Feature Identification for Colon Tumor Classification

Melody Lim, Anthony Hou, Natalie Congdon, and Janine Chua

Advisors: Dr. Fred Park, Dr. Ernie Esser, and Anna Konstorum
Interdisciplinary Computational and Applied Mathematics Program
University of California, Irvine
NSF PRISM grant DMS-0928427
October 8, 2013

Abstract. Hepatocyte Growth Factor (HGF) has been shown to be increased in the tumor microenvironment due to increased secretion by cancer-associated stromal cells. Qualitatively, high extracellular HGF has been correlated with increased growth and dispersiveness of a tumor. In this study, we develop quantitative methods to measure HGF-induced tumor growth and dispersion. Using image processing and machine learning techniques, we effectively classify images of colon cancer tumor spheroids cultured in +/-HGF conditions. Our goals are to define features that are effective for classification and to further help biologists quantify the effect of HGF on tumor spheroids.

1 Introduction

Hepatocyte Growth Factor (HGF) has been shown to be secreted by cancer-associated fibroblasts surrounding solid tumors, thereby increasing the HGF concentration in the colon tumor microenvironment *in vivo* [6, 9]. Moreover, increase in HGF correlates with the increased growth and dispersive behavior of colon tumor spheroids *in vitro* [15]. Currently, scientists can identify qualitative differences between tumor spheroids that have been exposed to HGF *in vitro* and those that have not, but are unable to quantify the effect. Due to increased interest in and use of such 3D *in vitro* tumor models, image processing tools have been developed for multicellular tumor spheroids, including size [12], shape [2], and invasion [16] analysis. In this study, we not only develop and utilize features for measuring spheroid growth and spread with respect to addition of HGF, but we also develop a computational method to select features that can discriminate between spheroids that have been treated with HGF and those that have not. The process of classification and feature selection can help to distinguish which biological phenomena are most altered in response to a chemical when this is not completely known *a priori*, allowing not just for a measure of a chemical's action on a spheroid, but also a better understanding of the biological action of the exogenous factor.

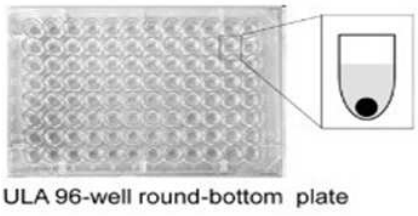
2 Methods

In this section, we will explore the experimental methods used to prepare the tumor spheroids for quantitative analysis. MATLAB and ImageJ were used to preprocess the data. Then

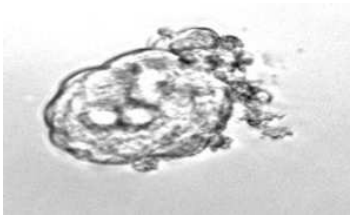
we selected features that we thought would be useful in classifying the two groups of tumor spheroids. We developed a classification scheme by using Fisher's linear discriminant.

2.1 Experimental Data

Experiments were carried out in the laboratory of Dr. Marian Waterman in the Department of Microbiology and Molecular Genetics at UC Irvine, in collaboration with Dr. John Lowengrub of the Department of Mathematics at UC Irvine. Colon cancer initiating cells (CCICs) were cultured as spheroids in ultra-low attachment flasks (Corning) using DMEM/F12 50:50 supplemented with N2, B12, EGF, bFGF, heparin, sodium pyruvate, and penicillin/streptomycin [13]. Unlike typical cell lines, CCICs are multipotent and capable of regenerating heterogeneous tumors with characteristics analogous to those found in primary tumors, from which they are derived [13, 11]. CCICs were trypsinized using a no-serum trypsin inhibitor. Single cells were counted and plated in 96 well ultra-low attachment plates (Corning) using the previously mentioned media with or without HGF at various concentrations. CCICs were imaged at 10x resolution once each day, see Figure 1.



(a) Single cells plated in 96 well ultra-low attachment plates with DMEM.



