

Implementation of a Systems Biology Data Integration Platform

Sam Ansari^{1†}, Leandro Hermida², Svenja Diehl¹, Carine Poussin³, Alain Sewer³, Yang Xiang³, Filipe Bonjour³, Komanda Phanuz³, Sylvain Gubian³, Bruce O'Neel³, Julia Hoeng³ and Manuel Peitsch³

¹Philip Morris International R&D, Philip Morris Research Laboratories GmbH, Cologne, Germany

²Leandro Hermida Consulting, Rüschlikon, Switzerland

³Philip Morris International R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

†E-mail: sam.ansari@pmintl.com

Abstract

To enable and support a systems biology approach to research, one requires an underlying infrastructure to manage, integrate, and share high-throughput functional genomics data and workflows from data production through annotation, analysis, and knowledge acquisition. At its core, there should be a comprehensive data management and annotation system and data repository that fully support publicly established standards for storing and reporting high-throughput functional genomics investigations. Such a system will serve as the platform's central hub and it will integrate with data analysis, visualization and mining tools. It will also enable collaboration between internal teams, publishing of internal investigations and data to public repositories, and incorporation of public investigations and data into the platform for internal comparison and analysis.

Here, the implementation of a systems biology data integration and knowledge management platform to support experimental and computational workflows, examining *in vivo* and *in vitro* generated systems response profiles (gene expression, microRNA, comparative genomic hybridization, and reverse-phase protein array proteomics data) is reported. The platform utilizes open-source, freely available components where suitable, featuring caArray and caGrid [1] from the National Cancer Institute Biomedical Informatics Grid (NCI caBIG®) [2] software family as its core data management and annotation infrastructure. For data exchange, the community standard MAGE-TAB [3] format was used. caArray is integrated with GenePattern [4], an open-source bioinformatics workflow management system for integrative genomics, which is used for quality control, data analysis, and visualization purposes. caArray is also integrated with several commercial data analysis and biological pathway inferring systems. Gene-centric, cross-investigation data mining capabilities are provided by the BioMart and InterMine open-source data warehouse systems. Under development are several other modules to integrate the platform with existing laboratory information management systems, as well as additional features to contribute to the open-source caArray project.

Implementation

- Data production platforms: Affymetrix®, Zeptosens, Exiqon
- Functional genomics data types: Gene expression, microRNA expression, aCGH, SNPs, protein expression
- Data exchange: MAGE-TAB
- Data management and annotation: caArray
- LIMS with planned full integration with caArray
- Analysis tools: R/Bioconductor [5], GenePattern
- Data mining and integration of public databases: BioMart, InterMine
- Programmatic integration: caGrid API, caArray API
- Sharing and publishing: caGrid, MAGE-TAB export ArrayExpress [6], SOFT export GEO [7]
- Various integrated proprietary and commercial tools for analysis and pathway inferring

Objective and Requirements

The objective is to establish and support a systems biology data integration and analysis environment. The following requirements are defined for the core infrastructure:

- Open-source
- Standards compliant (MIAME)
- Easy-to-use and user friendly (e.g., wizards for experiment creation)
- Customizable user interfaces
- Scalable, enterprise-ready architecture
- Support for standard data exchange format import/export (e.g., MAGE-TAB, MAGE-ML) for publishing to public repositories and incorporating public data
- Comprehensive programmatic API for data access and integration with other systems
- Data analysis and visualization tool integration
- Under active development and support
- Comprehensive documentation

Selection of Major Components

After evaluation of various open source systems biology data management systems, caArray, along with the caBIG® infrastructure and integrated tools, was determined to best meet these requirements and will serve as the core of the systems biology data integration platform. caArray is an open-source, web and programmatically accessible array data management system. It is developed and currently supported by the U.S. National Cancer Institute as part of caBIG®, an initiative to build an information and bioinformatics framework to enable collaborative and integrative approaches to biomedical research. caArray provides full support for MAGE-TAB, curation capabilities as well support for ontologies and controlled vocabularies to ensure that the functional genomics data stored in its repository are of high quality and abide by community standards for data sharing and exchange.

References

- [1] Seltz *et al.*, *Bioinformatics* 22:1960-6, 2006
- [2] Osters *et al.*, *J Am Med Inform Assoc.* 15:138-49, 2008
- [3] Rayner *et al.*, *BMC Bioinformatics* 7:489, 2006
- [4] Reich *et al.*, *Nat Genet.* 38:500-1, 2006
- [5] Gentleman *et al.*, *Genome Biol.* 5:R80, 2004
- [6] Brazma *et al.*, *Nucleic Acids Res.* 31:68-71, 2003
- [7] Edgar *et al.*, *Nucleic Acids Res.* 30:207-10, 2002

Abbreviations

aCGH: Array comparative genomic hybridization
API: Application Programming Interface
caArray: Array data management system
caGrid: Network architecture providing the basis for connectivity of caBIG® tools
ChIP-Chip: chromatin immunoprecipitation with microarray technology
LIMS: Laboratory Information Management System
MAGE-TAB: MicroArray Gene Expression Tabulator
MIAME: Minimum Information About a Microarray Experiment
QC: Quality Control
RPA: Reverse Phase Protein Array by Zeptosens
SNP: Single-nucleotide polymorphism

