Integrated tools for next generation Bio-Imaging

Joaquin Correa[1,2]
JoaquinCorrea@lbl.gov
Manfred Auer[1,3]
MAuer@lbl.gov
David Skinner[1,2]
DESkinner@lbl.gov

[1] Lawrence Berkeley National Laboratory, Berkeley, USA
[2] National Energy Research Scientific Computing Center (NERSC), USA
[3] Life Sciences Division

Abstract

Exponential data scaling in Bio-Imaging, combined with the data fusion demands of increasingly multi-modal imaging approaches, require R&D in computational Bio-Imaging solutions that extend the ways in which scientists can leverage image based data. Bio-Imaging is leaving an era of images-as-results and entering an era where results come in the form of biological models built on big images and vast image collections. These models require scalable analysis, integration, and dissemination of image data delivered from advanced Bio-Imaging instrumentation.

In Big Data Bio-Imaging, images are the feedstock rather than the end result of the instruments we use to observe biological systems.

This work presents a web-based HPC-enabled one-stop-shop solution for producers and consumers of models built on imaging data.

1 Introduction

A fundamental problem currently for the biological community is to adapt computational solutions known broadly in data-centric science toward the specific challenges of data scaling in Bio-Imaging. This is of particular significance and urgency as recent developments in image data acquisition have moved the data set size from gigabytes per imaging session to terabytes, making it impossible for biologists to continue their common practise to manually analyse the data sets. In this work we target software solutions fit for these tasks, which leverages success in large-scale data-centric science outside of Bio-Imaging.

2 Architecture

2.1 Science gateways

A science gateway is a web-based interface to access high performance computing (HPC) resources and storage systems. Gateways allow science teams to access data, perform shared computations and generally interact with HPC resources over the web. Common gateway goals are to improve ease of use in HPC so that more scientists can

© 2014. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

benefit from creating collaborative workspaces around data and computing and make data accessible and useful to the broader scientific community.

Science gateways can use a REST-based web API such as NEWT [7] to access HPC centers, including authentication, file management, job submission and accounting interfaces. These interfaces allow users to perform tasks through the web [8].

2.2 Image processing software stack

This work illustrates the deployment of custom-tailored image-processing algorithms and successful integration of a suite of universally used processing software among the biology community into an OMERO-based [1] science gateway [8] using HPC resources from the National Energy Research Scientific Computing Center (NERSC) [11].

The shared, scalable web service presented in this paper provides a modular and flexible one-stop-shop solution for producers and consumers of models built on imaging data by refining pixel data into actionable knowledge resources. As a result, this work offers a platform that involved co-design, has multi-modal capabilities, and is big-data ready.

