CHAPTER 1

ABOUT caARRAY

This chapter describes briefly the microarray process as well as the caArray software.

Topics in this chapter include:

caArray Overview

I've taken most of this from the Vision document. While I think it's good, I'd like to condense the text in this chapter. I doubt that readers of this user's guide will take the time to read this in its entirety the way it is.

If either of you can edit it, that would be great. With track changes on, you can just edit the text in the Vision doc (cause it's in Word :) and I'll make the appropriate changes here.

If neither of you have the time to do this, I could work it over. It will probably be easier for one of you because you know what is important enough to leave in and what can be excised.

caArray is an open-source, user-driven, role-based, web and programmatically accessible data management system. caArray guides the annotation and exchange of array data using a federated model of local installations whose results are shareable across the cancer Biomedical Informatics Grid (caBIG™) workspaces/Architecture/ caGrid (https://cabig.nci.nih.gov/). Identified primarily as a data service on the Grid, caArray furthers translational cancer research through acquisition, dissemination and aggregation of semantically interoperable array data to support subsequent analysis by

tools and services on and off the Grid. As array technology advances and matures, caArray will extend its logical library of assay management.

At the highest level, the following services are provided by caArray:

- Managing experiment annotation and array data upload, validation, and import
- Searching across experiments using a web browser
- Programmatic querying and retrieval of data over the grid or via a java API
- Managing of experiment security and access.

Relationship of caArray to caBIG

The National Cancer Institute (NCI) has launched the caBIG™ (cancer Biomedical Informatics Grid™) initiative to accelerate research discoveries and improve patient outcomes by linking researchers, physicians, and patients throughout the cancer community. caBIG™ serves as the cornerstone of NCI's biomedical informatics efforts to transform cancer research into a more collaborative, efficient, and effective endeavor. In 2007, caBIG™ has moved from the pilot phase to the enterprise and is looking to evolve its set of maturing assets.

As the scientific community begins to better understand cancer at the molecular levels and personalized medicine is implemented in cancer patient care, researchers and clinicians will require more rapid access to—and easier methods to analyze—the multiple types of information involved. To this end, the vision of caBIG™ is a full cycle of integrated cancer research, extending from bench to bedside, and back again.

caArray, an integral component of the caBIG array of tools, has a committed following, years of user feedback and contribution on which to build from, sufficient resources, and an expansive vision to support integrative cancer research. It

Expression profiling is now a standard tool set with which to interrogate biological systems. Parallel advances in computing and new array technology provide an opportunity for collaboration and discovery within the scientific community and across traditional boundaries to reach clinicians and ultimately patients. The insistence on open source development provides the community with the greatest opportunity to gain access to the tools they desperately need to execute their respective mission. caArray was initially developed with expression profiling in mind, using the caBIG Compatibility Guidelines, as well as the Microarray Gene Expression Data (MGED) society standards for microarray data. Compatibility with these standards and guidelines was and remains required. However, the ability to add new standards that are developing is also necessary to facilitate data exchange and analysis across domains. A number of analytical tools and services that connect to caArray are already available - including geWorkbench and GenePattern - that provide a variety of analysis, visualization and annotation functions for microarray data.

The primary goal of caArray is to further translational cancer research through acquisition, dissemination and aggregation of high quality array data to support subsequent analysis. Initially envisioned to be the de facto repository for cancer related expression data, other applications – primarily the Gene Expression Omnibus (GEO) – has assumed the role. However, the quality of the expression data and the ability to integrate the use of array technology into clinical research remains at issue. Further,

the opportunity for caArray to evolve to handle other types of array data will provide greater impetus for use among the Cancer Centers and their collaborators which will ultimately benefit the cancer community. Data from expression, SNP, methylation, and protein arrays are anticipated for inclusion in caArray. Also under consideration are the burgeoning platforms of RNAi and CHip-chip data. Tissue microarrays represent another array-based technique that require a significantly different business process in their creation but are being considered for inclusion in the repository.

