Image-based profiling: a powerful and challenging new data type

Gregory P. Way
Center for Health Artificial Intelligence,
University of Colorado Anschutz Medical School,
1890 N Revere Ct. Aurora CO 80045, USA
Email: gregory.way@cuanschutz.edu

Hannah Spitzer
Institute of Computational Biology,
Helmholtz Center,
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
Email: hannah.spitzer@helmholtz-muenchen.de

Philip Burnham
Kanvas Biosciences,
1 Deer Park Drive, South Brunswick, NJ 08852, USA
Email: phil@kanvasbio.com

Arjun Raj
Department of Bioengineering,
University of Pennsylvania,
220 S 33rd St, Philadelphia, PA 19104, USA
Email: arjunraj@seas.upenn.edu

Fabian Theis
Institute of Computational Biology,
Helmholtz Center,
Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
Email: fabian.theis@helmholtz-muenchen.de

Shantanu Singh and Anne E. Carpenter
Imaging Platform,
Broad Institute of Harvard and MIT,
415 Main Street, Cambridge, MA 02142 USA
Email: shsingh@broadinstitute.org, anne@broadinstitute.org

Software has provided cell biologists the power to quantify specific cellular features in cell images at scale. Before long, these biologists also recognized the potential to extract much more biological information from the same images. From here, the field of image-based profiling, the process of extracting unbiased representations that capture morphological cell state, was born. We are still in the early days of image-based profiling, and it is clear that the many opportunities to interrogate

biological systems come with significant challenges. These challenges include building expressive and biologically-relevant representations, adjusting for technical noise, writing generalizable software infrastructure, continuing to foster a culture of open science, and promoting FAIR (findable, accessible, interoperable, and reusable) data. We present a workshop at the Pacific Symposium on Biocomputing 2022 to introduce the field of image-based profiling to the broader computational biology community. In the following document, we introduce image-based profiling, discuss current state-of-the-art methods and limitations, and provide rationale for why now is the perfect time for the field to expand. We also introduce our invited speakers and agenda, which together provide an introduction to the field complemented by in-depth application areas in industry and academia. We also include five lightning talks to complement the invited speakers on various methodological and discovery advances.

Keywords: Computational Biology; Morphology; Systems Biology; Cell State; Cell Structure; Functional Genomics; Data Integration; Drug Discovery; Single-cell; Perturbation Biology.

1. Introduction

Microscopy images contain a wealth of phenotypic information about the sample and treatment. Typically, biologists extract only a small amount of information from the images (e.g. a single viability readout, or the intensity of a biomarker). In an image-based profiling experiment, however, a researcher measures an unbiased and high dimensional readout representing an overall morphological state of the imaged organism, tissue, cell, or organelle.[1-2] In consequence, image-based profiling extends the processing of microscopy data readouts into a data-intensive endeavor requiring method development, improved analytical approaches, and flexibility to maximize value from the extreme diversity of microscopy experimental designs.[3-4]

Scientists have been developing image-based profiling strategies for more than a decade.[5] Only recently, however, are various academic and industry labs intensely scaling up their own analysis platforms to use image-based profiling for various pursuits; including drug discovery, functional genomics, and perturbation toxicity characterization. Fueled by a dramatic increase in accessible compute power, widespread publicly available benchmark datasets, expressive algorithms, and the success of several industry-driven platforms, image-based profiling has enjoyed renewed excitement.[6-9] A recent comprehensive review discusses the power of image-based profiling and the pressing need for advancing computational method development.[10]

2. The time is now

In addition to methodological advances, image-based profiling would also benefit from a cultural rebirth. The image-based profiling discipline emerged from microscopy labs, and has benefitted relatively little from in-depth interdisciplinary influence with the broader computational biology community. Certain current challenges and open questions, such as genomic data-type integration, single cell normalization, and batch effect correction, would benefit from a cross-pollination of ideas. Additionally, existing systems biology labs who focus exclusively on genomic data would benefit from incorporating morphology as a new, complementary dimension closely tied to biological function.

In 2016, the CytoData Society formed to grow and foster the image-based profiling community. In addition to providing community resources, the society organizes a yearly conference to discuss recent research and emerging software tools. In these conferences, researchers from around the world gather to discuss common challenges and recent discoveries. However, the society has been mostly inwardly focused; on developing a strong community around a new data type capable of probing biology in new ways. Our workshop at the Pacific Symposium on Biocomputing will provide an outward focus, highlighting new computational approaches, methods, applications, and pressing biological questions in our field, while broadening the conversation to include a community of scientists from diverse disciplines.

3. Workshop agenda

The three-hour workshop will provide an introduction to the image-based profiling field and discuss important applications in industry and academia.

3.1. *Workshop invited speakers*

We have invited six speakers, who will each discuss either the historical context of our field or their own research applications. The first three speakers will introduce image-based profiling and discuss its challenges. The next three speakers will discuss image-based profiling applications.

