BIOINFORMATICS EDITORIAL Vol. 30 no. 10 2014, page 1353
doi:10.1093/bioinformatics/btu158

# A new era in bioimage informatics

Robert F. Murphy[1,2]

[1]Lane Center for Computational Biology, and Departments of Biological Sciences, Biomedical Engineering, and Machine Learning, Carnegie Mellon University and [2]Faculty 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 Wied, 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 (Puniyani 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

Bartels,P.H. and Wied,G.L. (1977) Computer analysis and biomedical interpretation of microscopic images: Current problems and future directions. *Proc. IEEE*, **65**, 252–261.

Boland,M.V. *et al.* (1998) Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, **33**, 366–375.

Buck,T.E. *et al.* (2012) Toward the virtual cell: automated approaches to building models of subcellular organization "learned" from microscopy images. *Bioessays*, **34**, 791–799.

Eaves,G.N. (1967) Image processing in the biomedical sciences. *Comput. Biomed. Res.*, **1**, 112–123.

Eliceiri,K.W. *et al.* (2012) Biological imaging software tools. *Nat. Methods*, **9**, 697–710.

Giuliano,K. *et al.* (1997) High-Content Screening: A new approach to easing key bottlenecks in the drug discovery process. *J. Biomol. Screen*, **2**, 249–259.

Kaman,E.J. *et al.* (1984) Image processing for mitoses in sections of breast cancer: a feasibility study. *Cytometry*, **5**, 244–249.

Murphy,R.F. *et al.* (2003) Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *J VLSI Signal Process.*, **35**, 311–321.

Nattkemper,T.W. *et al.* (2003) Human vs machine: evaluation of fluorescence micrographs. *Comput. Biol. Med.*, **33**, 31–43.

Patten,S.F. Jr *et al.* (1996) NeoPath, Inc. NeoPath AutoPap 300 Automatic Pap Screener System. *Acta Cytol.*, **40**, 45–52.

Puniyani,K. and Xing,E.P. (2013) GINI: from ISH images to gene interaction networks. *PLoS Comput. Biol.*, **9**, e1003227.

Sailem,H. *et al.* (2014) Cross-talk between Rho and Rac GTPases drives deterministic exploration of cellular shape space and morphological heterogeneity. *Open Biol.*, **4**, 130132.

van Driel-Kulker,A.M. and Ploem,J.S. (1982) The use of LEYTAS in analytical and quantitative cytology. *IEEE Trans. Biomed. Eng.*, **29**, 92–100.

Welch,C.M. *et al.* (2011) Imaging the coordination of multiple signalling activities in living cells. Nature reviews. *Mol. cell biol.*, **12**, 749–756.