

## **INTEGRATED BIOMEDICAL DATABASES: MOLECULAR BIOLOGY INTEGRATION, MICROARRAY, AND NATURAL LANGUAGE PROCESSING**

Phillips, Jeremy L.<sup>1</sup>; Santos, Carlos F.<sup>2</sup>; Gao, Jian; States, David J.<sup>2</sup>

<sup>1</sup>College of Engineering, University of Michigan, Ann Arbor, MI;

<sup>2</sup>Department of Human Genetics, University of Michigan Medical School, University of Michigan, Ann Arbor, MI

**Keywords:** sequence annotation, BioNLP, relational databases

The biomedical sciences are diverse and generate many different data types ranging from primary gene sequence to transcript expression, proteomics identifications and text.

Publicly available biological data resources are expanding rapidly, but these resources are often weakly interconnected. For instance, a data set derived from a microarray experiment is unlikely to contain links to relevant contextual information in the biomedical literature. In addition, a single type of biological data is often distributed over several heterogeneous databases. Moreover, because web sites are most often the primary gateway into biological data repositories, users of these repositories are unable to extract data with the same flexibility that is available in a standard query language such as SQL. Further, different data sources may overlap and often are incomplete or even contradictory. As a partial solution to these problems in three specific domains, we have created three closely interconnected in-house biological relational databases: The Molecular Biology Integration Database (MBI), the NCIBI Microarray Repository, and a Biological Natural Language Processing Database (BioNLP).

MBI is a repository of human, mouse, and chimpanzee sequences and associated annotations from several public sequence databases. By joining sequence annotations from these disparate sources into single, unique sequence records, MBI serves as a crossing point between both public and local data sources. The Microarray Repository is a collection of raw and normalized data from microarray experiments available in the public domain and from private contributors, covering several technologies and platforms. BioNLP contains a collection of biological named entities gleaned from the biomedical literature along with relationships between these entities, as an endpoint for the NCIBI natural language processing pipeline. These databases are all built using standard relational database technology (Microsoft SQL Server), allowing users to easily construct queries to gather data from all three databases. In addition, the databases are built to maximize query simplicity, allowing for easy and flexible access via standard SQL.

Using MBI, the Microarray Repository, and BioNLP, NCIBI investigators will be able to easily cross reference biological named entities and expression data with biological sequences and sequence annotation.

This research was supported by the National Institutes of Health Grant # U54-DA021519, National Center for Integrative Biomedical Informatics.