COMPUTER VISION FOR SYSTEMS BIOLOGY

Gaudenz Danuser
Laboratory for Computational Cell Biology, The Scripps Research Institute, La Jolla, CA 92037

ABSTRACT

This short paper introduces the scope of our special emphasis session on Computer Vision for Systems Biology. It attempts to define the needs for computer vision based readouts in systems biological research and to shed light on some of the challenges computer vision researchers should tackle for the systems biology community. Finally, it will give a short overview of the invited and contributed papers that will be presented in this session.

1. SYSTEMS BIOLOGY

Systems biology has become a central buzzword in the discussion of science in the 21st century. While debated broadly, there is no consensus what systems biology actually means. I suspect, as many definitions have been formulated as self-proclaimed systems biologists exist. Instead of adding yet another definition, I confine this introduction to an attempt of explaining the origin of the current hype with a brief historic perspective.

With the discovery of DNA as the code of life, a new era in biology began: molecular biology [1, 2]. Over the past 50 years most efforts in biomedical research have been devoted to capitalizing on the new genetic data with the goals to increase our ability to probe and manipulate biological processes with molecular specificity and to exploit these methods to tackle disease. The genomes of several organisms have been solved, most importantly the human genome [3]. Large initiatives are underway to routinely solve the genome of human individuals with the hope that diseases can be understood based on genetic differences between healthy and sick people (e.g. the Broad Institute http://www.broad.mit.edu). Structural genomics represents the next attempt to obtain genome-wide data for an entire organism. Here, the goal is to solve the structure of every protein that is encoded by the genome (http://www.structuralgenomics.org). In parallel, fantastic technology has been developed to mutate and clone genes and to express their products, i.e. the proteins, in large quantities, allowing researchers to directly and specifically interact with the code and the building blocks of life. Very recently, so-called RNA interference (RNAi) has enabled scientists to control protein synthesis in cells [4]. And chemical biologists have produced libraries of small synthetic molecules that inhibit molecular pathways inside living cells.

Common to all these technologies is that they target one molecule class at the time. Research in cell and molecular biology has thus been conducted with the following agenda: i) discovery of a gene involved in a certain biological function; ii) identification and cloning of the gene; iii) mutation and expression of the gene inside the target organism; or externally, e.g. in bacteria, for *in vitro* experiments and/or injection of the associated protein into the target organism; iv) *in vitro* biochemical and biophysical characterization of the protein; v) *in situ* characterization of the role of the gene based on behavioral changes of cells and organisms induced by overexpression, inhibition, or mutation of the gene product. Particularly this last step is very limited in providing reliable information about the function of a specific gene. Genes are embedded in large molecular networks with multiple and possibly context-specific dynamics. Networks overlap, and branches of many networks have redundancies. Thus, the manipulation of one gene, even when the specificity is at the level of a single amino-acid encoded by it, is generally accompanied by broad and completely unknown side effects, precluding the identification of the actual role of a gene in a pathway.

The difficulties in understanding pathways on a molecule-by-molecule basis defined the need for system-wide investigations. Systems biologists seek to understand biological functions with the perspective of how genes and their products interact in large networks [5, 6]. This idea is by no means new, but essentially revives the goals of cellular and organismal physiology that dominated biological research before the advent of molecular biology.

What has changed for the physiologists of today, now called systems biologists, is the experimental toolbox. System-wide analyses of biological processes can build on all the before-mentioned advances made during the era of molecular biology. In addition, new devices have been developed to measure physical and chemical parameters of molecular systems with ever increasing throughput and sensitivity, and numerical techniques have been devised, which allow us to analyze large data sets in the frameworks of computational system models.

However, before the systems biologists can begin to fully exploit these methods, substantial ground work is required to link the vast pools of primary data with high-level models of systems behavior. The bioinformatics community is well underway to tackling this problem for

genome data. Not only do we have complete genomes for several organisms, every researcher world-wide can access this information in standardized databases and use it for modeling. The genome sequences are generally considered of high quality. Robust algorithms are in place to convert the raw information acquired by a sequencer into searchable databases, which are used on a day-to-day basis by thousands of biologists who are unfamiliar with the technicalities of the data processing methods.