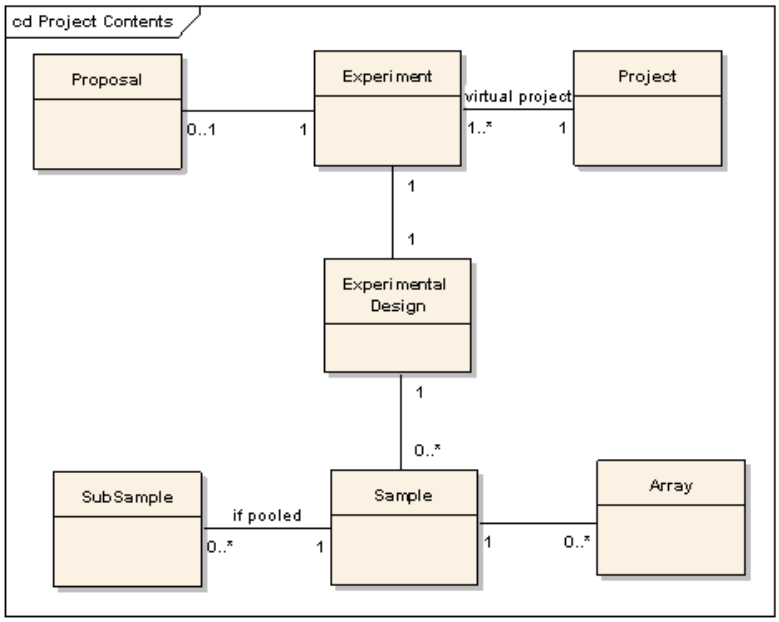
A significant challenge is to find common ground for the annotation, upload and extraction of these array data platforms to support meaningful analysis for cancer research. The need for logical and expedient approaches to storing, querying, retrieving and reporting on array–based data has increased over the last several years due to the contribution of the following factors:

- a significant increase in available array information (expression and others)
- decreases in the cost to generate array data
- improvements in the technology to generate this data
- advances in the approaches to analyze this data
- the need to increase the numbers of samples to enhance discovery and ensure the validity of results once found
- the need to ease the administration of the data and results by giving the community a single point of reference to perform the most essential tasks of array-based research.

The initial generation of caArray (1.0) and subsequent point releases (currently 1.5) have provided an interface for annotating scientifically significant meta-data along with the independent ability to upload and download data through an applet or accessing the MAGE-OM API over the grid service provided by caBIG. caArray 2.0 will improve upon this starting point by providing these features:

- Organizing the application around the natural workflow between investigators and the array labs that serve them
- Improve the user experience for storing and retrieving the data produced,
- Query the data through an easier to comprehend API
- Bridge the gap between the analysis tools in heavy use in the community today and the data they need to consume.

A simplified organizational structure for basic experiment relationships that fundamentally represent "annotation" versus the resulting array "data" is shown in Figure 1. This is an overall theme of the application.



caArray will look to perform intra-platform, intra-experiment queries for local adopters, via the NCICB instance and potentially across the grid collectively to increase the availability of quality data.

caArray will be built to scale with an open architecture and supportive documentation to allow for future enhancements, particularly with regard to interfacing with additional analysis tools, the ability to query across platforms, and poised to exploit web services and/or databases from other components of caBIG when available and prudent. The desire to create an extensible array system that is non-platform-specific and potentially customizable is a theme that should influence the building of caArray.

caBIG Microarray Data Analysis Tools

Several tools are currently able to analyze data stored in caArray from local installations.

- **geWorkbench** - http://wiki.c2b2.columbia.edu/workbench/index.php/Home

  Developed at the Columbia University, geWorkbench is an extendible and flexible desktop tool for microarray data analysis and visualization. The release 1.0 includes several data analysis and visualization functions, including hierarchical clustering, self organizing maps, color mosaic images and biological pathways.

- **Gene Pattern –** http://www.broad.mit.edu/cancer/software/genepattern/

  GenePattern, developed by the Broad Institute, is a powerful analysis workflow tool developed to support multidisciplinary genomic research programs and designed to encourage rapid integration of new techniques. puts sophisticated computational methods into the hands of the biomedical research community. A simple application interface gives a broad audience access to a growing

repository of analytic tools for genomic data, while an API supports computational biologists.

