Xiaobo Zhou and Stephen T.C. Wong

Informatics issues in cellular and molecular imaging

Informatics Challenges of High-Throughput Microscopy

High-throughput analysis of cellular and molecular images is increasingly becoming a powerful tool for drug target validation and compound lead selection. Currently, scientists resort to slow, manual analysis to extract and analyze information from large amounts of microscopic image data. Image informatics has become the rate-limiting factor in realizing the full potential of dynamic cellular and molecular imaging studies. In this article, we give examples of high-throughput microscopy applications in cell and molecular biology and review image analysis and data modeling techniques used in high-throughput image screening. We also illustrate a number of these techniques with four biological application examples: compounds screening, RNAi whole genome screening, time-lapse cell cycle screening, and clathrin particle detection and tracking (with more detailed discussion in the time-lapse imaging application). Informatics of high-throughput microscopic imaging is an interdisciplinary field that has the potential to expand our knowledge of biological mechanisms and accelerate the process of drug discovery.

1053-5888/06/$20.00©2006IEEE

BACKGROUND

HIGH-THROUGHPUT SCREENING

High-throughput screening (HTS) via automated fluorescence microscopy, also known as high-content screening (HCS), is becoming an important research tool to assist scientists in understanding the complex processes of cell division or mitosis [1], [2], disease diagnosis and prognosis, drug target validation, and compound lead selection. Its power comes from the sensitivity and resolution of automated light microscopy with multiwell plates, combined with the availability of fluorescent probes that are attached to specific subcellular components, such as chromosomes and microtubules, for visualization of cellular phenotypes and activities (such as mitosis) using standard epifluorescence microscopy techniques. Biologists have acquired large volumes of light microscopy images from cells and tissues for studying cellular dynamics at different biological levels of complexity and resolution, including cell movement, changes in cell shape in response to the environment, intracellular traffic of vesicles, pathogens, nucleic acids, proteins and lipids, biogenesis of organelles, and, more recently, the movement and behavior of single molecules within a cell.

There exist, however, significant challenges in high-throughput bioimaging, including accurate segmentation, tracking, and modeling of either the dynamic cellular behavior of cells in a large population or of thousands of fast-moving molecular particles within a living cell. Existing imaging analysis tools are limited in their scope and capacity to analyze high-throughput, live-cell images. Currently, scientists resort to slow, manual analysis to extract information. The manual approach is time consuming, subject to inter- and intra-observer variance, and not scalable for large studies. Informatics has thus become the rate-limiting factor in realizing the potential of dynamic cellular and molecular imaging studies. The ability to visualize, trace, quantify, and model cellular morphological features with high spatial and temporal resolution is key to the understanding of biological processes and the development of effective therapeutic agents.

MARKERS FOR CELL FUNCTIONS

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. First, mammalian cell functions are highly ordered around specialized subcellular organelles. DNA replication and mRNA transcription are carried out inside the nucleus. Membrane and secreted proteins are synthesized and processed in endoplasmic reticulum (ER) and Golgi apparatus, whereas cell shape and movement are supported by dynamic structure elements, such as microtubules and actin. The cytological properties of these subcellular structures and their responses to the outside environment reflect both the intrinsic property of cells and how cellular functions are regulated. These properties can be recorded by using fluorescence microscopy to study residential proteins of the organelles. Cellular image analysis methods are being developed to quantify the properties of these structures. Size, shape, intensity, and texture moments are among the common descriptors that can be quantitated using computational algorithms [6].

Second, to carry out proper cellular functions, many proteins shuttle between different intracellular locations. The most noticeable cases are kinases and transcription factors that transmit the signals from cell surface to nucleus, turning on transcription control. Phosphorylation often accompanies translocation, which sometimes includes only the movement of a small pool of phosphorylated proteins. There are also many signaling adaptor proteins that accumulate on the cell surface or other organelles when activated. Cell surface receptors, on the contrary, often move to endocytotic structures to signal down-regulation. All these movements can be used as signatures for turning on and off certain signaling pathways, which cannot effectively be scored by other HTS methods except automated fluorescence microscopy. Recently, protein locations have been studied on a genome-wide scale and shown to be closely associated with their functions and interacting partner networks [7]. This adds to the importance of monitoring the locations of proteins within cells.

Third, there are indicators of cellular physiological states, such as intracellular ion levels, pH, redox states, membrane potentials, and live/dead status. Many chemical probes are available to convert these physiological states into fluorescence signals that can be read out by a microscope. Not only spatial and cell individuality will contribute to the information that are not available with bulk readouts before, but the temporal changes of these states under different conditions can also provide important information about the cells.

AUTOMATED FLUORESCENT MICROSCOPY

The two major roadblocks that prevent high-throughput fluorescence microscopy from becoming more widely used are the lack of high-acquisition throughput and the difficulty of handling and analyzing large amounts of image datasets generated. Automated microscopy has recently become available. Traditionally, fluorescence microscopes are highly sophisticated pieces of equipment that depend entirely on thoroughly trained research operators. To support unattended data acquisition, an automated microscopy would generate digital images from multiwell plates, e.g., 96 and 384 plates, and would automatically focus when moved to a new well or plate and change filters for multiplexing colors.

Technical developments in the past five years, such as fast and reliable digital cameras for data acquisition, increased computational power, automated motorized microscopes, new fluorophores [such as the Alexa series, enhanced green fluorescent protein (EGFP)] and related fluorescent proteins, and quantum dots, have dramatically increased the ability to acquire multispectra data during time-lapse live cell imaging in two and three spatial dimensions. Consequently, an explosive growth has occurred in both the number and complexity of the images being acquired. With the time-lapse acquisition capability being added into automated florescence microscopy, the amount of image data reaches yet another scale of complexity. The tools for analyzing the resulting images have not kept pace with these developments, however. Current imaging tools such as NIH Image (available as Scion Image or Image/J),

MetaMorph, UTHSCSA ImageTool, and QED Image, while reasonably satisfactory for standard image processing, are extremely limited in their scope and capacity for HTS image data analysis, particularly with respect to complex shapes or multispectral and temporal correlations. The challenge lies in how to convert the images showing functions and interactions of macromolecules in live cells and tissues into quantitative descriptors that can be analyzed statistically.

Automated analysis of cellular microscope images depends on the development and integration of computerized segmentation, feature extraction, pattern recognition, and statistical modeling for this new class of image data types. In this article, we briefly describe a number of informatics issues arising from high-throughput microscopy analysis, emphasizing the discussion of quantitative data modeling.

