Xiaobo Zhou
and Stephen T.C. Wong

High Content Cellular Imaging for Drug Development

As the pharmaceutical industry battles the escalating cost and time in drug development, there is an urgent need for better decisionmaking in early stage targets, lead selection, and late-stage attribution. High content screening (HCS) allows functional analysis of targets and pathway modulation in cells by drug compounds in a high-throughput manner. It has recently emerged as a promising solution to improve the quality of decisionmaking in drug development. However, tools required for processing and analyzing HCS data are rather immature. The purpose of this article is to review major image analysis techniques in HCS and illustrate these techniques with different screening applications, including compounds screening, ribonucleic acid interference (RNAi) genome-wide screening, time-lapse cell cycle screening, and neuron-based assay screening.

BACKGROUND

Cell-based assays are assays prepared in multiwell formats, such as 96-well and 384-well plates, for high-throughput screening to study responses of a population of cells under different chemical, genetics, or radiation perturbations. They are widely used for the development of new drugs starting from primary screening to in vitro toxicology. Extracting quality information in bioassay development and screening is enabled by a powerful combination of multidye fluorescence imaging, flexible analysis algorithms, and full system automation. HCS is a powerful tool for disease diagnosis and prognosis, drug target validation, and compound lead selection. A roadblock that prevents HCS from becoming widely used is the difficulty of handling and analyzing large amounts of image datasets generated.

Recent technological advancements, such as fast digital cameras, automated motorized microscopes, new fluorophores (such as the enhanced green fluorescent protein (EGFP) and related fluorescent proteins), and quantum dots, as well as increased computational power, have dramatically enhanced the ability to acquire multispectra data during time-lapse microscopy imaging. Consequently, an explosive growth has occurred in both the number and the complexity of the images acquired. Existing imaging tools, such as the ImageJ from the National Institute of Health or commercial software packages such as MetaMorph from Molecular Devices Corporation (Sunnyvale, California), while satisfactory for simple cellular image processing, are extremely limited in their scope and capacity for high content cellular analysis. The challenge lies in how to convert all the images showing functions and interactions of macromolecules in live cells and tissues into quantitative numbers that can be analyzed statistically.

The ability to visualize, trace, and quantify cellular morphology at high spatial and temporal resolution is essential to understanding biological processes and the development of effective therapeutic agents. Three types of biomarkers can be used to characterize the spatial features of cells: subcellular structures, location of signaling proteins, and indicators of physiological states. Compounds that affect spatial arrangement of signaling proteins or cellular structures can provide important hints for studying biological processes and providing therapeutic intervention. To identify such compounds using microscopy is part of the "forward chemical genetics" approach that has been used in our center at Harvard Medical School. We are equipped with a GE IN Cell 1,000 cell analyzer and can also access several other high-throughput fluorescence imagers at Harvard. These microscopy imagers can automatically acquire cell images from transparent bottom 96- or 384-well plates, using objectives from 2x to 60x. Since drug treatment can induce changes in cellular organelles and protein localization that are readily detected by microscopy, and these changes may reflect cytotoxicity, true mechanisms of action, or both, it is important to continue the few studies performed so far. This motivates us to develop automated high content cellular analysis tools to address challenging HCS applications. These bioinformatics tools will improve the selection of drug targets and compound leads and reduce the cost of drug development. In this work, we review the challenges and data modeling of high-throughput automated microscopy in [1], and then focus on the image processing steps of the bioimage informatics pipeline of HCS.

