Towards Automated Cellular Image Segmentation for RNAi Genome-Wide Screening

Xiaobo Zhou[1,2], K.-Y. Liu[1,2], P. Bradley[3], N. Perrimon[3], and Stephen TC Wong[1,2,*]

[1] Center for Bioinformatics, Harvard Center for Neurodegeneration and Repair, Harvard Medical School, 3rd floor, 1249 Boylston, Boston, MA 02215
2 Department of Radiology, Brigham and Women's Hospital, Boston, MA 02115
3 Department of Genetics, Howard Hughes Medical Institute, Harvard Medical School, MA 02115, USA
wong@crystal.harvard.edu

Abstract. The Rho family of small GTPases is essential for morphological changes during normal cell development and migration, as well as during disease states such as cancer. Our goal is to identify novel effectors of Rho proteins using a cell-based assay for Rho activity to perform genome-wide functional screens using double stranded RNA (dsRNAs) interference. We aim to discover genes could cause the cell phenotype changed dramatically. Biologists currently attempt to perform the genome-wide RNAi screening to identify various image phenotypes. RNAi genome-wide screening, however, could easily generate more than a million of images per study, manual analysis is thus prohibitive. Image analysis becomes a bottleneck in realizing high content imaging screens. We propose a two-step segmentation approach to solve this problem. First, we determine the center of a cell using the information in the DNA-channel by segmenting the DNA nuclei and the dissimilarity function is employed to attenuate the over-segmentation problem, then we estimate a rough boundary for each cell using a polygon. Second, we apply fuzzy c-means based multi-threshold segmentation and sharpening technology; for isolation of touching spots, marker-controlled watershed is employed to remove touching cells. Furthermore, Voronoi diagrams are employed to correct the segmentation errors caused by overlapping cells. Image features are extracted for each cell. K-nearest neighbor classifier (KNN) is employed to perform cell phenotype classification. Experimental results indicate that the proposed approach can be used to identify cell phenotypes of RNAi genome-wide screens.

1 Introduction

High content screening by automated fluorescence microscopy is becoming an important and widely used research tools to assist scientists in understanding the complex cellular processes such as mitosis and apoptosis, as well as in disease diagnosis and prognosis, drug target validation, and compound lead selection [1]. Using images

* Corresponding author.

J. Duncan and G. Gerig (Eds.): MICCAI 2005, LNCS 3749, pp. 885 – 892, 2005.

acquired by automated microscopy, biologists visualize phenotypic changes resulting from reverse-functional analysis by the treatment of Drosophila cells in culture with gene-specific double-stranded RNAs (dsRNAs) [2]. In a small scale study by manual analysis [3], biologists were able to observe a wide range of phenotypes with affected cytoskeletal organization and cell shape. Nonetheless, without the aid of computerized image analysis tools, it becomes an intractable problem to characterize morphological phenotypes quantitatively and to identify genes as well as their dynamic relationships required for distinct cell morphologies on a genome-wide scale.

In this paper, we will study image-based morphological analysis in automatic high-content genome-wide RNAi screening for novel effectors of Rho family GTPases in Drosophila cells. About 21,000 dsRNAs specific to predicted Drosophila genes are robotically arrayed in 384-well plates. Drosophila cells are plated and take up dsRNA from culture media. After incubation with the dsRNA, expression of Rac1V12, RhoAV14 or Cdc42V12 is induced. Cells are fixed, stained, and imaged by automated microscopy. Each screen will generate ~400,000 images, or millions if including replicates. Clearly, there is a growing need for automated image analysis as high throughput technologies are extended to visual screens. Biologists have developed a cell-based assay for Rho GTPase activity using the Drosophila Kc167 embryonic cell line. Three-channel images are obtained by labeling F-actin, GFP-Rac and DNA. Fig. 1 gives an example of RNAi cell images of one well acquired with three channels for phenotypes of (a) DNA, (b) Actin, and (c) Rac. It is tough to segment the cells from (b) or (c). The three phenotypes are shown in Fig. 2. They are: S-spikey, R-ruffling, and A-actin acceleration at edge. The question is how to identify the three phenotypes automatically for each image. To reach this aim, we propose the following three steps: first, we segment each cell, then we calculate the morphological and textural features for each cell and built training data sets, finally we perform feature reduction and classify cellular phenotypes.

