

DEB: a Data Entry and Browsing Tool for Entering and Linking Whole Genome Microarray Metadata from Multiple Data Sources

Arie Shoshani¹, Victor Havin¹, Vijaya Natarajan¹, Tony Martino², Jerilyn A. Timlin², Katherine Kang³, Ian Paulsen³, Brian Palenik⁴, Thomas Naughton⁵

¹Lawrence Berkeley National Laboratory (LBNL), ²Sandia National Laboratories (SNL), ³The Institute for Genomic Research (TIGR), ⁴Scripps Institution of Oceanography (SIO), ⁵Oak Ridge National Laboratory (ORNL)

The process of generating and analyzing microarray data for the *Synechococcus* Sp. WH8102 whole genome in the Sandia National Laboratories-Oak Ridge National Laboratory Genomics:GTL project, “Carbon Sequestration in *Synechococcus* Sp.: From Molecular Machines to Hierarchical Modeling,” involves three collaborators who each generate metadata about their operation as well as data files. The *Synechococcus* Sp. microbes are cultured at the Scripps Institution of Oceanography (SIO) in San Diego, California and then the sample pool is sent to The Institute for Genomics Research (TIGR) in Rockville, Maryland for microarray hybridization, 2-color scanning, and analysis. The scanned files and slides are then sent to Sandia National Laboratories (SNL) in Albuquerque, New Mexico for analysis and additional scanning with a hyperspectral imaging instrument. Each of the laboratories has an independent system for keeping track of metadata about their part of the operation and unfortunately these systems do not facilitate easy sharing of metadata details between institutions. This situation is typical of many biology projects, and it begs for a solution.

In this sub-project, we are developing a single system where such metadata can be easily collected and linked in an orderly fashion. The key need is to have an easy-to-use, intuitive system that integrates the metadata of all the related activities in the overall Genomics:GTL project. We developed a web-based Data Entry and Browsing (DEB) tool that can capture the metadata from the laboratories and store them in a database in a computer-searchable form. The system is built upon the Oracle database system. The design of the DEB tool is based on insight from the biologists on the project and contains features that a biologist will find intuitive and useful, such as a familiar laboratory notebook interface design. The notebook interface design shows the biologist a single “object” and its attributes as a page that can be “turned.” The interface provide links between the objects in a simple fashion. An example of such a web-based screen is shown in the figure below.

The unique capability of the DEB system is that it is *schema-driven*; that is, all the interfaces to support all of its features are generated automatically from the schema definition. Therefore, new metadata schemas can quickly be used to generate DEB interfaces as well as the underlying oracle database for them. This feature makes this tool immediately applicable to new and/or changing databases. This allowed us quickly to generate databases based on schemas designed by the biologists. Specifically, the scientists from the three sites have defined schemas for the Nucleotide Pool of microbes, for the Microarray Hybridization (based on the MIAME concepts), and for the Hyperspectral Imaging and analysis system. The design includes the ability to link these schemas and thus allow a researcher from any area of the Genomics:GTL project to

extract metadata from the various parts of the experiment. For example, the Microarray Hybridization schema has “probe_source” that links (points to) the “nucleotide_pool_id” in the Nucleotide Pool schema.

Data entry into the databases is done in two different modes: 1) on-line, web-based data entry and 2) automated data uploading from another database source. The on-line mode is used by the SIO and SNL for entering data about the Nucleotide Pool and the Hyperspectral Imaging, respectively. The automated data uploading is used by TIGR for entering the Microarray Hybridization metadata because they have their own well-developed, internal electronic database system. The automated metadata loading is performed by exporting the metadata into simple formatted files (similar to spreadsheet output format) and imported into the common database using schema-driven tools.