BIOIMAGE PROCESSING AND ANALYSIS

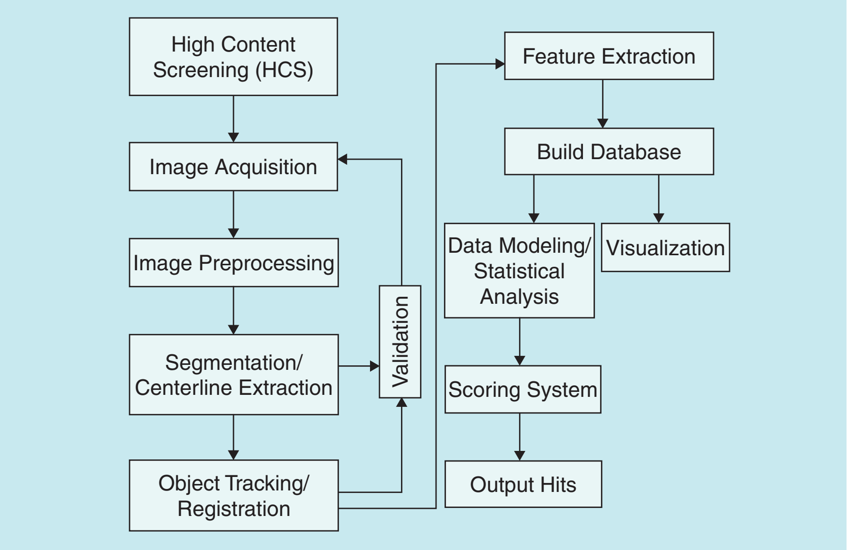
In drug development, chemists and biologists first design hypotheses and then screen a large number of compounds to identify hits or targets. Different microscopes are deployed to acquire imaging screens, and hundreds and thousands of images are generated. As shown in Figure 1, the bioimage informatics pipeline is composed of image processing, segmentation, feature extraction, database management, data visualization,

1053-5888/06/$20.00©2006IEEE

data modeling and statistical analysis, and compound scoring. The left portion of Figure 1 consists of the image processing modules, which are the focus of our discussion. The three major steps in HCS image processing are: 1) image preprocessing, 2) segmentation and centerline extraction, and 3) cell tracking and registration.

IMAGE PREPROCESSING

The image preprocessing module consists of a series of image processing methods to improve the quality of raw images for image segmentation and feature extraction. The module includes three parts: image restoration, noise removal, and contrast enhancement. The image restoration block deconvolves the degraded image from the microscope using the point spread function provided by the manufacturer. The noise removal block uses a median filter to remove the pepper-noise generated by the CCD detector in optical florescent microscopy, as the median filter preserves high-frequency information of the cell edges. The image contrast enhancement block uses a contrast-adjustment algorithm that compensates for the nonuniform image intensity from uneven light illumination. In microscopy image processing, deconvolution is often employed to preprocess the images. Nearest neighbors and no neighbor deblurring filters have been shown to be more effective than other deconvolution filters.


High Content Screening (HCS)
Image Acquisition
Image Preprocessing
Segmentation/ Centerline Extraction
Object Tracking/ Registration
Validation
Feature Extraction
Build Database
Data Modeling/ Statistical Analysis
Visualization
Scoring System
Output Hits

[FIG1] The bioimage informatics pipeline, which consists of image processing, feature extraction, database construction, data modeling and visualization. The left part in this figure describes the image processing modules that are our focus of discussion in this article.

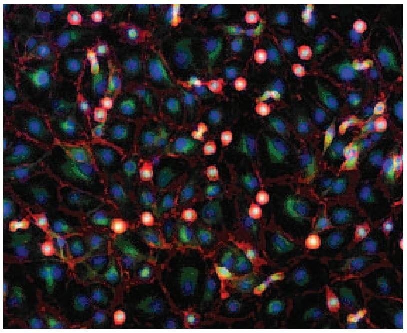
SEGMENTATION AND CENTERLINE EXTRACTION

Image segmentation is the most critical step in cellular image analysis. Well-known segmentation methods include histogram-based, edge-detection-based, watershed, morphological, and stochastic techniques. Commonly used histogram-based methods typically select a threshold via maximizing the variance between objects and background or minimizing the interclass variance of objects and background. The threshold may be computed globally (using the entire image) or locally (using image regions). Global thresholding methods sometimes may fail due to uneven background and illumination. Locally adaptive thresholding methods are often more effective when there exist spatially changing background and varying illumination conditions. More often than not, cell segmentation methods use the thresholding and edge-detection operators. However, when cells and spots of interest overlap or touch each other, some of the segmentation methods mentioned above may oversegment the image [2]. To correctly identify the cell phase and other phenotypes in such cases, improved methods have been proposed. Methods of segmentation and centerline extractions in applications such as drug screening, genome-wide RNAi screening, time-lapse cell-cycle analysis, and automated neuron-based assay analysis for Alzheimer's disease drug screening are discussed next.

