



## Message from NCICB

We welcome the release of calIntegrator as a significant contribution to translational research. This innovative informatics platform supports the retrieval, aggregation, analysis and sharing of data from a variety of heterogeneous operational repositories that house microarray, genomic, tissue array, imaging and clinical data. It provides a user friendly tool for clinicians and scientists to conduct translational analysis of study specific data. calIntegrator is also an important contribution to the caBIG™ initiative, enhancing our own translational research activities that will emerge from the Translational Informatics RFP process currently underway.

## In the Spotlight

### *calIntegrator*

By Subha Madhavan, Ph.D.

*In this edition of the caBIG™ Program Update, we spotlight calIntegrator, an informatics tool designed to integrate clinical trials and high throughput molecular analysis to support translational research and the transition to tailored therapy. calIntegrator was developed by the National Cancer Institute, Center for Bioinformatics (NCICB), in collaboration with the Neuro-Oncology branch (NOB) of the Center for Cancer Research (CCR), the National Institutes of Neurological Disorders and Stroke (NINDS) and the Carol Franc Buck Breast Care Center at the University of California, San Francisco (UCSF). calIntegrator reflects the caBIG™ initiative's commitment to provide modular and interoperable tools and interfaces that provide for integration between various biomedical data generated via cancer clinical trials. This tool will ultimately enable translational and integrative research by providing for the integration of clinical and basic research data. calIntegrator version 1.0 was released in January 2006 and is available at <http://ncicb.nci.nih.gov/download/>. It is expected to be submitted for review for caBIG™ Silver-compatibility in February, 2006, as part of the release of version 1.0.*

Progress in finding better therapies for cancer treatment has been hampered by the lack of opportunity to integrate biomedical data from disparate sources to enable translation of critical information from bench to bedside and back. Hence, a critical factor in the advancement of biomedical research and delivery is the ease with which data can be integrated, redistributed and analyzed both within and across functional domains. The NCICB, in collaboration with the NOB of the NCI CCR, the NINDS and the Carol Franc Buck Breast Care Center at the UCSF, has developed a novel translational informatics application called calIntegrator that allows physicians, researchers, and bioinformaticians to access and analyze clinical and experimental data across multiple clinical studies (*article continued on page 2*).

caBIG.nci.nih.gov

Contact:

Mark Adams, PhD

adamsm@mail.nih.gov



# calIntegrator

< from page 1

The calIntegrator framework provides a mechanism for integrating and aggregating biomedical research data (as shown in Figure 1 below) and provides access to a variety of data types (e.g. microarray-based gene expression, immunohistochemistry, Single Nucleotide Polymorphisms (SNPs), clinical trials data etc.) in a cohesive fashion. This knowledge framework offers a paradigm for rapid sharing of information and accelerates the process of analyzing results from various biomedical studies with the ultimate goal of rapidly changing routine patient care.

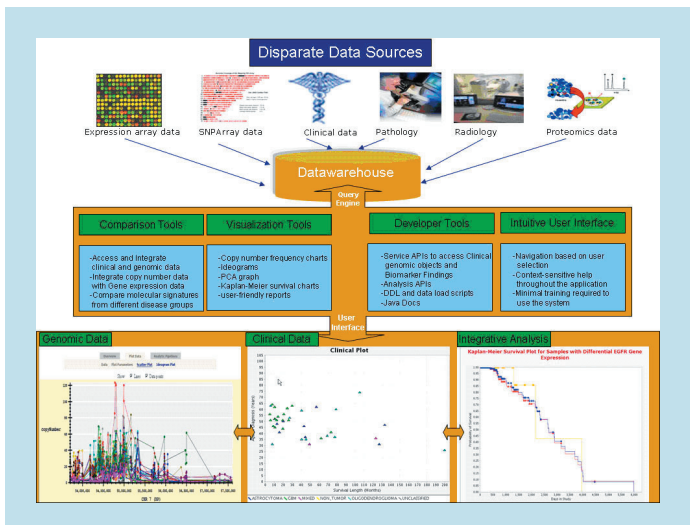


Figure 1: Data Integration via calIntegrator application

The goal for the calIntegrator platform is to interface with NCICB's transactional databases, such as caArray and C3D to create a stand-alone data warehouse that can be accessed by physicians and researchers. Depending on the particular user's role, he or she is permitted access to specific sets of the study data. A series of intuitive tools enable the user to easily analyze and interact with the integrated data to achieve greater insight into molecular signatures that characterize each tumor and correlate these signatures with clinical outcome.