The key issue is how to automatically segment cells of cell-based assays in a cost-effective manner, as such fast screening generate rather poor image quality and tens and hundreds of millions of cells in each study. There exist a number of publications on nuclei segmentation and cell segmentation. For example, Wahlby, et. al., [4] proposed a cytoplasm segmentation based on watershed segmentation and rule-based merging and splitting of over-segmented and under-segmented objects. Marker-controlled watershed segmentation is a popular method in cell segmentation [5-7]. In the literature, watershed methods with or without seeds are extensively studied. Although the oversegmentation caused by watershed can be reduced by rule-based merging of fragmented objects, it is difficult to devise reliable rules to merge the example which consists of one cell with three nuclei inside the cytoplasm. Lindblad, et. al., recently studied a similar problem about automatic segmentation of cytoplasms and classification of Rac1 activation [7]. There are several different points between our work and Lindblad's work. First, their data source is Chinese hamster ovary hIR (human insulin receptor) cells transfected with GFP-Rac1 reporter protein, and ours is Drosophila Kc167 embryonic cell line transfected with an inducible GFP-RacV12 protein. Second their data is two-channel (nucleus and GFP-Rac1 channels) 3D images, while our data is three-channel (DNA, F-actin and GFP-RacV12 channels) 2D images from larger scale genome-wide screening. Third, the quality of their images is better as they employed automated confocal laser scanning microscopy (see Fig.2 in [7]) while we used more commonly available, standard automated epi-fluorescence

microscopy. So it is much more challenging to segment RNAi genome-wide images. To address this hard problem, we propose a two-step segmentation approach. We then quantitate the tens of millions of cells and classify the cell phenotypes.

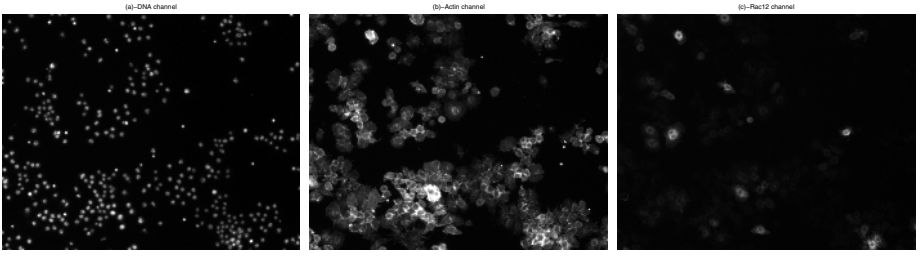

Fig. 1. RNAi cell images with three channels

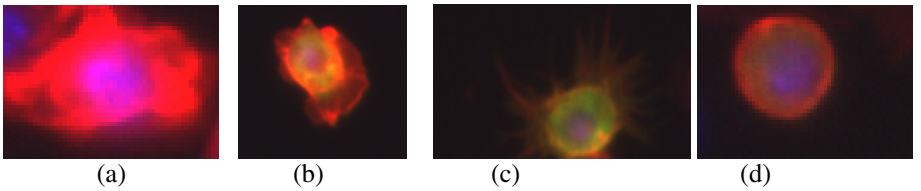

Fig. 2. Four different RNAi cell phenotypes: (a)-A-A, (b) N, (c) R and (d)

2 A Two-Step Segmentation Approach

Extracting rough boundary for each cell is the first step of our approach. It consists of two sub-steps: determine the center of each cell and determine the polygon for each cell. We then propose a fuzzy c-means based segmentation and sharpening; marker-controlled watershed is employed to extract each cell, and the Voronoi diagrams are further employed to correct errors due to some overlapping cells.

2.1 Extracting Rough Boundary for Each Cell

Large scale intensity variations as well as shading and shadowing effects in our images are often caused by uneven illumination over the field of view. A data-driven approach is employed to deal with this problem [4]. The algorithm works by iteratively making better distinction of the background of the image. A cubic B-spline surface is employed to model the background shading. After removing the shading, we adopt morphological transformation to enhance the image's contrast. Morphological filtering for enhancing images has been proposed by [8]. The first step of this method is to find peaks and valleys from original images. Peaks represent brighter spots of original image, and valleys represent the darker spots. Peaks are obtained by subtracting the morphologically opened image from the original image and valleys by subtracting the original image from morphologically closed image. The former is the Top-Hat transformation, and the later is the Bottom-Hat transformation. The contrast-enhanced image is obtained by the summation of the original, the peak and the nega-

tive valley image. Once the above two steps are completed, we then apply the ISODATA algorithm [9] to segment the nuclei in the DNA channel. After the above image processing, the touching spots are reduced, compared with the threshold methods traditionally used after preprocessing.

