J Biomol Screen OnlineFirst, published on May 20, 2010 as doi:10.1177/1087057110370894

# Automated Image Analysis for High-Content Screening and Analysis

AABID SHARIFF,[1,2] JOSHUA KANGAS,[1,2] LUIS PEDRO COELHO,[1,2] SHANNON QUINN,[1,3] and ROBERT F. MURPHY[1,2,3,4,5]

The field of high-content screening and analysis consists of a set of methodologies for automated discovery in cell biology and drug development using large amounts of image data. In most cases, imaging is carried out by automated microscopes, often assisted by automated liquid handling and cell culture. Image processing, computer vision, and machine learning are used to automatically process high-dimensional image data into meaningful cell biological results. The key is creating *automated* analysis pipelines typically consisting of 4 basic steps: (1) image processing (normalization, segmentation, tracing, tracking), (2) spatial transformation to bring images to a common reference frame (registration), (3) computation of image features, and (4) machine learning for modeling and interpretation of data. An overview of these image analysis tools is presented here, along with brief descriptions of a few applications. (*Journal of Biomolecular Screening* XXXX:xx-xx)

## INTRODUCTION

THE TERM *HIGH-CONTENT ANALYSIS* (HCA) IS FREQUENTLY USED to describe the combining of approaches from image processing, computer vision, and machine learning to provide fast and objective methods for analyzing large amounts of bioimage data. It includes the image analysis methods used for high-content screening (HCS), which focuses on identifying compounds or genes that can produce a desired cellular behavior. The field's origins in the mid-1990s were in the development of automated microscope systems that included hand-constructed automated analysis algorithms[1] and the successful application of machine learning methods to recognize subcellular patterns.[2] The algorithms behind HCA are firmly rooted in signal processing, providing a sound theoretical foundation for using machine learning techniques to extract meaningful information from large sets of bioimages. Spatiotemporal events within a cell can be captured by microscopy and quantified through image processing and machine learning methods to produce meaningful conclusions about the data within the experimental context. HCS methods aimed at drug discovery include applications that are based on screening for chemicals and toxins in lead discovery[3,4] and gene function discovery using RNAi screens.[5] The focus is on the basic understanding of the physiology of *target proteins* (such as G-protein-coupled receptors or protein kinases) and *target behaviors* (such as cell motility or secretion) and on the screening of candidate compounds or genes to find those that selectively act on the *target*. Example applications of screens include cell assays characterizing significant cellular phenotypes in response to small molecule or RNAi in areas such as stem cell differentiation, apoptosis, tumor biology, neurodegenerative disorders, arterial hypertension, and many others. The bioimage data for HCS are typically collected using fluorescent tags or stains to identify points of interest within the cells being imaged. By combining high-throughput cell biology and automated image analysis, a much larger number of experiments can be performed, and the subjectivity of the experimental observer can be minimized.

For example, HCS has had a tremendous impact on neuroscience drug discovery, enabling researchers to examine large amounts of drug and neuronal cell interactions at different time intervals and at a spatial resolution. This permits levels of sensitivity and objectivity not previously possible with spectrophotometric experimentation. Automated image analysis methods such as tracing, which we discuss in this article, have proven to be advantageous over manual methods for studying phenotypic changes in neurite extensions in response to drugs.[6]

[1]Lane Center for Computational Biology and Center for Bioimage Informatics, Carnegie Mellon University, Pittsburgh, PA.
[2]Joint Carnegie Mellon University–University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh, PA.
[3]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA.
[4]Departments of Biomedical Engineering and Machine Learning, Carnegie Mellon University, Pittsburgh, PA.
[5]Freiburg Institute for Advanced Studies, Albert Ludwig University of Freiburg, Freiburg, Germany.

Received Jan 25, 2010, and in revised form Mar 30, 2010. Accepted for publication Apr 1, 2010.