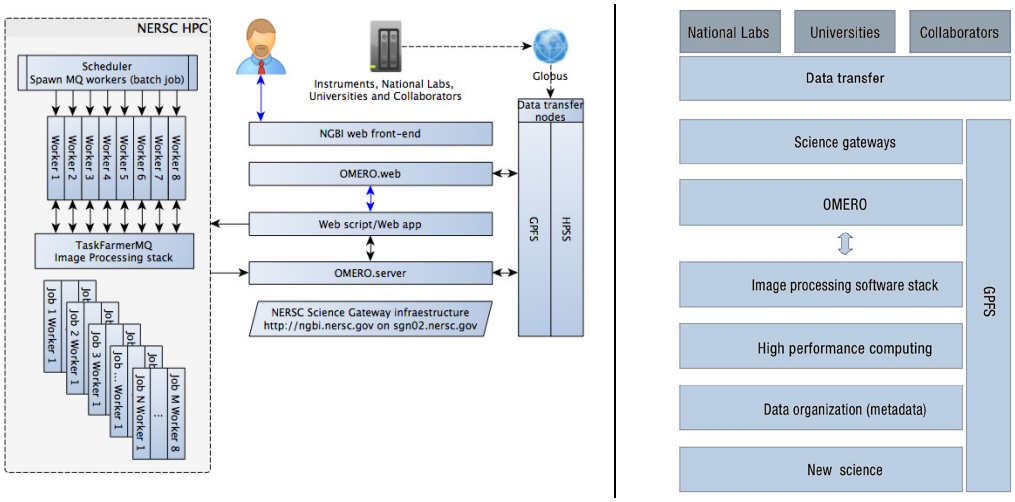
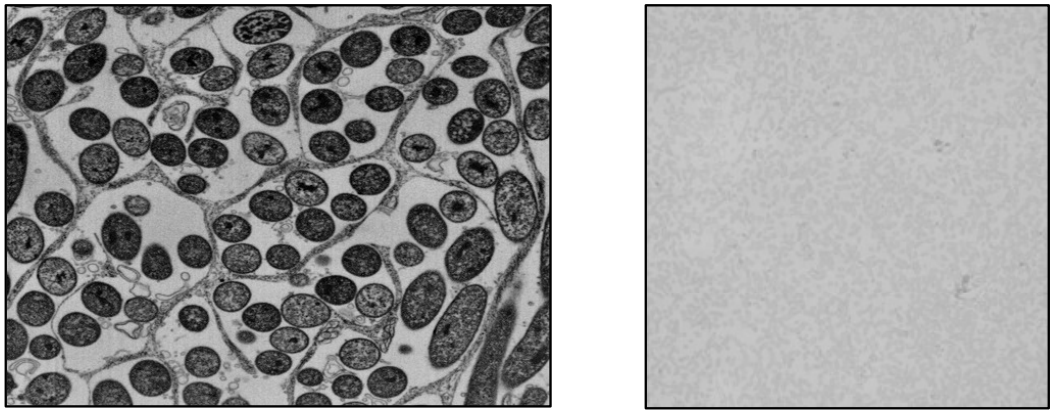


Figure 1: Architectural design. From top to bottom: (a) Data generation: National Laboratories, Universities and Collaborators (b) Data transfer to HPC facility (c) Web infrastructure: Science gateways [6] and OMERO [1] (d) Image processing stack: FIJI [3], Weka [3], Sci-kit learn [5], VLFeat [4], IMOD [9], CellProfiler [10] and UCSF Chimera [15] (e) Outcome: Data organization and New science.

3 Methodology

The authors demonstrate the capabilities of the platform presented in this work in the context of two monoculture bacterial biofilms, which are communities of bacterial cells with often-complex architectural organization. Studying biofilm organization is central to our understanding of microbial physiology [12], given that community organization represents the predominant bacterial lifestyle. Furthermore biofilms play an important

role for chronic infections and other microbial-mediated processes [13]. 3D datasets of Myxococcus xanthus [12] and Desulfovibrio vulgaris RCH1 (unpublished) microbial communities were used to illustrate segmentation, classification and visualization capabilities.



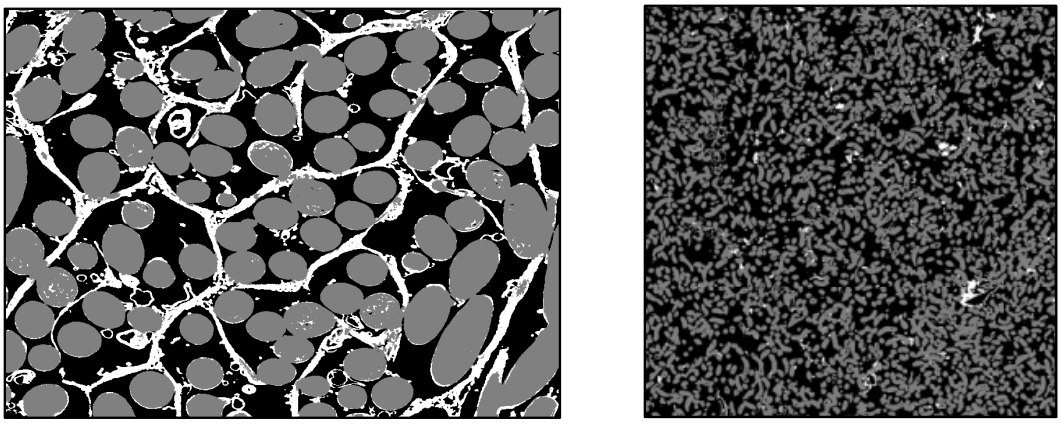
(a) (b)

Figure 2: Raw images: Showing a slice of the volume (a) Myxococcus xanthus (heavy metal stained) showing thin-walled partitions between cells (b) Desulfovibrio vulgaris RCH1 (unstained) showing extracellular metal deposits.