However, some nuclei may remain touching each other, we propose a method to segment touching cells. The aspect ratio of a particle image p is defined as $r(p) = w_{\min}(p)/w_{\max}(p)$, where $w_{\min}(p)$ and $w_{\max}(p)$ are the minimum and maximum diameters of the particle area. Denote the average size of the particles in the whole image to be $\aleph$, and size of cell p to be $s(p)$. The following condition is employed to isolate the touching cells: $r(p) < 0.5$, $s(p) > \aleph$, and $w_{\max}(p)$ is bigger than or equal to the pre-defined value 90 by experience. If the above condition is satisfied, then split the touching spots into two spots at the location of $w_{\min}(p)$. The merging criterion is different from the traditional approach. As an example, take a case where three nuclei can be seen within a single original cell. However, when watershed over-segments the original object, it is extremely difficult to find a proper rule to merge the fragments back into a single object. From the watershed point of view, the large object should be separated into three smaller objects. On the other hand, biologically, the three small nuclei belong to one single cell such that the large object should not be separated at the first place. Here we adopt the Hue transformation. Hue is a useful attribute in color segmentation since it is influenced by non-uniform illumination such as shade and shadow. The objective function used here is the square error of the piecewise constant approximation of the observed hue image H which is the hue transformation of the original three channels. Denote two cells regions as Ω_i and Ω_j. Define the mean Hue value as $\mu(\Omega_i)$ and $\mu(\Omega_j)$ and the following dissimilarity function:

$$f(\Omega_i, \Omega_j) = \frac{\| \Omega_i \| \cdot \| \Omega_j \|}{\| \Omega_i \| + | \Omega_j \|} \left[\mu(\Omega_i) - \mu(\Omega_j) \right]^2 I(i, j)$$

where $I(i, j)$ is 1 if Ω_i, Ω_j are neighbors; infinity otherwise. Finally if the $f(\Omega_i, \Omega_j)$ is less than a threshed 1.5, then merge the two cells. Fig. 2 shows an example the segmentation result of the DNA channel based on the original image.

Our aim in this step is to find a rough boundary or polygon that encloses the entire cell whose center is the nuclei. The assumption is that one cell shape area cannot reach other nuclei's area. Assume that we are studying the polygon of one cell whose center of the nuclei is denoted by $P_0(x_0, y_0)$, we pick up those cells whose distances between the centers of their nuclei and the (x_0, y_0) is less than a pre-defined threshold $T_0 = 100$. Denote those centers as $P_1(x_1, y_1), ..., P_N(x_N, y_N)$. Now the question is how to find certain points to be composed as vertices of a polygon. We define eight regions along the center $P_0(x_0, y_0)$ as follows:

$$R_k = \left\{ \alpha \left| -\frac{\pi}{8} + \frac{\pi}{4}(k-1) \le \alpha < \frac{\pi}{8} + \frac{\pi}{4}(k-1) \right. \right\}, \quad k = 1,2,...,8$$

We then calculate the slope and the angle between $P_i(x_i, y_i)$ and $P_0(x_0, y_0)$ as:

$$\theta_i = \begin{cases} \arctan\left(\dfrac{y_i - y_0}{x_i - x_0}\right) & if \quad x_i - x_0 >= 0 \\ \pi + \arctan\left(\dfrac{y_i - y_0}{x_i - x_0}\right) & if \quad x_i - x_0 < 0 \end{cases}, \qquad i = 1,2,...,N$$

Then it is easy to determine in which region each point of $P_1, P_2, ..., P_N$ is located. Without loss of generality, assume $P_1, P_2, ..., P_M$ points are located in region 7. Pick up the center whose distance is the closest to $P_0(x_0, y_0)$, say $P_1(x_1, y_1)$. Denote the Euclidean distance between the two centers as d_{p_0,p_1}, and the fitted radius of nuclei of $P_1(x_1, y_1)$ as r_1. Denote a new radius d as $d = d_{p_0,p_1} - r_1/2$. The point whose distance being d to the center point $P_0(x_0, y_0)$ in direction $3\pi/2$ is the right vertex in region 7. The obtained 8 vertex points are the vertex points of the expected polygon. The method proposed in this section gives us a rough boundary of each cell. Next we focus on how to segment each cell in this boundary.

