HPC Enabled Data Analytics for High-Throughput High-Content Cellular Analysis

Ross A. Smith, Rhonda J. Vickery, Jack Harris
Engility Corporation
U.S. Air Force Research Laboratory WPAFB, OH 45433, U.S.A.

Sara Gharabaghi, Thomas Wischgoll, David Short
Dept. of Computer Science
Wright State University,
Dayton, OH 45435, U.S.A.

Robert Trevino, Steven A. Kawamoto, Thomas J. Lamkin, Kevin Schoen
U.S. Air Force Research Laboratory WPAFB, OH 45433, U.S.A.

Eric E. Bardes, Scott C. Tabar, Bruce J. Aronow
Division of Biomedical Informatics,
Cincinnati Children's Hospital Medical Center,
Cincinnati, OH 45229, U.S.A.

***Abstract*—Biologists doing high-throughput high-content cellular analysis are generally not computer scientists or high performance computing (HPC) experts, and they want their workflow to support their science without having to be. We describe a new HPC enabled data analytics workflow with a web interface, HPC pipeline for analysis, and both traditional and new analytics tools to help them transition from a single workstation mode of operation to power HPC users. This allows the processing of multiple plates over a short period of time to ensure timely query and analysis to match potential countermeasures to individual responses.**

Keywords—High-Throughput High-Content Screening, Pipeline.

The INSIGHTS project discovers beneficial chemical or genetic interrogations for human performance, medical intelligence, or therapeutic application [1]. The envisioned use cases are force protection of U.S. warfighters from natural exposure or biological attack. Rapid screening of numerous possible countermeasures decreases the time-to-decision from years to months or even weeks. The methodology is based on high-throughput, high-content biological screening and matching potential countermeasures to individual response. Individual experiments are done within wells of an experimental plate (see Fig. 1). After incubation, each well is observed at multiple sites, and resulting images are submitted to the pipeline for cell segmentation (CS), generating image masks to identify each cell in a site. Each cell then goes through featurization generating 11,000+ features. Feature selection removes all but the most informative features. Finally, well scoring identifies interrogations for further investigation.

Each experimental batch has multiple plates where each well is imaged at multiple sites at frequencies optimized with commonly used biological fluorescent stains or when using non-fluorescent phase contrast imaging. This approach typically yields 4 images per site and 12,000+ images per plate. After the images have been collected, they are uploaded to a file system mounted by the HPC head nodes as shown in Fig. 2. The researcher ensures batch metadata is linked via a project database and launches analysis jobs via an easy to use μBatch web

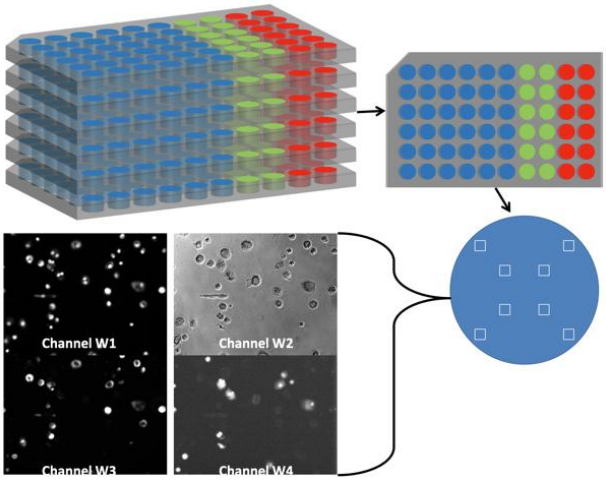

Fig. 1. An experimental batch can have multiple plates. Each plate has 384 wells. Each well is sampled at eight sites. Four images are captured at each site based on filters and dye frequencies.

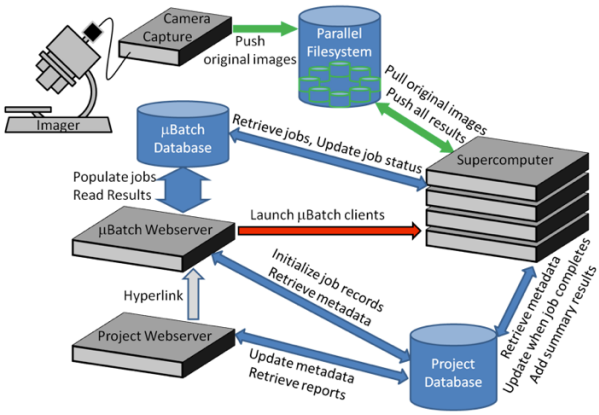

Fig. 2. Data movement through the system, including files, database accesses, job launches and user input. The 12,000+ images from a single plate take up 33 GB. Compressed feature data takes up 3 GB or more depending on the number of cells.

portal, where clients on HPC nodes retrieve µBatch jobs following a "bag of tasks" paradigm. The clients execute the data analytics pipeline, updating both the project database and the µBatch server.