3.1 Segmentation and Classification

For the use cases the authors present; each dataset was segmented into three different classes (background, bacteria and biofilm), manually curated classifier models served as an input to the workflow.

Pre-processing enhancement operations were performed based on the image characteristics followed by image segmentation and classification achieved by using Weka [3] Fast Random Forest (FRF) implementation and a software layer that enables HPC by parallelizing serialized tasks.

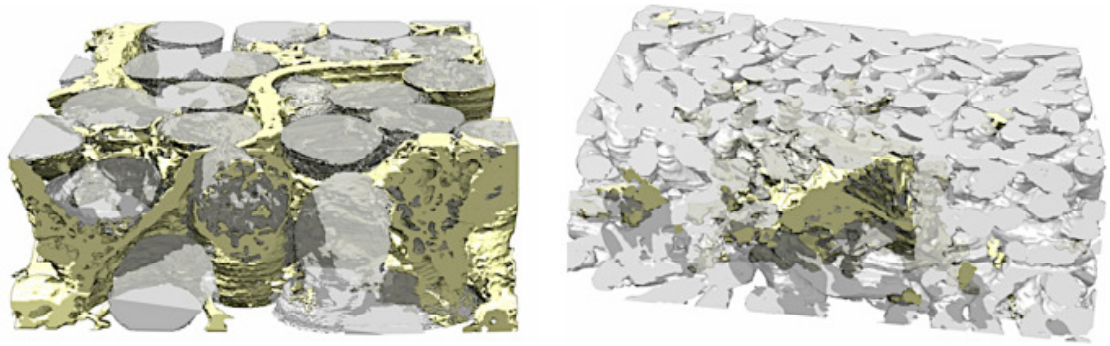


(a) (b)

Figure 3: Segmented datasets showing a slice of the volume: Background in black, bacteria in grey and partitions and metal deposits, respectively, in white (a) Myxococcus xanthus (b) Desulfovibrio vulgaris RCH1

3.2 Visualization

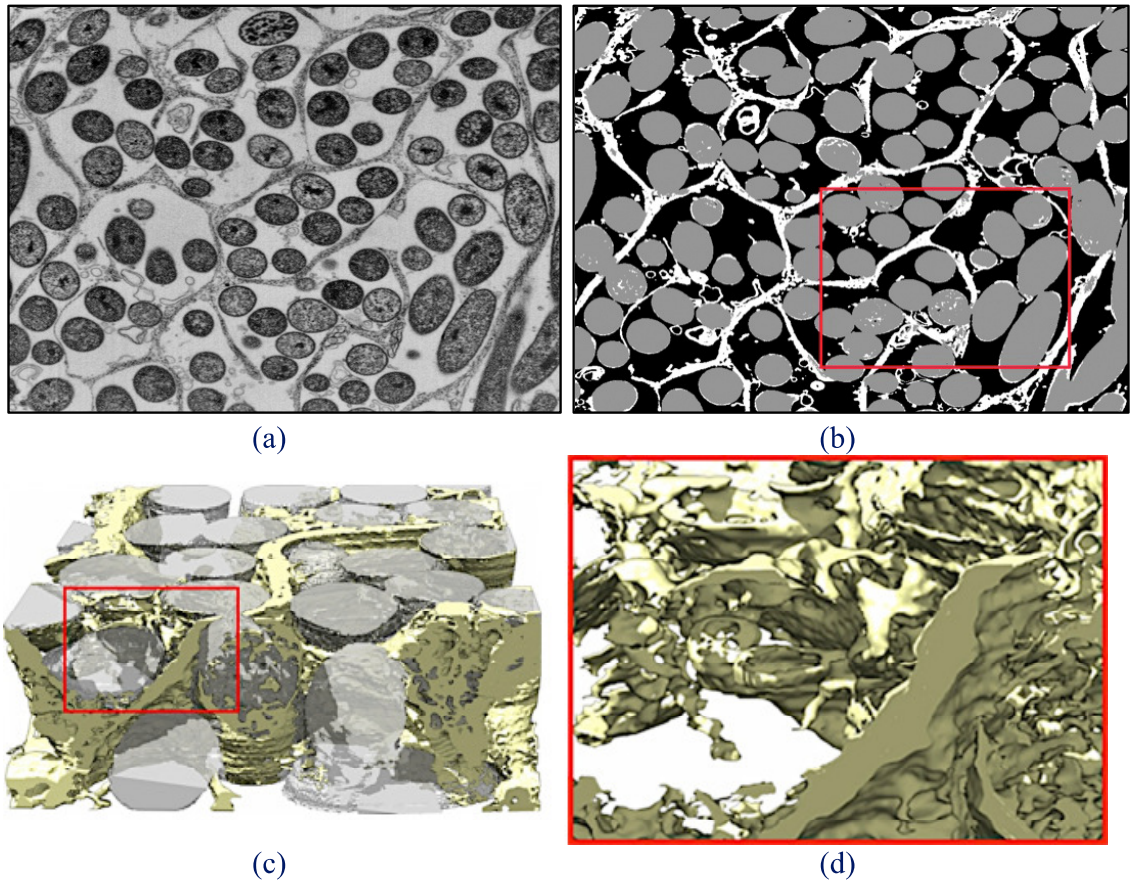
Corresponding segmented masks are used to produce 3D representation of the data using UCSF Chimera [15] as a running backend application accessed based on demand.



