EXTRACTING BIOMEDICALLY IMPORTANT INFORMATION FROM LARGE, AUTOMATED IMAGING EXPERIMENTS

Anne E. Carpenter

Imaging Platform, Broad Institute of Harvard and MIT

ABSTRACT

Major challenges remain in the extraction of rich information from high-throughput microscopy experiments. In this paper, I describe some of these challenges, particularly those that are the subject of ongoing research in my laboratory. The challenges include segmenting neurons, co-cultures of different cell types, and whole organisms; segmenting and tracking cells in time-lapse images; quantifying complex phenotypic changes; and discovering biologically relevant subpopulations of cells.*

Index Terms— high-throughput, screening, fluorescence microscopy, co-cultures, *C. elegans*

1. INTRODUCTION

Due to advancements in robotic systems, biologists in pharmaceutical companies and academic screening centers are now able to efficiently create hundreds of thousands of biological samples in parallel. Each sample tests the effects of a particular gene or potential drug on a disease-relevant biological system, such as cells or small organisms. The goal is to "screen" the samples to identify those with desired effects. When each sample is imaged by microscopy, extracting the relevant, quantitative information from each image in an automated fashion becomes the main challenge.

Image processing algorithms and machine learning tools have been successfully employed to score increasingly complex phenotypes over the past decade. Here, some major challenges in this field are reviewed as part of an ISBI special session drawing attention to this growing area in biomedical imaging. I highlight ongoing research areas of my group, the Imaging Platform of the Broad Institute of Harvard and MIT, where we focus on quantifying and mining the rich information present in high-throughput images (100,000–1,000,000 images per experiment) probing a variety of biological processes and diseases of interest.

* This is an invited paper in the special session on "Current challenges in image analysis for high-throughput microscopy."

2. NEURONAL CELL TYPES

Scoring samples for a specific phenotypic change has become fairly routine in most mammalian cell types as well as similar-looking non-mammalian cells. This is true even for complex phenotypes, where machine learning has become indispensable [1-3]. One major exception is neuronal cell types (Figure 1A). Because the thin neurites that protrude from the cell bodies are often very weakly stained, automated algorithms often fail to accurately trace each neurite unless sample preparation and imaging conditions are optimal. The state of the art is often to fall back on interactive guidance from the user, but this is infeasible for high-throughput experiments. Foreground–background segmentation alone can be challenging, and tracing individual neurites that cross or are entangled is even more difficult. Furthermore, important information about neuron connectivity can only be gained by three-dimensional imaging, making experiments involving neurons a computational challenge in many respects. Neuronal cell types are one of the "final frontiers" of two-dimensional mammalian cell image analysis [4].

3. MIXTURES OF CELL TYPES

High-throughput experiments are inherently artificial in that they usually involve cells grown out of their native environment, typically in plastic multi-well plates. Biologists studying many different biological processes and diseases are increasingly making the extra effort to preserve natural cell–cell interactions by growing mixtures of physiologically relevant cell types together. This type of co-culturing is also required for proliferation of some cell types.

Developing computational approaches to analyze the complex images resulting from mixtures of two visually distinctive cell types is challenging but worthwhile. For example, human hepatocytes proliferate and maintain their native liver-specific functions much better when grown in the presence of fibroblasts. If the fibroblasts are derived from mouse cells, the nuclei of the two cell types are distinctive enough to be distinguished by supervised machine learning. This enables large-scale experiments to identify chemicals that promote liver regeneration. The system should also be useful to assess a potential drug's human liver toxicity in order to prevent clinical trial failures. It is