## Key Translational Application Features

The overall goal of the calIntegrator project is to provide a framework with the infrastructural components needed to develop enterprise level translational applications. One such reference implementation at NCICB is REMBRANDT (Repository of Molecular BRAin Neoplasia DaTa) – <http://rembrandt.nci.nih.gov>. REMBRANDT is a powerful and intuitive informatics system designed to integrate genetic and clinical information from brain tumor clinical trials for improved research, disease diagnosis, and treatment. Figures 2-4 illustrate the key features in REMBRANDT implemented using the calIntegrator framework.

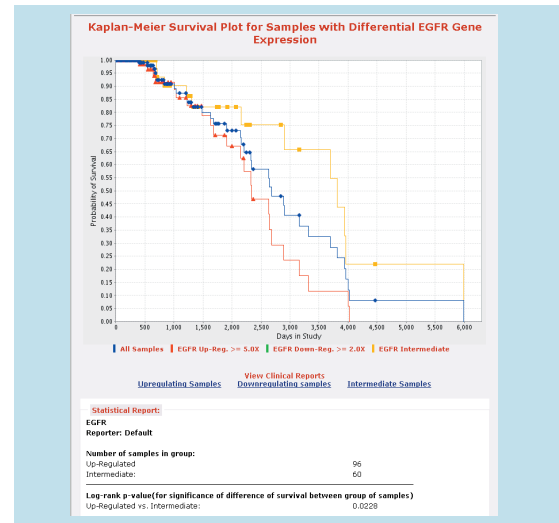


Figure 2: Kaplan-Meier survival plot based on gene expression data integrate gene expression data with clinical outcome

Users can query gene expression and graph changes in survival rate at each time point on the study. Kaplan-Meier (K-M) estimates are calculated based on the last follow-up time and the censor status (0=alive, 1=dead) from the samples of interest. The Kaplan-Meier estimates are then plotted against the survival time. Users can dynamically modify the fold change (up and down regulation) thresholds and redraw the plot. A log-rank p-value is provided as an indication of significance of the difference in survival between any two groups of samples segregated based on gene expression of the gene of interest.

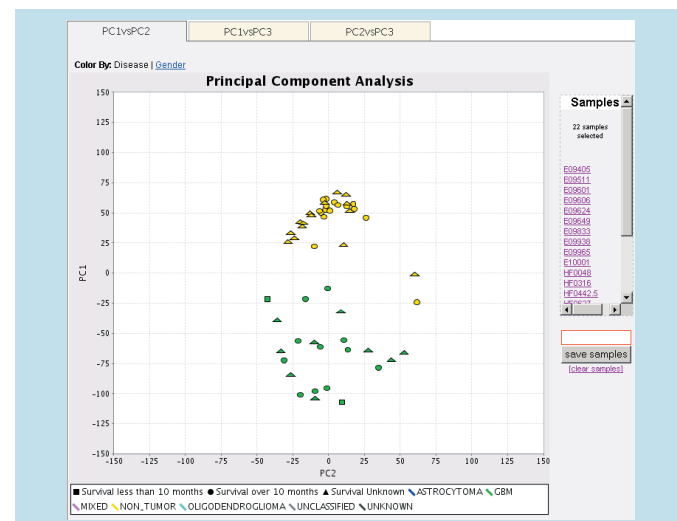


Figure 3: Perform higher-order statistical analysis on genomic and clinical datasets

This figure shows an example Principal Component Analysis report from the REMBRANDT application. This two-dimensional graph plots the various principal components from the gene expression PCA analysis. Various analysis options are provided to the user to select from an array of gene/reporter filtering and sample selection settings.

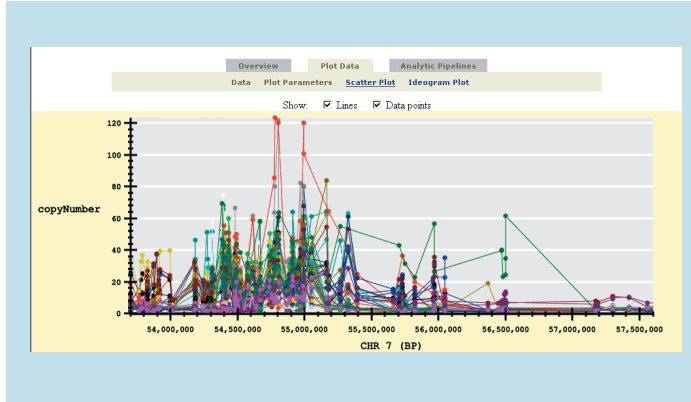


Figure 4: Plot copy number data from patient DNA samples against physical genomic location

Scatter plots (shown above) display measured copy number against physical genome location in an application called webGenome, which has been seamlessly integrated with calIntegrator. These plots are context-sensitive to the copy number reports generated from the copy number queries in the calIntegrator application. You can view data at arbitrary resolutions from the entire genome on down.

