GPLBrowse: An Interactive Platform Browser for NCBI GEO

Kay A. Robbins[1,*] and Cory Burkhardt[2]

[1]Department of Computer Science, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA.

ABSTRACT

Summary: The variability in normalization and labeling of data in NCBI GEO and other publicly available databases is a barrier to large-scale data mining of microarray data. Because platform-wide summary tools are not available, researchers must download and curate data without an overview of the available samples. GPLBrowse is an interactive web tool for investigating the characteristics of samples and series for microarray platforms in NCBI GEO. The first release of GPLBrowse provides statistical overviews of all samples and series in the top 18 oligonucleotide microarray platforms of GEO. GPLBrowse allows users to select and save samples, search based on keywords, and view additional metadata or link to GEO. GPLBrowse uses an AJAX-based split-client-server architecture to achieve highly responsive user interaction.

Availability: GPLBrowse is available at http://visual-charts.cs.utsa.edu/GPLBrowse.

Contact: krobbins@cs.utsa.edu.

1 INTRODUCTION

The emergence of large-scale microarray databases such as NCBI GEO (Barrett *et al*., 2007) raises the hope that data mining techniques can be applied to improve the reliability of *in silico* research and to discover new relationships among cell components. A number of recent studies have demonstrated the promise of these techniques (Hibbs *et al*., 2007; Pan *et al*., 2007a).

A recent snapshot of NCBI GEO shows 177,413 microarray samples (GSM) from 3,965 platforms (GPL) organized into 7,005 series (GSE). A platform is a particular type of microarray chip. A sample refers to a set of microarray expression measurements made from a single chip. Samples are grouped into series representing a single experiment. Samples may appear in more than one series, and a series may contain samples from different platforms.

Differences in platform technology, normalization techniques, and metadata labeling of the deposited data are barriers to large-scale data mining. NCBI GEO, for example, does not require that raw data be deposited. Thus, data miners do not have the option of applying a standard normalization scheme to all data for comparison. GEO also puts few requirements on sample labeling.

To overcome these difficulties, NCBI has engaged in a manual curation effort to better document coherent groups of samples within individual series. Datasets consist of hand-curated samples extracted from a reference series and assigned a common set of control variables. NCBI GEO currently contains 2,085 datasets (GDS). NCBI provides visualization tools for evaluating relative expression as a function of control variables for these datasets.

Most web tools for microarray analysis, such as Gene Aging Nexus (Pan *et al*., 2007b) use hand-curated subsets, often augmented by integration with other information. A few tools such as GS-LAGE (Yoon, *et al*., 2006) have recently emerged for platform-wide analysis. GS-LAGE uses datasets rather than series and eliminates samples as outliers if the mean gene expression is three times the platform mean. GS-LAGE assumes that the resulting expression distribution for each gene is normal and eliminates genes whose expression does not follow a normal distribution. GS-LAGE also assumes that input samples have been log-transformed.

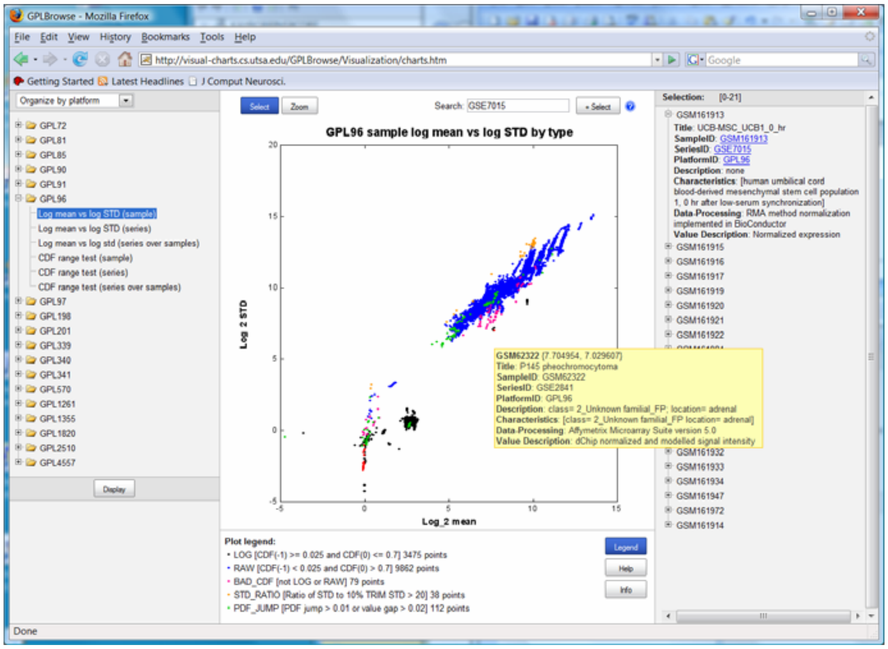
GS-LAGE server provides a motivation and an illustration for why platform-wide access to sample and series characteristics is important and useful. NCBI GEO does not provide information about whether or not a sample has been log-transformed. Our platform-wide studies have shown that most platforms have a mix of raw samples and samples that have undergone a log-like normalization. To by-pass this problem, some researchers calculate statistical characteristics such as correlation within series and then apply meta-analysis for platform-wide studies [Hibbs *et al*., 2006; Huttenhower *et al*., 2007]. However, even within the same series sample distributions can vary considerably and series-wide characterizations may not be valid without hand-curation.

