Valuation study of available genomic data storage platforms

Viti F (1), Merelli I (1), Porro I (2), Papadimitropoulos A (2), Milanesi L (1)

(1) Institute for Biomedical Technologies, National Research Council, Milan
(2) Department of Computer Science, Control Systems and Telecommunications,
University of Genoa, Genoa

Motivation

One of the most important problems in microarray research is the storage of large amounts of data, keeping track of the information about experiments, samples and projects in matter. Nowadays, there are three recognized public repositories for microarray experiments: Gene Expression Omnibus (GEO - NCBI), ArrayExpress (EBI) and the Center for Information Biology Gene Expression Database (CIBEX - DDBJ). These infrastructures can be used to deal with highthroughput experimental data in gene expression research, and they are all MIAME (Minimum Information About Microarray Experiment) compliant. Scientists who work with microarray data need more than an infrastructure for data sharing: the technique is difficult to handle not only because it produces thousands of data, but also because it requires many biological and technological steps that must be recorded. So they need a secure local storage system to manage and integrate broad genomic data, in order to freely insert, change and modify designs and protocols of their experiments, always following strict standards approved by all of the microarray community. In this context our aim is to test infrastructures the bioinformatics world proposes and to evaluate the most efficient genomic platform available at the moment. Considering that some important infrastructures (i.e. GEO and CIBEX) do not have a local downloadable version, we analysed ArrayExpress and compared it with another important platform, maybe less known but with great potentialities: GUS, the Genomics Unified Schema. It is an integrated databases system, developed by the enormous contribution of University of Pennsylvania, non completely avaliable as web service but on which schema of important on line databases like RAD (RNA Abundance Database) are based.

Methods

To make an evaluation of the capabilities of the data warehousing of these open source products we installed both of them in order to test their features in our microarray data and our experimental protocols. The two platforms could seem comparable, but analyzing them in detail some important differences emerge. GUS shows interesting aspects about user interface and scheme customization, which are not so rigid as in ArrayExpress. Moreover, its data integration in genomics, transcriptomics and proteomics fields presents good features for future bioinformatics studies. Concerning requirements, they are similar on various aspects: they both need a Unix Operative System; both suggested Oracle as RDBMS, even if, while ArrayExpress is a combination of two different databases (i.e. Oracle and MySQL) to separate data storage from query data warehouse, GUS uses a unique database. Loading proper data or information from external databases is different in the two systems. GUS perform this process by using and creating new Perl plug-ins (which generate Perl objects), while ArrayExpress provides the MAGEloader/MAGEvalidator, a java program which converts data into MAGE-ML (Microarray Gene Expression Markup Language), language derived from MAGE-OM (MicroArray Gene Expression Object Model) and similar to XML. MAGE-OM is, in fact, ArrayExpress fixed logic schema, while GUS as a proper schema that can be modified and also enlarged by single developers. Regarding data submission standards, both are MIAME-compliant, and GUS follows developed standards also for proteomics (MAIPE - Minimum Information About a Proteomics Experiment) and tissue gene expression localization (MISFISHIE - Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments) experiments. Both platforms can be deployed, through the Apache Tomcat servlet container, and visualized as web applications: an example of this possibility in GUS is RAD website, which makes accessible a part of the integrated databases system of GUS. Even if these two infrastructures seem to be quite similar, there are also important differences, particularly inherent database schema and UI structure. Query and data submission in ArrayExpress are pre-configured and cannot be modified, while GUS platform offers the possibility to manage, through a Web Development Kit, the user interface, with the aim of creating user-friendly applications and websites with advanced query capabilities. This is not a secondary aspect because bioinformatics databases are often handled by biologists who need a simple, linear schema to insert and make queries on information.

Results

In order to valuate a powerful platform to handle biologically consistent results, we noticed that GUS is highly customizable in structure and flexible in the user interface development. Moreover, ArrayExpress can store data from all microarray technologies and from array-based chromatin immunoprecipitation and array CGH, while GUS permits submission of different kinds of data, from genomics, to transcriptomics and proteomics, in an aim to improve and support broad genomics data integration.