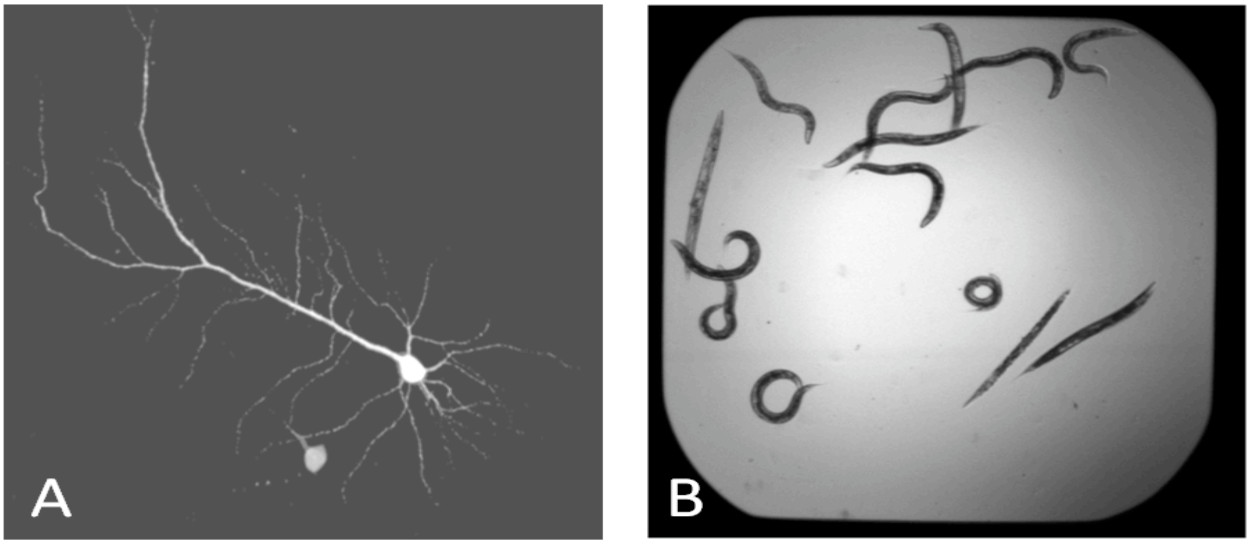

Figure 1. (A) Projection of a 3D image of a fluorescently stained neuron (kindly provided by collaborator Mehmet Fatih Yanik). Thin structures and low signal to noise make image segmentation and neuron tracing difficult, particularly when cells are more crowded than in this simplified example. (B) *C. elegans* worms cultured in 384-well plates imaged by bright field microscopy (kindly provided by collaborator Frederick M. Ausubel). Touching and overlapping worms complicate feature extraction from individual animals.

also increasingly common to co-culture adult hematopoietic or leukemic stem cells with stroma. Such experiments can identify chemicals that promote the proliferation of hematopoietic cells or transition leukemic cells to a harmless differentiated state.

4. WHOLE ORGANISMS

Some model organisms are small enough to fit in multi-well plates, yet complex enough to model many aspects of human physiological function. These organisms enable the study of biological processes and diseases that involve organ systems and multicellular interactions. Major challenges in high-throughput whole-organism image analysis include distinguishing between organisms and artifacts such as debris, distinguishing between organisms that clump or cross over each other in the population in each sample, and quantifying complex phenotypes in each individual in a population of organisms.

We have begun to address these challenges for *Caenorhabditis elegans* worms imaged in high throughput [5-7]. Here, separating touching and overlapping worms, as seen in Figure 1B, is one of the major challenges before measurements can be made on individual animals. To identify much-needed novel classes of antibiotics and anti-infectives, we have developed algorithms that can identify individual worms and detect whether a potential drug has cured the worms from pathogen infection. We are also testing these algorithms to quantify fat storage in individual worms, allowing regulators of metabolism to be studied in the context of a whole, living organism.

5. TIME-LAPSE IMAGES

Many biological questions can only be answered by collecting time-lapse movies [8, 9]. In such images, one challenge is accurately identifying cells and their compartments given the low signal-to-noise that is common in time-lapse images where cells must be exposed to as little light as possible in order to prevent phototoxicity and photobleaching. It is also challenging to track cells accurately from one frame of the movie to the next because the frame rate is usually minimized to reduce the cells' exposure to light. Many methods in the extensive tracking literature are unsuited for high-throughput images, and often require manual intervention.

We have been working to accurately extract information from time-lapse movies to identify, for example, novel cell cycle landmarks and motor protein regulators. We are also integrating time-lapse data with flow cytometry data to quantify unusual cell cycle outcomes. Ongoing experiments that track the motion and behavior of swimming zebrafish over time combine the challenges of whole-organism image analysis and time-lapse imaging.