3.1.1. *Niranj Chandrasekaran, Ph.D.*

Dr. Chandrasekaran is our first speaker, and he will introduce the field of image-based profiling. He is currently a postdoctoral scientist in the Carpenter-Singh laboratory at The Broad Institute of MIT and Harvard, and has been working in the image-based profiling field for two years. His background is in biophysics and genomics, and he has provided significant leadership for the Joint Undertaking in Morphological Profiling (JUMP) Cell Painting Consortium, bringing together a dozen pharmaceutical and non-profit entities to create a large, publicly available Cell Painting dataset of chemical and genetic perturbations. He is one of the first software architects of the image-based profiling pipeline, which include the necessary steps to process and normalize readouts derived from cell images.

3.1.2. *Juan C. Caicedo, Ph.D.*

Dr. Caicedo is our second speaker, and he will discuss emerging strategies for how to extract morphology representations from cell images. As a Schmidt Fellow at The Broad Institute of MIT and Harvard, his laboratory works on computer vision applications using images from a variety of sources. His research is primarily focused on using machine learning to model and extract information from biological images. He is developing representation learning algorithms for extracting image-based, single-cell embeddings to characterize cellular phenotypes at large scale, to support applications in high-throughput screening, drug discovery and functional genomics projects.

3.1.3. *Paula A. Marin Zapata, Ph.D.*

Dr. Marin Zapata is our third speaker of the workshop. She will discuss the nitty-gritty technical challenges for working with image-based data. She will discuss these challenges in the context of a particular large-scale project, the Joint Undertaking for Morphological Profiling (JUMP) Cell Painting Consortium, which is a collaborative effort involving over a dozen industry and academic partners. Dr. Marin Zapata is a research scientist, currently working at the Machine Learning Research group of Bayer Pharmaceuticals in Berlin, with a background in imaging data, deep learning, and cell biology.

3.1.4. *Imran Haque, Ph.D.*

After a short break, we will introduce Dr. Haque as our fourth speaker to begin the application-oriented portion of our workshop. He is currently the Vice President of Data Science at Recursion Pharmaceuticals, and he will discuss focused application areas in various stages of development at the company. Recently undergoing its initial public offering and announcing a large expansion, Recursion has continued to push boundaries in the field of image-based profiling and has four clinical trial candidates in the pipeline so far. Dr. Haque has a background in computer science and electrical engineering, and ten years of industry leadership experience.

3.1.5. *Kyogo Kawaguchi, Ph.D.*

Dr. Kawaguchi is the fifth speaker of our workshop, and will discuss his application of predicting cell fate decisions in tissues using graph neural networks and image-based profiling. Dr. Kawaguchi is currently a team leader in the RIKEN Cluster for Pioneering Research interested in collective cell dynamics and phase transitions. Dr. Kawaguchi has a background in non-equilibrium statistical mechanics and biophysics.

3.1.6. *Susanne Rafelski, Ph.D.*

Dr. Rafelski is the sixth and final invited speaker of our workshop. She will discuss ongoing image-based profiling applications at the Allen Institute for Cell Science (AICS) aimed at understanding the interplay between cell organization, dynamics, behavior, and fate. AICS is generating publicly available stem cell lines and producing open access tools and resources for computational image analysis. Dr. Rafelski is currently the Deputy Director overseeing the Scientific Programs at AICS and has a background in quantitative cell biology, live-cell imaging and analysis, molecular genetics, and computational biology.

3.2. *Lightning talks from emerging methods and applications*

We will also select up to five in-person lightning talks from submitted abstracts to provide a short, six-minute description of their methodological or research findings and implications. We will select these speakers to complement one another to balance the information presented in the full workshop.

4. Conclusion

Images are complex and often difficult to handle, but they provide an enormous amount of untapped information. To begin extracting this value, we must develop the field of image-based profiling. One major hurdle to expansion is a general lack of awareness by the peripheral computational biology community. Our goal for the workshop is to increase awareness of the power, potential, and challenges of image-based profiling data. We aim to excite a wide computational biology audience and to increase participation, recruitment, and impact.

References

1. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
2. Scheeder, C., Heigwer, F. & Boutros, M. Machine learning and image-based profiling in drug discovery. *Curr Opin Syst Biol* **10**, 43–52 (2018).
3. Thorn, K. A quick guide to light microscopy in cell biology. *Mol. Biol. Cell* **27**, 219–222 (2016).
4. Lee, J.-Y. & Kitaoka, M. A beginner's guide to rigor and reproducibility in fluorescence imaging experiments. *Mol. Biol. Cell* **29**, 1519–1525 (2018).
5. Perlman, Z. E. et al. Multidimensional drug profiling by automated microscopy. Science 306, 1194–1198 (2004).
6. Weigle, S., Martin, E., Voegtle, A., Wahl, B. & Schuler, M. Primary cell-based phenotypic assays to pharmacologically and genetically study fibrotic diseases in vitro. *J Biol Methods* **6**, e115 (2019).
7. Simm, J. *et al.* Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chem Biol* **25**, 611–618.e3 (2018).
8. Johnson, K. *et al.* A stem cell-based approach to cartilage repair. *Science* **336**, 717–721 (2012).
9. Gibson, C. C. *et al.* Strategy for identifying repurposed drugs for the treatment of cerebral cavernous malformation. *Circulation* **131**, 289–299 (2015).
10. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat. Rev. Drug Discov. 20, 145–159 (2021).