Journal of Biomolecular Screening XX(X); XXXX
DOI: 10.1177/1087057110370894

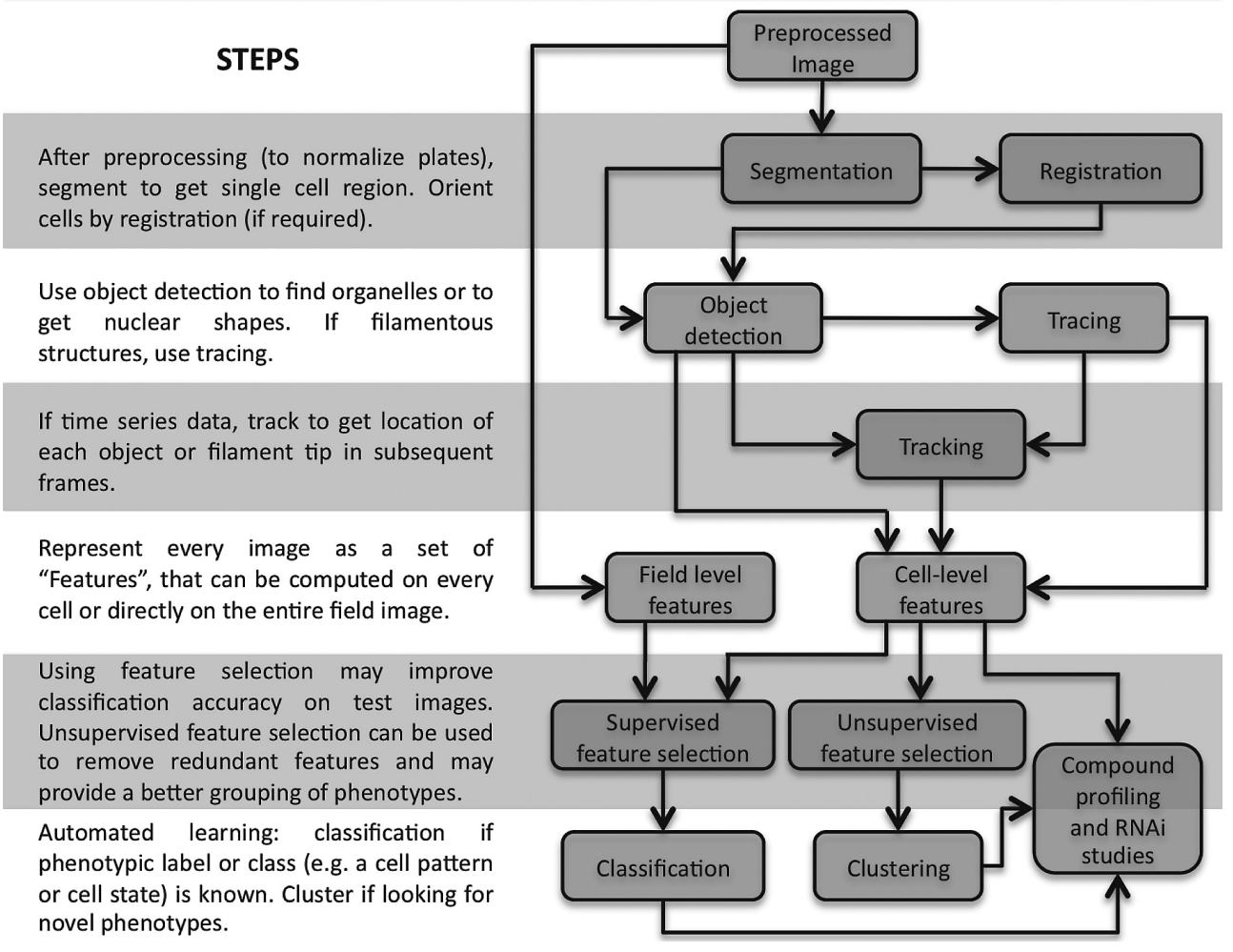


**FIG. 1.** Steps in typical image analysis pipelines for high-content analysis (HCA).

### *Overview of the image analysis tools*

This review focuses on the fundamental image analysis methods used in HCA. **Figure 1** shows a general overview of the common steps in HCA pipelines. The reader is encouraged to frequently refer to this pipeline while reading this article. Images are typically first normalized by preprocessing and then processed by segmentation to identify single cell regions. Depending on the experiment, tracing or tracking may be required. Features are then extracted from each region of the segmented image and/or from the results of tracing or tracking. Each cell is thus represented by a multidimensional feature vector. In some cases, features may be calculated on the whole image field without segmentation. In either case, the features are then used for classification or clustering or both depending on the application.

## NOTES ON DATA ACQUISITION FOR HCA

When optimizing an HCA process, multiple options must be weighed for their costs and benefits. One must consider the type of microscopy (transmitted light, widefield fluorescence, or confocal fluorescence), the number of imaging channels, the objective magnification and camera pixel size, whether to acquire 2D or 3D images, whether to acquire a single time point or a time series, and so on. The answers to these options lie in the biology of the cell assay and the trade-offs between speed and quality of the image acquisition. The number of imaging channels can have an effect on the segmentation methods used (see segmentation section) and on the ability to use colocalization information. The objective magnification and camera pixel size should be determined based on whether the goal is to image small objects in cells or individual cells or populations of cells. The acquisition of 2D images can be more rapid than capturing 3D images while yielding less information.

One challenge of HCS is to determine how many images should be acquired. The amount of information required can vary greatly from task to task. In general, it is best to be able to accurately characterize the variability for a given cell type under a given condition. Hence the more data, the better. In a task where the desire is to assign a label to a protein location

