Methods 96 (2016) 33–39

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Building cell models and simulations from microscope images

Robert F. Murphy *

Computational Biology Department, Center for Bioimage Informatics, and Departments of Biological Sciences, Biomedical Engineering and Machine Learning, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA
Freiburg Institute for Advanced Studies and Faculty of Biology, Albert Ludwig University of Freiburg, Germany

ARTICLE INFO

Article history:
Received 1 August 2015
Received in revised form 15 October 2015
Accepted 16 October 2015
Available online 17 October 2015

Keywords:
Computational biology
Image-based modeling
Cell organization
Cell shape
Generative models

ABSTRACT

The use of fluorescence microscopy has undergone a major revolution over the past twenty years, both with the development of dramatic new technologies and with the widespread adoption of image analysis and machine learning methods. Many open source software tools provide the ability to use these methods in a wide range of studies, and many molecular and cellular phenotypes can now be automatically distinguished. This article presents the next major challenge in microscopy automation, the creation of accurate models of cell organization directly from images, and reviews the progress that has been made towards this challenge.

1. Introduction

Biochemistry and structural biology were revolutionized by the ability to replace rough approximations of molecular shape and interactions, such as "rods," "sheets," and "globules" with spatially accurate models of protein structure directly learned from experimental data (such as from X-ray crystallography). Since molecules rarely have only a single structure, this led to probabilistic models for structures and structural transitions. This further enabled a critical advance: the ability to computationally simulate expected behaviors of molecules without requiring further experiments [1].

Cell biology has only begun to appreciate the need for a similar revolution in the way in which cell structure is represented. Currently, an explicit representation of organelle structure is avoided entirely; words, such as Genome Ontology terms, are used to refer to organelles with the assumption of a shared understanding of the structures they display. Communicating that understanding is done by hand-drawn cartoons or example images. Example images may include high-resolution reconstructions for a single organelle or cell, but these are only instances and do not capture the expected variation in that structure. Variation observed in images of a structure may be intrinsic (e.g., endosomes vary in size and shape) or due to measurement noise (e.g., coated vesicles may appear to vary in size or shape due to digital imaging of a low fluorescence signal). By assuming intrinsic variation is small, reconstruction methods have been used on many images to produce refined structures (primarily at the micron level) [2,3]. However, for most organelles, intrinsic variation is dramatic.

A natural question then becomes how we can represent the structure of cellular components that show significant intrinsic variation in their size or shape. This leads to a larger question: how can we create predictive models that capture variation in the organization of entire cells? In order to enable prediction, such models need to be *generative* rather than *descriptive*. The distinction can easily be seen by considering the task of distinguishing pictures of apples from oranges. This can be done using a single feature such as color, combined with the rule that an object is red if and only if it is an apple. However, if the task is to *create* an apple, knowing that apples are red is not nearly enough. As discussed below, generative models require some choices regarding the completeness or effectiveness of the description.

Automation of descriptive analysis of high resolution/high content cell images has progressed dramatically in the past twenty years [4–8]. The direct creation of generative models from cell images represents the next major challenge in high content analysis. There are a number of reasons why such generative models would be useful. First, as just discussed, they would capture the underlying spatial relationships in a collection of images, that is, they estimate not only the most probable reconstruction of each individual cell or organelle (e.g., removing noise in imaging) but also the modes of variation between individual instances. As such they are well-suited for representing the large collections of

[*] Address: Computational Biology Department, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.
E-mail address: murphy@cmu.edu

http://dx.doi.org/10.1016/j.ymeth.2015.10.011
1046-2023/© 2015 Elsevier Inc. All rights reserved.

images enabled by the development of high content screening and automated microscopy [9]. Second, models learned separately for different organelles or structures (i.e., from different sets of images) could potentially be combined to synthesize a cell containing all of those organelles in the same cell (assuming that the organelles do not affect each others position or shape). Third, such models could be used to predict the distribution of an organelle in a new cell type (e.g., with a different cell and nuclear arrangement), and those predictions could be rapidly confirmed (or modified) using feature-based approaches without having to build a separate generative model for all organelles in each cell type. Fourth, generative models could provide a better framework for connecting morphology to the mechanisms that produce it, since biochemistry could be directly linked to the parameters of the generative model. Fifth, databases of generative spatial models could provide an important complement to Genome Ontology terms, providing a spatial definition of those terms. Lastly, instances drawn from generative models could be used as the basis for spatially realistic simulations of cellular biochemistry. There are a number of powerful systems for performing such simulations, such as MCell [10], VirtualCell [11], Simmune [12] and SmolDyn [13], but most simulations currently performed with those systems use a very limited number of manually-segmented and manipulated images to provide compartment geometries. Generative models can provide, without manual editing, large numbers of cell geometries with the closed structures that are needed for such simulations.