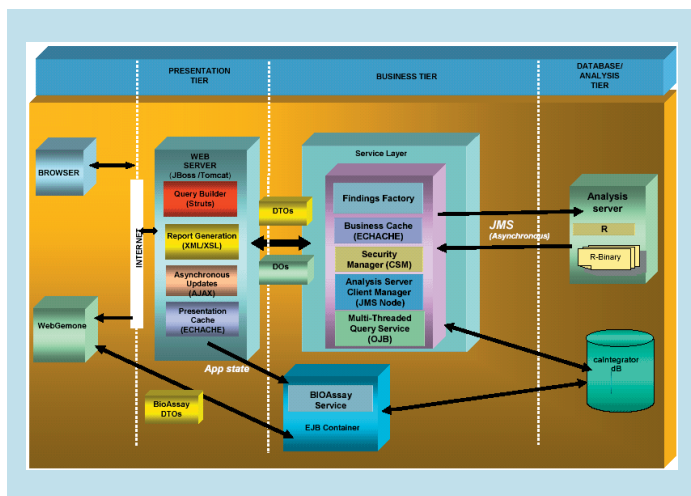


Figure 5: calIntegrator architecture

The calIntegrator application framework comprises an *n*-tier service oriented architecture that allows pluggable web-based graphical user interfaces, a business object layer, server components that process the queries and result sets, a data access layer and a robust data warehouse.

Following are some of the high-level features of the calIntegrator framework, which are reflected in Figure 5:

- ❖ A common set of interfaces (APIs) and specification objects that define the clinical genomic analysis services. In other words, they act as templates for the calIntegrator-based translational application(s), which will extend and implement these interfaces and specification objects. The application's user interface communicates with its calIntegrator-based middle-tier services via domain as well as business objects.
- ❖ A generic, real-time analytical service that currently supports class comparison analysis, principle component analysis and hierarchical analysis. It is designed to easily incorporate other types of analysis in the future, and scale to provide performance.
- ❖ The calIntegrator hybrid data system consists of a star schema database which contains the clinical and annotation data as dimensions, pre-calculated gene expression copy number data as facts, and Common Security Model (CSM) tables for user provisioning data. For performance reasons, normalized gene expression data used by the real time analysis module is stored as R-binary files.
- ❖ A generic interface to allow visualization of genomic data (copy number scatter plot and ideogram plots) via the WebGenome application.

## Analytical Server

The Analytical Server provides an on-the-fly computational analysis capability for calIntegrator-based applications. The Analytical Server communicates asynchronously with the calIntegrator's middle-tier via the Java Messaging Service (JMS). The Analytical Server utilizes the Java 1.5 ThreadPoolExecutor.

JMS allows calIntegrator to abstract the statistical packages being utilized for the heavy computational tasks. The current release utilizes R (a free software environment for statistical computing and graphics) to implement the statistical methods. The Rserve package (see: <http://stats.math.uni-augsburg.de/Rserve/>) is used to interface the R system with Java. Rserve provides Java classes to execute R commands and to retrieve results as Java objects. The overall architecture of the Analytical Server allows the user to plug in any other statistical package such as SAS (<http://www.sas.com/index.html>) as long as it exposes an API.

## WebGenome Interface

One of the main requirements of caIntegrator is to help researchers better analyze and visualize the results in a user-friendly manner. Since webGenome (aka webCGH) was already plotting some of these graphs, an interface between caIntegrator and webGenome was created and extended WebGenome's capabilities with new plots and functionality. WebGenome now exposes a set of interfaces to help populate these plots. caIntegrator framework implements these interfaces as Enterprise Java Beans (EJB). caIntegrator remotely invokes the WebGenome application via the Java Naming and Directory Interface (JNDI).