The goal of GPLBrowse is to provide a highly-interactive web-based application that allows users to examine the characteristics of samples and series across a platform. By offering a visual, platform-wide context, GPLBrowse allows users to better understand the properties of selected samples in relation to the available samples as well as how these samples relate to their respective series.

2 FEATURES

Fig. 1 shows a sample scatter-plot for platform GPL96, the most popular platform in NCBI GEO. The 13,566 samples and 469 series have been individually normalized to have zero mean and unit standard deviation (STD). The figure plots $\log_2$(mean) versus $\log_2$(STD) for samples. The points are independently color-coded to indicate a distribution type based on the form of their cumulative probability distributions (CDFs). Blue points indicate raw data, while black points indicate $\log_2$ normalized data. Other colors specify outliers with various characteristics.

[*]To whom correspondence should be addressed.


GPLBrowse - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://visual-charts.cs.utsa.edu/GPLBrowse/Visualization/charts.htm
Google
Getting Started
Latest Headlines
J Comput Neurosci.
Organize by platform
GPL72
GPL81
GPL85
GPL90
GPL91
GPL96
Log mean vs log STD (sample)
Log mean vs log STD (series)
Log mean vs log std (series over samples)
CDF range test (sample)
CDF range test (series)
CDF range test (series over samples)
GPL97
GPL198
GPL201
GPL339
GPL340
GPL341
GPL570
GPL1261
GPL1355
GPL1820
GPL2510
GPL4557
Display
Select
Zoom
+Select
GPL96 sample log mean vs log STD by type
Log 2 STD
Log_2 mean
20
15
10
5
0
-5
-5
0
5
10
15
Legend
Help
Info
Done

Search: GSE7015

Selection: [0-21]

GSM161913
Title: UCB-MSC_UCB1_0_hr
SampleID: GSM161913
SeriesID: GSE7015
PlatformID: GPL96
Description: none
Characteristics: [human umbilical cord blood-derived mesenchymal stem cell population 1, 0 hr after low-serum synchronization]
Data Processing: RMA method normalization implemented in BioConductor
Value Description: Normalized expression

GSM161915
GSM161916
GSM161917
GSM161919
GSM161920
GSM161921
GSM161922
GSM161932
GSM161933
GSM161934
GSM161947
GSM161972
GSM161914

GSM62322 [7.704954, 7.029607]
Title: P145 pheochromocytoma
SampleID: GSM62322
SeriesID: GSE2841
PlatformID: GPL96
Description: class= 2_Unknown familial_FP; location= adrenal
Characteristics: [class= 2_Unknown familial_FP location= adrenal]
Data Processing: Affymetrix Microarray Suite version 5.0
Value Description: dChip normalized and modelled signal intensity

Plot legend:
- LOG [CDF(-1) >= 0.025 and CDF(0) <= 0.7] 3475 points
- RAW [CDF(-1) < 0.025 and CDF(0) > 0.7] 9862 points
- BAD_CDF [not LOG or RAW] 79 points
- STD_RATIO [Ratio of STD to 10% TRIM STD > 20] 38 points
- PDF_JUMP [PDF jump > 0.01 or value gap > 0.02] 112 points

Fig. 1: GPLBrowse displays $\log_2$(mean) vs $\log_2$(STD) for the samples for top-ranked platform GPL96. Blue points correspond to raw data; black points correspond to $\log_2$ normalized data; orange points have STDs that are 20 times greater than their 10% trimmed STDs; green points have large jumps in their probability distributions; and pink points have abnormally shaped probability distributions. Red points have been selected using GSE7015 in the keyword search. A tooltip for GSM62322 also appears.