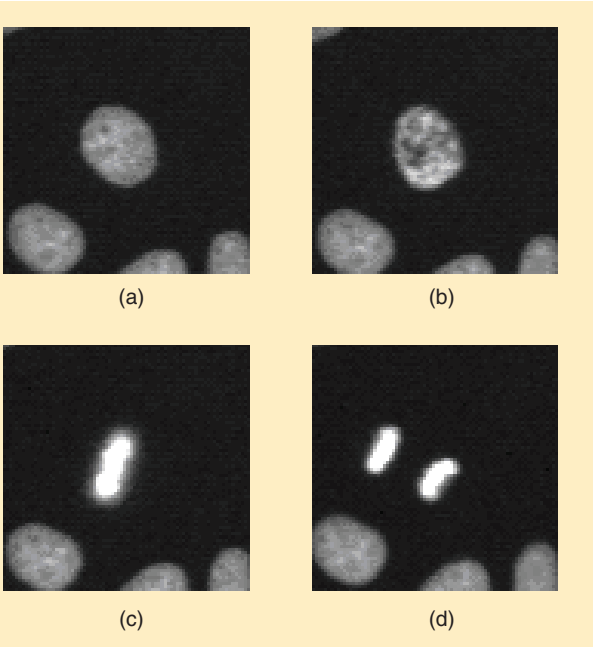
EXAMPLE HTS APPLICATIONS

Automated, high-throughput microscopy can be considered as consisting of six modules: 1) biological experiments and image acquisition, 2) image processing and analysis, 3) quantitative feature extraction and database storage, 4) validation, 5) data modeling and statistical analysis, and 6) visualization. We have reviewed the challenging issues in automated image processing for high-throughput cellular microscopy in [26]. In this section, we present four examples of biological experiments of HTS, focusing on the aspects of data modeling and statistical analysis. We focus on the discussion on the high-throughput cell cycle imaging studies and briefly review the other three applications.

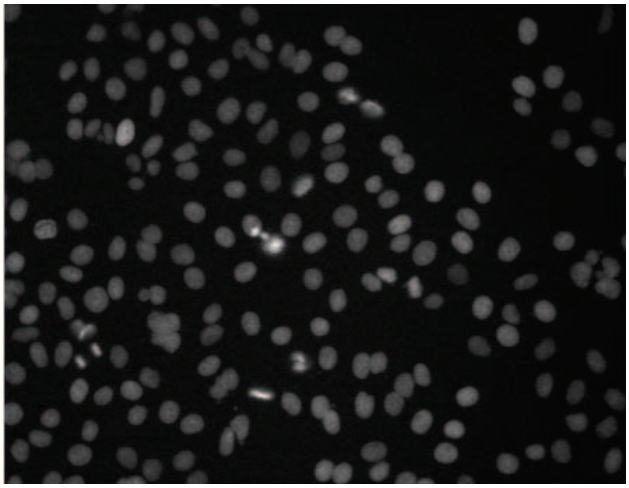
CELL CYCLE ANALYSIS FOR TIME-LAPSE MICROSCOPY SCREENS

To better understand how apoptosis is induced by antimitotic drugs and how drug resistance might arise, it would be ideal to be able to measure cell cycle progression, in particular, the mitotic phases that include prophase, metaphase, and anaphase (see Figure 1); to distinguish cells with normal and abnormal interphase; and to detect the initiation of apoptosis in individual cells as a function of time. Time-lapse microscopy has become an important means to dynamically quantitate the response of individual cells amid a population to drug treatment with far richer information content than the traditional fixed cell microscopy. It also has the potential to make significant contribution to the field of cellular biology by providing more precise, quantitative, and multiparametric characterizations of cell cycle mechanisms under different perturbations than the existing manual laboratory methods. We are developing a tool to identify individual cell phase changes [11], [22] over time and to discover special distributions that the cell dwelling time follows in each phase. This requires the ability to segment, track, and classify the nuclei of a large number of image sequences acquired by time-lapse microscopy (see an example of an HTS image in Figure 2). It also requires the development of novel data models to answer those questions.

An automated system for cell phase identification would therefore have advantages over current practice, such as speed, objectivity, reliability, and repeatability. The computational architecture consists of four steps: 1) image acquisition, 2) image preprocessing and analysis, 3) feature extraction and database, and 4) phase identification data modeling. In step 1, we acquire digital microscope images from high-throughput cellular screening. In step 2, we perform automated cell segmentation and cell tracking for tracing cellular drug response. We extract cellular features from HTS images and archive images, extracted cell features, and drug response results into an image database in step 3. And, finally, in step 4, we perform cell-cycle phase identification and data modeling for detecting cellular changes with drug treatments. From the perspective of informatics, we have to address two challenges: automated cell segmentation and cell tracking, as well as data modeling. We will focus on the issues related to the second challenge.



[FIG1] Different cell phases of the cell located in the center of each image: (a) interphase, (b) prophase, (c) metaphase, and (d) anaphase.



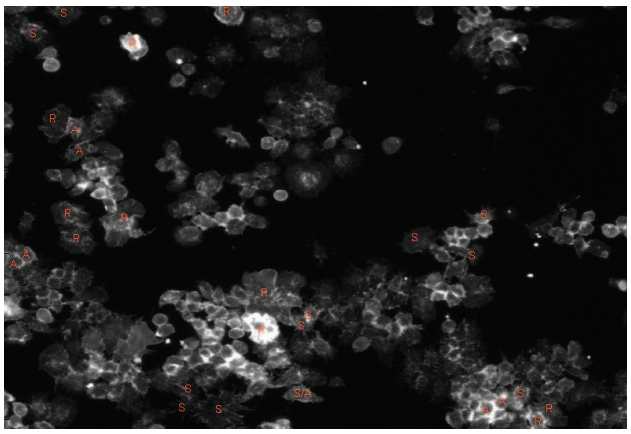
[FIG2] HTS image of nuclear DNA.

