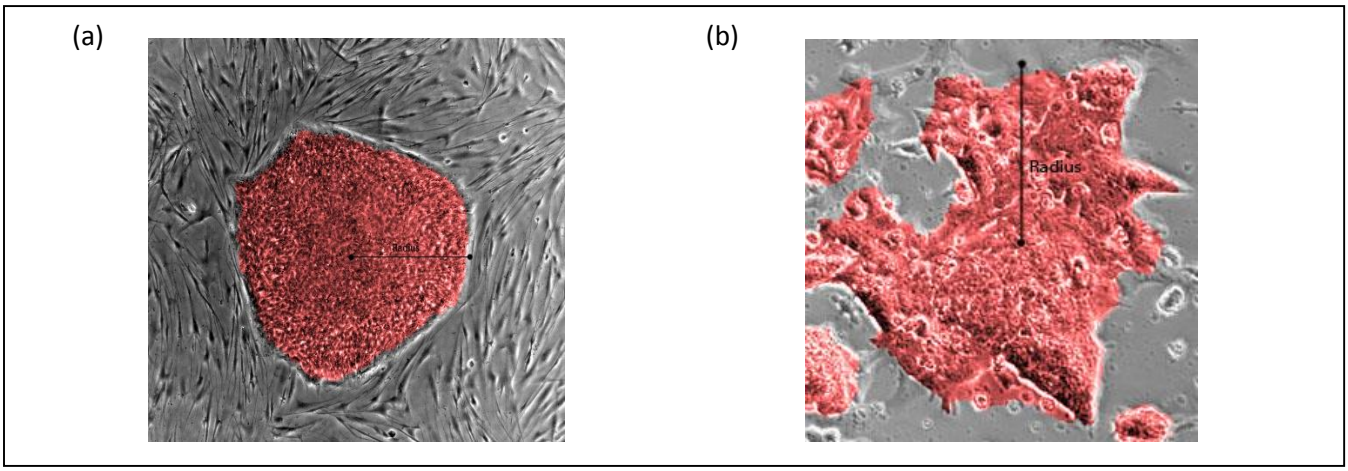


Figure 2. (a) Edge/object detection of the positive colony image for the differentiated colony. (b) Edge/object detection of the negative colony image for the undifferentiated colony. Note the morphological differences between the two.

Feature (Colony level)	Measurement
Shape closeness to perfect circle	OpenCV library to fit circular objects on the image. The closeness of the colony shape to a perfect circle determined by the OpenCV detector. Range: 0 <= closeness <= 1
Edge definition level	Edge definition level measured as the ratio of the number of edge pixels in the boundary of the colony to the total number of boundary pixels. The higher the ratio (number of edge pixels:total number of boundary pixels), the higher the edge definition is. Calculated using the OpenCV detector. Range: 0 <= edge definition level <= 1
Size	The size of the colony measured as the ratio of the radius of the edge-detected colony to the dimension of the cell image. Note that all cell colony images were taken under the same condition (size scale, light intensity, etc.) Range: 0 <= size <= 1
Texture density	Texture density of the cell colony measured as the *edgeness* per unit area, which is defined by F_edgeness = |{p | gradient_magnitude(p) > T}| / N where p is a pixel in the colony segment, T is some threshold, N is the total number of pixels in the colony segment, and gradient_magnitude computed from OpenCV detector. Range: 0 <= Texture density <= 1
Color intensity distribution	Color intensity distribution level measured by the standard deviation of the color intensity histogram for all pixels within the cell colony region divided by the range of the color intensity. Range: 0 <= Color intensity distribution <= 1

Figure 3. The feature vector extracted from each hES cell colony image

II. Superpixel feature extraction within a single colony

In addition extracting features of a colony to train SVM to perform binary classification of whether the colony is differentiated or undifferentiated in general, we also extracted features for each individual superpixels within each hES cell colony (superpixel was defined as 10x10 pixel-sqaure within the colony). A total of 121 samples of hES cell colonies were manually tagged for their undifferentiated region within them. Figure 4 displays the examples of the tagged (the red ROI) image input of hES colony, and Figure 5 shows the feature vector extracted from each superpixel (4 features extracted from each superpixel).