(b) Spheroid grown in media +50ng/ml HGF, day 8.

Figure 1: Illustrating spheroid generation procedure for the image processing in this study.

A total of 4 time-series for control experiments and 9 time-series for +HGF experiments were available for analysis. In order to increase the size of the data set for the classification procedure, 20 images from the control time-series and 22 images from the +HGF time-series were used. Images were considered to be independent of each other due to variability in plane of focus and brightness during image acquisition. Moreover, since spheroids were cultured in ultra-low attachment flasks, movement and rotation of spheroids resulted in additional intra-image variability within a time-series.

2.2 Preprocessing the Data

In order to focus on the main tumor, we hand-segmented the images using the image processing program ImageJ, since the petri dish and dead cells were not directly connected to the main tumor (Figure 2 (a), top panel). Thresholding was used to convert the segmented images into binary images, and to detect points that were not part of the main cluster of cells. To reduce the salt and pepper noise but still preserve the edges, we applied the 2-D

median filter MATLAB function *medfilt2* to the binary images. We then utilized the MATLAB function *bwboundaries* to find the exterior boundaries of the tumor and applied these boundaries to the binary images (Figure 2).

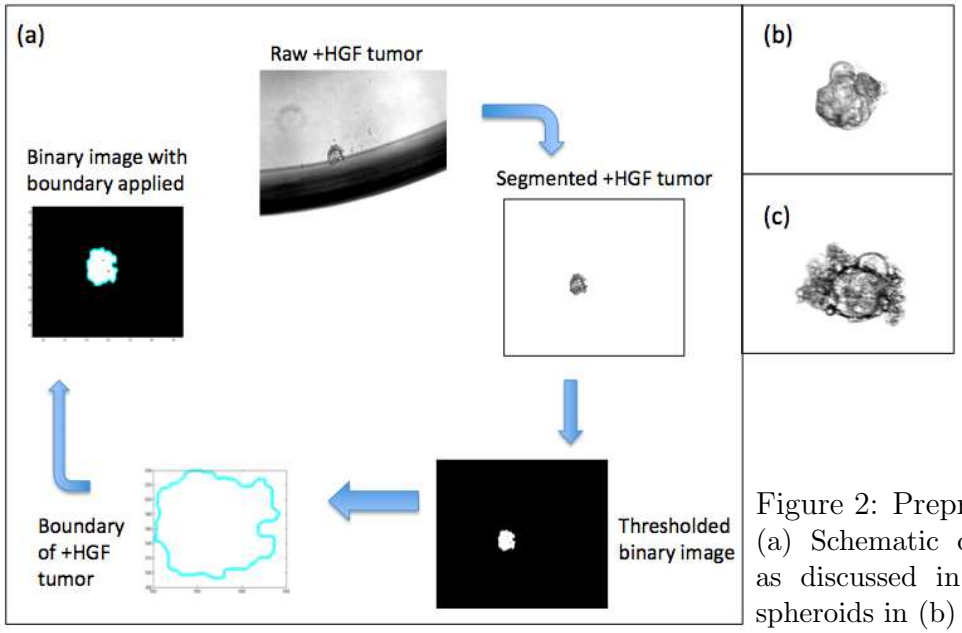


Figure 2: Preprocessing the images
(a) Schematic diagram of preprocessing steps, as discussed in the text. Sample segmented spheroids in (b) control and (c) +HGF media.