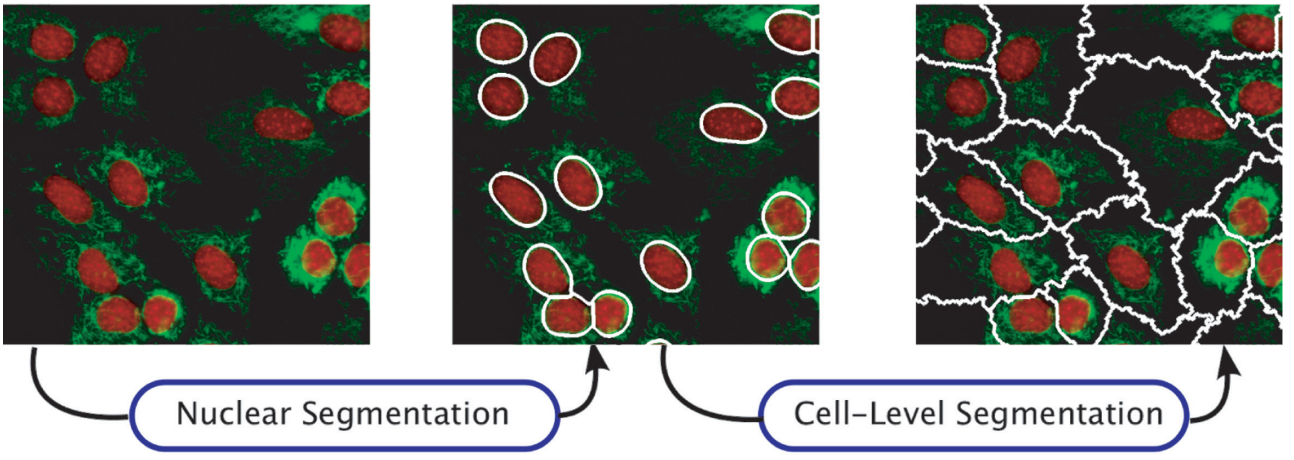


**FIG. 2.** Two-step methods in cell segmentation. Nuclei boundaries are first detected using a model-based approach[15] and then smoothed to use as inputs to seeded watershed segmentation for separating cells.

pattern (see supervised learning), images for 1 to 10 cells can be sufficient to place a protein into a known location class. However, it may require images for as many as 50 to 100 cells to adequately learn a new category.[7]

Although current HCA applications typically involve acquiring images of a fixed size for a fixed number of fields at a fixed number of time points, it is also possible to vary these parameters during the acquisition process based on analysis of the images acquired previously. The interested reader is pointed to intelligent acquisition algorithms that vary when and where to acquire images.[8,9]

## PREPROCESSING

Images often exhibit variations due to uneven illumination across the image because of imperfections in the optical system or differences between different runs of the imaging pipeline. These effects are uninteresting but can lead to processing problems. For example, if the top of the image is more brightly lit, this could lead a naive algorithm to wrongly assign cells that happen to lie on that region of the image a higher marker expression value.

Pixel-level preprocessing can ameliorate some of these artifacts. For estimating the uneven illumination, one typically computes the mean (or median) pixel value at each location across a large number of fields followed by smoothing (either Gaussian smoothing or fitting a plane or a parabola to the data). Then each pixel is normalized by dividing it by the mean pixel value at that location.

Another typical preprocessing step is to perform contrast stretching so that every pixel value lies between a predetermined interval. This is especially important if different dyes are being compared against each other as their relative brightness might be irrelevant.

## CELL SEGMENTATION AND OBJECT DETECTION

Most HCA analyses use image segmentation to separate cells in a field. Some commonly used methods for segmentation are Voronoi segmentation, model-based methods, seeded watershed, active contour-based approaches, graphical model segmentation, and active masks.[10-15] If tissues with both normal and cancerous phenotypes need to be separated, then graph partitions, clustering, or histogram-based segmentation methods are useful.

Most segmentation approaches are 2-step methods. Voronoi segmentation and seeded watershed require seed regions. In some traditional settings, a human operator would define the seeds manually, but this is not a solution for the large numbers of images generated by a screening study. Thus, automated methods are used. Nuclear-level segmentation, on a separately acquired nuclear channel, is often used to provide seeds for cell-level segmentation. **Figure 2** shows an example of a 2-step segmentation approach where nuclei are used as seeds for seeded watershed cell segmentation. Similarly, active contour methods require a window around the cell to be segmented, and again, a coarse boundary of the nuclei can be used. The coarse boundary is deformed iteratively to output the boundary of the cell. Graph-partitioning methods perform well when the regions to be segmented are coarsely labeled.

The best segmentation results are obtained by complex methods that take multiple aspects into account (active contours and graphical models fall into this category). However, these methods are also often very expensive computationally (taking from seconds to hours per image). Therefore, the full algorithm is sometimes only approximated, or on other occasions, it is even preferable to use a faster method such Voronoi or watershed (which takes less than 1 s per image). Some judicious filtering, such as removing objects that are too small or

too large to be a cell, can improve the quality of the results at the cost of losing some areas of the image. For many high-throughput screens, this might be the right trade-off.

Object detection can be used to get the shapes of objects such as cell, nucleus, vesicle, and other organelle boundaries. Segmentation approaches described here or Ridler-Calvard thresholding can be used to get the boundaries from their respective channels. For example, nuclear boundaries and nuclear spots can be acquired from the nucleus channel, coarse filamentous objects from the actin or microtubule channel, and small objects from vesicle channels such as lyososomes and endosomes.

## TRACING

In some experiments, it may be useful to quantify the numbers, lengths, and relative sizes of branching structures in images at multiple scales. In HCA applications, these may include vasculature, neurites, and microtubules.