## Database Design

caIntegrator employs basic star schema with modification for the study data warehouse design that supports the integration of clinical and genomic data. It is a generic, query optimized schema that contains fact tables such as "Differential\_Gene\_Expression\_Fact" and "Genomic\_Abnormality\_Fact", etc. Look-up entities such as Genes, Biosample, and Disease type make up the dimensions in the schema. This schema provides a highly de-normalized view of the data and a data neutral framework from which queries can be executed with quick retrieval time.

## caIntegrator Resources

### caIntegrator Information Site:

<http://caIntegrator.nci.nih.gov>

### Rembrandt Application:

<http://rembrandt-db.nci.nih.gov>

### Rembrandt Informational Site:

<http://rembrandt.nci.nih.gov>

### caIntegrator Clinical Genomics Object Model:

<http://caIntegrator.nci.nih.gov/content/caIntegratorIfs/caIntegratorEA1.0docs/index.htm>

### caIntegrator Clinical Genomics Data Model:

[http://caIntegrator.nci.nih.gov/developers/images/db\\_model2.jpg](http://caIntegrator.nci.nih.gov/developers/images/db_model2.jpg)

## Contact Information and Resources

For further information on caIntegrator, please contact :

Subha Madhavan, Ph.D.

Biomedical Informatics Specialist, NCI Center for Bioinformatics

6116 Executive Blvd. Suite 403

Rockville, Maryland 20852

(301) 451-2882

[madhavas@mail.nih.gov](mailto:madhavas@mail.nih.gov)

<http://ncicb.nci.nih.gov>

## caBIG™ Program Highlights

Answers to written questions submitted for the Translational Informatics RFP (issued 12/20/05) are now available and can be downloaded from:

[http://207.36.232.130/rfp\\_management/](http://207.36.232.130/rfp_management/)

As a reminder, proposals in response to this RFP must be received no later than 5:00pm ET on February 7, 2006, and must be completed in accordance with the Proposal Instructions.

## Featured Meetings and Announcements

### caBIG™ 2006 Annual Meeting

Call for Poster Board & Technology Demonstration Abstracts and Exhibitor Applications is now OPEN!!!

The National Cancer Institute Cancer Bioinformatics (NCICB) welcomes submission of abstracts for posters and technology demonstrations at the caBIG™ 2006 Annual Meeting that reflect contributions to the caBIG™ initiative. Preference will be given to projects that reflect collaboration across organizations and disciplines. The deadline for submitting poster board and technology demonstration abstracts is Monday, February 6, 2006.

In addition the 2006 Annual Meeting Exhibit Hall will feature organizations that are collaborating with the caBIG™ initiative or whose products are (or will become) caBIG™ compatible. Please visit:

[https://cabig.nci.nih.gov/guidelines\\_documentation](https://cabig.nci.nih.gov/guidelines_documentation)

to learn about caBIG™ compatibility. The exhibitor registration form is due Friday, February 28, 2006. For more information and to access the forms please visit:

[https://cabig.nci.nih.gov/2006\\_Annual\\_Meeting](https://cabig.nci.nih.gov/2006_Annual_Meeting)

If you have any question please feel free to contact us at [caBIG2006@esi-dc.com](mailto:caBIG2006@esi-dc.com).

# Featured Meetings and Announcements

< from page 4

## Upcoming caBIG™ Face-to-Face Meetings

### *Joint Architecture & VCDE Workspaces*

- ❖ February 1-2, 2006
- ❖ NIH Campus, Bethesda, MD
- ❖ For more information, please contact Mike Keller at [keller\\_michael@bah.com](mailto:keller_michael@bah.com) or 301-998-6846.

### *Data Sharing & Intellectual Capital Workspace*

- ❖ March 6-7, 2006
- ❖ Fox Chase Cancer Center, Philadelphia, PA.
- ❖ For more information, please contact Brian Bullock at [bullock\\_brian@bah.com](mailto:bullock_brian@bah.com) or 703-289-5944.

### *Strategic Planning Workspace*

- ❖ March 8-9, 2006
- ❖ University of Pennsylvania, Philadelphia, PA
- ❖ For more information, please contact Brooke Hatcher at [hatcher\\_brooke@bah.com](mailto:hatcher_brooke@bah.com) or 301-998- 8774.

## Other caBIG™ Meetings and Activities

To find out about other caBIG™ meetings or to become part of a Workspace or Working Group, please visit the caBIG™ Web site at: <https://caBIG.nci.nih.gov>.

---

