BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 23 2008, pages 2798–2800
doi:10.1093/bioinformatics/btn520

Data and text mining

GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus

Yuelin Zhu[1,2], Sean Davis[1], Robert Stephens[2], Paul S. Meltzer[1] and Yidong Chen[1,*]

[1]Genetics Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892 and [2]Advanced Biomedical Computing Center, National Cancer Institute-Frederick/SAIC-Frederick Inc., Frederick, MD 21702, USA

Received on April 22, 2008; revised on August 7, 2008; accepted on October 3, 2008

Advance Access publication October 7, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

The NCBI Gene Expression Omnibus (GEO) represents the largest public repository of microarray data. However, finding data in GEO can be challenging. We have developed GEOmetadb in an attempt to make querying the GEO metadata both easier and more powerful. All GEO metadata records as well as the relationships between them are parsed and stored in a local MySQL database. A powerful, flexible web search interface with several convenient utilities provides query capabilities not available via NCBI tools. In addition, a Bioconductor package, GEOmetadb that utilizes a SQLite export of the entire GEOmetadb database is also available, rendering the entire GEO database accessible with full power of SQL-based queries from within R.

Availability: The web interface and SQLite databases available at http://gbnci.abcc.ncifcrf.gov/geo/. The Bioconductor package is available via the Bioconductor project. The corresponding MATLAB implementation is also available at the same website.

Contact: yidong@mail.nih.gov

1 INTRODUCTION

The NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) represents the largest public repository of microarray data in existence (Edgar *et al*., 2002; Barrett *et al.*, 2007). The Bioconductor project (Gentleman *et al.*, 2004) contains hundreds of state-of-the-art methods for the analysis of microarray and genomics data. Previously we published software, called GEOquery (Davis and Meltzer, 2007), which effectively establishes a bridge between GEO microarray data and Bioconductor and facilitates reanalysis using novel and rigorous statistical and bioinformatic tools. However, a difficulty that remains in dealing with GEO is to find, based on the experimental metadata, the microarray data that are of interest especially for large-scale and programmatic access of GEO data. As part of the NCBI Entrez search system, GEO can be searched online via web pages or using NCBI eUtils (http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html). However, the NCBI/GEO web search is not yet full featured, particularly for programmatic access. NCBI eUtils offers another option for finding data within the vast stores of GEO, but it is cumbersome to use, often requiring multiple complicated eUtils queries to get the relevant information. GEOmetadb was developed in an attempt to make querying the GEO metadata both easier and more effective. GEOmetadb includes a web-based query engine, supported by a MySQL database backend, with several convenient utilities and a Bioconductor package, called GEOmetadb, which queries a locally installed GEOmetadb SQLite database that we update regularly and supply for download; each can be used independently of the other.

2 RESULTS

2.1 GEO metadata parsing

GEO has an open, adaptable design that can handle variety and a minimum information about a microarray experiment (MIAME)-compliant (Brazma *et al.*, 2001) infrastructure that promotes fully annotated submissions. The basic record types in GEO include Platforms (GPL), Samples (GSM), Series (GSE) and DataSets (GDS), of which GDS records are assembled by GEO curators and others are supplied by submitters. Essentially, information in each GEO record can be divided into two parts, a metadata part and the data part. The information in metadata part is critical for finding GEO microarray data of interest. NCBI offers several different methods to access GEO records, which we utilize to capture all GEO metadata for different GEO data types accordingly. Hypertext preprocessor (PHP, http://www.php.net) functions were written to parse, extract, reformat, construct data elements and interact with a MySQL database (http://www.mysql.com/) for storage and querying. The PHP function for parsing GDS SOFT files was adopted from the EzArray software (Zhu *et al*., 2008). The GEOmetadb MySQL database was designed to store parsed GEO metadata and relationships between them (Fig. 1). All data in GEOmetadb are faithfully parsed from GEO and no attempt is made to curate, semantically recode, or otherwise clean up GEO data. All field names are also taken from GEO records except for minor changes to improve usability in SQL queries. Fields containing multiple records are generally stored as delimited text within the same record; this denormalization significantly reduces complexity and improves efficiency of queries. SQLite 3 database (http://www.sqlite.org/) is a widely used, cross-platform SQL database engine which is a self-contained, embeddable, serverless, transactional SQL database engine. The RSQLite package (James, 2008) includes an embedded SQLite database engine and can interact with any SQLite database; each database exists as a simple file, which is easily exchanged and is platform independent. An R script converts the

*To whom correspondence should be addressed.