6. DISCOVERING MORPHOLOGICAL SIGNATURES

High-throughput imaging studies have been mostly limited to identifying samples that exhibit one or more specific, known phenotypic changes. However, a few labs have begun to explore the prospect of profiling samples using a more comprehensive set of automatically-discovered phenotypes, thereby revealing similarities and differences between samples that may be unexpected or even undetectable to the human eye [2, 10-12]. This is beginning to bring microscopy experiments into the realm of systems biology, such that image-based phenotypes may soon be routinely interrogated side by side with gene-expression and proteomic data.

High-throughput imaging experiments generate extremely large, high-dimensional data sets with quantifiable phenotypic information for every individual cell. Shifts in the distribution of phenotypes within a heterogeneous population are often difficult to detect by eye, but may be highly relevant from a biological perspective. We are using this rich, latent information to identify patterns in chemical or genetic perturbations in order to distinguish genes and chemicals with related cellular effects and to discover chemical targets and side effects.

7. BRIDGING THE GAP BETWEEN BIOLOGISTS AND ADVANCED COMPUTATIONAL TOOLS

Solving a particular image-processing challenge is only the first step towards extracting useful information from biological images. A published description of an algorithm is very different from a working implementation. Even if an algorithm was successfully applied to a real problem and the implementation made available, it may not be useful for others because making the implementation work on a different experiment and integrating it with other necessary processing steps requires much time and image-processing/software-development expertise as well as intimate knowledge of the new experiment. Ideally, algorithms are packaged in a user-friendly form with sufficient documentation so that the biologist, who is most familiar with the scientific questions at hand, can configure the algorithm and integrate it with other analysis steps.

Algorithms developed in our group are made available to the scientific community as modules for CellProfiler (http://www.cellprofiler.org), our open-source software package for quantifying a variety of phenotypes in biological images [13, 14]. Since we first released it in 2005, it has become widely used, with more than 10,000 downloads and over 250 citations so far. The software evolves within an active research environment involving dozens of diverse image-based assays, resulting in rich functionality as we continue to improve its capabilities, interface, and support.

Recognizing the real challenge in exploring high-dimensional data from hundreds of thousands of images, we are also developing tools and workflows for data analysis, exploration, and quality control. These are released under the name CellProfiler Analyst.

Several open-source software projects are beginning to be interfaced in ways that allow researchers to conveniently use the most suitable tool for each step in an experiment. CellProfiler has recently been interfaced with the pixel-based classification tool Ilastik (http://www.ilastik.org/) and the popular image-analysis tool ImageJ (aka. NIH Image, http://rsbweb.nih.gov/ij/). CellProfiler uses the NumPy and SciPy scientific-computing libraries [15]. For reading and writing various image file types, it uses the BioFormats library (http://www.loci.wisc.edu/software/bio-formats), developed as part of the Open Microscopy Environment project (http://www.openmicroscopy.org).

Finally, we are building up a collection of freely downloadable microscopy image sets through the Broad Bioimage Benchmark Collection (http://www.broadinstitute.org/bbbc/). In addition to the images themselves, each set includes a description of the biological application and some type of "ground truth" (expected results). Researchers are encouraged to use these image sets as reference points when developing, testing, and publishing new image analysis algorithms for the life sciences. We hope that the BBBC will lead to a better understanding of which methods are best for various biological image analysis applications.

8. CONCLUSION

Microscopy is one of the most powerful and informative ways to analyze experiments designed to uncover phenotypic variations in response to chemical, genetic, and other perturbations, but extracting biomedically important information from large, automated image experiments is challenging. High-throughput data requires segmentation algorithms that are robust to experimental variations, and complex as well as subtle phenotypes require advanced feature extraction and data mining. The challenges briefly summarized here are relevant for a wide range of biomedical research areas. With this review of our current challenges we hope to encourage not only additions to the wealth of powerful image analysis algorithms already available, but more specifically encourage development and dissemination of algorithms that solve biomedically relevant problems.

