GEO IMPORT

Abstract:

In less than a decade, microarrays have grown in popularity as a widespread technology for the comprehensive analysis of gene expression. The use of microarrays by biological researchers has increased as the cost of the technology has fallen, and as the complexity of the arrays has increased the volume of data has grown. Microarray technology is used in diverse disciplines such as toxicology, gene expression, genotyping and phylogenetic analysis. There is a now a large volume of data in the public domain in databases, on websites and in speciality databases. However, in spite of the adoption of the MIAME (Minimum Information About a Microarray Experiment) standard by the major public databases this data is still not available from a single portal or in a consistent format with consistent annotation. So data integration between microarray data repositories has become a necessity.

GEO (Gene Expression Omnibus) at National Center for Biotechnology Information (NCBI) and ArrayExpress at European Bioinformatics Institute (EMBL-EBI) are the two biggest public MIAME compliant microarray data repositories for gene expression data.

GEO centralizes gene expression data and gives guarantees of existence of those data and thus makes it easily accessible publicly. GEO is more like a catalogue of datasets to the user. ArrayExpress shares the same philosophy like GEO with some more enhanced and extended features. It archives the dataset as well as gene expression results.

Compared with GEO ArrayExpress is a well structured database and the data submission process is more controlled and standards based. ArrayExpress data is free of ambiguity and more distinct for future use. So integrating GEO data into ArrayExpress will enhance the scope of analysis for the users.

The aim of this project is to analyze data in GEO and design a set of rules how data can be mapped to ArrayExpress infrastructure, in particular the data warehouse (the query-oriented part of the ArrayExpress project which requires less structured metadata than the repository); to import some GEO data into ArrayExpress and work on partially automating this process.

The development process is custom-designed to suit the purpose of the project. The main goal of the program is generation of a Tab2MAGE spreadsheet from a SOFTtext file. It has been designed to work as a black box whose input is a SOFTtext file and some complementary files and output is a Tab2MAGE spreadsheet used for further processing and downloadable supplementary files. Further a tool is also being developed which converts Tab2MAGE spreadsheet to a new submission method called MAGE TAB which makes submission easier for the submitters.

Now we can only convert GEO data to Tab2MAGE spreadsheet which is later turned into MAGE-ML and imported to ArrayExpress repository and from there to data warehouse. A pipeline will be established which will integrate the developed import tools with present environment. Some batch processing will be introduced to convert the GEO data into spreadsheet and load it in the repository after turning it into MAGE-ML consecutively.