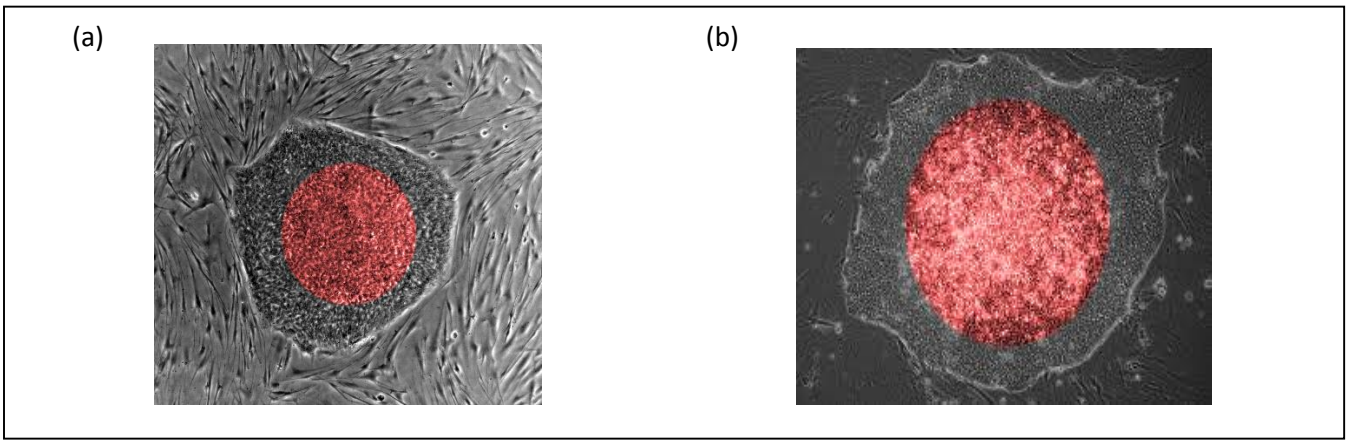
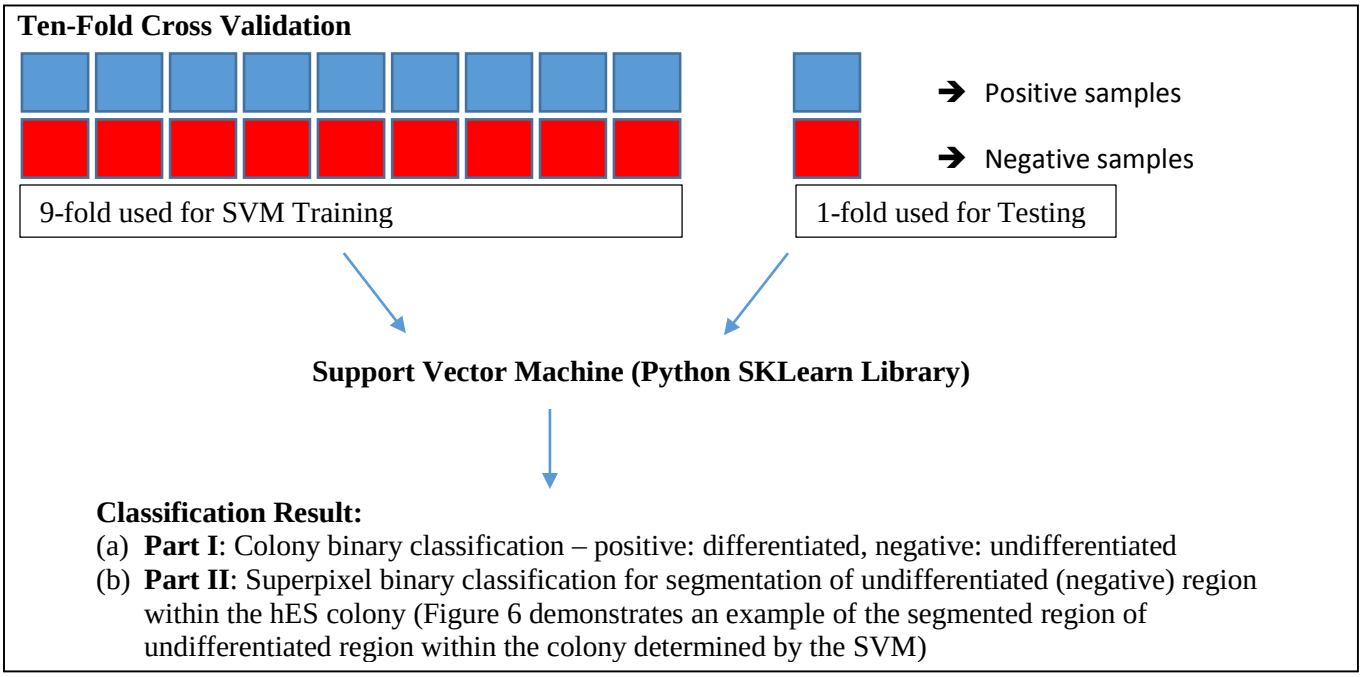