2.2 RNAi Cell Segmentation

After using the above method to determine the boundary of each cell, then we can focus on this region and try to extract this cell. For our goal, we will first binarize the gray-level image. To effectively binarize the RNAi cells, we proposed the fuzzy c-means segmentation with sharpening method. Obviously, the segmentation can be treated as an unsupervised classification problem. In all clustering algorithms, the fuzzy c-means [10] is an attractive algorithm because its convergence property and low complexity, and thus is efficient to implement to screen large number of images. We first use the fuzzy c-mean clustering to partition the image pixels into $K = 3$ classes. Assume the class k is the right class we are interested in, and then we sharpen the pixels in this class by using fuzzy c-means clustering again. Here we present the sharpening technique. Because of the low contrast, it is necessary to adjust the membership values of $u_{i,k}$ (the *i*th pixel in the *k*th class) of the output from fuzzy system. Let $u_p(y)$ be the fuzzy membership value that indicates how a possible pixel y belongs to the set containing the notion of the measure of fuzziness to sharpen the fuzzy region of interest defined as 1 when $u_p(y) \geq u_0$, and 0 otherwise, where u_0 is a fuzzy membership threshold. Pham, et. al., [11] proposed to select $u_0 = u_{c^*}(y^*)$, where y^* is the pixel with maximum intensity value, and c^* is the right class which we are interested in. This choice of u_0 does not work in our images. Principally, the misclassified pixels mainly come from the closed membership values between the biggest and the second biggest values. If they are too close, the classification results are not reliable. Denote the difference between the biggest and the second biggest membership values in the class c^* to be v, i.e.,

$v_i = \max_k(u_{i,k}) - \max_{k \neq k^0}(u_{i,k}), \quad i = 1,...,\zeta$, where $k^0 = \arg\max_k(u_{i,k})$ and ζ is the total number of pixels in class c^*. Here we propose to estimate the threshold u_0 using fuzzy c-means again. We first partition the values of $v_i, i = 1,...,\zeta$ into two classes. Denote the minimal of the bigger class as u_1, and the maximal of the smaller class as u_2. Then the u_0 is defined as the average of these two parameters.

Isolating the touched cells is extremely challenging in automatic RNAi screening. If we adopt the watershed method, it can easily generate many false positives due to oversegmentation. It is difficult to remove the oversegmentation simply by applying certain heuristic rules or criteria [4, 6]. Since we already know the rough region and center of each cell, we propose to modify the marker function so that pseudo minima can be removed while the center of each cell can be kept. The catchment's basins associated to the pseudo minima are filled by morphological reconstruction and then transformed into plateaus. Then those minima will not generate different regions when watersheds are obtained. This method is called marker-controlled watershed algorithm. The markers are the centers of cells which are obtained in the first step. As the boundary of many cells is weak, it is hard to extract their boundary by using intensity gradient of pixels. Thus, after we extract the binarized cells, we then calculate the Euclidian Distance map of the binary image developed with the proposed fuzzy c-means segmentation and sharpening. We impose the markers to this Euclidian distance map, and then we applied watershed algorithm to segment the cells. Note that the above procedure is for segmenting a single cell. After we process all cells, it still could cause overlapping between a small number of cells. We finally apply the Voronoi diagrams [12] to correct the overlapped cells. Fig. 3 is the segmentation result by using the proposed approach.

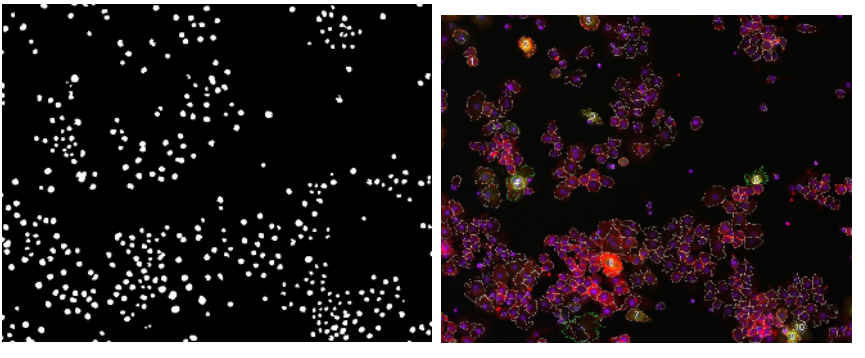

Fig. 3. left-DNA channel segmentation result; right- segmentation result

3 RNAi Cell Phenotype Recognition

In RNAi genome-wide screen, our goal is to identify all the possible cell phenotypes derived in the screening. In the meantime, we also need to consider some location features such as roundness and eccentricity. The first set of features was based on

Zerike moments. We extracted 49 features. The second set are co-occurrence contexture features. We extracted 28 textural features. The third set was based on cell hull location features [13]. Here we choose the three features which are the fraction of the convex hull area occupied by protein fluorescence, the roundness of the convex hull, and the eccentricity of the convex hull. We totally extracted 15 common image features. Since the feature values have completely different ranges, an objective scaling of features was achieved by calculating z-scores. Principle component analysis (PCA), commonly used in microarray research as a cluster analysis tool, is employed to reduce features. KNN classifier is a non-parametric classifier, which is easy to implement. Hence we use KNN to finish up the cell phenotype classification. Four classes are defined which are shown in Fig. 2. The four classes are Spikey, Ruffling, Actin acceleration at edge, and Normal cells. A training set is obtained for each class. We use KNN to classify test data.