GENOME-WIDE RNAi SCREENING IN DROSOPHILA CELLS

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 [27]. In our recent work [28], we studied image-based morphological analysis in high-throughput genome-wide RNAi screening for novel effectors of Rho family GTPases in Drosophila cells. About 21,000 dsRNAs specific to the predicted Drosophila genes are robotically arrayed in 384-well plates. Drosophila cells are plated in each well and take up the dsRNA from the culture media. After incubation with the dsRNA, expression of Rac1V12, RhoAV14, or Cdc42V12 is induced. Cells are fixed, stained, and imaged by automated fluorescence microscopy. Each screen generates about 400,000 images, or millions if replicates are included. If one person analyzes this manually with the rate of 60 seconds per image set, it will take about 6,600 hours or 825 workdays to complete. Clearly, there is a demand for automated solution [27]. In this application, a cell-based assay has been developed for Rho GTPase activity using the Drosophila Kc167 embryonic cell line. Three-channel images are obtained by labeling F-actin, GFP-Rac, and DNA. Figure 3 provides an example of RNAi cell images acquired from the actin channel. Three cellular phenotypes under investigation are spiky (S), ruffling (R), and actin-acceleration-at-edge (A) [28]. The challenge is how to identify the three phenotypes automatically for each image.

HTS OF MONASTROL SUPPRESSOR DRUGS

The mitotic spindle is a temporary and critical nuclear structure that serves as the machinery for chromosome segregation. During every cycle of cell division, chromosomes must be faithfully replicated and condensed, while remaining attached to one another in preparation for the action of the microtubule spindle [1], [2]. The role of the spindle is to accurately partition the sister chromatids equally into the two daughter cells. The tubulin-based structure of the spindle is the target of many anti-cancer 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.



[FIG3] An example of an RNAi cell image in the Actin channel.

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 is a lack of effective methods for automatic recognition of cell phases. It is also difficult to identify various phases in a cell cycle from only one phenotypic image, as a single fluorescent channel simply marks components of a specific kind. We have used the multiphenotypic mitotic analysis approach [14] to identify the monastrol suppressor.

DYNAMIC SUBCELLULAR PARTICLE DETECTION AND TRACKING

Clathrin-coated pits and vesicles are found in all nucleated cells, from yeast to humans. They represent an important means by which proteins and lipids are removed from the plasma membrane (endocytosis) and transported to an internal compartment (endosome). Fluorescently labeled versions of a variety of marker proteins have given us a tantalizing glimpse of the dynamics of the system in living cells. The clathrin pathway has thus acquired special status for analyzing molecular mechanisms in membrane traffic. It is difficult to perform global analysis required for a statistically significant interpretation of the data using the existing software tools. A single image may show thousands of coated pits and vesicles forming at the surface of a cell, which do not all behave in the same fashion in time and space. Information about these objects may need to be extracted and assembled at different time points and then correlated with information from other images obtained at different spectral frequencies. We can acquire long-time-series data using wide-field, confocal, and total internal reflection fluorescence microscopy. If software with the capabilities that we envision for automated imaging analysis were available, we would be able to ask biological questions previously not possible; for example, are there "hot spots" for the formation of clathrin-coated pits, or do pits and arrays form randomly on the plasma membrane? To answer these questions, we have to first track many hundreds of individual pits as they form and study their spatial distribution dynamically [4], [29].

Since each of the imaging applications is seeking answers to a different biological question, the image processing and data modeling methods used thus vary across applications.

QUANTITATION OF HTS IMAGES

BIOIMAGE PROCESSING

Major steps in HTS image processing include: filtering, cell and spot segmentation, cell and spot tracking and registration, as well as neuron tracing and labeling. Image filters, such as deconvolution, Gaussian filtering, nonlinear filters, and anisotropic diffusion filters, are popular methods [12]. Low-pass filtering can be used to remove high-frequency noise. Nonlinear filters, such as the median filter, can be used to remove impulse noise while keeping the edge information. Anisotropic diffusion filters can filter noise without degenerating image edge information, whereas spline and polynomial fitting are often

employed to correct systematic errors of background bias. Image segmentation is the most critical step in HTS image analysis. Most histogram-based methods select a threshold via either maximizing variance between objects and background or minimizing the interclass variance of objects and background [25]. A threshold that varies over different image regions is often used to fit the spatially changing background and illumination conditions, i.e., adaptive thresholding. Other image segmentation methods involve edge detection, watershed, morphological operators, and stochastic image processing.

Although the watershed algorithm [13] is a natural way to separate touching spots and is a popular algorithm for cellular image analysis, oversegmentation is a serious drawback of using this approach. In cell biology, the cell segmentation and tracking have been extensively studied [2], [13]. Scientists can now track individual particles automatically; the review paper in [15] summarizes five algorithms, namely, the centroid method, the Gaussian fit method, the correlation method, the sum-absolute difference method, and the interpolation method for individual particle tracking. Note that those methods cannot be simply deployed in high-throughput cell or spot tracking; for example, their single spot tracking methods do not address the ambiguous association—a difficult issue in multiple particles tracking. Improved methods, such as parametric active contour [16], maximal mutual information, Bayesian method [12], and the shape-and-size based method [11], have been developed to deal with the ambiguous association.

FEATURE EXTRACTION

After cellular image segmentation, relevant image features of every cell will be extracted into numerical descriptors. Extensive work in the Murphy lab at Carnegie Mellon University on single cell structure analysis using static cell imaging techniques [17], [18] to extract a list of image features, such as area, shape, size, perimeter, intensity, texture, moment, Zernike moment, and Haralick texture [17]. On the other hand, for dynamic, population-based cellular imaging, we found that the addition of temporal parameters—such as the change of the size and shape [11] 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. Other kinds of features, such as Fourier transform and wavelet transformation [18], are also important tools for feature extraction. After extracting a large number of features for HTS images, it is important to design a proper structured database system to archive and organize these meta-data. Image database systems [21], such as Protein Subcellular Location Image Database (PSLID) [8] and Open Microscopy Environment (OME) [5], can be used to classify, organize, and represent different classes of high-throughput 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. Those features or descriptors will enable the user to query individual objects by their cellular or subcellular features or relationships.

VALIDATION

There are two categories of methods used in validation: one is based on root mean square error, and the other is statistical hypothesis testing. The concept of the first category of methods is straightforward, and such methods are widely in use. We elaborate on the second category of methods here. Given the manual result denoted by vector V_1 and the automatic result V_2, the question concerns the statistical difference between V_1 and V_2. Obviously, we can test their distributions, means, variances, and any other statistics to see if there is a significant difference between the two groups. The difference between the automatic and manual methods is often tested by using paired student t-test or one-factor repeated analysis of variance (ANOVA). Distribution tests, such as quantile plot design and Kolmogorov-Simirnov (KS), are also well-established techniques for verifying the similarity of the distributions of any two data sets. Wilcoxon matched-pairs signed-ranks test (WMPSR) is often performed for skewed datasets to compare methods' difference. Nonparametric methods, such as Mann-Whitney-Wilcoxon test or Kruscal-Wallis test, may be performed for skewed datasets.