2.3 Features Considered

We considered one given feature, time, and six mathematical features which can be described as geometry- and intensity-based features:

- Area
- Perimeter-to-Area Ratio
- Eccentricity
- Circularity Ratio
- Total-Variation-to-Area Ratio
- Average Intensity

We included time as a feature because the tumor spheroids from different classes might appear similar at different stages. Thus, time information would be needed to distinguish between the two groups. The geometry-based features are area, perimeter-to-area ratio, eccentricity, and circularity ratio. Instead of using only perimeter as a feature, we divided the perimeter by area using the hypothesis that a tumor with a large perimeter-to-area ratio will likely have a more jagged boundary, while a tumor with a small perimeter-to-area ratio will likely have an edge that is smooth. Eccentricity is defined as the ratio of the eigenvalues of the covariance matrix that corresponds to a binary image of the shape. To calculate eccentricity, we used the eccentricity argument from the MATLAB function regionprops,

which determines the ratio of the distance between the foci of the ellipse fitted to the shape and its major axis length. The output is between 0 and 1, with 0 representing a circle and 1 a line [8]. Circularity ratio is defined by

$$C_1 = (A_s)/(A_C)$$

where A_s is the area of the shape and A_C is the area of the circle having the same perimeter as the shape [17]. We considered circularity ratio since by the dispersive action of HGF, control tumor spheroids could be more similar to circles than spheroids grown in high HGF conditions. These geometry-based features were applied to the binary images. The intensity-based features were the total-variation-to-area ratio and average intensity. Total variation is defined by

$$\sum_i \sum_j \sqrt{(u_{i+1,j} - u_{ij})^2 + (u_{i,j+1} - u_{ij})^2}$$

where u_{ij} is the segmented image with the original intensities. Total variation can be used to indicate when the images have a textured appearance, which would likely correspond to a highly variable density. The discrete total variation of an image is the summation of the magnitudes of intensity jumps between neighboring pixels. Since the image intensity is inversely related to cell density, high total variation indicates a cluster of cells where the density greatly varies. Average intensity is the sum of the image intensities over the shape divided by the area and is inversely related to density. Smaller values indicate less light passing through the tumor, which suggests a denser object. Using the raw intensity data, these intensity-based features were applied to the segmented images.

2.4 Classification Procedure

The purpose of the classification approach is to learn how to effectively distinguish the control from the +HGF group using the features considered. This is used to provide insight into which features and combinations of features are more discriminating. Before classifying the data, the images are represented as feature vectors and can be thought of as points in R^n where n is the number of features. The data is separated into training and testing sets. The training set is used to learn the classifier, which is then evaluated on the testing set. Throughout our test trials with various combinations of features, we used Fisher's Linear Discriminant (FLD), which can be described as a specific choice of direction for projection of the data in one dimension [1]. This direction can be thought of as the normal to the hyperplane which best separates the two classes. Given the training set, we used FLD to learn an optimal hyperplane for separating the feature vectors for the control and +HGF groups. FLD computes a vector of weights w in R^n that is used to define a linear discriminant function of the form $y(x) = w^T x + y_0$, where y_0 is the threshold chosen to minimize classification error on the training set. This function can be used to classify a feature vector x as being control or +HGF by checking whether $y(x) > 0$ or $y(x) < 0$. FLD computes w to maximize the ratio of the inter-class variance to intra-class variance of the training feature vectors after projecting onto the line spanned by w. The idea is to maximize

a function that will give a large separation between the two class means while giving a small variance within each class [1]. The linear classifier can be described as defining a hyperplane consisting of all points x such that $y(x) = 0$, where w is perpendicular to this hyperplane. New points are classified according to which side of the hyperplane they are on.

Classification results were computed by using Repeated Random Sub-Sampling Cross Validation, during which we randomly selected half of our data to form the training set and then designated the remaining half as our testing set. Through computing FLD on our training data, we were able to define the function y and then measure how accurately y classified the data in the testing set. We repeated this cross validation procedure fifteen times consecutively for the set of features, averaging fractions correctly classified each time. By applying this classification scheme to different subsets of features of the colon tumor spheroids grown in media with and without HGF, we were able to select the features that were most useful for classification.

3 Results

After selecting the features, we used Fisher's linear discriminant on various sets of features to determine whether it produced a clear distinction between the control and +HGF group. First, we ran the classification procedure using six features. Adding time as another feature gave us enough information to effectively classify the data. Thus, we repeated the classification procedure using these seven features. We then decided to use smaller subsets of features to find the most effective features.

