caGrid 1.0 : A Major Milestone for the cancer Biomedical Informatics Grid™

The cancer Biomedical Informatics Grid (caBIG^{TM}) is creating a national-scale network to leverage the combined strengths and expertise of research laboratories, cancer centers, and investigator projects to accelerate the development of effective patient therapies for cancer. To accomplish this goal, the caBIG™ program has developed common applications, tools, information standards, and data and analytical resources, and a Grid software infrastructure to link applications and resources and facilitate sharing of information within established guidelines and policies.

The release of **caGrid version 1.0** represents a major milestone in the caBIG™ program. Driven primarily by scientific use cases from the cancer research community, caGrid implements the core Grid architecture of caBIG™. It provides the technology that enables collaborating institutions to share information and analytical resources efficiently and securely, and that allows investigators easily to contribute to and leverage the resources of a national-scale, multi-institutional environment. The caGrid effort supports and enhances several related research activities: 1) *discovery* through precisely targeted searches that return attributes and values from heterogeneous collections of data sources; 2) *coordinated research* through controlled access to, and sharing of, data and analytical resources among collaborating institutions; and 3) *focused large dataset analysis* though controlled access to distributed analytical services.

Technical Architecture

The caGrid infrastructure builds on the Model Driven Architecture (MDA) paradigm and is implemented as a service-oriented architecture on top of Web Services Resource Framework (WSRF) standards. All resources in caGrid are implemented and accessed as WSRF-compliant Web services. A distinguishing feature of caGrid is the emphasis on syntactic and semantic interoperability across heterogeneous resources through use of common data elements, published information models, controlled vocabularies, and well-defined application programming interfaces. Following the metadata- and model-driven caGrid approach, each service advertises itself using standard caGrid service metadata. The rich semantic information in metadata is used to discover resources in the environment. caGrid services represent an *object-oriented view* of resources to the environment and are *strongly typed*. That is, client APIs and service interfaces are object-oriented and operate on well-defined and curated data types based on common data elements and controlled vocabularies.

caGrid 1.0 is layered on the Globus Toolkit (developed by the Globus Alliance) and is based on several existing middleware components including the NCICB caCore infrastructure and the Ohio State Mobius middleware. The caGrid infrastructure contains many innovative features augmenting these core components. These features include innovative metadata management services, along with workflow management, and fast, secure bulk data transport services for applications such as radiology. The caGrid distribution also includes a robust security infrastructure, called GAARDS which extends traditional grid security to provide enterprise security services. caGrid also provides an interactive development environment (IDE) tool called Introduce, which greatly simplifies the construction of new services. Several of these core components, including both GAARDS and Introduce, have been contributed to Globus to enable broad dissemination.

Community Development

In keeping with the overall open and community-based development strategy of caBIG™, the creation of caGrid was itself an act of effective collaboration. caGrid was conceived by the National Cancer Institute for Bioinformatics (NCICB); design and development have been carried out by the caGrid 1.0 team, which includes:

- Ohio State University/OSU Comprehensive Cancer Center - Biomedical Informatics Department
- University of Chicago/Argonne National Laboratory
- Duke Comprehensive Cancer Center
- ScenPro, Inc
- SemanticBits, LLC
- Science Application International Corporation (SAIC)
- Booz Allen Hamilton
- National Cancer Institute Center for Bioinformatics (NCICB)

Each caGrid participant is responsible for developing one or more key components of the caGrid infrastructure. In addition to core developer team responsibilities, the OSU Biomedical Informatics Department team has provided overall technical leadership as the lead developer site, providing technical oversight in close coordination with the caBIG™/caGrid leadership on the overall architecture to ensure delivery of a high-quality, integrated system. The Booz Allen Hamilton team has been responsible for the overall program management of the caGrid development effort. The NCICB team leads have provided coordination, oversight and general directions to ensure the caGrid effort would meet the overarching goals of the caBIG™ program.

End-User Applications

Seven reference implementations have been developed to accompany the caGrid 1.0 release. The reference implementations make use of caGrid 1.0 infrastructure as well as expose caGrid 1.0 services. The first three reference implements to be grid-enabled are caTRIP, GeneConnect, and GridIMAGE. caTRIP provides a metadata-driven graphical interface to pose translational questions and query caGrid data services. It exposes four data services, which include tumor registry, tissue banking, pathology, and SNP data sources. GeneConnect is a mapping service that facilitates caBIG interoperability by interlinking VCDE workspace-approved genomic identifiers. It exposes a caGrid data service for searching the pair wise connections and all-to-all relationships in the genomic identifier space. GridIMAGE allows radiologists to review geographically distributed cancer images stored in DICOM PACS or in caGrid data services. This effort is in collaboration with the QARC, CALGB, ACRIN, ATC, RTOG clinical cooperative groups, and it provides functionality for retrieving, analyzing, and storing image data.

The remaining four reference implementations, caArray, caBioconductor, GenePattern, and geWorkbench, specifically target gene expression data. caArray is a microarray database with open interfaces, strong security, and a user interface that is designed to make MIAME 1.1 level annotations as easy as possible. It exposes a caGrid data service that leverages the MAGE-OM open standard. The caBioconductor module allows R package developers to expose the functionality of their package as analytic services on caGrid. They expose functions for normalization of gene expression data derived from the Affymetrix platform using the caAffy package, one of the many open-source software components found in the Bioconductor project. GenePattern is a flexible analysis platform developed to support multidisciplinary biomedical research, providing an environment for rapid development and deployment of new analytic techniques. In caGrid, GenePattern exposes an analytical service for analyzing gene expression data. geWorkbench is an open source bioinformatics platform written in Java that makes sophisticated tools for data management, analysis, and visualization available to the community in a convenient fashion.

For more information on caBIG: please visit https://cabig.nci.nih.gov/
For more information on caGrid: please visit
https://cabig.nci.nih.gov/workspaces/Architecture/caGrid
caGrid Overview published in Bioinformatics (Open Access) please see:
http://bioinformatics.oxfordjournals.org/cgi/content/full/22/15/1910
For more information on Globus: please see www.globus.org