A database infrastructure for microarray data

Emerson A. , Rossi E. , Giuliani S.

High Performance Systems Divsion, CINECA, Casalecchio di Reno.

Motivation

The usual procedure employed in the analysis of microarray data typically involves little more than a desktop workstation and a spreadsheet program. However, managing data in the form of spreadsheet files is not convenient, particularly when submitting results to public databases prior to publication. In addition, many researchers are discovering that data coming from the latest generation of microarray chips, which may contain many tens of thousands of gene probes, cannot be processed even with the most powerful personal computer. A further deficiency is that normally there is no provision for the systematic recording of information related to the experiment themselves, e.g. platform design, sample hybridization or protocols; such data are critical in checking reproducibility and for comparing with other experiments. Motivation for a more sophisticated and rigorous approach to microarray analysis data has come from researchers in the Hormone Responsive Breast Cancer Network (HRBC, http://www.hrbc-genomics.net/), where the analysis and sharing of microarray data with other members of the project, as well as comparison with relevant data in public repositories, are essential requirements

Methods

The most reliable and robust data storage systems usually involve a UNIX server and a relational database system, such as mysql or Oracle. The need to record the information about the experimental set-up has in some part already been addressed in the formulation of the MIAME (Minimum Information About Microarray Experiments) standard which lists the minimum information considered to be essential for a correct description of the experiment and "unambiguous interpretation of the results" [1]. The Microarray Gene Expression Data Society which designed MIAME has also proposed a file-format, MAGE-ML, which can be used to store "MIAME compliant" data [2]. We should mention that adoption of the MIAME standard is not universal, the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) database does not output MAGE-ML files for example, but in the absence of any other accepted standard we have decided to make MIAME compliance a feature of our database system. Thus, in order to design our database infrastructure we have performed a survey of the available microarray database packages which are UNIX-based, use a relational database for data storage and are MIAME compliant, with possible MAGE-ML support. We finally decided to orient our infrastructure around the BASE program [3] for the following reasons: 1. It adheres well to the MIAME standard. 2. It is open source and so can be modified if necessary. 3. It uses a simple web interface for manual insertion of experimental data. 4. It is possible to interface BASE to programs such as the R statistical package and the popular GeneSpring software. 5. It has the capability of outputting MAGE-ML files. One of the main tasks of the project was to archive public data from the GEO repository which lists many breast cancer related experiments. This provides data in the form of simple text files ("soft files") so a Perl converter or "wrapper" program was written in order to extract the experimental meta-data. Instead of inserting the data directly into BASE, the converter creates a MAGE-ML file and then another wrapper is used to read the MAGE-ML file in order to connect to the BASE database. In this way our system is not tied to BASE, since we can keep all the microarray data in the form of platform-independent MAGE-ML files.

Results

We have installed and implemented the Base microarray database system and made it available to researchers in the HRBC consortium. In order to provide access to publicly available data from the GEO database we have created wrapper programs capable of reading GEO’s soft files and converting them into the MAGE-ML format and programs capable of reading MAGE-ML and transferring the data into BASE. The batch conversion of a large number soft files relating to experiments of interest to the HRBC consortium is currently underway. Further conversion programs for other database systems are planned. References [1] A. Brazma et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature Genetics, vol 29 (December 2001), pp 365 - 37 [2] P. Spellman at al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biology, 3(9), (2002), research0046.1-0046.9 [3] L. H. Saal, C. Troein, J.Vallon-Christersson, S. Gruvberger, A. Borg and C. Peterson. BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data. Genome Biology 2002 3(8): software0003.1-0003.6

Contact email: a.emerson@cineca.it
URL: http://www.biogate.it/base