Microarray Information and Data Integration Using SAMIDI

Juan M. Gómez
Universidad Carlos III de Madrid, Spain

Ricardo Colomo
Universidad Carlos III de Madrid, Spain

Marcos Ruano
Universidad Carlos III de Madrid, Spain

Ángel García
Universidad Carlos III de Madrid, Spain

INTRODUCTION

Technological advances in high-throughput techniques and efficient data gathering methods, coupled computational biology efforts, have resulted in a vast amount of life science data often available in distributed and heterogeneous repositories. These repositories contain information such as sequence and structure data, annotations for biological data, results of complex computations, genetic sequences and multiple bio-datasets. However, the heterogeneity of these data, have created a need for research in resource integration and platform independent processing of investigative queries, involving heterogeneous data sources.

When processing huge amounts of data, information integration is one of the most critical issues, because it's crucial to preserve the intrinsic semantics of all the merged data sources. This integration would allow the proper organization of data, fostering the analysis and access the information to accomplish critical tasks, such as the processing of micro-array data to study protein function and medical researches in making detailed studies of protein structures to facilitate drug design (Ignacimuthu, 2005). Furthermore, DNA micro-array research community urgently requires technology to allow up-to-date micro-array data information to be found, accessed and delivered in a secure framework (Sinnot, 2007).

Several research disciplines, such as Bioinformatics, where information integration is critical, could benefit from harnessing the potential of a new approach: the Semantic Web (SW). The SW term was coined by Berners-Lee, Hendler and Lassila (2001) to describe the evolution of a Web that consisted of largely documents for humans to read towards a new paradigm that included data and information for computers to manipulate. The SW is about adding machine-understandable and machine-processable metadata to Web resource through its key-enabling technology: ontologies (Fensel, 2002). Ontologies are a formal explicit and shared specification of a conceptualization. The SW was conceived as a way to solve the need for data integration on the Web.

This article expounds SAMIDI, a Semantics-based Architecture for Micro-array Information and Data Integration. The most remarkable innovation offered by SAMIDI is the use of semantics as a tool for leveraging different vocabularies and terminologies and foster integration. SAMIDI is composed of a methodology for the unification of heterogeneous data sources from the analysis of the requirements of the unified data set and a software architecture.

BACKGROUND

This section introduces Bioinformatics and its need to process massive amounts of data; the benefit of the integration of the existing data sources of biological information and semantics, a tool for integration.

Bioinformatics

The term Bioinformatics was coined by Hwa Lim in the late 1980s, and later popularized through its association with the human genome project (Goodman, 2002). Bioinformatics is the application of information science and technologies for the management of biological data (Denn & MacMullen, 2002) and it describes any use of computers to store, compare, retrieve, analyze or predict the composition of the structure of biomolecules (Segall & Zhang, 2006). Research on Biology requires Bioinformatics to manipulate and discover new biological knowledge at several levels of increasing complexity. Biological data are produced through high-throughput methods (Vyas & Summers, 2005), which means that they have to be represented and stored in different formats, such as micro-arrays.

Micro-Array Data Sources

A DNA micro-array is a collection of microscopic DNA spots attached to a solid surface forming an array for the purpose of expression profiling, which monitors expression levels for thousands of genes simultaneously. Those features are read by a scanner that measures the level of activation, and the data is downloaded onto a computer for subsequent analysis (Cohen, 2005). Micro-arrays allow investigating million of genes simultaneously (Segall & Zhang, 2006). A biological experiment may require hundreds of micro-array, where a single micro-array generates up to million fragments of data (Murphy, 2002). This fact makes data analysis and management a major challenge for gene expression studies using micro-arrays (Xu, Maresh, Giardina & Pincus, 2004).

The need to manage data generated from Bioinformatics is crucial. Understanding biological processes necessitates access to collections of potentially distributed, separately owned and managed biological data sets (Sinnott, 2007). These data sources reside in different storages, hardware platforms, data base management systems, data models and data languages (Chen, Promparmote & Maire, 2006), which makes impossible their integration. To make things worse, this incompatibility is not limited to the use of different data technologies, but also because of its incompatibility in terms of semantics. This heterogeneity can be of two sorts: *syntactic* and *semantic* (Verschelde, Dos Santos, Deray, Smith & Ceusters, 2004). Syntactic heterogeneity refers to differences in data models and data languages and can be easily resolved. Semantic heterogeneity refers to the underlying meanings of the data represented. It gives origin to naming conflicts and structural conflicts.

This incompatibility, and the necessity of sharing and aggregating information among the existing micro-array data sources leads researchers to seek for data integration.

Micro-Array Data Integration

Data analysis and management represent a major challenge for gene expression studies using micro-arrays (Xu, Maresh, Giardina & Pincus, 2004). Micro-array technology is still rather new and standards are not established (Murphy, 2002). This lack of standardization impedes micro-array data exchange. However, several projects have been started with a common goal: facilitate the exchange and analysis of micro-array data. MIAME (Minimum Information About Micro-array Experiment) is an XML based standard for the description of micro-array experiments. It's gaining importance because is required by numerous journals for the submission of articles providing micro-array experiments results. The purpose of MIAME is to define the core information needed for the description of an array based gene expression monitoring experiment. MAGE (Micro-Array Gene Expression) is a standard micro-array data model and exchange format that is able to capture information specified by MIAME.

Integration and Semantics

The ambiguity of terms, both within and between different databases and terminologies, makes integrating bioinformatics data task highly error prone (Verschelde *et al*, 2004). Converting all this information into a common data format will likely never be achieved and, therefore, the solution to the effective information management problem will necessarily go through the establishment of a common understanding. At this point is where semantics comes into play, bridging nomenclature and terminological inconsistencies to comprehend underlying meaning in a unified manner. The key elements that enable semantic interoperability are ontologies; semantic models of the data and they interweave human understanding of symbols with their machine processability (Della Valle, Cerizza, Bicer,