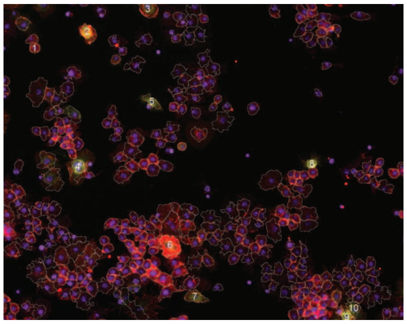
DRUG SCREENING

The role of the spindle in a cell is to accurately partition the sister chromatids equally into two daughter cells. The tubulin-based structure of the spindle is the target of many anticancer drug studies, since the microtubule spindle can modify the process and outcomes of cell divisions. Cell-permeable small molecules that rapidly activate or inactivate the function of the spindle can be useful probes of dynamic cellular processes. However, quantitative assessment of the effects of small molecules and compounds on the spindle function has so far seen relatively little use for drug discovery because manual counting makes the process extremely time consuming and there are no effective methods for automatic recognition of cell phases. It is also difficult to recognize various phases in a cell cycle from only one phenotypic image obtained using a single fluorescent channel. In what follows, we describe an example of segmentation proposed in our work to quantitatively analyze the cell mitosis under a variety of conditions, including wild drug treatment and gene knockout.

In our application for mitotic analysis, DNA, microtubules, and actin are stained with three different dyes for nuclei, mitosis, and cytoplasm. The combination of the three stains renders cellular nuclei stained deep blue, microtubules stained green, and cytoplasmic components stained red, as indicated in Figure 2. Our multiphenotypic mitotic analysis (MMA) system [3] distinguishes and labels cells at different phases in cell division, e.g., interphase, metaphase, and anaphase. This system can be used to calculate the numbers of cells in these three phases represented in an image, archive the extracted information and image data, and analyze the mitotic data to provide an index of cell division. We applied the MMA system to evaluate HCS data of Monastrol suppressor screening to demonstrate its utility in drug screening (Monastrol was discovered as the first cell permeable, small molecule inhibitor of the mitotic machinery that does not target tubulin [4]). In MMA, the image segmentation algorithm employed is largely edge-detection based, while most image processing procedures are binary morphological operations, except the Laplacian of Gaussian (LoG) edge detection [3]. The LoG method converts a gray-level image into a binary edge-enhanced image so that it is easy to use morphological operations to extract a final mask of the objects from the background. After cell recognition, the ratio of bipolar cells (metaphase cells and anaphase cells) over monoasters is calculated to index the effectiveness of each compound suppressing Monastrol. We calculated the ratio of bipolar cells over monoasters versus a selection of 320 compounds. The most effective compound to suppress Monastrol was found to be the compound 149, which stands out with the maximum bipolar/monoaster ratio of 3:1 [3].



[FIG2] A superimposed image of HCS using the three channels [2].



[FIG3] The segmented cells of RNAi genome-wide screening [3].

GENOME-WIDE RNAi SCREENING IN DROSOPHILA CELLS

Another example of segmentation is that involved in the work of our collaborators at Perrimon Lab of Harvard Medical School, who have developed a cell-based assay for Rho GTPase activity using the Drosophila (fruit fly) Kc167 embryonic cell line. GTPases are a large family of enzymes that can bind and hydrolyze the Guanosine triphosphate (GTP), which is a nucleotide. The Rho family of guanine nucleotide (GTP)-binding proteins consists of Rho, Rac, and Cdc42 subfamilies. The Rho family of small GTPases is essential for cell shape changes during normal development and cell migration, as well as during disease states such as cancer [5]. Additionally, Rho proteins regulate many other facets of cell behavior, such as endocytosis (a process in which a substance gains entry into a cell without passing through the cell membrane), vesicle trafficking (a common locus of the origin or manifestation of many human pathologies), cell polarity, and cell cycle. Understanding how Rho proteins cause various cellular responses is an area under intense investigation. In their active state, Rho proteins interact with effector molecules and modulate their activities to relay or implement downstream responses. As the number of genes implicated in Rho pathways rapidly increases, it becomes clear that Rho proteins interact with a myriad of effectors to orchestrate the varied cellular outcomes. Constitutively active forms of Rho proteins cause distinct morphological changes in cells both in culture and in vivo. It is this property that was used by our collaborators mentioned above. The double strain RNAs (dsRNAs) that result in an alteration or loss of RacV12 cell morphology will identify genes encoding in downstream components of Rac signaling. These genes will be further studied to elucidate the mechanisms by which they function in Rac signaling in vivo. Expression of RhoAV14 and Cdc42V12 in Kc167 cells also induces specific cellular morphologies. Comparison of the three screens will be useful to identify common and distinct downstream targets and to further elucidate the complex signaling networks in which these GTPases function.

