Function:Troilkatt/GEO

From FunctionWiki

(Redirected from Function:Troilkatt/Data Sources)

This article describes the **GEO** data source.

Contents

Downloaded Data

From GEO we *mirror*:

- All sample series (identified by GSExxx identifiers) in the SOFT file format. A series file is user submitted and summarizes an experiment. It contains a text description of the elements that may be detected and measured (platform description in GEO terminology), and a set of samples (identified by GSMxxx identifiers). The samples have a textual description of how the sample was handled in the experiment and the measurements for each element in the sample.
- All raw data for the sample series (identified by GSExxx_RAW identifiers).
- All datasets (identified by GSDxxx identifiers) in the SOFT file format. A dataset represents a curated collection of statistically and biologically comparable samples. The samples may be reassembled from multiple series but they are from the same platform and the values are similarly calculated and normalized.
- All *full* datasets (identified by GSDxxx_full identifiers). These differ from the datasets in that they contain additional up-to-date gene annotation information.

For GEO we *ignore* the GPLxxx files that contain samples and series organized by platform, since our analysis tools currently does not use this organization. We also ignore the SeriesMatrix files that contain values extracted from the series files, since our data processing pipeline does a similar extraction. All MINiML files are also ignored since we prefer the SOFT format. Finally, we ignore all supplementary files that contain the raw data output by the microarray machines since this data is of no interest to our data integration tools.

The mirroring is implemented by periodically retrieving a list of series and datasets files from the GEO FTP server and then comparing the list of files to a list of previously downloaded files. We are not able to detect updates to series (that may occur but are quite rare). In order to detect updates we can either use the GEO database query tools, or we could use the rsync server where fingerprints are compared to detect updates. The later cannot be used directly since we save the files in the Hadoop filesystem possibly with a different compression algorithm.

SOFT File Format

The series, dataset, and full dataset files are in the SOFT file format.

"Simple Omnibus Format in Text (SOFT) is designed for rapid batch submission (and download) of data. SOFT is a simple line-based, plain text format, meaning that SOFT files may be readily generated from common spreadsheet and database applications. A single SOFT file can hold both data tables and accompanying descriptive information for multiple, concatenated Platforms, Samples, and/or Series records." [1] (http://www.ncbi.nlm.nih.gov/geo/info/soft2.html) . Refer to ibid for additional details.

SOFT to PCL Conversion

The conversion of a SOFT file to a PCL file consists of finding the *sample table* in the SOFT file, identifying the columns with the gene names and the *normalized* expression values and writing these to the PCL file.

This step is done by the SeriesFamulyParser.java for series (GSEX.soft) files and convertSoft2Pcl.rb for dataset files (GSDX.soft).

Raw Data Format

The raw data contains platform specific data. For example, for the AffyMetrix the raw data includes the CEL files produced by the Microarray instrument.

The corresponding SOFT file for the dataset is created using the raw data as input. However, the normalization and other processing may be suboptimal such that the quality of the data can be improved by redoing the processing. This processing can be greatly simplified by using libraries provided by using libraries provided for the R statistical processing environment [2] (http://www.r-project.org/) . The following sections describe how this processing is done for the most important platforms.

AffyMetrix CEL to PCL Conversion

The AffyMetrix CEL to PCL conversion is implemented by an R script (troilkatt/scripts/R/ProcessCEL.R) that uses the affy and affyio libraries from Bioconductor [3] (http://www.bioconductor.org/) [4] (http://www.bioconductor.org/help/bioc-views/release/bioc/html/affyio.html) , and ENTREZG annotation files from BrainArray [5] (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/13.0.0/entrezg.asp) .

In order to use the script first install the Bioconductor libraries as described in [6] (http://www.bioconductor.org/help/bioc-views/release/bioc/html/affyio.html) , then download all "CDF Map Seq Map Desc" files from Brainarray [7] (http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/13.0.0/entrezg.asp) and unpack these into the R library directory, and then run the ProcessCEL.R script. This will do the normalization of the CEL files and create one output file for each platform type, provided that an BrainArray mapping exists for that platform.