The main features of the DEB system are (see also “bubble” annotation in the figure below):

- It supports multiple interrelated object-classes, such as experiment, materials, nucleotide-pool, samples, arrays, etc.
- For each object-class, it displays a page that mimics a notebook, with pages that can be “turned” (i.e., selected by previous-next, or by number).
- Objects can be linked to each other by simple connectors, such as a “probe_source” object linked to its “nucleotide_pool”.
- Any file types (document, images, spreadsheet, etc.) can be uploaded to the system and related to the metadata.
- Pages of the metadata can be printed for entry into a physical notebook – a requirement that makes sure the information is physically recorded.
- Recording a new entry can be based on a previous entry, thus avoiding the re-entry of existing entries.
- Security features to protect the metadata can be controlled by users/groups; each experiment or related object can be assigned insert, read, write, and delete permission to other users/groups.
- A query feature is available for searching the metadata based on conditions on the attributes of the objects.
- The metadata can be browsed by traversing connected objects with a single click of a button, such as finding the probe used in a hybridization.

The image shows a screenshot of a web-based application titled "MicroArray Pipeline metadata" running in a Microsoft Internet Explorer browser. The interface includes a navigation menu on the left with categories like SID_Metadata, TIGR_Metadata, and Global_Items. The main content area displays a "Probe Source" form with fields for probe source id, sample name, issue, nucleotide pool link, and date creation. Callouts provide detailed information about the system's capabilities:

- For each object-class, it displays a page that mimics a notebook, with pages that can be "turned" (i.e. selected by previous-next, or by number)** (pointing to the top navigation buttons).
- Pages of the metadata can be printed for entry into a physical notebook – a requirement that makes sure the information is physically recorded** (pointing to the "Print" button).
- A query feature to search the metadata based on conditions on the attributes of the objects** (pointing to the "Build Query" button).
- Ability to select entries to view by individual users, group, or Query selection** (pointing to the "Show" dropdown menu).
- Objects can be linked to each other by simple connectors, such as a "probe_source" object linked to its "nucleotide_pool"** (pointing to the "nucleotide pool link" field).
- Use of controlled vocabularies prevents entry errors and promotes standard terminology** (pointing to the "Nucleotide Pool" dropdown).
- Security features to protect the metadata can be controlled by users and groups; authorization can be assigned read, write, and delete** (pointing to the "Set Permissions" button).

The importance of an easy-to-use system for capturing metadata in the Genomics:GTL project cannot be overlooked, especially as an ever-growing number of experiments are conducted and a large number of datasets are collected. The ability to quickly and automatically generate metadata systems from a schema description is essential for this evolving field with multiple sources of data gathered independently. Below is the feedback provided Dr. Jerilyn Timlin, SNL, who helped in the design and usage of the system and who is using the system on a regular basis.

“The Data Entry and Browsing (DEB) tool has been tremendously useful for our microarray subproject. The benefits of this tool are many, including:

- Metadata entry is simple and efficient. I can enter data directly from my lab notebook and the templates reduce the need for re-typing common entries.
- I can perform queries to locate data with characteristics I desire, such as all phosphate experiments or everything from a given batch of RNA.
- The links from the metadata to the actual files on the ORNL storage server allow efficient retrieval of files.
- There are links to all steps of the data-generation process. For example, if one researcher is performing an analysis and sees an anomaly, he/she can search the slide barcode number or sample name, directly view the analysis files original TIFF images and hyperspectral images to look for anomalies, and obtain all the pertinent microarray information conforming to the MIAME standards.

These capabilities are exciting to me as a data analyst, and I expect they would be exciting to anyone using the variety of data in Genomics:GTL pipeline. This access to detail provides researchers increased confidence in the data and easy accessibility.”

The main user from the TIGR team, Katherine Kang, added: “Another useful feature of the DEB tool is the automated data uploading to populate TIGR metadata to the DEB database. This really expedites the data-populating process and save so much time from our end.”

DEB is currently running on LBNL’s development server and at ORNL’s Genomics:GTL project operational server. While this system is designed for this project, its schema-driven architecture means that it can be applied to other Genomics:GTL program efforts.

Acknowledgement: This project is supported by the U.S. Department of Energy’s Genomes to Life Program under project “Carbon Sequestration in *Synechococcus* Sp: From Molecular Machines to Hierarchical Modeling” (<http://www.genomes-to-life.org>)