To summarize, the cellular morphology of RacV12 expressing cells (cells expressing RacV12) must be identified and defined; the effect of the double strain RNA (dsRNA) on the normal distribution of RacV12 morphologies must be determined. As the first step, the three cell shapes must be identified; the following labels can be used: A (the shape has Actin accumulation at edge), ruffling (the shape is like ruflling), and spiky (the shape is like a spiky) [3]. For screening the whole genome of Drosophila with the three assays (rac, rho, and cdc42), three-channel imaging, and 22,000 genes, over a million images are generated in the study. Since the number of images is prohibitive for manual image analysis, a recently proposed approach that integrates a novel two-step segmentation, feature extraction, and phenotype classification is employed [6].

The key issue in this example is how to automatically segment cells of cell-based assays cost effectively, as fast screening often generates poor image quality. In [7], a cytoplasm segmentation based on the watershed segmentation and rule-based merging and splitting of oversegmented and undersegmented objects was proposed. As a typical example in Figure 3 shows, the segmentation of RNAi genome-wide images is challenging. The two-step segmentation approach mentioned earlier, followed by processing of tens of millions of cells and classification of cell phenotypes, yields the results shown in Figure 3. In this method that we proposed, the first step consists of the extraction of a rough boundary for each cell. The boundary is extracted by determining the center of each cell and the polygon for each cell. The next steps consist of a fuzzy c-means-based algorithm for segmentation and sharpening, a marker-controlled watershed algorithm for extracting each cell, and the Voronoi diagrams to correct errors due to overlapping cells. An alternative to this method is that based on level sets. Although it yields a good boundary of the cytoplasm, the associated computational cost is too high for HCS applications.

TIME-LAPSE CELL CYCLE ANALYSIS

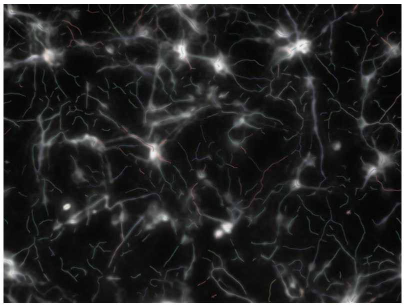
To better understand how apoptosis (cell death) is induced by antimitotic drugs and how drug resistance might arise, it is important to measure cell cycle progression, in particular, the mitosis (M) phase; to distinguish cells with normal and abnormal interphase; and to detect the initiation of apoptosis in individual cells [2]. In addition, cancer drug treatment can induce changes in apoptosis and protein localization, which are readily detected by the arrival of the time-lapse microscopy. Time-lapse microscopy has become an important method to evaluate the response of individual cells amid a population to drug treatment dynamically with far richer information content than the conventional fixed-cell microscopy. The availability of this new dynamic imaging modality gives rise to analytical challenges in extracting information efficiently from large volumes of time-lapse images.

In time-lapse fluorescence microscopy images, nuclei are bright, usually elliptic objects protruding out from a relatively uniform dark background. The nuclei can be segmented by applying global thresholding methods such as the Otsu algorithm. The algorithm segments most of the isolated nuclei correctly, except for nuclei that are overlapping or touching each other. To resolve the issue of touching nuclei, we used a watershed algorithm. To correct the oversegmentation of some of the nuclei, a postprocessing step is needed to correct and merge the oversegmented nuclei fragments using (for instance) a roughness criterion [2].

