Chapter 7
A Data Warehousing Approach for Genomics Data Meta-Analysis

Martine Collard
INRIA Sophia Antipolis, France and University of Nice-Sophia Antipolis, France

Leila Kefi-Khelif
INRIA Sophia Antipolis, France

Van Trang Tran
I3S laboratory, University of Nice-Sophia Antipolis, France

Olivier Corby
INRIA Sophia Antipolis, France

ABSTRACT

DNA micro-array is a fastest-growing technology in molecular biology and bioinformatics. Based on series of microscopic spots of DNA sequences, they allow the measurement of gene expression in specific conditions at a whole genome scale. Micro-array experiments result in wide sets of expression data that are useful to the biologist to investigate various biological questions. Experimental micro-arrays data and sources of biological knowledge are now available on public repositories. As a consequence, comparative analyses involving several experiments become conceivable and hold potentially relevant knowledge. Nevertheless, the task of manually navigating and searching for similar tendencies in such huge spaces is mainly impracticable for the investigator and leads to limited results. In this context, the authors propose a semantic data warehousing solution based on semantic web technologies that allows to monitoring both the diversity and the volume of all related data.

INTRODUCTION

DNA micro-arrays are now widely used for mRNA expression profiling and have applications in a variety of biological issues. Numerous laboratories have collected micro-array data that are available on public databases and web sites. For instance, the Gene Express Omnibus[1] (GEO) or the ArrayExpress[2] repositories provide public availability of data on

DOI: 10.4018/978-1-60566-816-1.ch007

gene profiles for the entire scientific community. Thus, it has recently appeared useful for biologists to take advantage of these archives of responses for different purposes.

Despite the genome wide dimension of micro-arrays and their expression data, results published in scientific media generally focus on the hundred first differentially expressed genes among thousands of a whole genome, and real discussions are on ten of them only. So novel statistical analyses may be led on archived data in order to explore them more deeply, and confirm original results or discover new knowledge.

Another use of these public archives is for comparative analyses. New micro-array experimental data may be compared to previous ones in order to highlight similar and specific responses to a particular biological test.

A third use of these expression datasets is to involve multiple data sets in a new meta-analysis. In order to highlight similar and specific biological responses to a particular biological test, it seems promising to transversally analyze the largest set of related data. Combined analyses of multiple data sets and their issues have focused either on differential expressions (Rhodes et al., 2002, Choi et al., 2003) or on co-expressed genes (Eisen et al., 1998, Lee et al., 2004). (Hong and Breitling, 2008) evaluated three statistical methods for integrating different micro-array data sets and concluded that meta-analyses may be powerful but have to be led carefully.

Nevertheless, biologists that are interested in studying micro-array data and finding novel knowledge face a very complex task. Navigating manually into huge amounts of diverse data stored in these public repositories is such a tedious task that they finally lead restricted studies and make limited conclusions. Systems like GEO for the NCBI[3], ArrayExpress[4] for the EBI[5], Gemma[6] or Genepattern[7] allow investigators to share data and analyses results, they provide user-friendly tools allowing the analysis of global expression data, as collected by DNA micro-array experiments. There are still critical points, on one hand to combine directly data sets derived from different experimental processes and micro-arrays, and on another hand, to take benefit from the whole set of related information.

In this context, our approach is to enable meta-analyses involving multiple types of source data including aggregated or synthetic data and semantic aspects.

This chapter presents the semantic data warehousing approach AMI (*Analysis Memory for Immunosearch*) that we designed in order to facilitate storage and intelligent querying of:

- gene expression data from multiple experiments,
- refined data (aggregate or synthetic) resulting from statistical analyses and data mining methods,
- data and metadata representing all related information from the biological domain.

All these different kinds of information may be considered as dimensions of the semantic data warehouse. Refined data may be considered as facts in a standard data warehouse. One idea is to take advantage of semantic relationships among metadata for querying this data warehouse and provide relevant comparative analyses. Technical solutions in AMI knowledge base and search engine are based on semantic web techniques such as semantic annotation languages and underlying ontologies.

The work realized within the AMI project aims in a final stage at providing the scientist user with semi-automatic tools facilitating navigation and comparative analyses into a whole set of comparable experiments and multiple sources of information related to a particular bi- 3 Data warehousing approach for Genomics Data Analysis biological process. This work was done in collaboration with the Immunosearch company[8] whose projects focus on human biological responses to chemicals. In a first step, AMI is devoted to human skin biological reactions only.

Related Content