3.1 Classification of the Six Features

We first ran the FLD code on all six features: area, circularity ratio, average intensity, eccentricity, perimeter-to-area ratio, and total-variation-to-area ratio. We obtained a 91.50% classification accuracy for the control group and a 90.99% classification accuracy for the +HGF group. Next, we attempted to obtain the same results utilizing fewer features. To identify the classification capabilities of fewer features, we ran the classification scheme for each feature separately (Figure 3), and in a number of feature combinations (Table 1).

The area and perimeter-to-area ratio features classified most of the tumor spheroids while the circularity ratio, eccentricity, average intensity, and total variation-to-area ratio features were not as effective. The area and perimeter-to-area ratio features were most effective because of the area and perimeter differences between the control and +HGF groups. Due to the relatively circular tumor spheroids from both groups, the circularity ratio and eccentricity features were not as effective. The average intensity and total-variation-to-area ratio features also performed poorly because there were some overlapping of densities between the +HGF and control groups.

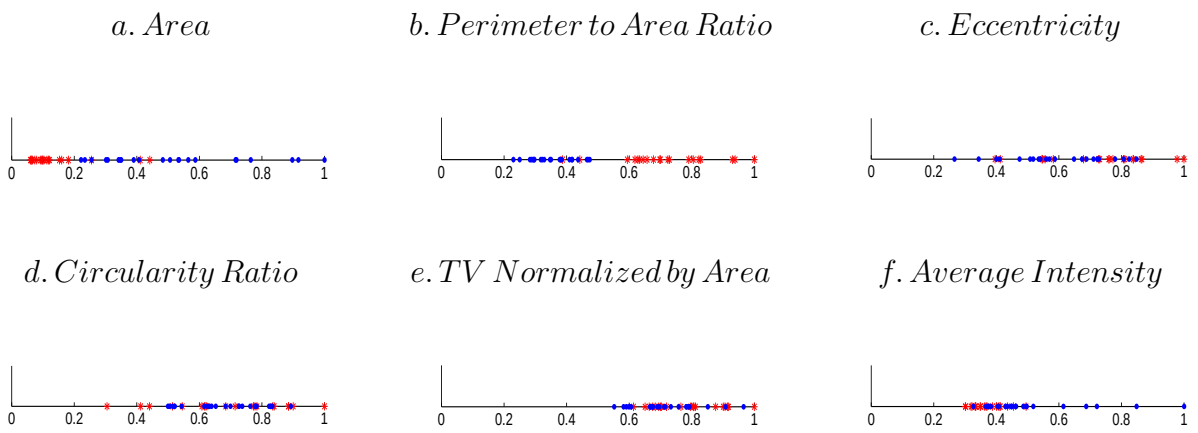


Figure 3: Classification results for individual features. The red and blue dots represent the +HGF and control group, respectively.

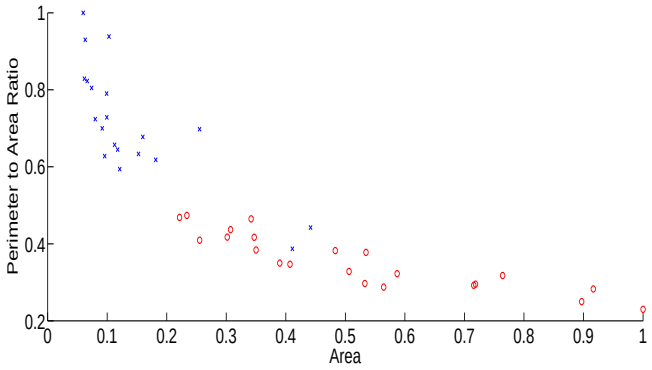
3.2 Addition of Time Component

We were able to increase classification accuracy by including the time (day) as a feature. The day descriptor was able to separate spheroids with similar shape and size into different groups. These similarities occurred since a few larger control tumor spheroids from a later phase had similar areas and perimeters to earlier-stage +HGF tumor spheroids. Without adding time as a feature, it would be difficult to distinguish between the two classes. Thus, our set of seven features included the original six features along with the addition of the day feature, and had a 98.88% classification accuracy for the control group and a 100% classification accuracy for the +HGF group.