However, when we detect or track lots of cells, proteins, and neuron spines, how can we validate the extracted results of automated analysis since it is almost impossible to perform similar manual analysis on a high-throughput scale? Manual methods can only be used to validate tens or at most up to hundreds of images. They would make lots of counting errors when analyzing hundreds and thousands of images and cannot be used to extract more detailed features of a large population of high-throughput images, such as volumes and rate of change. Thus, simulation of complex biological processes and images becomes an urgent issue. There are certain efforts trying to simulate the growth and interaction of neurons, but to the best of our knowledge, there is no equivalent level of work in simulating large populations of cells and proteins under drug treatment or RNAi perturbation. Should such simulation models exist, it would be invaluable in validating HTS image analysis algorithms without relying on laborious and costly manual validation.

Another biological validation method is to duplicate the experiments and then average the duplicates to reduce system errors, such as background noise, poor sample preparation, or bias of the imaging instruments. Scientists often resort to repeating the same experiments several times to correct such system errors. HTS replicates differ from microarray replicates, however. For microarray experiments, one can be certain of the same number of variables or genes in duplicates, but for HTS experiments, it is difficult to achieve the same number of cells or molecules in duplicate screens.

DATA MODELING AND STATISTICAL ANALYSIS

After extracting and quantitating the image features into alphanumeric metadata and archiving them into a persistent data model, one can then deploy a rich array of techniques

from statistics and data mining to analyze the accumulated metadata in the database system. Use of such analytical techniques in studying high-throughput cellular image information is growing, as large volumes of quantitative cellular image metadata become available. With the image data generated from our HTS experiments quantitated and organized in a structural data model, we are actively applying sophisticated analytic methods to further mine the database to increase the knowledge of the biological processes of interest; for instance, the understanding of cell-cycle behaviors under different drug perturbations or the improving the predictability of anti-mitotic drug treatments on cancer cells. Furthermore, as the database grows, the statistical power of the database increases. Data modeling is a crucial step in HTS because, via data modeling, we can take advantage of the large amounts of quantitated meta-data of images to investigate biological questions beyond the capability of individual laboratory experiments.

PREPROCESSING AND NORMALIZATION

Filtering is a necessary step to filter certain noise of the features. We employed an autoregression model to model the quantitative time-lapse data. There exists strong noise in the quantitative features due to imperfect segmentation; thus, modeling of the change of shape is important. In [22], we model the time-lapse data using an autoregression model, namely, we assume that the current input sample $y(n)$ is approximated by a linear combination of the past samples of the input signal. In our study, $\{y(n)\}_{n=1}^{N}$ represents the N time points of a cell feature. The prediction of $y(n)$ is computed using a finite impulse response filter. Another advantage of this model is that we do not need to perform any special normalization to the coefficients for trace identification.

A number of popular normalization techniques involve, for example, normalized by means, medial, or log (or log2) transformation. Usually, we take the log transformation if the original feature numbers are too big or their distributions approximate to a log normal distribution. Z-score transformation and normalization to interval are two popular methods used in HTS. Since Z-score transformation normalizes the data into zero means and unit standard variance, scientists usually adopt this method. Discussions of the different normalizations can be found in recent literature [3].

CLUSTER ANALYSIS

Cluster analysis received extensive investigation in microarray experiments, especially relating to the study of gene function. It is a straightforward method to capture the relationship among different variables [24]. Cluster analysis in HTS includes clustering compounds and ordered clustering for time-lapse cell traces. In drug discovery, clustering the compounds in cell-treatment experiments based on cytological phenotypes is important and useful in finding leads in new drugs. For example, good clustering would separate the effective and ineffective compounds, thus affecting the cell division under study. Also, the compounds in the same cluster may indicate certain similarity in functions [2], [14]. In time-lapse cell-cycle analysis, clustering techniques can help to identify which and when cells in the population are in apoptosis [22]. In cell shape analysis, clustering [6] can distinguish among the different cell phenotypes under various conditions. We briefly review relevant clustering algorithms and applications for HTS.

We need to denote each compound as a vector and find a way to calculate the correlation or distance between compounds before it is possible to perform cluster analysis. When comparing drug mechanisms, changes in specificity, such as phenotype, are relevant while changes in affinity, such as primary effective concentration, are not. Perlman et al. [2] developed a titration-invariant similarity score (TISS) to enable comparison between dose-response profiles independent of starting dose. TISS values were generated from N compounds that showed significant signals, and these were used for unsupervised clustering. TISS was successful at grouping compounds according to their similarities.

KS NONPARAMETRIC STATISTICS USED TO DENOTE A COMPOUND

An untreated population contains cells spread throughout the cell cycle, so measurements of nuclear area are not drawn from a normal distribution. The function $KS(f, g)$ computes $f - g$ at the point $|f - g|$ reaches the maximum, where f and g are two cumulative density functions. To assess the effectors of a compound c at a given titration t, we compute for each descriptor d the KS statistic $KS_{c,d,t} = KS_{c,d,t}(p_{c,d,t}, q_d)$, providing a quantitative measurement of a population response $p_{c,d,t}$ compared with the control population q_d. To assign a significance to the $KS_{c,d,t}$ and to normalize for descriptor variability, a z-score is computed by $z_{c,d,t} = KS_{c,d,t}/std(q_d(n))$, where n is the population size of cells used to determine $p_{c,d,t}$.