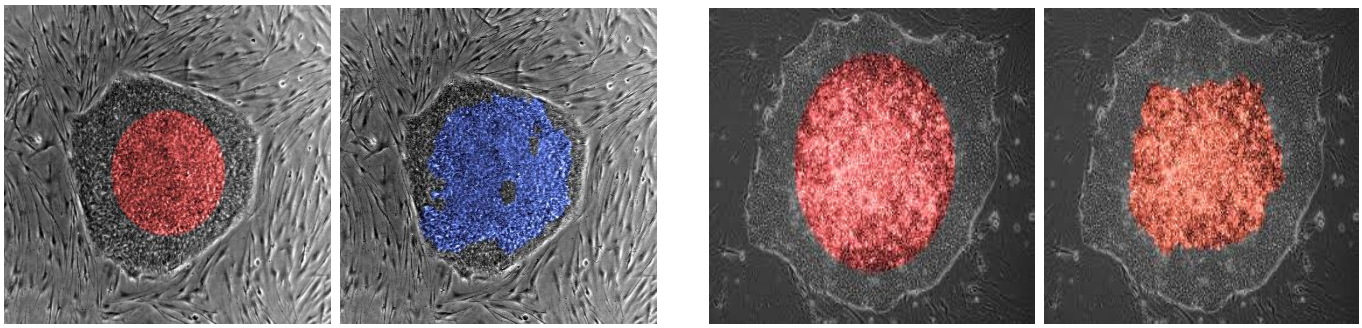
Figure 4. The undifferentiated region ROI of colony tagged input images. (Manually tagged)

Feature (Superpixel 10x10 level)	Measurement
Edge definition level	Edge definition level measured similar to the edge definition level from Figure 3, but measured from only within a particular superpixel. Range: 0 <= edge definition level <= 1
Distance from Centroid	The distance from Centroid measured as the ratio of pixel distance from the Centroid to the radius of the detected colony. Range: 0 <= size <= 1
Texture density	Texture density of the super pixel measured as the *edgeness* per unit area, similar to as described from Figure 3 feature vector for colony features. Range: 0 <= Texture density <= 1
Color intensity distribution	Color intensity distribution level measured by the standard deviation of the color intensity histogram for all pixels within the superpixel divided by the range of the color intensity. Range: 0 <= Color intensity distribution <= 1

Figure 5. The feature vector extracted from each superpixel with in the detected hES cell colony image

Part II. Training Support Vector Machine (SVM) and Logistic Regression Model with the extracted features





(a.1) Training (a.2) Test (b.1) Training (b.2) Test

Figure 6. The superpixel binary classification used to segment the undifferentiated (negative) region of the hES cell colony (a.2 and b.2 are the SVM-classification-induced ROI, while a.1 and b.1 are manually tagged ROI).

Part III. Result

I. Support Vector Machine on hES Colony – Colony level classification

For each of ten iterations of ten-fold Cross Validation, we recorded the positive and negative classified samples for both our prediction and the manually tagged result. By aggregating all iterations, we were able to deduce the following table.

		Real		
		+	-	total
Prediction	+	203	78	281
	-	65	190	255
	total	268	268	536

Figure 7. SVM prediction result using ten-fold Cross Validation

Using ten-fold Cross Validation for 10 times and aggregating every result gave us **73.3%** accuracy (Figure 7).

II. Support Vector Machine on Superpixel – Superpixel level classification

For each of ten iterations of ten-fold Cross Validation, we computed the numbers of predicted samples where the overlap between the predicted ROI of undifferentiated region and the manually tagged ROI is higher than 75% (that is, the ratio of the overlap between the two ROI's to the size of the manually tagged ROI exceeds 0.75). By aggregating all iterations, we found that our SVM achieves **68.5%** accuracy (82 out of 121 total samples had over 75% of its ROI predicted by our SVM model).

Discussion

While our SVM model to predict the differentiate-ness of the hES cell colony and the segmentation of the undifferentiated region in the colony achieved a degree of accuracy (around 75%), there are definitely sources or errors that hampered our study to be more accurate. First of all, our feature extraction and the normalization of the extracted features might not completely capture the differences between positive and negative samples. Most importantly, the size of our training set is not large enough to have achieve higher level of accuracy.

The future effort to increase the performance of the learning machine includes the following: first of all, use more data so that the machine gets trained better. Second, use more sophisticated image processing techniques to extract more features and also find out which features are the most important factors in determining the differentiated-ness of the colony. Lastly, we will be able to improve the segmentation algorithm used in this project by constructing more sophisticated and intuitive way to form superpixels. Currently, superpixel is constructed based on chosen constant. By allowing superpixels to have various size and shape, our analysis on the colony will have more accuracy since we will be able to more precisely depict the characteristics of colonies and cell.