- **Bioconductor –** http://www.bioconductor.org/

  Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data. Bioconductor is primarily based on the R programming language, but contributions in any programming language can be accepted. The Bioconductor core team is based primarily at the Fred Hutchinson Cancer Research Center. Other members come from various United States and international institutions.

- **webGenome** – http://webgenome.nci.nih.gov/

  WebGenome is a web application for creating various types of graphical plots of microarray-based data, especially aCGH data. The system can be configured to read data from relational database systems or from in-memory data structures from client applications. Developer??

- **VISDA -** https://cabig.nci.nih.gov/tools/VISDA

  VISDA (VIsual Statistical Data Analyzer) is an analytical tool for cluster modeling, visualization, and discovery. Developer??

NOTE: To utilize the open source tools available with caArray, you can access the source code and application programming interfaces (APIs) on the NCICB download site at http://ncicb.nci.nih.gov/download/.

caArray Standards Is this still true?

As research and technology expand, it is of critical importance that established standards be used to ensure better and more accurate data collection and experiment results. Removing ambiguity by using standardized terminologies is one of the cornerstones of caArray. caArray is designed to support the international standard for microarray data, MAGE, and caBIG compatibility guidelines.

MAGE-ML — Microarray Gene Expression Markup Language is a standard data exchange format for microarray data. MAGE-ML uses eXtensible Markup Language (XML), which provides a consistent way to label data. MAGE-ML can be automatically derived from Microarray Gene Expression Object Model (MAGE-OM), which is developed and described in Unified Modeling Language (UML). A broad range of computer programs can process XML-labeled data, making it a valuable tool for both analysis and input. For more information about MAGE-ML, refer to the website: http://www.mged.org/Workgroups/MAGE/introduction.html.

MIAME 1.1 — Minimum Information About a Microarray Experiment describes the minimum information necessary to enable unambiguous interpretation of microarray experiments. Although details for particular experiments may differ, the MIAME objective is to define common elements among most experiments such as information on experiment and Array Design, samples, protocols and measurements. MIAME is not a formal specification, but a set of guidelines. For more information about MIAME 1.1, refer to the website: http://www.mged.org/miame

caBIG Silver-Level Compatibility — This term refers to the National Cancer Institute Bioinformatics (NCICB) Cancer Biomedical Informatics Grid (caBIG) Compatibility Guidelines. caBIG is a common, extensible informatics platform that integrates diverse data types and supports interoperable analytic tools. This platform allows research groups access to the extensible collection of emerging cancer research data, while supporting their individual research. For more information about caBIG, refer to the website: https://cabig.nci.nih.gov/. caBIG-Silver is a compatibility reference indicating that the software being used to submit and share cancer-related data meets or exceeds the Silver level of compliance. A description of the various caBIG guidelines and compliance levels such as bronze, silver and gold can be found at this website: https://cabig.nci.nih.gov/workspaces/VCDE/Documents/caBIG_Compatibility.pdf.

Security Structure Reword to reflect 2.0

caArray is designed to provide maximum flexibility in the collaborative sharing of microarray data, while allowing users and data managers to make choices regarding who can access the data submitted. Three levels of security and visibility are available in caArray:

- **Public** permissions or visibility gives anyone the ability to view the data, but with Read-Only access. The only person who can modify Public data is the owner of the data. Once your microarray experiment has been published, you may choose to make your experimental data public; anyone can view the data, but after publishing, ownership of the data is moved to the NCICB. Only the NCICB data curator can modify the data at that point.
- **Private** access or visibility restricts access to the data, allowing only the owner the ability to view the data.

  Example: if you are in the initial information-gathering stage of a microarray experiment, you may choose to make the information available to only yourself.

  Note: In caArray, if no visibility is selected when you are submitting or modifying an object, the default visibility is "Private".
- **Group-level/Consortium** access or visibility allows only specified individuals or groups, as specified by the owner to view the microarray data. Groups have "read-only" privileges with exceptions granted to persons assigned "curation" status; they are allowed "modify" privileges. Curation status must be requested through NCICB Application Support.