TITRATION-INVARIANT SIMILARITY SCORE FOR COMPARING COMPOUND VECTORS

Titration-invariant similarity score (TISS) [2] is developed to assess the similarity of compounds independent of the starting points of their titration series. The TISS between two compounds is calculated in three steps: 1) define the mutation of titration subseries for each compound to account for different possible starting concentrations, 2) define a correlation for pairs of these subseries, and 3) define a similarity measure derived from the strongest correlation over a determined range of these subseries. Assume that there are N compounds under study. First, for each compound c, $X_c = (z_{c,1,1}, \ldots, z_{c,D,1}, \ldots, z_{c,1,T}, \ldots, z_{c,D,T})$, where D is the number of descriptors and T is the number of titrations. To allow comparisons of compounds with different titration starting points, the titration subseries is defined as $X_c(s) = (z_{c,1,1}, \ldots, z_{c,D,1}, \ldots, z_{c,1,T-s}, \ldots, z_{c,D,T-s})$ and $X_c(-s) = (z_{c,1,s}, \ldots, z_{c,D,s}, \ldots, z_{c,1,T}, \ldots, z_{c,D,T})$. Truncating starting or ending titration allows us to shift the starting point for the titration series. Second, define s-correlation as

$$R_{ij}(s) = \frac{\langle X_i(s), X_j(-s) \rangle}{||X_i(s)|| \bullet ||X_j(-s)||} \qquad 1 \le i, j \le N.$$

For each s, the correlation matrix $X(s) = (R_{ij}(s))_{N\times N}$ using all of the compounds from each of the two replicates. Third, given a range, $-S \le s \le S$, we want to look for the value of s that gives the highest correlation between two vectors. A nonparametric ranking is employed to normalize the similarity score as $\phi_{ij}(s) = (\#\text{entries in } X(s) \ge x_{ij}(s) - 1)/N^2$. Finally, the TISS between two compound vectors is defined to be their highest correlation over all truncations. Compound clustering was restricted to the compounds whose vectors show significant difference compared to compounds of controlled experiments using hierarchical clustering. For the detailed clustering results, we refer the readers to [2].

CLUSTERING CELL MORPHOLOGICAL SHAPES

There are two issues in clustering cell morphological shapes: determining the true number of clusters and clustering cell morphological shapes. The challenging problem is to determine how many phenotypes are hidden in the data sets. Determining the true number of clusters, known as the cluster validation problem, is a fundamental problem in cluster analysis. Approaches to this problem have recently been proposed. Two different approaches have been used in validation: one based on relative criteria and the other based on external and internal criteria. The first approach is to choose the best result from a set of clustering results according to a predefined criterion, whereas the second approach is based on statistical tests and involves computation of both intercluster and intracluster quality to determine the true number of clusters. Recently, the gap statistic has been proposed as a method for estimating the number of clusters [10]. This method considers the total sum of within-class dissimilarity for a given number of clusters, data set, and the clustering solution. Since it presupposes spherically distributed clusters, it contains structural bias. Other methods, such as graphic-based cluster validation, could be used to improve this method in cell phenotype clustering analysis.

CLASSIFICATION USING STATISTICAL MODELING AND CLASSIFIERS

There are two issues in HTS classification: feature reduction and classifier design. The objectives of feature selection are: 1) improving the predictive performance of classifiers and predictors, 2) providing faster and more effective computation in the classifiers and predictors, and 3) providing a better understanding of the underlying biological process that generated the data. Two approaches for feature reduction, including transformation-based method and feature selection, are studied in [18] and related works. The former includes principal component analysis and independent component analysis. This type of approach, however, provides a poor understandability of each feature. For the latter approach, which is an active area of research in pattern recognition, the success of feature ranking depends heavily on the choice of ranking criteria. Commonly used criteria are based on correlation, goodness-of-linear-fit, or information theory. Successful choices of such criteria include ANOVA, support vector machine (SVM) and decision tree, neural networks, and mutual information. For the heuristic searching methods to search for the optimal subset of feature, specific successful applications of such methods include genetic algorithms, the minimum description length principle for model selection, and so on. Statistical hypothesis tests can be employed to implement this aim for two-class-based feature selection. There are numerous publications and methods applying statistical hypothesis tests and statistical modeling in gene microarray studies. Nevertheless, we cannot simply apply these methods to HTS studies as the data structure is inherently different; microarray typically is of small sample and large variables, whereas the reverse is true for HTS, e.g., large sample and small (a few to tens of) variables.

Regarding the classification issue, there are many well-developed approaches and strategies [18], [22] to build a classifier; not all of them are useful in HTS applications, however. K-nearest neighbor classier (KNN), artificial neural networks, SVM, ensemble methods, Gaussian mixture models, and decision trees have been compared in Murphy's lab at CMU [18] and Wong's lab at Harvard University [22], [28], respectively. Some significant work on performance comparison with different classifiers has been done for classification of cell shape, cell phase, and compound classification. From our experience, we next review a number of relevant approaches, for example, neural networks, decision tree, SVM, and random forest [30] for HTS.

SUBCELLULAR PATTERN CLASSIFICATION

The most widely used neural network classifier today is the multilayer perceptron (MLP) network, which has also been extensively analyzed and for which many learning algorithms have been developed. The principle of the network is that when data from an input pattern is presented at the input layer, the network nodes perform calculations in the successive layers until an output value is computed at each of the output nodes. The training usually adopts the backward propagation algorithm [17].

SVMs are linear classifiers that find a hyperplane between two classes, which maximizes the minimum distance between the hyperplane and data points. The hyperplane [18] can be sparsely represented by a small amount of data lying on the boundary of the maximum margin, namely, the so-called support vectors. SVM is broadly defined for two classes. Max-win classification [18] trains N SVMs, each of which separates class k from other classes. The predicted class from the machine generating the highest output score is selected as the prediction.

All current machine learning models have constraints in certain applications given limited training samples, which could result from the fact that different classifiers might be optimized for different applications. Hence, an ensemble of classifiers is a set of classifiers whose individual decisions are combined to produce an overall decision. A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of individual members is that the classifiers are accurate and diverse. AdaBoost is an example of ensemble methods that create new and improved base classifiers by iteratively manipulating the training datasets. It generates base classifiers, such as a neural network or decision tree, at each iteration. Bagging is a sort of bootstrap aggregation for ensemble methods. In each bagging iteration, the training examples are

bootstrapped to generate different training sets for the base classifiers. The authors [18] applied the above classifier to proteome-wide determination of subcellular location. High accuracy is reached in their classification systems using either manually or automatically chosen cell boundaries [18]. In the RNAi genome screening project [28], the cell phenotype recognition accuracy using KNN is between 64% and 75% when we adopted automatic segmentation. Automatic and accurate cell image segmentation and phenotype identification are still unresolved problems today.

CELL PHASE IDENTIFICATION

The authors [22] compared the mixture model, two-layer neural networks, KNN, and perceptron classifiers. The results show the mixture model and KNN have better recognition accuracy than others in cell phase identification. In [11], by combining a KNN classifier with knowledge-driven heuristic rules, the classification module successfully identifies 99.8% of the interphase cells, 83% of prophase cells, 95.5% of metaphase cells, and 95.7% of anaphase cells. In time-lapse cell cycle studies, certain context-based classifiers should be effective in dealing with such classification problems. So, we introduce hidden Markov model (HMM) in this application. Gallardo et al. [19] applied this method to study mitotic cell recognition.