Some basic criteria for the creation of such generative models of cellular structure and organization have been proposed previously [14]. These were that the models be

(i) automated: learnable automatically from images;
(ii) generative: able to synthesize new, simulated images displaying the specific pattern(s) learned from images;
(iii) statistically accurate: able to capture pattern variation between cells; and
(iv) compact: representable by a small number of parameters and communicable with significantly fewer bits than the training images.

As with most modeling efforts, satisfying the latter two criteria requires balancing between the complexity and the completeness of the models (a version of the bias-variance tradeoff [15]). An illustration is the choice of whether to model the shape of an individual organelle (such as a mitochondrion), using an ellipse, which is very compact, or a mesh, which captures every surface irregularity. Converting these representations into generative models differs greatly in the amount of training data required – learning a statistical model of the variation of two or three axis lengths requires far less data than accurately capturing the relationships between hundreds or thousands of minor surface variations.

2. Overview of learning and use of generative models of cell organization

Over a number of years and contributions from a number of participants, the open source CellOrganizer system has been created as a step towards meeting the need for learning and using image-based generative cell models [14,16–23]. The basic principles of the CellOrganizer pipeline are illustrated in Fig. 1, and are generally applicable to efforts in this area. The input is a collection of cell images, most frequently of cells tagged with fluorescent probes specific for one or more proteins or organelles. We begin creation of models from those images by starting with the major geometric components of the eukaryotic cell, the overall cell and nuclear shape as reflected by the positions of the plasma and nuclear membranes. This choice is made not only because these components provide a logical starting point but also because they are easy to define even when specific probes are not include to delineate them. For example, we can get a reasonably good estimate of the plasma membrane position from the autofluorescence in fluorescent channels used to image other proteins; in the rare cases where a nuclear marker is not present, we can frequently make a good estimate of the position of the nuclear membrane from the "hole" present in the pattern of other markers. Having constructed a model of cell and nuclear shape (which we refer to as a framework model), we next construct other models (e.g., for organelles) that are conditional on those shapes.

The parameters of the learned model can then be readily compared with those of other, previously learned models, e.g., for different cell types or conditions. We can also use one or more learned models to synthesize an idealized cell instance free from blur or noise from imaging. This spatial representation of a cell instance can then be used to provide the geometries of compartments for use with biochemical models involving different organelles.

3. Constructing models

We next turn to some specifics on how generative cell models can be created, including general principles of how CellOrganizer creates models from image collections. CellOrganizer is a Matlab package that is accessed by a small number of interface functions that learn models from images or movies, compare models, and synthesize images or movies from models. Control over the operation of those functions is achieved by setting various parameters in a control structure. The starting point for learning a model is a collection of images. Given the many tools available to segment images into individual cell regions and the frequent need to tailor segmentation to a specific collection, CellOrganizer assumes either that input images either contain single cells or that masks are provided to define the region corresponding to each cell. It is called with strings specifying the paths (including optional wildcards specifying subsets of files in those paths) to images of a cytoplasmic or cell boundary marker, a nuclear marker, and any specific organelle markers or tagged proteins. An optional path can be given to provide mask images. The algorithms to be used (and any parameters that they require) are specified through a parameter structure. The primary output is a file containing the learned model of the cellular components and the relationships between them, and additional outputs, such as files containing the model parameters for each cell, can be requested.

An important consideration in constructing cell models is that they can only reflect the properties of the cells in the image collection used for training. Thus, models learned from a collection of fully differentiated cells cannot of course capture behaviors shown by cells during the differentiation process (i.e., they cannot synthesize "green apples" if trained with only images of red ones). Given a collection of images of cells at various stages of differentiation, the model can learn variation in organization associated with that differentiation. Similarly, movies rather than static images are needed as input to learn dynamic behaviors (an example of this is discussed below). However, it should be noted that models learned from static images can with some assumptions be used to simulate dynamics.

3.1. Cell framework models

3.1.1. Nuclear shape models

To illustrate the process, we turn first to modeling the most basic structural elements of cells, the nuclear and cell membranes.