(a) (b)

Figure 4: Visualization of ROI with bacteria shown in grey and (a) partitions shown in yellow for Myxococcus xanthus or (b) extracellular metal deposits shown in yellow for Desulfovibrio vulgaris RCH1

4 Workflow summary

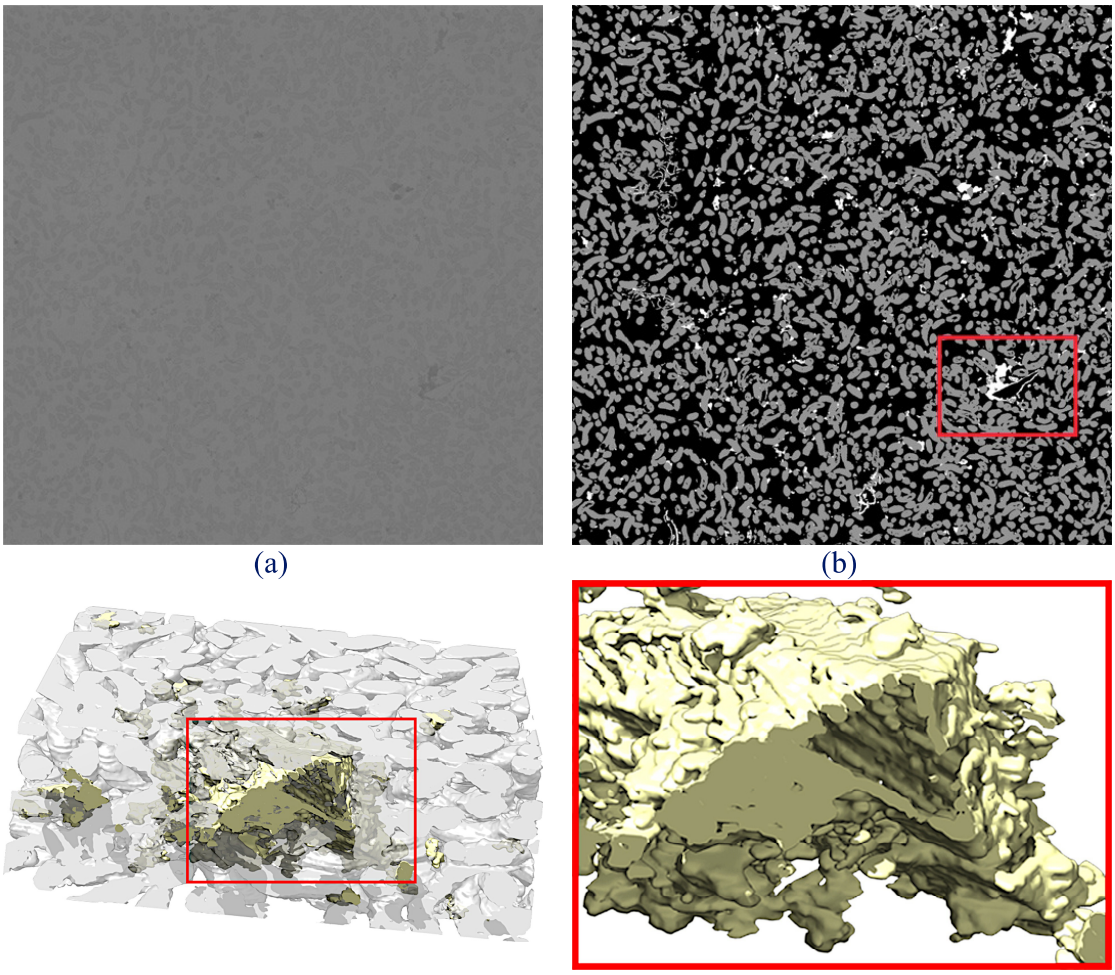


(a) (b)

(c) (d)

Figure 5: Myxococcus xanthus (a) Raw image showing a slice of the volume (b)

Segmented dataset showing a slice of the volume and ROI (c) 3D rendering of ROI showing bacteria and biofilm (d) higher magnification of ROI sub-region of (c) showing the 3D organization of the thin walls, with bacteria not being displayed.



(a) (b) (c) (d)

Figure 6: Desulfovibrio vulgaris RCH1 (a) Raw image showing a slice of the volume (b) Segmented dataset showing a slice of the volume and ROI (c) 3D rendering of ROI showing bacteria and extracellular metal deposits (d) higher magnification of ROI sub-region of (c) showing the 3D organization of the metal deposits, with bacteria not being displayed

5 Conclusions

The platform we have presented provides new qualitative understanding of the 3D organization of the *Myxococcus xanthus* and *Desulfovibrio vulgaris RCH1* microbial communities [12] [13] and holds promise for image analysis of the big data flood, which is just beginning. Among the data challenges before the bio-imaging community are high-throughput, high-content and multi-modal image data, which are currently being produced at unprecedented scales.

Acknowledgment

This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231.

References

[1] G. J. Kleywegt, G. Zanetti, J. R. Swedlow, et al., OMERO: flexible, model-driven data management for experimental biology. *Nature Methods 9, 245–253,* 2012.