Basic concepts and description of HMM have been reviewed by Rabiner [20]. Consider N distinct states of the cell cycle, e.g., interphase, prophase, and metaphase. Let q be a sequence of states with q_t being the state at time t, and let the observation sequence $O = \mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_T$, where x_t is a vector composed of the quantitative features of the cell at time t. If it is reasonable to assume that the probability of being in any one of these states is determined only by the predecessor state, then the sequence q can be considered as a first-order HMM. A matrix of state transition probabilities A with coefficients a_{ij} is defined as $a_{ij} = P[q_t = S_j | q_{t-1} = S_i]$, $1 \le i, j \le N$ where N is the number of states; S_j and S_i are different states. The coefficients a_{ij} satisfy $\Sigma_{j=1}^{N} a_{ij} = 1$ and $a_{ij} \ge 0$. The initial state of the process is a set of probabilities, denoted by $pi_j = P[q_1 = S_j]$, $1 \le j \le N$. The observation $\mathbf{x}$ is a probabilistic function of the state. Given one state j, the probability of $\mathbf{x}$ is often defined as $b_j(\mathrm{x}) = \Sigma_{m=1}^{M} c_{jm} \phi(\mathrm{x}; \mu_{jm}, \Sigma_{jm})$, $1 \le j \le N$, where $\Sigma_{m=1}^{M} c_{jm} = 1$ and $a_{jm} \ge 0$; $\phi(\cdot)$ represents the multivariate Gaussian density function and μ_{jm}, Σ_{jm} are the mean vector and covariance matrix for the m th mixture. For computational simplicity, we often set Σ_{jm} as a diagonal matrix. The standard expectation maximization (EM) algorithm is used for estimating the parameters. Given A, $\{b_j\}$, and π, an HMM can be applied to the time-lapse sequence $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_T$. The HMM model achieves the accuracy of 68% for dead cells, 88% for cell edges, and 77% for dividing cells in [19]. In summary, phase identification is still a challenging problem in HTS.

COMPOUND CLASSIFICATION

Random forest, proposed by Breiman [23], shows promising recognition accuracy. A random forest is created by growing trees, without pruning, on bootstrap samples of the data, selecting the best split at each node among a random selection of the explanatory variables. For quantitative outcomes, the forest is made of regression trees, where the tree predictor is the mean value of the training set observations in each terminal leaf. The random forest predictor is computed by averaging the tree predictors over trees for which the given observation was out-of-the-bug, i.e., not included in the bootstrap sample used to build the tree. The authors [30] compared random forest, artificial neural networks, SVM, KNN, partial least squares, decision tree, multiple linear regression, and linear discriminant analysis in predicting the categorical activities of a compound based on a quantitative description of that compound's molecular structure. They have showed that random forest has the highest accuracy among all classifiers [30].

STATISTICAL DISTRIBUTION ANALYSIS OF BIOLOGICAL EVENTS IN HTS

Perhaps the most interesting subject is the study of the statistical distribution of dwelling time and other variables or biological events in modeling time-lapse microscopy data. The complex dynamics of cell cycle and mitosis are controlled by regulatory networks consisting of numerous interacting protein assemblies. Qualitatively, much has been known about the different stages of the cell cycle and the different phases of mitosis. To the contrary, quantitatively, little is known about fundamental questions such as how long a cell within a cell population may stay in a specific state or phase and how the dwelling time may change with different perturbations [26].

The cell cycle consists of a sequence of events occurring in a strict order. The randomness in the cell cycle period found by numerous studies in the field of cell biology suggests that the length of certain phases of the cell cycle must also be random. To gain more detailed information about the dynamics occurring in each phase of the cell cycle, it is crucial to break the entire cell cycle into different phases and study what happens in each phase. We have screened five HeLa cell lines, two of which were treated with 100 nanomolar taxol, using fluorescence microscopy imaging technology. We studied the distributions for the interphase, prophase, metaphase, anaphase, and arrested metaphase, as well as the entire cell cycle, e.g., interphase. We found that the distributions are well characterized by truncated power laws. Unlike monomodal distributions such as normal and log-normal distributions, the distributions for the five phases of the cell cycle and the cell cycle period are heavy-tailed (see Figure 4). It is given by the following equation:

$$P(T > t) = \left(\frac{t}{b}\right)^{-\alpha}, \qquad t \ge b \ge 0,$$

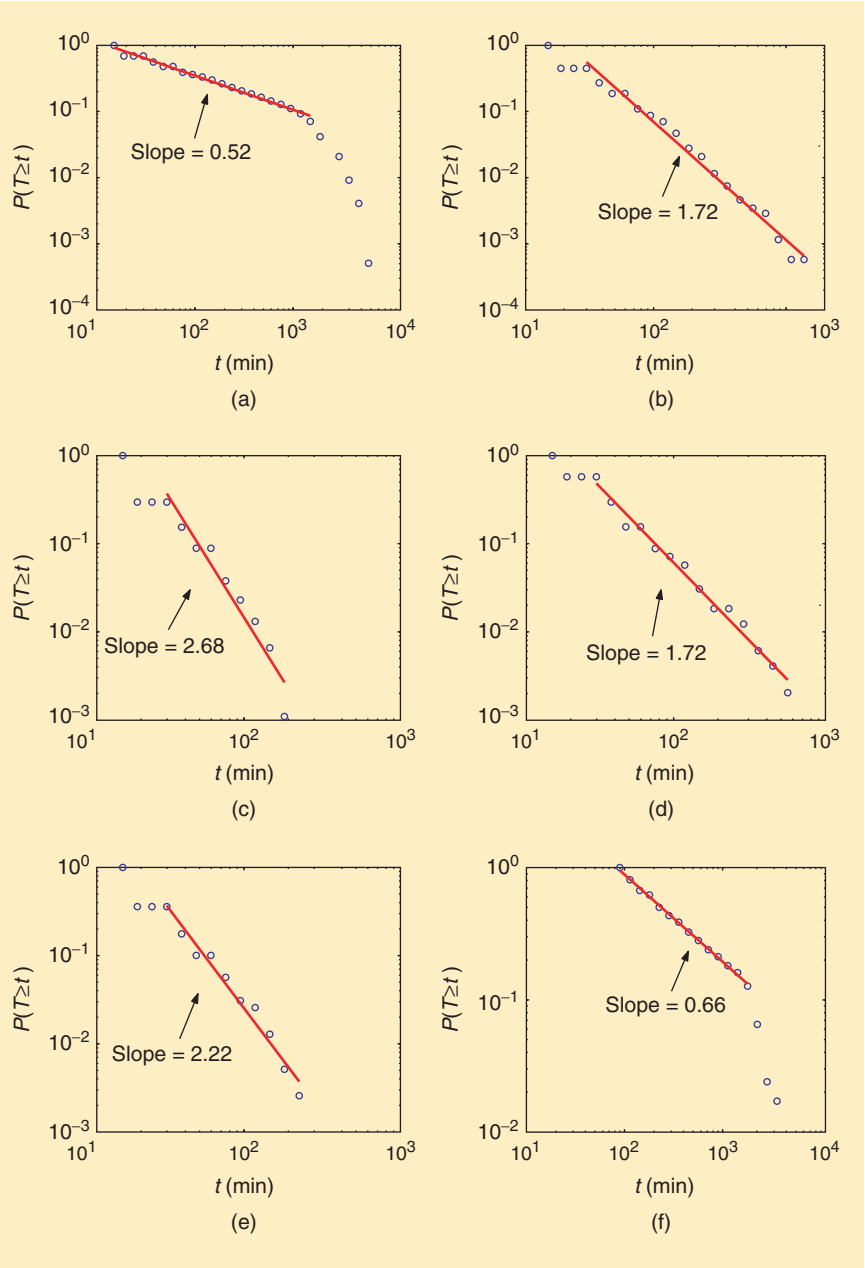
where α is the shape parameter and b is the location parameter. The equation is called the complementary cumulative distribution function.

