doi:10.1017/S1431927616003548
Microsc. Microanal. 22 (Suppl 3), 2016

High-Throughput, Automated Image Processing for Large-Scale Fluorescence Microscopy Experiments

Allen Goodman[1] and Anne E. Carpenter[1]

[1.] Imaging Platform, Broad Institute of Harvard and MIT, Cambridge USA

Biological images are a very rich source of information. Biologists have long relied on their eyes to inform them about the state of cells, tissues, and organisms. Visually examining samples can reveal signaling pathways, biological mechanisms of action, disease states, and more, if the cells are stained appropriately. Especially with fluorescent proteins and exquisitely sensitive antibodies, researchers can gain a tremendous amount of knowledge about cells and their status simply by looking at them.

With the help of automated image analysis algorithms, the information sought by a biologist can be quantified. Automated image analysis reduces subjectivity and increases throughput but can also detect changes too subtle, or too tedious, for the human visual system to assess. It can also be used to capture hundreds of morphological aspects of a sample simultaneously, allowing an unbiased assessment of changes in, for example, staining localization and patterning, protein levels, and cell morphology.

Our laboratory is particularly interested in using computers to turn images into numerical measurements because we work with large-scale biological experiments involving thousands to millions of images (with millions to billions of cells), far too many to examine visually. Working with collaborators with expertise in a certain biological or disease area, we help design an assay system and method of quantification suited to the question at hand. Algorithms are made user-friendly by incorporation into CellProfiler [1], open-source software for measuring properties of objects in images in high throughput.

A typical image analysis workflow involves building a pipeline out of individual modules that each perform a discrete function. For example, in the typical usage of CellProfiler, images are pre-processed (to correct illumination anomalies, for example), then segmentation algorithms identify individual cells. Lastly, features are extracted from every cell in every image, including categories such as counts, size, shape, intensity, texture, correlation, and neighborhood relationships.

The segmentation (identification) step is typically the most crucial in any image-processing pipeline, particularly when cells clump or overlap. In our experience, the vast majority of biological assay systems (most notably, cultured cells and yeast colonies), can be handled well by standard automatic thresholding variants and properly configured model-based segmentation algorithms. Once a pipeline is adjusted on some representative test images, the pipeline can be run to automatically identify and measure cells in thousands to millions of images using a computing cluster.

In cases where model-based segmentation algorithms do not achieve sufficient accuracy, pixel-based machine learning algorithms can be very powerful in identifying regions of interest in an automated manner. In brief, the researcher uses software such as the open-source ilastik [2] to mark some regions of representative images as belonging to different classes of interest (e.g., nucleus, cytoplasm, background). The underlying machine-learning algorithm assesses the neighborhood around each marked pixel and learns properties that distinguish the various classes. This learning, summarized in a machine-learning classifier, can then be applied to a large set of images *en masse*.

After individual biological objects have been identified, particular features of interest can be measured directly. Alternately, object-based supervised machine-learning algorithms can use a large number of measured morphological features in an unbiased way to identify a complex phenotype of interest. The software tools for this can be quite user-friendly; for example, our own CellProfiler Analyst [3] allows researchers to drag and drop individual cells having phenotype(s) of interest into bins, in order to train the computer to recognize those phenotypes and score them in a large image set.

These approaches have been used for a huge variety of biological goals. Often, each member of a small-molecule library is tested to identify potential therapeutics. In our own collaborations, measuring GFP-labeled *Mycobacterium tuberculosis* (Mtb) led to host-targeted compounds that inhibit pathogen growth in the lungs of infected mice [4]. Using machine learning to distinguish cobblestoned cells in an *in vitro* co-culture system identified a compound that extended the survival of mice given leukemic bone marrow cells [5]. Precisely measuring the DNA content of individual nuclei in images led to a clinical trial for acute megakaryoblastic leukemia, AMKL [6]. Genetic perturbations are also often screened. RNA interference screens have used time-lapse microscopy to identify novel regulators of lysosome motility [9], primary neurons to identify novel regulators of synaptogenesis [10], and irradiation to find genes involved in the regulation of DNA damage [7]. New potential targets for glioblastoma were found using machine learning to classify cells as having an adherent versus neurosphere phenotype [8].

Rich information is present in images beyond that which the biologist already seeks to measure. Computational approaches can extract this latent information to identify unusual phenotypes, or similarities and differences among samples, in an unbiased way. Often, the morphological profile resulting from a particular genetic or chemical perturbation is remarkably specific. This strategy, termed morphological profiling, seeks to detect patterns in measured morphological phenotypes. This can be useful, for example, to detect whether cells derived from diseased patients show any phenotypic differences versus controls, as a diagnostic and as a potential assay system to identify therapeutics, even personalized therapeutics. Morphological profiling can also be used to identify drug targets and mechanisms of action, determine the functional impact of disease-related alleles, create performance-diverse chemical libraries, and categorize mechanisms of drug toxicity [11].

References:

[1] AE Carpenter *et al*, Genome Biology **7** (2006) R100. (http://www.cellprofiler.org)
[2] C Sommer *et al*, Biomedical Imaging: Nano-Macro, IEEE Internat. Symposium (2011) p. 230–233.
[3] TR Jones *et al*, Proc Natl Acad Sci **106** (2009) p. 1826–1831.
[4] SA Stanley *et al*, PLoS Pathogens **10** (2014) e1003946.
[5] KA Hartwell *et al*, Nat Chem Biology **9** (2013) p. 840–848.
[6] Q Wen *et al*, Cell. **150** (2012) p. 575–589.
[7] SR Floyd *et al*, Nature **498** (2013) p. 246–250.
[8] Y Chudnovsky *et al*, Cell Reports **6** (2014) p. 313–324.
[9] AL Jolly *et al*, Cell Reports **14** (2016) p. 611–620.
[10] TJF Nieland *et al*, PLoS One. **9** (2014) e91744.
[11] Funding to support this work is, in part, from the National Institutes of Health (R01 GM089652, to AEC). The authors thank current and former members of the Imaging Platform as well as our many collaborators who have provided interesting and important driving biological questions.