3.3 Selection of Best Features

Based on the individual results, we combined the two strongest features, area and perimeter-to-area ratio, and plotted them both using a scatter plot to obtain an idea of their classifying potential.

Indeed, using just these two features, we managed to classify the majority of the tumor spheroids. If we were to draw a line separating the two groups, only two of the control tumor spheroids would be inaccurately categorized. After running the FLD code, we acquired a 89.03% classification accuracy for the control group and a 96.92% classification accuracy for the +HGF group. Since the area and perimeter-to-area features were the two strongest features, combining the time descriptor with these two other features gave us a 100% classification accuracy for both the control and the +HGF groups. We investigated whether the highly effective features prevented us from detecting whether the other features had any potential as classifiers. Thus, we grouped the less effective features (circularity ratio, total-variation-to-area ratio, average intensity, and eccentricity) into their own set of features and



(a) Area vs perimeter-to-area ratio

then ran the FLD code. These features classified 75.33% for the control group and 55.27% for the +HGF group. Hence, the less effective features were not useful, even in combination, to effectively classify the data. However, we note that since different experimental conditions may yield spheroids with different textures and shapes, these classifiers could be more effective in alternative conditions.

4 Discussion

Three-dimensional *in vitro* experimental models of tumor growth are beginning to supplant more traditional monolayer cell culture models as the former better replicate the spatial dynamics of the *in vivo* environment in which tumors develop [4]. But, analysis of experiment outcomes is difficult, as spheroids grown using different cell types and under different conditions can exhibit a variety of growth morphologies and cell spread [16]. Hence, image processing tools can play a large role in furthering outcome analysis as they can help to quantify the complex shape geography that spheroids possess. In this study, we used images from experiments on tumor spheroids composed of colon cancer initiating cells (CCICs) grown in media with HGF, an extracellular molecule that is present in the tumor microenvironment and is associated with tumor growth and spread [9].

We developed a set of image-based features that can be used to measure phenotypic changes in *in vitro* tumor spheroids grown in different conditions. We identified the features that differ most strongly between spheroids grown in control vs. +HGF media. Specifically, our results show that circularity ratio, total variation-to-area ratio, eccentricity, and average intensity were not effective classifiers for our unique data set. Solely based on computational results, the area and perimeter-to-area ratio, in conjunction with the day feature, were the most effective features. The effectiveness of the area feature is in accordance with the biological hypothesis that HGF increases cellular mitosis rate via activation of the c-MET receptor, which is over-expressed on colon cancer cells, and leads to nuclear localization of cytosol-anchored β-catenin, where it can potentiate cellular clonogenic activity [14, 3].

Number of Features Used	Features	Average Percentage Correct for Control	Average Percentage Correct for +HGF
3	Average Intensity, Circularity Ratio, Total Variation to Area Ratio	75.33%	55.27%
4	Average Intensity, Circularity Ratio, Total Variation to Area Ratio, Day	81.73%	89.04%
3	Area, Eccentricity, Perimeter to Area Ratio	92.29%	91.79%
2	Area and Perimeter to Area Ratio	89.03%	96.92%
3	Area, Perimeter to Area Ratio, Day	100%	100%
6	Area, Average Intensity, Circularity Ratio, Eccentricity, Perimeter to Area Ratio, Total Variation to Area Ratio	91.5%	90.99%
7	Area, Average Intensity, Circularity Ratio, Eccentricity, Perimeter to Area Ratio, Total Variation to Area Ratio, Day	98.88%	100%

Table 1: Comparisons of the feature classification results using different feature combinations