As illustrated by Fig. 1, beyond the relatively simple genomic data types the availability and quality of information deteriorates with increasing complexity.

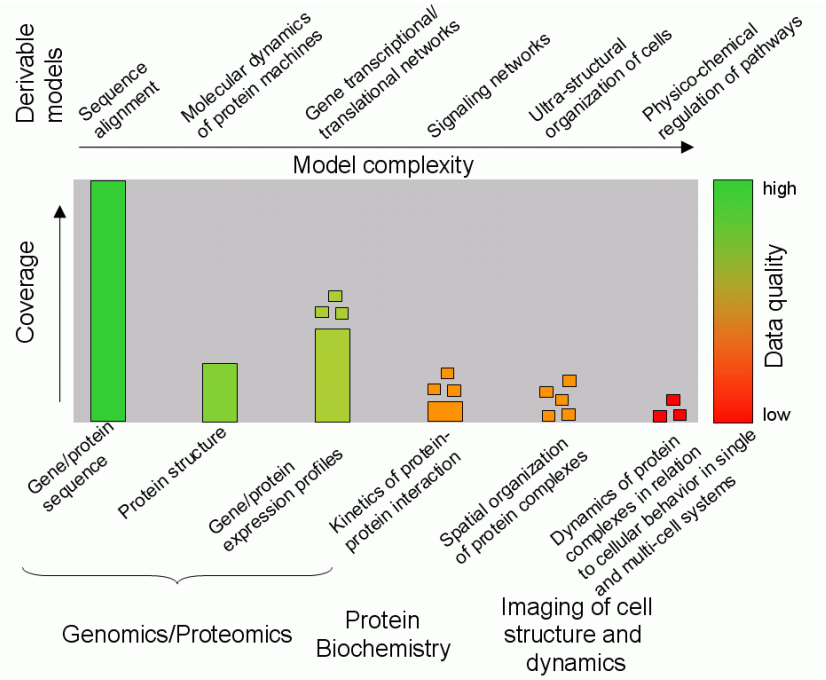

Fig. 1 **Data coverage problem of systems biology**: The field is rich in simple data types such as gene sequences and expression profiles, but poor in complex data types such as spatial and temporal protein dynamics as related to cellular outputs. The problem is amplified by the fact that complex data types are generally available only in isolated pools of relatively low quality that cannot be mined systematically and quantitatively. Significant challenges are awaiting the Computer Vision community as Systems Biology seeks to move from relatively simple models of, e.g., gene expression profiles to integral descriptions of cellular and multi-cellular systems, combining physical and chemical principles. New algorithms have to be devised for the complete and automated characterization of molecular and cellular processes in space and time. These data have to be seamlessly integrated with the physico-chemical models created by systems biologists. On the other hand, these models may be critical to increase the level of robustness and automation of the image analysis methods.

While we can expect organism-wide coverage of high-quality structural data as well as complete proteomic profiles perhaps in 10 – 20 years, the data needed to model molecular processes in cell and multi-cellular systems with spatial and temporal resolution are largely absent.

One of the most important approaches for acquiring such data is optical imaging. Electron microscopes (EMs) deliver spatial relationships of protein complexes with molecular resolution and partially molecular specificity. The limitation of this device is that the sample can not be observed under live conditions. Thus, no temporal information is extracted. However, sophisticated computational approaches are emerging which permit indirect reconstruction of ultrastructural dynamics of cellular machinery based on statistical processing of large EM data sequences. These provide excellent examples of the power of image analysis when tightly integrated with, in this case structural, modeling.

Driven by the demand of probing the dynamics of molecular systems in situ and in relation to functional outputs at the scale of cells or tissues the light microscope has seen an enormous revival over the last decade [7-9]. Optimization of the optical parts of the conventional light microscope has been paralleled with the development of very sensitive CCDs and of efficient fluorescent probes with high molecular specificity. Today the filming of the fate of a single molecule inside a cell is a standard approach to studying dynamic molecular processes, both in basic research labs and the biotechnological and pharmaceutical industry.