AUTOMATED NEURON-BASED ASSAY ANALYSIS FOR ALZHEIMER'S DISEASE DRUG SCREENING

An important application of HCS is to help identify possible drug leads in treating Alzheimer's disease (AD). The mechanism of neuronal dysfunction in AD remains unclear so far. Using multichannel image data obtained from hippocampal neurons, we integrated and developed techniques to screen for drug leaders in AD treatment. Our approach is to simultaneously analyze and quantify two types of images. Images of nuclei are stained by nuclei using various nuclear stains, e.g., Sytox Green and Hoechst, while images of neurites are stained using neuronal specific antibodies conjugated to Cy3. For each different dosage of Amyloid Beta (Aβ), we look for the changes in neurites and nuclei to find how the cells and neurites respond to Aβ. The result helps us elucidate how Aβ affects the appearance or disappearance of neurites and cells. At each experiment, hundreds of wells are treated with various dosages of Aβ followed by nuclear/neurite staining. The large number of images allows us to use powerful statistical models to elucidate the underlying relationship between Aβ, neurite loss, and apoptosis of neuronal cells.

Despite considerable efforts to automate the analysis of the images acquired as described above, only a few semiautomatic and automated approaches have been proposed so far. The common direct exploratory tracing algorithms automatically detect the initial points and extract the center lines of the line structures iteratively. This class of algorithms requires semiautomatic selection of parameters such as the length of the templates and the maximum expected width of the vessels, which may limit its possible applications to other types of images. In addition, the tracing performance is mainly determined by the selection of the parameters and the stop conditions. The line-pixel detection algorithms [8] model the local geometric properties of the lines using each pixel in the image, followed by linking the successive line pixels that are most likely to represent the center lines of the neurites. Although computationally expensive, these algorithms yield highly precise extraction of the line structures. Currently, there is no software available for fully automated neuron-based assay high-content screening.


[FIG4] The extracted centerlines (in white) of neuron-based assay drug screening [1].

Unlike the previously reported neurite tracking method in [8], our goal was to develop fully automated algorithms for neurite marking, which can detect all neurites together in the same image without any user interaction and involving only a few parameters [9]. Our approach consists of two phases: 1) a marking phase, in which a multiscale curvilinear structure detector yields a single and connected response for each neurite, and 2) a linking phase, in which breaks near branching structures are connected. One disadvantage of this method is that it is hard to detect the centerline when its signal is too weak, as shown in Figure 4.

TRACKING AND REGISTRATION

Object tracking has been well studied and often involves Kalman filters. In cell biology, scientists aim to track several individual particles automatically. For such a task, the centroid method, the Gaussian fit method, the correlation method, the sum-of-absolute differences method, and the interpolation method have been quite popular for individual particle tracking. However, these methods cannot be simply deployed in high-content cell or spot tracking, because their single spot tracking method does not address the ambiguous association—a difficult problem for multiple particles tracking. Improved methods that use the maximal mutual information, Bayesian theory, as well as our shape-and-size-based methods, have been proposed [2]. To analyze the circulatory patterns of blood cells, a cell tracking system called CellTracker was developed in [10]. This system utilized the direction of the blood flow and the location and velocities of circulating cells to solve the motion correspondence. Since cancer cell migration study is an important research area in cell biology and drug discovery, some works focused on tracking cell migration and used the mean-shift method.

DATABASE AND SCORING SYSTEM IN DRUG DEVELOPMENT

After cellular image segmentation, relevant image features of every cell are extracted into numerical descriptors. Substantial work has been done at Carnegie Mellon University on single cell structure analysis using fixed cell imaging. A list of image features such as area, shape, size, perimeter, intensity, and texture has been considered, using the features' associated descriptors. For dynamic and large population cellular imaging, we found that additional temporal parameters—such as the change of the size and shape of nuclei during and after the mitosis and the duration between different shapes—can be used to track and identify the progression of a cell or its offspring during the mitotic process and over a large population of cells simultaneously. After extracting a large number of features for high-content images, it is important to design a database system to archive and organize the meta data. A logical database model helps classify, organize, and represent different classes of high-content images, image descriptor meta data, and associated textual information to support cell line recognition, drug-treatment response analysis for drug discovery, data mining for experiment design and refining, and content-based image search. These capabilities further enable the user to query individual cellular or subcellular objects, as well as to model the data.