Moreover, the accuracy of the perimeter-to-area feature, which quantifies contiguous cell spread in the sense that the cells are leaving the tumor spheroid, verifies the well-known characteristic of HGF as an inducer of cell scatter and motility [5]. In the 3D spheroid, HGF can maximally act on the periphery of the spheroid, causing cells at the tumor-host interface to dissociate from the body of the tumor. This induced phenotype results in an image of a spheroid with a greater perimeter-to-area ratio than in control conditions. The finding that addition of the day feature increased classification accuracy can be attributed to the fact that given a relatively longer period of time, tumor spheroids can achieve growth and spread characteristics that are very similar to the phenotype seen in spheroids grown for a shorter period of time with HGF.

It was surprising to us that the circularity ratio did not serve as an efficient classifier. We initially hypothesized that +HGF spheroids would have a lower circularity ratio than the controls based on our knowledge the dispersive effects of HGF. But our calculations showed that the spheroids did not lose circularity in presence of HGF (Figure 3(d)). This failed classification feature showed us that the tumor spheroids' growing pattern with added HGF

was evenly dispersed among the perimeter of the spheroid, indicating that the proliferative effect of HGF outweighed the dispersive effect in the context of the spheroids.

In order to recover more data from the experiment, it will be necessary in future work to develop quantitative methods to analyze cell spread for cells no longer attached to the tumor. There have been several techniques developed, including segmentation of invading cells and calculation of scattered tumor area [16], calculation of nearest-neighbor distance measures between cells [10], and cellular motility [7]. As in our current work, we will develop several features that will measure different aspects of cell spread, and use these features to identify which feature is most altered in +HGF spheroids vs. controls. Additionally, we plan to develop better segmentation schemes for both the tumor and invading cells in heterogeneous media. For the current study, we were not able to derive a scheme that was able to effectively segment the main tumor from the background. Potentially, we would rerun the experiments, paying better attention to providing a homogeneous background in the images. In that regard, future experiments will allow us to obtain more metrics on HGF action via the following changes in experimental design: (i) run experiments with varying concentrations of HGF, to determine how increase in [HGF] correlates to the various feature outputs that we developed and (ii) stain the spheroids for markers of stemness. Recent data has shown that spheroids grown in high HGF conditions have increased stem cell populations at the tumor-host boundary [15]. An ability to visualize this outcome would provide another feature that would improve classification accuracy and give greater insight into HGF-driven tumor growth. Finally, we would also work to directly incorporate the time course data by considering how our selected features may change over time under different conditions. For example, change in the area feature over time would result in a growth rate for each experiment that can be used as a feature itself to classify the different conditions.

In summary, we have developed a robust feature-selection and classification scheme for use on tumor spheroids experiments that has the potential to be developed in parallel with experimental design to better quantify, and hence assess and understand, action of external agents on tumor growth.

5 Acknowledgements

We would like to thank Dr. Jack Xin, Dr. Hongkai Zhao, Dr. Sarah Eichhorn, Dr. Frederick Park, Dr. Ernie Esser, Anna Konstorum, Dr. Marian Waterman from the Department of Microbiology and Molecular Genetics, Dr. Lowengrub from the Department of Mathematics at UC Irvine, Stephanie Sprowl, and the NSF.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[2] Gang Cheng, Janet Tse, Rakesh K Jain, and Lance L Munn. Micro-environmental me-

chanical stress controls tumor spheroid size and morphology by suppressing proliferation and inducing apoptosis in cancer cells. *PLoS One*, 4(2):e4632, 2009.

[3] Riccardo Fodde and Thomas Brabletz. Wnt/beta-catenin signaling in cancer stemness and malignant behavior. *Curr Opin Cell Biol*, 19(2):150–158, Apr 2007.

[4] Juergen Friedrich, Reinhard Ebner, and Leoni A Kunz-Schughart. Experimental anti-tumor therapy in 3-d: spheroids–old hat or new challenge? *Int J Radiat Biol*, 83(11-12):849–71, 2007.