We can test whether or not a spatial particle pattern is random by using the spatial statistics. The K function [9] captures the spatial patterns between different parts of the region where sampling takes place. In [4], the authors studied the particle spatial distribution and found the dwell time of particles in each cluster following an exponential distribution. Currently, there is

no widespread use of data mining and biostatistics methodology in studying particle morphology and dwell time since quantitative image data often are not available. Here we present an example of how to track and analyze the data in our subcellular analysis application. Assume there are N particles under study. The K function is defined to be

$$K(h) = \frac{E(d_{ij} < h)}{\lambda}, \quad 1 \le i, j \le N,$$

where $E(d_{ij} < h)$ denotes the average of the number of particles within a distance h and λ is an estimator of the intensity. K function has been used for modeling the spatial cluster structure and exponential distribution of many particles [4]. K function is particularly powerful for testing spatial randomness, i.e., whether the spatial distribution of the events is consistent with a homogeneous Poisson process. In that case, $K(h) = \pi h^2$ for all h. To simulate the direct comparison with Poisson process, first define $L(h) = (K(h)/\pi)^{1/2}$. The $L(h)$ was calculated for every data set, consisting of all observed clathrin clusters in one particular experiment. The task was then to determine whether $L(h) \approx h$. This was proved in [4]. That means the formation of clathrin-coated vesicles follows a Poisson process.



[FIG4] Log-log plot of complementary cumulative distributions of the dwelling time of (a) interphase, (b) prophase, (c) metaphase, (d) anaphase, (e) arrested metaphase, and (f) the entire circle. (The main contributor of this plot is Dr. Jinbao Gao of the University of Florida, Gainsville.)

CONCLUSIONS

In this article, we discussed the emerging informatics issues of HTS using automated fluorescence microscopy technology, otherwise known as HCS in the pharmaceutical industry. Optimal methods of scoring biomarkers and identifying candidate hits have been actively studied in academia and industry, with the exception of data modeling topics. To find candidate hits, we need to score the images associated with different compound interventions. In the application example of RNAi genome-wide screening, we aim to find the candidate effectors or genes which correspond to the images acquired using the three channels. Scoring the effectors is equivalent to scoring the images based on the number of phenotypes existing in those images. Our ultimate objective of studying HTS is to model the relationship between gene networks and cellular phenotypes, investigate cellular communication via protein interaction, and study the disease mechanism beyond the prediction based on the molecular structure of the compound.

Finally, computational image analysis has become a powerful tool in cellular and molecular biology studies. Signal processing and modeling for high-throughput image screening is an emerging filed that requires novel algorithms for dynamical system analysis, image processing, and statistical modeling. We hope that this article will motivate the signal processing communities to address challenging data modeling and other informatics issues of HTS.

ACKNOWLEDGMENTS

The authors would like to express their appreciation of the rewarding collaborations with our biological colleagues: Prof.

Tim Mitchison, Department of Systems Biology, Harvard Medical School; Prof. Tom Kirchhausen, CBR Center for Biomedical Research, Harvard Medical School; Prof. Randy King, Department of Cell Biology, Harvard Medical School; and Prof. Norbert Perrimon, Department of Genetics, Harvard Medical School. The raw image data described in this article were obtained from our biological collaborators' laboratories. The image processing and computational modeling work are the contribution of the authors with the input of other members of the life science imaging group of HCNR-Center for Bioinformatics, notably Mr. Xiaowei Chen, Dr. Xiaoyin Xu, Dr. Jinmin Zhu, Dr. Kuang-Yu Liu, and Dr. Xinhua Cao. This research is funded by the HCNR Center for Bioinformatics Research Grant, Harvard Medical School, and a NIH R01 LM008696 Grant to STCW.

AUTHORS

***Xiaobo Zhou** (zhou@crystal.harvard.edu)* received the B.S. degree in mathematics from Lanzhou University, Lanzhou, China, in 1988, and the M.S. and the Ph.D. degrees in mathematics from Peking University, Beijing, China, in 1995 and 1998, respectively. From 2003–2005, he was a research fellow with the Harvard Center for Neurodegeneration and Repair in Harvard Medical School and Radiology Department in Brigham and Women's Hospital. Since 2005, he has been an instructor and faculty member with the Harvard Center for Neurodegeneration and Repair at Harvard Medical School and Radiology Department at Brigham and Women's Hospital. His current research interests include image and signal processing for high-content molecular and cellular imaging analysis, informatics for integrated multiscale and multimodality biomedical imaging analysis, molecular imaging informatics, neuroinformatics, and bioinformatics for genomics and proteomics.