With these new possibilities, the pressure for image analysis methods that convert electron and light microscopical raw data into relevant information that can be used to build models of biological systems is rapidly increasing. As alluded to by Fig. 1 the quality of image-derived data is extremely poor, mainly due to lacking tools for robust computer vision. Whereas methods of low level vision, e.g. filtering or deconvolution, are incorporated in every commercial image analysis package for electron and light microscopy, the quantitation and interpretation of image contents, i.e. the tasks of intermediate and high-level computer vision, are essentially left to the investigator's visual inspection. Today's practice in a microscopy lab is that after an imaging experiment of one day, researchers spend weeks analyzing movies. It is intrinsic to manual measurements that the resulting data is subjective, incomplete, and not standardized. Images and derived data are very rarely exchanged between labs, and it happens too often that even within one lab years of image analysis work get lost when a lab member leaves. Consequently, efforts of systems biologists in beginning to model the

integrated physico-chemical regulation of molecular networks are hampered by the absence of appropriate data. In most cases it is not the raw images that are missing, but the systematic and complete image-derived information which is required to feed mechanistic models. Here is the call for the image processing and computer vision community to duplicate the achievements of the bioinformatics community on genomic data, but with images. It should be mentioned that the size of particularly live cell light microscopic data sets will soon exceed any other type of image data the computer vision community has ever dealt with, if it has not already done so. Therefore, developments of image analysis systems for light microscopy will be a well-received and very rewarding investment by our community.

2. CHALLENGES FOR COMPUTER VISION IN SYSTEMS BIOLOGY

Algorithms are required to quantify, in three dimensions, morphology, morpho-dynamics, and motion of very complex structures at the sub-cellular, cellular, and tissue-scale. The state-of-the art in image segmentation, texture analysis, motion tracking, etc. sets an excellent platform to start. However, novel challenges arise with the extremely low signal to noise ratio of electron and light microscope images, the geometrical anisotropy of image volumes, and the large variability of image features both in space and time. Computational image measurements need to be complemented with data mining methods that categorize image events. To achieve this goal, substantial challenges have to be tackled in terms of complexity and volume of image-derived data. For example, time-lapse sequences of light microscopic images easily contain millions of events per spectral channel. Multi-spectral imaging in two or three channels is already norm today, and we will soon see data sets with tens of channels. Systems biological models will require that several multi-spectrally resolved processes will be coupled in space and time. This will demand new methods to correlate complex data of hitherto unseen combinatorial dimensions.

Measurement and data mining algorithms should cope not only with huge and complex data sets, but they must be designed for detection of statistically very rare image events. There is increasing consensus in the biology field that molecular defects which cause pathologies at the macroscopic scale have only small effects (so-called weak phenotypes) at the cellular scale, where they can be studied experimentally. Thus, computer vision methods must be developed with near-zero tolerance for failures in identifying even the rarest image events. Literally, the search for the needle in the haystack is on, a new aspect for most computer vision systems.

In the same vein, biological processes exhibit a high level of heterogeneity. Heterogeneity is the signature of so-called homeostasis, i.e. the tendency of biological systems to counterbalance fluctuations in order to maintain a stable equilibrium state. The natural variation and transitions between states thus contains critical information about the auto-regulation of molecular systems. Computer vision analysis must be sensitive to distinguish all relevant states of a fluctuating biological system and to separate meaningful state fluctuations from noise. Classical training methods of algorithms that rely on pools of expected image events have to be redesigned to turn computer vision systems into self-tuned, adaptive processing pipelines that reliably accommodate previously "unseen" events.

Another aspect of image measurements for systems biological analyses is the integration of rigorous error propagation methods. Knowing the intrinsic uncertainty of image-derived variables is critical to classifying system states. Since fluctuations in the image measurements reflect both meaningful state fluctuations and measurement errors, an independent estimation of the effect of noise in the raw image signal on the derived variables must be implemented throughout the entire image analysis pipeline. Uncertainty estimates are also required to asses the uniqueness, the determinability and the sensitivity of a model derived from experimental data. These are important measures of model quality that are more and more evaluated with systems biological descriptions.

