CellSium – Versatile Cell Simulator for Microcolony Ground Truth Generation

Christian Carsten Sachs[1], Karina Ruzaeva[1,2,†], Johannes Seiffarth[1,3,†], Wolfgang Wiechert[1,3], Benjamin Berkels[2], Katharina Nöh[1]

[1]Institute of Bio- and Geosciences, IBG-1: Biotechnology, Forschungszentrum Jülich GmbH, Jülich, Germany

[2]Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Aachen, Germany

[3]Computational Systems Biotechnology (AVT.CSB), RWTH Aachen University, Aachen, Germany

[†]Equal contribution

Abstract

Summary: To address the need for ground truth data for deep learning based segmentation algorithms in microfluidic live-cell imaging, we present *CellSium*, a cell simulator primarily aimed at synthesizing realistic image sequences of bacterial microcolonies suitable for training neural networks.

Availability and Implementation: *CellSium* is free and open source software under the BSD license, implemented in Python, available at https://github.com/modsim/cellsium (DOI: 10.5281/zenodo.6193033), along with documentation, usage examples and Docker images.

Contact: k.noeh@fz-juelich.de

Supplementary information: Supplementary data are available online.

Introduction

Deep learning (DL) based segmentation methods have become standard for bioimage analysis, largely surpassing traditional approaches (Jeckel and Drescher, 2021). A bottleneck in the application of DL techniques is the need for comprehensive ground truth (GT) data for training and validation. For the task of cell segmentation, for example, pixel-perfect masks discerning individual cells from the background are crucial. Such data is, however, problem-specific and laborious to produce (Jeckel and Drescher, 2021). Synthesizing data along with its GT is a tried and tested way to alleviate this bottleneck. Cell simulators strive to form a better understanding of biological phenomena, by modeling molecular aspects, rather than generating photorealistic images (e.g., Gutiérrez *et al.*, 2017). While for eukaryotic cells, data augmentation using (fluorescence) image generators is established (e.g., Lehmussola *et al.*, 2007; Svoboda and Ulman, 2017), apart from an example for brightfield yeast image generation (Kruitbosch *et al.*, 2022), the benefits of synthetic imagine generation have not been exploited for microbial image analysis.

To serve the needs of training data hungry DL-approaches to analyze phase contrast image stacks, we have developed the bacterial microcolony simulator *CellSium*. We demonstrate its applicability for object detection (YOLOv5) as well as semantic segmentation (Mask R-CNN) using synthetic data and verify the resulting neural networks with real image data. Further usage scenarios beyond image synthetization are also showcased.

Approach

CellSium is an agent-based simulator. Internally, each cell is represented as a Python object, with individual properties, such as shape or growth behavior, implemented as mixins following object-oriented programming paradigms. Various unicellular geometries are implemented, such as straight and bent rods, coccoid (circular) and ellipsoid shapes (SI Fig. S.1). For each simulated time step, the `grow()` method is called, implementing phenomenological cell size homeostasis (growth) models (SI Sec. 4). Geometries are used with a physics engine to calculate realistic cell placements. Various output routines exist, such as realistic phase contrast images, TrackMate XML (Tinevez *et al.*, 2017), or simulation-ready input geometries (STL) for CFD simulations (SI Sec. S.3).

The image generation process is shown in Fig. 1A: A “perfect” (noiseless) image is generated (A1-A5), which is deteriorated by adding different kinds of noise. The typical phase contrast “halo” around the cells is generated by Gaussian blurring of the image (A2-4), which is combined to mimic a phase contrast image (A5). Next, uneven illumination is added (A6), along with additive/multiplicative noise (A7,9), yielding a realistic phase contrast bacterial cell image (A10). Optionally, fluorescence can be generated using a Gaussian point spread function (SI Fig. S.2).

