Why bioimage informatics matters

Gene Myers

Driven by the importance of spatial and physical factors in cellular processes and the size and complexity of modern image data, computational analysis of biological imagery has become a vital emerging subdiscipline of bioinformatics and computer vision.

By the end of 2001, the human genome and those of the major model organisms—*Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*—had been sequenced[1–5]. With the genetic code in hand, the imperative question became 'what is encoded therein, and how does it orchestrate life?' There were many obvious avenues to pursue: comparative sequencing, expression analyses of various sorts, chromatin immunoprecipiation analyses, determination of epidemiological state and so on. In overall effect, one could argue that what having these sequences enabled was not 'systems biology' but rather 'systems genetics', which refers to the ability to systematically modify, measure or instrument every putative genomic element of interest.

Two early examples of systems genetic explorations were the study of O'Shea and Weissman[6] for which they created a protein-GFP fusion of every protein in *Escherichia coli* to map the localization of every protein, and the study of Hyman and Escheverri[7] who systematically knocked down every gene in *C. elegans* to observe which had a critical effect on mitosis. In both studies, the 'readout' was images obtained via light microscopy—effectively direct *in vivo* observations of the mesoscale of the cell. Even though the researchers captured the images digitally, they only examined them by eye. In large part, this was because the software needed to more precisely and quantitatively analyze the images was nowhere to be found.

Ten years later, molecular biologists have come to clearly appreciate that the cell is a complex set of nanoscale machines that self-aggregate and dissolve according to phase transitions, that the cell is spatially organized and has many membrane-enclosed subcompartments and that cells adhere to each other and generate forces, which in turn can create patterns of expression[8]. None of this is captured in the interaction networks that seem to be the primary output of genomics, transcriptomics and proteomics, and indeed these networks, although necessary, are insufficient to explain many phenomena satisfactorily, if at all. For example, the establishment of cell polarity cannot be explained just in terms of interactions between proteins. A satisfying intellectual explanation of the process requires understanding the biophysics and spatial organization of the actomyosin cortex[9].

Because of this increasing recognition of the importance of spatial and physical factors in the control and function of cellular processes, the number of systems genetics explorations and efforts involving direct *in vivo* observations of function is increasing and driving not only the nascent field of bioimage informatics (the analysis of images and stacks of cells and cellular collections obtained by light or electron microscopy) but also continued advances in transgenics (our ability to instrument the cell) and exciting developments in the throughput and resolution of microscopes (our ability to observe the instrumented cell).

Bioimage informatics is a specialization of computer vision and image analysis. Most of the larger community is typically focused on trying to interpret high-contrast digital captures of complex natural scenes and people. This is in contrast to the difficulties of our specialty, which has to deal with low signal-to-noise ratios and limited resolution, offset to some extent by the fact that the scene is generally very simple and there is often much prior information about the object(s) under observation.

In broad terms, researchers in this specialty can be categorized—based on the size of the objects being imaged—into those who work on cell images and try to quantify and model mesoscale phenomena in cells, to those who observe collections of cells both *in vivo* and *in situ* to understand development trajectories and to build digital anatomical atlases, to those who observe and analyze the behavior of entire organisms.

The field is still in its early days, and there is no such thing as a typical bioimage informatician: they are either computer vision experts looking for new problems, classic sequence-based bioinformaticians looking for the new thing or physicists and molecular biologists whose experiments require them to bite the informatics bullet. There are as yet no established large-scale forums for the work, both in terms of meetings and journals. Young people in the field present challenges to established academic evaluation committees. From my perspective, it is very reminiscent of the state of bioinformatics in the early 1980s: the exciting, somewhat chaotic free-for-all that is potentially the birth of something new.

Also reminiscent of sequence-based informatics, is the fact that image-based sys-

Gene Myers is at the Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany.
e-mail: myers@mpi-cbg.de

tems genetics involves a shift from one-off experiments to pipelines that require (i) the due diligence of pilot studies before starting in earnest, and (ii) a strong process engineering discipline centered on optimized protocols, absolutely reproducible and constant acquisition conditions and strong quality-control systems. One might ask how many widely used staining protocols that one might consider using in a systems genetics pipeline have actually been optimized by examining the effect of varying concentration, times and temperatures, and be somewhat disappointed in the answer. And like in the early days of 'high-throughput genomics', there have already been a few projects that have been less than successful precisely because of a failure to adhere to the caveats just mentioned. A great systems geneticist has to appreciate and incorporate great engineering into their projects.

The need for the computational analysis of images is also being driven by the rate at which such data are being produced by the newer microscopes and the increasingly ambitious scale of the projects being undertaken. Application of recently described light-sheet microscopes can result in 30 terabytes of data per day[10]. Attempts to image an entire fly brain using electron microscopy are estimated to deliver a 150 terabyte data set (http://www.janelia.org/team-project/fly-em/), and current light-based approaches to mapping the anatomy of a fly's brain involve tens of thousands of three-dimensional stacks each of which is 1–3 gigabytes in size (http://www.janelia.org/team-project/fly-light/).

Described in this issue of *Nature Methods* are the nascent efforts of bioimage informaticians to build systems that facilitate the transformation of images into information[11–13]. These toolkits solve a useful range of use cases and can be used in small- to medium-scale projects. Large-scale projects that require fully automated and highly tuned performance still require dedicated informatics efforts and to some degree always will. Nevertheless, the described tools highlight the need for informatic extraction and make it possible for investigators to begin to familiarize themselves with what can and cannot be done by a computer, and what can and cannot be extracted from an image or stack. All in all, a good start.

In summary, bioimage informatics increasingly matters because of the increasing scale in the production of imagery and because of the increasing number of systems genetics explorations aimed at understanding the crucial physical and spatial nature of proteomic signals and machinery. In effect, the genome sequences, in conjunction with advances in microscopy and computation, are enabling the direct, *in vivo* observation of any genetic entity of interest. Frankly, it does not get any better than that.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. *C. elegans* Sequencing Consortium. *Science* **282**, 2012–2018 (1998).
2. Adams, M.D. *et al. Science* **287**, 2185–2195 (2000).
3. Venter, J.C. *et al. Science* **291**, 1304–1351 (2001).
4. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
5. Mural, R.J. *et al. Science* **296**, 1661–1671 (2002).
6. Ghaemmaghami, S. *et al. Nature* **425**, 737–741 (2003).
7. Sonnichsen, B. *et al. Nature* **434**, 462–469 (2005).
8. Howard, J., Grill, S.W. & Bois, J. *Nat. Rev. Mol. Cell Biol.* **12**, 400–406 (2011).
9. Goehring, N. *et al. Science* **334**, 1137–1141 (2011).
10. Keller, P.J., Schmidt, A.D., Wittbrodt, J. & Stelzer, E.H. *Cold Spring Harb. Protoc.* **10**, 1235–1243 (2011).
11. Schindelin, J. *et al. Nat. Methods* **9**, **XXX–XXX** (2012).
12. de Chaumont, F. *et al. Nat. Methods* **9**, **XXX–XXX** (2012).
13. Kankaanpää, P. *et al. Nat. Methods* **9**, **XXX–XXX** (2012).