4 Results and Discussions

In our study, four images of cells that had been visually classified are used to establish the training data sets. The limited amount of training data was due mainly to the tedious visual classification. We first used the proposed automated segmentation algorithm to segment each image, and then we asked the biologists to mark the cell with different phenotypes. Note that we actually have four classes in the training data sets as the other one is the class of normal cell which does not have significant change in their morphological shape. Cross-validation is employed to test the performance of the proposed automatic screening approach. We randomly split the data sets into training data set and test data sets, where 70% of cells are treated as training data and 30% of cells are treated as testing data. The feature reduction and normalization are first done with the training data sets, and then applied the transformation matrix of feature reduction to the test data set and similar procedure applied to normalization. Then we run 100 times and calculate the mean of the recognition accuracy based on the test data. The recognition accuracy is listed in Table 1. It is seen that the recognition accuracy is between 62% and 75%. Our biological collaborators reckoned that 70% accuracy should be adequate for the purpose of automatic screening. To further improve the accuracy and specificity, we continue to improve the segmentation algorithm, the phenotype definition, and the specific image features for extraction.

In this study, we proposed a two-step segmentation approach to segment high content cell images automatically. Certain regular image features, Heraik contextual feature and Zerick moment features are extracted for each cell. KNN is employed to perform cell phenotype classification. Experiments show that the proposed approach can automatically and effectively identify cell phenotype. Although we have built an initial workable system for automated RNAi genome-wide screening, there are certain problems remains. For example, we conjecture that the snake model in segmentation might be more effective than the marker-controlled watershed algorithm and Voronoi diagrams as the cells in GFP-Rac12 and F-actin have many different kinds of closed curves. We are in the processing of testing this conjecture. For feature extraction, additional image features specific to the different cell phenotypes, such as spiky region, ruffling region and actin acceleration region, would need to be identified. Furthermore, we would study how to automatically extract phenotypes hidden in cell

shapes by using cluster analysis. The ultimate goal is to score each image by establishing robust mathematic models to map the number of different phenotypes in each image to a scoring system which would let biologists easily find the novel candidate genes in their screens.

Table 1. Recognition accuracy for RNAi cell phenotypes

Phenotype	Actin-A	Ruffling-R	Spiky-S	Normal-N
Actin	72.87%	8.6%	1.69%	16.8%
Ruffling	8.68%	75.14%	6.13%	10.06%
Spikey	6.97%	20.36%	62.83%	9.84%
Normal	14.5%	6.54%	4.79%	74.17%

References

1. Z.E. Perlman, et.al., “Multidimensional drug profiling by automated microscopy,” Science, 306 (2004) 1194-1198.
2. M. Boutros, et.al., "Genome-Wide RNAi analysis of growth and viability in Drosophila cells," Science, 303 (2004), 832-835.
3. A.A. Kiger, et. al., “A functional genomic analysis of cell morphology using RNA-interference,” Journal of Biology, 2(4): 27, 2003.
4. C. Wählby, et.al., “Algorithms for cytoplasm segmentation of fluorescence labeled cells,” Analytical Cellular Pathology, 24(2002):101-111.
5. N. Malpica, et. al., “Applying watershed algorithms to the segmentation of clustered nuclei,” Cytometry, 28(1997):289-97.
6. G. Lin, et.al., “Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei,” Cytometry Part A. 63(2005):20-33, 2005.
7. J. Lindblad, et.al., “Image Analysis for Automatic Segmentation of Cytoplasms and Classification of Rac1 Activation,” Cytometry Part A, 57A(1):22-33, 2004.
8. S. Chen and M. Haralick, “Recursive erosion, dilation opening and closing transform,” IEEE Transactions on Image Processing, 4(1995):335-345.
9. N. Otsu, “A threshold selection method from gray level histogram,” IEEE Trans. on System man Cybernetics, 8(1978): 62-66.
10. 10 X. Zhou, et.al., “Gene clustering based on cluster-wide mutual information,” Journal of Computational Biology, 11(2004):151-165.
11. 11 T.D. Pham, et.al., “Extraction of fluorescent cell puncta by adative fuzzy segmentation,” Bioinformatics, 20(2004):2189-2196.
12. A. Okabe, et. al., “Nearest neighbourhood operations with generalized Voronoi diagrams: a review,” International Journal on Geographical Information Systems, 8(1994):43-71.
13. M. V. Boland and R.Murphy, Bioinformatics, 17(2001):1213-1223.