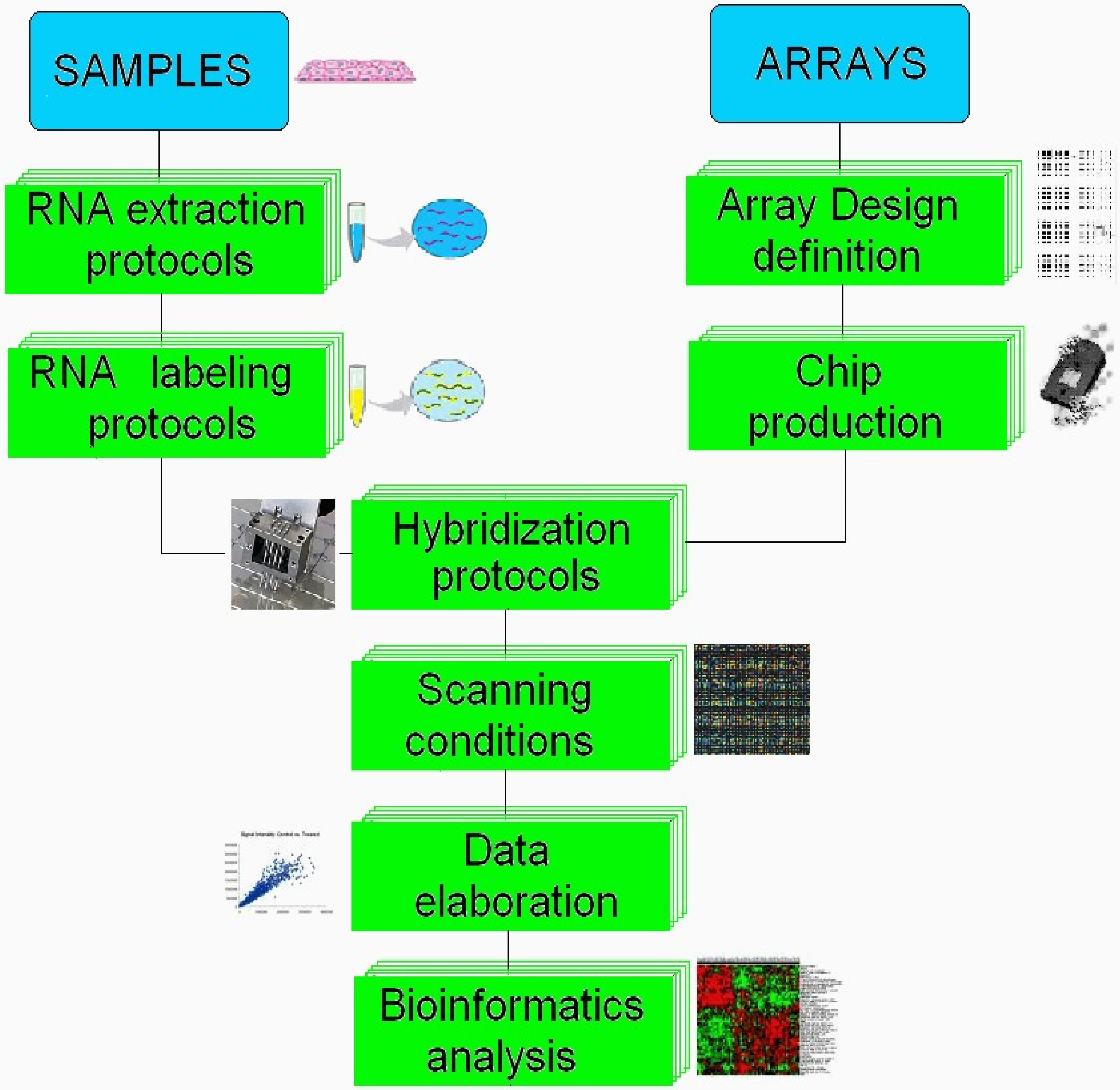
Availability: http://www.gusdb.org/

Contact email: federica.viti@itb.cnr.it

Supplementary informations

GUS is available at http://www.gusdb.org/
ArrayExpress is available at http://www.ebi.ac.uk/arrayexpress/
GEO is available at http://www.ncbi.nlm.nih.gov/geo/
CIBEX is available at http://cibex.nig.ac.jp/index.jsp

MICROARRAY EXPERIMENT COMPLEXITY



SAMPLES

RNA extraction protocols

RNA labeling protocols

ARRAYS

Array Design definition

Chip production

Hybridization protocols

Scanning conditions

Data elaboration

Bioinformatics analysis