A new era in bioimage informatics

Robert F. Murphy1,2

1Lane Center for Computational Biology, and Departments of Biological Sciences, Biomedical Engineering, and Machine Learning, Carnegie Mellon University and 2Faculty of Biology and Freiburg Institute for Advanced Studies, Albert Ludwig University of Freiburg

Bioimage informatics arose from efforts to automate pathology and cytology tasks (Eaves, 1967). With few exceptions, much of the software developed during these early days, whether in academic or commercial institutions, was proprietary. The primary paradigm was production of hand-tuned engineered systems that could reproduce human performance, and visualization was emphasized for interpreting results or providing assistance to clinicians (Bartels and Wind, 1977; Kaman, et al., 1984; van Driel-Kulker and Ploem, 1982). The computational resources available at the time were frequently limiting. Essentially, no successful commercial systems came from these efforts for many years, until the US Food and Drug Administration’s approval of automated Pap smear analysis in the mid 1990s (Patten et al., 1996).

Beginning around this time, a new era began with automation of tasks associated with more basic measurement of cellular and subcellular phenomena, such as detection of drug effects (Giuliano et al., 1997) and recognition of subcellular patterns (Boland et al., 1998). The field gained significant exposure, and a new name, with the founding of two National Science Foundation-sponsored Centers for Bioimage Informatics at the University of California, Santa Barbara and Carnegie Mellon University in 2003. This led to the first Bioimage Informatics conference held in Santa Barbara in 2006. The primary paradigm in this era became supervised and semisupervised machine learning, and automated systems were reported that were able to outperform humans at recognizing cell positions in tissues (Nattkemper et al., 2003) and recognizing subcellular patterns (Murphy et al., 2003). Cutting-edge systems during this period increasingly avoided visualization, hand-tuning or user intervention (with the exception of marking things like cell or nuclear boundaries for training purposes), with a primary goal to produce a simple actionable answer. Examples include identifying which compounds or inhibitory RNAs produce particular effects, and answers typically require further follow-up work, usually not involving imaging, to produce a final biological result. Another shift has been the increased prevalence of open-source software (Eliceiri et al., 2012), although the frequency with which investigator-provided software has been usable by others has varied extensively. Papers have often emphasized the further refinement of methods for benchmark datasets, such as the original 2D HeLa dataset, with the risk that systems become overly tuned to the benchmark.

Bioimage informatics is entering its third era, in which the goal is the fully automated production of models of biological systems. This includes a shift toward unsupervised machine learning, and especially toward structure learning methods. Examples of initial steps include building models of signaling networks from multiprobe images (Welch, et al., 2011), identification of regulatory modules from gene expression images of embryos (Punyani and Xing, 2013), initial attempts at building image-derived generative models of cells (Buck et al., 2012) and identification of proteins involved in cell shape regulation (Sailem et al., 2014). Provision of data and software in reproducible research archives, begun already in the early 2000s, is growing. Other significant trends are the use of image datasets from different sources, and integration with non-imaging data such as from genomics, transcriptomics, proteomics and metabolomics studies.

In this third era, image analysis papers published in Bioinformatics should reflect these trends and follow the journal’s focus on analysis of systems at a molecular level. Incremental refinements to algorithms, methods requiring hand-labeling and -tuning and studies in which the primary outputs are visualizations are all expected to give way to those producing verifiable models that can be combined to produce multiscale, multicomponent representations of the molecular basis and dynamics of cell and tissue organization and behavior.

REFERENCES