[2] K. Eliceiri, P. Tomancak, A. Cardona, et al., Fiji: an open-source platform for biological-image analysis, *Nature Methods 9(7): 676-682.* 2012.

[3] B. Pfahringer, P. Reutemann, I. H. Witten, et al., *The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.* 2009.

[4] A. Vedaldi and B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008), *URL: http://www.vlfeat.org/,* 2008.

[5] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

[6] NERSC, Bio-Imaging: High Performance Computing Facility Operational Assessment for the National Energy Research Scientific Computing Center 2013. 2014.

[7] NEWT, URL: https://newt.nersc.gov/

[8] National Energy Research Scientific Computing Center (NERSC) Science gateways, URL: http://www.nersc.gov/users/science-gateways/

[9] Kremer J.R., D.N. Mastronarde and J.R. McIntosh (1996) Computer visualization of three-dimensional image data using IMOD. J. Struct. Biol. 116:71-76.

[10] J. Moffat, P. Golland, DM. Sabatini, et al., *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome Biology 7:R100. PMID: 17076895. 2006.

[11] National Energy Research Scientific Computing Center (NERSC), URL: http://www.nersc.gov/about/

[12] JP. Remis, JW. Costerton, M. Auer. Biofilms: structures that may facilitate cell-cell interactions. ISME J. 2010 Sep;4(9):1085-7.

[13] JD. Wall, LR. Krumholz (2006) Uranium reduction, Annu Rev Microbiol, 60: 149-166

[14] Integrated tools for Next Generation Bio-Imaging (NGBI), URL: http://ngbi.nersc.gov

[15] EF. Pettersen, TD. Goddard, CC. Huang, et al., UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605-12.