Three major methods are used for tracing. The first method is called skeletonization.[16,17] In this approach, an image is first segmented or coarsely thresholded. The remaining pixels (or voxels in 3 dimensions) are then removed systematically by considering the surrounding neighborhood. This process leaves a skeleton structure of the image. This skeletal structure can then be analyzed for the features of interest. The second method is often referred to as vectorizing.[18] This method is an exploratory method in that only a relatively small section of the image is analyzed in a step. A starting point is discovered in the image either manually or automatically. Once this starting point is identified, the algorithm recursively explores the region of interest. This method is significantly faster than the previous method because it considers only the regions of interest and not the entire image. The third method uses superellipsoids,[19] which are generally cylindroidal (cylinders with an elliptical cross section) as a model for the structures of interest. By modeling the structure in this manner, the structures can be represented compactly, and important features of the objects can be easily calculated. These methods have been shown to be effective with significant levels of noise in the images.

The general strategy for picking the method of choice is to compare the results from each of these methods with ground truth acquired by manual tracing of the filaments by an expert.

## TRACKING

To study the dynamics of movement inside cells, objects may be tracked from one image frame to the next. Even when intracellular movement itself is not the object of study, it may be necessary to track *cells* from one frame to the next to study their behavior. Sigal et al.[20] studied the cell cycle in an unsynchronized population by computationally aligning the trajectories. State-of-the-art methods for tracking in fluorescence imaging have been reviewed recently.[21]

The traditional approach for tracking is to separate object detection and tracking steps. In the first step, a list of objects is generated. The simplest method is to threshold the image and then characterize all contiguous above-threshold regions as objects. The tracking (or linking) step consists of assigning objects in one frame to objects in the adjacent frame.

Objects are characterized by a set of measurements $\{\boldsymbol{x}_i\}$, that include the position $(x,y)$ and any other measurements considered valuable (such as the object size, brightness, and shape features). Linking is done by defining a distance between 2 objects $d(x_i,x_j)$, typically using the normalized Euclidean distance. Minimizing the total distance involved in linking objects in 2 adjacent frames leads to the Hungarian algorithm, which is simple, deterministic, and very fast.

For harder problems, one needs to take into account multiple frames to model inertia in movement. State-of-the-art approaches are based on scoring a whole set of tracks at once[22] or particle filtering, a model-based probabilistic approach.[23]

## REGISTRATION

Image registration is the application of a geometric transformation to align an object in one image to a template object in another image. Registration methods include point-based, surface-based, or intensity-based methods.[24] Point-based methods align corresponding pairs of feature points that can be found a priori. Surface-based methods compute and align the 3D boundary surfaces of the 2 objects in the images. Intensity-based methods are increasingly becoming the most popular among the registration approaches. They transform the pixel or voxel values in an iterative fashion by optimizing a similarity score between the 2 images. There are many similarity scores published in the literature: some common examples are based on least squares, cross-correlation, and mutual information.[24] Intensity-based approaches are generally a good method to use as a starting point for many of the high-content applications.

This step in automated analysis is necessary if the features computed are not rotation invariant (see Image Features section). An important application of this tool is for alignment of successive slices in a 3D image.[25] This step is especially important if acquisition time of a single 3D image is long for live-cell imaging, where artifacts such as cell movement are possible. Another application would be in the alignment of immunohistochemistry tissue samples that have a high slice-to-slice variation because of tissue damage during slice preparation.[26]

## IMAGE FEATURES

Image features, numerical descriptors that can be computed directly from an image to represent its important aspects, form

the heart of HCS and HCA systems. These features can be computed from 2D or 3D images or 2D or 3D time series. They can be derived from a single fluorescence channel or from 2 or more channels collected for the same field. Some features require presegmentation of the image into the single-cell region, whereas field-level features do not. Field-level features can be computed when the patterns in different cells within the field are fairly homogeneous. For analysis of cell patterns, features computed should preferably have properties such as invariance to image rotation or translation. If not, the images must be registered (see Registration section) before features can be computed. Example features include Haralick texture features, Zernike moment features, morphological features, object-based features, wavelet and frequency transform coefficients, threshold adjacency statistics, features from multiresolution subspaces, and others.[27-30] For any given HCA application, morphological and Haralick texture features generally serve as a good starting set of features because they often yield good classification accuracies. However, not all features are important for every application, and feature selection can be used to identify them (see below). Despite the vast set of features mentioned above, problem-specific features need to be designed for cell biology problems to improve classification accuracies (see Automated Learning Paradigms sections)—for example, the extent of overlap between a protein and a nuclear marker or edge features for microtubules.[31] Some features can also be parameters of a generative model that comprehensively describes the pattern in an image. For example, the parameters of object-based subcellular pattern models have been demonstrated to be capable of distinguishing major subcellular patterns nearly as well as descriptive features.[32]

### *Feature subset selection and recombination*

Not all features that can be computed may be useful for a desired task. For some machine learning algorithms, the presence of large numbers of uninformative or redundant features may inhibit performance. In such cases, feature selection methods can be used to select a subset of the features that are most informative in discriminating the various classes. Stepwise discriminant analysis (SDA) is one such method where the criteria for selection are based on statistical tests at every step as the number of features selected is increased.[33] A number of other methods have been described, and a comparison of their performance for subcellular pattern classification has been presented.[34] Since that study, additional methods such as minimum redundancy maximum relevance have been described.[35]

An alternative to feature selection is to create new sets of features by recombining the original features. The basic idea is to project the feature data to a lower dimensional space whose bases are computed by solving an optimization problem. Linear supervised approaches are called linear discriminant analysis, and a popular method is Fisher linear discriminant, where the features are weighted to output a lower dimensional feature vector. If the manifold of feature space is nonlinear, higher order features can also be created by using kernel methods with the goal to improve accuracy of classification between the various classes of images. In addition to supervised approaches, unsupervised approaches such as principal components analysis (PCA) or independent components analysis (ICA) can be used when labeled data are not available. Depending on the data, many variations of feature recombination algorithms can be designed by modifying the objective function to be optimized. Examples include maximum variance unfolding, nonlinear PCA, and Isomap.