***Stephen T.C. Wong** (stephen-wong@hms.harbard.edu)* is the director of the Center for Bioinformatics, Harvard Center of Neurodegeneration and Repair (HCNR), director of the Functional and Molecular Imaging Center, and an associate professor of radiology, Harvard Medical School and Brigham and Women's Hospital. His research interests has been focused on the application of advanced technology to pragmatic biomedical problems and is based on the belief that problems of importance involve the interplay between theory and application. He is a hybrid scientist. He has published over 170 peer-reviewed papers and holds six patents in biomedical informatics. He was a key member of UCSF PACS effort, founded product development departments of Philips Medical Systems, and directed the Web trading development and rearchitecturing of Schwab.com, one of the largest secured eCommerce sites. He received his executive education from MIT Sloan School of Management, Stanford University School of Business, and Columbia University School of Business.

REFERENCES

[1] J.C. Yarrow, Y. Feng, Z.E. Perlman, T. Kirchhausen, and T.J. Mitchison, "Phenotypic screening of small molecule libraries by high throughput cell imaging," *Comb. Chem. High Throughput Screen.*, vol. 6, no. 4, pp. 279–286, 2003.

[2] Z.E. Perlman, M.D. Slack, Y. Feng, T.J. Mitchison, L.F. Wu, and S.J. Altschul, "Multidimensional drug profiling by automated microscopy," *Science*, vol. 306, no. 5699, pp. 1194–1198, 2004.

[3] D. Kevorkov, "Statistical analysis of systematic errors in high-throughput screening," *J. Biomolec. Screen.*, vol. 10, no. 6, pp. 557–567, 2005.

[4] M. Ehrlich, W. Boll, A. Van Oijen, R. Hariharan, K. Chandran, M.L. Nibert, and T. Kirchhausen, "Endocytosis by random initiation and stabilization of clathrin-coated pits," *Cell*, vol. 118, no. 5, pp. 591–605, 2004.

[5] I.G. Goldberg, C. Allan, J.-M. Burel, D. Creager, A. Falconi, H. Hochheiser, J. Johnston, J. Mellen, P.K Sorger, and J.R Swedlow, "The open microscopy environment (OME) data model and XML file: Open tools for informatics and quantitative analysis in biological imaging." *Genome Biol.,* vol. 6, no. 5, p. R47, 2005.

[6] X. Chen and R. F. Murphy, "Objective clustering of proteins based on subcellular location patterns," *J. Biomed. Biotech*. vol. 2, no. 2, pp. 87–95, 2005.

[7] W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, and E.K. O'Shea, "Global analysis of protein localization in budding yeasts," *Nature*, vol. 425, no. 6959, pp. 686–691. 2003.

[8] K. Huang, J. Lin, J.A. Gajnak, and R.F. Murphy, "Image content-based retrieval and automated interpretation of fluorescence microscope images via the protein subcellular location image database," in *Proc. IEEE Int. Symp. Biomedical Imaging*, Washington, DC, 2002, pp. 325–328.

[9] B.D. Ripley, "The second-order analysis of stationary point processes," *J. Appl. Probab.*, vol. 13, pp. 255–266, 1976.

[10] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a dataset via the gap statistic" *J. Royal Stat. Soc.* B, vol. 63, pp. 411–423, 2001.

[11] X. Chen, X. Zhou, and S.T.C. Wong, "Automated segmentation, classification, and tracking cancer cell nuclei in time-lapse microscopy," *IEEE Trans. Biomed. Eng.,* to be published.

[12] R. Eils and C. Athale, "Computational imaging in cell biology," *J. Cell Biol.*, vol. 161, no. 3, pp. 477–481, 2004.

[13] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. Image Processing*, vol. 2, no. 2, pp. 176–201, 1993.

[14] X. Zhou, X. Cao, Z.E. Perlman, and S.T.C. Wong, "A high content image analysis system for Monastrol suppressor screening based on the biological spindle model," *J. Biomed. Inform.*, to be published.

[15] M.K Cheezum, W.F. Walker, and W.H. Guilford, "Quantitative comparison of algorithms for tracking single fluorescent particles," *Biophys. J.,* vol. 81, no. 4, pp. 2378–2388, 2001.

[16] Z. Christophe, L. Elisabeth, M.Y. Vannary, G. Nancy, and O.M. Jean-Christophe, "Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: A tool for cell-based drug testing," *IEEE Trans. Med. Imag.*, vol. 21, no. 10, pp. 1212–1221, 2002.

[17] M.V. Boland and R.F. Murphy, "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells," *Bioinform.*, vol. 17, no. 12, pp. 1213–1223, 2001.

[18] K. Huang and R.F. Murphy, "Boosting accuracy of automated classification of fluorescence microscope images for location proteomics," *BMC Bioinform.,* vol. 5, no. 5, p. 78, 2004.

[19] G Gallardo, F. Yang, M.A. Mackey, F. Ianzini, and M. Sonka, "Mitotic cell recognition with hidden Markov models," *Medm Imagm Proc. SPIE*, vol. 5337, pp. 661–668, 2004.

[20] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[21] S.T.C. Wong, Ed. *Medical Image Databases*. Norwell, MA: Kluwer, 1998.

[22] X. Zhou, X. Chen, K.L. Liu, and S.T.C. Wong, "Time-lapse cell cycle quantitative data analysis using Gaussian mixture models," in *Life Science Data Mining*, S.T.C. Wong and C.S. Li, Eds. Singapore: World Scientific, to be published.

[23] L. Breiman, "Random forests," *Machine Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[24] M. Schena and S. Knudsen, *Guide to Analysis of DNA Microarray Data*. Wilmington, DE: Wiley-Liss, 2002.

[25] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Elect. Imag.*, vol. 13, no. 1, pp. 146–165, 2004.

[26] X. Zhou and S.T.C. Wong, "High content cellular imaging for drug development," *IEEE Signal Processing Mag.*, to be published.

[27] M. Boutros, et al., "Genome-wide RNAi analysis of growth and viability in drosophila cells," *Science*, vol. 303, no. 5659, pp. 832–835, 2004.

[28] X. Zhou, K.L., Liu, and S.T.C. Wong, *Towards Automated Cellular Image Segmentation for RNAi Genome-Wide Screening*, vol. 3749 (Lecture Notes in Computer Science). Berlin: Springer, 2005, pp. 885–892.

[29] X. Zhou and S.T.C. Wong, "Dynamic sub-cellular behavior study in high content imaging using a novel approach," in *Proc. IEEE Conf. Circuits and Systems*, Kobe, Japan, May 2005.

[30] V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan, and B.P. Feuston, "Random forest: A classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci*. vol. 43, no. 6, pp. 1947–1958, 2003. SP