An XML message broker framework for exchange and integration of microarray data

Donny Tjandra1,∗, Stephen Wong2,3, Weimin Shen1, Brian Pulliam4, Elaine Yu5 and Laura Esserman1,5,6

1 UCSF Department of Radiology, 505 Parnassus Avenue, Box 0628, San Francisco, CA 94143-0628, USA, 2Harvard Center for Neurodegeneration and Repair (HCNR), Harvard Medical School, Boston, MA 02115, USA, 3Department of Radiology, Brigham & Women Hospital, Harvard Medical School, Boston, MA 02115, USA, 4Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143, USA, 5Carol Franc Buck Breast Care Center, University of California, San Francisco, CA 94115, USA and 6Department of Surgery, School of Medicine, University of California, San Francisco, CA 94143, USA

Received on January 6, 2003; revised on March 18, 2003; accepted on April 17, 2003

ABSTRACT
Motivation: Microarrays are an important research tool for the advancement of basic biological sciences. However this technology has yet to be integrated with clinical decision making. We have implemented an information framework based on the Microarray Gene Expression Markup Language (MAGE-ML) specification. We are using this framework to develop a testbed integrated database application to identify genomic and imaging markers for diagnosis of breast cancer.

Results: We developed extensible software architecture for retrieving data from different microarray databases using MAGE-ML and for combining microarray data with breast cancer image analysis and clinical data for correlation studies. The framework we developed will provide the necessary data integration to move microarray research from basic biological sciences to clinical applications.

Availability: Open source software will be available from SourceForge (http://sourceforge.net/projects/microsoap/).

Contact: donny.tjandra@radiology.ucsf.edu

INTRODUCTION
Microarray is a promising technology in the area of functional genomics. Researchers are able to use microarrays to study the expression patterns of several thousand genes in a single experiment. This technology is an important research tool for the basic biological sciences. The utilization of microarrays for clinical setting is also rapidly emerging. Clinical applications of microarrays, such as finding correlations between individual genetic variations and risk factors, identifying diagnostic markers, and predicting response to treatment, can potentially improve disease diagnosis and therapy strategies.

Currently, one of the unresolved issues is how to best represent and share information about microarray experiments.

In this paper, we describe an information framework based on Microarray and Gene Expression Markup Language (MAGE-ML) (Brazma et al., 2001; Spellman et al., 2002) to support the management and exchange of experimental microarray data. We present the system architecture and methodology for the MAGE-ML software framework, including the middleware architecture, user interface, and data integration methods. Finally, we conclude that the software tools we developed will greatly improve the transition of microarray data from basic biological research to clinical applications.

IMPLEMENTATION
MAGE-ML server
The primary function of the MAGE-ML server is to obtain and encode data from proprietary microarray databases into MAGE-ML messages. The server handles the mapping from the proprietary database schema to the MAGE-ML schema. These operations are published using the Web Service Description Language (WSDL) and are accessible through the Simple Object Access Protocol (SOAP). The MAGE-ML Document Type Definition (DTD) is used to verify the resulting MAGE-ML message, while the MAGE-ML client is responsible for the actual display of the MAGE-ML message. In the future, the middleware database will also be designed to include clinical data in addition to microarrays, such as medical images, labs and reports. This will be accomplished using data warehousing technology similar to a system previously developed for another disease application (detailed in Wong et al., 2002).

∗To whom correspondence should be addressed.
MAGE-ML client

A basic MAGE-ML client (Fig. 1) was developed to manage the overall workflow of the microarray data exchange and to browse the data. The client communicates with the MAGE-ML server using a series of XML messages wrapped in the SOAP specification. A listing of microarray labs with MAGE-ML servers is stored in the client’s database. This list includes the given name of the lab and the corresponding Uniform Resource Locator (URL) of the MAGE-ML server. It is presented as a drop-down box. When an entry from this list is selected, a login prompt is shown. After a successful login, a list of arrays available at the lab is shown. The user can select an array from this list to get the complete experiment details. A text area is provided on the MAGE-ML client for users to directly edit the MAGE-ML message. Researchers can then upload the updated MAGE-ML message to any other MAGE-ML servers on the list. Uploading data can be used to create a new experiment, or to upload arrays associated with an experiment. The MAGE-ML client was implemented in Java with the Swing toolkit to allow cross-platform execution on several platforms. This MAGE-ML client is not intended to be a full functional microarray viewer and analysis tool. Several tools exist for full-featured microarray viewing and analysis such as the MultiExperiment Viewer from the Institute for Genomic Research (TIGR). It should be noted that several microarray analysis tools, including TIGR MultiExperiment Viewer, currently support (or will support) viewing of data in MAGE-ML format. The MAGE-ML client is meant to be a tool for on-line retrieving or uploading array data from multiple laboratories in remote locations.

DISCUSSION

Effective data sharing and integration are important steps for the successful application of microarray technology to integrative clinical research. The XML based software tools based on recognized data exchange standards such as DICOM (Digital Imaging and Communication in Medicine; http://medical.nema.org/) for medical images and HL7 (Health Level 7; http://www.hl7.org/) for clinical and administrative text are commercially available. The MAGE-ML broker framework we developed will provide another component to allow researchers and clinicians to access microarray profile data from individual lab databases via standard XML/SOAP protocol. We are integrating the MAGE framework with existing XML based tools of DICOM and HL7 to develop an integrated database system for correlative studies in breast cancer management.

REFERENCES