  Example: If you are working on a collaborative effort with colleagues in your lab as well as with investigators at a different research facility, you may choose to make your specified data available to a group/consortium; the consortium has read-only access to the referenced objects. Only you and others in the group specified as Curators can modify the data.

Visibility can be designated with multiple choices. For example, if you assign Public visibility as well as a specific group for an object, the object is available for viewing to everyone, but only the members of the group assigned curation privileges can modify the object. The advantage to this arrangement is that a user with appropriate permissions can search for and display all objects assigned specified curation privileges.

caArray Workflow

Important! After going through the use cases, I believe it would still be helpful to have a workflow section. I've left in some of the text. Could this be reworked to reflect 2.0, illustrating a suggested workflow? Perhaps just expand on the class diagram on page 8?

I would be willing to work on coming up with a workflow, but Brent, I'd like to talk it through with you first. Or if you come up with a workflow, I could draft the illustration.

Because there are dependencies between models of information in caArray that comprise an Experiment, it is important that you create objects in caArray in the order discussed and displayed in this section shown in *Figure 1.1* If you do not do so, as you create an Experiment, you may find yourself thwarted by lack of a previously defined data element, and required to abandon partially filled out screens to work backwards.

Follow the numbered steps to create the listed objects. Chapters in this caArray User's Guideonline help correspond to each of these headings:

As you define the details of all of these components of a microarray experiment, associations between objects are identified and relationships are established. Then an Experiment object is created, incorporating the associated objects that comprise the experiment. These components of an Experiment are then used to annotate microarray hybridization data and raw data files that are uploaded and parsed into the caArray database.

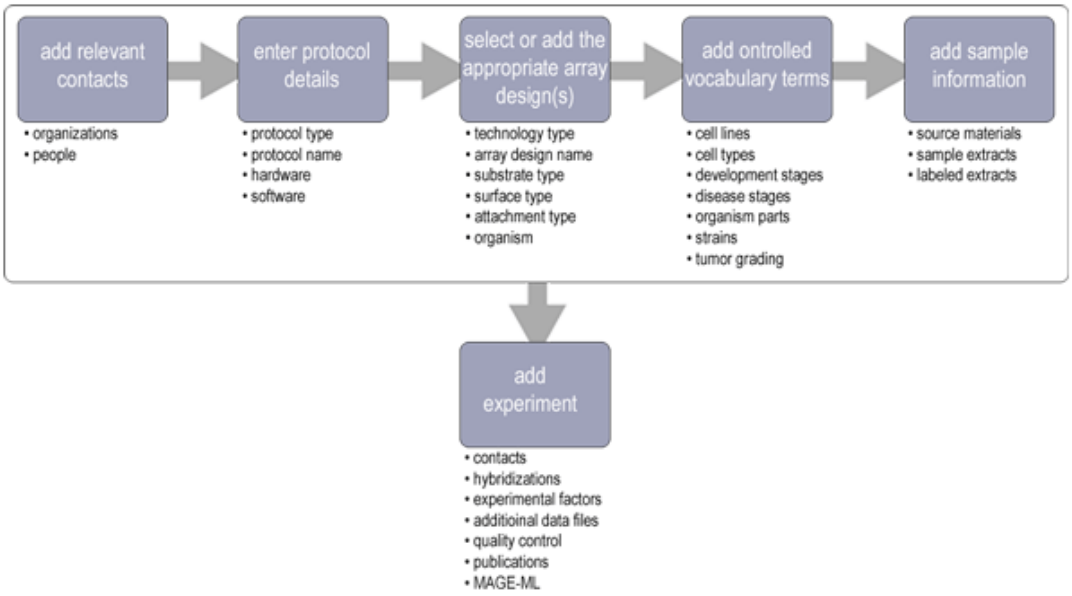


Figure 1.1 caArray Workflow. Each step in the workflow is described in its text and also serves as a hypertext link to a corresponding caArray viewer. Some steps in the workflow have dependencies on previous tasks. They are described, where appropriate, in this user's guide.