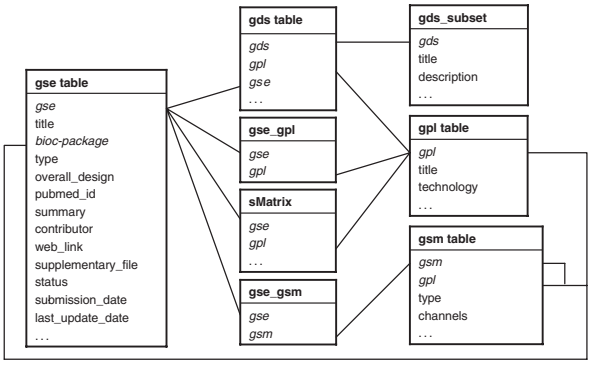

Fig. 1. Diagram of GEO entity relationships in GEOmetadb.

GEOmetadb MySQL database to an SQLite 3 database file that contains data identical to those in the GEOmetadb MySQL database. The SQLite version of GEOmetadb is maintained and distributed for local installation.

2.2 GEOmetadb bioconductor package

The GEOmetadb Bioconductor package is simply a thin wrapper around the GEOmetadb SQLite database. The package also includes extensive documentation and example queries. The function *getSQLiteFile* is the standard method for downloading and unzipping the most recent GEOmetadb SQLite file from the server. The function *geoConvert* performs conversion of one GEO entity to other associated GEO entities, providing a very fast, convenient mapping between GEO types. To convert 'GPL96' to other possible GEO entities in the GEOmetadb.sqlite:

```
> conversion <- geoConvert("GPL96")
> sapply(conversion, dim)
      gse gsm gds sMatrix
[1,] 532 15923 236 539
[2,]   2     2   2   2
```

The example provided below utilizes RSQLite function *dbGetQuery* to extract all affymetrix GeneChips that have .CEL supplementary submission to GEO.

```
> rs <- dbGetQuery(con, paste("select
+ gpl.bioc_package,gsm.gpl,",
+ "gsm,gsm.supplementary_file",
+ "from gsm join gpl on gsm.gpl=gpl.gpl",
+ "where gpl.manufacturer='Affymetrix'",
+ "and gsm.supplementary_file like
   '%CEL.gz'"))

> rs[1:3,]
  bioc_package   gpl    gsm
1      hu6800 GPL80 GSM575
2      hu6800 GPL80 GSM576
3      hu6800 GPL80 GSM577

supplementary_file
1 ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/
supplementary/samples/GSMNnnn/GSM575/GSM575.
cel.gz
```

Fig. 2. Screen-capture of GEOmetadb online search: combined GSE-GPL-GSM search.

```
2 ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/
supplementary/samples/GSMNnnn/GSM576/GSM576.
cel.gz
3 ftp://ftp.ncbi.nlm.nih.gov/pub/geo/DATA/
supplementary/samples/GSMNnnn/GSM577/GSM577.
cel.gz
```

2.3 The GEOmetadb online search tool

The GEOmetadb online search tool is a web-based search interface for searching, viewing and downloading GEO metadata stored in the GEOmetadb MySQL database. GEO metadata records can be searched by individual data type or by a flexible, efficient, powerful combined GSE-GPL-GSM search, as shown in Figure 2, where GEO entities in the tables are linked by relationships between them. Essential query fields are provided with drop-down menu for popular entries, and keyword search for full-text querying from multiple text fields in GEO. Other features include multiple field query, query within results, creating lists, flexible display options, downloading and detailed views of any results.

3 CONCLUSIONS

With the continued growth in the volume and complexity of microarray data available via NCBI GEO, it is critical that researchers have efficient, flexible, powerful methods for querying those data. While GEO offers several options for finding microarray data, GEOmetadb provides an alternative, yet much more flexible and efficient, set of tools for both online and programmatic access to GEO metadata. We expect that improved access to GEO metadata will not only enhance researchers' abilities to find data of interest, but also provide a possibility for users to create a customized GEO metadata database, e.g. annotating experiments with controlled vocabulary and integrating with other biological data sources.

ACKNOWLEDGEMENTS

We would like to thank BioInforx for the BxAF search functionality used on the web query pages and the NCBI GEO staff for valuable input and support during the development process.

Funding: Intramural Research Program of the NIH, National Cancer Institute, Bethesda, USA.

Conflict of Interest: none declared.

REFERENCES

Edgar,R. *et al*. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.

Barrett,T. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res*., **35**, D760–D765.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*., **29**, 365–371.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.1-R80.16.

Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and Bioconductor. *Bioinformatics*, **23**, 1846–1847.

Zhu,Y. *et al*. (2008) EzArray: a web-based highly automated Affymetrix expression array data management and analysis system. *BMC Bioinformatics*, **9**, 46–55.

James,D.A. (2008) RSQLite: SQLite interface for R. R package version 0.6-8, http://cran.r-project.org/web/packages/RSQLite/index.html.