GPLBrowse uses a three-column layout for its main display: the first column lists the available visualizations, the second column displays the selected scatter-plot, and the third column shows metadata for user-selected points. The scatter-plots can display multiple datasets using different shapes and colors.

In *Select* mode, users select points by dragging a box around the points of interest. When the user releases the mouse, GPLBrowse selects the boxed points and displays their metadata in the third column for additional browsing. Users can save selected points or load previously selected points by right-clicking on the plot area and choosing from options on a pop-up menu. Selected points can also be cleared or partially cleared. In *Zoom* mode, users drag a box to designate the zoomed area. GPLBrowse rescales the display. GPLBrowse also supports tooltips. If the user rests the mouse for 0.5 seconds over a data point, a tooltip displays the metadata corresponding to that point.

Users can search by typing keywords in the Search box on the upper right. Data satisfying the search criteria are immediately highlighted. Users can add these points to the current selection by pressing the +Select button to the right of the Search. GPLBrowse bases its search on all available content words in the sample and series metadata, not on a predefined list of keywords. For example, to find the samples associated with a particular series, the user just types the series name (e.g., GSE7015). GPLBrowse immediately highlights those samples. The search handles standard search logic with operators such as AND and OR.

3 IMPLEMENTATION

GPLBrowse uses a split-client-server architecture based on AJAX (asynchronous JavaScript and XML) technology to achieve user-interactivity close to that of desktop applications. The server responds to a selection of visualization by sending an image of the plot. The web browser draws the axes and maintains the locations of the points for displaying tooltips and for performing selections. GPLBrowse uses the AJAX-enabled YUI (Yahoo User Interface) Toolkit for its widgets and Apache Lucene for search. All menus and plot information are generated from XML files to allow additional visualizations to be added without programmatic change to GPLBrowse. The server side is implemented using Java Servelets. All code is standards-based and runs on the latest versions of Internet Explorer and Mozilla Firefox.

ACKNOWLEDGEMENTS

This work was supported by NIH Research Centers in Minority Institutions 2G12RR1364-06A1 and NIH-NSF CRCNS EIA-0217884.

REFERENCES

Barrett,T. Troup,D.B. Wilhite,S.E. Ledoux,P. Rudney,D. Evangelista,C. Kim,I.F. Soboleva,A. Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression. *Nucleic Acids Research*, **35**, D760–D765.

Hibbs,M.A. Hess,D.C. Myers,C.L. Huttenhower,C. Li,K. Troyanskaya,O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23(20)**, 2692–2699.

Huttenhower,C. Hibbs,M. Myers,C. Troyanskaya,O.G: (2006) A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, **22**, 2890–2897.

Pan,F. Kamath,K. Zhang,K. Pulapural,S. Achar,A. Nunez-Iglesias,J. Huang,Y. Yan,X. Han,J. Hu,H. Xu,M. Hu,J. Zhou,X.J. (2007a) Integrative Array Analyzer: a software package for analysis of cross-platform and cross-species microarray data. *Bioinformatics*, **22**(3), 1665–1667.

Pan,F. Chiu,C.-H. Pulapura,S. Mehan,M.R. Nunez-Iglesias,J. Zhang,K. Kamath,K. Waterman,M.S. Finch,C.E. Zhou,X.J. (2007b) Gene Aging Nexus: a web database and data mining platform for microarray data on aging. *Nucleic Acids Research*, **35**, D756–D759.

Yoon,S. Yang,Y. Choi,J. Seong,J. (2006) Large scale data mining approach for gene-specific standardization of microarray gene expression data. *Bioinformatics*, **22(23)**, 2898–2904.