[5] Stefan Grotegut, Dietrich von Schweinitz, Gerhard Christofori, and Francois Lehembre. Hepatocyte growth factor induces cell scattering through mapk/egr-1-mediated upregulation of snail. *EMBO J*, 25(15):3534–3545, Aug 2006.

[6] Annette Kaminski, Jens Claus Hahne, El-Mustapha Haddouti, Alexandra Florin, Axel Wellmann, and Nicolas Wernert. Tumour-stroma interactions between metastatic prostate cancer cells and fibroblasts. *Int J Mol Med*, 18(5):941–950, Nov 2006.

[7] Dinah Loerke, Quint le Duc, Iris Blonk, Andre Kerstens, Emma Spanjaard, Matthias Machacek, Gaudenz Danuser, and Johan de Rooij. Quantitative imaging of epithelial cell scattering identifies specific inhibitors of cell motility and cell-cell dissociation. *Sci Signal*, 5(231):rs5, Jul 2012.

[8] 2013 MathWorks. *Image Processing Toolbox: User's Guide (R2013b)*, Retrieved September 15 2013.

[9] Kunio Matsumoto and Toshikazu Nakamura. Hepatocyte growth factor and the met system as a mediator of tumor-stromal interactions. *Int J Cancer*, 119(3):477–483, Aug 2006.

[10] Melissa D Pope, Nicholas A Graham, Beijing K Huang, and Anand R Asthagiri. Automated quantitative analysis of epithelial cell scatter. *Cell Adh Migr*, 2(2):110–6, 2008.

[11] Lucia Ricci-Vitiani, Dario G Lombardi, Emanuela Pilozzi, Mauro Biffoni, Matilde Todaro, Cesare Peschle, and Ruggero De Maria. Identification and expansion of human colon-cancer-initiating cells. *Nature*, 445(7123):111–5, Jan 2007.

[12] Bjoern Rodday, Franziska Hirschhaeuser, Stefan Walenta, and Wolfgang Mueller-Klieser. Semiautomatic growth analysis of multicellular tumor spheroids. *J Biomol Screen*, 16(9):1119–24, Oct 2011.

[13] Shaheen S Sikandar, Kira T Pate, Scott Anderson, Diana Dizon, Robert A Edwards, Marian L Waterman, and Steven M Lipkin. Notch signaling is required for formation and self-renewal of tumor-initiating cells and for repression of secretory cell differentiation in colon cancer. *Cancer Res*, 70(4):1469–78, Feb 2010.

[14] Hiroya Takeuchi, Anton Bilchik, Sukamal Saha, Roderick Turner, David Wiese, Maki Tanaka, Christine Kuo, He-Jing Wang, and Dave S B Hoon. c-met expression level in primary colon cancer: a predictor of tumor invasion and lymph node metastases. *Clin Cancer Res*, 9(4):1480–8, Apr 2003.

[15] Louis Vermeulen, Felipe De Sousa E Melo, Maartje van der Heijden, Kate Cameron, Joan H de Jong, Tijana Borovski, Jurriaan B Tuynman, Matilde Todaro, Christian Merz, Hans Rodermond, Martin R Sprick, Kristel Kemper, Dick J Richel, Giorgio Stassi, and Jan Paul Medema. Wnt activity defines colon cancer stem cells and is regulated by the microenvironment. *Nat Cell Biol*, 12(5):468–476, May 2010.

[16] Maria Vinci, Sharon Gowan, Frances Boxall, Lisa Patterson, Miriam Zimmermann, William Court, Cara Lomas, Marta Mendiola, David Hardisson, and Suzanne A Eccles. Advances in establishment and analysis of three-dimensional tumor spheroid-based functional assays for target validation and drug evaluation. *BMC Biol*, 10:29, 2012.

[17] Ronsin Joseph Yang Mingqiang, Kpalma Kidiyo. A survey of shape feature extraction techniques. *Pattern Recognition*, pages 43–90, 2008.