## AUTOMATED LEARNING PARADIGMS

Over the past 30 years, there has been tremendous growth in the computational methods for automated learning and discovery. The discipline of machine learning emerged from the field of artificial intelligence that had previously been dominated by rule-based, knowledge capture approaches. The essential characteristic of machine learning systems is their ability to improve their performance with experience. There are 3 basic paradigms: supervised, unsupervised, and semi-supervised learning. Once features have been extracted and computed, supervised classification methods can be used to recognize different classes of samples, such as drugs that do and do not cause a desired change or normal and diseased phenotypes in pathology studies. Unsupervised clustering, on the other hand, can identify novel phenotypes. Semi-supervised learning employs both supervised and unsupervised methods.

### *Supervised*

Supervised learning is a paradigm of machine learning concerned with performing classification and regression on labeled data to build a concise model of the distribution of class labels.[36] Example labels or classes for a cytotoxicity assay include (1) normal, (2) necrotic, and (3) apoptotic. Given a set of classes a priori and example members of such classes, supervised learning techniques can learn a classifier (a function) that can assign new data points to one of the classes.

This sort of approach is appropriate, for example, in classifying shape. The properties of shape, such as width, circumference, or convexity, can be quantified, and these morphological features can be used to create a classifier. For example, in nuclear shape analysis, intensity and nuclear *spot* features were computed and used to classify and profile nuclear phenotypes.[37] Another example involves determining protein subcellular location patterns: given a set of subcellular protein location patterns observed through microscopy, supervised classification algorithms such as neural networks or support vector machines can assign a location within the cell to the protein pattern.[7] Classification can also be used to reject out-of-focus

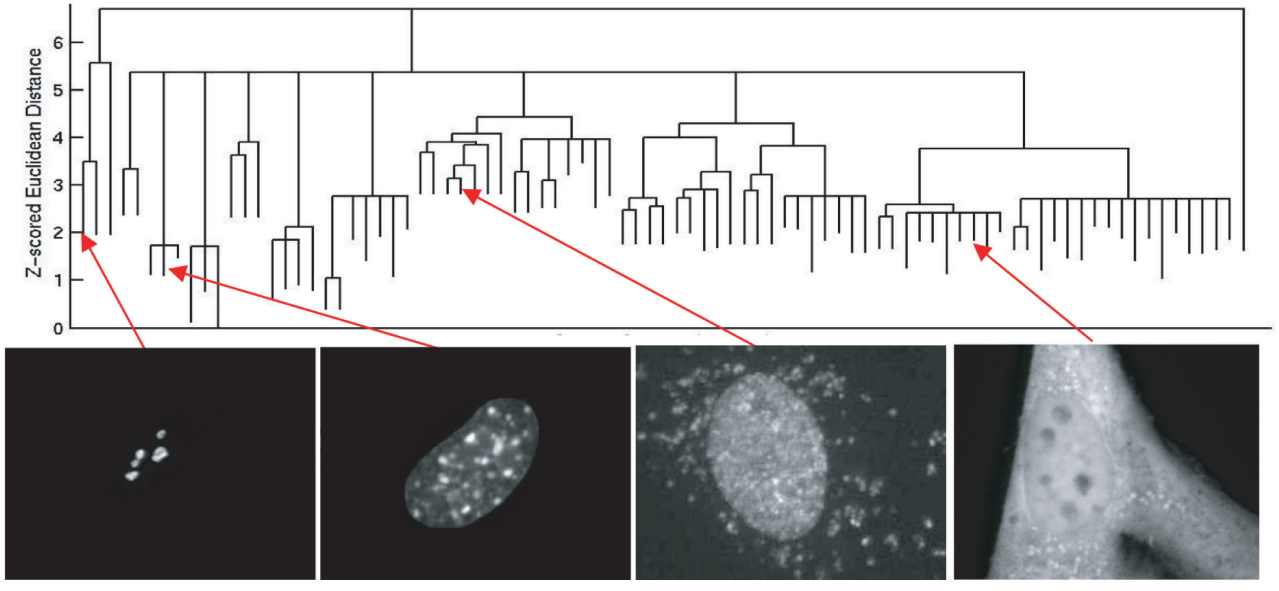


**FIG. 3.** Hierarchical clustering of subcellular localization patterns in CD-tagged 3T3 cells. Four representative images from different clusters in the tree are shown. Adapted from Murphy.[56]

images, by labeling such images as a separate class in a classifier.