9. ACKNOWLEDGEMENTS

The author acknowledges the entire Imaging Platform of the Broad Institute of Harvard and MIT for their contributions to the hundreds of projects represented here. Equally important, the Platform acknowledges the scientific contributions and images provided by our collaborators. Presentation of this research is supported by the NIH (R01 GM089652-01).

10. REFERENCES

[1] T. R. Jones, A. E. Carpenter, M. R. Lamprecht, J. Moffat, S. J. Silver, J. K. Grenier, A. B. Castoreno, U. S. Eggert, D. E. Root, P. Golland, and D. M. Sabatini, "Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning," *Proc Natl Acad Sci U S A,* vol. 106, pp. 1826-1831, Feb 2 2009.

[2] E. Glory and R. F. Murphy, "Automated subcellular location determination and high-

throughput microscopy," *Developmental Cell,* vol. 12, pp. 7-16, Jan 2007.
[3] B. Misselwitz, G. Strittmatter, B. Periaswamy, M. Schlumberger, S. Rout, P. Horvath, K. Kozak, and W.-D. Hardt, "Enhanced CellClassifier: a multi-class classification tool for microscopy images," *BMC Bioinformatics,* vol. 11, p. 30, 2010.
[4] E. Meijering, "Neuron tracing in perspective," *Cytometry A,* vol. 77, pp. 693-704, 2010 Jul 2010.
[5] T. Riklin Raviv, V. Ljosa, A.L. Conery, F.M. Ausubel, A.E. Carpenter, P. Golland and C. Wählby, "Morphology-Guided Graph Search for Untangling Objects: C. elegans Analysis" in *Proceedings of MICCAI 2010, LNCS* 6363:634-641, Springer 2010.
[6] C. Wählby, T. Riklin-Raviv, V. Ljosa, A. L. Conery, P. Golland, F. M. Ausubel, and A. E. Carpenter, "Resolving clustered worms via probabilistic shape models," in *IEEE International Symposium on Biomedical Imaging (ISBI): From Nano to Macro*, Rotterdam, The Netherlands, 2010.
[7] T. I. Moy, A. L. Conery, J. Larkins-Ford, G. Wu, R. Mazitschek, G. Casadei, K. Lewis, A. E. Carpenter, and F. M. Ausubel, "High-throughput screen for novel antimicrobials using a whole animal infection model," *ACS Chem Biol,* vol. 4, pp. 527-33, Jul 17 2009.
[8] E. Meijering, I. Smal, O. Dzyubachyk, and J. C. Olivo-Marin, "Time-Lapse Imaging," in *Microscope Image Processing,*, F. A. M. Q. Wu, K. R. Castleman, Ed. Burlington, MA: Elsevier Academic Press, 2008, pp. 401-440.
[9] J. F. Dorn, G. Danuser, and G. Yang, "Computational processing and analysis of dynamic fluorescence image data," *Methods Cell Biol,* 2008 2008.
[10] Z. E. Perlman, M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler, "Multidimensional drug profiling by automated microscopy," *Science,* vol. 306, pp. 1194-8, Nov 12 2004.
[11] L. H. Loo, L. F. Wu, and S. J. Altschuler, "Image-based multivariate profiling of drug responses from single cells," *Nat Methods,* vol. 4, pp. 445-53, Apr 1 2007.
[12] C. L. Adams, V. Kutsyy, D. A. Coleman, G. Cong, A. M. Crompton, K. A. Elias, D. R. Oestreicher, J. K. Trautman, and E. Vaisberg, "Compound classification using image-based cellular phenotypes," *Methods Enzymol,* vol. 414, pp. 440-68, 2006.
[13] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, "CellProfiler: image analysis software for identifying and quantifying cell phenotypes," *Genome Biol,* vol. 7, p. R100, 2006.
[14] M. R. Lamprecht, D. M. Sabatini, and A. E. Carpenter, "CellProfiler: free, versatile software for automated biological image analysis," *Biotechniques,* vol. 42, pp. 71-5, Jan 2007.
[15] T. E. Oliphant, "Python for Scientific Computing," *Comput. Sci. Eng.,* vol. 9, pp. 10–20, 2007.