Optimal methods of scoring biomarkers and identifying candidate hits have been actively studied in academia and industry. To find candidate hits, we need to score the images associated with different compound interventions. In the first application example of compound screening, we need to score each compound. In the application example of RNAi genome-wide screening, we aim to find the candidate effectors or genes that correspond to the images acquired using the three color channels. Scoring the effectors is equivalent to scoring the images based on the number of phenotypes that exist in the images. To do so, we need to first obtain the number of different phenotypes and then build a modeling system to predict each gene's score. In the neuron-based assay screening for treating AD, the score can be defined as the ratio of the total lengths of neuritis to the number of neurons or nuclei. In the time-lapse cell cycle study, the score can be defined as the ratio between the number of cells in arrested metaphase and the number of cells in interphase. We believe that the scoring approach will become a permanent component of HCS.

SUMMARY

In this article, we discussed the bioinformatics issues related to HCS and presented a bioinformatics pipeline for HCS drug development. We reviewed advanced methods for bioimage processing, cell segmentation, neurite centerline extraction, and cell tracking and registration in the context of HCS of different bioassays. These methods prove that HCS is potentially a powerful tool in quantitative cell biology and drug discovery. There remain, however, many challenging issues that require the development of new bioimage analysis algorithms to fulfill the potential of HCS.

ACKNOWLEDGMENTS

The authors acknowledge rewarding collaborations with biology colleagues Timothy Mitchison, Junying Yuan, Randall King, and Norbert Perrimon at Harvard Medical School. The images in this article were supplied by our collaborators. The image processing and computational modeling work is the original contribution of the authors with the input from L. Ji and G. Xiong at Tsinghua University, and other members of life science imaging group of HCNR-Center for Bioinformatics, notably Yong Zhang, Xiaowei Chen, Jun Yuan, Xiaoyin Xu, Jinmin Zhu, Kuang-Yu Liu, and Xinhua Cao. This research is funded by the HCNR Center for Bioinformatics, Harvard Medical School, and the NIH R01 LM008696 Grant.

AUTHORS

Xiaobo Zhou is an instructor of radiology at Brigham and Women's Hospital, Harvard Medical School. His current research interests include high-content molecular and cellular imaging, computational systems bioinformatics, systems biology, and neuroinformatics.

Stephen Wong is the director of HCNR Center for Bioinformatics, executive director of the Functional and Molecular Imaging Center, and an associate professor of Radiology, Harvard Medical School and Brigham & Women's Hospital. He has 20 years of R&D experience with HP, Bell Labs, Japanese 5th generation computer systems project, Royal Philips Electronics, Charles Schwab, UCSF, and Harvard.

REFERENCES

[1] X. Zhou and S.T.C. Wong, "Informatics challenges of high-throughput microscopy," *IEEE Signal Processing Mag.*, to be published.

[2] X. Chen, X. Zhou, and S.T.C. Wong, "An image analysis system for segmentation, classification, and tracking of cell cycle phases in time-lapse fluorescence microscopy," *IEEE Trans. Biomed. Eng.*, to be published.

[3] X. Zhou, X. Cao, Z. Perlman, and S.T.C. Wong, "A computerized cellular imaging system for high content analysis in monastrol suppressor screens," *J. Biomed. Inform.*, to be published.

[4] T.U. Mayer et al., "Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen," *Science*, vol. 286, no. 5441, pp. 971–974, 1999.

[5] R.W. Carthew, "Gene silencing by double-stranded RNA," *Curr. Opin. Cell. Biol.*, vol. 13, no. 2, pp. 244–248, 2001.

[6] X. Zhou, K.L. Liu, P. Bradley, N. Perrimon, and S.T.C. Wong, "Towards automated cellular image segmentation for RNAi genome-wide screening," *Lect. Notes Comp Sci.*, vol. 3749, pp. 885–892, 2005.

[7] C. Wahlby et al., "Algorithms for cytoplasm segmentation of fluorescence labelled cells," *Anal. Cell. Pathol.*, vol. 24, no. 2–3, pp. 101–111, 2002.

[8] E. Meijering et al., "Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images, *Cytometry A*, vol. 58, no. 2, pp. 167–176, 2004.

[9] G. Xiong, X. Zhou, L. Ji, A. Degterev, and S.T.C. Wong, "Automated neurite labeling and analysis in fluorescence microscopy images," *Cytometry A*, to be published.

[10] E. Eden et al., "An automated method for analysis of flow characteristics of circulating particles from in vivo video microscopy," *IEEE Trans. Med. Imag.*, vol. 24, no. 8, pp. 1011–1024, 2005.