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

Ross A. Smith[1], Rhonda J. Vickery[1], Sara Gharabaghi[2], Thomas Wischgoll[2], Jack J. Harris[1], Robert P. Trevino[3], Steven A. Kawamoto[3], Thomas J. Lamkin[3], Eric E. Bardes[4], Scott C. Tabar[4], Bruce J. Aronow[4], David Short[2], Kevin Schoen[3]

[1] Engility Corporation, U.S. Air Force Research Laboratory, Wright Patterson Air Force Base, OH 45433, U.S.A.; [2] Wright State University, Dayton, OH 45435, U.S.A.; [3] U.S. Air Force Research Laboratory, Wright Patterson Air Force Base, OH 45433, U.S.A.; [4] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, U.S.A.;

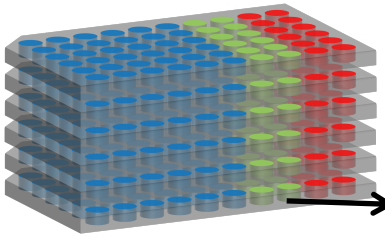
Motivation and Overview

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.

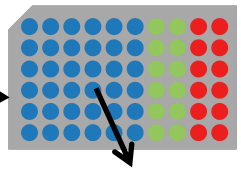
- The INSIGHTS project discovers chemical or genetic interrogations that are beneficial for human performance, medical intelligence, or therapeutic application [1].
- Envisioned use cases are force protection of U.S. warfighters from natural exposure or biological attack.
- Rapid screening of a large number of countermeasures decreases the time-to-decision from years to months or even weeks.
- Based on high-throughput, high-content biological screening.
- Individual experiments are done on wells of an experimental plate.
- After incubation, each well is observed at multiple sites, and resulting images are submitted to the pipeline for featurization that generates 11,000+ features per cell, feature reduction to select those most useful features, and well scoring to identify interrogations for further investigation.

High-Throughput High-Content Cellular Analysis

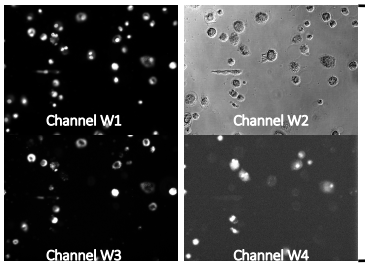
An experimental batch is comprised of multiple experimental plates. Each plate within the batch is subjected to the same procedure.

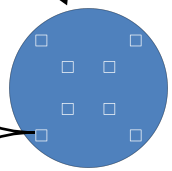


Experimental plates have test/experiment wells (blue), uninfected negative control wells (green), and infected positive control wells (red). A typical plate has 384 wells.



Each site is imaged at multiple frequencies and/or phase contrasts, referred to as channels. The captured images are in 16-bit grayscale. The images below have been scaled for improved visibility.





Each well is imaged at eight sites.

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.

Analytics Tools

The biologist has additional flexibility to do post analysis to facilitate verification and validation of results.

- One example is traditional dose response curve plots from downloaded Kolmogorov–Smirnov reports (Fig. 3) with the first 25 of 32 plots displayed for a plate. Dimple, based on Data Driven Documents (D3), was used for the visualization, and NW.js was used to prototype the web page as a desktop application.
- Provides common analysis for comparison of individual treatments or responses.
- A D3 visualization provides a more comprehensive comparison of features by logical grouping using parallel coordinates plots with controls to determine which treatments cause responses that are dissimilar to uninfected cells (Fig. 4) [2].
- Provides both familiar and more comprehensive tools.

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

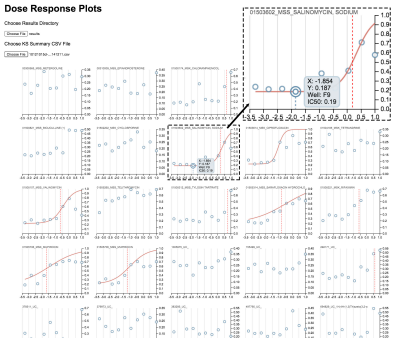


Fig. 3: Dose response curves with single point highlighted. Blue circles indicate actual points from the data and hovering over one of these shows the data values for that point, as shown in the center plot. The red curves indicate where dose response curves were found using the Hill equation, and a red dashed line marks the half maximal inhibitory concentration (IC50).

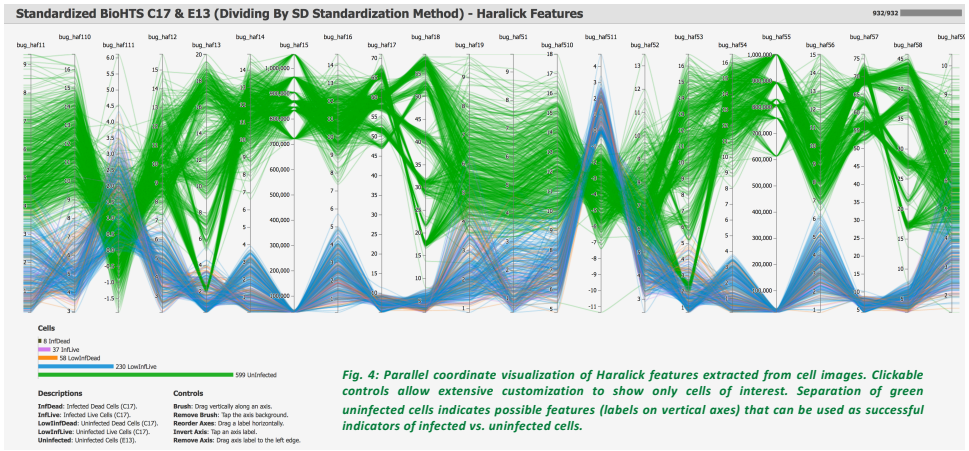


Fig. 4: Parallel coordinate visualization of Haralick features extracted from cell images. Clickable controls allow extensive customization to show only cells of interest. Separation of green uninfected cells indicates possible features (labels on vertical axes) that can be used as successful indicators of infected vs. uninfected cells.

Data Flow

- Each well is imaged at eight sites at frequencies optimized with commonly used biological fluorescent stains or when using non-fluorescent phase contrast imaging.
- Process yields four images per site and 3,000 sites produce 12,000 images per plate.
- Images are uploaded to a file system on the HPC (fig. 2).
- Batch metadata is linked via a project database.
- Biologists use a web portal to launch analysis jobs.
- HPC clients retrieve batch jobs following a "bag of tasks" paradigm.
- Other HPC clients written in C++, Python, Java, and C# execute the data analytics pipeline, updating both the project database and batch server.
- Although sites have around 50 cells, more than 300 cells are not uncommon. Processing 11,000 features for hundreds of plates, including additional analytics, requires HPC and produces millions of files.

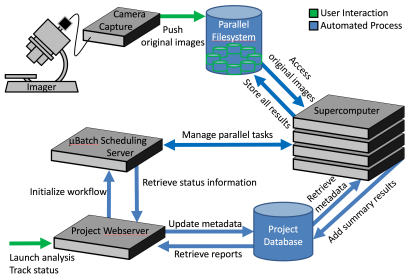


Fig. 2: Data movement through the system, including files, database accesses, job launches and user input. The 12,000+ images from a single plate take up 33 GB. Compressed feature data takes up 3 GB or more depending on the number of cells.

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.

Acknowledgements

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.

Scan the QR code for more information about this poster, including a video describing the processing.