### *Unsupervised*

Whereas supervised learning deals only with labeled data and seeks to either classify or regress against the data, unsupervised learning makes no initial assumptions about how the data are related and instead seeks to discover and characterize the hidden distribution of data. In unsupervised learning, or clustering, the data are unlabeled, and algorithms such as self-organizing maps, k-means, or hierarchical clustering are used to observe how the data group together and measure the distances between data points in the feature space. While clustering, 2 important considerations must be taken into account: (1) the distance metric used in feature space and (2) the number of clusters, where the Akaike information criterion (AIC) is often used in combination with the model likelihood to estimate the optimal number of clusters. A helpful discussion of using AIC to pick the number of clusters has been presented recently.[38]

These methods can be extremely useful to distinguish images in which the general phenotypes, or the effects of added drugs, are unknown. Early examples of the application of unsupervised methods in HCA include building subcellular location trees for large numbers of randomly tagged proteins[39] and learning how centrosome duplication is affected by a number of different drugs.[40] In these cases, the unsupervised approach is given a set of images and clusters them according to the fundamental patterns they contain. **Figure 3** shows an example in which a tree of subcellular protein locations is created by hierarchical clustering of subcellular patterns, based on their z-scored Euclidean distance.

### *Semi-Supervised*

Unlike supervised learning methods, where all data are labeled, and unsupervised learning methods, where all data are unlabeled, the semi-supervised learning paradigm addresses the presence of both labeled and unlabeled data. In some cases, semi-supervised learning methods can be used to augment supervised learning algorithms where data are scarce and can also improve unsupervised learning by incorporating a small amount of known data.[41] Although there may be many known patterns in the data being analyzed, the possibility remains that some have yet to be observed. Using a small amount of labeled data, the large quantities of unlabeled images available in HCA can be used as training and testing instances along with the labeled data, identifying and classifying known patterns and possibly exposing previously unobserved and novel patterns.

Most current HCA applications use supervised methods, but semi-supervised methods are now being applied. For example, a method relying on semi-supervised learning but also on transductive learning, to distinguish subcellular organelle patterns, has been described.[42]

## APPLICATIONS

Automated image analyses have been reported for many HCS applications that are based on gene expression, RNAi, and small-molecule screens, and many more unpublished studies

have been carried out within the pharmaceutical and biotechnology industries.

In small-molecule screens, the goal is to identify a set of small molecules that cause a phenotypic change.[43] However, an additional challenge would be to identify the biochemical target of that small molecule. An example of a small-molecule screen using automated image analysis is done by Tanaka et al.[44] Furthermore, image analysis methods can also be used to profile the drug dosage phenotypic response of various drugs.[45]

The goal of RNAi screens is identifying a set of genes that express mutant phenotypes when inhibited by siRNA interference.[5] Depending on the pathway that is of interest, cells with appropriate biomarkers are imaged under a fluorescence microscope (e.g., tagged tubulin would be an appropriate marker for studying cytoskeleton reorganization) after treatment with one of a library of siRNAs. The images are segmented, registered, and features extracted from every cell and summarized for every siRNA as *well parameters*. Example well parameters based on *number of nuclear spots* include fraction of cells with varying numbers of nuclear spots. Using these parameters, genes (or siRNA) are scored and statistical tests are performed to identify unique genes that could be involved in the pathway.[46] Researchers have reported using this technology for identifying genes involved in mitotic spindle assembly,[47,48] cell morphology,[49] viral infection,[50] and others. Recently, regression modeling was proposed for scoring images to predict the biological relevance of genes in RNAi screens.[51]

## DISCUSSION

This review presents an overview of automated image analysis methods used in HCA. HCA is a relatively new approach to life sciences that adds a spatial dimension to vast amounts of cell biology data for drug discovery that have been made possible because of advancements in the throughput of transmitted light and fluorescence microscopy. Because of this explosion in the amounts of image data, image analysis has become the bottleneck in the HCA process. Recent advances in microscopy image analysis tools that are based on the framework of machine learning have provided approaches that yield high accuracy.[28]

However, one important analysis step to be performed after every automated image analysis section is *validation*. Each of the sections discussed here can be analyzed using many approaches mentioned in this review, but the best one is picked on the basis of validation where most strategies are based on quantifying accuracy by comparing predictions with the ground truth.

As more high content data are available, online strategies for analysis must be made available as well as database methods to query new data or retrieve existing data. Operations that image database systems must provide include choosing a set of images based on image metadata, picking a representative image from a set of images, finding the most similar image to a given query image, comparing distributions of images under different conditions, and clustering images by their pattern. Many database systems for HCA have been described such as the Protein Subcellular Location Image Database,[52] Open Microscopy Environment,[53] or the Cell Centered Database.[54] In addition, the Human Protein Atlas provides a major source of cell and tissue images showing the patterns of thousands of proteins.[55]

The image analysis tools briefly reviewed here can be expected to be used increasingly in new HCA applications to minimize human effort, improve accuracy, and, most important, provide the structured information necessary for the success of systems biology and personalized medicine.