In summary, systems biology requires the contribution of the computer vision community to overcome the lack of spatially and temporally resolved data. Many of the existing algorithms will serve us well in fulfilling this demand. New challenges arise with the hitherto unseen complexity and variation of the data and with the need for robustness and completeness in the analysis. To achieve these goals, the computer vision community will have to devise domain-specific solutions. A new category of computer vision researchers will have to be grown whose focus is not to develop computer vision methods, but to address systems biological problems using computer vision approaches. Of all the challenges, this kind of marriage between the communities seems to be the toughest. Therefore, we have decided to present a session with outstanding examples of computer vision systems that have been developed in response to systems biological problems, and where innovative solutions have been implemented which set high standards both in terms of computer vision technology and its application to biological questions.

3. OVERVIEW OF THE SESSION

Since the systematic collection of spatially and temporally resolved data for systems biological modeling is still in its beginnings, an important application of computer vision is in the basic discovery of genes involved in a specific cell

function. This is achieved by identifying altered system behavior in response to molecularly specific manipulations. The assumption underlying such screens is that if perturbation of gene A does alter cell behavior and perturbation of gene B does not, gene A belongs, at least indirectly, to the pathways mediating the function while gene B does not. System responses are read out from microscopic image data using morphological and dynamic measures. Data sets of hundreds of thousands of cells must be evaluated to accumulate evidence for positive hits. Our session will feature an example of a genome-wide screen for factors involved in mitosis, the process in which replicated DNA is distributed from the dividing mother cell into the two daughter cells [10].

Similar screening strategies are also applied in drug discovery. The primary goal here is to identify chemical compounds which alter cell functions in a specific manner. The molecular modes by which the drug affects the behavior are of secondary importance only. Besides the obvious application of such screens to the search for new pharmacological agents, systems biology gains from them specific inhibitors that can be used to probe the mechanisms of a pathway. Our session will feature two presentations of computer vision systems [11, 12] that have been implemented for fully automated, large-scale drug profiling and one talk that takes image-based screening one step further to identify cell therapies [13].

Finally, the session will contain two talks which discuss novel algorithms for motion tracking and semantic analysis of cell biological images that directly feed mechanistic models of cellular functions [14, 15].

4. REFERENCES

1. Clark, D.P. and L.D. Russell, *Molecular Biology Made Simple and Fun*. 2000, Carbondale: Cache River Press.
2. Alberts, B., et al., *Molecular biology of the cell*. 4 ed. 2002, New York: Garland Science.
3. Venter, J.C., et al., *The Sequence of the Human Genome.* Science, 2001. **291**: p. 1304-1351.
4. Elbashir, S.M., et al., *Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells.* Nature, 2001. **411**(6836): p. 494-498.
5. Palsson, B., *Systems Biology : Properties of Reconstructed Networks*. 2006, Cambridge: Cambridge University Press.
6. Kitano, H., ed. *Foundations of Systems Biology*. 2001, The MIT Press: Cambridge.
7. Weijer, C.J., *Visualizing Signals Moving in Cells.* Science, 2003. **300**(5616): p. 96-100.
8. Lippincott-Schwartz, J. and G.H. Patterson, *Development and Use of Fluorescent Protein Markers in Living Cells.* Science, 2003. **300**(5616): p. 87-91.
9. Stephens, D.J. and V.J. Allan, *Light Microscopy Techniques for Live Cell Imaging.* Science, 2003. **300**(5616): p. 82-86.
10. Ellenberg, J. *Time-lapse microscopy based genome wide RNAi screening in live human cells*. in *ISBI'06*. 2006. Arlington: IEEE.
11. Loo, L.-H. *Automated Drug Effect Profiling Using Fluorescent Microscopy Images*. in *ISBI'06*. 2006. Arlington: IEEE.
12. Price, J. *Cell-Image-Based Drug, Genomics and Toxicity Screening*. in *ISBI'06*. 2006. Arlington: IEEE.
13. Olivo-Marin, J.-C. *Color image analysis for automated cell culture screening*. in *ISBI'06*. 2006. Arlington: IEEE.
14. Manjunath, B. *Towards Automated Bioimage Analysis: from features to semantics*. in *ISBI'06*. 2006. Arlington: IEEE.
15. Sibarita, J.-B. *Quantification of membrane trafficking on a 3D cytoskeleton network in living cells*. in *ISBI'06*. 2006. Arlington: IEEE.