The biologist has additional flexibility to do post analysis to facilitate verification and validation of results. An example is the ability to create traditional dose response curve plots from downloaded Kolmogorov–Smirnov reports, as shown in Fig. 3 with the first 15 of 32 plots displayed for a plate. The blue circles indicate actual points from the data and hovering over one of these shows the data values for that point, as shown in the center plot. The red curves indicate where dose response curves were found using the Hill equation, and a red dashed line marks the half maximal inhibitory concentration (IC50). This is a common analysis for comparison of individual treatments or responses. This is supplemented with a more comprehensive

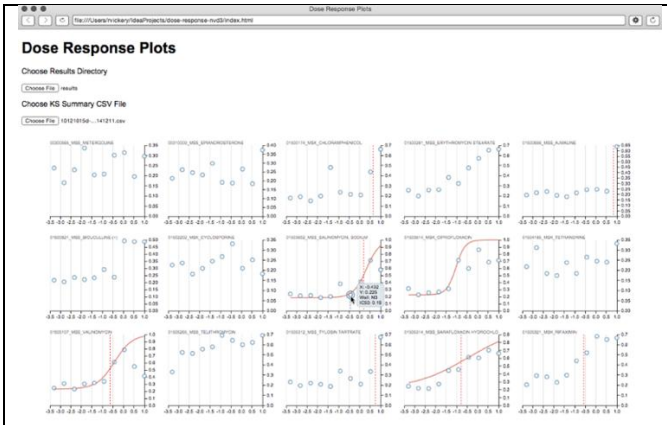

Fig. 3. Dose response curves with single point highlighted.

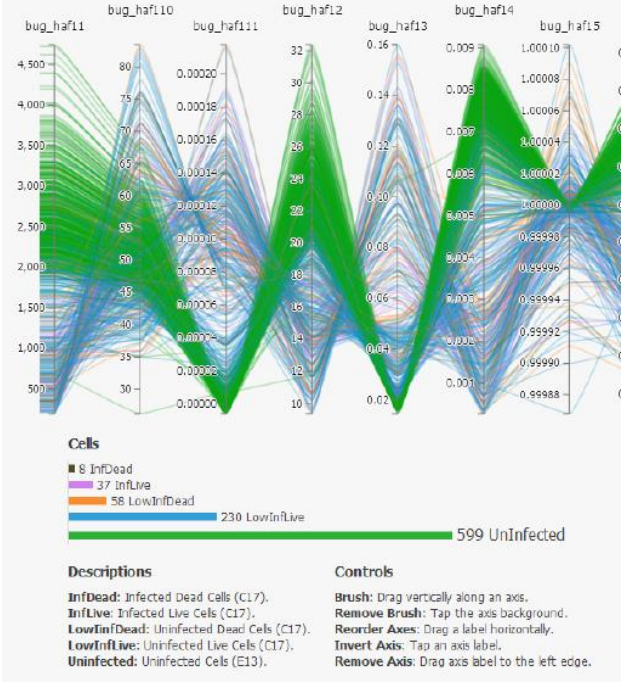

Fig. 4. Parallel coordinate visualization of Haralick features extracted from cell images.

comparison of features by logical grouping using parallel coordinates plots with controls to determine which treatments cause responses that are similar to uninfected cells, as shown in Fig. 4 [2]. In this way, both familiar and more comprehensive tools are available.

We are currently adapting the capability to other biology and chemistry oriented problem datasets. A public facing portal is also being developed.

ACKNOWLEDGMENTS

The authors would like to acknowledge the computational resources and PETTT software support from the DoD High Performance Computing Modernization office under Contract No. GS04T09DBC0017 and biological support for molecular signatures from the U.S. Air Force under Contract No. FA8650-14D-6516.

REFERENCES

[1] Vickery, R.J., Smith, R., Bardes, E., Tabar, S., Short, D., and Trevino, R., Utilization of Hybrid Computing for High Throughput Identification of Beneficial Chemicals or Biological Interrogations for Human Effectiveness, PETTT PP-ACE-KY06-003-P3 Final Technical Report, 21 August 2015.

[2] Wischgoll, T., Vickery, R., and Smith, R., Extending the Visual Analytics Framework for High Throughput Screening of Biological Infectious Agents Project, PETTT PP-ACE-KY05-005-P3 Final Technical Report, 22 August 2014.