## REFERENCES


1. Giuliano K, DeBiasio R, Dunlay T, Gough A, Volosky J, Zock J, et al: High-content screening: a new approach to easing key bottlenecks in the drug discovery process. *J Biomol Screen* 1997;2:249-259.
2. Boland MV, Markey MK, Murphy RF: Classification of protein localization patterns obtained via fluorescence light microscopy. *Proc IEEE Int Conf EMBS Soc* 1997:594-597.
3. Bleicher KH, Bohm HJ, Muller K, Alanine AI: Hit and lead generation: beyond high-throughput screening. *Nat Rev Drug Discov* 2003;2:369-378.
4. Taylor DL: Past, present and future of high content screening and the field of cellomics. *Methods Mol Biol* 2007;356:3-18.
5. Carpenter AE, Sabatini DM: Systematic genome-wide screens of gene function. *Nat Rev Genet* 2004;5:11-22.
6. Daub A, Sharma P, Finkbeiner S: High-content screening of primary neurons: ready for prime time. *Curr Opin Neurobiol* 2009;19:537-543.
7. Glory E, Murphy RF: Automated subcellular location determination and high-throughput microscopy. *Dev Cell* 2007;12:7-16.
8. Jackson C, Murphy RF, Kovacevic J: Efficient acquisition and learning of fluorescence microscopy data models. *IEEE Int Conf Image Proc* 2007;6:245-248.
9. Jackson C, Murphy RF, Kovacevic J: Intelligent acquisition and learning of fluorescence microscope data models. *IEEE Trans Image Proc* 2009;18: 2071-2084.
10. Chan TF, Vese LA: Active contours without edges. *IEEE Trans Image Proc* 2001;10:266-277.
11. Jones TR, Carpenter AE, Golland P: Voronoi-based segmentation of cells on image manifolds. In Liu Y, Jiang T, Zhang C (eds): *Computer Vision for Biomedical Image Applications*. Berlin: Springer, 2005:535-543.
12. Chen S-C, Zhao T, Gordon GJ, Murphy RF: A novel graphical model approach to segmenting cell images. *Proc IEEE Symp Comput Intell Bioinform Comput Biol* 2006:1-8.
13. Srinivasa G, Fickus MC, Guo Y, Linstedt AD, Kovacevic J: Active mask segmentation of fluorescence microscope images. *IEEE Trans Image Proc* 2009;18:1817-1829.
14. Gould S, Gao T, Koller D: Region-based segmentation and object detection. In *Advances in Neural Information Processing Systems (NIPS*

