Cytegeist: A framework for biological data analysis
Adam Treister¹, Anja Mirenska², Christian Hennig²
1) Gladstone Institutes, San Francisco, CA 2) Zellkraftwerk GmbH, Hannover, Germany

A New Cellular Metaphor

Complex experimental analysis is broken down into individual steps executed by individual pieces of software. The traditional process of chaining analytical steps together is known as pipelining. The plumbing metaphor is accepted and ubiquitous, but also reflects the limitations of an industrial age, where architectures are static and changes require intervention by skilled laborers to fundamentally reconstruct the pipeline. We propose here a new approach to combining tools in a cellular metaphor. We claim this architecture is more appropriate to modern (Big Data) programming problems, where small amounts of work are performed in parallel by independent work cells.

This is not a new computer architecture. It is merely a rebranding of existing technology to better fit the mental model of the biologist. The Service Oriented Architecture (SOA) approach to enterprise computing is well accepted as a scalable abstraction that can distribute problems across a sea of virtual machines. Well-defined service contracts are required to move data between services, analogous to pathway diagrams used to explain biological mechanisms.

As with cellular processes, there is a strict distinction between inside and outside the cell, as well as a nucleus of essential data that is translated into pieces of output. Internal pathways are used to:

- normalize data
- map terms and abbreviations
- visualize data sets
- track workflow protocols.

Intermediate results and internal regulatory mechanisms float in a protected environment inside the cell. At the same time a wide array of services can access external databases and ontologies, interact with robots, outsource extended calculations, share data, and publish reports. As select analyses mature, they can be cloned to scale up to the limits of available resources.

Cytegeist is fundamentally written in Java 1.8+. It is platform-agnostic, data type-independent, and open source. Through the CyRest interface, it inherits state transfer access to Python, Ruby and R scripting. Interface elements are coded in JavaFX for desktop and Javavis for web clients. Data interchange is generally pushed through JSON, XML and RDF channels, but specific formats are easy to plug into the extensible architecture.

Summary

We propose a framework of analysis tools that interact through a service API resembling an antigen binding model. The goal is to provide a context that captures a modern service architecture in a format comfortable to the biologist. We are looking for partners in all aspects of developing open methods of analysis and visualization.

Contact at: adam.treister@gladstone.ucsf.edu

References


The National Resource for Network Biology is an NIH Biomedical Technology Research Center (BTRC) supported by the National Institute of General Medical Sciences (NIGMS) under grant P41GM103504.