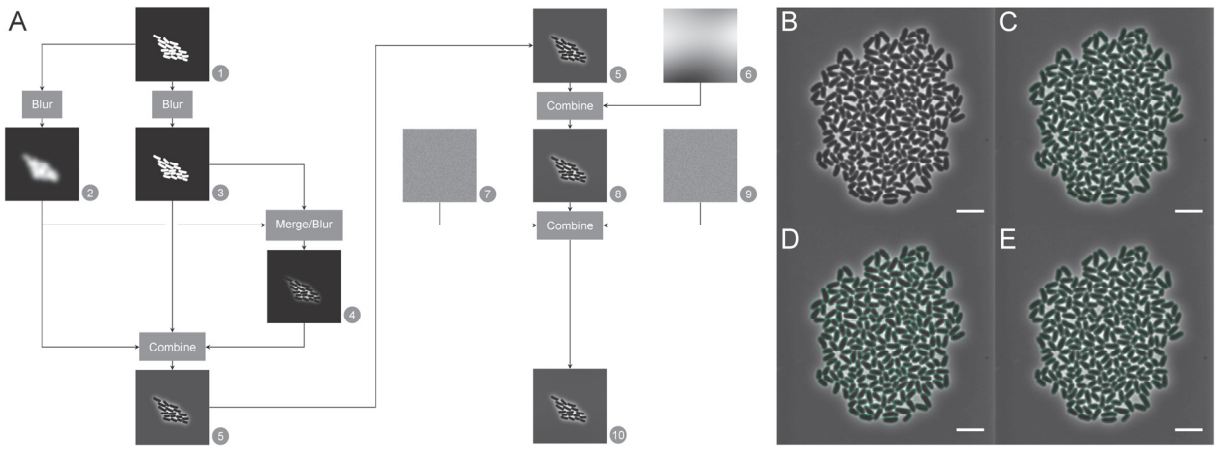

Figure 1. A: Image generation flow to produce realistic phase contrast images. **B**-**E**: Prediction results of neural networks trained with data simulated using *CellSium* and evaluated with real data. **B**: input image (a *Corynebacterium glutamicum* microcolony), **C:** manually corrected GT, **D:** YOLOv5 object detector result, and **E:** Mask R-CNN segmentation network result.

Methods

CellSium is implemented in the Python programming language, using the NumPy and SciPy libraries, matplotlib and OpenCV for image generation. Physical placement simulation is performed using the off-the-shelve physics library PyMunk. Source code, PyPI/ Anaconda packages, and a Docker image are available, enabling platform-independent usage. The documentation includes API documentation, as well as usage examples for configuring time-lapse simulations, 3D cell geometries, and parametrizable cell models (SI Sec. S1).

Use case: *CellSium* as GT generator

To assess the applicability of *CellSium* as a GT generator, a dataset was generated using the *YOLOOutput* and *COCOOutput* modules (128 images, 512x512, 0.09 µm/pixel, 0 to 512 cells per frame). The outputs were directly used to train the object detector/segmentation frameworks YOLOv5 (Jocher et al, 2020) and the Mask R-CNN (He *et al.*, 2017) module of MMDetection (Chen *et al.*, 2019). As test data, a microcolony image of *Corynebacterium glutamicum* was used (Fig. 1B), which was interactively segmented using the Trainable Weka Segmentation and then hand-corrected (Fig. 1C). YOLO and Mask R-CNN yielded a $mAP_{0.5}$ of 0.994 and 0.987, respectively, for bounding box object detection. Configuration parameters and further result metrics are given in SI Tab. S.1; all files to reproduce the evaluation are available in the GitHub repository.

Conclusion

CellSium is a microcolony simulator primarily aimed at bacterial image dataset GT generation. Simulated training data were proven useful for state-of-the-art object detector/segmentation frameworks (YOLOv5, Mask R-CNN), yielding networks capable of producing competitive results with real time-lapse microscopy images. Compared to the laborious manual annotation of image sequences, taking an in silico approach allows generating training data with desired characteristics at any amounts, which benefits training and method verification. *CellSium* can both be easily used as-is, or, due to its pluggable architecture, serve as a base for implementing other cell types, developing custom cell behavior within growing microcolonies, or embedding the image generation, for example directly in a DL training loop for continuous procedural training data generation. *CellSium*'s current image generation capabilities are well-suited for training networks tailored to one imaging modality; a future step is to implement generative adversarial networks (GAN) to achieve a high level of realism, independent of the modality.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [WI 1705/16-2], the Helmholtz Imaging Platform HIP [SATOMI, ZT-I-PF-04-011], and the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

References

Chen,K. *et al.* (2019) MMDetection: Open MMLab detection toolbox and benchmark. *arXiv:1906.07155*.

Gutiérrez,M. *et al.* (2017) A new improved and extended version of the multicell bacterial simulator gro. *ACS Synth. Biol.*, **6**, 1496–1508.

He,K. *et al.* (2017) Mask R-CNN. In, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.

Jeckel,H. and Drescher,K. (2021) Advances and opportunities in image analysis of bacterial cells and communities. *FEMS Microbiol. Rev.*, **45**, 1–14.

Jocher,G. (2020) ultralytics/yolov5: v3.1 - Bug fixes and performance improvements.

Kruitbosch,H.T. *et al.* (2022) A convolutional neural network for segmentation of yeast cells without manual training annotations. *Bioinformatics*, **38**, 1427–1433.

Lehmussola,A. *et al.* (2007) Computational framework for simulating fluorescence microscope images with cell populations. *IEEE Trans. Med. Imaging*, **26**, 1010–1016.

Svoboda,D. and Ulman,V. (2017) MitoGen: A framework for generating 3D synthetic time-lapse sequences of cell populations in fluorescence microscopy. *IEEE Trans. Med. Imaging*, **36**, 310–321.

Tinevez,J.-Y. *et al.* (2017) TrackMate: An open and extensible platform for single-particle tracking. *Methods*, **115**, 80–90.