*2009)*. La Jolla, CA: Neural Information Processing Systems Foundation, 2009.
15. Lin G, Adiga U, Olson K, Guzowski JF, Barnes CA, Roysam B: A hybrid 3D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. *Cytometry Part A* 2003;56A:23-36.
16. Cohen AR, Roysam B, Turner JN: Automated tracing and volume measurements of neurons from 3-D confocal fluorescence microscopy data. *J Microsc* 1994;173:103-114.
17. He W, Hamilton TA, Cohen AR, Holmes TJ, Pace C, Szarowski DH, et al: Automated three-dimensional tracing of neurons in confocal and bright-field images. *Microsc Microanal* 2003;9:296-310.
18. Al-Kofahi K, Lasek S, Szarowski DH, Page CJ, Nagy G, Turner JN, et al: Rapid automated three-dimensional tracing of neurons from confocal image stacks. *IEEE Trans Inform Tech Biomed* 2002;6:171-187.
19. Tyrrell JA, di Tomaso E, Fuja D, Tong R, Kozak K, Jain RK, et al: Robust 3-D modeling of vasculature imagery using superellipsoids. *IEEE Trans Med Imaging* 2007;26:223-237.
20. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Alaluf I, et al: Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nat Methods* 2006;3:525-531.
21. Genovesio A, Olivo-Martin J-C: Particle tracking in 3D+t biological imaging. In Rittscher J, Machiraju R, Wong STC (eds): *Microscopic Image Analysis for Life Science Applications*. Norwood, MA: Artech House, 2008:223-282.
22. Li K, Miller ED, Chen M, Kanade T, Weiss LE, Campbell PG: Cell population tracking and lineage construction with spatiotemporal context. *Med Image Anal* 2008;12:546-566.
23. Smal I, Draegestein K, Galjart N, Niessen W, Meijering E: Particle filtering for multiple object tracking in dynamic fluorescence microscopy images: application to microtubule growth analysis. *IEEE Trans Med Imaging* 2008;27:789-804.
24. Fitzpatrick JM, Hill DLG, Maurer CR Jr: Image registration. In Fitzpatrick JM, Sonka M (eds): *Handbook of Medical Imaging: Volume 2. Medical Image Processing and Analysis*. Bellingham, WA: SPIE—The International Society for Optical Engineering, 2000:447-513.
25. Yang S, Kohler D, Teller K, Cremer T, Le Baccon P, Heard E, et al: Nonrigid registration of 3-d multichannel microscopy images of cell nuclei. *IEEE Trans Image Process* 2008;17:493-499.
26. Mosaliganti K, Pan T, Sharp R, Ridgway R, Iyengar S, Gulacy A, et al: Registration and 3D visualization of large microscopy images. *Proc SPIE Ann Med Imaging Meetings* 2006;6144:923-934.
27. Boland MV, Murphy RF: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics* 2001;17:1213-1223.
28. Huang K, Murphy RF: From quantitative microscopy to automated image understanding. *J Biomed Opt* 2004;9:893-912.
29. Chebira A, Barbotin Y, Jackson C, Merryman T, Srinivasa G, Murphy RF, et al: A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinform* 2007;8:210.
30. Hamilton N, Pantelic R, Hanson K, Teasdale R: Fast automated cell phenotype image classification. *BMC Bioinform* 2007;8:110.
31. Newberg J, Murphy RF: A framework for the automated analysis of subcellular patterns in human protein atlas images. *J Proteome Res* 2008;7:2300-2308.
32. Zhao T, Murphy RF: Automated learning of generative models for subcellular location: building blocks for systems biology. *Cytometry Part A* 2007;71A:978-990.
33. Jennrich RI: Stepwise discriminant analysis. In Ralston A, Wilf HS, Enslein K (eds): *Statistical Methods for Digital Computers*. Vol. 3. New York: John Wiley, 1976:79-95.
34. Huang K, Velliste M, Murphy RF: Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proc SPIE* 2003;4962:307-318.
35. Peng H, Long F, Ding C: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 2005;27:1226-1238.
36. Kotsiantis SB: Supervised machine learning: a review of classification techniques. *Informatica* 2007;31:249-268.
37. De Vos WH, Van Neste L, Dieriks B, Joss GH, Van Oostveldt P: High content image cytometry in the context of subnuclear organization. *Cytometry Part A* 2010;77A:64-75.
38. Burnham KP, Anderson DR: Multimodel inference: understanding AIC and BIC in model selection. *Sociol Method Res* 2004;33:261-304.
39. Chen X, Velliste M, Weinstein S, Jarvik JW, Murphy RF: Location proteomics: building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins. *Proc SPIE* 2003;4962:298-306.
40. Perlman ZE, Mitchison TJ, Mayer TU: High-content screening and profiling of drug activity in an automated centrosome- duplication assay. *ChemBioChem* 2004;2004:1-8.
41. Zhu X, Goldberg AB: *Introduction to Semi-Supervised Learning*. San Rafael, CA: Morgan & Claypool, 2009.
42. Lin Y-S, Huang Y-H, Lin C-C, Hsu C-N: Feature space transformation for semi-supervised learning for protein subcellular localization in fluorescence microscopy images. *IEEE Intl Symp Biomed Imaging* 2009:414-417.
43. Eggert US, Mitchison TJ: Small molecule screening by imaging. *Curr Opin Chem Biol* 2006;10:232-237.
44. Tanaka M, Bateman R, Rauh D, Vaisberg E, Ramachandani S, Zhang C, et al: An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol* 2005;3:e128.
45. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ: Multidimensional drug profiling by automated microscopy. *Science* 2004;306:1194-1198.
46. Birmingham A, Selfors LM, Forster T, Wrobel D, Kennedy CJ, Shanks E, et al: Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* 2009;6:569-575.
47. Goshima G, Wollman R, Goodwin SS, Zhang N, Scholey JM, Vale RD, et al: Genes required for mitotic spindle assembly in Drosophila S2 cells. *Science* 2007;316:417-421.
48. Harder N, Mora-Bermudez F, Godinez WJ, Ellenberg J, Eils R, Rohr K: Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. *Med Image Comput Comput Assist Interv* 2006;9:840-848.
49. Bakal C, Aach J, Church G, Perrimon N: Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* 2007;316:1753-1756.
50. Matula P, Kumar A, Worz I, Erfle H, Bartenschlager R, Eils R, et al: Single-cell-based image analysis of high-throughput cell array screens for quantification of viral infection. *Cytometry Part A* 2009;75A:309-318.

51. Wang J, Zhou X, Li F, Bradley PL, Chang S-F, Perrimon N, et al: An image score inference system for RNAi genome-wide screening based on fuzzy mixture regression modeling. *J Biomed Inform* 2009;42:32-40.
52. Huang K, Lin J, Gajnak JA, Murphy RF: Image content-based retrieval and automated interpretation of fluorescence microscope images via the protein subcellular location image database. *IEEE Intl Symp Biomed Imaging* 2002:867-870.
53. Swedlow JR, Goldberg I, Brauner E, Sorger PK: Informatics and quantitative analysis in biological imaging. *Science* 2003;300:100-102.
54. Martone ME, Tran J, Wong WW, Sargis J, Fong L, Larson S, et al: The cell centered database project: an update on building community resources for managing and sharing 3D imaging data. *J Struct Biol* 2008;161: 220-231.
55. Berglund L, Bjorling E, Oksvold P, Fagerberg L, Asplund A, Szigyarto CA, et al: A genecentric Human Protein Atlas for expression profiles based on antibodies. *Mol Cell Proteomics* 2008;7:2019-2027.
56. Murphy RF: Location proteomics: a systems approach to subcellular location. *Biochem Soc Trans* 2005;33:535-538.

Address correspondence to:
*Robert F. Murphy*
*Lane Center for Computational Biology*
*Carnegie Mellon University*
*5000 Forbes Avenue, Pittsburgh, PA 15213*

*E-mail:* murphy@cmu.edu